Lecture 24: The Standard Linear Model: Estimation

Relevant textbook passages:

Larsen–Marx [2]: Sections 11.1, 11.2, 11.3.

Larsen and Marx [2, Chapter 11] treat the standard linear model (or the simple linear model) for the special cease where there are two variates X and Y (and a constant term). This is unfortunate since it obscures the simplicity and symmetry of the general model with variates $X_1, \dots, X_K$ and Y. Since you have had some linear algebra in Ma 1b, I will use the matrix approach to the more general problem, which is sometimes referred to as **multiple regression**, since there is more than one X_k variate.

24.1 The standard model

Larsen–Marx [2]: Section 11.3

The **standard linear model** is so called, not because standard statistical problems satisfy the assumptions, but because the standard assumptions make the model nice to deal with. The basic premise of the model is that the expected value of the **left-hand side variate** Y conditional on the value of a K-vector $\boldsymbol{X}$ of **right-hand side** variates $X_1, \dots, X_K$ is a linear function of the X_k's.

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_K\beta_K + \varepsilon.$$

The right-hand side variates may be random variables, or they may be chosen by the experimenter. For instance, if we are interested in the effects of irrigation (X) on the sugar content of grapes (Y), there are two ways to collect data on the amount of water the grapes receive. One is to set out a rain gauge and see how much it rains, the other is to control the amount of water by using an irrigation system. From the point of view of the experimenter, the rainfall is a random variable, while the amount of irrigation is not random, it is chosen.

The $\boldsymbol{E}(Y \mid \boldsymbol{X})$ as a function of $\boldsymbol{X}$ is called a **regression function** and the component variates $X_1, \dots, X_K$ are called **regressors**.

The values of the regressors $X_1, \dots, X_K$ are viewed as determining the value of Y up to some constant plus a random error. One frequent interpretation of the random error is that it is the sum of many omitted variables that we cannot/will not observe or measure.

Examples:

- The first time I ever heard of regression analysis was a few decades back. When I was a teen-ager in the sticks of Riverside County, I regularly read a gearhead magazine called *Road & Track*. (Today I only read it at my dentist's office.) They tested unaffordable exotic cars and dutifully reported on such things as quarter-mile times, 0–60 times, braking distances, lateral g-forces, etc. At some point, they got new testing gear for measuring performance that included a bicycle wheel strapped to the rear bumper connected to state-of-the-art *transistorized* electronic sensors. This new gear either weighed more or less than their old gear, so that any new tests they ran would not be comparable to their old test results. To maintain their journalistic integrity, they took all their old results and performed a regression analysis of quarter-mile times on pounds-per-horsepower. I was impressed by the plot of their data and the regression line. They

then used the results to "correct" the old results so they would be comparable to the new ones, or maybe it was vice-versa. Some day I may be able to go back and find this out because in 2012 the *Road & Track* archives were donated to the Stanford University Library.

- Hedonic pricing. Cf. zillow.com

$$\text{price} = \text{const} + \beta_1 \text{ sq. ft.} + \beta_2 \text{ no. rooms} + \cdots + \varepsilon$$

- Kepler's 3rd Law.

The square of the orbital period of a planet is directly proportional to the cube of the semi-major axis of its orbit.

$$P^2 = cA^3.$$

or

$$2 \ln P = \ln c + 3 \ln A$$

- Hubble's Law.

$$\text{red shift} = c \cdot \text{distance}$$

- Newtons's Law of Gravity:

$$F = G\frac{M_1 M_2}{d^2}$$

$$\ln F = \ln G + \ln M_1 + \ln M_2 - 2 \ln d$$

That is, the random variate Y satisfies

$$Y = X_1\beta_1 + \cdots + X_K\beta_K + \varepsilon \tag{1}$$

where ε is the **error term**.

The assumption of linearity is less restrictive than it may seem.

For instance, a polynomial in x is a linear function of x, x^2, x^3, etc.

Economists are for a variety of reasons (see, e.g., [1]) fond of weighted geometric means:

$$y = cx_1^{b_1} \cdots x_K^{b_K},$$

which upon taking logarithms can be written as a linear relationship

$$\log y = \log c + b_1 \log x_1 + \cdots + b_K \log x_K.$$

This is perhaps why Newcomb's library's table of logarithms were so noticeably worn.

We can allow for discrete categorical variates with **dummy variables**, which are indicators that assume the value one if the observation fits the category and zero otherwise. This allows the different categories to have different intercepts. [a]

[a] If you use dummy variables for each category, you cannot have a constant term—as this will make the regressors linearly dependent.

The variates X_k may be fixed constants chosen by an experimenter or they may be random variables themselves. It is very common to treat the constant term β_0 as a coefficient on the variate X_0, which always takes on value 1. Larsen and Marx [2] include a constant term without mentioning it as a variate. The constant term β_0 reduces to the expected value of Y conditional on all the X_k's being zero.

It is quite often the case that the investigator has a special interest in the effect of X_1 on Y, but it is known that the other $X_2, \dots, X_K$ have an effect on Y, and the are included in order to "control for the effects of $X_2, \dots, X_K$."

The standard linear model takes as its data a set of N **observations** of the values $x_1, \dots, x_K$ and y.

$$y_t = \beta_0 + x_{t,1}\beta_1 + \cdots + x_{t,K}\beta_K + \varepsilon_t \qquad (t = 1, \dots, N)$$

where the ε_t's are unobserved errors. The relationship among these data are usually summarized in a matrix equation[1]

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \text{ is a } N \times 1 \text{ column vector}$$

$$\boldsymbol{X} = \begin{bmatrix} x_{1,0} & \cdots & x_{1,K} \\ \vdots & \ddots & \vdots \\ x_{N,0} & \cdots & x_{N,K} \end{bmatrix} \text{ is a } N \times (K+1) \text{ matrix,}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_K \end{bmatrix} \text{ is a } (K+1) \times 1 \text{ column vector,}$$

and

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix} \text{ is a } N \times 1 \text{ column vector.}$$

The statistical assumptions are that the error vector $\boldsymbol{\varepsilon}$ satisfies

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) &= 0, \\ \boldsymbol{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}) &= \sigma^2 \boldsymbol{I}_{N\times N}. \end{aligned} \tag{3}$$

This last assumption is known as **homoskedasticity**.[2] It is possible to deal with more general error structures, but that takes us into the realm of the **generalized linear model**. When the observations correspond to different time-periods (days, weeks, months) it is unlikely that the ε_t and ε_{t+1} are uncorrelated. Special techniques have been developed to deal with **time series** that exhibit **serial correlation**.

24.2⋆ Least Squares Estimation

Regression analysis, or simply **regression**, is concerned with estimating the components of $\boldsymbol{\beta}$ and testing hypotheses regarding them.

[1] The matrix $\boldsymbol{X}$ has $(K+1)$ columns, columns 0 is a vector of ones for the constant term, plus there are K columns of values for the regressors of interest.

[2] This is sometimes written as *homoscedaticity*, but there is a convincing case to be made for the use of *k*, not *c*, see, e.g., [3].

The method of **least squares (LS)**, also known as **ordinary least squares**, estimates $\beta_0, \dots, \beta_K$ by minimizing the sum over t of squared deviations or **residuals** of y_t from a linear combination of the $x_{t,k}$'s. Computing such an estimate is usually called **regressing $\boldsymbol{Y}$ on $\boldsymbol{X_1, \dots, X_K}$**, or **running a regression of $\boldsymbol{Y}$ on $\boldsymbol{X_1, \dots, X_K}$**.

For instance, when there is one regressor of interest and a constant term, minimizing the sum of squared residuals fits a straight line through the set of points (x_t, y_t), $t = 1, \dots, N$, so as to minimize the sum of the squares of the vertical distances from the line. See Figure 24.1.

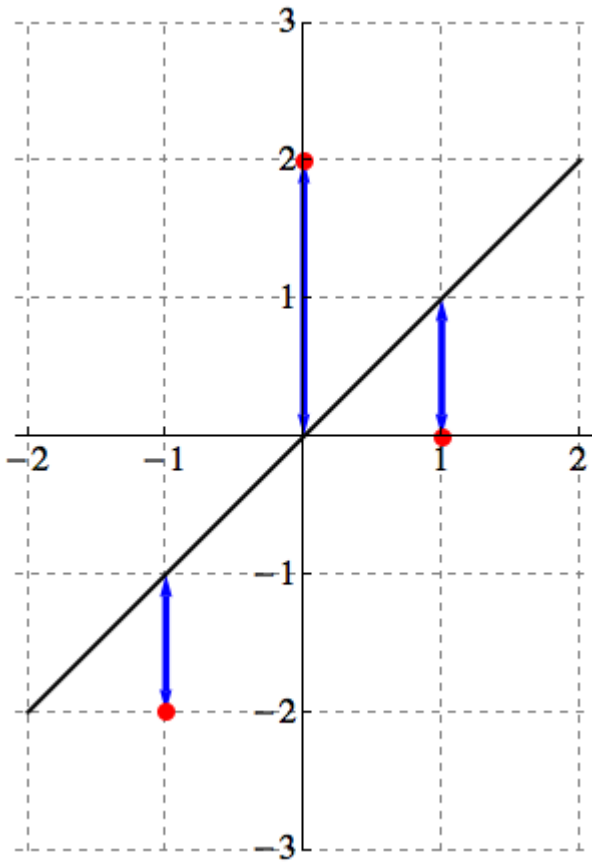

Figure 24.1. The line $y = x$ minimizes the sum of the squares of the vertical distances from the three points $(-1, -2)$, $(0, 2)$, $(1, 0)$.

Given a column $(K+1)$-vector $\boldsymbol{b}$,

$$\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}$$

is the vector of **residuals**, or differences of y_t from $\sum_{k=0}^{K} x_{t,k} b_k$. The **sum of squared residuals (SSR)** is thus $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}$ dotted with itself:

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}). \tag{4}$$

Expanding (4) yields

$$\operatorname{SSR}(\boldsymbol{b}) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{b} + \boldsymbol{b}'\boldsymbol{x}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b},$$

which is a convex quadratic function in the components of $\boldsymbol{b}$.[3] The gradient of this function is

$$\nabla \operatorname{SSR}(\boldsymbol{b}) = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}.$$

[3] To see that it is convex, note that its Hessian is $2\boldsymbol{X}'\boldsymbol{X}$, which is positive semidefinite as $\boldsymbol{x}'(2\boldsymbol{X}'\boldsymbol{X})\boldsymbol{x} = 2(\boldsymbol{X}\boldsymbol{x}) \cdot (\boldsymbol{X}\boldsymbol{x}) \geqslant 0$.

By convexity, the minimum occurs whenever the gradient equals zero. Thus the minimizer $\hat{\boldsymbol{\beta}}_{\text{LS}}$ satisfies the first-order condition $\nabla \text{SSR}(\hat{\boldsymbol{\beta}}_{\text{LS}}) = 0$, or

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}}. \tag{5}$$

This matrix equation is known as the **normal equation** for $\hat{\boldsymbol{\beta}}_{\text{LS}}$. The reason for the terminology will become clear in a bit.

On the hypothesis that $\boldsymbol{X}'\boldsymbol{X}$ (a $(K+1)\times(K+1)$ matrix) is nonsingular, we then have that

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{6}$$

minimizes the sum of squared residuals.

This $\hat{\boldsymbol{\beta}}_{\text{LS}}$ is called the **least squares (LS) estimator** of β.

24.2.1 Remark What is the matrix $\boldsymbol{X}'\boldsymbol{X}$? The i,j-entry is the i^{th} row of $\boldsymbol{X}'$ dotted with the j^{th} column of $\boldsymbol{X}$, which is just the dot product of the i^{th} and j^{th} columns of $\boldsymbol{X}$. That is,

$$(\boldsymbol{X}'\boldsymbol{X})_{ij} = \sum_{t=1}^{N} x_{ti}x_{tj}.$$

24.2.2 Remark What if $\boldsymbol{X}'\boldsymbol{X}$ is singular? This happens only if the rank of $\boldsymbol{X}$ is less than $(K+1)$, which means that there is a nonzero linear combination of the columns that sums to zero. Let $\boldsymbol{X}^k$ denote the k^{th} column of $\boldsymbol{X}$, and let

$$a_0\boldsymbol{X}^0 + a_1\boldsymbol{X}^1 + \cdots + a_K\boldsymbol{X}^K = 0,$$

where not all a_k are zero. Then if

$$\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{X}^1 + \cdots + \beta_K\boldsymbol{X}^K + \boldsymbol{\varepsilon},$$

we also have

$$\begin{aligned}\boldsymbol{y} &= \beta_0\mathbf{1} + \beta_1\boldsymbol{X}^1 + \cdots + \beta_K\boldsymbol{X}^K + \varepsilon + c(\underbrace{a_0\boldsymbol{X}^0 + a_1\boldsymbol{X}^1 + \cdots + a_K\boldsymbol{X}^K}_{=0})\\ &= (\beta_0 + ca_0)\mathbf{1} + (\beta_1 + ca_1)\boldsymbol{X}^1 + \cdots + (\beta_K + ca_k)\boldsymbol{X}^K + \varepsilon\end{aligned}$$

for any value of c. Whenever a_k is nonzero (and there is at least one), the coefficient on $\boldsymbol{X}^k$ is not unique—in fact, by choosing the proper c, it can be whatever we want. That is, the data cannot tell us what the coefficient β_k is, *even if every error term is zero.*

Note that $\hat{\boldsymbol{\beta}}_{LS}$ *is a random vector.* This is because by (2) and (6) we have

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}, \tag{7}$$

where $\boldsymbol{\varepsilon}$ is a random vector.

An important property of the LS estimator is that the vector $\boldsymbol{e}$ of **LS residuals**,

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}},$$

is orthogonal to each k^{th} column vector of the values of the **regressor** X_k. In matrix terms, this can be written

$$\boldsymbol{X}'\boldsymbol{e} = \mathbf{0}. \tag{8}$$

To see this, observe that

$$\begin{aligned}\boldsymbol{X}'\boldsymbol{e} &= \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}})\\ &= \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}}\\ &= \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\\ &= \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{0}.\end{aligned}$$

24.2.3 Remark If the regressors include a constant term, then the fitted "plane" passes through the sample means. That is,

$$\bar{y} = \hat{\beta}_0 + \bar{x}_1\hat{\beta}_1 + \cdots + \bar{x}_K\hat{\beta}_K.$$

To see why, observe that

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}} + \boldsymbol{e},$$

so

$$\boldsymbol{1}'\boldsymbol{y} = \boldsymbol{1}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}} + \boldsymbol{1}'\boldsymbol{e}, \tag{9}$$

where $\boldsymbol{1}$ is a N-vector of ones. Since $\boldsymbol{1}$ is also a column of $\boldsymbol{X}$ (column 0, corresponding to the constant term), (8) implies $\boldsymbol{1}'\boldsymbol{e} = 0$. Dividing (9) by N gives $\bar{y} = \hat{\beta}_0 + \bar{x}_1\hat{\beta}_1 + \cdots + \bar{x}_K\hat{\beta}_K$.

24.3 ⋆ The geometry of LS estimation

There is a simple geometric interpretation of the LS estimator. In general, the vector $\boldsymbol{y}$ of the T observations on Y is not an exact linear combination of the columns $\boldsymbol{X}^k$ of the observations on the X_k's. The vectors all belong to the T-dimensional space Euclidean space, but the columns of $\boldsymbol{X}$ span an at most $(K+1)$-dimensional subspace M. What LS estimation does is project $\boldsymbol{y}$ orthogonally onto the subspace M. That is, it decomposes $\boldsymbol{y}$ into two parts

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}} + \boldsymbol{e},$$

where $\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}}$ belongs to M (it is a linear combination of the columns of $\boldsymbol{X}$) and $\boldsymbol{e}$ is orthogonal to all the vectors in M. This latter is what (8) says. If we rewrite (8), we obtain

$$\boldsymbol{0} = \boldsymbol{X}'\boldsymbol{e} = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}}) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}},$$

which is the **normal equation** (5). The normal equations just state the LS residuals are orthogonal the column space of X. The term "normal" is sometimes a synonym for orthogonal.[4]

If $\boldsymbol{X}$ has rank K, then the columns of $\boldsymbol{X}$ constitute a basis for M so $\boldsymbol{y}$ can be written as unique linear combination of the columns. But even if $\boldsymbol{X}$ is not of full rank, any solution to the normal equations will minimize the sum of squared residuals.

24.4 The textbook case: $K = 1 + 1$

Larsen and Marx [2, Chapter 11] treat the standard linear model (or the simple linear model) for the special case where there are two variates X and Y (and a constant term). This is unfortunate since it obscures the simplicity and symmetry of the general model. Here we show that the analysis I just did agrees with theirs.

In this case, let's call the coefficient on the first regressor, the constant 1, β_0, and the second regressor simply X with coefficient β_1, and let us rewrite the model as they did:

Simplify this with $\sum_t x_t = N\bar{x}$, etc.

[4] This is a plausible etymology, but I haven't tracked down whether it is correct.

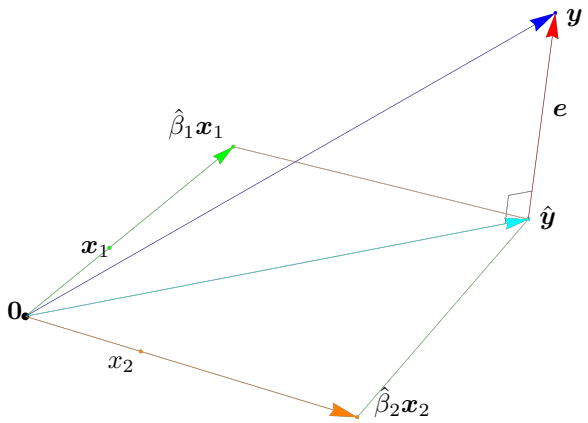

Figure 24.2. Geometry of LS.

$$y_t = \beta_0 + \beta_1 x_t.$$

This translates to

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, \qquad \boldsymbol{X}' = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix},$$

so

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} N & \sum_t x_t \\ \sum_t x_t & \sum_t x_t^2 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \sum_t y_t \\ \sum_t y_t x_t \end{bmatrix}.$$

The determinant of $\boldsymbol{X}'\boldsymbol{X}$ is simply $N\sum_t x_t^2 - \left(\sum_t x_t\right)^2$, which by the Pythagorean Theorem for Data, Proposition S2.4.4, can be rewritten as

$$\det \boldsymbol{X}'\boldsymbol{X} = N\sum_t (x_t - \bar{x})^2.$$

Thus the 2×2 inverse $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is simply

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{N\sum_t (x_t - \bar{x})^2} \begin{bmatrix} \sum_t x_t^2 & -\sum_t x_t \\ -\sum_t x_t & N \end{bmatrix}$$

so

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \frac{1}{N\sum_t (x_t - \bar{x})^2} \begin{bmatrix} (\sum_t x_t^2)(\sum_t y_t) - (\sum_t x_t)(\sum_t y_t x_t) \\ -(\sum_t x_t)(\sum_t y_t) + N(\sum_t y_t x_t) \end{bmatrix}$$

which agrees with the formula for $\hat{\beta}_1$ in [2, Theorem 11.3, p. 557] (their n is my N), namely

$$\hat{\beta}_1 = \frac{N(\sum_t y_t x_t) - (\sum_t x_t)(\sum_t y_t)}{N(\sum_t x_t^2) - (\sum_t x_t)^2}. \tag{10}$$

Unfortunately this masks a much simpler expression for $\hat{\beta}_1$. We know from Remark 24.2.3 that when a constant term is included, the LS regression line passes through the sample means $(\bar{x}, \bar{y})$, so consider the **centered variables**

$$\tilde{x}_t = x_t - \bar{x}, \qquad \tilde{y}_t = y_t - \bar{y}.$$

The slope of the regression coefficient from regressing the centered $\tilde{y}$ on the centered $\tilde{x}$ will be exactly the same as the slope of the regression line from regressing y on x. This means that in (10) we may replace x_t and y_t by $\tilde{x}_t$ and $\tilde{y}_t$. But by construction,

$$\sum_t \tilde{x}_t = \sum_t \tilde{y}_t = 0,$$

so

$$\hat{\beta}_1 = \frac{\sum_t \tilde{y}_t \tilde{x}_t}{\sum_t \tilde{x}_t^2} = \frac{\sum_t (y_t - \bar{y})(x_t - \bar{x})}{\sum_t (x_t - \bar{x})^2}. \tag{11}$$

24.5 Regression and correlation

Recall from Lecture 9.15 that the correlation between random variables X and Y is defined to be

$$\operatorname{Corr}(X, Y) = \frac{\boldsymbol{Cov}(X, Y)}{(\operatorname{SD} X)(\operatorname{SD} Y)}$$

It is also equal to

$$\operatorname{Corr}(X, Y) = \boldsymbol{Cov}(X^*, Y^*) = \boldsymbol{E}(X^* Y^*),$$

where X^* and Y^* are the standardization of X and Y.

A natural estimate of the correlation was proposed by Karl Pearson. Given pairs (x_t, y_t), $t = 1, \dots, N$, of observations, define the sample **correlation coefficient** r by

$$r = \frac{\frac{\sum_{t=1}^N (x_t - \bar{x})(y_t - \bar{y})}{N}}{\sqrt{\frac{\sum_{t=1}^N (x_t - \bar{x})^2}{N}} \sqrt{\frac{\sum_{t=1}^N (y_t - \bar{y})^2}{N}}}$$

$$= \frac{\sum_{t=1}^N (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^N (x_t - \bar{x})^2} \sqrt{\sum_{t=1}^N (y_t - \bar{y})^2}}$$

which is the sample analog of the correlation. It is also known as the **Pearson product-moment correlation coefficient**. Note that in many regression applications, the variate X is not necessarily a random variable, but the sample correlation coefficient is still defined.

Define

$$s_x = \sqrt{\frac{\sum_t (x_t - \bar{x})^2}{N}} \quad \text{and} \quad s_y = \sqrt{\frac{\sum_t (y_t - \bar{y})^2}{N}}$$

Use (11) to rewrite the formula for the correlation coefficient as

$$r = \frac{\sum_{t=1}^N (x_t - \bar{x})(y_t - \bar{y})/N}{s_x s_y} = \hat{\beta}_1 \frac{s_x}{s_y}. \tag{12}$$

Among other things this implies that $r = 0$ if and only the slope $\hat{\beta}_1$ of the regression line is zero. (If $s_x = 0$, then all the x_t are the same, and $\hat{\beta}_1$ is undefined and the slope is not identifiable.)

If we go one step further and "standardize" the sample,

$$x_t^* = \tilde{x}_t / s_x, \qquad y_t^* = \tilde{y}_t / s_y,$$

then the slope of the regression line of y^* on x^* is the same as the slope of the regression line of x^* on y^*, and both are equal to the correlation coefficient.

24.6 Regress Y on X or X on Y?

When there are only two variates of interest, X and Y, should you "regress Y on X," that is, estimate the mode $Y = \beta_0 + \beta_1 X + \varepsilon$; or should you "regress X on y," that is, estimate the model $X = \alpha_0 + \alpha_1 X + \eta$? This is not a statistical question, it is a scientific question. The statistical model $Y = \beta_0 + \beta_1 X + \varepsilon$ implicitly assumes that Y is influenced by X, but not vice-versa. Your scientific theory should tell you which is the appropriate model. For instance, the spring rainfall influences the tonnage of the grape harvest in October, not the other way around.

Occasionally, there might be a case that each variate influences the other. For instance, if Y is the crime rate and X is the hours of police patrols in a neighborhood, then a high crime rate might cause the police department to dispatch more patrols, but a higher police presence may deter crime. In situations like this, we need a more sophisticated model with equations for the demand for policing as a function of the crime rate, and the supply of patrols as response to the crime rate. You need to find a way to estimate the parameters of the equations separately. This problem is common in economics and political science and effort has gone into developing method for dealing with such "simultaneous equations" models. They are beyond the scope of this course, but I can recommend **Ec 122: Econometrics** to learn more about it.

But whatever you decide, if you use only two variates X and Y, then (11) tells us that the ratio of the two slopes satisfies

$$\frac{\hat\beta_{Y\text{on}X}}{\hat\beta_{X\text{on}Y}} = \frac{\sum_t (y_t - \bar y)^2}{\sum_t (x_t - \bar x)^2},$$

so there is little we can say a priori. In particular, even if the constant term is zero in each regression, it is not necessarily the case that $\hat\beta_{Y\text{on}X} = 1/\hat\beta_{X\text{on}Y}$.

24.7 LS and MLE

> When the error vector $\boldsymbol{\varepsilon}$ has a multivariate normal distribution $N(\mathbf{0}, \sigma^2 \boldsymbol{I})$ distribution, then the LS estimator of $\boldsymbol{\beta}$ is also the Maximum Likelihood Estimator.

To see this, write $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$. Then the density of $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ is the multivariate normal density $N(\mathbf{0}, \sigma^2 \boldsymbol{I})$

$$\left(\frac{1}{\sqrt{2\pi}}\right)^N \frac{1}{\sqrt{\det \sigma^2 \boldsymbol{I}}}\, e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})} = \left(\frac{1}{\sqrt{2\pi}}\right)^N \left(\frac{1}{(\sigma^2)^N}\right)^{\frac{1}{2}}\, e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}$$

Taking logs, we find the log likelihood function is

$$-\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Maximizing this with respect to β amounts to minimizing $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, which is exactly what LS does.

The first order condition for the maximum with respect to σ^2 is

$$-\frac{N}{2}\frac{1}{\widehat{\sigma^2}} + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}})\frac{1}{(\widehat{\sigma^2})^2} = 0.$$

Then as in Lecture 18, multiply by $2(\widehat{\sigma^2})^2$ to get

$$-N\widehat{\sigma^2} + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}}) = 0,$$

so

$$\hat\sigma^2_{\text{MLE}} = \frac{\boldsymbol{e}'\boldsymbol{e}}{N},$$

where

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}}.$$

24.8 ⋆ Statistics of the LS estimator

We now compute the mean and covariance matrix of the random vector $\hat{\boldsymbol{\beta}}_{\text{LS}}$. By (6)

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon},$$

so

$$\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

which is multivariate normal if $\boldsymbol{\varepsilon}$ is. Taking the expectation we see

$$\boldsymbol{E}(\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta}) = \boldsymbol{E}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\boldsymbol{E}\,\boldsymbol{\varepsilon} = 0.$$

In other words, $\hat{\boldsymbol{\beta}}_{\text{LS}}$ is **unbiased**,

$$\boldsymbol{E}\,\hat{\boldsymbol{\beta}}_{\text{LS}} = \boldsymbol{\beta}.$$

Furthermore,

$$(\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta})' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

so the covariance matrix is

$$\begin{aligned}\boldsymbol{Var}(\hat{\boldsymbol{\beta}}_{\text{LS}}) &= \boldsymbol{E}(\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{\text{LS}} - \beta)' \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

24.9 ⋆ Gauss–Markov Theorem

This does not depend on the normality of the error terms; only that they mean zero, so we should move this to an earlier section.

24.9.1 Gauss–Markov Theorem *In the standard linear model, if $\boldsymbol{X}$ has rank $(K+1)$, then the LS estimator $\hat{\boldsymbol{\beta}}_{LS}$ is the **Best Linear Unbiased Estimate (BLUE)** of $\boldsymbol{\beta}$ in the following sense. Given any other estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$ which is linear in $\boldsymbol{y}$ and which satisfies $\boldsymbol{E}\,\boldsymbol{b} = \boldsymbol{\beta}$ for any possible value of $\boldsymbol{\beta}$, then $\boldsymbol{Var}\,\boldsymbol{b} = \boldsymbol{Var}\,\hat{\boldsymbol{\beta}}_{LS} + \boldsymbol{P}$, where $\boldsymbol{P}$ is positive semidefinite.*

Proof: Let $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{y}$ be a linear function of $\boldsymbol{y}$ where $\boldsymbol{A}$ is a $(K+1) \times N$ matrix. Define the $(K+1) \times N$ matrix $\boldsymbol{D}$ to be the difference between $\boldsymbol{A}$ and the matrix for the LS estimator $\hat{\boldsymbol{\beta}}_{\text{LS}}$. That is,

$$\boldsymbol{D} = \boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

Then we may write $\boldsymbol{A} = \boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ so that

$$\begin{aligned}\boldsymbol{b} = \boldsymbol{A}\boldsymbol{y} = \big(\boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\big)\boldsymbol{y} &= \big(\boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}X'\big)(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta} + \big(\boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\big)\boldsymbol{\varepsilon},\end{aligned}$$

so

$$\boldsymbol{b} - \boldsymbol{\beta} = \boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta} + \big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D}\big)\boldsymbol{\varepsilon}. \tag{13}$$

So in expectation, since **expectation is a positive linear operator**,

$$\boldsymbol{E}\,\boldsymbol{b} - \boldsymbol{\beta} = \boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta} + \underbrace{\big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D}\big)\,\boldsymbol{E}\,\boldsymbol{\varepsilon}}_{=0}.$$

Thus $\boldsymbol{b}$ is unbiased if and only if $\boldsymbol{DX\beta} = \mathbf{0}$. But since we require this result to hold for all possible values of $\boldsymbol{\beta}$, it follows that

$$\boldsymbol{DX} = \mathbf{0},$$

in which case (13) becomes

$$\boldsymbol{b} - \boldsymbol{\beta} = \big(\boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\big)\boldsymbol{\varepsilon}.$$

So for an unbiased linear estimator $\boldsymbol{b}$,

$$\begin{aligned}
\boldsymbol{Var}\,\boldsymbol{b} &= \boldsymbol{E}(\boldsymbol{b} - \boldsymbol{\beta})(\boldsymbol{b} - \boldsymbol{\beta})' \\
&= \big(\boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\big)\,\boldsymbol{E}\,\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\big(\boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\big)' \\
&= \sigma^2\big(\boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\big)\big(\boldsymbol{D}' + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\big) \\
&= \sigma^2\big(\boldsymbol{D}\boldsymbol{D}' + \underbrace{\boldsymbol{D}\boldsymbol{X}}_{=\mathbf{0}}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}'\boldsymbol{D}'}_{=\mathbf{0}}\big) + (\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2\boldsymbol{D}\boldsymbol{D}' + \boldsymbol{Var}\,\hat{\boldsymbol{\beta}}_{\text{LS}}.
\end{aligned}$$

Now $\boldsymbol{P} = \sigma^2\boldsymbol{D}\boldsymbol{D}'$ is positive semidefinite. ∎

Among other things, this implies that $\boldsymbol{Var}\,b_k = \boldsymbol{Var}\,\hat{\boldsymbol{\beta}}_{\text{LS}k} + P_{kk} \geqslant \boldsymbol{Var}\,\hat{\boldsymbol{\beta}}_{\text{LS}k}$ for each k.

24.9.1 Estimating σ^2

Let

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LS}}$$

be the vector of LS residuals. Then we can rewrite this as

$$\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{\varepsilon},$$

where

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}',$$

the **annihilator** of the column space of $\boldsymbol{X}$. (That is, $\boldsymbol{M}$ is the projection operator onto the orthogonal complement of the column space of $\boldsymbol{X}$.) The matrix $\boldsymbol{M}$ is symmetric and idempotent, that is, $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$, and $\operatorname{trace}\boldsymbol{M} = N - (K+1)$. (See Section S2.8.)

Thus

$$\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{\varepsilon}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}.$$

Since $\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}$ is $1\times$, it is equal to its trace, and since trace is a linear operator, the expected value of the trace of a random matrix is the trace of the expected matrix. Thus

$$\begin{aligned}
\boldsymbol{E}(\boldsymbol{e}'\boldsymbol{e}) &= \boldsymbol{E}(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}) \\
&= \boldsymbol{E}\big(\operatorname{trace}(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon})\big) \\
&= \boldsymbol{E}\big(\operatorname{trace}(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\big) \\
&= \operatorname{trace}\big(\boldsymbol{M}\,\boldsymbol{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\big) \\
&= \operatorname{trace}\big(\boldsymbol{M}(\sigma^2 I)\big) \\
&= \sigma^2 \operatorname{trace}(\boldsymbol{M}) \\
&= (N - (K+1))\sigma^2
\end{aligned}$$

Therefore $\sigma^2 = \dfrac{\boldsymbol{E}(\boldsymbol{e}'\boldsymbol{e})}{N-(K+1)}$. Define

$$s^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{N-(K+1)}.$$

24.9.2 Theorem *If $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$, then $\hat{\boldsymbol{\beta}}_{LS} \sim N\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$, and*

$$\frac{(N-(K+1))s^2}{\sigma^2} \sim \chi^2(N-(K+1))$$

Also, $\hat{\boldsymbol{\beta}}_{LS}$ and s^2 are independent.

Therefore for any $(K+1)$-vector $\boldsymbol{w}$ of weights,

$$\boldsymbol{w}'(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}) \sim N\left(\mathbf{0}, \sigma^2 \boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w}\right),$$

so

$$\frac{\boldsymbol{w}'(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})}{s\sqrt{\boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w}}} \sim t(N-(K+1)). \tag{14}$$

The unbiased estimator s^2 of σ^2 is

$$s^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{N-(K+1)}.$$

The square root s of the estimated variance,

$$s = \sqrt{\frac{\boldsymbol{e}'\boldsymbol{e}}{N-(K+1)}}$$

is called the **residual standard error**. (Note that since s^2 is an unbiased estimate of σ^2, Jensen's Inequality implies that s is *not* an unbiased estimate of σ.)

We also saw that the covariance matrix of the random vector $\hat{\boldsymbol{\beta}}_{\text{LS}}$ is

$$\boldsymbol{Var}\,\hat{\boldsymbol{\beta}}_{\text{LS}} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{15}$$

To see the last assertion, note that

$$\frac{\boldsymbol{w}'(\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta})}{\sqrt{\sigma^2\boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w}}} \sim N(0,1),$$

and is independent of the $\chi^2(N-(K+1))$ random variable $\frac{(N-(K+1))s^2}{\sigma^2}$, we have that

$$\frac{\frac{\boldsymbol{w}'(\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta})}{\sqrt{\sigma^2\boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w}}}}{\sqrt{\frac{\frac{(N-(K+1))s^2}{\sigma^2}}{N-(K+1)}}} \sim t(N-(K+1)),$$

which reduces to (14).

From (14) with $\boldsymbol{w}$ being the k^{th} coordinate vector we have the following:

24.9.3 Corollary

$$\frac{\hat{\beta}_k - \beta_k}{s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}} \sim t(N-(K+1)).$$

Recall that (15) *implies that $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk} = \boldsymbol{Var}\,\hat{\boldsymbol{\beta}}_{LSk}$, so $s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}$ is the estimated standard deviation of $\hat{\boldsymbol{\beta}}_{LSk}$, and is called the* ***standard error*** *of $\hat{\boldsymbol{\beta}}_{LSk}$.*

That means

$$P\left(-t_{\alpha/2,N-(K+1)} \leqslant \frac{\hat\beta_k - \beta_k}{s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}} \leqslant t_{\alpha/2,N-(K+1)} \right) = 1-\alpha,$$

or equivalently,

$$P\left(-t_{\alpha/2,N-(K+1)} s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}} \leqslant \hat\beta_k - \beta_k \leqslant t_{\alpha/2,N-(K+1)} s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}} \right) = 1-\alpha$$

or equivalently,

$$P\left(\hat\beta_k - t_{\alpha/2,N-(K+1)} s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}} \leqslant \beta_k \leqslant \hat\beta_k + t_{\alpha/2,N-(K+1)} s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}} \right) = 1-\alpha$$

The $1-\alpha$ confidence interval for β_k is

$$\left(\hat\beta_k - t_{\alpha/2,N-(K+1)} s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}},\quad \hat\beta_k + t_{\alpha/2,N-(K+1)} s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}} \right) \tag{16}$$

Recall $s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}$ is the estimated standard deviation or **standard error** of $\hat{\boldsymbol{\beta}}_{\text{LS}k}$.

24.10 Testing common hypotheses

The typical hypothesis test associated with a linear model is a hypothesis regarding some β_k. So let the null hypothesis be

$$H_0 \colon \beta_k = \beta_k^0.$$

We know from Corollary 24.9.3

$$\frac{\hat\beta_k - \beta_k}{s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}} \sim t(N-(K+1)).$$

So to test the null hypothesis against the two-sided alternative

$$H_1 \colon \beta_k \neq \beta_k^0$$

at the α level of significance we find the $1-\alpha/2$ quantile $t_{\alpha/2,N-(K+1)}$ of the t-distribution with $N-(K+1)$ degrees of freedom. We next compute the test statistic

$$t = \frac{\hat{\boldsymbol{\beta}}_{\text{LS}k} - \beta_k^0}{s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}}$$

We reject the null hypothesis if $t < -t_{\alpha/2,N-(K+1)}$ or if $t > t_{\alpha/2,N-(K+1)}$ Equivalently we reject the null hypothesis if $|t| > t_{\alpha/2,N-(K+1)}$.

One-sided alternatives are treated similarly, *mutatis mutandis*.

A common null hypothesis is even more specific, namely

$$H_0 \colon \hat\beta_k = 0.$$

The test statistic for this is even simpler:

$$t = \frac{\hat{\boldsymbol{\beta}}_{\text{LS}k}}{s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}}.$$

This is what statistical software (including R and Mathematica) will use to compute a "t-value" for your estimated coefficients.

Bibliography

[1] C. W. Cobb and P. H. Douglas. 1928. A theory of production. *American Economic Review* 18(1):139–165. Supplement, Papers and Proceedings of the Fortieth Annual Meeting of the American Economic Association http://www.jstor.org/stable/1811556

[2] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[3] J. H. McCulloch. 1985. Miscellanea: On heteros*edasticity. *Econometrica* 53(2):483–483. http://www.jstor.org/stable/1911250