Graduate Econometrics Review

Paul J. Healy
pj@pjhealy.com

April 13, 2005

Contents

Preface

This document was created with help from fellow graduate students Ben Klemens, Ming Hsu, Brian Rogers, and Isa Hafalir. Nearly all of the knowledge contained within is not my own. Instead, I used a wide variety of valuable resources which I hope to completely list in the bibliography.

At this point, this is not a complete document. The word "blah" is inserted as a flag that much more could be written in a particular section. Eventually, this document will be completed and the "blah"'s will be replaced with meaningful text.

Introduction

This document was created as a study guide for first year students in Social Sciences at Caltech to help prepare for the infamous Econometrics Preliminary Examination. It is a summary of the major topics learned throughout the first year of graduate study. The typical student will have a variety of texts and other resources at their disposal, and this is only meant to be a compact supplement to those more in-depth sources. This document is by no means a substitute for those texts.

Many proofs are given, though the reader should make an effort to prove theorems on their own before reading the proof given. Most of the proofs contained within are not terribly difficult and make for good test questions.

So find a comfortable chair, relax, and enjoy your review.

Part I

Probability

Chapter 1

Probability Theory

1.1 Counting

We begin with a brief but important discussion on the counting of events. Although a formal definition of an *event* will be given later, it is sufficient to think of an event as a string of outcomes. For example, if we flip three coins, the event "heads, heads, tails" may be observed. A convenient label for this event is "HHT.".

1.1.1 Ordering & Replacement

When couting events, it is crucial to understand whether we are sampling with or without replacement and whether or not ordering of outcomes matters. This is best understood through an example.

Given 5 cans of paint (R,B,Y,O,G), how many ways can you paint 3 boxes? In general notation, $n = 5,\ k = 3$. Your job is to assign paint to each box. If you sample paint colors with replacement, then a color may appear on two different boxes. If you sample colors without replacement, then each color can be assigned to at most one box - once you've used a color, you can't put it back into the set of colors from which you can choose for the next box. On the second dimension, if you assume that the ordering of your boxes matters, then painting Red, Blue, and then Green is *different* than painting Blue, Green, and then Red. One way to think of this is that the boxes are somehow distinguishable (perhaps by size) and painting the largest one red is a distinct event from painting the smallest one red (all else constant, or "ceteris peribus.")

An important way to clarify these problems so that you can correctly solve them is by identifying which set of items is getting assigned to the other set. Are you assigning paint to boxes, or are you assigning boxes to paint? Both make intuitive sense, but to answer this, use the following logic. Imagine one of your sets of elements as being a set of buckets, or bins, while the other set is comprised of balls. You want to assign balls to buckets and let n be the number of buckets and k be the number of balls. It makes sense to assign multiple balls to a single bucket, but it's crazy to assign multiple buckets to a single ball. For our painting example, it makes sense to assign multiple boxes to a single color of paint, but it's not allowable to assign multiple colors to a single box. Therefore, boxes are our ”balls,” k is 3, paints are our ”buckets,” and n is 5.

	With Replacement	Without Replacement
Order Matters	RRB≠BRR $5^3 = n^k$	RBG≠BGR $\frac{5!}{2!} = \frac{n!}{(n-k)!}$
Order Doesn't Matter	RRB=BRR $\binom{5+3-1}{3} = \binom{n+k-1}{k}$. ”arrange (n+k-1) interior walls & k balls”	RBG=BGR $\binom{5}{3} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$

Example 1.1 *If you receive exactly 12 calls per week, but they are randomly distributed among the 7 days with equal probability, how many ways can you have at least one call every day this week? How many total arrangements of calls are there possible in a week?*

The first question is to ask whether calls are assigned to days or days are assigned to calls. It makes sense to have multiple calls in a day, but not vice versa. So, calls are like balls and days are like buckets into which we assign the calls. Therefore, we assign n as our number of days (buckets) and k as our number of calls (balls.)

The second question is whether we have replacement and ordering. Since we'll allow multiple calls on any given day, we are choosing our days (buckets) with replacement. We weren't given any explicit information about calls being distinct, so it's safe to assume that the ordering of the calls does not matter. We could claim that the calls are in fact distinct and solve the problem correctly using that assumption, but we proceed assuming that order doesn't matter.

The second question is easier. We know that $n = 7$ *and* $k = 12$, *so our answer is* $\binom{7+12-1}{12} = \binom{18}{12}$.

To answer the first question, assume that one call has been distributed already to each of the 7 days. Therefore, we have 7 days and 5 calls remaining. So, $n = 7$ *and* $k = 5$. *Our answer is* $\binom{7+5-1}{5} = \binom{11}{5}$.

1.2 The Probability Space

The goal of developing a probability space is to take the natural set of possible outcomes from an experiment and develop it into a framework in which we are able to use the tools of mathematics to apply a method of measuring probabilities of collections of outcomes.

1.2.1 Set-Theoretic Tools

Probability theory uses the concepts of the union ($\cup$), the intersection ($\cap$), and the compliment $\left(A^C\right)$ quite often. Make sure you know how to manipulate these standard operators. DeMorgan's Laws define two ways to manipulate the compliment operator.

Theorem 1.2 *(DeMorgan's Laws)*

1. $(A \cup B)^C = A^C \cap B^C$
2. $(A \cap B)^C = A^C \cup B^C$

Make sure you are comfortable with basic set theory and its intuition. It is very useful in understanding the problems you'll be asked to solve. We will discuss some of these topics in a bit more detail in Section 1.2.4. The following are useful theorems and definitions as well.

Theorem 1.3 *For any sets A, B, and C, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$*

Proof. *Recall that $(B \cup C) = \{x : x \in B \text{ or } x \in C\}$ and $(A \cap B) = \{x : x \in A \text{ and } x \in B\}$*

$$A \cap (B \cup C) = \{x : x \in A \text{ and } (x \in B \text{ or } x \in C)\}$$
$$= \{x : (x \in A \text{ and } x \in B) \text{ or } (x \in A \text{ and } x \in C)\}$$
$$= (A \cap B) \cup (A \cap C)$$

■

Theorem 1.4 *(General set theory) For any sets A and B, $(A \cup B) \cup (A \cap B) = A \cup B$*

Proof. *Using the definitions of $\cap$ and $\cup$ from Theorem 1.3, we have that*

$$(A \cup B) \cup (A \cap B) = \{x : (x \in A \text{ or } x \in B) \text{ or } (x \in A \text{ and } x \in B)\}$$
$$= \{x : x \in A \text{ or } x \in B\} = A \cup B$$

This is because any x in both A and B must also be in A or B. Therefore, the second condition can be removed as it is satisfied by the first. ■

Definition 1.1 *Sets A and B are **mutually exclusive** if $A \cap B = \emptyset$*

Definition 1.2 *A **partition** of a set B is a countable (possibly infinite) set of sets $\{A_1, A_2, ...\}$ such that $A_i \cap A_j = \emptyset \; \forall i, j$ and $A_1 \cup A_2 \cup ... = B$.*

Definition 1.3 *The **compliment** of a set A_i is defined as $A_i^C = \{A_1 \cup A_2 \cup ... \cup A_{i-1} \cup A_{i+1} \cup ...\} = X \setminus A_i$, where $\{A_1, A_2, ..., A_{i-1}, A_i, A_{i+1}, ...\}$ form a partition of X.*

1.2.2 Sigma-algebras

Note that this section is dense and very mathematical. The typical user of econometrics needs only a superficial understanding of these concepts. The goal is to formally develop those concepts needed to define our probability space, which culminates in the results of Theorem 1.20. A reader not interested in developing these tools should focus only on those definitions marked with a double asterix (**) and then proceed to subsection 1.2.4.

We begin with a large collection of sets, X. From this set, we define families of subsets of X that have desirable properties. Eventually, we develop a way to measure elements from those families of subsets. This measure provides us with a way to assign probabilities to abstract outcomes.

Definition 1.4 *A collection of sets $\mathcal{A}$ is called a **semiring** if it satisfies the following properties*

1. $\emptyset \in \mathcal{A}$
2. $A, B \in \mathcal{A} \Longrightarrow A \cap B \in \mathcal{A}$
3. $A, B \in \mathcal{A} \Longrightarrow$ *there exists a collection of sets* $C_1, C_2, ..., C_n \in \mathcal{A}$ *such that* $A \setminus B = \cup_{i=1}^{n} C_i$.

This definition is not very intuitive. Fortunately, we soon develop Theorem 1.7 that enables us to ignore the technical properties of semirings.

Definition 1.5 *A collection of sets is called an **algebra** if it contains the empty set and is closed under compliments and unions. If*

1. $A \in \mathcal{A} \Longrightarrow A^C \in \mathcal{A}$ *(or, "$\mathcal{A}$ is closed under complimentation")*
2. $(A \in \mathcal{A}$ & $B \in \mathcal{A}) \Longrightarrow (A \cup B) \in \mathcal{A}$ *(or, "$\mathcal{A}$ is closed under finite unions")*

then $\mathcal{A}$ is an algebra.

Remark 1.5 *Algebras are also called "fields."*

Definition 1.6 *A collection $\mathcal{E}$ of sets is called a **σ-algebra** if it satisfies*

1. *If $A \in \mathcal{E}$,then $A^C \in \mathcal{E}$ (closed under complementation)*
2. *If $A_1, A_2, ... \in \mathcal{E}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{E}$ (closed under countable unions)*

Remark 1.6 *DeMorgan's laws together with properties 1 and 2 above imply that σ-algebras are also closed under countable intersections. Similarly, this implies that algebras are closed under finite intersections.*

Note that the set $\mathcal{E} = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a,b\}, \{a,c\}, \{b,c\}, \{a,b,c\}\}$ is an algebra and a σ-algebra.

Sigma algebras are a subset of algebras in the sense that all σ-algebras are algebras, but not vice versa. Algebras only require that they be closed under pairwise unions while σ-algebras must be closed under countably infinite unions. Consider the following collection of sets:

$$X = \mathbb{N}, \text{ the natural numbers} \tag{1.1}$$

$$\text{Let } n \text{ represent any single number in } \mathbb{N} \tag{1.2}$$

$$\mathcal{A} = \left\{ \begin{array}{c} \{n_i\}_{i=1}^{\infty}, \{\mathbb{N}\backslash n_i\}_{i=1}^{\infty}, \\ \left\{\cup_{i=1}^{k} n_i : 0 \leq k < \infty\right\}, \\ \left\{\mathbb{N}\backslash \cup_{i=1}^{k} n_i : 0 \leq k < \infty\right\} \end{array} \right\} \tag{1.3}$$

In words, this collection $\mathcal{A}$ contains all single numbers n_i, all compliments of single numbers ($\mathbb{N}\backslash n_i$), all finite unions of single numbers, and the compliment of each finite union of single numbers. Note that the union of 0 numbers is the empty set and that the unions and compliments of every element in the collection is also in the collection (verify this yourself.) Therefore, this collection is an algebra. However, if we take the countable union of all single numbers n_i, we get the entire set of natural numbers $\mathbb{N}$. Since $\mathbb{N}$ is not in this collection, the collection is not a σ-algebra. This famous example is known as the **finite-cofinite algebra**. This leads into the following theorem, which is left unproven.

Theorem 1.7 *All σ-algebras are algebras, and all algebras are semi-rings.*

Therefore, if we require a set to be a semiring, it is sufficient to show instead that it is a σ-algebra or algebra. This will be useful later.

Sigma algebras can be generated from abritrary sets. This will be useful in developing the probability space.

Theorem 1.8 *For some set X, the intersection of all σ-algebras $\mathcal{A}_i$ containing X (meaning that $x \in X \implies x \in \mathcal{A}_i \ \forall i$) is itself a σ-algebra, denoted $\sigma(X)$. This is called the **σ-algebra generated by** X.*

Proof. *(Intuitive, not formal)*

Take any element $x \in X$. For any arbitrarily chosen σ-algebra $\mathcal{A}_i$ such that $X \subseteq \mathcal{A}_i$, we know that $x \in \mathcal{A}_i$. Since $\mathcal{A}_i$ is closed under compliments, we know that $x^C \in \mathcal{A}_i$. For any $x, y \in X$, we know that $x, y \in \mathcal{A}_i$ as well. Therefore, $\{x\} \cup \{y\} \in \mathcal{A}_i \ \forall x, y \in X$.

Note that for each element, we have its compliment in $\mathcal{A}_i$ as well as the union of all elements, the unions of the compliments, the compliments of all unions of two elements, and so on. Clearly, we can take the elements of X and extend the set to include those elements that are needed to satisfy the properties of a σ-algebra. In fact, we do so without knowing anything about the σ-algebra in which these elements belong. We only know that the σ-algebra is going to *include them because of the properties of σ-algebras. Therefore, this "extended set" we've developed must be in* ***every*** *σ-algebra including X.*

Furthermore, notice that all σ-algebras contain $\emptyset$, so their intersection also contains $\emptyset$.

In conclusion, we know that every σ-algebra will contain both the "extended set" we generated from the elements of X as well as $\emptyset$. By construction, this extended set and $\emptyset$ satisfy the properties of a σ-algebra. ■

In our probability space, we begin with our sample space.

Definition 1.7 *(**) The* ***sample space*** *Ω is the set of all possible unique outcomes of the experiment at hand.*

If we were tossing a coin, $\Omega = \{Heads, Tails\}$.

In the probability space, the σ-algebra we use is $\sigma(\Omega)$, the σ-algebra generated by Ω. So, as in the above proof, take the elements of Ω and generate the "extended set" consisting of all unions, compliments, compliments of unions, unions of compliments, etc. Include $\emptyset$ with this "extended set" and the result is $\sigma(\Omega)$, which we denote as Σ.

Definition 1.8 *(**) The σ-algebra generated by Ω, denoted Σ, is the collection of possible* ***events*** *from the experiment at hand.*

If the experiment had a sample space of $\Omega = \{10, 15, 20\}$, then
$\Sigma = \{\emptyset, \{10\}, \{15\}, \{20\}, \{10, 15\}, \{10, 20\}, \{15, 20\}, \{10, 15, 20\}\}$.

Each of the elements of Σ is an event. Events can be thought of as descriptions of experiment outcomes. For example, the description "less than 18" is represented by the event $\{10, 15\}$. Similarly, the description "any outcome happens" is given by $\{10, 15, 20\}$. Note that $\emptyset$ is included to complete the definition of a σ-algebra, but can be thought of as the event "nothing happens." We will see that $\emptyset$ is usally assigned a probability of 0.

Note that σ-algebras can be defined over the real line as well as over abstract sets. To develop this notion, we first need to develop the concept of a topology.

Definition 1.9 *A* ***topology*** *τ on a set X is a collection of subsets of X satisfying*

1. *$\emptyset, X \in \tau$*
2. *τ is closed under finite intersections*
3. *τ is closed under arbitary unions*

Definition 1.10 *Any element of a topology is known as an* ***open set****.*

Definition 1.11 *The **Borel σ-algebra** (or, Borel field,) denoted $\mathcal{B}$, of the topological space (X, τ) is the σ-algebra generated by the family τ of open sets. Its elements are called **Borel sets**.*

Lemma 1.9 *Let $\mathcal{C} = \{(a, b) : a < b\}$. Then $\sigma(\mathcal{C}) = \mathcal{B}_{\mathbb{R}}$ is the Borel field generated by the family of all open intervals $\mathcal{C}$.*

What do elements of $\mathcal{B}_{\mathbb{R}}$ look like? Take all possible open intervals. Take their compliments. Take arbitrary unions. Don't forget to include $\emptyset$ and $\mathbb{R}$. Clearly, $\mathcal{B}_{\mathbb{R}}$ contains a wide range of intervals including open, closed, and half-open intervals. It also contains disjoint intervals such as $(2, 7] \cup (19, 32)$. Roughly speaking, it contains (nearly) every possible collection of intervals that are easily imagined.

1.2.3 Measure Spaces & Probability Spaces

We now develop the concepts needed to measure the "size" of our experiment outcomes.

Definition 1.12 *A pair (X, Σ) is a **measurable space** if X is a set and Σ is a nonempty σ-algebra of subsets of X.*

By defining a measurable space, we guarantee that we can define a function that assigns real-numbered values to the abstract elements of Σ.

Definition 1.13 *A set function $\mu : \mathcal{S} \longrightarrow [0, \infty)$ is a **measure** if:*

1. *$\mathcal{S}$ is a semiring*

2. ***(σ-additivity)** For any sequence of pairwise disjoint sets $\{A_n\} \in \mathcal{S}$ such that $\cup_{n=1}^{\infty} A_n \in \mathcal{S}$, we have $\mu\left(\cup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$*

3. *μ assumes at most one of the values $-\infty$ and ∞*

4. *$\mu(\emptyset) = 0$*

Definition 1.14 *A triplet (X, Σ, μ) is a **measure space** if (X, Σ) is a measurable space and $\mu : \Sigma \longrightarrow [0, \infty)$ is a measure.*

Definition 1.15 *A measure space is a **probability space** if $\mu(X) = 1$. In this case, μ is a **probability measure**.*

1.2.4 The Probability Measure *P*

We now develop our standard probability measure $\mathbb{P}$. Andrei Kolmogorov's axioms define a nice list of properties for a probability measure, and they are the standard in use by most statisticians.[1]. Let $\mathbb{P} : \mathcal{E} \longrightarrow [0,1]$ be our probability measure and $\mathcal{E}$ be some sigma-algebra of events generated by X.

Axiom 1.16 $\mathbb{P}[A] \leq 1 \; \forall A \in \mathcal{E}$

Axiom 1.17 $\mathbb{P}[X] = 1$

Axiom 1.18 $\mathbb{P}[A_1 \cup A_2 \cup ... \cup A_n] = \mathbb{P}[A_1] + \mathbb{P}[A_2] + ... + \mathbb{P}[A_n]$, *where* $\{A_1, A_2, ..., A_n\}$ *are disjoint sets in* $\mathcal{E}$.

These three basic axioms imply the following

Theorem 1.10 $\mathbb{P}[A_i^C] = 1 - \mathbb{P}[A_i]$

Proof. *Using the definition of a compliment,*

$$\begin{aligned}\mathbb{P}[A_i^C] &= \mathbb{P}[A_1 \cup A_2 \cup ... \cup A_{i-1} \cup A_{i+1} \cup ...] \\ &= \mathbb{P}[A_1] + \mathbb{P}[A_2] + ... + \mathbb{P}[A_{i-1}] + \mathbb{P}[A_{i+1}] + ... + \mathbb{P}[A_n] \\ &= \mathbb{P}[\mathcal{E}] - \mathbb{P}[A_i] = 1 - \mathbb{P}[A_i]\end{aligned}$$

■

Theorem 1.11 $\mathbb{P}[\emptyset] = 0$

Proof. *Simple.* $\emptyset = X^C$ *and* $\mathbb{P}[X^C] = 1 - \mathbb{P}[X] = 1 - 1 = 0$ ■

Theorem 1.12 $\mathbb{P}[A_i] \in [0,1]$

Proof. *We know that* $A_i^C \in \mathcal{E}$ *and that* $\mathbb{P}[A_i^C] = 1 - \mathbb{P}[A_i]$. *Rearranging gives* $\mathbb{P}[A_i] = 1 - \mathbb{P}[A_i^C]$. *Since* $\mathbb{P}[A_i^C] \geq 0$, *then* $\mathbb{P}[A_i] \leq 1$. *Since* $\mathbb{P}[A_i] \geq 0$, *we have that* $\mathbb{P}[A_i] \in [0,1]$ ■

Theorem 1.13 $\mathbb{P}[B \cap A^C] = \mathbb{P}[B] - \mathbb{P}[A \cap B]$

Proof. *Using Theorem 1.3,*

$$\begin{aligned}\mathbb{P}[B] = \mathbb{P}[B \cap X] = \mathbb{P}[B \cap (A \cup A^C)] &= \mathbb{P}[(B \cap A^C) \cup (A \cap B)] \\ &= \mathbb{P}[B \cap A^C] + \mathbb{P}[A \cap B]\end{aligned}$$

Simply subtract the $\mathbb{P}[A \cap B]$ *from both sides to get the desired result.* ■

Theorem 1.14 $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$

Proof. *Starting with Axiom 1.18 and using Theorem 1.4,*

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] \tag{1.4}$$

$$\mathbb{P}[(A \cup B) \cup (A \cap B)] = \mathbb{P}[A] + \mathbb{P}[B] \tag{1.5}$$

$$\mathbb{P}[A \cup B] + \mathbb{P}[A \cap B] = \mathbb{P}[A] + \mathbb{P}[B] \tag{1.6}$$

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] \tag{1.7}$$

■

[1] Different sources vary in what they define as Kolmogorov's axioms, but the versions not given here are at least implied by the three given.

Theorem 1.15 $A \subseteq B \Rightarrow \mathbb{P}[A] \leq \mathbb{P}[B]$

Proof. *From Theorem 1.13 we know that*

$$\mathbb{P}[B] = \mathbb{P}[B \cap A^C] + \mathbb{P}[A \cap B] \tag{1.8}$$

Since $A \subseteq B$, then $A \cap B = A$. Therefore,

$$\mathbb{P}[B] = \mathbb{P}[B \cap A^C] + \mathbb{P}[A] \tag{1.9}$$

$$\mathbb{P}[B] \geq \mathbb{P}[A] \tag{1.10}$$

■

Theorem 1.16 $A = B \Rightarrow \mathbb{P}[A] = \mathbb{P}[B]$

Proof. *If $A = B$, then $\forall\ x \in B$, it must be true that $x \in A$.*
Equivalently, $\nexists x : x \in B$ and $x \notin A$
Or, $\nexists x : x \in B$ and $x \in A^C$
Which implies that $B \cap A^C = \emptyset$
From Theorem 1.11 we know that if $B \cap A^C = \emptyset$, then $\mathbb{P}[B \cap A^C] = 0$.
Using the proof of Theorem 1.15 gives $\mathbb{P}[B] = \mathbb{P}[B \cap A^C] + \mathbb{P}[A]$
Since $\mathbb{P}[B \cap A^C] = 0$, we are done. ■

Theorem 1.17 $\mathbb{P}[A] = \sum_{i=1}^{\infty} \mathbb{P}[A \cap C_i]$*, where $\{C_1, C_2, ...\}$ forms a partition of $\mathcal{E}$.*

Proof. *Left as an exercise.* ■

Theorem 1.18 *If $\{A_1, A_2, ...\}$are pairwise disjoint sets in $\mathcal{E}$, then $\mathbb{P}[\cup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$*

Proof. *Since $\mathcal{E}$ is a σ-algebra, then $\cup_{i=1}^{\infty} A_i \in \mathcal{E}$, so $\mathbb{P}[\cup_{i=1}^{\infty} A_i]$ is well defined. Use Axiom 1.18 to show that the sum of the probabilities equals the probability of the union... but this is not complete. Axiom 1.18 needs to be extended to infinite unions... blah* ■

Theorem 1.19 *(Boole's Inequality, aka "Countable Subadditivity") $\mathbb{P}[\cup_{i=1}^{\infty} A_i] \leq \sum_{i=1}^{\infty} \mathbb{P}[A_i]$ for* any *set of sets $\{A_1, A_2, ...\}$*

Proof. *Also an exercise. Hint: $\{A_1, A_2, ...\}$ is not necessarily a partition. See Casella & Berger for a proof.* ■

1.2.5 The Probability Space $(\Omega, \Sigma, \mathbb{P})$

We now have all the tools required to establish that Ω, Σ , and $\mathbb{P}$ for a probability space.

Theorem 1.20 *Define Ω as the sample space of outcomes of an experiment, Σ as the σ-algebra of events generated from Ω, and $\mathbb{P} : \Sigma \longrightarrow [0, \infty)$ as a probability measure that assigns a nonnegative real number to each event in Σ. The space $(\Omega, \Sigma, \mathbb{P})$ satisfies the definition of a probability space.*

Proof. *Ω is a set and Σ is a nonempty σ-algebra of subsets of Ω. Therefore, (Ω, Σ) is a measurable space by Definition 1.12.*

Since Σ is a σ-algebra, it is also a semiring by Theorem 1.7. By Theorem 1.18, $\mathbb{P}$ satisfies σ-additivity. $\mathbb{P}$ never assumes values $-\infty$ and ∞. By Theorem 1.11, $\mathbb{P}[\emptyset] = 0$. These four properties imply that $\mathbb{P}$ is a measure by Definition 1.13.

Since (Ω, Σ) is a measurable space and $\mathbb{P}$ is a measure, then $(\Omega, \Sigma, \mathbb{P})$ is a measure space by Definition 1.14.

By Axiom 1.17, $\mathbb{P}[X] = 1$. Therefore, the measure space $(\Omega, \Sigma, \mathbb{P})$ is a probability measure. ■

In summary, the sample space is the list of all possible outcomes. Events are groupings of these outcomes. The σ-algebra Σ is the collection of all possible events. To each of these possible events (or, groupings of outcomes,) we assign some "size" using the probability measure using $\mathbb{P}$. An example will further clarify these concepts.

Example 1.21 *Consider the tossing of two fair coins. The sample space is $\{HH, HT, TH, TT\}$. One possible event would be "the coins have different sides showing," which is $\{HT, TH\}$. Another possible event is "at least one head," which is $\{HH, HT, TH\}$. Our sigma algebra generated from the sample space (defined in the next subsection) will be the collection of all possible such events:*

$$\left\{ \begin{array}{c} \emptyset, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \\ \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \\ \{HT, TH\}, \{HT, TT\}, \{TH, TT\}, \\ \{HH, HT, TH\}, \{HH, HT, TT\}, \\ \{HH, TH, TT\}, \{HT, TH, TT\}, \\ \{HH, HT, TH, TT\} \end{array} \right\}$$

For example, the event "first flip heads or second flip tails" is $\{HH, HT, TT\}$.

The probability measure $\mathbb{P}$ assigns a number from 0 to 1 to each of those events in the sigma algebra. If we had a fair coin, we would want to assign the following probabilities to the events in the above sigma algebra:

$\{0, 1/4, 1/4, 1/4, 1/4, 1/2, 1/2, 1/2, 1/2, 1/2, 1/2, 3/4, 3/4, 3/4, 3/4, 1\}$

However, we need not use those values. Perhaps an experimenter has reason to believe the coin is biased so that heads appears $3/4$ of the time. Then the following values for $\mathbb{P}$ would be appropriate:

$\{0, 9/16, 3/16, 3/16, 1/16, 3/4, 3/4, 5/8, 3/8, 1/4, 1/4, 15/16, 13/16, 13/16, 7/16, 1\}$

As long as the values of the probability measure are consistent with Kolmogorov's axioms and the consequences of those axioms, then we consider the probabilities to be mathematically acceptable, even if they aren't reasonable for the given experiment. This opens the door for philosophical comment on whether or not the probability values assigned can even be considered unreasonable as long as they're mathematically acceptable.

1.2.6 Random Variables & Induced Probability Measures

A random variable is a convenient way to express the elements of Ω as numbers rather than abstract elements of sets. We use random variables (which are really functions) to map the elements of Ω into the real number line. However, we must first touch on the concept of measurability of such functions.

Definition 1.19 *Let $\mathcal{A}_X, \mathcal{A}_Y$ be nonempty families of subsets of X and Y, respectively. A function $f : X \longrightarrow Y$ is $(\mathcal{A}_X, \mathcal{A}_Y)$-**measurable** if $f^{-1}(A) \in \mathcal{A}_X$ $\forall A \in \mathcal{A}_Y$.*

Remark 1.22 *If a function is $(\mathcal{A}_X, \mathcal{A}_Y)$-measurable, but $\mathcal{A}_X$ and $\mathcal{A}_Y$ are understood from context, we simply say the function is **measurable**.*

Definition 1.20 *A **random variable** X : is a measurable function from the probability space $(\Omega, \Sigma, \mathbb{P})$ into the probability space $(\mathcal{X}, \mathcal{A}_{\mathcal{X}}, \mathbb{P}_X)$, where Ω, Σ, and $\mathbb{P}$ are the sample space, sigma-algebra, and probability measure as defined above, $\mathcal{X} \subseteq \mathbb{R}$ is the range of X (which is a subset of the real number line,) $\mathcal{A}_{\mathcal{X}}$ is a Borel field of $\mathcal{X}$, and $\mathbb{P}_X$ is the probability measure on $\mathcal{X}$ induced by X. Specifically, $X : \Omega \longrightarrow \mathcal{X}$.*

Remark 1.23 *Always keep in mind that random variables, despite their name, are really functions. Also, remember the distinction between the probability measures and the random variable. Both are functions, but each performs a very different task.*

Definition 1.21 *The inverse of a random variable maps sets in $\mathcal{A}_{\mathcal{X}}$ back to sets in Σ. Specifically, $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$.*

Remark 1.24 *Random variables take* single elements *in Ω and map them to* single points *in $\mathbb{R}$. The inverse of random variables maps* sets *in $\mathcal{A}_{\mathcal{X}}$ back to* sets *in Σ.*

Definition 1.22 *A random variable is **discrete** if there exists a countable set $S \in \mathcal{X}$ such that $\mathbb{P}_X\left[S^C\right] = 0$. Otherwise, the random variable is **continuous**.*

The induced measure $\mathbb{P}_X$ is just a way of relating measure on the real line (the range of X) back to the original probability measure over the abstract events in the σ-algebra of the sample space. Specifically, $\mathbb{P}_X[A] = \mathbb{P}\left[X^{-1}(A)\right] = \mathbb{P}\{\omega \in \Omega : X(\omega) \in A\}$. In effect, we take the probability weights associated with events and assign them to real numbers. Remember that when we deal with probabilities on some random variable X, we're really dealing with the $\mathbb{P}_X$ measure.

There are some notational concerns involving the induced probability measure $\mathbb{P}_X$. We often encounter statements like $\mathbb{P}[2] = 0.5$. This statement is completely uninformative (and meaningless) unless in some context. What is meant by this statement is that the probability of the event in $\mathcal{E}$ which are assigned a value of "2" by the random variable in question is 0.5. Remember, the probability measure on the range of the random variable is induced by the probability measure on the sample space.

In general, if we want to look at some condition $X \in A$ (which might be something like $X = 2$), we may write $\mathbb{P}[X \in A]$, but we really mean $\mathbb{P}\left[X^{-1}(A)\right] = \mathbb{P}\{\omega \in \Omega : X(\omega) \in A\}$. In other words, we're measuring the "size" of the set (using $\mathbb{P}$ as our measure) of ω's such that the random variable X returns values in A. Remember that $\mathbb{P}$ is the probability measure over the sample space and $\mathbb{P}_X$ is the probability measure over the range of the random variable. Therefore, when we write $\mathbb{P}[A]$ (where A is a subset of the range of X,) what we really mean is $\mathbb{P}_X[A]$, which is equivalent to $\mathbb{P}\{\omega \in \Omega : X(\omega) \in A\}$. This notational shortcut of using $\mathbb{P}[A]$ instead of $\mathbb{P}_X[A]$ is extremely common, but can be misleading if there's confusion about whether A is in the sample space or in the range of X.

One standard probability space is the **Borel field** over the **unit interval** of the real line under the **Lebesgue measure** λ. Notationally, that's $([0,1], \mathcal{B}, \lambda)$. The Borel field over the unit interval gives us a set of all possible intervals taken from $[0,1]$. The **Lebesgue measure** (denoted by $\lambda[\cdot]$ and pronounced "le-BAYG") is just a way to measure the size of any given interval. For any interval $[a,b] \subseteq [0,1]$ with $b \geq a$, $\lambda[[a,b]] = b - a$. As you can see, Lebesgue measure is just a fancy name for a very intuitive method of measuring the size of intervals.

Despite the confusing math rhetoric, this probability space is one of the most common and well-known. It's the "uniform distribution", where the probability of any interval of values is simply the size of that interval. For example, the probability that some random variable with a uniform distribution lies in the interval $[1/3, 1/2]$ is $1/2 - 1/3 = 1/6$. The concept of a random variable will be defined shortly. One interesting observation is that any point $x \in \mathbb{R}$ has Lebesgue measure zero since $x = [x,x]$ and $\lambda[[x,x]] = x - x = 0$. Another observation is that Lebesgue measure does not depend on the closedness of the interval. In other words, $\lambda[[x,y]] = \lambda[(x,y)] = y - x$.

Of course, as we learn different probability distributions, we'll implicitly be learning different probability measures other than the simple Lebesgue measure λ.

Example 1.25 *Refer back to Example 1.21 where two coins are tossed. We defined the sample space (Ω) as all possible outcomes and the sigma algebra ($\mathcal{E}$) of all possible subsets of the sample space. A simple probability measure ($\mathbb{P}$) was applied to the events in the sigma algebra.*

Define the random variable X to be "the number of heads." Recall that X takes Ω into $\mathcal{X}$ and induces $\mathbb{P}_X$ from $\mathbb{P}$. In this example, $\mathcal{X} = \{0,1,2\}$ and $\mathcal{A} = \{\emptyset, \{0\}, \{1\}, \{2\}, \{0,1\}, \{0,2\}, \{1,2\}, \{0,1,2\}\}$. The induced probability measure $\mathbb{P}_X$ from the measure defined above would look like:

Prob. of 0 heads $= \mathbb{P}_X[0] = \mathbb{P}[\{TT\}] = 1/4$
Prob. of 1 heads $= \mathbb{P}_X[1] = \mathbb{P}[\{HT, TH\}] = 1/2$
Prob. of 2 heads $= \mathbb{P}_X[2] = \mathbb{P}[\{HH\}] = 1/4$
Prob. of 0 or 1 heads $= \mathbb{P}_X[\{0,1\}] = \mathbb{P}[\{TT, TH, HT\}] = 3/4$
Prob. of 0 or 2 heads $= \mathbb{P}_X[\{0,2\}] = \mathbb{P}[\{TT, HH\}] = 1/2$
Prob. of 1 or 2 heads $= \mathbb{P}_X[\{1,2\}] = \mathbb{P}[\{TH, HT, HH\}] = 3/4$
Prob. of 1, 2, or 3 heads $= \mathbb{P}_X[\{0,1,2\}] = \mathbb{P}[\{HH, TH, HT, TT\}] = 1$
Prob. of "nothing" $= \mathbb{P}_X[\emptyset] = \mathbb{P}[\emptyset] = 0$
The empty set is simply needed to complete the σ-algebra. Its interpretation is not important since $\mathbb{P}[\emptyset] = 0$ for any reasonable $\mathbb{P}$.

If all of the above was confusing, rely on the following "executive summary." We have defined a probability space, $(\Omega, \mathcal{E}, \mathbb{P})$, where:

- Ω is the **sample space** - the set of possible outcomes from an experiment.
  - An **event** A is a set containing outcomes from the sample space.
- Σ is a **σ-algebra** of subsets of the sample space. Think of Σ as the collection of all possible events involving outcomes chosen from Ω.
- $\mathbb{P}$ is a **probability measure** over Σ. Remember that $\mathbb{P}$ assigns a number to each event in Σ.

We also have random variables that allow us to look at real numbers instead of abstract events in Σ. For each random variable X, there exists a new probability measure $\mathbb{P}_X$. $\mathbb{P}_X[A]$ where $A \in \mathbb{R}$ simply relates back to $\mathbb{P}\{\omega \in \Omega : X(\omega) \in A\}$. So, if we ask for $\mathbb{P}_X[A]$, we are really interested in the probability $\mathbb{P}\{\omega \in \Omega : X(\omega) \in A\}$, where A simply represents $\{\omega \in \Omega : X(\omega) \in A\}$ through the inverse transformation X^{-1}.

1.3 Conditional Probability & Independence

1.3.1 Conditional Probability

Conditional probability is the probability measure of an event in question given that another event from the same sample space has occurred. For example, the example of 2 coins being tossed was introduced previously. Suppose we know that the event "at least one head" occurred. If we now want to measure the probability of the event "two heads" given that "at least one head" has occurred, we have effectively reduced our σ-algebra of events down to those that satisfy "at least one head." Notationally, we write this as $\mathbb{P}[\textit{two heads} \mid \textit{at least one head}]$.

To calculate conditional probabilities, we use the following formula:

Definition 1.23 *The* ***conditional probability*** *of A given B is*

$$\mathbb{P}\left[A|B\right] = \frac{\mathbb{P}\left[A \cap B\right]}{\mathbb{P}\left[B\right]} \tag{1.11}$$

This definition leads directly to a useful theorem about partitions of the sample space.

Theorem 1.26 *If $A_1, A_2, ..., A_n$ form a partition of the sample space, then* $\mathbb{P}\left[B\right] = \sum_{i=1}^{n} \mathbb{P}\left[B|A_i\right] \ \mathbb{P}\left[A_i\right]$

Proof. *Rearranging Definition 1.23 gives $\mathbb{P}\left[A \cap B\right] = \mathbb{P}\left[A|B\right] \mathbb{P}\left[B\right]$. Recall from Theorem 1.17 that $\mathbb{P}[B] = \sum_{i=1}^{n} \mathbb{P}[B \cap A_i]$, where $\{A_1, A_2, ..., A_n\}$ forms a partition of $\mathcal{E}$. Substituting the definition of conditional probability into this theorem gives the result.* ■

1.3.2 Warner's Method

Warner's Method is a clever way of extracting sensitive survey data using the results of the above theorem. Subjects enter a private booth and choose one of two cards, each containing a "yes/no" question. The subjects cannot see the questions before they choose a card. One of the questions (which we'll call Question 1) will always generate a "yes" response (for example, "Are you human?") The other will contain the sensitive question of interest (for example, "Do you use cocaine?") The subject answers the question on the chosen card and gives only his answer to the experimenter without saying which card was chosen. The experimenter has no way of knowing which question was being answered by this subject. However, if this experiment is done with many subjects, the proportion of people who answer the sensitive question can be accurately determined.

Let "Y" indicate a "yes" response, Q_1 indicate that question 1 was chosen, and Q_2 indicate that question 2 was chosen. After the experiment is run, the proportion of "yes" answers seen by the experimenter is p. Since Q_1 and Q_2 form a partition, we know that

$$\mathbb{P}\left[Y\right] = \mathbb{P}\left[Y|Q_1\right] \mathbb{P}\left[Q_1\right] + \mathbb{P}\left[Y|Q_2\right] \mathbb{P}\left[Q_2\right] \tag{1.12}$$

This can be rearranged to give the proportion of "yes" answers to the more sensative question (Question 2):

$$\mathbb{P}\left[Y|Q_2\right] = \frac{\mathbb{P}\left[Y\right] - \mathbb{P}\left[Y|Q_1\right] \mathbb{P}\left[Q_1\right]}{\mathbb{P}\left[Q_2\right]} = \frac{p - (1)\,(1/2)}{(1/2)} = 2p - 1 \tag{1.13}$$

Note that this method does assume that the questions are chosen with equal probability and that every subject always answers "yes" to Question 1.

1.3.3 Independence

In some cases, the knowledge of one event has no effect on another. When this is true, we say the events are **independent**.

Definition 1.24 *Events A and B are **indpendent** if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$.*

Definition 1.25 *$A_1, A_2, ..., A_n$ are **mutually independent** if $P\left[\cap_{j=1}^{k} A_{i_j}\right] = \prod_{j=1}^{k} \mathbb{P}\left[A_{i_j}\right]$ for any collection $A_{i_1}, A_{i_2}, ..., A_{i_k}$ of $A_1, A_2, ..., A_n$.*

Theorem 1.27 *If A and B are independent, then $\mathbb{P}[A|B] = \mathbb{P}[A]$*

Proof. *Just use the definitions of conditional probability and independence:*

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[A]\ \ \mathbb{P}[B]}{\mathbb{P}[B]} = \mathbb{P}[A] \tag{1.14}$$

■

Theorem 1.28 *If A and B are independent, then*

1. *A and B^C are independent*

   Proof. *We know from Theorem 1.17 that $\mathbb{P}[A] = \mathbb{P}[A \cap B] + \mathbb{P}\left[A \cap B^C\right]$ since B and B^C form a partition.*

$$\begin{aligned} \mathbb{P}[A] &= \mathbb{P}[A \cap B] + \mathbb{P}\left[A \cap B^C\right] && (1.15)\\ \mathbb{P}\left[A \cap B^C\right] &= \mathbb{P}[A] - \mathbb{P}[A \cap B] && (1.16)\\ &= \mathbb{P}[A] - \mathbb{P}[A]\,\mathbb{P}[B] && (1.17)\\ &= \mathbb{P}[A]\,(1 - \mathbb{P}[B]) && (1.18)\\ &= \mathbb{P}[A]\,\mathbb{P}\left[B^C\right] && (1.19) \end{aligned}$$

   ■

2. *A^C and B are independent*

   Proof. *Same as the previous proof.* ■

3. *A^C and B^C are independent.*

   Proof. *From DeMorgan's Laws, $A^C \cap B^C = (A \cup B)^C$.*

$$\begin{aligned} \mathbb{P}\left[A^C \cap B^C\right] &= \mathbb{P}\left[(A \cup B)^C\right] && (1.20)\\ &= 1 - \mathbb{P}[A \cup B] && (1.21)\\ &= 1 - (\mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]) && (1.22)\\ &= 1 - \mathbb{P}[A] - \mathbb{P}[B] + \mathbb{P}[A]\,\mathbb{P}[B] && (1.23)\\ &= (1 - \mathbb{P}[A])\,(1 - \mathbb{P}[B]) && (1.24)\\ &= \mathbb{P}\left[A^C\right]\mathbb{P}\left[B^C\right] && (1.25) \end{aligned}$$

   ■

1.3.4 Philosophical Remarks

Conditional probability and independence are here introduced as mathematical formulae. However, they have serious philosophical meanings and difficulties that open questions about the very basics of probability theory and cause-effect relationships. A few quick examples are given as thoughtful distractions.

Example 1.29 *Events with probability 1 and probability 0 are independent of themselves. Let* $\mathbb{P}[A] = 0$ *and* $\mathbb{P}[B] = 1$. $\mathbb{P}[A \cap A] = \mathbb{P}[A]\ \mathbb{P}[A] = 1$ *and* $\mathbb{P}[B \cap B] = \mathbb{P}[B]\,\mathbb{P}[B] = 0$. *How can an event be independent of its own occurrence? How can a zero-probability event still have zero probability given that it has occurred - or does result reflect the fact that it couldn't have occurred in the first place?*

Example 1.30 *The statement "A is independent of B" makes no reference to the probability measure. If we change probability measures, we can remove independence. What does this say about the causal independence of these two real-world events?*

Example 1.31 *Imagine tossing a fair coin three times. Let* A *be the event "at least two heads" and* B *be the event "the first two flips are the same." Note that* A *and* B *are independent.* $\mathbb{P}[A \cap B] = 1/4 = 1/2 * 1/2 = \mathbb{P}[A]\,\mathbb{P}[B]$.

Now toss a fair coin four times. Define A *and* B *as before. We know that* $\mathbb{P}[B]$ *will be* $1/2$ *regardless of how many flips are made. Now* A *and* B *are not independent.* $\mathbb{P}[A \cap B] = 10/32 \neq 11/32 = 11/16 * 1/2 = \mathbb{P}[A]\ \mathbb{P}[B]$.

Try two tosses. $\mathbb{P}[A \cap B] = 1/4 \neq 1/8 = 1/4 * 1/2 = \mathbb{P}[A]\,\mathbb{P}[B]$. *Not independent.*

Try any number of tosses other than three and we will always find dependence.

It is a mathematical oddity that independence only comes with three tosses of the coin. Does this imply a true independence of these events only when the coin is tossed three times? Can it be that two events will have the intersection of their probability measures coincidentally equal to the product of those measures while maintaining an underlying causal dependence?

1.4 Important Probability Tools

Theorem 1.32 *(Principle of Inclusion & Exclusion)*

$$\begin{aligned} P\left[\cup_{i=1}^{n} A_i\right] = & \sum_{i=1}^{n} \mathbb{P}\left[A_i\right] \\ & - \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{P}\left[A_i \cap A_j\right] \\ & + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} \mathbb{P}\left[A_i \cap A_j \cap A_k\right] \\ & - \ldots \\ & + (-1)^{n+1}\, \mathbb{P}\left[\cap_{i=1}^{n} A_i\right] \end{aligned} \tag{1.26}$$

Proof. *Before considering the general proof, look first at the first few steps of the $n = 4$ case.*

$$\begin{aligned} \mathbb{P}\left[A_1 \cup A_2 \cup A_3 \cup A_4\right] = & \mathbb{P}\left[A_1\right] + \mathbb{P}\left[A_2 \cup A_3 \cup A_4\right] - \mathbb{P}\left[A_1 \cap (A_2 \cup A_3 \cup A_4)\right] \qquad (1.27) \\ & = \mathbb{P}\left[A_1\right] + \mathbb{P}\left[A_2 \cup A_3 \cup A_4\right] - \\ & \mathbb{P}\left[(A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_1 \cap A_4)\right] \\ & = \mathbb{P}\left[A_1\right] + \mathbb{P}\left[A_2\right] + \mathbb{P}\left[A_3 \cup A_4\right] - \mathbb{P}\left[A_2 \cap (A_3 \cup A_4)\right] - \\ & \mathbb{P}\left[(A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_1 \cap A_4)\right] \\ & = \mathbb{P}\left[A_1\right] + \mathbb{P}\left[A_2\right] + \mathbb{P}\left[A_3 \cup A_4\right] - \mathbb{P}\left[(A_2 \cap A_3) \cup (A_2 \cap A_4)\right] - \\ & \mathbb{P}\left[(A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_1 \cap A_4)\right] \\ & = \mathbb{P}\left[A_1\right] + \mathbb{P}\left[A_2\right] + \mathbb{P}\left[A_3\right] + \mathbb{P}\left[A_4\right] - \mathbb{P}\left[A_3 \cap A_4\right] - \\ & \mathbb{P}\left[(A_2 \cap A_3) \cup (A_2 \cap A_4)\right] - \\ & \mathbb{P}\left[(A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_1 \cap A_4)\right] \end{aligned}$$

The strategy is to expand unions by separating the first term out. The resulting subtracted term contains a single term intersected with a sequence of unions. We distribute that intersetion through the unions. The first of these unions is then separated and the process is repeated until all union operators are *eliminated.*

$$
\begin{aligned}
\mathbb{P}\left[\cup_{i=1}^{n} A_i\right] &= \mathbb{P}\left[A_1\right] + \mathbb{P}\left[\cup_{i=2}^{n} A_i\right] - \mathbb{P}\left[A_1 \cap \left(\cup_{i=2}^{n} A_i\right)\right] \\
&= \mathbb{P}\left[A_1\right] + \underbrace{\mathbb{P}\left[A_2\right] + \mathbb{P}\left[\cup_{i=3}^{n} A_i\right] - \mathbb{P}\left[A_2 \cap \left(\cup_{i=3}^{n} A_i\right)\right]}_{\mathbb{P}\left[\cup_{i=2}^{n} A_i\right]} - \mathbb{P}\left[A_1 \cap \left(\cup_{i=2}^{n} A_i\right)\right] \\
&\vdots \\
&= \sum_{i=1}^{n} \mathbb{P}\left[A_i\right] - \sum_{i=1}^{n} \mathbb{P}\left[A_i \cap \left(\cup_{j=1+1}^{n} A_j\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{P}\left[A_i\right] - \sum_{i=1}^{n} \mathbb{P}\left[\cup_{j=1+1}^{n} \left(A_i \cap A_j\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{P}\left[A_i\right] - \sum_{i=1}^{n} \begin{pmatrix} \mathbb{P}\left[A_i \cap A_{i+1}\right] + \mathbb{P}\left[\cup_{j=i+2}^{n} \left(A_i \cap A_j\right)\right] \\ -\mathbb{P}\left[\left(A_i \cap A_{i+1}\right) \cap \left(\cup_{j=i+2}^{n} \left(A_i \cap A_j\right)\right)\right] \end{pmatrix} \\
&= \sum_{i=1}^{n} \mathbb{P}\left[A_i\right] - \sum_{i=1}^{n} \begin{pmatrix} \mathbb{P}\left[A_i \cap A_{i+1}\right] + \mathbb{P}\left[A_i \cap A_{i+2}\right] + \mathbb{P}\left[\cup_{j=i+3}^{n} \left(A_i \cap A_j\right)\right] \\ -\mathbb{P}\left[\left(A_i \cap A_{i+2}\right) \cap \left(\cup_{j=i+3}^{n} \left(A_i \cap A_j\right)\right)\right] \\ -\mathbb{P}\left[\left(A_i \cap A_{i+1}\right) \cap \left(\cup_{j=i+2}^{n} \left(A_i \cap A_j\right)\right)\right] \end{pmatrix} \\
&\vdots \\
&= \sum_{i=1}^{n} \mathbb{P}\left[A_i\right] - \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{P}\left[A_i \cap A_j\right] + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{P}\left[\left(A_i \cap A_j\right) \cap \left(\cup_{k=j+1}^{n} \left(A_i \cap A_k\right)\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{P}\left[A_i\right] - \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{P}\left[A_i \cap A_j\right] + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{P}\left[\cup_{k=j+1}^{n} \left(\left(A_i \cap A_j\right) \cap \left(A_i \cap A_k\right)\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{P}\left[A_i\right] - \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{P}\left[A_i \cap A_j\right] + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{P}\left[\cup_{k=j+1}^{n} \left(A_i \cap A_j \cap A_k\right)\right] \\
&\vdots \\
&= \sum_{i=1}^{n} \mathbb{P}\left[A_i\right] - \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{P}\left[A_i \cap A_j\right] + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} \mathbb{P}\left[A_i \cap A_j \cap A_k\right] - \\
&\sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} \mathbb{P}\left[\left(A_i \cap A_j \cap A_k\right) \cap \left(\cup_{l=k+1}^{n} \left(A_i \cap A_j \cap A_l\right)\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{P}\left[A_i\right] - \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{P}\left[A_i \cap A_j\right] + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} \mathbb{P}\left[A_i \cap A_j \cap A_k\right] - \\
&\sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} \mathbb{P}\left[\cup_{l=k+1}^{n} \left(A_i \cap A_j \cap A_k \cap A_l\right)\right] \\
&\vdots \\
&= \sum_{i=1}^{n} \mathbb{P}\left[A_i\right] - \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{P}\left[A_i \cap A_j\right] + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} \mathbb{P}\left[A_i \cap A_j \cap A_k\right] - \\
&... + (-1)^{n+1}\, \mathbb{P}\left[\cap_{i=1}^{n} A_i\right]
\end{aligned}
\tag{1.28}
$$

■

Theorem 1.33 *(Bonferroni's Inequality - Simple)* $\mathbb{P}[A \cap B] \geq \mathbb{P}[A] + \mathbb{P}[B] - 1$

Proof.

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] \tag{1.29}$$

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cup B] \geq \mathbb{P}[A] + \mathbb{P}[B] - 1 \tag{1.30}$$

$$\mathbb{P}[A \cap B] \geq \mathbb{P}[A] + \mathbb{P}[B] - 1 \tag{1.31}$$

■

Theorem 1.34 *(Bonferroni's Inequality - General)* $\mathbb{P}\left[\cap_{i=1}^n A_i\right] \geq \sum_{i=1}^n \mathbb{P}\left[A_i\right] - (n-1)$

Proof. *Starting with Boole's Inequality applied to A_i^C, we have that*

$$\mathbb{P}[\cup_{i=1}^n A_i^C] \leq \sum_{i=1}^n \mathbb{P}[A_i^C] \tag{1.32}$$

$$\mathbb{P}[(\cap_{i=1}^n A_i)^C] \leq \sum_{i=1}^n \mathbb{P}[A_i^C] \tag{1.33}$$

$$1 - \mathbb{P}[\cap_{i=1}^n A_i] \leq \sum_{i=1}^n (1 - \mathbb{P}[A_i]) \tag{1.34}$$

$$-\mathbb{P}[\cap_{i=1}^n A_i] \leq -1 + n + \sum_{i=1}^n (-\mathbb{P}[A_i]) \tag{1.35}$$

$$\mathbb{P}[\cap_{i=1}^n A_i] \geq \sum_{i=1}^n \mathbb{P}[A_i] - (n-1) \tag{1.36}$$

■

Theorem 1.35 *(Bonferroni's Inequality - Alternative Version)* $\mathbb{P}[\cap_{i=1}^\infty A_i] \geq 1 - \sum_{i=1}^n \mathbb{P}\left[A_i^C\right]$

Proof. *This proof uses many of the tools used in the general version of Bonferroni's Inequality and the final step requires Boole's Identity.*

$$\mathbb{P}[\cap_{i=1}^n A_i] = 1 - \mathbb{P}\left[([\cap_{i=1}^n A_i)^C\right] = 1 - \mathbb{P}\left[\cup_{i=1}^n A_i^C\right] \tag{1.37}$$

$$\geq 1 - \sum_{i=1}^n \mathbb{P}[A_i^C] \tag{1.38}$$

■

Theorem 1.36 *(Bayes' Rule)* $\mathbb{P}[A_i|B] = \frac{\mathbb{P}[B|A_i]\mathbb{P}[A_i]}{\sum_{j=1}^\infty \mathbb{P}[B|A_j]\mathbb{P}[A_j]}$, *where* $A_1, A_2, ...$ *partition the sample space.*

Proof. $\mathbb{P}[A_i|B] = \frac{\mathbb{P}[A_i \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B \cap A_i]\mathbb{P}[A_i]}{\mathbb{P}[A_i]\mathbb{P}[B]} = \frac{\mathbb{P}[B|A_i]\mathbb{P}[A_i]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B|A_i]\mathbb{P}[A_i]}{\sum_{j=1}^\infty \mathbb{P}[B|A_j]\mathbb{P}[A_j]}$ since $\{A_n\}$ is a partition. ■

Remark 1.37 *Bayes' Rule is* very *important in many economic problems, particularly in the field of game theory.*

1.4.1 Probability Distributions

A random variable induces a probability measure on the real line from the probability measure over the sample space. This measure assigns measure (or weight) to real numbers in the range of the random variable. The term for this induced measure on the real line is the **distribution** of the random variable. From this distribution, we can define a function representing the relationship between a real number and its assigned probability measure. This function is a **distribution function**.

Definition 1.26 *The* ***cumulative distribution function*** F_X *of a random variable* X *(also known as its* ***cdf****) is*

$$F_X(x) = \mathbb{P}[X \leq x] \tag{1.39}$$

In all of statistics, it is crucial to note the difference between the capital letter X and the small letter x. The capital letter refers to the random variable itself, while the small letter refers to one particular value that the random variable might take. We subscript the cdf function with the random variable to remind us which measure we're looking at. If it is obvious, the subscript is ignored.

Also, we use the notation $X \sim F(\theta)$ to indicate that $F_X(x;\theta)$ is the cdf of X, which might take some parameter θ. For example, the Normal distribution takes as its parameters the mean of the distribution, μ, and the variance, σ^2. So, if X were distributed with a normal distribution, we would say $X \sim N(\mu, \sigma^2)$.

Theorem 1.38 *Any cumulative distribution function* F *has the following properties*

1. F *is nondecreasing*
2. $\lim_{x\to\infty} F(x) = 1$
3. $\lim_{x\to-\infty} F(x) = 0$
4. F *is right continuous (*$\lim_{x\to x_0^+} F(x) = F(x_0)$*)*

Theorem 1.39 *Any function* G *satisfiying the four conditions of Theorem 1.38 is the CDF of some random variable.*

Definition 1.27 *The* ***probability distribution function*** *(or, in discrete random variables, the* ***probability mass function****), (also known as the* ***pdf****) is*

$$f_X(x) = \mathbb{P}[X = x] \tag{1.40}$$

Corollary 1.40 *The cdf is the integral of the pdf in the following way*

$$F_X(y) = \mathbb{P}[X \leq y] = \int_{-\infty}^{y} f_X(t)dt \tag{1.41}$$

where t is a dummy variable of integration.

Theorem 1.41 *Any probability distribution function f has the following properties*

1. $f_X(x) \geq 0 \; \forall x \in \mathcal{X}$
2. $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$

Theorem 1.42 *Continuous random variables have continuous cdf's while discrete random variables have right-continuous step-function cdf's.*

In the Subsection 2.1 we consider probability distributions in more detail, but these definitions are needed at this point to understand the concepts of expected value and variance of random variables.

1.4.2 The Expectation Operator

The expectation operator defines the mean (or population average) of a random variable or expression.

We develop the expectation operator in terms of the Lebesgue integral. First recall that (roughly speaking) the Lebesgue measure $\lambda(A)$ for some set A gives the length/area/volume of the set A. If $A = (2, 5)$, then $\lambda(A) = 2 - 5 = 3$. The Lebesgue integral of f on $[a, b]$ is defined in terms of $\sum_{i=1}^{n} y_i \lambda(A_i)$, where $0 = y_1 \leq y_2 \leq ... \leq y_n$, $A_i = \{x : y_i \leq f(x) < y_{i+1}\}$, and $\lambda(A_i)$ is the Lebesgue measure of the set A_i. The value of the Lebesgue integral is the limit as the y_i's are pushed closer and closer together. Essentially, we break the y-axis into a grid using $\{y_n\}$ and break the x-axis into the corresponding grid $\{A_n\}$ where $A_i = \{x : f(x) \in [y_i, y_{i+1})\}$. We then calculate the area under the function by taking the sum of the rectangles given by $\lambda(A_i)\, y_i$.

Consider our probability space. Take an event (a set A of $\omega \in \Omega$) and a random variable that assigns real numbers to each $\omega \in A$. If we were to take an observation from A without knowing which $\omega \in A$ was to be drawn, we may want to ask what value of $X(\omega)$ we should *expect* to see. Since each of the $\omega \in A$ has been assigned a probability measure $\mathbb{P}[\omega]$ (which induces $\mathbb{P}_X[x]$,) we can use this to weight the values $X(\omega)$. Since $\mathbb{P}$ is a probability measure, these weights sum to 1, so the weighted sum provides us with a weighted average of $X(\omega)$. If our measure $\mathbb{P}$ actually gives the "correct" likelihood of ω being chosen, then this weighted average of $X(\omega)$ (which we'll call $\mathbb{E}[X]$) gives an indication of what values of $X(\omega)$ we should expect to draw.

Using the Lebesgue integral concept, we can take the possible values $\{x_i\}$ and construct a grid on the y-axis, which gives a corresponding grid on the

x-axis in A, where $A_i = \{\omega \in A : X(\omega) \in [x_i, x_{i+1})\}$. Denote elements in this x-axis grid as A_i. The weighted average is

$$\sum_{i=1}^{n} x_i \mathbb{P}[A_i] = \sum_{i=1}^{n} x_i \mathbb{P}_X[X = x_i] = \sum_{i=1}^{n} x_i f_X(x_i) \tag{1.42}$$

As we shrink this grid infinitessimaly, A_i will become an infinitessimal. Denote the infinitessimal set A_i by $d\omega$. The Lebesgue integral becomes

$$\mathrm{lim}_{n\to\infty} \sum_{i=1}^{n} x_i \mathbb{P}[A_i] = \int_{-\infty}^{\infty} x\mathbb{P}[d\omega] = \int_{-\infty}^{\infty} x\mathbb{P}_X[X = x] = \int_{-\infty}^{\infty} x f_X(x)\, dx \tag{1.43}$$

We now have our definition of expected values.

Definition 1.28 *The* ***expected value*** *of a continuous random variable X is*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\mathbb{P}[d\omega] = \int_{-\infty}^{\infty} x f_X(x) dx \tag{1.44}$$

The expected value of a discrete random variable Y is

$$\mathbb{E}[Y] = \sum_{y\in Y} y f_Y(y)$$

Remark 1.43 *We sometimes denote $\mathbb{E}[\cdot]$ as $\mathbb{E}_X[\cdot]$ to indicate that the expectation is being taken over $f_X(\cdot)dx$.*

The following results can be easily shown using the definition of expected value and some simple integral manipulation. Therefore, proofs are omitted.

- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$

An alternative measure of the "center" of the distribution is the median.

Definition 1.29 *The* ***median*** *of a random variable X is the unique number m that solves*

$$\int_{-\infty}^{m} f_X(x)\, dx = \int_{m}^{\infty} f_X(x)\, dx \tag{1.45}$$

Theorem 1.44 *For any random variable X with median m, $F_X(m) = \frac{1}{2}$.*

Proof. *Simply integrate out the definition of the median and solve for $F_X(m)$.*

$$\int_{-\infty}^{m} f_X(x)\, dx = \int_{m}^{\infty} f_X(x)\, dx \tag{1.46}$$

$$F_X(m) - F_X(-\infty) = F_X(\infty) - F_X(m) \tag{1.47}$$

$$F_X(m) - 0 = 1 - F_X(m) \tag{1.48}$$

$$2F_X(m) = 1 \tag{1.49}$$

$$F_X(m) = \frac{1}{2} \tag{1.50}$$

■

1.4.3 Variance and Coviariance

Definition 1.30 *The* ***variance*** *of a random variable X, denoted* $\text{Var}\,[X]$*, is*

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] \tag{1.51}$$

Definition 1.31 *The* ***covariance*** *between two random variables X and Y, denoted* $\text{Cov}\,[X, Y]$*, is*

$$\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{1.52}$$

The following results can be shown using the definitions of variance and covariance given above. Therefore, proofs are again omitted.

- $\text{Var}[XY] = \text{Var}[X] + \text{Var}[Y] - \text{Cov}[X, Y]$
- $\text{Var}[aX + b] = a^2\,\text{Var}[X]$

Theorem 1.45 $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}^2[X]$

Proof. *We can use simple algebra to show this result*

$$\begin{aligned}\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] &= \mathbb{E}[X^2 - 2X\mathbb{E}[X]^2 + \mathbb{E}[X]^2] \qquad (1.53)\\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X\mathbb{E}[X]^2] + \mathbb{E}[X]^2\end{aligned}$$

■

Theorem 1.46 $\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Proof. *From Definition 1.31 we know that*

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \qquad (1.54)\\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]]\\ &= \mathbb{E}[XY] - \mathbb{E}[X\mathbb{E}[Y]] - \mathbb{E}[Y\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]\mathbb{E}[Y]]\end{aligned}$$

■

Chapter 2

Probability Distributions

2.1 Density & Mass Functions

Recall from above the definitions of cumulative distribution functions and probability distribution functions.

$$F_X(x) = \mathbb{P}[X \leq x] \tag{2.1}$$

$$f_X(x) = \mathbb{P}[X = x] \tag{2.2}$$

The probability that X satisfies some condition $X \in E$, is

$$\mathbb{P}[X \in E] = \int_{x \in E} f_X(x)dx \text{ for continuous distributions} \tag{2.3}$$

$$\mathbb{P}[X \in E] = \sum_{x \in E} f_X(x)dx \text{ for discrete distributions} \tag{2.4}$$

Therefore, $\mathbb{P}[X \leq c] = \int_{-\infty}^{c} f_X(t)dt = F_X(c)$ (or, $\sum_{t=-\infty}^{c} f_X(t) = F_X(c)$ for discrete distributions.)

2.1.1 Moments & MGFs

Definition 2.1 *For each integer n, the n^{th} **moment** of X, called μ'_n, is $\mu'_n = \mathbb{E}[X^n]$ and the n^{th} **central moment** of X, called μ_n, is $\mu_n = \mathbb{E}[X - \mu]^n$*

Remark 2.1 $\text{Var}[X] = \mu_2$

Definition 2.2 *The **moment generating function (mgf)** of X, denoted $M_X(t)$, is*

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx \textit{ if } X \textit{ is continuous}$$

$$M_X(t) = \sum_x e^{tx} \mathbb{P}[X = x] \textit{ if } X \textit{ is discrete}$$

Remark 2.2 $\mathbb{E}[X^n] = M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)|_{t=0}$

Theorem 2.3 *If $F_X(x)$ and $F_Y(y)$ are two cdfs all of whose moments exist, then*

1. *If X and Y have bounded support, then $F_X(u) = F_Y(u)\ \forall u\ iff\ \mathbb{E}[X^r] = \mathbb{E}[Y^r]\ \forall r = 0, 1, 2, ...$*
2. *If the mgfs exist and $M_X(t) = M_Y(t)\ \forall t$ in some neighborhood of 0, then $F_X(u) = F_Y(u)\ \forall u$.*

Theorem 2.4 *The mgf of the random variable $aX + b$ is given by*

$$M_{aX+b}(t) = e^{bt} M_X(at)$$

2.1.2 A Side Note on Differentiating Under and Integral

Theorem 2.5 *(Leibnitz's Rule) If $f(x, \theta), a(\theta)$, and $b(\theta)$ are differentiable with respect to θ, then*

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)dx = f(b(\theta),\theta)\frac{d}{d\theta}b(\theta) - f(a(\theta),\theta)\frac{d}{d\theta}a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta} f(x,\theta)dx$$

and if $a(\theta)$ and $b(\theta)$ are finite constant functions, then

$$\frac{d}{d\theta}\int_a^b f(x,\theta)dx = \int_a^b \frac{\partial}{\partial\theta} f(x,\theta)dx$$

Theorem 2.6 *Suppose $f(x, \theta)$ is differentiable at θ_0 and there exists a function $g(x, \theta_0)$ and a constant $\delta_0 > 0$ such that*

1. $\left|\frac{f(x,\theta_0+\delta)-f(x,\theta_0)}{\delta}\right| \leq g(x, \theta_0)\ \forall x\ \forall\ |\delta| \leq \delta_0$
2. $\int_{-\infty}^{\infty} g(x, \theta_0)dx < \infty$

   then
$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)dx|_{\theta=\theta_0} = \int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta} f(x,\theta)|_{\theta=\theta_0}\right) dx$$

In short, you can move derivatives across integrals if and only if

1. The limits of integration are constant and finite, or
2. The limits of integration are infinite and there is some function that weakly dominates the slope of the function inside the integral whose infinite integral is itself finite.

2.2 Commonly Used Distributions

2.2.1 Discrete Distributions

Binomial(n,p)

Description	n Bernoulli trials, x successes
PMF	$\binom{n}{x}p^x(1-p)^{n-x}$
Mean	np
Variance	$np(1-p)$
Notes	Bernoulli Dist'n is Binomial(1,p)

Multinomial(n,m,p)

Description	m trials with n possible outcomes, x_i=# times i^{th} outcome occurs
PMF	$f(x_1, x_2, ..., x_n) = \frac{m!}{x_1!x_2!...x_n!}p_1^{x_1}p_2^{x_2}...p_n^{x_n} = m!\prod_{i=1}^{n}\frac{p_i^{x_i}}{x_i!}$
Mean	?
Variance	?

Discrete Uniform(N)

Description	uniform w/ X a discrete variable
PMF	$\frac{1}{N}$
Mean	$\frac{N+1}{2}$
Variance	$\frac{(N+1)(N-1)}{12}$
Notes	Bernoulli Dist'n is Binomial(1,p)

Geometric(p)

Description	X=# of trials until 1st success
PMF	$p(1-p)^{x-1}$
Mean	$\frac{1}{p}$
Variance	$\frac{1-p}{p^2}$
Notes	Special case of negative binomial

Hypergeometric(N,M,K)

Description	Draw K balls from an urn of N, with M of them red. How many red drawn?
PMF	$\frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$
Mean	$\frac{KM}{N}$
Variance	$\frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}$
Notes	

Negative Binomial(r,p)

Description	Number of failures before the r^{th} success
PMF	$\binom{r+x-1}{x}p^x(1-p)^x$
Mean	$\frac{r(1-p)}{p}$
Variance	$\frac{r(1-p)}{p^2}$
Notes	Can be derived as a gamma mixture of Poissons

Poisson(λ)

Description	Number of occurrence in a time or space interval
PMF	$\frac{e^{-\lambda}\lambda^x}{x!}$
Mean	λ
Variance	λ
Notes	

2.2.2 Continuous Distributions

Uniform(a, b)

PDF	$\frac{1}{b-a}$
Mean	$\frac{b+a}{2}$
Variance	$\frac{(b-a)^2}{12}$

Normal(μ, σ^2)

PDF	$\frac{1}{\sqrt{2\pi}\sigma} e^{\left(\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)}$
Mean	μ
Variance	σ^2
Notes	Member of exponential family of dist'ns.

Student's t(ν)

Description	ν degrees of freedom
PDF	$\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\nu\pi}} \frac{1}{(1+(\frac{x^2}{\nu}))^{\frac{\nu+1}{2}}}$
Mean	0
Variance	$\frac{\nu}{\nu-2}$
Notes	$F_{1,\nu} = t_\nu^2$, $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} \sim t_{n-1}$

Gamma(α, β)

Description	
PDF	$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{(\alpha-1)} e^{\left(\frac{-x}{\beta}\right)}$ $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$
Mean	$\alpha\beta$
Variance	$\alpha\beta^2$
Notes	$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$

Snedecor's F Distribution(ν, υ)

PDF	(too messy!)
Mean	$\frac{\upsilon}{\upsilon-2}$
Variance	(too messy!)
Notes	$F_{\nu,\upsilon} = \frac{\frac{\chi^2_\nu}{\nu}}{\frac{\chi^2_\upsilon}{\upsilon}}$

Exponential(β)

Description	Models things like lifetimes
PDF	$\frac{1}{\beta}e^{\left(\frac{-x}{\beta}\right)}$
Mean	β
Variance	β^2
Notes	Memoryless property

Chi-Squared(p)

Description	Sum of squared standard normals, $p = d.f.$
PDF	$\frac{1}{\Gamma\left(\frac{p}{2}\right)}x^{\left(\frac{p}{2}-1\right)}e^{\left(\frac{-x}{2}\right)}$
Mean	p
Variance	$2p$
Notes	Special case of the gamma distribution

Beta(α, β)

Description	Number of occurrences in a time or space interval
PDF	$\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\;\; B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
Mean	$\frac{\alpha}{\alpha+\beta}$
Variance	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

2.2.3 Useful Approximations

Normal Approximation of Binomial

Binomail random variables can be difficult to compute with large numbers (without the aid of a computer.) However, the normal distribution serves as a good approximation for large n and p not too far from 0.5.

Theorem 2.7 *Let* $X\,\tilde{}\,b\left(n,p\right)$ *and* $Y\,\tilde{}\,N\left(np, np\left(1-p\right)\right)$. *Then*

$$\mathbb{P}\left[X \leq x\right] \approx \mathbb{P}\left[Y \leq x + 0.5\right] \tag{2.5}$$

$$\mathbb{P}\left[X \geq x\right] \approx \mathbb{P}\left[Y \geq x - 0.5\right] \tag{2.6}$$

Chapter 3

Working With Multiple Random Variables

3.1 Random Vectors

A random variable is a function from the sigma-algebra on the sample space to the real line. An n-dimensional random vector is a function from the sigma algebra on the sample space to $\mathbb{R}^n$. For example, we may roll a pair of dice and (X, Y) may represent the numerical values of the dice. Or, perhaps they represent the sum and difference of the two dice.

3.2 Distributions of Multiple Variables

In many cases the rules of probability discussed above that applied to events and their probabilites have nearly identical analogues in the realm of random variables and their distributions.

3.2.1 Joint and Marginal Distributions

If we have multiple random variables (as in a random vector,) the probability that they all take on a particular vector of values is described by the **joint pdf**.

Definition 3.1 *The **joint probability distribution function** is*

$$f_{X,Y}(x, y) = \mathbb{P}\left[X = x, Y = y\right] \tag{3.1}$$

Strictly speaking, if X and Y are continuous, we should define the probabilities over intervals since the probability of a continuous variable being a single number is zero. So, $f_{X,Y}((X, Y) \in A) = \int_{y \in A} \int_{x \in A} f_{X,Y}(x, y) dx dy$.

Definition 3.2 *The* ***joint cumulative distribution function*** *is*

$$F(x,y) = \mathbb{P}[X \le x, Y \le y] \tag{3.2}$$
$$= \int_{-\infty}^{x} \int_{-\infty}^{y} f(s,t)\,dt\ ds$$

Definition 3.3 *The* ***expected value*** *of a function of discrete random variables $g(X,Y)$ is*

$$\mathbb{E}[g(X,Y)] = \sum_{(x,y)\in\mathbb{R}^2} g(x,y)\, f_{X,Y}(x,y) \tag{3.3}$$

and the expected value of a function of continuous random variables $h(X,Y)$ is

$$\mathbb{E}[h(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)\, f_{X,Y}(x,y)\ \ dx\ dy$$

Definition 3.4 *The* ***marginal distribution function*** *is the pdf of only one of the variables in a joint distribution function, and can be calculated by*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy \tag{3.4}$$

if Y is continuous and

$$f_X(x) = \sum_{y\in\mathcal{Y}} f_{X,Y}(x,y) \tag{3.5}$$

if Y is discrete, where $\mathcal{Y}$ is the range of Y.

In words, we "integrate out" the unwanted variable.

3.2.2 Conditional Distributions and Independence

The conditional probability of X given Y is given by $\mathbb{P}[X|Y] = \mathbb{P}[X\cap Y]/\mathbb{P}[Y]$. We define a similar concept for distributions.

Definition 3.5 *The* ***conditional pmf of* X *given* Y = y** *is*

$$f_{X|Y}(x|y) = \mathbb{P}[X = x|Y = y] = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{3.6}$$

At this point, we will relax notation a bit. Since $f(x,y)$ clearly represents $f_{X,Y}(x,y)$ and $f(x|y)$ represents $f_{X|Y}(x|y)$, we often drop the subscripts on the f whenever the distinction is obvious.

The conditional pmf satisfies the properties of a regular pmf. For instance, if X and Y are discrete,

$$\sum_{x\in\mathcal{X}} f(x|y) = \sum_{x\in\mathcal{X}} \frac{f(x,y)}{f_Y(y)} = \frac{\sum_{x\in\mathcal{X}} f(x,y)}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1 \tag{3.7}$$

Furthermore, the expectation and variance operators are similar to as before.

$$\mathbb{E}\left[g\left(Y\right)|X=x\right]=\int_{-\infty}^{\infty}g\left(y\right)f\left(y|x\right)dy \tag{3.8}$$
$$Var\left[Y|x\right]=\mathbb{E}\left[Y^2|x\right]-\left(\mathbb{E}\left[Y|x\right]\right)^2$$

The notion of independence of random variables is also a transparent extension of independence of events.

Definition 3.6 *Random variables X and Y are* ***independent*** *if*

$$f\left(x,y\right)=f_X\left(x\right)f_Y\left(y\right) \tag{3.9}$$

or equivalently, if

$$f\left(y|x\right)=f_Y\left(y\right) \tag{3.10}$$

3.2.3 The Bivariate Normal Distribution

The bivariate normal distribution has several interesting properties and is also a popular prelim and exam topic.

Recall the pdf of the normal distribution for $X\sim N\left(\mu,\sigma^2\right)$

$$f_X\left(x\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{\left(x-\mu\right)^2}{\sigma^2}\right) \tag{3.11}$$

If there exist two random variables $X\sim N\left(\mu_X,\sigma_X^2\right)$ and $Y\sim N\left(\mu_Y,\sigma_Y^2\right)$ such that $\mathrm{Corr}\left[X,Y\right]=\rho$, where μ_X and μ_Y are finite, σ_X^2 and σ_Y^2 are positive, and $\rho\in(-1,1)$, then their joint distribution is the bivariate normal distribution with parameters μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ, which has the following pdf

$$f_{X,Y}\left(x,y\right)=\frac{\exp\left(\frac{-1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2-2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right)\right)}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \tag{3.12}$$

which is abbreviated by

$$\left(X,Y\right)\sim BVN\left(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho\right) \tag{3.13}$$

Although this pdf has many terms, its roots in the univariate normal pdf are obvious upon inspection. Take the product of $f_X\left(x\right)$ and $f_Y\left(y\right)$:

$$f_X\left(x\right)f_Y\left(y\right)=\frac{1}{\sqrt{2\pi\sigma_X^2}}\exp\left(-\frac{1}{2}\frac{\left(x-\mu_X\right)^2}{\sigma_X^2}\right)\frac{1}{\sqrt{2\pi\sigma_Y^2}}\exp\left(-\frac{1}{2}\frac{\left(y-\mu_Y\right)^2}{\sigma_Y^2}\right) \tag{3.14}$$

$$=\frac{1}{2\pi\sigma_X\sigma_Y}\exp\left(-\frac{1}{2}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right) \tag{3.15}$$

and the result is the bivariate normal when $\rho = 0$. This serves as the proof of Theorem 3.3.

Theorem 3.1 *If $f_{X,Y}(x,y)$ is $BVN\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho\right)$, then $X \sim N\left(\mu_X, \sigma_X^2\right)$ and $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$.*

Proof. *We prove the result for X only, which is surprisingly difficult. The proof for Y is symmetric.*

To avoid the extensive notation, we define a few variables.

$$W = \frac{x - \mu_X}{\sigma_X} \tag{3.16}$$

$$V = \frac{y - \mu_Y}{\sigma_Y} \tag{3.17}$$

$$Z \sim N\left(\rho W, \left(1 - \rho^2\right)\right) \tag{3.18}$$

Note that $\partial V / \partial y = 1/\sigma_Y$, which implies that $dy = \sigma_Y dV$.

The BVN pdf is now given by

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(\frac{-\left(W^2 - 2\rho WV + V^2\right)}{2\left(1-\rho^2\right)}\right) \tag{3.19}$$

$$= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(\frac{-W^2}{2\left(1-\rho^2\right)}\right) \tag{3.20}$$

$$\times \exp\left(\frac{-\left(V^2 - 2\rho WV + \rho^2 W^2\right) + \rho^2 W^2}{2\left(1-\rho^2\right)}\right) \tag{3.21}$$

$$= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(\frac{-W^2}{2\left(1-\rho^2\right)} + \frac{\rho^2 W^2}{2\left(1-\rho^2\right)}\right) \tag{3.22}$$

$$\times \exp\left(\frac{-\left(V - \rho W\right)^2}{2\left(1-\rho^2\right)}\right) \tag{3.23}$$

Taking the integral of the joint pdf gives the marginal.

$$f_X(x) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(\frac{-W^2(1-\rho^2)}{2(1-\rho^2)}\right) \tag{3.24}$$

$$\times \int_{-\infty}^{\infty} \exp\left(\frac{-(V-\rho W)^2}{2(1-\rho^2)}\right) \tag{3.25}$$

$$= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(\frac{-W^2}{2}\right) \tag{3.26}$$

$$\times \int_{-\infty}^{\infty} \frac{\sqrt{2\pi}\sqrt{1-\rho^2}}{\sqrt{2\pi(1-\rho^2)}} \exp\left(\frac{-(V-\rho W)^2}{2(1-\rho^2)}\right) dy \tag{3.27}$$

$$= \frac{\sqrt{2\pi}\sqrt{1-\rho^2}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(\frac{-W^2}{2}\right) \tag{3.28}$$

$$\times \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(\frac{-(V-\rho W)^2}{2(1-\rho^2)}\right) \sigma_Y dV \tag{3.29}$$

$$= \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(\frac{-W^2}{2}\right) \tag{3.30}$$

$$\times \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(\frac{-1}{2}\left(\frac{V-\rho W}{\sqrt{1-\rho^2}}\right)^2\right) dV \tag{3.31}$$

$$= \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(\frac{-1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right) \int_{-\infty}^{\infty} f_Z(V)\, dV \tag{3.32}$$

$$= \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(\frac{-1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right) = N\left(\mu_X, \sigma_X^2\right) \tag{3.33}$$

■

Remark 3.2 *The converse of Theorem 3.1 is not true.*

Theorem 3.3 *If X and Y are uncorrelated and $f_{X,Y}(x,y)$ is $BVN(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$, then X and Y are independent.*

Proof. *The discussion at the beginning of this section shows that if $\rho = 0$, then $f_{X,Y}(x,y) = f_X(x) f_Y(y)$, which defines independence of X and Y.* ■

This is a unique result. We know independence implies zero correlation, but the converse is certainly not true. However, in the BVN case, the converse is always true. This makes for a clever test question.

Theorem 3.4 *If Z_1 and Z_2 are independent standard normal variables $(N(0,1))$ and A is a 2×2 matrix of real numbers, then $(X,Y)' = A(Z_1, Z_2)'$ have a bivariate normal joint pdf.*

Proof. *Left as an exercise.* ■

We now consider the conditional distribution of $X|Y$ when X and Y are normal random variables with a BVN joint density function.

Theorem 3.5 *If X and Y have a bivariate normal joint distribution, then the variable $X|Y$ is distributed $N\left(\mu_X + \rho\left(\frac{\sigma_X}{\sigma_Y}\right)(y-\mu_Y), \sigma_X^2\left(1-\rho^2\right)\right)$*

Proof. *Recall the notation of $W = (x-\mu_X)/\sigma_X$ and $V = (y-\mu_Y)/\sigma_Y$.*

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(\frac{-\left(W^2 - 2\rho WV + V^2\right)}{2\left(1-\rho^2\right)}\right) \tag{3.34}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(\frac{-1}{2}V^2\right) \tag{3.35}$$

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} \tag{3.36}$$

$$= \frac{\sqrt{2\pi\sigma_Y^2}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \frac{\exp\left(\frac{-1}{2}\frac{\left(W^2-2\rho WV+V^2\right)}{\left(1-\rho^2\right)}\right)}{\exp\left(\frac{-1}{2}V^2\right)} \tag{3.37}$$

$$= \frac{1}{\sqrt{2\pi\sigma_X^2\left(1-\rho^2\right)}} \exp\left(\frac{-1}{2}\left(\frac{\left(W^2-2\rho WV+V^2\right)}{\left(1-\rho^2\right)} - V^2\right)\right) \tag{3.38}$$

$$= \frac{1}{\sqrt{2\pi\sigma_X^2\left(1-\rho^2\right)}} \exp\left(\frac{-1}{2}\left(\frac{W^2-2\rho WV+\rho^2V^2}{\left(1-\rho^2\right)}\right)\right) \tag{3.39}$$

$$= \frac{1}{\sqrt{2\pi\sigma_X^2\left(1-\rho^2\right)}} \exp\left(\frac{-1}{2}\left(\frac{W-\rho V}{\sqrt{1-\rho^2}}\right)^2\right) \tag{3.40}$$

$$= \frac{1}{\sqrt{2\pi\sigma_X^2\left(1-\rho^2\right)}} \exp\left(\frac{-1}{2}\left(\frac{\frac{x-\mu_X}{\sigma_X} - \rho\frac{y-\mu_Y}{\sigma_Y}}{\sqrt{1-\rho^2}}\right)^2\right) \tag{3.41}$$

$$= \frac{1}{\sqrt{2\pi\sigma_X^2\left(1-\rho^2\right)}} \exp\left(\frac{-1}{2}\left(\frac{x-\left(\mu_X+\rho\left(\frac{\sigma_X}{\sigma_Y}\right)(y-\mu_Y)\right)}{\sigma_X\sqrt{1-\rho^2}}\right)^2\right) \tag{3.42}$$

$$= f(x|y) \sim N\left(\mu_X + \rho\left(\frac{\sigma_X}{\sigma_Y}\right)(y-\mu_Y), \sigma_X^2\left(1-\rho^2\right)\right) \tag{3.43}$$

■

3.2.4 Useful Distribution Identities

Theorem 3.6 *If $\{X_i\}_{i=1}^n$ is a sequence of normal random variables with $X_i \sim N\left(\mu_i, \sigma_i^2\right)$, then $\left(\sum_{i=1}^n X_i\right) \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$*

3.3 Transformations (Functions of Random Variables)

3.3.1 Single Variable Transformations

Given some random variable X, we may wish to study some function $g(X)$. This function will itself be a random variable $Y = g(X)$. To see this, recall Definition 1.20 that defined X as a transformation from $(\Omega, \mathcal{E}, \mathbb{P})$ into $(\mathcal{X}$,$\mathcal{A}, \mathbb{P}_X)$. Y is therefore a transformation from $(\mathcal{X}$,$\mathcal{A}, \mathbb{P}_X)$ into $(\mathcal{Y}$,$\mathcal{F}, \mathbb{P}_Y)$, where $\mathcal{Y} \subseteq \mathbb{R}$ is the range of Y, $\mathcal{F}$ is a Borel field of $\mathcal{Y}$, and $\mathbb{P}_Y$ is the induced probability measure over $\mathcal{F}$. For this to be of any use, we need to discover how the probability measure $\mathbb{P}_Y$ relates to the measure $\mathbb{P}_X$.

For some $F \in \mathcal{F}$, we have that $\mathbb{P}_Y[F] = \mathbb{P}_X\{x \in \mathcal{X} : Y(x) \in F\} = \mathbb{P}\{\omega \in \Omega : g\,(X(\omega)) \in F\} =$

$\mathbb{P}\{\omega \in \Omega : X(\omega) \in g^{-1}(F)\} = \mathbb{P}_X[g^{-1}(F)]$

With non-transformed variables, we step "backwards" from the values of the random variable to the set of events in Ω. In the transformed case, you have to make two steps "backward" - once from the range of the transformation back to the values of the original random variable, and then again back to the set of events in Ω. The only problem one might encounter in this process is going backwards through the transformation $g(x)$ (which means you need to work with $g^{-1}(x)$) will not give you unique results if $g(X)$ is not monotonic (or, "1 to 1 and onto".) We will first look at the case where $g(X)$ is monotonic.

Theorem 3.7 *For some transformation $Y = g(X)$, where $g(X)$ is a differentiable function that is strictly monotonic, we have that*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{\partial}{\partial y} g^{-1}(y) \right| \tag{3.44}$$

*where $\left| \frac{\partial}{\partial y} g^{-1}(y) \right|$ is referred to as the **Jacobian** of the transformation.*

Proof. *First we will look at the cdf of Y and then differentiate to find the pdf.*

$$F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[g(X) \leq y] = \left\{ \begin{array}{c} \mathbb{P}[X \leq g^{-1}(y)] \text{ if } \frac{d}{dx}g(X) > 0 \\ \mathbb{P}[X \geq g^{-1}(y)] \text{ if } \frac{d}{dx}g(X) < 0 \end{array} \right.$$

$$= \left\{ \begin{array}{c} F_X[g^{-1}(y)] \text{ if } \frac{d}{dx}g(X) > 0 \\ 1 - F_X[g^{-1}(y)] \text{ if } \frac{d}{dx}g(X) < 0 \end{array} \right.$$

$$f_Y(y) = \left\{ \begin{array}{c} f_X[g^{-1}(y)]\frac{d}{dy}g^{-1}(y) \text{ if } \frac{d}{dx}g(X) > 0 \\ -f_X[g^{-1}(y)]\frac{d}{dy}g^{-1}(y) \text{ if } \frac{d}{dx}g(X) < 0 \end{array} \right.$$

$$= f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

The reasoning for the last equality is that if $g(X)$ is monotonic $\frac{d}{dx}g(X) > 0$, then $\frac{d}{dy}g^{-1}(y) > 0$ and if $\frac{d}{dx}g(X) < 0$, then $\frac{d}{dy}g^{-1}(y) < 0$. If this isn't intuitive, think of it in terms of correlations. If $Y = g(X)$ is increasing in X, then higher values of X yield higher values of Y. Therefore, going backwards, higher values of Y give higher values of X. The symmetric argument holds for decreasing functions. The correlation between the variables is maintained through inverting the function. ■

To look at transformations more generally, we still require that $g(x)$ be invertible, but we now do not require it to be monotonic over its entire domain. Instead, we only require it to be piecewise strictly monotonic, which means that the domain of the transformation (usually the real line) can be partitioned into intervals over which the transformation is strictly monotonic.

A good example is the 3^{rd} order polynomial $g(x) = \frac{3}{4}x^3 + 2x^2 - x + 20$, which looks like:

REMOVED

Note that $g'(x) = \frac{9}{4}x^2 + 4x - 1$, which has roots at $\{-2, \frac{2}{9}\}$. Therefore, the transformation $g(x)$ is monotonic over the intervals $\{(-\infty, -2), [-2, \frac{2}{9}], (\frac{2}{9}, \infty)\}$, but is not monotonic over its entire domain. Note that these three intervals form a partition of the real line. To look at the probability measure over the values of $g(X)$, we must divide it into this partition in order to get unique results. Before proceeding, we first introduce the notion of an indicator function.

Definition 3.7 *An **indicator function** is a function evaluated over a statement S, where*

$$\chi_{\{S\}} = \begin{cases} 1\, if\ S\ is\ true \\ 0\, if\ S\ is\ false \end{cases} \tag{3.45}$$

Example 3.8 $\chi_{\{x\leq 2\ \&\ y>4\}} = \begin{cases} 1\ if\ x \leq 2\ \&\ y > 4 \\ 0\ if\ x > 2\ or\ y \leq 4 \end{cases}$

Theorem 3.9 *The expected value of an indicator function containing a random variable is equal to the probability that the condition inside the indicator function is true.*

Proof. *Using the definition of the expectation, we have*

$$\mathbb{E}[\chi_{\{X\in E\}}] = \int_{-\infty}^{\infty} \chi_{\{x\in E\}} f_X(x)dx = \int_{x\in E} 1 f_X(x)dx + \int_{x\notin E} 0 f_X(x)dx =$$
$$\int_{x\in E} f_X(x)dx = \mathbb{P}[X \in E]$$

■

Continuing, we define a partition of $\mathcal{X} = \{A_1, A_2, ...\}$ such that $g(X)$ is monotonic over $A_i \forall i$. We now define $g_i(X) = g(X)\chi_{\{x \in A_i\}}$. Therefore,

$$g(X) = g_1(X) + g_2(X) + ... \tag{3.46}$$

Now, to look at the distribution of $Y = g(X)$, consider the cdf.

$$\begin{aligned} F_Y(y) = \mathbb{P}_Y[Y \leq y] = \mathbb{P}_Y[g(X) \leq y] &= \mathbb{P}_Y[g_1(X) \leq y] + \mathbb{P}_Y[g_2(X) \leq y] + ... \\ &= \mathbb{P}_X[X \leq g_1^{-1}(y)] + \mathbb{P}_X[X \leq g_2^{-1}(y)] + ... \end{aligned}$$

We've now transformed the problem into one that looks like the globally monotonic transformation problem. Using the results from that theorem gives us the following.

Theorem 3.10 *If $X \sim F_X(x)$ and $Y = g(X)$, where $g(X)$ is piecewise monotonic over the partition $\{A_1, A_2, ...\}$ of $\mathcal{X}$, $g_i^{-1}(y)$ is differentiable on y, and $\mathcal{Y}_i = \{y : g^{-1}(y) \in A_i\}$, then*

$$f_Y(y) = \sum_{i=1}^{\infty} f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right| \chi_{\{y \in \mathcal{Y}_i\}} \tag{3.47}$$

Remark 3.11 *Note that $\{\mathcal{Y}_i\}_{i=1}^{\infty}$ does not form a proper partition of $\mathcal{Y}$ since it may be true that $\mathcal{Y}_i \cap \mathcal{Y}_i \neq \emptyset$. However, it will be true that $\cup_{i=1}^{\infty} \mathcal{Y}_i = \mathcal{Y}$.*

Example 3.12 *Let $X \sim U[-2, 2]$ and $Y = X^2\chi_{\{X \leq 0\}} + \log[X]\chi_{\{X > 0\}}$.*

We must be careful with this transformation since $\exists y \in [-2, 4]$, $x \in [-2, 2] \ni \log[x] = y$ & $x^2 = y$. Therefore we must partition the domain of the transformation. Although we don't have an identification problem until $x \geq 1$ (where $\log[x] \geq 0$), we do have a monotonicity problem if we look at $[-2, 1]$. Therefore, we must partition the domain into $\{[-2, 0), [0, 2]\}$.

Over $[-2, 0)$ we have $Y = g(X) = X^2$. Therefore, $g^{-1}(Y) = -\sqrt{Y}$ (negative since we're restricted to $X < 0$.) Therefore, $\frac{dg^{-1}}{dy} = -Y^{-1/2}$. We know that since $X \sim U[-2, 2]$, the pdf will be $f_X(x) = \frac{1}{2+2} \forall x \in [-2, 2]$. Therefore $f_X(g_1^{-1}(y)) \left| \frac{d}{dy} g_1^{-1}(y) \right| \chi_{\{y \in \mathcal{Y}_1\}} = \frac{1}{4} \left| -y^{-1/2} \right| \chi_{\{y \in (0,4]\}}$. Similarly, over $[0, 2]$, we have $f_X(g_2^{-1}(y)) \left| \frac{d}{dy} g_2^{-1}(y) \right| \chi_{\{y \in \mathcal{Y}_2\}} = \frac{1}{4} \left| e^y \right| \chi_{\{y \in (-\infty, \log[2]]\}}$.

Therefore, we have our result that $f_Y(y) = \frac{1}{4}\left(-y^{-1/2}\chi_{\{0<y\leq 4\}} + e^y\chi_{\{-\infty<y\leq \log[2]\}}\right)$.

To double-check that this is in fact a proper pdf, integrate it out.

$$\begin{aligned} &\int_{-\infty}^{\infty} \frac{1}{4}\left(-y^{-1/2}\chi_{\{0<y\leq 4\}} + e^y\chi_{\{-\infty<y\leq \log[2]\}}\right) = \\ &\qquad \frac{1}{4}\left(\int_0^4 -y^{-1/2}dy + \int_{-\infty}^{\log[2]} e^y dy\right) = \frac{1}{4}\left((2-0) + (2-0)\right) = 1, \textit{ QED} \end{aligned}$$

3.3.2 Bivariate Transformations

blah - this is a very big blah since bivariate transformations are both difficult and important!

Chapter 4

Famous Inequalities

The following inequalities are all commonly used in various proofs. The proofs of these inequalities also serve as instructive exercises for the reader.

4.1 Bonferroni's Inequality

First, recall this basic inequality of probability theory that first appeared on page 21.

Theorem 4.1 $\mathbb{P}[A \cap B] \geq \mathbb{P}[A] + \mathbb{P}[B] - 1$

4.2 A Useful Lemma

Lemma 4.2 *If* $\frac{1}{p} + \frac{1}{q} = 1$*, then* $\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$

Remark 4.3 *Almost all of the following inequalities are consequences of this lemma.*

4.3 Holder's Inequality

Theorem 4.4 *For* p, q *satisfying Lemma 4.2*

$$|\mathbb{E}[XY]| \leq \mathbb{E}\,|XY| \leq \left(\mathbb{E}\,|X|^p\right)^{1/p} \left(\mathbb{E}\,|Y|^q\right)^{1/q} \tag{4.1}$$

4.4 Cauchy-Schwarz Inequality

Theorem 4.5 *(Holder's inequality with* $p = q = 2$*)*

$$|\mathbb{E}[XY]| \leq \mathbb{E}\,|XY| \leq \sqrt{\mathbb{E}\,|X|^2\,\mathbb{E}\,|Y|^2} \tag{4.2}$$

4.5 Covariance Inequality

Theorem 4.6 *(application of Cauchy-Schwarz)*

$$\mathbb{E}\left|(X-\mu_X)(Y-\mu_Y)\right| \leq \sqrt{\mathbb{E}(X-\mu_X)^2\mathbb{E}(Y-\mu_Y)^2} \tag{4.3}$$
$$(\mathrm{Cov}[X,Y])^2 \leq \sigma_X^2\sigma_Y^2$$

4.6 Markov's Inequality

Theorem 4.7 *If $\mathbb{E}\left[X\right] < \infty$ and $t > 0$ then*

$$\mathbb{P}\left[|X| \geq t\right] \leq \frac{\mathbb{E}\left[|X|\right]}{t} \tag{4.4}$$

4.7 Jensen's Inequality

Theorem 4.8 *If $g(x)$ is convex, then*

$$\mathbb{E}[g(x)] \geq g(\mathbb{E}[x]) \tag{4.5}$$

If $g(x)$ is concave, then

$$\mathbb{E}[g(x)] \leq g(\mathbb{E}[x]) \tag{4.6}$$

Equality holds iff, for every line $a+bx$ that is tangent to $g(x)$ at $x=\mathbb{E}[x]$, $\mathbb{P}[g(x)=a+bX]=1$

4.8 Chebyshev's Inequality

Theorem 4.9 $\mathbb{P}\left[\mu_X - k\sigma_X \leq X \leq \mu_X + k\sigma_X\right] \geq 1 - \frac{1}{k^2}$

This theorem sets a weak lower bound on the probability that a random variable falls within a certain confidence interval. For example, the probability that X falls within two standard deviations of its mean is at least $1-1/4 = 3/4$. Note that the Chebyshev inequality usually undershoots the actual probability by a long measure. In the normal distribution, the probability of a random variable falling within two standard deviations of its mean is 0.95.

Part II

Statistics

Chapter 5

Parameter Estimation

5.1 Statistics - Definitions and Properties

Any time you have an unknown parameter of a model, a distribution, or whatever, if you can observe data that gives you information about that unknown parameter, then you can estimate it.

A **statistic** is any function of observed data. A sample mean ($\sum x_i/n$) is a good example of a statistic. Since data is all we can work with, we form statistics to estimate our unknown parameters. Really, any statistic can claim to be an estimate of any parameter. For example, $(x_1 + x_5 - 7)$ is by definition a statistic and we could claim that it estimates the population mean of the variable y. However, this is probably a really bad estimate! Therefore, we'd like our estimators to have certain desirable properties.

Definition 5.1 *A statistic $T(X)$ is* ***sufficient*** *for a parameter θ if the conditional distribution of the sample taken given the value $T(X)$ does not depend on θ. In other words, all information in the sample based on θ is captured by $T(X)$.*

Theorem 5.1 *(**Sufficiency Principle**) If $T(X)$ is sufficient for θ, then any inference about θ using the sample X will depend only on $T(X)$ and not on θ.*

Definition 5.2 *An estimator $\hat{\theta}$ of θ is* ***unbiased*** *if $\mathbb{E}\left[\hat{\theta}\right] = \theta$.*

Definition 5.3 *An estimator $\hat{\theta}$ of θ is* ***consistent*** *if $\operatorname{plim} \hat{\theta}_n = \theta$ (see Section 7.1 for details)*

Remark 5.2 *Note that you can define* ***asymptotic unbiasedness*** *as $\lim_{n\to\infty} \mathbb{E}\left[\hat{\theta}_n\right] = \theta$, which is different than consistency. However, consistency is more common, and the two are roughly substitutes.*

Definition 5.4 *An estimator is* ***efficient*** *if the variance of the estimator is minimized.*

Definition 5.5 *An estimator is* ***asymptotically efficient*** *if the* $\text{Var}\left[\hat{\theta}\right] \xrightarrow{p}$ *CRLB (the Cramer-Rao Lower Bound, defined in Theorem 5.4.)*

Definition 5.6 *An estimator is* ***BUE****, or* ***B****est* ***U****nbiased* ***E****stimate, if it is the estimator with the smallest variance among all unbiased estimates.*

Definition 5.7 *An estimator is* ***BLUE****, or* ***B****est* ***L****inear* ***U****nbiased* ***E****stimate, if it is the linear estimate with the smallest variance among all unbiased linear estimates.*

One very popular estimation technique is the Ordinary Least Squares regression. This process (described in detail later) provides the most efficient linear unbiased estimates (under certain assumptions) of the coefficients that would form the best linear relationship (i.e., closest in the average squared distance from the actual relationship) between a given variable and a given set of possibly related variables.

5.2 Method of Moments

If your underlying distribution is described by k parameters, generate sample analogues of each of the first k moments of the distribution. This gives you k equations and k unknowns. For example, normal distributions are specified by two parameters, μ and σ^2. So, the equations are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \tag{5.1}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \tag{5.2}$$

5.3 Maximum Likelihood Estimation (MLE)

MLE is a very useful estimation technique because it is relatively simple to use, has nice properties under fairly general assumptions, and can be applied to virtually any estimable parameter.

To form an MLE of a parameter θ, we will first gather a set of data that gives us information about θ. We than ask what the probability is that a certain value of θ would have generated the data we've gathered. The value of θ that is the most likely to have given us our data is our maximum likelihood estimate.

To find this value, we must first form the probability that the value of the underlying parameter was θ. This is a conditional probability - conditional on the observed data. This forms the **likelihood function** L.

$$L(\theta) = \Pr[\theta = c | X = x] = f(\theta | X) \tag{5.3}$$

If the data are **iid** (independent and identically distributed), then the probability is just

$$L(\theta) = \prod_{i=1}^{n} \Pr\left[X_i = x_i|\theta\right] = \prod_{i=1}^{n} f_{X_i}(x_i|\theta) \tag{5.4}$$

where X_i is a random variable and x_i is our i^{th} observation of that random variable.

Our goal now is to maximize $L(\theta)$ across all possible values of θ. However, in the iid case (which is very common,) we have a large product. The solution is to transform the maximization objective function using logarithms. Since logarithms are monotonic, maximal points are invariant to the log transformation. So, if $x = \arg\max\{f(x)\}$, then $x = \arg\max\{\log[f(x)]\}$. To prove this, notice that the first and second order conditions for maximization must hold for the first equation and then verify that they will not change under the log transformation.

Therefore, we define the log-likelihood **function** $\mathcal{L}$ to be

$$\mathcal{L}(\theta) = \log[L(\theta)] = \sum_{i=1}^{n} \log\left[f_{X_i}(x_i|\theta)\right] \tag{5.5}$$

The first-order condition for maximization will be

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{f_{X_i}(x_i|\theta)} \frac{\partial f_{X_i}(x_i|\theta)}{\partial \theta} = 0 \tag{5.6}$$

The solution to this maximization will be our solution $\hat{\theta}_{MLE}$.

Example 5.3 *Some example...*

5.3.1 Properties of ML Estimators

The following properties require some regularity conditions which are omitted here.

1. **Consistent** – $\operatorname{plim} \hat{\theta}_{MLE} = \theta$

2. **Asymptotically Normal and Efficient** – $\hat{\theta}_{MLE} \xrightarrow{d} N\left[\theta, I\left(\theta\right)^{-1}\right]$, where $I\left(\theta\right)$ is the information matrix defined in Definition 5.10.

3. MLE's have the **invariance property** - if $\hat{\theta}$ is the MLE of θ, then $T(\hat{\theta})$ is the MLE of $T(\theta)$ for any function T.

5.4 Bayes Estimation

In Bayes estimation, we have some **prior distribution** on the parameter, $\pi(\theta)$ which is known before any data is gathered. Data is then gathered and the **posterior distribution** of the parameter, $\pi(\theta|x)$ is calculated through the following "updating" process:

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{f_X(x)} = \frac{f(x,\theta)}{f_X(x)} \tag{5.7}$$

There is a good example of Bayes estimation given as a practice problem in the appendix.

5.5 The Cramer-Rao Lower Bound

In evaluating our estimates, we should (as a rough rule) require consistency, strive for unbiasedness, and hope for the estimate's variance to be as small as possible. Consistency and unbiasedness are generally easy to check by taking the plim or expected value, respectively. However, we would like to get some idea about what kind of estimate variance is acceptable.

The intuitive answer to the estimator variance question is that we'd like to define a class of acceptable estimators and then find the estimator in that class that has the smallest variance. We usually restrict our attention to the class unbiased estimates, so we will be searching for the **uniform minimum variance unbiased estimate** (**UMVUE**.) Formally,

Definition 5.8 *An estimator W^* is the UMVUE of the function $\tau(\theta)$ if $\mathbb{E}_\theta[W^*] = \tau(\theta)\ \forall\theta$ (unbiasedness) and* $\mathrm{Var}_\theta[W^*] \leq \mathrm{Var}_\theta[W]\ \forall W \ni \mathbb{E}_\theta[W] = \tau(\theta)\ \forall\theta$.

The UMVUE definition is unsatisfying because it implicitly requires us to test our estimator against all other possible unbiased estimators. So, we now turn to defining what the minimum variance might look like for a given estimator. If we can establish a lower bound on the variance possible, then we know that any unbiased estimator with that variance will not be "beaten" by some other unbiased estimate that has lower variance. This lower bound is well defined and is known as the **Cramer-Rao Lower Bound**.

Theorem 5.4 ***(Cramer-Rao Lower Bound)*** *Let $X_1, ..., X_n$ be a sample with pdf $f(x|\theta) = f(x_1, ..., x_n|\theta)$ and let $W(x) = W(x_1, ..., x_n)$ be any estimator such that*

$$\frac{d}{d\theta}\mathbb{E}_\theta[W(x)] = \int_X W(x)\left(\frac{\partial}{\partial\theta}f(x|\theta)\right)dx \tag{5.8}$$

and $\mathrm{Var}_\theta\left[W\left(x\right)\right] < \infty$*, then*

$$\mathrm{Var}_\theta\left[W\left(x\right)\right] \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta\left[W(x)\right]\right)^2}{\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\left[f\left(x|\theta\right)\right]\right)^2\right]} \tag{5.9}$$

Proof. *The proof of this important theorem uses the Cauchy-Schwarz Inequality, which states that* $\mathrm{Cov}\left[X,Y\right]^2 \leq \mathrm{Var}\left[X\right]\mathrm{Var}\left[Y\right]$.

Note that

$$\frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right] = \frac{1}{f(x|\theta)}\left(\frac{\partial}{\partial\theta}f(x|\theta)\right) = \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)} \tag{5.10}$$

By assumption,

$$\frac{d}{d\theta}\mathbb{E}_\theta\left[W\left(x\right)\right] = \int_X W\left(x\right)\left(\frac{\frac{\partial}{\partial\theta}f\left(x|\theta\right)}{f(x|\theta)}\right)f(x|\theta)dx = \mathbb{E}_\theta\left[W\left(x\right)\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right]$$

$$\overset{by\ Eq.\ 5.10}{=} \mathbb{E}_\theta\left[W\left(x\right)\frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right]$$

Note that

$$\mathrm{Cov}_\theta\left[W\left(x\right), \frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right] = \mathbb{E}_\theta\left[W\left(x\right)\frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right] - \mathbb{E}_\theta\left[W\left(x\right)\right]\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right] \tag{5.11}$$

We now focus on calculating the two expectations. If we refer to the above assumption and let $W(x) = 1$*, then*

$$\mathbb{E}_\theta\left[(1)\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right] = \mathbb{E}_\theta\left[(1)\frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right] = \frac{d}{d\theta}\mathbb{E}_\theta\left[(1)\right] = 0 \tag{5.12}$$

Therefore,

$$\mathrm{Cov}_\theta\left[W\left(x\right), \frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right] = \mathbb{E}_\theta\left[W\left(x\right)\frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right] = \frac{d}{d\theta}\mathbb{E}_\theta\left[W\left(x\right)\right] \tag{5.13}$$

Since $\mathrm{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$ *and* $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right] = 0$ *(by Equation 5.12) we can say that*

$$\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right)^2\right] - 0 \tag{5.14}$$

Rearranging the Cauchy-Schwarz Inequality from above gives

$$\mathrm{Var}\left[X\right] \geq \frac{\mathrm{Cov}\left[X,Y\right]^2}{\mathrm{Var}\left[Y\right]} \tag{5.15}$$

Here, let $X = W(x)$ and $Y = \frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]$ to get

$$\operatorname{Var}\left[W(x)\right] \geq \frac{\operatorname{Cov}\left[W(x), \frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right]^2}{\operatorname{Var}\left[\frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right]} \tag{5.16}$$

Plugging equations 5.13 and 5.14 into this inequality gives the result:

$$\operatorname{Var}_\theta\left[W(x)\right] \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta\left[W(x)\right]\right)^2}{\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\left[f(x|\theta)\right]\right)^2\right]} \tag{5.17}$$

■

5.6 Score, Information Matrix, & Information Equality

These concepts are nothing substantially new at this point. They are simply new names referring to values and functions we've already seen. However, since they appear time and time again, they are given special names.

Definition 5.9 *The **score** at a parameter value $\tilde{\theta}$ is defined as the gradient of the log-likelihood function $\mathcal{L}(\tilde{\theta})$ and is denoted $s(\tilde{\theta}) \equiv \frac{\partial}{\partial\tilde{\theta}}\mathcal{L}(\tilde{\theta})$.*

Definition 5.10 *The **information matrix** is defined as $I(\theta) \equiv \mathbb{E}\left[s(\theta)s(\theta)'\right]$, where the score is evaluated at the true parameter θ.*

Lemma 5.5 *The expected value of the score evaluated at the true parameter value is zero, or $\mathbb{E}_\theta\left[s(\theta)\right] = 0$*

Proof. *Let $f(x|\theta)$ be the multinomial pdf of the observed sample, which is also the likelihood function. Then*

$$\mathbb{E}\left[s(\theta)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\ln\left[f(x|\theta)\right]\right] = \mathbb{E}\left[\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right] = \int_{\tilde{\theta}}\frac{\partial}{\partial\theta}f(x|\theta)d\tilde{\theta} =$$

$$\frac{\partial}{\partial\theta}\int_{\tilde{\theta}}f(x|\theta)d\tilde{\theta} = \frac{\partial}{\partial\theta}(1) = 0$$

Trivia Question: What regularity conditions have we assumed here[1]? Are they valid? ■

[1] See Section 2.1.2 for the answer.

Theorem 5.6 ***(Information Equality)*** *Under regularity conditions*[2]*, the information matrix is equal to the negative of the expected value of the Hessian (matrix of 2^{nd} partials) of the log likelihood function:*

$$I(\theta) = -\mathbb{E}_\theta \left[H \left(\mathcal{L} \left(\theta \right) \right) \right] = -\mathbb{E}_\theta \left[\frac{\partial^2 \mathcal{L}(\theta)}{\partial \tilde{\theta} \, \partial \tilde{\theta}'} \right] \tag{5.18}$$

Proof. *The proof will be straightforward. We will work "right to left." Note that we are working with vectors, so we will often see terms of the form XX' instead of the scalar-equivalent X^2.*

$$\frac{\partial^2 \mathcal{L}(\theta)}{\partial \tilde{\theta} \, \partial \tilde{\theta}'} = \frac{\partial}{\partial \tilde{\theta}} \left(\frac{\partial}{\partial \tilde{\theta}} \ln \left[f(x|\theta) \right] \right) = \frac{\partial}{\partial \tilde{\theta}} \left(\frac{\frac{\partial}{\partial \tilde{\theta}} f\left(x|\theta\right)}{f\left(x|\theta\right)} \right) =$$

$$\frac{-\left(\frac{\partial}{\partial \tilde{\theta}} f\left(x|\theta\right) \right) \left(\frac{\partial}{\partial \tilde{\theta}} f\left(x|\theta\right) \right)'}{f\left(x|\theta\right) \; f\left(x|\theta\right)'} + \frac{\frac{\partial^2}{\partial \tilde{\theta} \, \partial \tilde{\theta}'} f\left(x|\theta\right)}{f(x|\theta)}$$

Taking the negative of the expectation over θ gives

$$-\mathbb{E}_\theta \left[\frac{-\left(\frac{\partial}{\partial \tilde{\theta}} f\left(x|\theta\right) \right) \left(\frac{\partial}{\partial \tilde{\theta}} f\left(x|\theta\right) \right)'}{f\left(x|\theta\right) \; f\left(x|\theta\right)'} \right] - \mathbb{E}_\theta \left[\frac{\frac{\partial^2}{\partial \tilde{\theta} \, \partial \tilde{\theta}'} f\left(x|\theta\right)}{f(x|\theta)} \right] =$$

$$\mathbb{E}_\theta \left[\left(\frac{\frac{\partial}{\partial \tilde{\theta}} f\left(x|\theta\right)}{f\left(x|\theta\right)} \right) \left(\frac{\frac{\partial}{\partial \tilde{\theta}} f\left(x|\theta\right)}{f\left(x|\theta\right)} \right)' \right] - \int_{\tilde{\theta}} \frac{\frac{\partial^2}{\partial \tilde{\theta} \, \partial \tilde{\theta}'} f\left(x|\theta\right)}{f\left(x|\theta\right)} f\left(x|\theta\right) d\theta =$$

$$\mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \tilde{\theta}} \ln \left[f(x|\theta) \right] \right) \left(\frac{\partial}{\partial \tilde{\theta}} \ln \left[f(x|\theta) \right] \right)' \right] - \frac{\partial^2}{\partial \tilde{\theta} \, \partial \tilde{\theta}'} \int_{\tilde{\theta}} f\left(x|\theta\right) d\theta =$$

$$\mathbb{E}_\theta \left[s(\theta) \; s(\theta)' \right] - \frac{\partial^2}{\partial \tilde{\theta} \, \partial \tilde{\theta}'} (1) = I\left(\theta\right) - 0 = I\left(\theta\right)$$

■

We can use these new terms to simplify the statement of the Cramer-Rao Lower Bound.

Theorem 5.7 ***(Cramer-Rao Lower Bound)*** *Let $X_1, ..., X_n$ be a sample with pdf $f\left(x|\theta\right) = f(x_1, ..., x_n|\theta)$ and let $W\left(x\right) = W\left(x_1, ..., x_n\right)$ be any estimator*[3] *such that*

$$\frac{d}{d\theta} \mathbb{E}_\theta \left[W\left(x\right) \right] = \int_X W\left(x\right) \left(\frac{\partial}{\partial \theta} f\left(x|\theta\right) \right) dx \tag{5.19}$$

and $\mathrm{Var}_\theta \left[W\left(x\right) \right] < \infty$, then

$$\mathrm{Var}_\theta \left[W\left(x\right) \right] \geq I\left(\theta\right)^{-1} \tag{5.20}$$

[2] These conditions guarantee that $E\left[\partial L(\theta)/\partial \theta\right] = \partial E\left[L(\theta)\right]/\partial \theta$.

[3] The CRLB seems to only apply to unbiased estimates. However, to "unbias" any statistic, subtract its bias and use the result as the new statistic.

So, the result is that the Cramer-Rao lower bound is the inverse of the information matrix.

Chapter 6

Hypothesis Testing

6.1 Overview

For the uninitiated, an hypothesis test is simply a test of a statement. Hypothesis tests always contain some null hypothesis $\mathbf{H}_0$ (usually the thing you hope *isn't* true) and an alternative hypothesis $\mathbf{H}_1$. It should be the case that $\mathbf{H}_1$ and $\mathbf{H}_0$ be logically mutually exclusive. In other words, if one of the statements if false, then the other must be true.

In most cases, tests are designed such that the null hypothesis is some very specific situation and the alternative hypothesis is just the compliment of the null. For example, if we are concerned that some parameter β might be exactly equal to 1, then our null hypothesis would be "$\mathbf{H}_0 : \beta = 1$". This leaves our alternative hypothesis to be "$\mathbf{H}_1 : \beta \neq 1$".

Since the null hypothesis is a very specific condition, our goal isn't to prove it to be true. In the above example, the event that β is *exactly equal* to 1 is a zero-probability event, in the sense that β could be any of an infinite number of values very close to 1, while 1 itself is just a single point. So, the correct approach is to try to statistically reject the null hypothesis. If we find strong evidence that β is almost certainly not 1 (or, in a neighborhood of 1,) then we can reject $\mathbf{H}_0$. If we are unable to make such a claim, we "Do Not Reject" (DNR) the null hypothesis. This distinction is somewhat trivial, but if you claim that you accept the null hypothesis to be true, then you claim to be certain that β is equal to 1 but definitely not equal to 1.00001, for example. Regardless, many scholars still ignore this distinction and either "accept" or "reject" the null hypothesis.

We will delve into hypothesis tests in more detail in their specific applications. For example, in Section 8.8 we examine various hypothesis tests available for ordinary least squares regressions.

Chapter 7

Large-Sample Results

7.1 Introduction

Properties of estimates such as unbiasedness are known as small-sample properties because they apply for reasonably small sample sizes. However, there often arise cases where the small-sample properties are unknown. In these cases, we can often derive some large-sample results by looking at what *would* happen if we had near-infinite sample sizes.

7.2 Notions of Convergence

Definition 7.1 *A random variable X_n* ***converges in probability*** *to a constant c if $\lim_{n\to\infty} \mathbb{P}\left[|X_n - c| > \varepsilon\right] = 0 \ \forall \varepsilon > 0$. We denote this by $\operatorname{plim} X_n = c$ and c is called the "probability limit" of X_n. An alternative notation is $X_n \xrightarrow{p} c$.*

We can use Chebychev's Inequality (Theorem 4.9) to derive a large-sample equivalent to the theorem.

Theorem 7.1 *If X_n is a random variable and c is a constant, then $\mathbb{P}\left[|X_n - c| > \varepsilon\right] \leq \mathbb{E}\left[(X_n - c)^2\right]/\varepsilon^2$*

The proof of the above theorem actually requires the following lemma, whose proof is given as it is a clever use of conditional probability.

Lemma 7.2 *(Markov's Inequality) If Y_n is a non-negative random variable and δ is a constant, then $\mathbb{P}\left[Y_n \geq \delta\right] \leq \mathbb{E}\left[Y_n\right]/\delta$*

Proof. *$\mathbb{E}\left[Y_n\right] = \mathbb{P}\left[Y_n < \delta\right]\mathbb{E}\left[Y_n | Y_n < \delta\right] + \mathbb{P}\left[Y_n \geq \delta\right]\mathbb{E}\left[Y_n | Y_n \geq \delta\right]$. Since $Y_n > 0$, all expected values of Y_n will be nonnegative. So, $\mathbb{E}\left[Y_n\right] \geq \mathbb{P}\left[Y_n \geq \delta\right]\mathbb{E}\left[Y_n | Y_n \geq \delta\right]$. Now note that $\mathbb{E}\left[Y_n | Y_n \geq \delta\right] \geq \delta$. So, $\mathbb{E}[Y_n] \geq \mathbb{P}\left[Y_n \geq \delta\right]\mathbb{E}\left[Y_n | Y_n \geq \delta\right] \geq \mathbb{P}\left[Y_n \geq \delta\right]\delta$. Finally, we divide both sides by δ to get our lemma.* ■

Convergence in probability is a useful property of estimates.

Definition 7.2 *An estimate $\hat{\theta}_n$ of θ is* ***consistent*** *if and only if* $\operatorname{plim} \hat{\theta}_n = \theta$.

We will show, for example, that a sample average is a consistent estimator for the population mean. Furthermore, it can be shown that

Theorem 7.3 *For any function $g(X)$, if $\mathbb{E}\left[g(X)\right]$ and $\operatorname{Var}\left[g(X)\right]$ are finite constants, then*

$$\operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} g(X_n) = \mathbb{E}\left[g(X)\right] \tag{7.1}$$

The plim operator has some very nice properties that make it easy to work with. This is partly why we can usually get consistency statements in the absence of small-sample statements. If $\operatorname{plim} X_n = c$, $\operatorname{plim} Y_n = d$, and $g(X_n)$ is a continuous function that does not depend on n, then

1. $\operatorname{plim}\left(X_n + Y_n\right) = c + d$

   The following are part of the **Slutsky Theorem**

2. $\operatorname{plim}\left(X_n Y_n\right) = cd$

3. $\operatorname{plim}\left(\frac{X_n}{Y_n}\right) = c/d \;\; (\forall d \neq 0)$

   The following is known as the **Mann-Wald Theorem**, and is often mistakenly called the Slutsky Theorem.

4. $\operatorname{plim} g\left(X_n\right) = g\left(\operatorname{plim} X_n\right) = g\left(c\right)$

   Furthermore, if W and Z are matrices, $\operatorname{plim} W_n = A$, and $\operatorname{plim} Z_n = B$, then

5. $\operatorname{plim} W_n^{-1} = A^{-1} \;\; (\forall \text{ nonsingular } A)$

6. $\operatorname{plim} W_n Z_n = AB$

Another notion of convergence is defined as follows.

Definition 7.3 *A random variable X_n* ***converges almost surely*** *to a constant c if* $\mathbb{P}\left[\lim_{n\to\infty} |X_n - c| > \varepsilon\right] = 0 \;\forall \varepsilon > 0$. *We denote this by $X_n \xrightarrow{a.s.} c$.*

Finally, we introduce our third notion of convergence.

Definition 7.4 *A random variable X_n with cdf $F_{X_n}(x)$* ***converges in distribution*** *to a random variable X^* with cdf $F_{X^*}(x)$ if $\lim_{n\to\infty} |F_{X_n}(x) - F_{X^*}(x)| = 0$ for all x where $F_{X^*}(x)$ is continuous. We denote this by $X_n \xrightarrow{d} X^*$, where $F_{X^*}(x)$ is the* ***limiting distribution*** *of X_n.*

When a variable converges in distribution to another, we can use the limiting distribution to approximate the finite distribution of the converging variable. For example, we will show that $\bar{X}_n \xrightarrow{d} N\left[\mu, \sigma^2/n\right]$. Therefore, we can approximate the distribution of the sample average $\bar{X}_n$ with the normal distribution. This is denoted as $\bar{X}_n \overset{a}{\sim} N\left[\mu, \sigma^2/n\right]$ and the normal distribution is called the **asymptotic distribution** of $\bar{X}_n$.

Some results to note are

1. $X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$, where X is a random variable
2. $X_n \xrightarrow{p} c \Leftrightarrow X_n \xrightarrow{d} c$, where c is a constant.
3. $X_n \xrightarrow{as} c \Rightarrow X_n \xrightarrow{p} c \Leftrightarrow X_n \xrightarrow{d} c$
4. If $X_n \xrightarrow{d} X$ and $\text{plim}\, Y_n = d$, then $X_n Y_n \xrightarrow{d} cX$.
5. If $X_n \xrightarrow{d} X$ and $g(X_n)$ is continuous, then $g(X_n) \xrightarrow{d} g(X)$.
6. If $Y_n \xrightarrow{d} Y$ and $\text{plim}(X_n - Y_n) = 0$, then $X_n \xrightarrow{d} Y$.

7.3 Laws of Large Numbers

From the different convergence notions come different laws of large numbers. It is somewhat important to remember that there do exist different laws of large numbers. Technically, you should always specify which you're using since they are not equivalent. They all show that sample averages converge to expected values, but each requires different assumptions or uses a different notion of convergence.

Theorem 7.4 ***(Chebychev's Weak Law of Large Numbers)*** *If $X_1, ..., X_n$ are independently drawn where each have mean μ_i, variance $\sigma_i^2 < \infty$, $\lim_{n\to\infty} \sum_{i=1}^{n} (\sigma_i/n)^2 < \infty$, and $\text{Cov}[X_i, X_j] = 0\ \forall i, j$, then $\bar{X} \xrightarrow{p} \bar{\mu}$, where $\bar{\mu} = (1/n)\sum_{i=1}^{n} \mu_i$.*

This theorem is very general since it allows for different means and different variances (heteroscedasticity,) but it does require finite variances and no covariances. Khinchine's Weak Law removes replaces some of these restrictions with an *iid* assumption, and is perhaps the most commonly used law of large numbers.

Theorem 7.5 ***(Khinchine's Weak Law of Large Numbers)*** *If $X_1, ..., X_n$ are independently drawn and identically distributed (iid) with mean μ, then $\bar{X} \xrightarrow{p} \mu$.*

We have a similar pair of Laws of Large Numbers for almost-sure convergence.

Theorem 7.6 ***(Kolmogorov's Strong Law of Large Numbers)*** *If $X_1, ..., X_n$ are independently drawn where each have mean μ_i, variance $\sigma_i^2 < \infty$, and $\sum_{i=1}^{\infty} (\sigma_i/n)^2 < \infty$, then $\bar{X} \xrightarrow{a.s.} \bar{\mu}$, where $\bar{\mu} = (1/n) \sum_{i=1}^{n} \mu_i$.*

Theorem 7.7 ***(Markov's Strong Law of Large Numbers)*** *If $X_1, ..., X_n$ are independently drawn and identically distributed (iid) with mean $\mu < \infty$ and $\exists \delta > 0 \ni \sum_{i=1}^{\infty} \mathbb{E}\left[|X_i - \mu_i|^{1+\delta}\right] / i^{1+\delta}$, then $\bar{X} \xrightarrow{a.s.} \mu$.*

7.4 Central Limit Theorems

Much like the Laws of Large Numbers, there exist various Central Limit Theorems that depend on the assumptions on the sample $X_1, ..., X_n$. However, since the theorems concern the convergence of a function of random variables to the normal distribution, we use only the convergence in distribution concept.

Theorem 7.8 ***(Univariate Linberg-Levy Central Limit Theorem)*** *If $X_1, ..., X_n$ are independently drawn and identically distributed (iid) with mean $\mu < \infty$ and variance $\sigma^2 < \infty$, then*

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N\left[0, \sigma^2\right] \tag{7.2}$$

Note that we could scale the left-hand side by σ to get a slightly more useful form

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N\left[0, 1\right] \tag{7.3}$$

Theorem 7.9 ***(Univariate Linberg-Feller Central Limit Theorem)*** *Let $X_1, ..., X_n$ be independently drawn with (possibly unique) means $\mu_i < \infty$ and variance $\sigma_i^2 < \infty$. If $\lim\limits_{n \to \infty} \left(n^{-1} \sum_{i=1}^{n} \sigma_i^2\right) < \infty$, then*

$$\sqrt{n}\left(\bar{X}_n - \bar{\mu}_n\right) \xrightarrow{d} N\left[0, \sigma^2\right] \tag{7.4}$$

Roughly equivalent statements exist for multivariate central limit theorems. We will only state the multivariate Linberg-Levy CLT here, but the extension is fairly obvious.

Theorem 7.10 ***(Multivariate Linberg-Levy Central Limit Theorem)*** *If random vectors $X_1, ..., X_n$ are independently drawn and identically distributed from a multivariate distribution with mean vector μ and finite positive definite covariance matrix Q, then*

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N\left[0, Q\right] \tag{7.5}$$

One particular generalization of the central limit theorems is the **delta method**.

Theorem 7.11 *If Y_n satisfies a central limit theorem for some θ (so, $\sqrt{n}\left(Y_n - \theta\right) \xrightarrow{d} N\left[0, \sigma^2\right]$), then for a function g such that $g'(\theta)$ exists and is nonzero, then*

$$\sqrt{n}\left(g\left(Y_n\right) - g\left(\theta\right)\right) \xrightarrow{d} N\left[0, \sigma^2\left(g'\left(\theta\right)\right)^2\right] \tag{7.6}$$

The proof of the delta method requires the Taylor series expansion of $g(Y_n)$ around $Y_n = \theta$.

There also exist central limit theorems for situations where the variables are drawn from distributions of different means and variances. These are less common.

Finally, there exists a related Theorem for the Maximum Likelihood Estimator, which is a direct consequence of property 2.

Theorem 7.12 *If $X_1, X_2, ..., X_n$ are iid $f\left(x|\theta\right)$ and $\hat{\theta}$ is the MLE estimate of θ, then (under some regularity conditions on $f\left(x|\theta\right)$)*

$$\sqrt{n}\left(\tau\left(\hat{\theta}\right) - \tau\left(\theta\right)\right) \xrightarrow{d} N\left[0, I\left[\theta\right]^{-1}\right] \tag{7.7}$$

where $I\left[\theta\right]^{-1}$ is the Cramer-Rao Lower Bound for the estimate θ.

Chapter 8

Ordinary Least Squares (OLS)

8.1 The Model

We have some **endogenous** or **dependent** variable Y which is explained by some collection of **exogenous** or **independent** variables $X_1, X_2, ..., X_k$, which we combine into columns of a matrix $\mathbf{X}$. For each variable we have T observations, so the variables are each vectors with 1 column and T rows. This means Y is a $Tx1$ vector and $\mathbf{X}$ is a Txk matrix. Finally, we admit that Y is probably not an exact linear combination of the X variables ($Y \notin Col.Space[\mathbf{X}],$) so we approximate Y by $\hat{Y} \in Col.Space[\mathbf{X}]$ plus the distance between Y and $\hat{Y}$, called the residual and denoted e.

We wish to find the coefficients on the X variables that minimize the residuals - the distance between Y and the "predicted values" $\hat{Y}$. We call these coefficients $\beta_1, \beta_2, ..., \beta_k$, and put them together in a $kx1$ vector called β. Putting this together gives our model

$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$

or, observation-by-observation,

$Y_t = \mathbf{X}_t\beta{+}\varepsilon_t$

Note that in a good textbook, vectors and matrices will be boldfaced. For simplicity, we shall drop the boldfacing shortly. In the OLS context, remember that Y_t, X_{jt}, and ε_t are scalars while Y, X_j, ε, and β are vectors and X is a Txk matrix. Also note that we use $t \in \{1, 2, ..., T\}$ to denote each observation, but it is also common to denote them as $n \in \{1, 2, ..., N\}$. The use of t may indicate observations over time (see Chapter 10) or some cross-sectional data taken at one moment in time but over different geographical reigons, for example.

8.2 Justifying the Error Term

There are four possible justifications for believing that Y is not an exact linear combination of the X variables, which is why we must introduce this error term ε.

The first is the fact that observed variables in the real world are very sensative to a large number of other factors. We would likely need a very large number of X variables to completely specify all of the factors that influence the value of Y. If we fail to include even one such X variable, then our included X variables will not perfectly predict our Y values. This is referred to as a model **misspecification**. If the effect of our misspecifications is fairly small and, in a sense, "random," then we will not have any problems in analyzing our model.

The second justification for using an error term is that the true relationship is likely to be some nonlinear equation and we are only estimating it using a linear approximation. If we knew the true nonlinear equation and all relevant X variables, we could write the model with no error term and Y would be completely determined by the X values.

Theoretical physicists and other critics of determinism reading this monograph would object to this reasoning with the belief that no real-world variables can be *completely* described by some other set of variables. Furthermore, they would argue that no real-world variables could be measured with complete accuracy. Therefore, we offer as a third and fourth justification the possibility that Y can't be an exact functional relationship of our X values and that our variables are measured with some error. In this "impossibility" setting, we can never find a model without an error term. In practice, all of these justifications are likely to be valid and we will always include error terms in our model.

If the physicists are wrong and we did find the magical deterministic equation relating our pefectly-measured Y variable to a set of perfectly-measured X variables, then our statistical analysis would find that this error term is everywhere equal to zero, thus making its inclusion (and the use of statistics as an analytical tool) harmlessly superflous.

One point of frustration is in inconsistent error-term notation across (and within) texts. It is common to use ε_t, u_t, or v_t for our error term. Like most texts, we vary our notation, although ε is the most commonly used throughout.

8.3 OLS Assumptions

The following assumptions are made when performing the OLS procedure. We will see that these assumptions appear frequently in the analysis of different problems one might encounter. Each property of the OLS estimates that we will derive are entirely dependent upon some subset of these assumptions. Therefore, as we violate an assumption, we will lose a certain set of desirable properties.

The assumptions are

1. Linearity of the Model

   The true model is $\mathbf{Y}_t = \alpha + \beta\mathbf{X}_t + \mathbf{u}_t$, which we will estimate with $\mathbf{Y}_t = \mathbf{a} + \mathbf{b}\mathbf{X}_t + \mathbf{e}_t$

2. $\mathbf{X}$ is known and nonrandom

   This implies that $\text{Cov}[\mathbf{X}_t, \mathbf{u}_t|\mathbf{X}] = \mathbb{E}[\mathbf{X}_t\mathbf{u}_t|\mathbf{X}] - \mathbb{E}[\mathbf{X}_t]\mathbb{E}[\mathbf{u}_t|\mathbf{X}] = \mathbf{X}_t\mathbb{E}[\mathbf{u}_t|\mathbf{X}] - \mathbf{X}_t\mathbb{E}[\mathbf{u}_t|\mathbf{X}] = 0$

3. No Multicollinearity

   $\mathbf{X}$ is an $n \times k$ matrix with $Rank[\mathbf{X}] = k$ (i.e., full rank,) where k is the number of exogenous (RHS) variables

   Note that in order to satisfy full rank, it must be that $n \geq k$.

   (a) $n > k$ - this modified assumption will be need for unbiased variance estimates. See Property 4 below.

4. Regression

   $\mathbb{E}[u_t|\mathbf{X}] = \mathbf{0}\ \forall t$

5. Homoscedasticity

   All u_t are identically distributed with $\text{Var}[u_t|\mathbf{X}] = \mathbb{E}[u_t^2|\mathbf{X}] = \sigma^2\ \forall t$

6. Serial Independence

   $\text{Cov}[\mathbf{u}_t, \mathbf{u}_s|\mathbf{X}] = \mathbb{E}[\mathbf{u}_t\mathbf{u}_s|\mathbf{X}] - \mathbb{E}[\mathbf{u}_t|\mathbf{X}]\mathbb{E}[\mathbf{u}_s|\mathbf{X}] = 0$ and with **A4**, we have that $\mathbb{E}[\mathbf{u}_t\mathbf{u}_s|\mathbf{X}] = 0$

7. $\mathbf{X}_t \neq \mathbf{X}_s$ for some t, s

   If all $\mathbf{X}_t$ were equal, then the regression could not be estimated. The slope of a line over a single x value is not defined (or, has infinite slope.)

8. Normality of Errors

   $\mathbf{u}_t \sim N(0, \sigma^2)\ \forall t \Rightarrow \mathbf{Y}|\mathbf{X} \sim N(\alpha + \beta\mathbf{X}, \sigma^2)$

   Note that this assumption is very strong and can often be dropped. Also note that it implies several of the other assumptions.

8.4 The OLS Solution

The goal of OLS, as the name indicates, is to find the ”least squares” estimator. The ”least squares” refers to minimizing the sum of the squared residuals. The solution is found as follows

$$\begin{aligned}
\min_\beta \sum_{i=1}^n e^2 &= \min_\beta\ (Y - X\beta)'(Y - X\beta) \\
&= \min_\beta\ (Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta) \\
&= \min_\beta\ (\beta'X'X\beta - \beta'X'Y - Y'X\beta)
\end{aligned}$$

Note that the $Y'Y$ term is irrelevant since we're maximizing over β. Now, take a minute to consider the dimensions of the terms in this last equation. Y is $nx1$, X is nxk, and β is $kx1$. Therefore, $Y'X\beta$ is $(1xn)(nxk)(kx1) = 1x1$, $\beta'X'Y$ is $(1xk)(kxn)(nx1) = 1x1$, and $\beta'X'X\beta$ is $(1xk)(kxn)(nxk)(kx1) = 1x1$. Most importantly, note that $\beta'X'Y = Y'X\beta$, so we can rewrite this minimization as

$$\min_{\beta} \left(\beta'X'X\beta - 2X'Y\beta\right) \tag{8.1}$$

Finally, recall from linear algebra that

$$\frac{\partial}{\partial\beta}\left(\beta'X'X\beta\right) = 2(X'X)\beta \tag{8.2}$$

$$\frac{\partial}{\partial\beta}\left(2X'Y\beta\right) = 2X'Y \tag{8.3}$$

So, the first order condition for our minimization problem is

$$2(X'X)\beta^* - 2X'Y = 0 \tag{8.4}$$

$$X'X\beta^* = X'Y \tag{8.5}$$

$$\beta^* = b = (X'X)^{-1}X'Y \tag{8.6}$$

The phrase "X-prime-X inverse X-primeY" should be permanently ingrained into your memory.

We now check the second order conditions.

$$\frac{\partial^2}{\partial\beta^2}\left(\beta'X'X\beta - 2X'Y\beta\right) = 2X'X > 0 \tag{8.7}$$

We know that $X'X > 0$ since $X'X$ is a kxk matrix with each term being $\sum_i X_{ik}^2$ which is nonnegative by construction. Further, the only way $X'X$ could be zero is if X is a matrix of zeros. However, this violates the assumption that not all X_{jt} are the same. Since the minimization problem is strictly convex, we know that our first order conditions imply minimization.

8.4.1 Linear Algebra Interpretation of the OLS Solution

The vector of the dependent variable, Y is located in $\mathbb{R}^n$. However, we want to explain Y as best we can using the observable independent variables $X = (\mathbf{1}, X_1, X_2, ..., X_k)$. In linear algebra terms, this means we want to project the vector Y onto the span of X, where the span is a subspace such that $span[X] = \{x : x = \lambda_0\mathbf{1} + \lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_k X_k\}$.

8.5 Properties of OLS

1. $\{A1, A2, A3\} \Rightarrow$ the LSE $\hat{\alpha}$ and $\hat{\beta}$ are unbiased - $\mathbb{E}[\mathbf{a}] = \alpha$ and $\mathbb{E}[\mathbf{b}] = \beta$.
2. $\{A1, A2, A3, A5, A6\} \Rightarrow \text{Cov}[\mathbf{b}, \mathbf{e}] = \mathbf{0}$
3. $\{A1, A2, A3, A5, A6\} \Rightarrow \text{Var}[\mathbf{b}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
4. $\{A1, A2, A3a, A5, A6\} \Rightarrow \mathbb{E}[s^2] = \sigma^2$

   Remark 8.1 *The Maximum Likelihood Estimator of σ^2, $s^2_{ML} = \frac{n-k}{n}s^2$ is biased since s^2 is unbiased. However, s^2_{ML} is consistent.*

5. $\{A2, A5, A6, A8\} \Rightarrow \varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$
6. $\{A1, A2, A3, A5, 6, A8\} \Rightarrow (\mathbf{b} - \beta) \sim N(\mathbf{0}, \sigma^2(\mathbf{X}'X)^{-1})$

   Remark 8.2 *This property is used to test $\mathbf{H}_0 : \beta_k = c \Rightarrow \beta_k - c = 0$. Simply set up*

$$z_k \equiv \frac{b_k - c}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim N(0,1) \tag{8.8}$$

$$t_k \equiv \frac{b_k - c}{SE[b_k]} = \frac{b_k - c}{\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim t_{(n-k)} \tag{8.9}$$

7. $\{A7, A4, A2\} \Rightarrow$ the LSE are consistent - $\lim_{n\to\infty} \mathbb{E}[\mathbf{b}] = \beta$, $\lim_{n\to\infty} \text{Var}[\mathbf{b}] = 0$[1]

8.5.1 Gauss-Markov Theorem

The final two properties given are perhaps the most significant in terms of evaluating the OLS estimators against any other possible estimation.

1. $\{A1, A2, A3, A5, A6\} \Rightarrow$ OLS estimators are **BLUE** - the most efficient among unbiased linear estimators. This is the Gauss-Markov Theorem.

   Theorem 8.3 *(Gauss-Markov) The least squares estimate b is the minimum variance linear unbiased estimator of β.*

   Proof. NOTE TO READER: I BELIEVE I FOUND A PROBLEM WITH THIS PROOF AT ONE TIME, BUT I DON'T REMEMBER WHAT IT WAS. READ CAREFULLY AND THINK FOR YOURSELF!!! blah

[1] My sources seem to disagree on which assumptions are actually needed to get consistent estimates. This needs to be double-checked.

This proof selects another hypothetical *linear and unbiased* estimate for β and shows that its variance must be greater than or equal to the OLS estimate for β.

If the model is indeed linear, then it is of the form $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, which can be estimated by

$$\begin{aligned} \mathbf{Y} &= \mathbf{Xb} \qquad (8.10)\\ \mathbf{X'Y} &= \mathbf{X'Xb} \\ (\mathbf{X'X})^{-1}\mathbf{X'Y} &= \mathbf{b} \\ \mathbf{PY} &= \mathbf{b} \end{aligned}$$

where $\mathbf{P}$ is a $k \times n$ projection matrix (as $\mathbf{X'X}$ is $k \times k$ and $\mathbf{X'}$ is $k \times n$). Now let us choose an arbitrary linear estimate of β, called $\mathbf{b}_0$. Let $\mathbf{b}_0 = \mathbf{CY}$. Since we assume $\mathbf{b}_0$ to be unbiased (because we're looking for the best *unbiased* estimate,) then

$$\mathbb{E}[\mathbf{CY}] = \mathbb{E}[\mathbf{CX}\beta + \mathbf{C}\varepsilon] = \beta \qquad (8.11)$$

which implies that $\mathbf{CX} = \mathbf{I}$ since $\mathbb{E}[\mathbf{C}\varepsilon] = \mathbf{0}$. The variance of our unbiased estimator will be

$$\begin{aligned} \text{Var}[\mathbf{b}_0] &= \mathbb{E}\left[(\mathbf{CY} - \beta)(\mathbf{CY} - \beta)'\right] \qquad (8.12)\\ &= \mathbb{E}\left[(\mathbf{CX}\beta + \mathbf{C}\varepsilon - \beta)(\mathbf{CX}\beta + \mathbf{C}\varepsilon - \beta)'\right] \qquad (8.13)\\ &= \mathbb{E}\left[\begin{array}{c} (\mathbf{CX} - \mathbf{I})(\mathbf{CX} - \mathbf{I})'\beta + \mathbf{C}\varepsilon(\mathbf{CX} - \mathbf{I})'\beta \\ +(\mathbf{CX} - \mathbf{I})\beta\varepsilon'\mathbf{C} + \mathbf{C}\varepsilon\varepsilon'\mathbf{C} \end{array}\right] \qquad (8.14)\\ &= (\mathbf{CX} - \mathbf{I})(\mathbf{CX} - \mathbf{I})'\beta + 0 + 0 + \mathbb{E}\left[\mathbf{C}\varepsilon\varepsilon'\mathbf{C}\right] \qquad (8.15)\\ &= 0 + 0 + 0 + \mathbf{CC'}\mathbb{E}\left[\varepsilon\varepsilon'\right] \qquad (8.16)\\ &= \sigma^2\mathbf{CC'} \qquad (8.17) \end{aligned}$$

Let $\mathbf{D} = \mathbf{C} - \mathbf{P}$ (the "distance" between our new projection matrix $\mathbf{C}$ and the OLS projection $\mathbf{P}$.) So, $\mathbf{C} = \mathbf{D} + \mathbf{P}$. Note that

$$\begin{aligned} \mathbf{CX} &= \mathbf{DX} + \mathbf{PX} = \mathbf{I} \qquad (8.18)\\ &= \mathbf{DX} + (\mathbf{X'X})^{-1}(\mathbf{X'X}) = \mathbf{I} \qquad (8.19)\\ &\Rightarrow \mathbf{DX} = \mathbf{0} \qquad (8.20) \end{aligned}$$

Using the fact that $\mathbf{C} = \mathbf{D} + \mathbf{P}$ and $\text{Var}[\mathbf{b}_0] = \sigma^2\mathbf{CC'}$, we have

$$\begin{aligned} \text{Var}[\mathbf{b}_0] &= \sigma^2[(\mathbf{D} + \mathbf{P})(\mathbf{D} + \mathbf{P})'] \qquad (8.21)\\ &= \sigma^2[(\mathbf{D} + (\mathbf{X'X})^{-1}\mathbf{X'})(\mathbf{D} + (\mathbf{X'X})^{-1}\mathbf{X'})'] \qquad (8.22)\\ &= \sigma^2[\mathbf{DD'} + (\mathbf{X'X})^{-1} + 2\mathbf{DX}(\mathbf{X'X})] \qquad (8.23)\\ &= \sigma^2\mathbf{DD'} + \sigma^2(\mathbf{X'X})^{-1} + 0 \qquad (8.24)\\ &= \sigma^2\mathbf{DD'} + \text{Var}[\mathbf{b}] \qquad (8.25) \end{aligned}$$

Since σ^2 and $\mathbf{X}$ are known values, we can only minimize the variance by reducing $\mathbf{DD}'$. If we set $\mathbf{C} = \mathbf{P} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, then we have that

$$\mathbf{DD}' = (\mathbf{C} - \mathbf{P})(\mathbf{C} - \mathbf{P})' = \mathbf{0} \tag{8.26}$$

and that $\mathbf{b}_0 = \mathbf{b}$. So, $\mathbf{b}$ is the minimum-variance unbiased estimate. ■

2. $\{A1, A2, A3, A5, A6, A8\} \Rightarrow$ the OLS estimate $\mathbf{b}$ is **BUE** - the best unbiased estimator. Adding the normality assumption gives the stronger condition that the OLS estimate is the smallest-variance estimate among *all* unbiased estimates.

8.6 Various Mathematical Results and Identities

The following commonly-used variables are useful in shortening notation.

$S_{xx} = \mathbf{X}'\mathbf{X} = \sum(\mathbf{X}_t - \bar{\mathbf{X}})^2 = \sum \mathbf{X}_t^2 - n\bar{\mathbf{X}}^2 = \sum \mathbf{X}_t^2 - \frac{1}{n}\left(\sum \mathbf{X}_t\right)^2$

$S_{xy} = \mathbf{X}'\mathbf{Y} = \sum(\mathbf{X}_t - \bar{\mathbf{X}})(\mathbf{Y}_t - \bar{\mathbf{Y}}) = \left(\sum \mathbf{X}_t\mathbf{Y}_t\right) - n\bar{\mathbf{X}}\bar{\mathbf{Y}} = \sum \mathbf{X}_t\mathbf{Y}_t - \frac{1}{n}\left(\sum \mathbf{X}_t \sum \mathbf{Y}_t\right)$

$SSE = \sum \left(Y_t - \hat{Y}\right)^2$ is the "sum of squares, errors"

$SSR = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$ is the "sum of squares, regression"

$SST = \sum \left(Y_t - \bar{Y}\right)^2$ is the "sum of squares, total"

Note that most of these results are derived from the "standard" OLS assumptions.

1. $\bar{Y} = a + b\bar{X}$
2. $E[Y] = X\beta$
3. $\mathbb{E}[\varepsilon\varepsilon'] = \mathrm{Var}[\varepsilon] = \sigma^2 I$ (spherical disturbances)
4. $b = \frac{S_{xx}}{S_{xy}} = (X'X)^{-1}X'Y$
5. $e = Y - X\left((X'X)^{-1}X'Y\right) = (I - X(X'X)^{-1}X')Y = MY$
6. $\frac{\varepsilon'}{\sigma^2} M \frac{\varepsilon}{\sigma^2} = \frac{e'e}{\sigma^2} \sim \chi^2_{Rank[M]=k}$ if $\varepsilon \sim N\left[0, \sigma^2\right]$
7. $X'M = MX = \mathbf{0}$
8. $SST = SSR + SSE$
9. $SST = Y'DY$, where D is an $n \times n$ idempotent matrix that transforms observations into deviation form

10. $SSE = e'e$
11. $Est\,\text{Var}\left[\varepsilon_i\right] = \hat{\sigma}^2 = \frac{SSE}{n-k} = \frac{e'e}{n-k}$
12. $R^2 = \frac{SSR}{SST} = \frac{b'XY}{SST}$
13. $\bar{R}^2 = 1 - \frac{SSE/(n-k)}{SST/(n-1)} = 1 - \frac{n-1}{n-k}(1-R^2) =$ "adjusted R^2"
14. $S_{xu} = \sum(X_t - \bar{X})(u_t - \bar{u}) = \sum(X_t - \bar{X})u_t \;\; \mathbb{E}[S_{xu}] = 0$
15. $\text{Var}[b] = \frac{\sigma^2}{S_{xx}} = \sigma^2(X'X)^{-1}$
16. $Est.\,\text{Var}[b] = s_b^2 = \frac{\hat{\sigma}^2}{S_{xx}} = \hat{\sigma}^2(X'X)^{-1} = \frac{e'e}{n-k}\left(X'X\right)^{-1}$
17. b is also the MLE estimator of β
18. Var $[b]$ achieves the Cramer-Rao Lower Bound of $\sigma^2\left(X'X\right)^{-1}$ and is therefore UMVUE.
19. $\text{Var}[a] = \sigma^2\frac{\sum X_t^2}{nS_{xx}} \quad Est.\,\text{Var}[a] = s_a^2 = \hat{\sigma}^2\frac{\sum X_t^2}{nS_{xx}}$
20. $\text{Cov}[a,b] = -\sigma^2\frac{\bar{X}}{S_{xx}} \quad Est.\,\text{Cov}[a,b] = s_{ab} = -\hat{\sigma}^2\frac{\bar{X}}{S_{xx}}$

8.7 MLE of The Linear Model

Assume that $(X_1,Y_1)...(X_n,Y_n)$ are drawn *iid* from the model

$$Y = X\beta + U \tag{8.27}$$

where $U \sim N\left[0.\sigma^2\right]$ with σ^2 and β unknown.

The likelihood function for a single observation is

$$L_i\left(\beta\right) = \mathbb{P}\left[U_i = Y_i - X_i\beta\right] = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{Y_i - X_i\beta}{\sigma}\right)^2\right] \tag{8.28}$$

The likelihood function for all N observations is

$$L(\beta,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(Y_i - X_i\beta\right)^2\right] \tag{8.29}$$

$$L\left(\beta,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{N}{2}}\exp\left[-\frac{1}{2\sigma^2}\left(Y - X\beta\right)'\left(Y - X\beta\right)\right] \tag{8.30}$$

The log-likelihood is

$$\mathcal{L}\left(\beta,\sigma^2\right) = -\frac{N}{2}\log\left[2\pi\right] - \frac{N}{2}\log\left[\sigma^2\right] - \frac{1}{2\sigma^2}\left(Y - X\beta\right)'\left(Y - X\beta\right) \tag{8.31}$$

Taking the FOC's gives

$$\frac{\partial \mathcal{L}}{\partial \beta} = -\frac{1}{2\sigma^2} 2 (Y - Xb)(-X') = 0 \tag{8.32}$$

$$0 = X'Xb - X'Y \tag{8.33}$$

$$b = (X'X)^{-1} X'Y = b_{OLS} \tag{8.34}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = -\frac{N}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} (Y - X\beta)' (Y - X\beta) = 0 \tag{8.35}$$

$$\frac{N}{\hat{\sigma}^2} = \frac{1}{(\hat{\sigma}^2)^2} (Y - X\beta)' (Y - X\beta) \tag{8.36}$$

$$\frac{(\hat{\sigma}^2)^2}{\hat{\sigma}^2} = \hat{\sigma}^2 = \frac{(Y - X\beta)' (Y - X\beta)}{N} = \frac{e'e}{N} \neq \hat{\sigma}^2_{OLS} \tag{8.37}$$

So, we have that the OLS coefficients are identical to the MLE coefficients, but the OLS estimate of σ is *not* equal to the MLE estimate. However, they are asymptotically equivalent.

8.8 Model Restrictions

8.8.1 Omitted Variables and OLS Estimates

If the "true" model contains more variables than the model estimated, then we have a model misspecification. As argued in the development of the error term in the OLS equation, any omitted variables get collected into the error term. If this misspecification is serious (in the sense that the omitted variable does explain variation in the dependent variable,) then we will have problems with our OLS estimates.

For example, consider the true model

$$Y_t = X_{1t}\beta_1 + X_{2t}\beta_2 + u_t \tag{8.38}$$

and the omitted variable model

$$Y_t = X_{1t}\beta_1 + v_t \tag{8.39}$$

Some problems with this new model are

$$\mathbb{E}[v_t] = \mathbb{E}[X_{2t}\beta_2 + u_t] = X_{2t}\beta_2 \neq 0 \text{ if } \beta_2 \neq 0 \tag{8.40}$$

$$\operatorname{Cov}[X_{1t}, (X_{2t}\beta_2 + u_t)] = \beta_2 \operatorname{Cov}[X_{1t}, X_{2t}] + \operatorname{Cov}[X_{1t,} u_t] = \beta_2 \operatorname{Cov}[X_{1t}, X_{2t}] \tag{8.41}$$

The estimate of $\hat{\beta}_1$ for the new model will be

$$\begin{aligned}\hat{\beta}_1 = (X_1'X_1)^{-1} X_1'Y = (X_1'X_1)^{-1} X_1' (\beta_1 X_1 + \beta_2 X_2 + u) \\ = \beta_1 + \beta_2 (X_1'X_1)^{-1} X_1'X_2 + (X_1'X_1)^{-1} X_1'u\end{aligned}$$

The expectation of this estimate gives $\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1 + \beta_2 \left(X_1'X_1\right)^{-1} X_1'X_2$, so we have a biased estimate. Specifically, the second term is called the **misspecification bias**. Even worse is the fact that β_2 appears in this equation, indicating that $\hat{\beta}_1$ is no longer an estimate of only the marginal effect of X_1.

Another misspecification error is the inclusion of excessive variables. Adding variables into the model that have a true coefficient of zero will not bias the coefficient estimates, but will increase estimate variance. Therefore, models with unnecessary variables will produce inefficient estimates. However, it can be shown that the estimated variance will still be unbiased, so that hypothesis tests are still valid. In general, these additional variables are generally harmless, but the regression will produce better results if they were omitted. Of course, if theory necessitates their inclusion, they should not be omitted.

These arguments give rise to the need for testing whether or not variables should be included or excluded in a given model.

8.8.2 Testing β Against a Given Value

$\mathbf{H}_0 : \beta_k = c$ (most often, $c = 0$, which implies that the corresponding X variable could be removed from the model.)

Property 6 can be used to test $\mathbf{H}_0 : \beta_k = c \Rightarrow \beta_k - c = 0$. Simply set up

$$z_k \equiv \frac{b_k - c}{\sqrt{\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim N(0,1) \tag{8.42}$$

$$t_k \equiv \frac{b_k - c}{SE\left[b_k\right]} = \frac{b_k - c}{\sqrt{\frac{e'e}{n-K}\left(X'X\right)^{-1}_{k,k}}} \sim t_{(n-k)} \tag{8.43}$$

Or, Do Not Reject (DNR) $\mathbf{H}_0$ if $c \in [b_k \pm SE[b_k] \cdot t_{\alpha/2,n-k}]$ ($\alpha/2$ is used for a 2-sided test)

8.8.3 General Setup for Model Restrictions Tests

$\mathbf{H}_0 : \mathbf{R}\beta = \mathbf{r}$

We require that $Rank[\mathbf{R}] = \dim[\mathbf{r}] = \#\mathbf{r}$, or that the rows of $\mathbf{R}$ (the restrictions) be linearly independent.

Example 8.4 *If you want to test* $\beta_2 = \beta_3$ and $\beta_4 = 0$ *as an "overall" restriction to your model, then the values of R and r will be*

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{8.44}$$

Note that the rows of **R** *are linearly independent, so this restriction is testable.*

The following techniques are all useful in testing model restrictions.

8.8.4 The F-Test

Under $\mathbf{H}_0 : \mathbf{R}\beta = \mathbf{r}$, the F-ratio is defined as

$$F^* \equiv \frac{\frac{(\mathbf{Rb}-\mathbf{r})'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{Rb}-\mathbf{r})}{\#\mathbf{r}}}{s^2} = \frac{(\mathbf{Rb}-\mathbf{r})'(\mathbf{R}\, s^2_{\mathbf{b}}\, \mathbf{R}')^{-1}(\mathbf{Rb}-\mathbf{r})}{\#\mathbf{r}} \sim F_{rank[R],n-k} \tag{8.45}$$

DNR if $F^* \; < F_{\alpha,\#r,n-k}$. Remember that F distributions are strictly positive and F tests are always one-sided tests.

An equivalent way to set up this problem (and easier to remember) is

$$F^* \equiv \frac{(SSR_R - SSR_U)/\#r}{SSR_U/(n-k)} \tag{8.46}$$

8.8.5 The Lagrange Multiplier Test

The Lagrange Multiplier (or, LM) Test is a very general test that can be used to examine model restrictions, to test for multiple insignificant variables, to build a model, or any general model-comparison procedure.

The first way to specify the LM test is simpler, but more specific. A general approach is added later.

Define your restricted model as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m + u \tag{8.47}$$

Define the unrestricted (or, full) model as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m + \beta_{m+1} X_{m+1} + ... + \beta_k X_k + v \tag{8.48}$$

The null hypothesis is that the added variables provide no explanatory power. Formally, $\mathbf{H}_0 : \beta_{m+1} = \beta_{m+2} = ... = \beta_k = 0$. Therefore, $\mathbf{H}_1$:*At least one of* $\beta_{m+1}, \beta_{m+2}, ..., \beta_k \neq 0$.

We first estimate the restricted model 8.47 to get $\hat{\beta}$'s. From this model we can calculate the residuals as

$$\hat{u}_R = Y - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - ... - \hat{\beta}_m X_m \tag{8.49}$$

Note that if the added variables in the *unrestricted* model are in fact significant, then they would be highly related to the residual $\hat{u}_R$ since their absence from Model 8.47 forces the effect on Y of those omitted variables into the error term u. So, if we regress $\hat{u}_R$ on the missing variables and find a good fit, then we know there is a suspicion that the omitted variables were "hidden" in the error term of the restricted model.

We now set up an **auxiliary regression** in which we regress the restricted model's residuals against the *full model* (all variables included.) From this model we can calculate $R^2_{AUX} = \frac{SSR_{AUX}}{SST_{AUX}}$. It can be shown that $nR^2_{AUX} \sim \chi^2_{k-m}$. Recall that the chi-squared distribution takes strictly positive values, so we form a one-tail test to test our hypothesis. Clearly if nR^2_{AUX} is large, then our auxiliary regression has a good fit and it is very likely that some of the omitted variable have an effect on the response variable Y.

This method is useful in iteratively "designing" a model. Start with a most basic model and continue to use the LM test procedure to test if a given variable should be added to the model. This is best done one variable at a time for obvious reasons. However, caution should be taken... you may end up adding variables that, although highly *correlated* with the response variable, may not have any place in your actual model. In other words, model-building can become dangerous if not done with a constant eye on whether or not each variable makes sense in the given model.

In the more general approach to the Lagrange Multiplier Test, we perform a likelihood maximization subject to an equality constraint. This procedure will require us to define a Lagrangian function with a Lagrange multiplier that gives this test its name. Reference section 5.3 for an introduction to likelihood functions and maximum likelihood estimation.

We wish to test the restriction that $c(\theta) - q = 0$, where c is some function of the parameters θ. In OLS, think of $\theta = \beta$. The Lagrangian for maximizing likelihood subject to this constraint is

$$\mathcal{L}^*(\theta) = \mathcal{L}(\theta) + \lambda(c(\theta) - q) \tag{8.50}$$

Maximizing this function over θ and λ gives our first order conditions

$$\frac{\partial \mathcal{L}^*}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \theta} + c'(\theta)\lambda = 0 \tag{8.51}$$

$$\frac{\partial \mathcal{L}^*}{\partial \theta} = c(\theta) - q = 0 \tag{8.52}$$

Note that if θ is a vector, then c is a vector-valued function and $c'(\theta) = \nabla c(\theta)$. From optimization theory, if a constraint on a maximization is not binding ("slack,") then the Lagrange multiplier will be zero. Therefore, we test $\mathbf{H}_0 : \lambda = 0$.

A simpler formulation of this test is to take the derivative of the log-likelihood function f the restricted model, which is

$$\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \hat{\theta}_R} = -c'(\hat{\theta}_R)\hat{\lambda} = \hat{g}_R \tag{8.53}$$

Here, if $\hat{g}_R = 0$, then the restricted model achieves the maximum likelihood, implying that the restriction is valid. This version of the test is often called the **score test** since the vector of first derivatives of the log-likelihood function is known as the **score**. The negative expected value of the matrix of second partials of $\mathcal{L}$ is known as the **information matrix** and is denoted $\mathbf{I}(\hat{\theta}_R) = -\mathbb{E}\left[\frac{\partial^2 \mathcal{L}}{\partial \theta^2}\right]$, which we saw in section 5.3. Finally, we define our test statistic to be

$$LM = \left(\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \hat{\theta}_R}\right)' \left[\mathbf{I}(\hat{\theta}_R)\right]^{-1} \left(\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \hat{\theta}_R}\right) \tag{8.54}$$

Under the null-hypothesis, LM has a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions.

When testing an OLS model, use the auxiliary regression method, but for more general procedures, the score test is appropriate. Keep in mind that LM on *converges* to a chi-squared distribution. For small samples, this test may not be valid.

8.8.6 The Likelihood Ratio Test

The Likelihood Ratio Test (or, LRT) is a very general procedure. We will introduce its applications to OLS with the goal of making the more general applications fairly transparent.

In subsection 5.3 we introduced the concept of maximum likelihood and the maximum likelihood function $L(\theta)$. Our goal is to compare to likelihood functions - that of a hypothesized set of coefficients, β_0, and the estimated set of coefficients, $\hat{\beta}$. Our null hypothesis will be $\mathbf{H}_0 : \beta = \beta_0$. We construct the following statistic

$$\lambda = \frac{L(\beta_0)}{L(\hat{\beta})} \tag{8.55}$$

Notice that $L(\hat{\beta})$ is guaranteed to be larger (or, less negative) than $L(\beta_0)$ since we know that $\hat{\beta}_{OLS} = \hat{\beta}_{MLE}$. Therefore, $\lambda \in [0, 1]$. If our null hypothesis is correct, then we expect λ to be very close to 1. Therefore, we will reject $\mathbf{H}_0$ unless λ is sufficiently large.

We first define the upper bound of our rejection region by specifying a desired level of significance, α, and then find some value K such that $\mathbb{P}[0 \leq \lambda \leq K | \beta = \beta_0] = \alpha$. In order to do this, we must know the distribution of λ. In some cases, this distribution may be known to be a t, F, or χ^2 distribution. Otherwise, we must rely on a large-sample test.

For large distributions, the test statistic $LR = -2 \ln \lambda = 2 \ln L(\hat{\beta}) - 2 \ln L(\beta_0) \sim \chi^2_{\#\beta_0}$, where $\#\beta_0$ represents the number of coefficients being restricted.

There are a lot of "warnings" that come with the LRT. For one, if you don't know the distribution of λ, then you can only use the asymptotic result that LR *converges* to a chi-squared distribution. For smaller samples, this test may not be appropriate. Secondly, when testing distributional assumptions (such as $\mathbf{H}_0 : X \sim N(\mu, \sigma^2)$,) this test becomes invalid as the underlying distribution between the restricted and unrestricted parameters must be equal. If not, the likelihood functions are not directly comparable and this test becomes meaningless.

8.8.7 The Wald Test

The Wald test is another likelihood test using the information matrix. We focus on the squared distance between our OLS estimate $\hat{\beta}$ and the value β_0 to which we restrict β. Clearly, if $\hat{\beta} = \beta_0$, then this restriction is justified. If $\left(\hat{\beta} - \beta_0\right)^2$ is large, then it is unlikely that this restriction is valid. We construct the following statistic

$$W = \left(\hat{\beta} - \beta_0\right)^2 \mathbf{I}(\hat{\beta}) \tag{8.56}$$

where $\mathbf{I}(\hat{\beta})$ is the information matrix calculated from the log-likelihood of our OLS estimates.

We know that $\hat{\beta} \sim N(\beta, \sigma^2/S_{xx})$. Therefore, $z = (\hat{\beta} - \beta)/\left(\sigma^2/S_{xx}\right) \sim N(0,1)$ and $z^2 \sim \chi^2_1$. If we were to test $\beta = 0$, then $W = \hat{\beta}^2 S_{xx}/\sigma^2$. Since $\hat{\beta} = S_{xy}/S_{xx}$, then $W = \hat{\beta} S_{xy}/\sigma^2$. Furthermore, we know that $\hat{\beta} S_{xy} = SSR$ and $\sigma^2 = SSE/n$, so

$$W = \frac{nSSR}{SSE} = \frac{nR^2}{1 - R^2} \tag{8.57}$$

For a large enough sample, W will be distributed as a chi-squared.

The more general Wald test (where $\hat{\theta}$ is a vector and the restriction is of the form $c(\theta) - q = 0$) is written as

$$W = \left[c(\hat{\theta}) - q\right]' \left(\mathrm{Var}[c(\hat{\theta} - q]\right)^{-1} \left[c(\hat{\theta}) - q\right] \tag{8.58}$$

where $W \sim \chi^2_{\#q}$, where $\#q$ represents the number of restrictions, which equals the number of elements in the q vector.

8.8.8 Summary of Tests

T-Test

$\mathbf{H}_0 : \beta_k = c$

$$t_k \equiv \frac{b_k - c}{SE\left[b_k\right]} = \frac{b_k - c}{\sqrt{s^2 \left(X'X\right)^{-1}_{k,k}}} \sim t_{(n-k)} \tag{8.59}$$

Do Not Reject (DNR) H_0 if $c \in [b_k \pm SE[b_k] \cdot t_{\alpha/2,n-k}]$ ($\alpha/2$ is used for a 2-sided test)

The F-Test

$\mathbf{H}_0 : \mathbf{R}\beta = \mathbf{r}$

$$F^* \equiv \frac{(\mathbf{Rb} - \mathbf{r})'(\mathbf{R}\, s^2_{\mathbf{b}}\, \mathbf{R}')^{-1}(\mathbf{Rb} - \mathbf{r})}{\#\mathbf{r}} \sim F_{rank[R],n-k} \tag{8.60}$$

or,

$$F^* \equiv \frac{(SSR_R - SSR_U)/\#r}{SSR_U/(n-k)} \tag{8.61}$$

DNR if $F^* \ < F_{\alpha,\#r,n-k}$.

The Lagrange Multiplier Test

$\mathbf{H}_0 : \beta_{m+1} = \beta_{m+2} = ... = \beta_k = 0$
$\mathbf{H}_1$:*At least one of* $\beta_{m+1}, \beta_{m+2}, ..., \beta_k \neq 0.$

$$\hat{u}_R = Y - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - ... - \hat{\beta}_m X_m \tag{8.62}$$

$$R^2_{AUX} = \frac{SSR_{AUX}}{SST_{AUX}} \tag{8.63}$$

$nR^2_{AUX} \ \sim \ \chi^2_{k-m}$
nR^2_{AUX} large implies omitted variables have an effect on the response variable Y.

More general approach:

Testing :$c(\theta) - q = 0$ for parameters θ. In OLS, think of $\theta = \beta$. The Lagrangian for maximizing likelihood subject to this constraint is

$$\mathcal{L}^*(\theta) = \mathcal{L}(\theta) + \lambda(c(\theta) - q) \tag{8.64}$$

Maximizing this function over θ and λ gives our first order conditions

$$\frac{\partial \mathcal{L}^*}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \theta} + c'(\theta)\lambda = 0 \tag{8.65}$$

$$\frac{\partial \mathcal{L}^*}{\partial \theta} = c(\theta) - q = 0 \tag{8.66}$$

$\mathbf{H}_0 : \lambda = 0.$

A simpler formulation:

$$\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \hat{\theta}_R} = -c'(\hat{\theta}_R)\hat{\lambda} = \hat{g}_R \tag{8.67}$$

If $\hat{g}_R = 0$, then the restriction is valid.

$$LM = \left(\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \hat{\theta}_R}\right)' \left[\mathbf{I}(\hat{\theta}_R)\right]^{-1} \left(\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \hat{\theta}_R}\right) \tag{8.68}$$

Under $\mathbf{H}_0$, LM has a limiting $\chi^2_{\#r}$ distribution.

The Likelihood Ratio Test

$\mathbf{H}_0 : \beta = \beta_0$.

$$\lambda = \frac{L(\beta_0)}{L(\hat{\beta})} \tag{8.69}$$

Reject $\mathbf{H}_0$ unless λ is sufficiently large.

Find K such that $\mathbb{P}[0 \le \lambda \le K | \beta = \beta_0] = \alpha$.

For large distributions, the test statistic $LR = -2\ln\lambda = 2\ln L(\hat{\beta}) - 2\ln L(\beta_0) \sim \chi^2_{\#r}$

The Wald Test

$$W = \left(\hat{\beta} - \beta_0\right)^2 \mathbf{I}(\hat{\beta}) = \frac{nR^2}{1-R^2} \tag{8.70}$$

$\mathbf{I}(\hat{\beta})$ is the info matrix calculated from the log-likelihood of our OLS estimates.

If we were to test $\beta = 0$, then $W = \hat{\beta}^2 S_{xx}/\sigma^2$

For a large enough sample, W will be distributed as a chi-squared.

The more general Wald test:

$$W = \left[c(\hat{\theta}) - q\right]' \left(\mathrm{Var}[c(\hat{\theta} - q]\right)^{-1} \left[c(\hat{\theta}) - q\right] \tag{8.71}$$

$W \sim \chi^2_{\#r}$

8.9 The ANOVA Table

The Analysis of Variance (ANOVA) table is the standard output of most computer programs that run regressions.

	Source	**Deg. of Freedom (df)**	**Mean Square = SS/df**	**F-stat**
Regression	SSR	1	SSR	$\frac{MSR}{MSE}$
Residual	SSE	$n-2$	$s^2 = \frac{SSE}{n-2}$	
Total	SST	$n-1$	$s_y^2 = \frac{SST}{n-1}$	
R^2	1-$\frac{SSR}{SST}$			

Chapter 9

Non-Spherical Disturbances

9.1 Introduction

Looking back to Section 8.3, we recall the following assumptions:

A5: Homoscedasticity

A6: Serial Independence

Together, these assumptions imply that $\mathbb{E}[\varepsilon'\varepsilon|X] = \sigma^2 I$. This is referred to as spherical disturbance because the errors should be evenly distributed, forming something that looks like a hypersphere in $\mathbb{R}^n$. If you really want to see this, try generating 100 random numbers from $\varepsilon_1 \sim N(0,1)$ and $\varepsilon_2 \sim N(0,1)$. Graph these on a 2-dimensional plane and you'll get a scattering of points mostly located within the unit circle. Of course there may be several outliers not within the unit circle. Extend this to higher dimensions and you roughly get a hypersphere. Now try the same exercise with $\varepsilon_1 \sim N(0,1)$ and $\varepsilon_2 \sim N(0,4)$. This forms an ellipse of points. Extending this to higher dimensions warps the sphere into a strange hyperellipse object.

If the error terms of the model are not "spherical," then the result isn't terribly disastrous. The normal OLS estimates will still be unbiased as long as $X'\varepsilon = 0$. To see this, recall that

$$\begin{aligned}\mathbb{E}[b] &= (X'X)^{-1}X'Y \qquad (9.1)\\ &= (X'X)^{-1}X'(X\beta + \varepsilon)\\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon\\ &= \beta + (X'X)^{-1}X'\varepsilon\\ &= \beta \; iff \; X'\varepsilon = 0\end{aligned}$$

However, the OLS estimates will no longer be efficient. In other words, their variance *and covariances* will be inflated. This can be seen by noting that the Gauss-Markov Theorem (Theorem 8.3) required both **A5** and **A6**. As a consequence, any sort of inferences or forecasts made with these estimates will be inefficient.

Although unbiasedness is in some ways more desirable than efficiency, we are restricted to making only weak statistical inferences about our estimates without efficiency. Therefore we examine procedures to achieve efficiency under these conditions.

9.2 Generalized Least Squares (GLS)

Consider a linear model in which T observations are taken, but the error-term variances of each observation are known to be unique. In other words, $\varepsilon_t \neq \varepsilon_s$ for some observations t, s. Therefore, we have non-spherical disturbances.

We introduce GLS with a specific (but instructive and useful) example.

Recall that the variables are also vectors of length T. We transform out model to

$$\frac{Y_t}{\sigma_t} = \frac{1}{\sigma_t}\beta_0 + \frac{X_{1t}}{\sigma_t}\beta_1 + \frac{X_{2t}}{\sigma_t}\beta_2 + ... + \frac{X_{kt}}{\sigma_t}\beta_k + \frac{u_t}{\sigma_t} \ \forall t \tag{9.2}$$

Relabelling for simplicity gives

$$Y_t^* = X_{0t}^*\beta_0 + X_{1t}^*\beta_1 + X_{2t}^*\beta_2 + ... + X_{kt}^*\beta_k + u_t^* \tag{9.3}$$

The variance of the error term becomes

$$\text{Var}[u_t^*] = \text{Var}\left[\frac{u_t}{\sigma_t}\right] = \frac{\text{Var}[u_t]}{\sigma_t^2} = 1 \tag{9.4}$$

We now have spherical disturbances and our estimates of regressing Y^* on $X_1^*, X_2^*, ..., X_k^*$ (without X_0^*) will be BLUE once again.

In general, assume $\mathbb{E}[\varepsilon'\varepsilon|X] = \sigma^2\Omega$, with $\mathbb{E}[\varepsilon|X] = 0$. Therefore, errors are non-spherical but still centered around zero. Assume that Ω is a known, positive definite, and symmetric matrix. Therefore, we can factor Ω into $C\Lambda C'$, where C is a matrix with eigenvectors as columns and Λ is a matrix with eigenvalues along the diagonal. Note that if we define $P' = C\Lambda^{-1/2}$, then $\Omega^{-1} = P'P$. This procedure is known as a Cholesky decomposition. Our model is $Y = X\beta + \varepsilon$. Premultiplying by P gives

$$PY = PX\beta + P\varepsilon \tag{9.5}$$

$$Y^* = X^*\beta + \varepsilon^* \tag{9.6}$$

As we did in our specific example above, consider the variance of our adjusted error term. $\mathbb{E}[\varepsilon^*\varepsilon^{*\prime}] = \sigma^2 I$ (can you verify this?) Our GLS estimate of the coefficients will be

$$\hat{\beta}_{GLS} = \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}Y^* = \left(X'\Omega^{-1}X'\right)^{-1} X'\Omega^{-1}Y \tag{9.7}$$

We now have an efficient, unbiased estimate.

Theorem 9.1 *(Unbiasedness) If* $\mathbb{E}[\varepsilon^*|X] = 0$*, then* $\mathbb{E}[\hat{\beta}_{GLS}] = \beta$

Proof. $\mathbb{E}[\hat{\beta}_{GLS}] = \mathbb{E}[(X^{*\prime}X^*)^{-1}X^{*\prime}Y^*] = \beta + \mathbb{E}[(X^{*\prime}X^*)^{-1}X^{*\prime}\varepsilon^*] = \beta$

Note that $\varepsilon^* = P\varepsilon$*, so* $\mathbb{E}[\varepsilon^*|X] = 0 \Longleftrightarrow \mathbb{E}[\varepsilon|X] = 0$ *since* P *is a matrix of known constants.* ■

Theorem 9.2 *(Efficiency)* $\hat{\beta}_{GLS}$ *is the minimum variance estimator among all linear unbiased estimates.*

Proof. *Simple extension of the Gauss-Markov Theorem.*

From this, we find that $\mathrm{Var}[\hat{\beta}_{GLS}] = \sigma^2(X'\Omega^{-1}X)^{-1}$ ■

Note that all of the results are essentially the same now as the OLS procedure, except we have Ω^{-1} appearing almost everywhere variables are multiplied.

An even more general setup is the **Weighted Least Squares (WLS)** model. Where GLS specifies the values of Ω, WLS allows this matrix to be determined by the researcher. There may be reasons for weighting certain observations, etc. However, careless WLS procedures will create undesirable properties (such as inefficiency,) so non-GLS versions of WLS should only be used when absolutely appropriate. Also note that OLS is a form of WLS where $\Omega = I$.

There exists a version of maximum likelihood estimation analogous to GLS that also uses Ω. However, we do not cover that topic here.

9.3 Feasible GLS

In setting up the GLS procedure, we made one particularly strong assumption. In real-world analysis, the value of Ω would almost never be known. Therefore, our next-best alternative is to estimate $\hat{\Omega}$.

To come up with an estimate for Ω, we must parametrize it as a function of X and θ, where θ is some vector of parameter that my include β.

The usual procedure is find some consistent estimator $\hat{\theta}$ of θ, which will depend on the question at hand. Using this, we can perform the Cholesky decomposition as described in the previous section to get $\hat{P}$. Premultiplying the variables in our model by $\hat{P}$ gives us the FGLS model, which can then be regressed using normal OLS techniques.

Since $\hat{\theta} \xrightarrow{d} \theta$ ("$\hat{\theta}$ *converges in distribution* to θ",) then we have that the asymptotic properties of the FGLS estimators are equivalent to the GLS estimators. Therefore, as $n \to \infty$, we have that the FGLS estimators are asymptotically efficient.

A common procedure in OLS is to first fit the unadjusted model and analyze the residuals against each of the model's variables. For example, if the residuals plotted against X_1 show a "megaphone" shape, then we'll want to look at $f(\varepsilon_i) = \varepsilon_i^2$ regressed against X_1. If the residuals plotted against Y show a megaphone shape, we'll want to look at $f(\varepsilon_i) = |\varepsilon_i|$ regressed against Y. Next, regress $f(\varepsilon)$ on the appropriate variables. Use the fitted values $\hat{\varepsilon}$ from this regression to form your vector weights w. Weight the OLS variables by w to get a FGLS model. Note that if your fitted values $\hat{\varepsilon}$ differ substantial from ε, the it is advisable to iterate the procedure to get "more convergent" values of $\hat{\varepsilon}$. This process is known as **iteratively reweighted least squares (IRLS)**.

We will introduce another, more specific way to perform FGLS in Section 9.4.2.

9.4 Heteroscedasticity

Heteroscedasticity means that $\text{Var}[\varepsilon_t] \neq \text{Var}[\varepsilon_s]$ for some s, t. In other words, each observation has its own error variance. For example, if data was gathered across different neighborhoods, then it may be unreasonable to assume that the error variance across neighborhoods is equal. This would introduce heteroscedasticity into the model[1].

We've already discussed the consequences of ignoring heteroscedasticity in Section 9 and how to derive BLUE estimates (or, to avoid redundancy, BLU estimates) in Section 9.2. Therefore we proceed to specific tests for heteroscedasticity.

9.4.1 Jackknife Estimator

It has been noted that if consistent estimates are possible for the variance-covariance matrix of the OLS estimates ($\text{Var}[b]$,) then we can make valid inferences in the presence of heteroscedasticity as long as we have sufficiently large samples. White (of the White test discussed below) suggests the **jackknife procedure** for devising consistent estimates of $\text{Var}[b]$, which he called a *heteroscedasticity consistent covariance matrix (HCCM) estimator*. The basic idea is to estimate your model over and over, each time dropping one (and only one) observation. This gives you a series of estimates, each with a unique variance-covariance matrix. The average of these variance-covariance matrices will be a consistent estimate for the true variance-covariance matrix.

The appeal of the jackknife procedure is that with a statistical software package capable of performing iterative loops, you can set up a fairly simple jackknife program to be run automatically.

[1] Heteroscedasticity is often spelled "heteroskedasticity." While the former is seen more often in literature, the latter is closer to the Greek word *skedastos*, which means "capable of being scattered." The term was coined in 1905 by statistics pioneer Karl Pearson, who spelled it with a "c." Interestingly, the phrase first appeared in the journal *Biometrika*, and we now have *Econometrica* - another case of the "k to c" trend.

9.4.2 Testing for Heteroscedasticity

There exist several tests for heteroscedasticity. Under the category of Lagrange Multiplier (LM) tests, there are the Breusch-Pagan test, the Harvey-Godfrey (or multiplicative heteroscedasticity) test, and the Park test, to name a few. There's also non-LM tests such as the Goldfield-Quandt test and the White test.

The **White test** is very general, but nonconstructive (i.e., does not suggest any hint of the nature of heteroscedasticity) and can also falsely classify a misspecification problem as a heteroscedasticity problem. For reference, the usual statistic is $V = s^2 \left(X'X\right)^{-1}$, which is asymptotically distributed as χ^2_{k-1}, where k is the number of regressors in the model, including the constant.

The **Goldfeld Quant test** is useful for testing whether the data can be separated into groups of equal variance, usually separated along the range of some X variable. To increase the power of the test, it is common to omit the "intermediate" values which lie on the borders of the two group, thus distinctly separating the groups. However, this represents a loss of information and is somewhat undesirable. The procedure is to estimate a regression for each of the two groups and then construct the statistic

$$F^*_{n_1-k,n_2-k} = \frac{e_1'e_1/(n_1-k)}{e_2'e_2/(n_2-k)} \tag{9.8}$$

where large values of the statistic lead to a rejection of the null hypothesis that the two groups have the same error variance.

The class of LM tests are similar and will be described jointly. Since we have to estimate n values of σ_t^2 and k values of β_i, we have $n+k$ free parameters, which is too many for our n observations to estimate. Therefore, we need to make simplifying restrictions. The suggested restrictions are

$$\sigma_t^2 = \alpha_1 + \alpha_2 Z_{2t} + \alpha_3 Z_{3t} + ... + \alpha_p Z_{pt} \quad \text{(Breusch-Pagan)} \tag{9.9}$$

$$\sigma_t = \alpha_1 + \alpha_2 Z_{2t} + \alpha_3 Z_{3t} + ... + \alpha_p Z_{pt} \quad \text{(Glesjer)} \tag{9.10}$$

$$\sigma_t^2 = \exp\left[\alpha_1 + \alpha_2 Z_{2t} + \alpha_3 Z_{3t} + ... + \alpha_p Z_{pt}\right] \quad \text{(Harvey-Godfrey)} \tag{9.11}$$

where Z_i is some variable of known (measurable) values. These equations are jointly known as **auxiliary equations** for the error variances. Note that the **Park test** is simply a special case of the Harvey-Godfrey test and will not be covered here. The procedure is as follows.

1. Regress Y against a constant and X. Obtain OLS estimates $\hat{\beta}$ and the residuals $\hat{u}_t$
2. Regress $\hat{u}_t^2$, $\hat{u}_t$, or $\ln[\hat{u}_t]$ against Z_t according to one of the above auxiliary equations.
3. Compute the test statistic $LM = nR^2$. As in Section 8.8.5, we find that this statistic is distributed as χ^2_{p-1}, where p is the number of Z variables used in the auxiliary regression.

4. The p-value will be $\mathbb{P}[\chi^2_{p-1} > LM]$, which is a one-tailed test as described previously.

5. Our $\mathbf{H}_0 : \alpha_i = 0 \; \forall 1 \leq i \leq p, i \in \mathbb{N}$ can be rejected for sufficiently large values of LM (or, sufficiently small p-values.)

Note that this auxiliary regression test is not actually what the original authors suggested. For example, the original Glesjer test is actually a Wald test. However, all of these tests are asymptotically equivalent and differ only in the assumed specification of the error term. We therefore restrict our attention to the more useful and computationally simple auxiliary regression technique.

Using the above auxiliary regressions, we can also perform a FGLS procedure. As above, run the normal OLS procedure to obtain $\hat{u}_t$. Run the desired auxiliary regression from the above three, where the Z variables will be the X variables, their squares, and their cross products. So, the set of variables will be $\{X_1, X_2, ..., X_k, X_1'X_1, X_1'X_2, ..., X_1'X_k, X_2'X_2, X_2'X_3, ..., X_2'X_k, ..., X_k'X_k\}$ Running the auxiliary regression gives estimates for the coefficients on each Z_i. Substituting these estimated coefficients back into the regression equation gives predicted values $\hat{\sigma}_t^2$. If using equation 9.9, we have $w_t = 1/\sqrt{\hat{\sigma}_t^2}$. If using equation 9.10, we have $w_t = 1/\hat{\sigma}_t$. If using equation 9.11, we have $w_t = 1/\sqrt{\hat{\sigma}_t^2}$. Note that in the first two equations we could observe negative predicted values of variance. In the third equation, our estimates are $\ln[\hat{\sigma}_t^2]$, so the exponentiation forces these values to be positive, which is desirable.

Chapter 10

Time-Series Models

10.1 Introduction

Many studies involve gathering data on certain variables at regular time intervals. For example, we may observe the variables $(Y_t, X_{1t}, ..., X_{kt})$ at points in time $t = 1, 2, ..., T$. As long as the OLS assumptions remain intact, time-series models present nothing different or unusual. However, time series data is quite likely in practice to violate a few of the standard assumptions. Therefore, we must examine those likely scenarios and how to deal with them.

10.2 Serial Correlation

Serial correlation (also called autocorrelation) is defined as correlation between error terms across time in a time series model. Recall from Section 8.2 the discussion on why we include an error term in the model. If we gather time-series data on $(Y, X_1, ..., X_k)$ but there exists some unobserved X_{k+1} that affects Y, then we have a model misspecification. That missing variable will be "embedded" into the error term. However, it is likely that X_{k+1} has some trend over time, particularly if the unit of time between observations is fairly small. Therefore, our error terms between observations will be correlated - particularly with smaller time intervals. As an example, models of asset prices over time (such as the price of a given stock) are almost serially correlated. This correlation between error terms violates OLS Serial Independence assumption ($A6$.)

In general, we believe that our error terms ε_t have the following property

$$\mathbb{E}\left[\varepsilon\varepsilon'\right] = \sigma^2\Omega \tag{10.1}$$

where Ω is a positive definite matrix with the constant $\sigma^2 = \text{Var}\left[\varepsilon_t\right]$ along the diagonal. Therefore, we assume constant variance of our error terms (homoscedasticity,) but admit correlation.

We define the **autocovariances** to be

$$\mathrm{Cov}\left[\varepsilon_t, \varepsilon_{t-s}\right] = \gamma_s \tag{10.2}$$

Note that γ is a function of the *distance in time* between the observations, but not of the *location in time* of the observations. This is referred to as **stationairty**. Stationarity means that the covariance between ε_1 and ε_3 is the same as the covariance between ε_{1001} and ε_{1003}. This assumption implies the value σ^2 from equation 10.1 is constant since $\sigma_t^2 = \mathrm{Var}\left[\varepsilon_t\right] = \mathrm{Cov}\left[\varepsilon_t, \varepsilon_t\right] = \gamma_0 = \mathrm{Cov}\left[\varepsilon_r, \varepsilon_r\right] = \mathrm{Var}\left[\varepsilon_r\right] = \sigma_r^2 \ \forall r, t \in [1, 2, ..., T]$. In fact, this analysis lets us rewrite equation 10.1 as

$$\mathbb{E}\left[\varepsilon\varepsilon'\right] = \gamma_0 R \tag{10.3}$$

where

$$R = \begin{bmatrix} 1 & \frac{\gamma_1}{\gamma_0} & \frac{\gamma_2}{\gamma_0} & \frac{\gamma_3}{\gamma_0} & \cdots & \frac{\gamma_{T-1}}{\gamma_0} \\ \frac{\gamma_1}{\gamma_0} & 1 & \frac{\gamma_1}{\gamma_0} & \frac{\gamma_2}{\gamma_0} & & \frac{\gamma_{T-2}}{\gamma_0} \\ \frac{\gamma_2}{\gamma_0} & \frac{\gamma_1}{\gamma_0} & 1 & \frac{\gamma_1}{\gamma_0} & & \frac{\gamma_{T-3}}{\gamma_0} \\ \frac{\gamma_3}{\gamma_0} & \frac{\gamma_2}{\gamma_0} & \frac{\gamma_1}{\gamma_0} & 1 & & \frac{\gamma_{T-4}}{\gamma_0} \\ & \vdots & & & \ddots & \vdots \\ \frac{\gamma_{T-1}}{\gamma_0} & \frac{\gamma_{T-2}}{\gamma_0} & \frac{\gamma_{T-3}}{\gamma_0} & \frac{\gamma_{T-4}}{\gamma_0} & \cdots & 1 \end{bmatrix} \text{ or } R_{ts} = \frac{\gamma_{|t-s|}}{\gamma_0} \tag{10.4}$$

R is referred to as the **autocorrelation matrix**.

10.3 Ignoring Serial Correlation

The proofs of unbiasedness and consistency of $\hat{\beta}_{OLS}$ did not depend on the Serial Independence assumption. Therefore, if we perform OLS estimates with serially correlated errors, we will still derive unbaised and consistent estimates. However, as in the case of general non-spherical disturbances, the Gauss-Markov Theorem (Theorem 8.3) is no longer valid. Therefore, OLS estimates will not be efficient. As a result, hypothesis tests based on the estimates will be invalid. In fact, under certain conditions, the t-statistic will be grossly overestimated, causing the experimenter to (possibly) reject $\mathbf{H}_0 : \beta = \mathbf{0}$ when it should not be rejected at any reasonable level of significance.

blah - see Grether's notes

10.4 Various Forms of Serial Correlation

Serial correlation can take on a variety of functional forms. In other words, the manner in which ε_t and ε_{t-s} are related could be different across different models. We now consider commonly assumed forms of correlation. The actual type encountered clearly depends on the underlying theory and beliefs about the dynamics of the data across time.

10.4.1 AR(1) - 1^{st}-Order Autoregressive Correlation

First-order autoregressive correlation (**AR(1)**) assumes that error terms follow the behavior given by

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \tag{10.5}$$

$$\rho \in (-1, 1) \tag{10.6}$$

$$\mathbb{E}\left[u_t\right] = 0 \; \forall t \tag{10.7}$$

$$\mathbb{E}\left[u_t u_t\right] = \sigma_\varepsilon^2 < \infty \tag{10.8}$$

$$\mathbb{E}\left[u_t u_{t-s}\right] = 0 \; \forall s \neq 0 \tag{10.9}$$

Notice that we assume properties of u that would normally be assumed about ε[1].

AR(1) has a fairly easy functional form and is the most commonly assumed among the variants. Mainly, AR(1) is often seen as a reasonable approximation of more complex correlation structures. However, its simplicity gains more than just tractability. Attempting to use more complex structures can be very sensative to the data and might produce drastically different predictions under very minor changes.

An important property of AR(1) is that $\operatorname{Cov}\left[\varepsilon_t, \varepsilon_{t-s}\right] \neq 0 \; \forall t, s$, although it can become quite small when $|\rho| < 1$. This is easily seen by iteratively substituting the functional form for the correlation into itself to get

$$\varepsilon_t = \rho\left(\rho\left(\rho\left(...\left(\rho\varepsilon_1 + u_1\right)...\right) + u_{t-2}\right) + u_{t-1}\right) + u_t = \rho^{t-1}\varepsilon_1 + \sum_{i=0}^{t-1} \rho^i u_{t+1} \tag{10.10}$$

To devise a test of the correlation structure, the parameter ρ can easily be estimated using

$$\hat{\rho} = \frac{\sum\limits_{t=2}^{T} e_t e_{t-1}}{\sum\limits_{t=1}^{T} e_t^2} \tag{10.11}$$

[1] These three assumption are often referred to as **white noise** (with zero mean) assumptions.

10.4.2 AR(p) - p^{th}-Order Autoregressive Correlation

Higher-order autoregressive correlation (**AR(p)**) assumes that error terms follow the behavior given by

$$\begin{aligned} \varepsilon_t &= \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_p \varepsilon_{t-p} + u_t && (10.12) \\ \theta_k &\in (-1,1)\ \forall k \in [1,2,...,p] && (10.13) \\ \mathbb{E}[u_t] &= 0\ \forall t && (10.14) \\ \mathbb{E}[u_t u_t] &= \sigma_\varepsilon^2 < \infty && (10.15) \\ \mathbb{E}[u_t u_{t-s}] &= 0\ \forall s \neq 0 && (10.16) \end{aligned}$$

10.4.3 MA(1) - 1^{st}-Order Moving Average Correlation

blah $\varepsilon_t = \rho u_{t-1} + u_t$

10.4.4 MA(p) - p^{th}-Order Moving Average Correlation

blah $\varepsilon_t = \theta u_{t-1} + \theta^2 u_{t-2} + ... + \theta^p u_{t-p} + u_t$

10.5 Testing for Serial Correlation

10.5.1 Introduction

By far the most common test statistic for serial correlation is the Durbin-Watson (DW) statistic. In fact, most computer packages generate the value of DW *regardless of whether or not the data is time-series*, probably because it's easy for a computer to calculate once the residuals are known. However, an uninformed researcher performing a regression on data that has no meaningful ordering may see the DW statistic in the computer output and be alarmed that there is a correlation problem. If there is no meaningful ordering to the data, then the data ordering can be "re-shuffled" and the DW statistic will completely change. The DW has other limitations which we will discuss shortly.

10.5.2 The Durbin-Watson Test

The DW test statistic is defined in terms of the residuals $\{e_t\}_{t\in\{1,2,...,T\}}$ as

$$DW = \frac{\sum\limits_{t=2}^{T} (e_t - e_{t-1})^2}{\sum\limits_{t=1}^{T} e_t^2} \tag{10.17}$$

In the AR(1) setting, we have[2]

$$DW = 2(1-\hat{\rho}) - \frac{e_1^2 + e_T^2}{\sum\limits_{t=1}^{T} e_t^2} \tag{10.18}$$

From this, we have that

$$\lim_{t\to\infty} DW = 2(1-\hat{\rho}) \tag{10.19}$$

$$DW \approx 2(1-\hat{\rho}) \tag{10.20}$$

Some texts will claim that $2(1-\hat{p})$ is a good approximation for the DW statistic. However, calculating $\hat{\rho}$ is nearly as difficult as calculating the true DW statistic. Therefore, this simplification should only be used to note that

$$\rho \in [-1,1] \overset{approx}{\Longrightarrow} DW \in [0,4] \tag{10.21}$$

$$\rho = 0 \Rightarrow DW = 2 \tag{10.22}$$

To test whether or not serial correlation exists, the null hypothesis is $\mathbf{H}_0 : DW = 2$. Since the distribution of DW isn't very nicely behaved, we frequently see charts of the upper and lower bounds (DW_U and DW_L) that depend on T and k. Charts give upper and lower bounds for a one-tail 5% or 1% significance level test since those are really the only levels for which charts have been tabulated. Unfortunately, we cannot perform a two-tailed test Therefore, we have only the following.

Possible Durbin-Watson Hypothesis Tests

- Testing for positive serial correlation ($\rho > 0$)

  $\mathbf{H}_0 : \rho = 0\ (DW = 2)$

  $\mathbf{H}_1 : \rho > 0\ (DW < 2)$

  Reject $\mathbf{H}_0$ at α significance level if $DW \le DW_L^{T,k,\alpha}$

  Do Not Reject $\mathbf{H}_0$ at α significance level if $DW \ge DW_U^{T,k,\alpha}$

  No conclusion if $DW_L^{T,k,\alpha} \le DW \le DW_U^{T,k,\alpha}$

- Testing for negative serial correlation ($\rho < 0$)

  $\mathbf{H}_0 : \rho = 0\ (4 - DW = 2)$

  $\mathbf{H}_1 : \rho < 0\ (4 - DW < 2)$

  Reject $\mathbf{H}_0$ at α significance level if $4 - DW \le DW_L^{T,k,\alpha}$

  Do Not Reject $\mathbf{H}_0$ at α significance level if $4 - DW \ge DW_U^{T,k,\alpha}$

  No conclusion if $DW_L^{T,k,\alpha} \le 4 - DW \le DW_U^{T,k,\alpha}$

[2] For a tedious exercise, try showing this.

10.5.3 Limitations of the DW Test

There are significant limitations to the DW test. These limitations make it more surprising that statistical software packages include the DW statistic so frequently. Since the DW test is somewhat often used in inappropriate situations, econometrics professors often will drill their students on the limitations of the DW test so that they don't make these mistakes.

1. If the data is not time-series (or order-dependent in some way,) then there is no meaning to serial correlation and no need to test for it.

2. There exists an inconclusive range in the hypothesis tests between the upper and lower limits in which nothing can be said about serial correlation. The Lagrange Multiplier (LM) Test should be used when the DW test is inconclusive.

3. The DW test is invalid if lagged independent variables appear in the model (see section 10.6.)

4. The DW test is powerful for testing ρ in the AR(1) setting. In other serial correlation structures (such as AR(p),) ρ may not be indicative of the true serial correlation pattern.

5. The DW test requires X to be nonstochastic.

For example, if the correlation structure were $\varepsilon_t = \rho\varepsilon_{t-2}+u$, the DW statistic would not indicate any correlation problems.

10.6 Lagged Independent Variables

For lagged independent variables, the model is of the form

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + ... + \beta_k X_{t-k} + u_t \tag{10.23}$$

10.7 Lagged Dependent Variables

blah

Chapter 11

Explanatory Variable Problems

11.1 Multicollinearity

Multicollinearity occurs when two explanatory variables have approximate linear relationships. In other words, if two of the independent variables are very highly correlated, then we have multicollinearity.

In the case of **exact multicollinearity**, two variables have an exact linear relationship. This violates OLS Assumption 3 since the rank of the X matrix would be less than k due to the fact that the columns would not be linearly independent. This is a very extreme case and, if it were to occur, OLS estimates simply wouldn't exist as the matrix $X'X$ would be singular.

The more realistic case is that of **near multicollinearity**, where two variables have an approximate linear relationship. This is also somewhat more problematic because the OLS procedure will yield unique estimates, so the investigator will not receive any sort of "warning" that the explanatory variables are problematic. From this point forward, when we speak of multicollinearity, we will mean near multicollinearity as exact multicollinearity is a somewhat trivial case.

What will be the properties of multicollinearity? Note that the proof of the Gauss-Markov Theorem (Theorem 8.3) required $A3$, which assumes no *exact* multicollinearity. However, the theorem is not affected by highly correlated explanatory variables - as long as they are not perfectly correlated. Therefore, the theorem holds and the OLS estimates will still be BLUE. Furthermore, the OLS estimates are still the Maximum Likelihood estimates.

The only major drawback to multicollinearity is the increased standard error of the regression coefficients. It can be shown that if X_2 and X_3 are correlated with $\rho_{2,3} < 1$, then

$$\operatorname{Var}\left[\hat{\beta}_2\right] = \frac{\sigma^2}{(1-R_2^2)S_{22}} \tag{11.1}$$

$$\operatorname{Var}\left[\hat{\beta}_3\right] = \frac{\sigma^2}{(1-R_3^2)\,S_{33}} \tag{11.2}$$

$$\operatorname{Cov}\left[\hat{\beta}_2, \hat{\beta}_3\right] = \frac{-\sigma^2\rho_{2,3}}{\left(1-\rho_{2,3}^2\right)\sqrt{S_{22}S_{33}}} \tag{11.3}$$

where R_j^2 is the R^2 of a regression of the j^{th} variable on the other variables, $S_{22} = S_{X_2X_2} = \sum_{i=1}^{n}\left(X_{2i} - \bar{X}_2\right)^2$ and similar for S_{33}.

Notice that these variances explode as the correlation between one X variable and the others in the model (R_j^2) nears 1. Also, note that since the estimates are BLUE, this increased variance is still the minimum variance possible. Therefore, using correlated explanatory variables will inflate your estimates, but there is guaranteed to be no better linear estimate than the OLS solution. Your choice is therefore to admit the higher variance or to remove one of the correlated variables.

In the absence of multicollinearity, we have that $R_j^2 = 0 \; \forall j$. As a consequence, $\hat{\beta}_j = S_{Yj}/S_{jj} \; \forall j$. Therefore, the estimates are completely independent of the inclusion of the other variable. So, the estimate of $\hat{\beta}_j$ when Y is regressed on all of the X's will be identical to the estimate of $\hat{\beta}_j$ when Y is regressed on X_j alone. This is known as **orthogonal regression**.

Obviously, datasets are going to contain some multicollinearity. It is generally up to the researcher to determine whether or not the degree of correlation between X variables is extreme. One measure often reported by software packages is the **Variance Inflation Factor**, or (**VIF**.) $VIF_k = 1/\left(1+R_k^2\right)$. If

Perhaps the most instructive lesson from multicollinearity comes from the practice of "model building" - where an investigator adds more and more explanatory variables until the regression results are satisfactory. This procedure is often abused as meaningless variables that happen to have some correlation with the endogenous variable can get thrown into a model in order to help the fit. However, as we pile on explanatory variables, we are likely to include variables which are highly correlated. The result is that the estimate variances drastically increase for *both* estimates, which drives up their t-statistics and makes both variables seem insignificant. The investigator can be confused when a variable that used to be highly significant suddenly becomes insignificant when a new variable gets added to the model.

On the other hand, a good researcher may realize that in their particular investigation two highly correlated explanatory variables should both be included in the model for underlying theoretical reasons *even though* multicollinearity will hurt their significance level.

11.2 Testing for Multicollinearity

There aren't really any formal tests for multicollinearity since its detection is straight-forward. The consequences of multicollinearity discussed above lead to certain unusual situations that can be "red flags" for multicollinearity.

1. **High Correlation Coefficients**

   Any good researcher would run summary statistics for their variables to look for red flags before attempting to fit their model. Correlation between X variables is the definition of multicollinearity and can be spotted easily through this analysis.

2. **High R^2 with Low Values of the *t*-statistic**

   We have seen that multicollinearity increases the estimate variances, which in turn raises the t-statistics. However, we also know that adding variables always increases the R^2 of a regression. So, if we see a high R^2 with very low significance levels, it is most likely that multicollinearity exists. In fact, a high R^2 should normally correspond to a very good fit of the data, so we should expect to see some fairly significant coefficient estimates, so this situation is especially curious.

3. **High F-Value for a Group of Coefficients that are Individually Insignificant**

   If we run a test of joint significance over a group of X variables and find that they are, as a group, highly significant while each coefficient is apparently insignificant, we have a red flag for multicollinearity.

4. **Coefficients Changing With Inclusion of New Variables**

   We saw that if $\rho_{2,3} = 0$, then $\hat{\beta}_2$ will be unaffected by the inclusion of X_3 in the model. The contrapositive of this statement is that if $\hat{\beta}_2$ is affected by the inclusion of X_3 in the model, then $\rho_{2,3} \neq 0$. So, if adding a new variable changes the existing coefficients, multicollinearity is likely.

11.3 Measurement Error & Random Regressors

First note that if Y is measured with error, no OLS violation results. To see this, note that

$$\begin{aligned} Y + u &= X\beta + \varepsilon \qquad (11.4) \\ Y &= X\beta + (\varepsilon - u) \end{aligned}$$

where $(\varepsilon - u) \sim N\left[0, \sigma_\varepsilon^2 + \sigma_u^2\right]$, which completely conforms to the OLS assumptions. Therefore, we will restrict our attention to models in which only the explanatory variables are measured with error.

We suppose that X is measured with error (or that it is random for some other reason,) so that

$$X = Z + \varepsilon \tag{11.5}$$

This problem can come about either when a variable is known to be measured with error or if a proxy variable is used that approximately measures a desired variable. For example, if we want to measure intelligence (”INTL”) but can only observe IQ scores, we have a ”noisy” proxy for intelligence. We may assume that $IQ = INTL + \varepsilon$, where $\varepsilon \sim N\left[0, \sigma_\varepsilon^2\right]$.

Assume the ”true” model is $Y = Z\beta + u$, with error it becomes our estimated model of the form

$$Y = (X - \varepsilon)\beta + u = X\beta + (u - \varepsilon\beta) = X\beta + w \tag{11.6}$$

Further assume that Z, u, and ε are mutually independent.

Since our error term now contains ε, we have correlation between X and the error term. Specifically, $\operatorname{Cov}[X, w] = \operatorname{Cov}[Z + \varepsilon, u - \varepsilon\beta] = -\beta\sigma_u^2$

The OLS estimate will become inconsistent in this scenario.

$$\begin{aligned} \operatorname{plim} b &= \frac{\operatorname{plim}(1/n)\, X'Y}{\operatorname{plim}(1/n)\, X'X} \qquad (11.7)\\ &= \frac{\operatorname{plim}(1/n)\,(Z' + \varepsilon')(Z\beta + u)}{\operatorname{plim}(1/n)\,(Z + \varepsilon)'(Z + \varepsilon)} \\ &= \frac{\operatorname{plim}(1/n)\,(Z'Z\beta + Z'u + \varepsilon'Z\beta + \varepsilon'u)}{\operatorname{plim}(1/n)\,(Z'Z + Z'\varepsilon + \varepsilon'Z + \varepsilon'\varepsilon)} \\ &= \frac{\beta \operatorname{plim}(1/n)\,(Z'Z)}{\operatorname{plim}(1/n)\,(Z'Z) + \sigma_\varepsilon^2} \\ &= \beta\left(\frac{1}{1 + \frac{\sigma_\varepsilon^2}{\operatorname{plim}(1/n)(Z'Z)}}\right) \end{aligned}$$

Therefore, b is inconsistent as long as $\operatorname{plim}(1/n)(Z'Z)$ is finite (which we assume.) If $\sigma_\varepsilon^2 > 0$, then β is asymptotically **attenuated** (downwardly-biased.)

Chapter 12

Panel Data

12.1 Introduction

When time-series data is gathered that can be broken into cross-sectional blocks, the result is **panel data**. In these data sets, researchers track distinct groups across time. Agriculture studies frequently use panel data to study various crops or fields across time.

12.2 The Model

Since our data are grouped both cross-sectionally and across time, it is appropriate to put both an i subscript and a t subscript on the variables. Furthermore, we add a "group-effect" variable to our model to capture a common effect among all observations within a given group i. This group-effect term will be an important aspect of our model, as we shall soon see. Formally, we have

$$Y_{it} = \alpha_i + X_{it}\beta + \varepsilon_{it} \tag{12.1}$$

If α_i is a constant term that depends on i, then we have a **fixed effects model** where each cross-sectional group is assumed to have some fixed effect on the response variable. For example, if a farmer has multiple fields, there may be one field that naturally has a higher yield than the others. If it is assumed that this difference is constant for the entire field, then a fixed effects model is reasonable.

If α_i is a random variable similar to the error term, but is drawn only once for each group (i.e., it does not have a t subscript.) This model is a **random effects model** since the group-level effect is considered random instead of fixed.

12.3 Fixed Effects Models

Since α_i is assumed constant, it should be thought of as a model parameter and not a variable. The subscript is prehaps misleading. We do not directly observe α_i, so it must be estimated. Instead of using the subscript notation, we will use dummy variables to give α the desired "behavior" in our model. We write

$$Y_{it} = D\alpha + X_{it}\beta + \varepsilon_{it} \tag{12.2}$$

where D is an $nT \times n$ matrix and α is a column vector described by

$$D = \left[\begin{array}{cccc} d_1 & d_2 & ... & d_n \end{array} \right] \tag{12.3}$$

$$\alpha' = \left(\begin{array}{cccc} \alpha_1 & \alpha_2 & ... & \alpha_n \end{array} \right) \tag{12.4}$$

Here, d_g is a column vector of nT ones and zeros indicating whether a given observation belongs to the g^{th} group. Formally, if $i_{i,t}$ is the group of the observation i, t, then d_g is a vector of indicator functions.

$$d_g = \left(\begin{array}{c} \chi_{\{i_{1,1}=g\}} \\ \chi_{\{i_{1,2}=g\}} \\ \vdots \\ \chi_{\{i_{1,T}=g\}} \\ \chi_{\{i_{2,1}=g\}} \\ \vdots \\ \chi_{\{i_{2,T}=g\}} \\ \vdots \\ \chi_{\{i_{n,T}=g\}} \end{array} \right) \tag{12.5}$$

Note that $\sum_{g=1}^{n} d_{git} = 1 \; \forall t$, so that for any given observation i, t, we have only one dummy variable "switched on."

This is more confusing that it perhaps needs to be, but dummy variables are a clever way to switch on and off our variable α_i as we scan through our observations i, t.

blah - a lot more can be said on this topic!

12.4 Random Effects Models

blah

Chapter 13

Systems of Regression Equations

13.1 Introduction

In this chapter, we consider the case of multiple regression equations that appear to be completely independent, but it is known that there exists some inherent "link" between the various equations. If we estimate the equations independently, we are effectively throwing away information about how the various equations are linked. Statisticians abhor the idea of discarding information (as well they should,) so we analyse what can be done with these "linked" equations.

13.2 The Seemingly Unrelated Regressions Model

We assume a list of regressions that appear to be independent, estimable equations. However, there exists an underlying correlation structure (for some theoretical reason) between the error terms of the models. For example, a series of regressions may be run on each cross-sectional unit of a given panel data set. If stock prices of two companies are tracked over time along with some explanatory variables, for example, this panel data could be used to generate two regressions – one for each company. However, there is likely to be correlation between the two regression error terms as stock prices tend to have natural correlations.

We take as an example the regressions of two firms, i and j

$$Y_i = X_i\beta + \varepsilon_i \tag{13.1}$$

$$Y_j = X_j\beta + \varepsilon_j \tag{13.2}$$

with T observations each and where we know (for some reason) that

$$\mathbb{E}\left[\varepsilon_i \varepsilon_j'\right] = \sigma_{ij} I^{(T)} \tag{13.3}$$

Notice that $\mathbb{E}\left[\varepsilon_{it}\varepsilon_{js}\right] = 0$ in this setup for all $t \neq s$.

At this point, the uninterested reader can skip the development that follows and proceed to the conclusions in subsection 13.2.1.

We diverge for a moment to introduce the **Kronecker product**. This matrix operator multiplies on matrix into another term-by-term. Formally, if

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1s} \\ b_{21} & b_{22} & \cdots & b_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{rs} \end{bmatrix} \tag{13.4}$$

then

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1m}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & a_{n2}\mathbf{B} & \cdots & a_{nm}\mathbf{B} \end{bmatrix} \tag{13.5}$$

Note that in the Kronecker product, the matrix on the right side of the $\otimes$ symbol gets inserted into the matrix on the left side, term-by-term. Also note that $\mathbf{A}$ and $\mathbf{B}$ had dimensions of $n \times m$ and $r \times s$, respectively, so $\mathbf{A} \otimes \mathbf{B}$ has dimensions $nr \times ms$. Finally note that Kronecker products can be applied to vectors, as well as to scalars[1].

Theorem 13.1 *For any nonsingular matrix A and any matrix B,*

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \tag{13.6}$$

Theorem 13.2 *For any nonsingular matrix A,and matricies C and D such that CD exists,*

$$C\,(A \otimes I)\,D = A \otimes CD \tag{13.7}$$

The proofs are omitted.

Stacking the regression equations gives

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_m \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix} \tag{13.8}$$

which can be written as

$$\mathbf{Y}^{(mT\times 1)} = \mathbf{X}^{(mT\times mk)}\beta^{(km\times 1)} + \varepsilon^{(mT\times 1)} \tag{13.9}$$

where the matrix dimensions are included as superscripts.

[1] The Kronecker product of two scalars is identically equal to their scalar product.

Note that this combined model is itself a nice regression equation. However, we know that there is correlation between observations, so we must use a GLS procedure. This means the covariance matrix for this new equation needs to be determined. For a given observation t, the covariance matrix would be

$$\mathbb{E}\left[\varepsilon_t\varepsilon_t'\right] = \mathbf{\Sigma}_t = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{bmatrix} \quad \forall t \tag{13.10}$$

However, within a single regression, the T observations are assumed to be homoscedastic. Therefore, putting it all together gives

$$\mathbb{E}\left[\varepsilon\varepsilon'\right] = \mathbf{\Sigma} \otimes I^{(T)} = \mathbf{V} \tag{13.11}$$

From Theorem 13.1 we know that $\mathbf{V}^{-1} = \mathbf{\Sigma}^{-1} \otimes I^{(T)}$. Recall that $\hat{\beta}_{GLS} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y$. Therefore, we have that

$$\hat{\beta}_{SUR} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y \tag{13.12}$$

Denote the i, j^{th} element of $\mathbf{\Sigma}^{-1}$ as γ_{ij}. Expanding $\hat{\beta}_{SUR}$ and using Theorem 13.2 gives

$$\hat{\beta}_{SUR} = \begin{bmatrix} \gamma_{11}X_1'X_1 & \gamma_{12}X_1'X_2 & \cdots & \gamma_{1m}X_1'X_m \\ \gamma_{21}X_2'X_1 & \gamma_{22}X_2'X_2 & \cdots & \gamma_{2m}X_2'X_m \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m1}X_m'X_1 & \gamma_{m2}X_m'X_2 & \cdots & \gamma_{2m}X_m'X_m \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^m \gamma_{1j}X_1'Y_j \\ \sum_{j=1}^m \gamma_{2j}X_2'Y_j \\ \vdots \\ \sum_{j=1}^m \gamma_{1j}X_m'Y_j \end{bmatrix} \tag{13.13}$$

It is noteworthy that the inverted matrix above is the asymptotic covariance matrix of the estimate. Also note that FGLS is the more practical procedure here since the correlation will rarely be known. The appropriate FGLS procedure is to run the separate regression equations and use the residuals of each equation to estimate the values of σ_{ij}. Specifically, $\hat{\sigma}_{ij}$ is estimated using the diagonal elements of e_ie_j', which is assumed to be constant but obviously may not be so. Therefore, we take their average to get $\hat{\sigma}_{ij} = (e_i'e_j)$[2]. This generates an estimate of $\mathbf{\Sigma}$ and the SUR model can be estimated.

[2] Yes, we have switched from e_ie_j' to $e_i'e_j$. Note that the trace of e_ie_j' equals $e_i'e_j$, which equals $\sum_{t=1}^T e_{it}e_{jt}$.

13.2.1 SUR Conclusions

We have used the fact that the original, homoscedastic equations have some inter-equation correlation structure to develop a larger GLS regression equation that does have heteroscedasticity. Therefore, the results of this combined regression will have all of the properties of a standard GLS model. Furthermore, by using the correlation structure, we may have gained some efficiency. In fact, we generally gain efficiency *unless* one of the following is true[3].

- The equations have no inter-equation correlation structure ($\mathbb{E}\left[\varepsilon_{ti}\varepsilon'_{jt}\right] = 0\ \forall t, i \neq j$.)
- The equations all have the same explanatory variables ($X_i = X_j$)

If there in fact does exist some efficiency gain, then

- The efficiency gain is larger as the correlations increase.
- The less correlated are the X variables across equations, the greater the efficiency increase.

[3] These are sufficient conditions for the SUR-OLS equivalence, but may not be necessary.

Chapter 14

Simultaneous Equations

14.1 Introduction

In simultaneous equations, we have multiple equations, but introduce the possibility of endogenous variables appearing in equations. In other words, the dependent variable of one equation may be appear as an explanatory variable in another.

In general, a system of M equations with M endogenous variables and K exogenous variables can be written as

$$\begin{aligned}
\gamma_{11}Y_{1t} + \gamma_{21}Y_{2t} + ... + \gamma_{M1}Y_{Mt} + \beta_{11}X_{1t} + ... + \beta_{K1}X_{Kt} &= \varepsilon_{1t} \\
\gamma_{12}Y_{1t} + \gamma_{22}Y_{2t} + ... + \gamma_{M2}Y_{Mt} + \beta_{12}X_{1t} + ... + \beta_{K2}X_{Kt} &= \varepsilon_{2t} \\
&\vdots \\
\gamma_{1M}Y_{1t} + \gamma_{2M}Y_{2t} + ... + \gamma_{MM}Y_{Mt} + \beta_{1M}X_{1t} + ... + \beta_{KM}X_{Kt} &= \varepsilon_{Mt}
\end{aligned} \tag{14.1}$$

Or, in matrix notation,

$$Y_t\Gamma + X_t\beta = \varepsilon_t \tag{14.2}$$

Y variables in the above equation are considered endogenous, while X variables are exogenous. The investigator must properly define endogeneity based on the underlying theory. For example, in a supply and demand model, variables such as income, rainfall, and taxes will usually be exogenous to the model while quantity and price are certainly endogenous. Note that any lagged variables are typically considered exogenous since their value is predetermined.

14.2 Ignoring Simultaneity

If we ignore the fact that the structural equations for a simultaneous system of equations and estimate each equation separately, we will generate biased and inconsistent estimates. This implies that all forecasts, predictions, and hypothesis tests will be invalid.

To see this, consider the structural equations for the following macroeconomic model of consumption C, income Y, and investment I.

$$C = \alpha + Y\beta + u \tag{14.3}$$
$$Y = C + I \tag{14.4}$$

Solving for the reduced form equations gives

$$C = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta}I + \frac{1}{1-\beta}u \tag{14.5}$$
$$Y = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta}I + \frac{1}{1-\beta}u \tag{14.6}$$

If we estimate Equation 14.3 on its own, then we violate the assumption that $Y \cdot u = 0$ since we know from the reduced form equations that Y is correlated with u. The violation of this assumption gives us biasedness since

$$\hat{\beta} = \left(Y'Y\right)^{-1} Y'C = \left(Y'Y\right)^{-1} Y'\left(Y\beta + u\right) = \beta + \left(Y'Y\right)^{-1} Y'u \tag{14.7}$$

Thus the expected value of the estimate is not equal to the true value. It can also be shown that

$$\lim_{n\to\infty} \hat{\beta} = \beta + \frac{\left(1-\beta\right)\sigma_u^2}{\sigma_I^2 + \sigma_u^2} \neq \beta \tag{14.8}$$

Therefore, $\hat{\beta}$ is inconsistent. The fraction in the above expression is known as the **simultaneous equaton bias**. Unbiasedness and inconsistency render the estimation useless in about every respect, so this error is severe.

14.3 Recursive Systems

As a side note, we introduce the concept of recursive systems. If Γ is an upper-triangular matrix, then we have a **recursive system** of the form

$$\begin{aligned} Y_{1t} &= f_1\left(X_t\right) + \varepsilon_{1t} \\ Y_{2t} &= f_2(Y_{1t}, X_t) + \varepsilon_{2t} \\ &\vdots \\ Y_M &= f_M\left(Y_{1t}, Y_{2t}, ..., Y_{M-1,t}, X_t\right) + \varepsilon_{Mt} \end{aligned} \tag{14.9}$$

These are nice identifiable systems since Y_1 is clearly determined by X, which are then combined to determine Y_2, and so on.

14.4 Indirect Least Squares (Reduced Form Estimation)

It may be that a certain model lists multiple equations that have common variables among them. For example, we might have a model for the corn market whose **structural equations** are

$$\begin{aligned} QD &= \alpha_0 + \alpha_1 P + \alpha_2 Y + u && (14.10)\\ QS &= \beta_0 + \beta_1 P + \beta_2 R + v && (14.11)\\ QS &= QD && (14.12) \end{aligned}$$

where P is price, Y is income, and R is rainfall.

In this above example, we can solve for P and plug this equation into the equation for $Q = QD = QS$. This gives the **reduced form equations**

$$\begin{aligned} P &= \left(\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1}\right) + \left(\frac{-\alpha_2}{\alpha_1 - \beta_1}\right) Y + \left(\frac{\beta_2}{\alpha_1 - \beta_1}\right) R + \left(\frac{v - u}{\alpha_1 - \beta_1}\right)\\ P &= \gamma_0 + \gamma_1 Y + \gamma_2 R + \varepsilon_1 && (14.13)\\ Q &= (\alpha_0 + \alpha_1 \gamma_0) + (\alpha_1 \gamma_1 + \alpha_2) Y + (\alpha_1 \gamma_2) R + \varepsilon_2\\ Q &= \mu_0 + \mu_1 Y + \mu_2 R + \varepsilon_2 && (14.14) \end{aligned}$$

where the γ and μ coefficients are equal to the equivalent terms in parentheses above.

The reduced form of the general system of equations represented in Equation 14.2 is

$$Y_t = X_t\left(-\beta\Gamma^{-1}\right) + \left(\varepsilon_t \Gamma^{-1}\right) = X_t\left(\Pi\right) + \left(\nu_t\right) \qquad (14.15)$$

This implies a **completeness condition** that states that Γ^{-1} must exist (Γ nonsingular) for this reduced-form solution to exist. We refer to Π as the reduced form coefficients and ν_t as the reduced form errors.

Since we assume $\mathbb{E}\left[\varepsilon_t\right] = 0$, $\mathbb{E}\left[\varepsilon_t \varepsilon_t'\right] = \Sigma$, and $\mathbb{E}\left[\varepsilon_t \varepsilon_s'\right] = 0\ \forall t \neq s$, then it will be true that $\mathbb{E}\left[\nu_t\right] = 0$ and $\mathbb{E}\left[\nu_t \nu_t'\right] = \Gamma^{-1\prime}\Sigma\Gamma^{-1} = \Omega$

It can be shown that the reduced form estimator $\hat{\Pi}$ is consistent. Therefore, plim $\hat{\Pi} = \Pi$. Furthermore, as long as X doesn't contain lagged jointly dependent variables, then $\mathbb{E}\left[\hat{\Pi}\right] = \Pi$, so we have unbiasedness as well. This means that the reduced form will be helpful in estimating the structural equation parameters *if we can solve for the structural estimates given only the reduced form estimates.*

In each equation we will have one exogenous Y variable be the *dependent variable*[1], so it will have a coefficient of 1. This is just a normalization and not

[1] Note that endogenous and dependent are *not* synonyms. Dependent variables are necessarily endogenous, but definitely not vice versa.

a true restriction. However, our theory may impose other restrictions on the model. In equations 14.10 through 14.12, we have two exogenous variables (Y and R) and two endogenous variables (Q and P,) but all four variables do not appear in both equations (as their coefficients are assumed to be zero.) Their absence is a result of the implicit assumption that quantity supplied does not depend on income and quantity demanded does not depend on rainfall. These underlying restrictions in fact make the equations estimable, as we will see.

14.4.1 The Identification Problem

The question at hand is if we estimate the reduced form equation and get estimates $\hat{\gamma}$ and $\hat{\mu}$, will we be able to uniquely solve for $\hat{\alpha}$ and $\hat{\beta}$? Can we go back to our structural form from our reduced form estimates? This question is a matter of computability, but is fundamental because if we know *a priori* that we will be unable to deduce our structural form parameters, then we should not spend time and energy to do so.

There are three possible outcomes when trying to go from reduced form estimates back to estimates of a given structural equation. It will either be impossible, uniquely possible, or there will exist an infinite number of possible solutions (although there will be some restrictions on the possible values.) These conditions, respectively, are called **unidentified** (or **underidentified**,) **exactly identified**, and **overidentified**. This problem in general is referred to as the **identification problem**.

One good example of underidentification is the familiar graph of linear supply and demand schedules. If we gather observations of (p_t, q_t), we are (by assumption) observing different equilibria of the system. As we observe different equilibria points, it could be that the demand line has shifted, the supply line has shifted, or both have shifted. In fact, the demand and supply lines could be shifted such that *any* point in $\mathbb{R}^2$ is an equilibrium point. Therefore, if we allow the slopes and intercepts of each line to be free, then we can never estimate all four parameters using only the two equations. Given some string of observed equilibrium points, we could construct a variety of supply and demand equations that shift in such a way to generate the data observed. These two possibilities cannot be separated using our data. Therefore, they are **observationally equivalent**.

The brute-force method of proceeding is to take the structural equation, write out the reduced form, and express the reduced form coefficients in terms of the structural form coefficients. Then, if you are able to solve for the structural coefficients in terms of the reduced form coefficients, you have identification. This method will highlight exactly which equations are identifiable.

A more analytical approach is to consider necessary and sufficient conditions for identifiability. Recall from equation 14.1 that we have M^2 coefficients on the endogenous variables to estimate using only M equations. This is impossible without further restrictions. Restrictions that can be made are

1. Normalizations - such as setting $\gamma_{ij} = 1$ for the dependent variable
2. Identities - equations such as $QS = QD$ restrict the model.
3. Exclusions - we know that rainfall won't affect quantity demanded, so we exclude R from the QS equation.
4. Other Restrictions - such as linear restrictions, restrictions on disturbance covariance matrix, and nonlinearities.

Given enough restrictions, we can whittle down the number of parameters until we have an estimable system of equations. This gives us the order condition.

14.4.2 The Order Condition

The order condition is a sufficient condition for avoiding unidentifiability but only a necessary condition for exact identifiability. So, if it is satisfied, we can exclude the possibility of the equation being unidentified, but we cannot guarantee a unique solution.

The order condition looks at how many of the total number of variables in the system are *not* present in a given structural equation. If we have enough variables omitted from an equation, then identification is possible. Let M_j and K_j be the number of *included* endogenous and exogenous variables in equation j, respectively, and let M_j^* and K_j^* be the number of *excluded* endogenous and exogenous variables in equation j.

Condition 14.1 *If an equation j is identified, then $K_j^* \geq M_j - 1$.*

Remark 14.1 *This is easy to remember. Just say "**ex**cluded **ex**ogenous variables $\geq$ **in**cluded **en**dogenous variables $-$ 1"*

It is important to stop here and note that some textbooks use a misleading statement of the order condition. Here, we have assumed that M_j includes the dependent variable. We subtract 1 from M_j to get the number of non-dependent endogenous variables. Some textbooks define M_j to be the number of non-dependent endogenous variables. For those books, $M = M_j + M_j^* + 1$. In that case, the order condition is $K_j^* \geq M_j$. When reading a text on this topic, make sure you understand which definition is being used.

Satisfying the order condition only guarantees that the equation is not unidentified. We require a further condition to achieve exact identification.

14.4.3 The Rank Condition

The **rank condition** is a sufficient condition for exact identifiability. It is conceptually more difficult than the order condition.

Recall from equation 14.15 that $\Pi = -\beta\Gamma^{-1}$. Therefore, $\Pi\Gamma = -\beta$. The j^{th} column of this matrix is $\Pi\Gamma_j = -\beta_j$, which applies to the j^{th} equation in our system. In Γ_j, we have one $\gamma_{ji} = 1$ and we have several $\gamma_{jk} = 0$. These are the restrictions in our model. For each of these restrictions, label π_j as the element of Π_j associated with the $\gamma_{ji} = 1$ restriction, $\bar{\Pi}_j$ as the elements of Π_j associated with the $\gamma_{jk} = 0$ restriction, and $\tilde{\Pi}_j$ as the remaining unrestricted elements. Let π_j^*, $\bar{\Pi}_j^*$, and $\tilde{\Pi}_j^*$ be defined equivalently, but for the omitted variables. Putting this all together, we have

$$\Pi\Gamma = -\beta \tag{14.16}$$

$$\Pi\Gamma_j = -\beta_j \tag{14.17}$$

$$\begin{bmatrix} \pi_j & \tilde{\Pi}_j & \bar{\Pi}_j \\ \pi_j^* & \tilde{\Pi}_j^* & \bar{\Pi}_j^* \end{bmatrix} \begin{bmatrix} 1 \\ -\gamma_j \\ 0 \end{bmatrix} = \begin{bmatrix} \beta_j \\ 0 \end{bmatrix} \tag{14.18}$$

Expanding this matrix equation gives

$$\beta_j = \pi_j - \tilde{\Pi}_j\gamma_j \tag{14.19}$$

$$\pi_j^* = \tilde{\Pi}_j^*\gamma_j \tag{14.20}$$

Finally, in this setup, we have our rank condition.

Condition 14.2 *If* Rank $\left[\tilde{\Pi}_j^*\right] = M_j$, *then equation j is exactly identified.*

This condition is hardly intuitive. There is an equivalent rank condition (with its own equivalent order condition) that is easier to use, particularly since it does not require taking the inverse of a big coefficient matrix.

Define

$$A = \begin{bmatrix} \Gamma \\ \beta \end{bmatrix} = \begin{bmatrix} 1 & A_1 \\ -\gamma_j & A_2 \\ 0 & A_3 \\ -\beta_j & A_4 \\ 0 & A_5 \end{bmatrix} \tag{14.21}$$

where we simply move the coefficients for the j^{th} equation to the first column and leave the other coefficients ambiguously labelled $A_1, ..., A_5$. Note that the columns of A are linearly independent. If they weren't, then two of the structural equations would be linearly dependent and one could be eliminated.

It can be shown that if there exists some non-trivial (i.e., non-zero) vector f_j such that

$$\begin{bmatrix} A_3 \\ A_5 \end{bmatrix} [f_j] = A_0 f_j = 0 \tag{14.22}$$

then the structural equation might be able to take on an alternate form and the two would be observationally equivalent (i.e., inseparable.) So, to make this impossible, we require

$$\operatorname{Rank} A_0 = M - 1 \tag{14.23}$$

so that f_j has only the trivial solution, since $f_j = A_0^{-1} 0 = 0$.

The number of rows of A_0 is the total number of excluded coefficients, $K_j^* + M_j^*$. The number of columns of A is M, so the number of columns of A_0 is $M - 1$. If the number of rows is less than the number of columns, then it must be that the rank of A_0 is less than the number of columns, which means the rank is less than $M - 1$. So, it is necessary for the number of rows to be greater than or equal to the number of columns. If $K_j^* + M_j^* \geq M - 1$, then $K_j^* \geq M_j - 1$, which is the order condition from above. Therefore, satisfying this rank condition also satisfies the necessary order condition.

If $K_j^* > M_j - 1$, then we have an overidentified equation. Therefore, for exact identification, we must satisfy the rank condition *and* have equality in the order condition.

To operationalize this concept, we will consider an example shortly. The following list summarizes the algorithm for checking the rank condition of the j^{th} equation.

1. Stack the Γ matrix on top of the β matrix.
2. Move the j^{th} column (which corresponds to the j^{th} equation) to the first column.
3. For any non-zero entry in the first column, remove the entire row.
4. Remove the first column, which now consists of all zeros.
5. If the rank of the resulting matrix is equal to the number of columns, then the rank condition is satisfied.

If we satisfy the rank (and order) conditions, then we have that the ILS procedure is consistent and efficient. If X contains no lagged jointly dependent variables, then ILS is also unbiased. Therefore, under exact identification, ILS provides us with BLUE estimates for the structural equations.

In reality, ILS procedures are not often used because most estimable models are overidentified. In these cases, the ILS procedure will only produce linear restrictions on the possible structural form coefficients. There exist other methods of estimation that do not have this sensitivity to identification. Before proceeding, we consider a useful ILS example.

14.4.4 Klein's Macro Model 1: An ILS Example

The following widely-used macro model serves as a good ILS example.

The structural equations for consumption, investment, wages, output, profits, and capital stock are

$$C_t = \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 (W_t^p + W_t^g) + \varepsilon_{1t} \tag{14.24}$$

$$I_t = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} + \varepsilon_{2t} \tag{14.25}$$

$$W_t^p = \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 A_t + \varepsilon_{3t} \tag{14.26}$$

$$X_t = C_t + I_t + G_t \tag{14.27}$$

$$P_t = X_t - T_t - W_t^p \tag{14.28}$$

$$K_t = K_{t-1} + I_t \tag{14.29}$$

The endogenous variables are

$$Y_t = \begin{bmatrix} C_t & I_t & W_t^p & X_t & P_t & K_t \end{bmatrix} \tag{14.30}$$

The exogenous variables are

$$X_t = \begin{bmatrix} 1 & W_t^g & G_t & T_t & A_t & P_{t-1} & K_{t-1} & X_{t-1} \end{bmatrix} \tag{14.31}$$

The combined model $Y\Gamma = X\beta + E$ gives the parameter matrices

$$\Gamma = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & -1 \\ -\alpha_3 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & -\gamma_1 & 1 & -1 & 0 \\ -\alpha_1 & -\beta_1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{14.32}$$

$$\beta = \begin{bmatrix} \alpha_0 & \beta_0 & \gamma_0 & 0 & 0 & 0 \\ \alpha_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & \gamma_3 & 0 & 0 & 0 \\ \alpha_2 & \beta_2 & 0 & 0 & 0 & 0 \\ 0 & \beta_3 & 0 & 0 & 0 & 1 \\ 0 & 0 & \gamma_2 & 0 & 0 & 0 \end{bmatrix} \tag{14.33}$$

We will look at the consumption equation only. First, consider the necessary order condition. Of the 6 endogenous variables, 3 are included. Of the 8 exogenous variables, 5 are excluded. Therefore, $5 = K_1^* \geq (M_1 - 1) = 2$. This means we apparently have an overidentified equation. Specifically, it is overidentified by (at least) 3 restrictions. However, since it is not unidentified, we proceed to demonstrate the sufficient condition.

We proceed to check the sufficient rank condition by identifying those submatrices of coefficients on variables that do *not* appear in the consumption equation. These are the rows from the two matrices above that have zeros in the first column (which corresponds to the first equation.) This gives the matrix

$$\left[\begin{array}{cccccc} 0 & 1 & 0 & -1 & 0 & -1 \\ 0 & 0 & -\gamma_1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & \gamma_3 & 0 & 0 & 0 \\ 0 & \beta_3 & 0 & 0 & 0 & 1 \\ 0 & 0 & \gamma_2 & 0 & 0 & 0 \end{array}\right] = \left[\begin{array}{cc} 0 & A_3 \\ 0 & A_5 \end{array}\right] \tag{14.34}$$

Of course, we can drop the first column to get

$$A_0 = \left[\begin{array}{c} A_3 \\ A_5 \end{array}\right] = \left[\begin{array}{ccccc} 1 & 0 & -1 & 0 & -1 \\ 0 & -\gamma_1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & \gamma_3 & 0 & 0 & 0 \\ \beta_3 & 0 & 0 & 0 & 1 \\ 0 & \gamma_2 & 0 & 0 & 0 \end{array}\right] \tag{14.35}$$

Since we have more rows in A_0 than columns, we (again) over-satisfy the order condition. For the rank condition, we know that the rank cannot be greater than 5 since we have only 5 columns. So, we need only to find 5 linearly independent rows. The other 3 will consequently be linearly dependent. Choosing rows 3 through 7 gives us the required number of independent rows (check this for yourself,) thus proving that Rank $A_0 = 5$.

We therefore have the rank condition satisfied and the order condition "over-satisfied." The conclusion is that the equation is overidentified, so if we were to estimate the reduced form coefficients and try to use them to solve for $\alpha_0, ..., \alpha_3$, we would only be able to specify a range of values that would work.

14.5 Lagrange Multiplier Test for Omitted Variables

blah

14.6 Conclusion

In simultaneous equations applications, we have seen that OLS is inconsistent and should thus be avoided. Furthermore, we demonstrated that ILS may be appropriate if we know our equations are exactly identified. Unfortunately, this is a stringent condition and is not normally met in real applications.

In the next section, we will examine general estimation procedures that are well adapted for the simultaneous equations application. However, they are general enough to merit a separate chapter. As you read on, remember that simultaneous equations is a specific application of the procedures introduced in the next section.

Chapter 15

Correlation Between Regressors and Error

15.1 Introduction

The OLS assumptions imply that X be uncorrelated with ε. If we violate OLS assumption $A2$ (non-random regressors,) then we lose the fact that $\text{Cov}[X_t \varepsilon_t] = 0$[1]. This could happen in a simultaneous equations setting or when we measure our explanatory variables with error (so that they are random variables.) We saw in the Section 14.2 that this nonzero covariance causes estimate inconsistency, which makes our estimates effectively useless. In that particular framework, endogenous variables have a random component as they are dependant on equations of other variables. Therefore, in some of the structural equations, we have random explanatory variables.

In any of these cases, we need to deal with this problem in order to make useful inferences. One particular answer for simultaneous equations is Indirect Least Squares. However, if we have an identification problem or if we have exogenous variables measured with error, we need an alternative solution.

15.2 Instrumental Variables

A desirable solution to the problem of regressors correlated with error would be to replace the "bad" regressors with some **instrumental variables** that are not correlated with the error term, but are highly correlated with the X variables they replace. We define Z to be the matrix of variables *after* the

[1] Violating this assumption may not be either sufficient or necessary for creating correlation between ε and X. It is presented as an example.

"bad" regressors have been replaced by their instrumental variables. So, if we had three X variables and X_2 was correlated with the error term, then

$$X = \left[X_1 \vdots X_2 \vdots X_3 \right] \quad Z = \left[X_1 \vdots Z_2 \vdots X_3 \right] \tag{15.1}$$

where Z_2 is the "instrument" for X_2. Note that Z will have at least as many variables as X.

If Z is uncorrelated with the error term, we will arrive at consistent estimates. Use of the IV procedure is quite general - it can be applied whenever correlation exists between regressors and the error term. Therefore, OLS is in fact a special case of IV estimation.

Unfortunately, the general setup of instrumental variables does not offer suggestions for what variables Z to use. It assumes that the investigator can find desirable "instruments" for their purpose. We will proceed with the assumption that the ideal instruments have been found. In two-stage least squares, we will see how this problem can be circumvented.

We proceed on the following assumptions:

1. We have L instrumental variable regressors Z and K original X variables, where $L \geq K$ (typically $L = K$.)

2. $\mathbb{E}\left[\varepsilon_i | X_i\right] = \eta_i$

3. $\mathbb{E}\left[\eta_i\right] = 0$

4. $\text{Var}\left[\eta_i\right] = \kappa^2 < \infty$

   This implies that $\text{Var}\left[\varepsilon_i\right] = \sigma^2 + \kappa^2$, so variation around ε is partly due to variation of the regressors.

5. $\text{Cov}\left[X_i, \varepsilon_i\right] = \text{Cov}\left[X_i, \eta_i\right] = \gamma$

   This implies that $\text{plim}\, \frac{1}{n} X'\varepsilon = \gamma$ by Khinchine's Weak Law of Large Numbers.

6. $\mathbb{E}\left[x_{ik}^2\right] = Q_{XX,kk}$

7. $\mathbb{E}\left[z_{il}^2\right] = Q_{ZZ,ll}$

   This implies that $\text{plim}\, \frac{1}{n} Z'Z = Q_{ZZ}$ (a PSD matrix)

8. $\mathbb{E}\left[z_{il} x_{ik}\right] = Q_{ZX,lk}$

   This implies that $\text{plim}\, \frac{1}{n} Z'X = Q_{ZX}$ (an $L \times K$ matrix of rank K)

9. $\mathbb{E}\left[\varepsilon_i | z_i\right] = 0$

   This implies that $\text{plim}\, \frac{1}{n} Z'\varepsilon = 0$

Note that the standard OLS model fits in the IV framework with $\eta = 0$, $\gamma = 0$.

If $\eta \neq 0$, then $\mathbb{E}\left[b|X\right] = \beta + \left(X'X\right)^{-1} X'\eta \neq \beta$, so we have unbiasedness of the standard (unadjusted) OLS estimate b. Furthermore, $\operatorname{plim} b = \beta + \operatorname{plim}\left(\frac{1}{n}X'X\right)^{-1} \operatorname{plim}\left(\frac{1}{n}X'\varepsilon\right) = \beta + Q_{XX}^{-1}\gamma \neq \beta$, so we don't have consistency either (which we've seen.)

We use the following to develop our IV estimator.

$$Y = X\beta + \varepsilon \tag{15.2}$$

$$Z'Y = Z'X\beta + Z'\varepsilon \tag{15.3}$$

$$\operatorname{plim}\left(\frac{1}{n}Z'Y\right) = \operatorname{plim}\left(\frac{1}{n}Z'X\beta\right) + \operatorname{plim}\left(\frac{1}{n}Z'\varepsilon\right) \tag{15.4}$$

$$\operatorname{plim}\left(\frac{1}{n}Z'Y\right) = \operatorname{plim}\left(\frac{1}{n}Z'X\right)\beta \tag{15.5}$$

$$\left[\operatorname{plim}\left(\frac{1}{n}Z'X\right)\right]^{-1}\left[\operatorname{plim}\left(\frac{1}{n}Z'Y\right)\right] = \beta \text{ (if } L = K) \tag{15.6}$$

$$\operatorname{plim}\left(\left(Z'X\right)^{-1} Z'Y\right) = \beta \tag{15.7}$$

$$\operatorname{plim}\left(\hat{\beta}_{IV}\right) = \beta \tag{15.8}$$

So, by choosing $\hat{\beta}_{IV} = \left(Z'X\right)^{-1} Z'Y$, we have developed a consistent estimator, as long as $L = K$.

The following are properties of $\hat{\beta}_{IV}$

1. $\hat{\beta}_{IV} \overset{a}{\sim} N\left[\beta, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1}\right]$
2. $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - X_i'\hat{\beta}_{IV}\right)^2$
3. $\mathbb{E}st.Asy.\operatorname{Var}\left[\hat{\beta}_{IV}\right] = \hat{\sigma}^2 \left(Z'X\right)^{-1}(Z'Z)(X'Z)^{-1}$

If we have more variables in Z than we do X, then $Z'X$ will not be nonsingular. We need to "shrink" Z down to having only K variables. Instead of throwing away information by simply removing $L - K$ variables from Z, we will instead project the columns of X into the column space of Z to get

$$\hat{X} = Z\left(Z'Z\right)^{-1} Z'X \tag{15.9}$$

which gives $\hat{\beta}_{IV} = \left(\hat{X}'X\right)^{-1}\hat{X}'Y = \left(X'Z\left(Z'Z\right)^{-1}Z'X\right)^{-1} X'Z\left(Z'Z\right)^{=1}Z'Y$

Note that the estimated asymptotic variance remains unchanged (check this yourself.)

15.3 Two-Stage Least Squares

Two-stage least squares is a method of developing instrumental variables from nothing more than the original variables. We use the OLS procedure to generate predicted values of the problem variables and then substitute those predicted values in as the instruments for the variables.

The procedure is as follows:

1. Identify all "problem" regressors W. These are explanatory variables that are either endogenous (in a simultaneous equations setting) or random.
2. For all of these problem variables, generate predicted values $\hat{W}$ using an OLS model with all non-problematic variables as regressors.
3. Substitute the predicted values into the original model and estimate the model using OLS (or whatever procedure is important.)

Using this method will generate consistent estimates since it can be shown that the predicted values are valid instrumental variables, and we know that the IV procedure produces consistent estimates.

Furthermore, 2SLS (2 Stage Least Squares) is identical to ILS when applied to an exactly identified equation.

Example 15.1 *The following supply & demand model is proposed*

$$\textit{Demand: } q = \alpha_0 + \alpha_1 p + \alpha_2 y + u \tag{15.10}$$

$$\textit{Supply: } q = \beta_0 + \beta_1 p + \beta_2 r + \beta_3 f = v \tag{15.11}$$

where y is income, p is price (endogenous,) r is rainfall, and f is fertilizer. Since p is endogenous in this model, we generate predicted values $\hat{p}$ using

$$p = \gamma_0 + \gamma_1 y + \gamma_2 r + \gamma_3 f + \varepsilon \tag{15.12}$$

We choose this particular regression for p because it contains all "non-problematic" exogenous variables in the model. The OLS estimate gives us a vector $\hat{p}$ of (non-random) predicted values. We then substitute these values into our structural equations.

$$\textit{Demand: } q = \alpha_0 + \alpha_1 \hat{p} + \alpha_2 y + u \tag{15.13}$$

$$\textit{Supply: } q = \beta_0 + \beta_1 \hat{p} + \beta_2 r + \beta_3 f = v \tag{15.14}$$

This procedure will yield consistent estimates of the structural equations.

Chapter 16

Qualitative Response Models

16.1 Introduction

Here we consider various models that have discrete variables as the dependent term. For example, we may want to model the probability that a smoker dies of lung cancer, but all we can observe is death by lung cancer or death by other causes. In order to study this in a statistical model, we must first "code" the observations and then develop a model with desirable properties from which we can derive desirable estimators of the model parameters and make inferences about the effect of other variables on our dependent term. We will find that there are several ways to develop desirable estimates for the parameters.

16.2 Linear Probability Model

The linear probability model is the "original" binary response model. It is basically an OLS extension and suffers from various problems. Most notably, these models can predict probabilities outside the unit interval. Furthermore, several OLS assumptions on the error variance are violated. Regardless, we will quickly develop this model to give an understanding of the problems that arise.

In a linear probability model, the dependent variable is binary, where $Y_t \in \{0, 1\}$. We observe only 1's and 0's in our data, but we're actually interested in $\mathbb{P}[Y_t = 1|X] = p_t$. Since our linear model is written as $Y_t = \alpha + \beta X_t + u_t$, then $\mathbb{P}[Y_t = 1] = \mathbb{P}[u_t = 1 - \alpha - \beta X_t]$. Note that $u_t \in \{1 - \alpha - \beta X_t, -\alpha - \beta X_t\}$. Therefore, u_t is properly modelled as a binomial random variable. As before, we assume that $\mathbb{E}[u_t] = p_t(1-\alpha-\beta X_t)+(1-p_t)(-\alpha-\beta X_t) = 0$. Solving for p_t gives $p_t = \alpha+\beta X_t$, with variance $\mathbb{E}[u_t^2] = \sigma_t^2 = p_t(1-\alpha-\beta X_t)^2+(1-p_t)(-\alpha-\beta X_t)^2 = p_t(1 - p_t) = (1 - \alpha - \beta X_t)(\alpha + \beta X_t)$.

Notice that the variance calculated directly depends on X_t, so we have that $\sigma_t^2 \neq \sigma_s^2$. Therefore, the linear probability model suffers from heteroscedasticity. As before, we can use a FGLS procedure to recover asymptotically efficient estimates. However, this model is not restricted to $\hat{Y}_t \in (0,1)$. Note that if $\hat{Y}_t \notin (0,1)$, then $\hat{\sigma}_t^2 = \hat{Y}_t(1-\hat{Y}_t) \leq 0$, which is nonsensical.

Another problem is that the coding of $\{0,1\}$ is somewhat arbitrary. If we used $\{0,2\}$ or $\{5,19\}$, we would get drastically different results. If our variable can take on multiple values, we may code them $\{0,1,2\}$, but again this is arbitrary and can lead to various problems.

The only advantage of the linear probability model is that the coefficients have a very natural interpretation. If $\hat{\beta}_i = 0.02$, then we claim that an increase of X_i by one unit (and all else constant) leads to a 0.02 increase in the response probability.

Because of the various problems in this model, we avoid using linear probability models in favor of the more desirable models described below.

16.3 The Log-Odds Ratio

Our goal is to estimate $\mathbb{P}[Y_i = j] = \theta_j$, which we assume is constant across observations i, but not across categories j..

Define the log-odds ratio as

$$\lambda_j = \log\left[\frac{\theta_j}{1-\theta_j}\right] \tag{16.1}$$

Assume that θ_j takes the logistic distribution, so

$$\theta_j = \frac{e^{X_j\beta}}{1+e^{X_j\beta}} \tag{16.2}$$

The log-odds ratio is therefore

$$\lambda_j = \log\left[\frac{\frac{e^{X_j\beta}}{1+e^{X_j\beta}}}{\frac{1}{1+e^{X_j\beta}}}\right] = \log\left[e^{X_j\beta}\right] = X_j\beta_j \tag{16.3}$$

We now have that the log-odds ratio is linear in our variables X and our parameters β.

Before considering how to estimate $\hat{\beta}_j$, let's look at how it will be interpreted. What we would like to know is how θ_i changes with a per-unit change in X_{ij}. Taking the derivative gives

$$\frac{\partial\theta_i}{\partial X_{ij}} = \theta_i(1-\theta_i)\beta_j \Rightarrow \beta_j = \frac{\partial\theta_i/\partial X_{ij}}{\theta_i(1-\theta_i)} \tag{16.4}$$

The best way to understand the coefficient is to pick an "average" data point X_i for some "average" individual i. For this "typical" data observation, we can say that $\partial\theta_i/\partial X_{ij} = \hat{\theta}_i(1-\hat{\theta}_i)\hat{\beta}_j$. A simulation of various X_i values could be used to study the effect on θ_i. Unfortunately, our analysis is sensitive to the individual and data point chosen, so we must analyze the results for only specific scenarios.

We don't observe θ_i or β. However, we do observe relative frequencies. Let $R_j = \sum_{i=1}^{n_j} \{Y_i = j\}$ be the number of times outcome j was observed. We estimate θ_j by $\hat{\theta}_j = R_j/n_j$, the relative frequency of "successes" of outcome j. The log-odds ratio can then be estimated by

$$\hat{\lambda}_j = \log\left(\frac{\hat{\theta}_j}{1-\hat{\theta}_j}\right) \tag{16.5}$$

and we can now use the linear model

$$\hat{\lambda} = X\beta + \varepsilon \tag{16.6}$$

to estimate β. However, we have heteroscedasticity among our error terms, so weighted least squares is the appropriate estimate technique. To do WLS, we divide by the square root of the variance.

This example is actually the binary logit model introduced in the next section. However, it is transformed into a linear model through the log-odds transformation.

16.4 Binary Logit & Probit Models

In a simple dichotomy (or, binary) model, our response variable takes one of two possible values, which we code $\{0, 1\}$. For example, if a certain dosage of a drug causes death in a rat, we label the observation a "0", otherwise we label it a "1". The standard assumption is that there exists some underlying (or, *latent*) variable Y^* that is truly driving behavior. In our rat example, Y_t^* would represent the maximum dosage that the t^{th} rat can handle before dying (also known as the rat's tolerance,) possibly plus or minus some constant. Of course, we can only observe Y and not Y^*. Our "true" model in this example will be $Y_t^* = X_t\beta - u_t$, where $Y_t = 1\{Y_t^* > 0\} + 0\{Y_t^* \leq 0\} = \{Y_t^* > 0\}$. If c as the actual dosage being given the rats, then a rat will survive if his tolerance is sufficiently high. We know that tolerance $T_t = Y_t^* + k$ for some constant k. If $T_t > c$, then the rat can withstand the dosage and will survive and we will observe $Y_t = 1$ (survival.) This is equivalent to $T - c > 0$, or $Y_t^* > 0$ (by setting the arbitrary constant $k = c$.) Therefore, the condition $Y_t^* > 0$ can be thought of as the rat withstanding the dosage.

The probability of a “success” (withstanding the dosage) is written as

$$\theta_t = \mathbb{P}[Y_t = 1] = \mathbb{P}\left[X_t\beta - u_t > 0\right] = \mathbb{P}[X_t\beta < u_t] = F_{ut}\left(X_t\beta\right) \tag{16.7}$$

where $F_{ut}(\cdot)$ is the cdf of the error terms u_t.

The difference between a **logit model** and a **probit model** is in the assumption of this distribution F.

In the logit model, we assume that u_t is distributed by the logistic distribution

$$\theta_t = F_{ut}(X_t\beta) = \frac{1}{1+e^{-X_i\beta}} = \frac{e^{X_t\beta}}{1+e^{X_t\beta}} \tag{16.8}$$

In the probit model, we assume that u_t/σ is distributed by the standard normal distribution

$$\theta_t = F_{ut}(X_t\beta) = \Phi(X_t\beta) = \int_{-\infty}^{X_t\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{\lambda^2}{2}} d\lambda \tag{16.9}$$

where $\Phi(\cdot)$ is the notation for the standard normal cdf function (which does not have a closed-form representation.)

Can you guess the distribution function of u_t in the linear probability model from Section 16.2[1]?

16.5 Binary Model Estimation

16.5.1 Maximum Likelihood Estimates

The likelihood of a single observation (or, the probability of observing a single observation) is simply

$$\mathbb{P}[Y_t = 1]^{Y_t}\mathbb{P}[Y_t = 0]^{(1-Y_t)} = F_u\left(X_t\beta\right)^{Y_t}\left(1 - F_u\left(X_t\beta\right)\right)^{(1-Y_t)} \tag{16.10}$$

Therefore our likelihood function and log-likelihood function for all T observations are

$$L(\beta|X) = \prod_{t=1}^{T} F_u\left(X_t\beta\right)^{Y_t}\left(1 - F_u\left(X_t\beta\right)\right)^{(1-Y_t)} \tag{16.11}$$

$$\mathcal{L}(\beta|X) = \sum_{t=1}^{T} Y_t \log\left[F_u(X_t\beta)\right] + \sum_{t=1}^{T}(1-Y_t)\log\left[1 - F_u\left(X_t\beta\right)\right] \tag{16.12}$$

[1] The answer is that u_t is modelled *as if* it were distributed uniformly, so that $F_u(X_t\beta) = X_t\beta$, meaning that $P_t = X_t\beta$. However, this is somewhat of a trick question since $X_t\beta$ does not necessarily lie in $[0, 1]$. Therefore $X_t\beta$ is not a proper pdf.

Note that in the Logit Model, these would be

$$L(\beta|X) = \prod_{t=1}^{T} \left(\frac{e^{X_t\beta}}{1+e^{X_t\beta}}\right)^{Y_t} \left(\frac{1}{1+e^{X_t\beta}}\right)^{(1-Y_t)} \tag{16.13}$$

$$\mathcal{L}(\beta|X) = \sum_{t=1}^{T} Y_t \log\left[\frac{e^{X_t\beta}}{1+e^{X_t\beta}}\right] + \sum_{t=1}^{T}(1-Y_t)\log\left[\frac{1}{1+e^{X_t\beta}}\right] \tag{16.14}$$

Our goal is to maximize $\mathcal{L}(\beta|X)$. The first-order conditions are that $\hat{\beta}$ solves

$$\frac{\partial\mathcal{L}}{\partial\beta} = \sum_{t=1}^{T} \frac{Y_t X_t f_u(X_t\hat{\beta})}{F_u(X_t\hat{\beta})} + \sum_{t=1}^{T} \frac{(1-Y_t)\, X_t \left(-f_u(X_t\hat{\beta})\right)}{1-F_u(X_t\hat{\beta})} =$$
$$\sum_{t=1}^{T}\left(\frac{Y_t}{F_{ut}} + \frac{Y_t - 1}{1-F_{ut}}\right) X_t f_{ut} = \sum_{t=1}^{T}\left(\frac{Y_t - F_{ut}}{F_{ut}(1-F_{ut})}\right) X_t f_{ut} = 0$$

where $F_{ut} = F_u(X_t\hat{\beta})$ and $f_{ut} = f_u(X_t\hat{\beta})$[2].

The value $\hat{\beta}_{MLE}$ solves this equation. However, we don't have a nice closed-form representation of this solution. Instead, there exist other methods of calculating estimates equivalent to $\hat{\beta}_{MLE}$, such as the Minimum χ^2 Method and the Method of Scoring.

Of course, we need to comment on the second-order conditions (sufficient conditions) for the maximization problem. This condition is equivalent to $\partial^2\mathcal{L}/\partial\beta^2$ being a negative definite matrix (i.e., concavity of the objective function.) The proof is rather technical, but it can be shown that global concavity exists for logit and probit models[3].

Remark 16.1 *The MLE estimator is consistent and asymptotically normal as long as we assume differentiability of $F_u(\cdot)$ (which must be a proper cdf) and that $\lim_{T\to\infty}\left[(1/n)\sum_{t=1}^{T} X_t' X_t\right]$ is finite and nonsingular*[4]*. Specifically*

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0, A^{-1}) \tag{16.15}$$

where

$$A = \lim_{T\to\infty}\left[\frac{1}{T}\sum_{t=1}^{T} \frac{f_u(X_t\beta)^2}{F_t(X_t\beta)\,(1-F_u(X_t\beta))} X_t' X_t\right] \tag{16.16}$$

[2] A useful sidenote is that for any distribution $F(x)$, the quotient $f/(1-F)$ is called the *hazard rate* and is typically denoted $H(x)$. Note that the hazard rate $H_u(X_t\hat{\beta})$ could be substituted into the first-order condition if desired. Hazard rates appear often in the mechanism design literature.

[3] See Amemiya's *Advanced Econometrics* Harvard Press 1985, Chapter 9 for a proof.

[4] Again, see Amemiya's 1985 book for a proof.

We now turn to alternate estimation procedures for the QR model that are equivalent to the MLE. These are useful since the MLE solution doesn't have a nice closed form.

16.5.2 Minimum χ^2 Estimates

blah There are actually several variations on the MINχ^2 method.

16.5.3 Method of Scoring

The **method of scoring** is an iterative procedure that generates estimates of β for the QR model that are equivalent to the MLE solution.

16.5.4 Non-Linear Weighted Least Squares (NLWLS)

blah

16.5.5 Method of Scoring and NLWLS Equivalence

blah. They're equivalent.

16.6 Multinomial Models

Multinomial models set up very similar to binary models, except that the response variable can take on more than two values. For example, we may want to model the factors that determine who drives to work, who rides a bus to work, and who takes the subway. In such a model, the dependent variable (mode of transportation) can take on three possible values.

In general, we will assume that we have T independent observations, each of which has m_t possible alternatives for the dependent variable. The model is thus

$$\mathbb{P}\left[Y_t = j\right] = F_{tj}\left(X\beta\right) \quad t \in \{1, 2, ..., T\} \; j \in \{1, 2, ..., m_t\} \tag{16.17}$$

Note that $j \neq 0$ since $F_{t0} = 1 - \sum_{j=1}^{m_t} F_{tj}$. Also note that m_t depends on t since not all alternatives will be available for all observations. Some commuters may not be able to take the subway, for example.

The log-likelihood function is

$$\mathcal{L}\left(\beta|X\right) = \sum_{t=1}^{T}\sum_{j=0}^{m_t}\{Y_t = j\}\log\left[F_{tj}\left(X\beta\right)\right] \tag{16.18}$$

To find the MLE, we set the score $(\partial\mathcal{L}/\partial\beta)$ equal to zero and solve for $\hat{\beta}$.

Most of the MLE properties that exist in the binary case will still be true here. Also, the equivalence between the method of scoring and NLWLS is maintained.

16.6.1 Independence of Irrelevant Alternatives

Definition 16.1 *A probability measure $\mathbb{P}$ satisfies* ***Independence of Irrelevent Alternatives****, or (**IIA**) if for any nonempty sets of alternatives S and T and some pair of alternative choices $(i, j) \in S$,*

$$\frac{\mathbb{P}_{i|S}}{\mathbb{P}_{j|S}} = \frac{\mathbb{P}_{i|S\cup T}}{\mathbb{P}_{j|S\cup T}} \tag{16.19}$$

where $\mathbb{P}_{i|S}$ is the probability of i being chosen when given only the set of alternatives S.

Theorem 16.2 *IIA implies that the probability of choosing some alternative i from the set S is identical to the conditional probability of choosing i from*

Theorem 16.3 *$S \cup T$ given that some alternative from $S \cup T$ will be chosen. Formally,*

$$\mathbb{P}_{i|S} = \frac{\mathbb{P}_{i|S\cup T}}{\sum\limits_{j\in S} \mathbb{P}_{j|S\cup T}} \tag{16.20}$$

Proof. *Since $(i, j) \in S \cup T$, then the IIA statement can be manipulated to get*

$$\frac{\mathbb{P}_{i|S\cup T}}{\mathbb{P}_{j|S\cup T}} = \frac{\mathbb{P}_{i|S}}{\mathbb{P}_{j|S}} \tag{16.21}$$

$$\frac{\frac{\mathbb{P}_{i|S\cup T}}{\sum_{k\in S} \mathbb{P}_{k|S\cup T}}}{\frac{\mathbb{P}_{j|S\cup T}}{\sum_{k\in S} \mathbb{P}_{k|S\cup T}}} = \frac{\mathbb{P}_{i|S}}{\mathbb{P}_{j|S}} \tag{16.22}$$

If the numerators (and denominators) of these two equations are equal, then we have our result. However, we can only conclude that they are proportional. So,

$$\frac{\mathbb{P}_{i|S\cup T}}{\sum_{k\in S} \mathbb{P}_{k|S\cup T}} = \delta \mathbb{P}_{i|S} \tag{16.23}$$

$$\frac{\mathbb{P}_{j|S\cup T}}{\sum_{k\in S} \mathbb{P}_{k|S\cup T}} = \delta \mathbb{P}_{j|S} \tag{16.24}$$

Summing over all alternatives in S for the first equation gives

$$\sum_{i\in S}\frac{\mathbb{P}_{i|S\cup T}}{\sum_{k\in S}\mathbb{P}_{k|S\cup T}} = \sum_{i\in S}\delta\mathbb{P}_{i|S} \tag{16.25}$$

$$\frac{\sum_{i\in S}\mathbb{P}_{i|S\cup T}}{\sum_{k\in S}\mathbb{P}_{k|S\cup T}} = \delta\sum_{i\in S}\mathbb{P}_{i|S} \tag{16.26}$$

$$1 = \delta \tag{16.27}$$

Therefore, the only possible constant of proportionality is $\delta = 1$. So, the numerators and denominators of equation 16.22 are each equal, proving the theorem. ■

In the logit model, we have that $\mathbb{P}_{i|S} = \frac{\exp(X_i\beta)}{\sum_{j\in S}\exp(X_j\beta)}$, so $\mathbb{P}_{i|S}/\mathbb{P}_{j|S} = \frac{\exp(X_i\beta)}{\exp(X_j\beta)}$, which doesn't depend on any other alternative k. Therefore, the IIA property holds for logit models. However, it does not hold for all models. The probit model and the nested logit violate this property, for example.

The ”Red Bus, Blue Bus” problem proposed by McFadden is often used as a strong argument *against* IIA, but is in reality an argument showing the importance of ”nesting” highly related choice alternatives. blah.

16.7 Ordered Models

blah

16.8 Hypothesis Tests

blah see Amemiya 1985 Ch9

Part III

Practice Problems

Chapter 17

Prelim Problems Solved

The following problems have been selected to be either instructive in some general way or to be good practice problems in preparation for the prelim exams[1].

17.1 Probability Problems

17.1.1 Drug Testing

Problem 1 *A particular heart disease has a prevalence of 1/1000 people. A test to detect this disease has a false positive rate of 5% (meaning 5% of healthy people incorrectly are tested as being ill.) Assume that the test diagnoses correctly every person who has the disease. What is the chance that a randomly selected person found to have a positive result actually has the disease?*

Solution 1.1 *The answer is just less than 2%. This counter-intuitive problem is often given to doctors to show how "bad" they are at statistics. Apparently, the question was given to a group of 60 Harvard Medical Students. Almost half said 95%, the average answer was 56%, and 11 students answered correctly. It is an example of "base rate neglect."*

*We use Bayes' Rule (1.36) to solve this problem. Let I indicate "**i**nfected", N indicate "**n**ot infected", and T indicate "positive **t**est result."*

$$\mathbb{P}[I|T] = \frac{\mathbb{P}[T|I]\,\mathbb{P}[I]}{\mathbb{P}[T|I]\,\mathbb{P}[I]\ +\ \mathbb{P}[T|N]\,\mathbb{P}[N]} = \frac{(1)\,(1/1000)}{(1)\,(1/1000) + (0.05)\,(999/1000)} =$$
$$\frac{1}{1 + 0.05 * 999} = \frac{1}{50.95} = 0.01\,962\,7 \approx 2\%$$

The reason so many people get this problem wrong is that they ignore the fact that the rarity of the disease (the "base weight") outweighs the seemingly

[1] Peter piper picked a peck of pickled peppers.

small error in the test. In a group of 100,000 people, we would expect 100 to be infected. From the other 99,900 people, we expect 4995 of them (5%) to return false positives. That's 5095 positive test results, of which only 100 are accurate. Therefore, 100 out of 5095 is our answer, which reduces to 1/50.95 and matches our above answer.

17.1.2 The Birthday Problem

Problem 2 *Assume the probability that a person is born on any given day is 1/365 (and exclude Feb. 29th.) Given a set of k individuals, what is the probability that at least two of them have the same birthday?*

Solution 2.1 *To solve this, consider the compliment event of no matches.* $\mathbb{P}\left[at\ least\ one\ match\right] = 1 - \mathbb{P}\left[no\ matches\right]$

The probability of no matches can be solved as a counting problem.

$\mathbb{P}\left[no\ matches\right] = \frac{\#\ ways\ to\ have\ no\ matches}{\#\ possible\ outcomes} = \frac{365 \cdot 364 \cdot \ldots \cdot\ (365-k+1)}{365^k} = \left(\frac{365!}{(365-k)!}\right) / \left(365^k\right)$

Here we are assigning birthdays to individuals, so birthdays are like buckets and individuals are like balls - we have n birthdays and k individuals. It makes sense to assign multiple individuals the same birthday, but it doesn't make sense to assign multiple birthdays to the same individual. The number of ways to assign birthdays to individuals without any matches is choosing birthdays without replacement. Since individuals are distinct, order matters. Therefore, we use the formula $n!/\left(n-k\right)!$. *For the denominator, we are looking at all possible ways to assign birthdays to individuals. Although order still matters (since individuals are still distinct,) we now allow replacement. Therefore, we use the formula* n^k.

17.1.3 The Monty Hall Problem

Problem 3 *Behind 3 labelled doors (A, B, and C) are randomly placed two goats and a car. A contestant (who values cars much more than goats) is asked to choose one of the three doors. After the choice is made, Monty Hall (who knows the location of the car) opens one of the two unchosen doors - making sure to open a door containing a goat. If both unchosen doors contain goats, he chooses between them with equal probability. He then asks the contestant to choose between the door first chosen or the remaining closed door. Whatever lies behind the chosen door will be given to the contestant Which option, if any, gives the highest probability of revealing the car? Should our contestant stay or switch?*

Solution 3.1 *This exercise in conditional probability emphasizes the importance of defining events properly. Let "A" be the event that the car is behind door A, and so on. Let* M_B *be the event that Monty opens door B, and so on.*

Assume (without loss of generality) that the contestant chooses door A. The probability that the prize is behind the chosen door given that it's not behind door B is

$$\mathbb{P}\left[A|M_B\right] = \frac{\mathbb{P}\left[A \cap M_B\right]}{\mathbb{P}\left[M_B\right]} = \frac{\mathbb{P}\left[M_B|A\right]\ \mathbb{P}\left[A\right]}{\mathbb{P}\left[M_B\right]}$$

$\mathbb{P}\left[A\right] = 1/3$ *since the prizes are placed randomly. If the car is behind door A, then it is assumed that Monty chooses the remaining doors with equal probability. Therefore,* $\mathbb{P}\left[M_B|A\right] = 1/2$.

To solve $\mathbb{P}\left[M_B\right]$, *we need to consider all possible outcomes:*
$\{A \cap M_B, A \cap M_C, B \cap M_C, C \cap M_B\}$

The event that B is chosen is $\{A \cap M_B, C \cap M_B\}$. *The probability of this event is* $\mathbb{P}\left[A \cap M_B\right] + \mathbb{P}\left[C \cap M_B\right]$.

$$\begin{aligned}\mathbb{P}\left[M_B\right] &= \mathbb{P}\left[A \cap M_B\right] + \mathbb{P}\left[C \cap M_B\right] \\ &= \mathbb{P}\left[M_B|A\right]\mathbb{P}\left[A\right] + \mathbb{P}\left[M_B|C\right]\mathbb{P}\left[C\right] \\ &= (1/2)\,(1/3) + (1)\,(1/3) = 1/2\end{aligned}$$

Putting all of this information together, we have

$$\mathbb{P}\left[A|M_B\right] = \frac{\mathbb{P}\left[M_B|A\right]\ \mathbb{P}\left[A\right]}{\mathbb{P}\left[M_B\right]} = \frac{(1/2)\,(1/3)}{(1/2)} = \frac{1}{3}$$

Given that Monty opens door B, the car can't be behind door B. Therefore,

$$\frac{2}{3} = \mathbb{P}\left[(A|M_B)^C\right] = \mathbb{P}\left[C|M_B\right]$$

So, if our contestant stays with his originally chosen door, his probability of winning the car is $1/3$, *but if he switches to the remaining unopened door, his probability of winning the car doubles to* $2/3$.

A good way to explain this problem without using statistics is to consider the following. What if there were 1,000,000 doors and, after the contestant chose one, Monty opened 999,998 other doors. Clearly, the chance that the contestant chose the right door on the first guess is 1/1,000,000. When Monty opens all but one other door, it becomes obvious that the remaining door almost certaintly contains the prize. In effect, when Monty opens doors, he "compresses" the probability weight on all the other doors into the one door he leaves closed.

17.1.4 Conditional Probability 1

Problem 4 *If* X *and* U *are independent random variables symmetric about zero, prove that* $\mathbb{P}\left[U > 0|X < U\right] \geq \mathbb{P}\left[U < 0|X < U\right]$

Solution 4.1 *For the first part, we use Baye's rule*

$$\mathbb{P}[U > 0|X < U] = \frac{\mathbb{P}[X < U|U > 0]\,\mathbb{P}[U > 0]}{\mathbb{P}[X < U|U > 0]\,\mathbb{P}[U > 0] + \mathbb{P}[X < U|U < 0]\,\mathbb{P}[U < 0]} \tag{17.1}$$

$$= \frac{\mathbb{P}[X < U|U > 0]}{\mathbb{P}[X < U|U > 0] + \mathbb{P}[X < U|U < 0]} \tag{17.2}$$

$$= \frac{\mathbb{P}[X < U|U > 0]}{\mathbb{P}[X < U]} \tag{17.3}$$

Similarly,

$$\mathbb{P}[U < 0|X < U] = \frac{\mathbb{P}[X < U|U < 0]}{\mathbb{P}[X < U]} \tag{17.4}$$

Substituting these values gives

$$\mathbb{P}[U > 0|X < U] \geq \mathbb{P}[U < 0|X < U] \tag{17.5}$$

$$\frac{\mathbb{P}[X < U|U > 0]}{\mathbb{P}[X < U]} \geq \frac{\mathbb{P}[X < U|U < 0]}{\mathbb{P}[X < U]} \tag{17.6}$$

$$\mathbb{P}[X < U|U > 0] \geq \mathbb{P}[X < U|U < 0] \tag{17.7}$$

$$\frac{\mathbb{P}[X < U \ \&\ U > 0]}{\mathbb{P}[U > 0]} \geq \frac{\mathbb{P}[X < U \ \&\ U < 0]}{\mathbb{P}[U < 0]} \tag{17.8}$$

$$\mathbb{P}[X < U \ \&\ U > 0] \geq \mathbb{P}[X < U \ \&\ U < 0] \tag{17.9}$$

$$\int_0^\infty \left(\int_{-\infty}^{u>0} f(x)\,dx\right) f(u)\,du \geq \int_{-\infty}^0 \left(\int_{-\infty}^{u<0} f(x)\,dx\right) f(u)\,du \tag{17.10}$$

$$\int_0^\infty \left(\int_{-\infty}^{u>0} f(x)\,dx\right) f(u)\,du \geq \int_{-\infty}^0 \left(\int_{-\infty}^{u<0} f(x)\,dx\right) f(u)\,du \tag{17.11}$$

Note that by symmetry of U *around zero,*

$$\int_0^\infty \left(\frac{1}{2}\right) f(u)\,du = \frac{1}{2}\int_0^\infty f(u)\,du = \frac{1}{4} = \frac{1}{2}\int_{-\infty}^0 f(u)\,du = \int_{-\infty}^0 \left(\frac{1}{2}\right) f(u)\,du \tag{17.12}$$

Similary, by symmetry of X *around zero,*

$$\int_{-\infty}^{u>0} f(x)\,dx > \frac{1}{2} > \int_{-\infty}^{u<0} f(x)\,dx \tag{17.13}$$

Therefore, we have that

$$\int_0^\infty \left(\int_{-\infty}^{u>0} f(x)\,dx\right) f(u)\,du \geq \int_0^\infty \left(\frac{1}{2}\right) f(u)\,du = \frac{1}{4}$$
$$= \int_{-\infty}^0 \left(\frac{1}{2}\right) f(u)\,du \geq \int_{-\infty}^0 \left(\int_{-\infty}^{u<0} f(x)\,dx\right) f(u)\,du \tag{17.14}$$

Of course, this proof requires symmetry of both *X and U.*

17.1.5 Conditional Probability 2

Problem 5 *Assume X and Y are independent random variables with cdfs of $F_X\left(x\right)$ and $F_Y\left(y\right)$ respectively where $\mathbb{P}\left[X \geq 0\right] = \mathbb{P}\left[Y \geq 0\right] > 0$ and $\mathbb{P}\left[X > a\right] \geq \mathbb{P}\left[Y > a\right]$ $\forall a \geq 0$ and with strict inequality for some $a \geq 0$. Show that $\mathbb{P}\left[X > Y | X, Y > 0\right] > 1/2$*

Solution 5.1 *First use the definition of conditional probability.*

$$\mathbb{P}\left[X > Y | X, Y > 0\right] = \frac{\mathbb{P}\left[X > Y \ \& \ X, Y > 0\right]}{\mathbb{P}\left[X, Y > 0\right]} \tag{17.15}$$

Next use the fact that the variables are independent.

$$\begin{align}
\mathbb{P}\left[X > Y | X, Y > 0\right] &= \frac{\mathbb{P}\left[X > Y > 0\right]}{\mathbb{P}\left[X, Y > 0\right]} \tag{17.16}\\
&= \frac{\mathbb{P}\left[X > Y > 0\right]}{\mathbb{P}\left[X > 0\right]\mathbb{P}\left[Y > 0\right]} \tag{17.17}\\
&= \frac{\int_0^\infty \int_y^\infty f_X\left(x\right) dx f_Y\left(y\right) dy}{\left(1 - F_X\left(0\right)\right)\left(1 - F_Y\left(0\right)\right)} \tag{17.18}\\
&= \frac{\int_0^\infty \left(1 - F_X\left(y\right)\right) f_Y\left(y\right) dy}{\left(1 - F_X\left(0\right)\right)\left(1 - F_Y\left(0\right)\right)} \tag{17.19}\\
&\geq \frac{\int_0^\infty \left(1 - F_Y\left(y\right)\right) f_Y\left(y\right) dy}{\left(1 - F_X\left(0\right)\right)\left(1 - F_Y\left(0\right)\right)} \tag{17.20}\\
&= \frac{\int_0^\infty f_Y\left(y\right) dy - \int_0^\infty F_Y\left(y\right) f_Y\left(y\right) dy}{\left(1 - F_X\left(0\right)\right)\left(1 - F_Y\left(0\right)\right)} \tag{17.21}\\
&= \frac{\left(1 - F_Y\left(0\right)\right) - \left[\frac{1}{2} F_Y\left(y\right)^2\right]_0^\infty}{\left(1 - F_X\left(0\right)\right)\left(1 - F_Y\left(0\right)\right)} \tag{17.22}\\
&= \frac{\left(1 - F_Y\left(0\right)\right) - \frac{1}{2}\left(1 - F_Y\left(0\right)^2\right)}{\left(1 - F_X\left(0\right)\right)\left(1 - F_Y\left(0\right)\right)} \tag{17.23}\\
&= \frac{\frac{1}{2} - F_Y\left(0\right) + \frac{1}{2} F_Y\left(0\right)^2}{\left(1 - F_X\left(0\right)\right)\left(1 - F_Y\left(0\right)\right)} \tag{17.24}\\
&= \frac{\frac{1}{2}\left(1 - F_Y\left(0\right)\right)^2}{\left(1 - F_X\left(0\right)\right)\left(1 - F_Y\left(0\right)\right)} \tag{17.25}\\
&= \frac{1}{2}\frac{\left(1 - F_Y\left(0\right)\right)}{\left(1 - F_X\left(0\right)\right)} \tag{17.26}\\
&= \frac{1}{2} \tag{17.27}
\end{align}$$

The inequality in Equation 17.20 comes from the fact that $\mathbb{P}\left[X \geq a\right] \geq \mathbb{P}\left[Y \geq a\right]$ $\forall a \geq 0$, which implies that $\left(1 - F_X\left(a\right)\right) \geq \left(1 - F_Y\left(a\right)\right) \forall a \geq 0$.

The fraction in Equation 17.26 is equal to one since $\mathbb{P}\left[X \geq 0\right] = \mathbb{P}\left[Y \geq 0\right]$ implies that $1 - F_X\left(0\right) = 1 - F_Y\left(0\right)$.

17.2 Transformation Problems

Problem 6 *Let X have pdf $f_X(x) = \frac{2}{9}(x+1)$, $-1 \le x \le 2$. Define $Y = X^2$. Find $f_Y(y)$.*

Solution 6.1 *First note the range of Y to be $y \in [0,4]$. Also note that X^2 is not monotonic over the entire domain of Y, so we will partition the problem accordingly.*

$$\begin{aligned}
F_Y(y) &= \mathbb{P}[Y \le y] && (17.28)\\
&= \mathbb{P}\left[X^2 \le y, -1 \le X < 0\right] + \mathbb{P}\left[X^2 \le y, 0 \le X \le 2\right] && (17.29)\\
&= \mathbb{P}[X \ge -\sqrt{y}, -1 \le X < 0] + \mathbb{P}[X \le \sqrt{y}, 0 \le X \le 2] && (17.30)\\
&= \mathbb{P}[\max[-\sqrt{y}, -1] \le X < 0] + \mathbb{P}[0 \le X \le \max(\sqrt{y}, 2)] && (17.31)\\
&= \mathbb{P}[\max[-\sqrt{y}, -1] \le X < 0] + \mathbb{P}[0 \le X \le \sqrt{y}] && (17.32)\\
&= F_X(0) - F_X(\max[-\sqrt{y}, -1]) + F_X(\sqrt{y}) - F_X(0) && (17.33)\\
&= \begin{cases} F_X(\sqrt{y}) - F_X(-\sqrt{y})\, if\ -\sqrt{y} > -1 \\ F_X(\sqrt{y}) - F_X(-1)\, if\ -\sqrt{y} \le -1 \end{cases} && (17.34)\\
&= \begin{cases} F_X(\sqrt{y}) - F_X(-\sqrt{y})\, if\ 0 \le y < 1 \\ F_X(\sqrt{y})\, if\ 1 \le y \le 4 \end{cases} && (17.35)\\
&= (F_X(\sqrt{y}) - F_X(-\sqrt{y}))\{0 \le y < 1\} + F_X(\sqrt{y})\{1 \le y \le 4\} && (17.36)
\end{aligned}$$

At this point, we have an acceptable expression of $F_Y(y)$. We differentiate to get

$$f_Y(y) = \left(\frac{f_X(\sqrt{y})}{2\sqrt{y}} - \frac{f_X(-\sqrt{y})}{2(-\sqrt{y})}\right)\{0 \le y < 1\} + \frac{f_X(\sqrt{y})}{2\sqrt{y}}\{1 \le y \le 4\} \tag{17.37}$$

$$= \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}\{0 \le y < 1\} + \frac{f_X(\sqrt{y})}{2\sqrt{y}}\{1 \le y \le 4\} \tag{17.38}$$

Finally, substituting in the expression for $f_X(x)$ gives the final answer.

$$f_Y(y) = \frac{\frac{2}{9}(\sqrt{y}+1) + \frac{2}{9}(1-\sqrt{y})}{2\sqrt{y}}\{0 \le y < 1\} + \frac{\frac{2}{9}(\sqrt{y}+1)}{2\sqrt{y}}\{1 \le y \le 4\} \tag{17.39}$$

$$= \frac{2}{9\sqrt{y}}\{0 \le y < 1\} + \left(\frac{1}{9} + \frac{1}{9\sqrt{y}}\right)\{1 \le y \le 4\} \tag{17.40}$$

17.3 $\mathbb{E}[Y|X]$ Problems

Problem 7 *If X and Y are continuous random variables, show that*

$$\mathbb{E}_Y[[Y - \mathbb{E}[Y|X]]\,|X] = 0 \tag{17.41}$$

Solution 7.1 *Breaking this apart and solving gives*

$$\mathbb{E}_Y[Y|X] - \mathbb{E}_Y\left[\mathbb{E}_Y\left[Y|X\right]|X\right] = \mathbb{E}_Y[Y|X] - \mathbb{E}_Y\left[\mathbb{E}_Y\left[Y|X\right]\right] = \\ \mathbb{E}_Y[Y|X] - \mathbb{E}_Y[Y|X] = 0$$

Problem 8 *If X and U are independent standard normal random variables and $Y = X + U$, then prove or disprove the following:*

Question 8.1 $\mathbb{E}\left[Y|X = t\right]$ *is increasing in t*

Question 8.2 $\mathbb{E}\left[X|Y = t\right]$ *is increasing in t*

Question 8.3 *What if X and U are nonconstant, independent random variables symmetric about zero, but not necessarily normal?*

Solution 8.1 *First be warned that we are asked to "prove or disprove." Never forget that this opens the door for false statements!*

The first part is the simplest. $\mathbb{E}\left[Y|X = t\right] = \mathbb{E}\left[X + U|X = t\right] = \mathbb{E}\left[t + U\right] = t + \mathbb{E}\left[U\right] = t$. *Clearly t is increasing in t, so we have our proof. Once again, we did not require normality, so this proof extends to the more general question.*

The second part is the trickiest. There are two traps here. First, it is tempting to say

$$\mathbb{E}\left[X|Y = t\right] = \mathbb{E}\left[X|X + U = t\right] = \mathbb{E}\left[X|X = t - U\right] \tag{17.42}$$
$$= \mathbb{E}\left[t - U\right] = t - \mathbb{E}\left[U\right] = t \tag{17.43}$$

However, this statement is incorrect. It is not true that $\mathbb{E}\left[X|X = t - U\right] = \mathbb{E}\left[t - U\right]$.

Fortunately, we have Theorem 3.5 that says $X|Y \sim N\left(\mu_X + \rho\left(\sigma_X/\sigma_Y\right)\left(y - \mu_Y\right), \sigma_X^2\left(1 - \rho^2\right)\right)$ *if X and Y are normal. Since Y is the sum of two standard normals, $Y \sim N\left(0, 2\right)$ by Theorem 3.6. Therefore,*

$$\mathbb{E}\left[X|Y = t\right] = \mu_X + \rho\left(\sigma_X/\sigma_Y\right)\left(y - \mu_Y\right) \tag{17.44}$$
$$= 0 + \rho\left(1/2\right)\left(t - 0\right) \tag{17.45}$$
$$= t\frac{\rho}{2} \tag{17.46}$$

which is clearly increasing in t. However, since Theorem 3.5 applies only to normal random variables, we have not answered the more general case.

Normally distributed variables have nice single-peaked and continuous density functions. If a counter-example exists to disprove this statement for the *general case, it would likely be an example with discrete, multi-peaked densities. In fact, a good counter example is the variables $X \in \{-1, 0, 1\}$ and $U \in \{-10, 0, 10\}$. For this example, we can assume that $\mathbb{P}$ assigns equal weight (1/3) to each of the 3 outcomes for each variable, but we really don't need any assumptions about $\mathbb{P}$ at all. We know that $Y \in \{-11, -10, -9, -1, 0, 1, 9, 10, 11\}$, therefore t can only take those values. Furthermore, $\mathbb{E}[X|Y = t]$ is completely deterministic. If $t = -11$, then we know that $X = -1$, so $\mathbb{E}[X|Y = -11] = -1$. Continue this procedure through the 9 possibilities.*

$\mathbb{E}[X|Y = -10] = 0$

$\mathbb{E}[X|Y = -9] = 1$

$\mathbb{E}[X|Y = -1] = -1$

$\mathbb{E}[X|Y = 0] = 0$

$\mathbb{E}[X|Y = 1] = 1$

$\mathbb{E}[X|Y = 9] = -1$

$\mathbb{E}[X|Y = 10] = 0$

$\mathbb{E}[X|Y = 11] = 1$

These expectations have been listed in order of increasing t, but the value $\mathbb{E}[X|Y = t]$ is certainly not increasing. Therefore, we have our counterexample. This is a very tough question because very few counterexamples do exist. Even in this example, $\mathbb{E}[U|Y = t]$ is (weakly) increasing in t.

17.4 Gauss-Markov Problems

Problem 9 *Prove that*

$$\min_{g(X)} \mathbb{E}\left[(Y - g(X))^2 |X\right] = \mathbb{E}[\mathrm{Var}[Y|X]] \tag{17.47}$$

Solution 9.1 *This is essentially a version of the Gauss-Markov Theorem proof. We use the usual trick of adding and subtracting $\mathbb{E}[Y|X]$ to our equation to proceed. Also note that the expectation operators in the given equation are expectations over Y with X being fixed.*

$$\mathbb{E}_Y\left[(Y - g(X))^2 |X\right] = \mathbb{E}_Y\left[\left((Y\underbrace{-\mathbb{E}(Y|X)) + (\mathbb{E}(Y|X)}_{\text{“}\textit{the usual trick}\text{”}} - g(X))\right)^2 |X\right] \tag{17.48}$$

$$
\begin{aligned}
&= \mathbb{E}_Y \left[\begin{array}{c} (Y - \mathbb{E}[Y|X])^2 + (\mathbb{E}[Y|X] - g(X))^2 \\ +2(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - g(X))|X \end{array} \right] \\
&= \mathbb{E}_Y \left[(Y - \mathbb{E}[Y|X])^2 |X \right] + \mathbb{E}_Y \left[(\mathbb{E}[Y|X] - g(X))^2 |X \right] \\
&+ 2(\mathbb{E}[Y|X] - g(X)) \underbrace{\mathbb{E}_Y [(Y - \mathbb{E}[Y|X])|X]}_{=0 \text{ by Problem } 7} \\
&= \mathbb{E}_Y \left[(Y - \mathbb{E}[Y|X])^2 |X \right] + \mathbb{E}_Y \left[(\mathbb{E}[Y|X] - g(X))^2 |X \right] \\
&> \mathbb{E}_Y \left[(Y - \mathbb{E}[Y|X])^2 |X \right] \ \forall g(X) \neq \mathbb{E}[Y|X]
\end{aligned}
$$

Therefore, to minimize $\mathbb{E}_Y \left[(Y - g(X))^2 \right]$*, we choose* $g(X) = \mathbb{E}[Y|X]$*, which is our least-squares solution. Substituting this into our expression and using Definition 1.30 gives*

$$
\mathbb{E}\left[(Y - g(X))^2 |X \right] = \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2 |X \right] = \mathbb{E}[\mathrm{Var}[Y|X]] \qquad (17.49)
$$

17.5 Basic Binary Models

Problem 10 *Consider the following regression model*

$$
Y = \begin{cases} 1 & \text{if } X\beta - u > 0 \\ 0 & \text{if } X\beta - u \leq 0 \end{cases} \qquad (17.50)
$$

Assume u *s independent of* X *and* $u \sim F_u(\cdot)$*, where* $F_u(\cdot)$ *is known. Find* $\mathbb{E}[Y|X]$ *and write out the sample analogue of* $\mathbb{E}[Y - \mathbb{E}[Y|X]]^2$ *for this model. Develop an estimator of* β *based on this statistic.*

Solution 10.1 *Before beginning, note that this looks suspiciously like the rat tolerance problem posed in Section 16.4. We know from that section that* $\mathbb{P}[Y = 1] = \mathbb{P}[u < X\beta] = F_u(X\beta)$*. Given this fact, we know that* $\mathbb{E}[Y|X] = \mathbb{P}[Y = 1] = F_u(X\beta)$*, which we will use in our sample analogue.*

The sample analogue of an expectation is an average, so we'll use as our analogue for $\mathbb{E}[Y - \mathbb{E}[Y|X]]^2$

$$
\frac{1}{n} \sum_{i=1}^{n} (Y_i - F_u(X_i\beta))^2 \qquad (17.51)
$$

Using the weak law of large numbers confirms that our sample analogue converges (in probability) to the expectation. This statistic we've generated is also *the MSE, or mean-squared error. Recall that the OLS estimate minimizes the mean-squared error, so we shall use the same technique here to develop our estimate.*

Goal: $\min_\beta \frac{1}{n}\sum_{i=1}^n \left(Y - F_u\left(X\beta\right)\right)^2$*. The FOCs are*

$$\frac{1}{n}\sum_{i=1}^n 2\left(Y_i - F_u\left(X_i\beta\right)\right)\left(-f\left(X_i\beta\right)X_i\right) = 0 \tag{17.52}$$

$$\sum_{i=1}^n X_i f(X_i\beta)\left(F\left(X_i\beta\right) - Y_i\right) = 0 \tag{17.53}$$

$$\sum X_i f(X_i\beta^*)F(X_i\beta^*) = \sum X_iY_i \tag{17.54}$$

Although we can't get a nice solution to this problem, we can say that β^* that satisifies these first-order conditions will be our estimate. Furthermore, it is the estimate that minimizes residuals and therefore appears to be a BLUE estimate, though this should be checked formally.

17.6 Information Equality - OLS

Problem 11 *Verify the Information Equality for β in the linear OLS model, assuming $\varepsilon \sim N\left[0, \sigma^2 I\right]$. Use this result and the fact that the OLS estimate is BLUE to show that* $\operatorname{Var}[b] = \sigma^2\left(X'X\right)^{-1}$*. Do NOT calculate this variance directly.*

Solution 11.1 *First recall the information equality:*

$$I\left(\theta\right) \equiv \mathbb{E}_\theta\left[s(\theta)s(\theta)'\right] = -\mathbb{E}\left[\frac{\partial^2\mathcal{L}(\theta)}{\partial\theta\,\partial\theta'}\right] \tag{17.55}$$

Since $\varepsilon \sim N\left[0, \sigma^2 I\right]$, then $Y \sim N\left[X\beta, \sigma^2 I\right]$. Therefore, the likelihood of y_i given x_i is

$$f(y_i|X_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{y_i - X_i\beta}{\sigma}\right)^2\right] \tag{17.56}$$

So, the likelihood function is

$$\begin{aligned} L(Y|X) = f(Y|X) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{y_i - X_i\beta}{\sigma}\right)^2\right] \\ &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n\left(y_i - X_i\beta\right)^2\right] \\ &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\left(Y - X\beta\right)'\left(Y - X\beta\right)\right] \end{aligned} \tag{17.57}$$

The log-likelihood function is

$$\mathcal{L}(Y|X) = -\frac{n}{2}\log\left[2\pi\right] - \frac{n}{2}\log\left[\sigma^2\right] - \frac{1}{2\sigma^2}\left(Y - X\beta\right)'\left(Y - X\beta\right) \tag{17.58}$$

We want to differentiate with respect to β, so we get the following

$$\begin{aligned} s(\beta) = \frac{\partial \mathcal{L}(Y|X)}{\partial \beta} &= -\frac{1}{2\sigma^2}\frac{\partial}{\partial \beta}\left(Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta\right) = \\ &- \frac{1}{2\sigma^2}\left(-2X'Y + 2X'X\beta\right) = \frac{1}{\sigma^2}X'(Y - X\beta) = \frac{1}{\sigma^2}X'\varepsilon \end{aligned}$$

Since we have that $\mathbb{E}\left[X'\varepsilon\right] = 0$, then we get our first result that $\mathbb{E}\left[s(Y|X)\right] = 0$, which agrees with Lemma 5.5.

Taking second derivatives of the log-likelihood gives

$$\frac{\partial^2 \mathcal{L}(Y|X)}{\partial \beta\, \partial \beta'} = \frac{1}{\sigma^2}\frac{\partial}{\partial \beta}X'(Y - X\beta) = -\frac{1}{\sigma^2}X'X \tag{17.59}$$

Taking the negative expectation over $Y|X$ is equivalent to multiplying the above by -1. So,

$$-\mathbb{E}_Y\left[\frac{\partial^2 \mathcal{L}(Y|X)}{\partial \beta\, \partial \beta'}|X\right] = \frac{1}{\sigma^2}X'X \tag{17.60}$$

So, the right side of the information equality will be

$$-\mathbb{E}\left[\frac{\partial^2 \mathcal{L}(Y|X)}{\partial \beta\, \partial \beta'}|X\right] = \mathbb{E}_Y\left[s(\beta)\; s(\beta)'\right] \tag{17.61}$$

The information matrix for the linear model is therefore

$$I(\beta) = \mathbb{E}\left[\left(\frac{1}{\sigma^2}X'\varepsilon\right)\left(\frac{1}{\sigma^2}X'\varepsilon\right)'|X\right] = \frac{X'X}{\sigma^4}\mathbb{E}\left[\varepsilon\varepsilon'\right] = \frac{1}{\sigma^2}X'X \tag{17.62}$$

Combining equations 17.62 and 17.60 gives the information equality

$$I(\beta) = \frac{1}{\sigma^2}X'X = -\mathbb{E}_Y\left[\frac{\partial^2 \mathcal{L}(Y|X)}{\partial \beta\, \partial \beta'}|X\right] \tag{17.63}$$

Finally, we know that the OLS estimate is BLUE. Therefore, it achieves a Cramer-Rao Lower Bound of

$$\mathrm{Var}\left[b\right] = CRLB = I\left(\beta\right)^{-1} = \sigma^2\left(X'X\right)^{-1} \tag{17.64}$$

which give the final result that $\mathrm{Var}\left[b\right] = \sigma^2\left(X'X\right)^{-1}$.

17.7 Information Equality - Binary Logit

Problem 12 *Define the score and Hessian for the binary logit model. Demonstrate the information equality. Do any special properties of the Hessian suggest an easy estimation method other than MLE?*

Solution 12.1 *To solve this, we will use $F(X_i\beta)$ in place of the more complicated logistic distribution, and substitute the actual distribution in when needed.*

The likelihood function is

$$L(\beta) = \prod_{i=1}^{2} \left(F(X_i\beta)\right)^{Y_i} \left(1 - F(X_i\beta)\right)^{1-Y_i} \tag{17.65}$$

$$L(\beta) = \prod_{i=1}^{2} \left(F(X_i\beta)\right)^{Y_i} \left(1 - F(X_i\beta)\right)^{1-Y_i} \tag{17.66}$$

$$\mathcal{L}(\beta) = \sum_{i=1}^{2} \left(Y_i \log\left[F(X_i\beta)\right] + (1 - Y_i)\log\left[1 - F(X_i\beta)\right]\right) \tag{17.67}$$

Taking the first derivative gives the score

$$s(\beta) = \sum_{i=1}^{2} \left(Y_i \frac{f(X_i\beta)}{F(X_i\beta)} X_i + (1 - Y_i) \frac{-f(X_i\beta)}{1 - F(X_i\beta)} X_i\right) \tag{17.68}$$

At this point, we substitute the logistic distribution back in to get a formula for the hazard rate $h(X_i\beta)$ and take its derivative

$$F(X_i\beta) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)} \tag{17.69}$$

$$f(X_i\beta) = \frac{(1 + \exp(X_i\beta))X_i' \exp(X_i\beta) - \exp(X_i\beta) X_i' \exp(X_i\beta)}{(1 + \exp(X_i\beta))^2} = \frac{X_i' \exp(X_i\beta)}{(1 + \exp(X_i\beta))^2} \tag{17.70}$$

$$\frac{f(X_i\beta)}{F(X_i\beta)} = \frac{1}{1 + \exp(X_i\beta)} X_i' = h(X_i\beta) \tag{17.71}$$

$$\frac{\partial h(X_i\beta)}{\partial\beta} = \frac{-1}{(1 + \exp(X_i\beta))^2} X_i' X_i \tag{17.72}$$

Also note that

$$\frac{f(X_i\beta)}{1 - F(X_i\beta)} = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)} X_i' = F(X_i\beta) X_i' \tag{17.73}$$

Substituting into the score gives

$$s(\beta) = \sum_{i=1}^{2} \left(Y_i h(X_i\beta) X_i - (1 - Y_i) F(X_i\beta) X_i'\right) \tag{17.74}$$

Taking the derivative

$$H(\beta) = \tag{17.75}$$

Solution 12.2 *Alternate approach (unfinished!) blah*

Let $f = f(X_i\beta)$, $f' = \partial f/\partial\beta$ and $F = F(X_i\beta)$. Taking second derivatives gives

$$H(\beta) = \sum_{i=1}^{2}\left(Y_i\frac{Ff' - f^2}{F^2}X_i'X_i - (1 - Y_i)\frac{(1-F)f' + f^2}{(1-F)^2}X_i'X_i\right) \tag{17.76}$$

17.8 Binary Choice Regression

Problem 13 *Define the Binary Choice Model as*

$$Y = \begin{cases} 1 \text{ if } X\beta - \varepsilon > 0 \\ 0 \text{ otherwise} \end{cases} \tag{17.77}$$

where ε is independent of X and has some known continuous cdf $F(\varepsilon)$.

Find $\mathbb{E}[Y|X]$ for this model. How would you estimate β? How would you estimate $\mathbb{E}[Y|X]$?

Solution 13.1 *To answer this question, refer back to the Binary Logit & Probit discussion in Section 16.4.*

$\mathbb{E}[Y|X] = \mathbb{E}\left[\chi_{\{X\beta-\varepsilon>0\}}\right] = \mathbb{P}[X\beta - \varepsilon > 0] = F(X\beta)$

The "correct" method for estimating β would be maximum likelihood estimation. This is covered in Section 16.5.1 in sufficient detail. Another possible option is to minimize the distance between Y_i and $\mathbb{E}[Y_i|X]$ across all i simultaneously. To do this, we would set up the equation

$$\min_{\beta}\sum_{i=1}^{n}(Y_i - F(X_i\beta))^2 \tag{17.78}$$

and solve for the first- and second-order conditions. This is equivalent to the normal equations in the OLS procedure.

If asked this on an exam, it would be best to use the MLE method as it is the technique actually used in practice.

17.9 MLE, Frequency, and Consistency

Problem 14 *Assume that $Y_i \in \{0,1\}$. Let $\mathbb{P}[Y_i = 1] = \theta$ and $\sum_{i=1}^n Y_i = R$.*

1. *Show that R/n is the MLE estimate of θ.*
2. *Calculate the variance of $\sqrt{n}\left(\hat{\theta} - \theta\right)$*
3. *Construct a consistent estimator of the variance of $\sqrt{n}\left(\hat{\theta} - \theta\right)$*
4. *Construct a simple Wald statistic to test $\mathbf{H}_0 : \theta = 1/2$*

Solution 14.1 *To perfom MLE, we set up the likelihood function, take its log, and look at the FOC's.*

$$L(\theta) = \theta^R (1-\theta)^{n-R} \tag{17.79}$$

$$\mathcal{L}(\theta) = R\log[\theta] + (n-R)\log[1-\theta] \tag{17.80}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{R}{\theta^*} - \frac{(n-R)}{1-\theta^*} = 0 \tag{17.81}$$

$$\frac{R - n\theta^*}{\theta^*(1-\theta^*)} = 0 \tag{17.82}$$

$$\theta^* = \frac{R}{n} = \hat{\theta}_{MLE} \tag{17.83}$$

The finite-sample variance calculation is more brute-force.

$$\begin{aligned}
\operatorname{Var}\left[\sqrt{n}\left(\hat{\theta}-\theta\right)\right] &= n\operatorname{Var}\left[\hat{\theta}-\theta\right] && (17.84)\\
&= n\operatorname{Var}\left[\sum_{i=1}^{n}\frac{Y_i}{n}\right] + n\operatorname{Var}[\theta] && (17.85)\\
&= \frac{n}{n^2}\sum_{i=1}^{n}\operatorname{Var}[Y_i] + 0 && (17.86)\\
&= \frac{1}{n}n\operatorname{Var}[Y_1] && (17.87)\\
&= \mathbb{E}\left[(Y_1-\theta)^2\right] && (17.88)\\
&= \mathbb{E}\left[Y_1^2 - 2Y_1\theta + \theta^2\right] && (17.89)\\
&= \mathbb{E}\left[Y_1^2\right] - 2\theta^2 + \theta^2 && (17.90)\\
&= \mathbb{P}\{Y_1 = 1\} - \theta^2 && (17.91)\\
&= \theta - \theta^2 && (17.92)\\
&= \theta(1-\theta) && (17.93)
\end{aligned}$$

17.10 Bayesian Updating

Problem 15 *A random variable p is known to have a Beta distribution with parameters α and β, denoted $B\left[\alpha,\beta\right]$. The pdf is given by*

$$f_p\left(p\right) = \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}p^{\alpha-1}\left(1-p\right)^{\beta-1} \tag{17.94}$$

$$\mathbb{E}\left[p\right] = \frac{\alpha}{\alpha+\beta} \tag{17.95}$$

$$\mathrm{Var}\left[p\right] = \frac{\alpha\beta}{\left(\alpha+\beta\right)^2\left(\alpha+\beta+1\right)} \tag{17.96}$$

We now observe k successes in n independent binary trials (say, flips of an unfair coin) with a probability of success in each flip of p. If our prior distribution for p is $B\left(\alpha,\beta\right)$, what is our posterior after the trials are observed? What happens as the number of trials goes to infinity?

Solution 15.1 *Recall the Bayesian formula*

$$\pi(p|k) = \frac{\pi(p)f\left(k|p\right)}{f_k(k)} = \frac{f\left(p,k\right)}{f_k(k)} \tag{17.97}$$

Here, our prior is a Beta distribution and our observed data follows a binomial distribution. k is our data and p is our parameter to estimate. We will use the following solution process:

1. *Use the two known distributions to calculate the joint distribution $f(p,k)$.*
2. *Integrate the joint distribution to get the marginal distribution $f_k(k)$*
3. *Divde the joint by the marginal to get the posterior distribution.*

To get the joint distribution $f(k,p)$, we multiply the prior and the observed distributions

$$\begin{aligned} f\left(k,p\right) &= \left(\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}p^{\alpha-1}\left(1-p\right)^{\beta-1}\right)\left(\binom{n}{k}p^k\left(1-p\right)^{n-k}\right) \\ &= \binom{n}{k}\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}p^{k+\alpha-1}\left(1-p\right)^{n-k+\beta-1} \end{aligned} \tag{17.98}$$

Integrating over $[0,1]$ (the possible values of p) gives

$$\begin{aligned} f_k\left(k\right) &= \int_0^1\binom{n}{k}\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}p^{k+\alpha-1}\left(1-p\right)^{n-k+\beta-1}dp \\ &= \binom{n}{k}\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\int_0^1 p^{k+\alpha-1}\left(1-p\right)^{n-k+\beta-1}dp \\ &= \binom{n}{k}\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\left(\frac{\Gamma\left(k+\alpha\right)\Gamma\left(n-k+\beta\right)}{\Gamma\left(n+\alpha+\beta\right)}\right) \end{aligned} \tag{17.99}$$

The last step in this set of equations is quite tricky. Recall that $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ *and work from there.*

We know have all the pieces of our puzzle.

$$\begin{aligned}\pi\left(p|k\right) &= \frac{f\left(k,p\right)}{f_k\left(k\right)} && (17.100)\\ &= \frac{\binom{n}{k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{k+\alpha-1}\left(1-p\right)^{n-k+\beta-1}}{\binom{n}{k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta)}\right)}\\ &= \frac{\Gamma\left(n+\alpha+\beta\right)}{\Gamma\left(k+\alpha\right)\Gamma\left(n-k+\beta\right)}p^{k+\alpha-1}\left(1-p\right)^{n-k+\beta-1}\\ &= B\left[\left(k+\alpha\right),\left(n-k+\beta\right)\right]\end{aligned}$$

So, our posterior is of the same form as our prior - a Beta distribution. However, the parameters have updated with the addition of new information. The new mean and variance are

$$\mathbb{E}\left[p\right] = \frac{k+\alpha}{n+\alpha+\beta} \qquad (17.101)$$

$$\mathrm{Var}\left[p\right] = \frac{\left(k+\alpha\right)\left(n-k+\beta\right)}{\left(\alpha+n+\beta\right)^2\left(\alpha+n+\beta+1\right)} \qquad (17.102)$$

As n increases to infinity (and k increases with it in proportion,) the mean converges to k/n and the variance converges to 0*. The important thing to notice is that as we gain more and more data, we rely less and less on our prior beliefs and update "completely" on the observed data.*

17.11 "Research" Problems

All of these problems pretend that you are a researcher performing some statistical analysis on some data you've gathered. The questions almost always have some discussion of model misspecification and endogeneity.

Problem 16 *You have the following variables:*

- y_{it} *= 1 if running for reelection, 2 if running for higher office, 0 if retiring*
  - x_{it1} *= 1 for Democrat, 0 for Republican*
  - x_{it2} *= 1 if scandal while in office, 0 otherwise*
  - x_{it3} *= amount of "pork barrel" benefits flowing to the incumbent's district while in office*
  - x_{it4} *= difference between ideology rating of candidate and district (rated by Americans for Democratic Action.)*
  - x_{it5} *= 1 if Senate seat is open in that state, 0 otherwise*

- $u_{it} \sim N\left[0, \sigma^2\right]$, $\mathbb{E}\left[u_{it}u_{is}\right] = 0 \; \forall t, s$

Answer the following questions

1. *You run the OLS model* $y_{it} = \alpha + x_{it1}\beta_1 + \varepsilon_{it}$, *but the "true" model should include both* x_{it1} *and* x_{it2}. *If, in the true model,* $\beta_1 > 0$, $\beta_2 > 0$, *and* $\sigma_{x_{it1}, x_{its}} = 0.5$, *then what effect will this have on your model?*

2. *Show how you would test to see if* x_{it2} *through* x_{it5} *have a significant impact on the model (and should therefore be included.)*

3. *If the true model contains all 5 variables, but* $\varepsilon_{it} = u_{it} - \rho\varepsilon_{i(t-1)}$, *and you ran a model including all 5 variables but ignoring this autocorrelation, what effect will this have on your ability to estimate the coefficients? What will happen to the standard errors of your coefficients? How can you correct for this?*

4. *A colleague argues that the "pork" variable is endogenous to the decision of an incumbent to run. In other words, candidates who know they're going to run again work harder to get more "pork" into their districts, while those retiring won't.*

   (a) *How would this affect your OLS results?*

   (b) *What can you do to eliminate these problems? What are the properties of your new estimators?*

5. *Another colleague warns that your response variable is discrete, not continuous..*

   (a) *Is this a problem for your OLS results?*

   (b) *Have you violated OLS assumptions?*

   (c) *How might you estimate the parameters in your model to fix these problems?*

Solutions:
First, demean each variable so that we can ignore the constant term.

Solution 16.1 *If we run the restricted model, the omitted variable* (X_2) *goes into the error term. So,* $\varepsilon_{it} = x_{it2}\beta_2 + u_{it}$. *Therefore, our estimate will be*

$$\begin{aligned}
\hat{\beta}_1 &= \left(X_1'X_1\right)^{-1} X_1' \left(X_1\beta_1 + X_2\beta_2 + u\right) \qquad (17.103)\\
&= \left(X_1'X_1\right)^{-1} X_1'X_1\beta_1 + \left(X_1'X_1\right)^{-1} X_1'X_2\beta_2 + \left(X_1'X_1\right)^{-1} X_1'u\\
&= \beta_1 + \left(X_1'X_1\right)^{-1} X_1'X_2\beta_2 + \left(X_1'X_1\right)^{-1} X_1'u
\end{aligned}$$

In expectation, the first term is β *and the third term is* 0, *but the middle term is positive (since* X_1 *is positively correlated with* X_2 *and* $\beta_2 > 0$, *all by assumption.) Therefore, we have an upward bias in our estimate of* β_1.

Solution 16.2 *There are several tests available. The most straight-forward is to run the restricted model (using only X_1) and get a vector of residuals $\hat{u}_R$. We then regress these residuals on the omitted variables to see if the "observed error" in our restricted model is highly correlated with the omitted variables. If nR^2_{AUX} (which is distributed χ^2_{k-m}, $k = 5, m = 4$) is large, we reject the hypothesis that $0 = \beta_2 = ... = \beta_5$.*

Solution 16.3 *This is a case of AR(1). Ignoring serial correlation will still give unbiased estimates (and consistent,) though they will be inefficient (because they won't acheive a Cramer-Rao Lower Bound.) The variance of the OLS estimate will be*

$$\begin{aligned}
\mathrm{Var}[b] &= \mathbb{E}\left[(b-\mathbb{E}[b])(b-\mathbb{E}[b])'\right] \qquad (17.104)\\
&= \mathbb{E}\left[\left((X'X)^{-1}X'(X\beta+\varepsilon)-\beta\right)\left((X'X)^{-1}X'(X\beta+\varepsilon)-\beta\right)'\right]\\
&= \mathbb{E}\left[\left(\beta+(X'X)^{-1}X'\varepsilon-\beta\right)\left(\beta+(X'X)^{-1}X'\varepsilon-\beta\right)'\right]\\
&= \mathbb{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right]\\
&= (X'X)^{-1}X'\Omega X(X'X)^{-1}
\end{aligned}$$

If there were no serial correlation, then $\Omega = \sigma^2 I^{(T)}$. Note that if we used the GLS estimator $\hat{\beta}$, we would have

$$\begin{aligned}
\mathrm{Var}\left[\hat{\beta}\right] &= \mathbb{E}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)'\right] \qquad (17.105)\\
&= \mathbb{E}\left[\left(X'\Omega^{-1}X\right)^{-1}X'\varepsilon\varepsilon'X\left(X'\Omega^{-1}X\right)^{-1}\right]\\
&= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}X\left(X'\Omega^{-1}X\right)^{-1}\\
&= \left(X'\Omega^{-1}X\right)^{-1}
\end{aligned}$$

Compare the two variances by the following

$$\mathrm{Var}\left[\hat{\beta}\right] \overset{?}{<} \mathrm{Var}[b] \qquad (17.106)$$

$$\left(X'\Omega^{-1}X\right)^{-1} \overset{?}{<} (X'X)^{-1}X'\Omega X(X'X)^{-1} \qquad (17.107)$$

$$I^{(T)} \overset{?}{<} (X'X)^{-1}X'\Omega X(X'X)^{-1}\left(X'\Omega^{-1}X\right) \qquad (17.108)$$

$$(X'X)^{-1}X'\Omega X(X'X)^{-1}\left(X'\Omega^{-1}X\right) - I^{(T)} \overset{?}{>} 0 \qquad (17.109)$$

This condition should be true since the matricies are all positive definite... but it's not established by this argument. Regardless, we know that the GLS estimate will produce unbiased, consistent, and efficient estimates and should therefore be used to correct this problem. If Ω is unknown, we can use a FGLS procedure instead.

Solution 16.4 *Ignoring endogeneity between variables is a serious offense.*

1. *It creates correlation between the endogenous explanatory variable and the error term in the model, so the estimates will be biased and inconsistent.*

2. *To correct this, consider either finding an instrumental variable to replace the endogenous variable, or perform a Two-Stage Least Squares procedure. To do this, regress the endogenous variable on all exogenous variables in the model to get predicted values for the endogenous variables. Those can be inserted into the original equation. This gives cosistent estimates that (usually) are asymptotically normally distributed.*

Solution 16.5 *Since we have a discrete model, this is a serious problem for our OLS model.*

1. *A predicted value of 1.5 is meaningless.*

2. *This introduces fairly serious heteroscedasticity into the model since the Y terms take very limited numbers of values. This means that $e = Y - X\beta$ can only take on 3 possible values for a given (X, Y) observation. Even more seriously, the assumption that $\mathbb{E}[\varepsilon|X] = 0$ forces the probability of each Y value to be seriously constrained. Formally,*

$$\mathbb{E}[\varepsilon|X] = P_0 (0 - X\beta) + P_1 (1 - X\beta) + (1 - P_0 - P_1)(2 - X\beta) = 0 \tag{17.110}$$

   where $P_i = \mathbb{P}[Y = i] \, \forall i \in \{0, 1, 2\}$. Solving for P_0 gives a function in terms of P_1 and thus forces these two probabilities to be related in a specific way. If they aren't, then the error term is not of zero conditional mean, violating an OLS assumption.

3. *We need to switch to a qualitative response (QR) model instead of a linear model. A trichotomous probit would be appropriate since the errors are assumed to be normally distributed. If they were of a logistic distribution, we could use logit. This would estimate the probabilities of each action being taken.*

Problem 17 *You have a dataset with the following variables:*

- V_{idt} = *% of voters in each district who voted for the incumbent (1972-1994)*
- S_{idt} = *campaign expenditures of incumbent*
- S_{cdt} = *campaign expenditures of challenger*
- Q_{idt} = *quality of the challenger. coded 1 if challenger held previous political office, 0 otherwise.*

Assume the "true" model is $V_{idt} = \beta_1 + \beta_2 S_{idt} + \beta_3 S_{cdt} + \beta_5 Q_{idt} + \mu_{idt}$. Assume $\beta_j > 0 \, \forall j$ and all variables have positive correlation.

1. *If you exclude Q_{idt}, will the coefficients you estimate allow you to make correct inferences about the effects of the other variables? Will your estimate of β_2 be correct? If not, will it be over- or under-estimated?*

2. *Now you estimate the correct model, but find a problem. The data comes from the same geographic area in 2-year intervals over 22 years. Is the OLS estimator still BLUE?*

3. *You also note that your data comes from very different geographic areas of different sizes and demographic characteristics. Is this a problem? Will OLS estimates still be BLUE?*

4. *You run the full model. A colleague tells you that the impact of incumbent expenditures on vote shares has a diminishing marginal return. How would you re-estimate the model to take this into account? Is the OLS estimator still appropriate?*

5. *A colleague says "studies show that there is endogeneity between incumbent expenditures and vote shares." What do they mean? Does this violate an OLS assumption in your model? How can you deal with this? Show how your solution will yield consistent estimates.*

Solution 17.1 *Excluding the variable will cause overestimation of the remaining coefficients*

Solution 17.2 *This indicates serial correlation. The OLS estimates are no longer efficient, so they're not BLUE.*

Solution 17.3 *This is heteroscedasticity. The OLS estimates are not BLUE because they're not efficient.*

Solution 17.4 *This is a non-linearity. You can transform the variable S_{idt} to something like $\log[S_{idt}]$. The OLS estimator will still be appropriate, but the meaning of the coefficient will not be as transparent.*

Solution 17.5 *Endogeneity causes unbiasedness. It violates the assumption that the explanatory viables are uncorrelated with the error term, which is also the cause of the unbiasedness. Use a two-stage least squares procedure to deal with this. We know that 2SLS is consistent.*

Part IV

Estimation Asymptotics

Chapter 18

Background

This chapter provides a toolbox of theorems and definitions that are frequently used in asymptotic analysis. Many of these topics also appear in the previous sections of this book, but are repeated here for easier reference.

18.1 Probability, Measures, and Integration

We know that $\mathbb{P}$ is a measure that assigns real numbers to elements in Ω. In effect, $\mathbb{P}$ measures the "size" of those elements. The measure $\mathbb{P}$ defined over a set of elements determines the collecive "size" of the whole set. Equivalently, a function $f(x) \in \mathbb{R}$ assigns a "size" to each value x and $\int_{I \subset \mathbb{R}} f(x)\,dx$ measures the collective "size" of the points in some interval I. Therefore, the measure $\mathbb{P}$ and the operator $\int$ perform the same function. For probabily measures defined over the real line, it is easy to see that they are identical.

Throughout this part of the book, we often interchange $\int$ and $\mathbb{P}$ between theorems and definitions. For example, a theorem that applies to the $\int$ operator will apply to the $\mathbb{P}$ operator.

18.2 Convergence

Recall the following definitions.

Definition 18.1 *A random variable X_n **converges in probability** to a constant c if $\lim\limits_{n\to\infty} \mathbb{P}[|X_n - c| > \varepsilon] = 0\ \forall \varepsilon > 0$. We denote this by $\operatorname{plim} X_n = c$ and c is called the "probability limit" of X_n. An alternative notation is $X_n \xrightarrow{p} c$. Another alternative notation is $X_n = c + o_p(1)$, which will be explained later.*

Definition 18.2 *A random variable X_n **converges almost surely** to a constant c if $\mathbb{P}\left[\lim\limits_{n\to\infty} |X_n - c| > \varepsilon\right] = 0\ \forall \varepsilon > 0$. We denote this by $X_n \xrightarrow{a.s.} c$.*

A good way to think of almost sure convergence is that there exists some "good set" $\Omega^* \subset \Omega$ such that X_n converges to c for all $\omega \in \Omega^*$, and this "good set" has probability 1. Therefore, the states of the world in which X_n does not converge to c have zero measure (or, are zero-probability events.)

Convergence almost surely is a stronger condition than convergence in probability, but it has the important property that increased sampling is guaranteed to maintain almost sure convergence, while convergence in probability may or may not be maintained as sampling is increased. To see this, define

$$\Omega_{n,\varepsilon} = \left\{\omega \in \Omega : \sup_{m \geq n} |X_m(\omega) - c| \leq \varepsilon\right\} = \left\{\omega \in \Omega : X_n \xrightarrow{a.s.} c\right\} \tag{18.1}$$

$$\Gamma_{n,\varepsilon} = \{\omega \in \Omega : |X_n(\omega) - c| \leq \varepsilon\} = \left\{\omega \in \Omega : X_n \xrightarrow{p} c\right\} \tag{18.2}$$

Note that $\Omega_{n,\varepsilon} \subseteq \Omega_{n+k,\varepsilon}$ $\forall k > 0$, but $\Gamma_{n,\varepsilon} \nsubseteq \Gamma_{n+k,\varepsilon}$ necessarily. The conclusion is that if a state of the world ω is drawn and, for some n, $|X_n - c| \leq \varepsilon$, then we know that as n increases, ω will still be an element of $\Omega_{n,\varepsilon}$, but ω may not necessarily be an element of $\Gamma_{n,\varepsilon}$. For example, if after 100 observations we find that X_n is within ε of c and we have almost-sure convergence, then X_m is guaranteed to be within ε of c for all $m > n$, but this may not be true if we only have convergence in probability.

Definition 18.3 *A random variable X_n* ***converges in distribution*** *to a random variable X (or, $X_n \rightsquigarrow X$) if $\mathbb{P}[f(X_n)] \longrightarrow \mathbb{P}[f(X)]$ for all bounded, continuous functions $f : \mathbb{R} \to \mathbb{R}$.*

Note that this definition is equivalent to our previous definition that required the cdf of X_n to converge to the cdf of X pointwise since the pdf functions are always bounded and continuous by assumption. This new definition will be more applicable to asymptotic analysis.

Definition 18.4 *Let $\{X_n\}$ be a sequence of random variables. The sequence is* ***bounded in probability*** *if, $\forall \varepsilon > 0,\ \exists M_\varepsilon > 0, N_\varepsilon \ni \forall n > N_\varepsilon$*

$$\mathbb{P}[|X_n| > M_\varepsilon] < \varepsilon \tag{18.3}$$

We say that $X_n = O_p(1)$ ("X_n is big oh p 1", as opposed to "little oh p 1" which we will see later.)

Theorem 18.1 ***(Continuous Mapping Theorem)*** *Let $X_n \rightsquigarrow X$, $T : \mathbb{R} \to \mathbb{R}$, and $\mathbb{P}_X\{x \in \mathbb{R} : T \text{ is continuous at } x\} = 1$. Then $T(X_n) \rightsquigarrow T(X)$.*

The proof of this theorem is a direct consequence of the definition of convergence in distribution.

Theorem 18.2 ***(Slutsky's Theorem)*** *If $X_n \rightsquigarrow X$ and $\operatorname{plim} Y_n = 0$ then $X_n + Y_n \rightsquigarrow X$.*

18.2.1 Limits and Integration

This subsection deals with conditions under which the limit of a sequence of integrated elements equals the integral of the limit of the sequence.

The following theorem gives conditions under which the limit of the integral of a sequence of functions is equal to the integral of the limit. In fact, the name of the theorem identifies those conditions (instead of identifying the result of the theorem.)

Theorem 18.3 ***(Monotone Convergence Theorem)*** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space (μ need not be a probability measure.) If $\{f_n\}$ is a sequence of functions such that $f_n : \Omega \to \mathbb{R}$ and*

1. $0 \leq f_1(\omega) \leq f_2(\omega) \leq \ldots \ \forall \omega \in \Omega$ *($\{f_n\}$ is monotone)*
2. $f_n(\omega) \longrightarrow f(\omega) \ \ \forall \omega \in \Omega$ *($\{f_n\}$ is convergent)*

then f is measruable and

$$\lim_{n\to\infty} \int_\Omega f_n(\omega)\, d\mu(\omega) = \int_\Omega f(\omega)\, d\mu(\omega) \tag{18.4}$$

For a sequence of sets (or events) $\{A_n\}$, we construct the following limit concept.

Definition 18.5 *For any sequence of sets $\{A_j\}_{j=1}^{\infty}$, define $\{A_n \ i.o.\} = \cap_{k\geq 1} \cup_{j\geq k} A_j$. This is the set "$A_n$ infinitely often."*

The set A_n infinitely often is the set of points that remains in the sets of the sequence as n goes to infinity.

Now consider an increasing or decreasing sequence $\{A_n\}$. We first prove a very convenient lemma, followed by a theorem about the probability measure defined on a monotone sequence of sets (sequences where each set includes or is included in all subsequent sets in the sequence.)

Lemma 18.4 *For any sequence of sets $\{A_j\}_{j=1}^{\infty}$, there exists a disjoint sequence $\{B_j\}_{j=1}^{\infty}$ of sets such that $B_j \subset A_j \ \forall j$, $\cup_{j=1}^{n} A_j = \cup_{j=1}^{n} B_j \ \forall n \geq 1$, and $\cup B_j = \cup A_j$.*

Proof. *See Helms p. 43* ■

Theorem 18.5 *Let $\{A_j\}_{j=1}^{\infty}$ is a sequence of events.*

1. *If $A_1 \subset A_2 \subset \ldots$ and $A = \cup_{j=1}^{\infty} A_j$, then $\mathbb{P}[A] = \lim_{n\to\infty} \mathbb{P}[A_n]$*
2. *If $A_1 \supset A_2 \supset \ldots$ and $A = \cap_{j=1}^{\infty} A_j$, then $\mathbb{P}[A] = \lim_{n\to\infty} \mathbb{P}[A_n]$*

Proof. *This is an application of the Monotone Convergence Theorem given above. Let* $f_n(\omega) = A_n$ *and* $f(\omega) = A$. *Let the measure in the MCT be* $\mathbb{P}$ *and the result is immediate.* ■

Theorem 18.6 ***(Borel-Cantelli Lemma)***[1] *If* $A_1 A_2, \ldots \in \mathcal{A}$ *and* $\sum_{i=1}^{\infty} \mathbb{P}[A_i] < \infty$ *then* $\mathbb{P}\{A_n \text{ i.o.}\} = 0$. *In other words, the limit set of* $\{A_n\}$ *is of measure zero w.r.t* $\mathbb{P}$.

Proof. *Since the sequence* $\{\cup_{j>k} A_j\}_{k=1}^{\infty}$ *is a decreasing sequence by construction and* $\{A_n \text{ i.o.}\} = \cap_{k=1}^{\infty} \cup_{j \geq k} A_j$,

$$\mathbb{P}[\{A_n \text{ i.o.}\}] = \lim_{k\to\infty} \mathbb{P}[\cup_{j\geq k} A_j] \tag{18.5}$$

by Theorem 18.5.

Using Boole's Inequality (Theorem 1.19), we have that

$$\mathbb{P}[\cup_{j\geq k} A_j] \leq \sum_{j=k}^{\infty} P[A_j] \tag{18.6}$$

Therefore, $\forall k \geq 1$

$$0 \leq \mathbb{P}[\{A_n \text{ i.o.}\}] \leq \sum_{j=k}^{\infty} \mathbb{P}[A_j] \tag{18.7}$$

Since $\sum_{i=1}^{\infty} \mathbb{P}[A_i]$ *converges, then the terms* $\mathbb{P}[A_n]$ *approach zero as* n *becomes very large. Consequently, the right side of the above equation tends to zero as* $k \to \infty$. *The inequalities of the equation are independent of* k, *and thus* $\mathbb{P}[\{A_n \text{ i.o.}\}] = 0$. ■

Theorem 18.7 ***(Dominated Convergence Theorem)*** *Let* $\{f_n\}$ *be a sequence of measurable functions on* Ω *where* $f(\omega) = \lim_{n\to\infty} f_n(\omega)$ *exists for each* $\omega \in \Omega$. *If there exists a measurable function* g *such that*

1. $\int_{\Omega} g(\omega)\, d\mu(\omega) < \infty$
2. $\sup_n |f_n(\omega)| \leq g(\omega) \;\; \forall \omega \in \Omega$

then

1. $\int_{\Omega} |f(\omega)|\, d\mu(\omega) < \infty$
2. $\lim_{n\to\infty} \int_{\Omega} f_n(\omega)\, d\mu(\omega) = \int_{\Omega} f(\omega)\, d\mu(\omega)$

[1] The Borel-Cantelli Lemma must have been a lemma in the original publication, but is now quite famous and most often appears as its own theorem despite the name.

18.3 Law of Large Numbers

There do exist several laws of large numbers. However we only recall the most important here.

Theorem 18.8 ***(Khinchine's Weak Law of Large Numbers)*** *If $X_1, ..., X_n$ are independently drawn and identically distributed (iid) with mean μ, then $\bar{X} \xrightarrow{p} \mu$.*

We can illustrate the weak law of large numbers with a computer simulation. For each n from 1 to 10,000, we generate a sample of n bernoulli observations X_i where $\mathbb{P}[X_i = 1] = p$ and $\mathbb{P}[X_i = 0] = (1 - p)$. For each n, we calculate $\bar{X}_n = (1/n) \sum X_i$. The resulting graph for $p = 0.5$ is

REMOVED

This picture clearly shows that averages converge to population means, but it illustrates the fact that convergence may be a slow process. Even with several thousand observations there may exist significant variation between observed sample averages. Another important thing to notice is that the rate of convergence slows as $n \to \infty$. This implies that if increased sampling has a constant positive marginal cost, then there exists some n^* where the cost of an additional observation outweighs the expected gains in reduced estimator variance.

We now focus on a more general law of large numbers especially applicable to estimation procedures such as maximum likelihood.

blah

18.4 Central Limit Theorems

Much like the Laws of Large Numbers, there exist various Central Limit Theorems that depend on the assumptions on the sample $X_1, ..., X_n$. However, since the theorems concern the convergence of a function of random variables to the normal distribution, we use only the convergence in distribution concept.

Theorem 18.9 ***(Univariate Linberg-Levy Central Limit Theorem)*** *If $X_1, ..., X_n$ are independently drawn and identically distributed (iid) with mean $\mu < \infty$ and variance $\sigma^2 < \infty$, then*

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N\left[0, \sigma^2\right] \tag{18.8}$$

Note that we could scale the left-hand side by σ to get a slightly more useful form

$$\sqrt{n}\left(\frac{\bar{X}_n-\mu}{\sigma}\right)\xrightarrow{d}N\left[0,1\right]\tag{18.9}$$

One particular generalization of the central limit theorems is the **delta method**.

Theorem 18.10 *If Y_n satisfies a central limit theorem for some θ (so, $\sqrt{n}\left(Y_n-\theta\right)\xrightarrow{d} N\left[0,\sigma^2\right]$), then for a function g such that $g'(\theta)$ exists and is nonzero, then*

$$\sqrt{n}\left(g\left(Y_n\right)-g\left(\theta\right)\right)\xrightarrow{d}N\left[0,\sigma^2\left(g'\left(\theta\right)\right)^2\right]\tag{18.10}$$

The proof of the delta method requires the Taylor series expansion of $g(Y_n)$ around $Y_n=\theta$.

There also exist central limit theorems for situations where the variables are drawn from distributions of different means and variances. These are less common.

Finally, there exists a related Theorem for the Maximum Likelihood Estimator, which is a direct consequence of property 2.

Theorem 18.11 *If $X_1,X_2,...,X_n$ are iid $f\left(x|\theta\right)$ and $\hat{\theta}$ is the MLE estimate of θ, then (under some regularity conditions on $f\left(x|\theta\right)$)*

$$\sqrt{n}\left(\tau\left(\hat{\theta}\right)-\tau\left(\theta\right)\right)\xrightarrow{d}N\left[0,I\left[\theta\right]^{-1}\right]\tag{18.11}$$

where $I\left[\theta\right]^{-1}$ is the Cramer-Rao Lower Bound for the estimate θ.

Theorem 18.12 ***(Glivenko-Cantelli)*** *Let $X_1,...,X_n$ be iid with measure P, which has cdf F. Then*

Proof. *For each $t\in(0,1)$, let $Q\left(t\right)=\inf\left\{x\in\mathbb{R}:F\left(x\right)\geq t\right\}$ be the "quantile transformation" function.*

Note that if, for some t, F has a "flat spot" from x_0 to x_1, then $Q\left(t\right)=x_0$. If, on the other hand, F "jumps over t'" at x ($\lim_{x_n\uparrow x}F\left(x_n\right)=t_0$ and $F\left(x\right)=t_1$, where $t_1>t'>t_0$,) then $Q\left(t'\right)=x$ since cdfs are right-continuous. In short, the interval $\left[Q\left(t\right),\infty\right)=\left\{x\in\mathbb{R}:F\left(x\right)\geq t\right\}$. So, for all $t\in(0,1)$, $F\left(x\right)\geq t\iff x\geq Q\left(t\right)$.

Take $Y\sim U\left[0,1\right]$. For each $x\in\mathbb{R}$, we have that $\mathbb{P}\left\{Q\left(Y\right)\leq x\right\}=\mathbb{P}\left\{Y\leq F\left(x\right)\right\}=F\left(x\right)$. Thus $Q\left(Y\right)$ has cdf F.

Draw $Y_1,...,Y_n$ iid $U\left[0,1\right]$ and let $z_i=Q\left(Y_i\right)\forall i=1,2,...,n$.

Thus,

$$\sup\nolimits_{x\in\mathbb{R}}\left|\mathbb{P}\right|\tag{18.12}$$

■

Bibliography

[1] Aliprantis, C. and K. Border. *Infinite Dimensional Analysis, A Hitchhiker's Guide* 2nd ed.

[2] Gourieroux. *Econometrics of Qualitative Dependent Variables.* English translation.

[3] Greene. *Econometric Analysis* 4th ed.

[4] Montgomery, D. *Design and Analysis of Experiments* 4th ed.

[5] Neter, Kutner, Nachtsheim, and Wasserman. *Applied Linear Statistical Models* 4th ed.

[6] Ramanathan, R. *Introductory Econometrics with Applications* 4th ed.

[7] Sherman, R. SS 200, SS 222, SS223 Class Notes. California Institute of Technology.