Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2021

Lecture 25: The Standard Linear Model: Hypothesis Testing

Relevant textbook passages:

Larsen–Marx [4]: Section 11.4. Chapter 12.

Theil [6]: Chapters 3–5.

25.1 The LS estimator

Let there be N observations on K **regressors** X plus a constant term, and a **response** y. The matrix X is $N \times (K+1)$, and assume it has rank $(K+1)$. A **constant term** is a column of all ones. The standard linear model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\boldsymbol{E}\,\boldsymbol{\varepsilon} = 0 \quad \text{and} \quad \boldsymbol{Cov}\,\boldsymbol{\varepsilon} = \boldsymbol{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{I}.$$

The LS estimator $\hat{\boldsymbol{\beta}}_{\text{LS}}$ of the **coefficients** $\boldsymbol{\beta}$ is given by

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

The coefficient on the constant term is called the **intercept**. Set

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

and

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

The vector $\boldsymbol{e}$ of **residuals** satisfies

$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}.$$

That is, $\boldsymbol{e}$ is orthogonal to each column vector of regressors. When there is a constant term in the regression, then the sum of residuals $\boldsymbol{1}'\boldsymbol{e} = 0$, so $\bar{e}$, the mean residual is also zero.

25.2 ⋆ More complicated hypotheses

Sometimes we entertain more complicated null hypotheses on our coefficients. For instance, economists are often interested in whether estimated shares of income add up to one. This material is not in Larsen and Marx [4], but can be found for instance, in Theil [6, pp. 143–144].

The general **linear hypothesis** of q linear constraints on the vector $\boldsymbol{\beta}$ can be written simply as

$$H_0\colon \boldsymbol{a} = \boldsymbol{A}\boldsymbol{\beta},$$

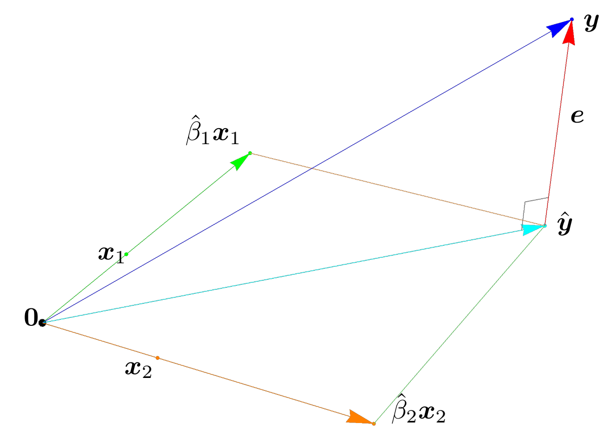


Figure 25.1. Geometry of LS.

where $\boldsymbol{A}$ is a $q \times (K+1)$ matrix and $\boldsymbol{a}$ is $q \times 1$ column vector. We assume that $\boldsymbol{A}$ has rank q.

For instance the single hypothesis that $\beta_1 + \cdots + \beta_K = 1$ (so $q = 1$) uses $a = [1]$ and $A = [0, 1, \ldots, 1]$.

The next result may be found in [6, Theorem 3.9, p. 144]. It may seem impossibly opaque now, but I will explain it in Section 25.4⋆.

25.2.1 Theorem *Under the null hypothesis, the test statistic*

$$F = \frac{1}{qs^2}(\boldsymbol{a} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_{\text{LS}})'\big[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\big](\boldsymbol{a} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_{\text{LS}})$$

has a Snedecor F-distribution with $\big(q, N - (K+1)\big)$ degrees of freedom.

Tests based on these statistics are called ***F*-tests**.

Many software packages, including R, compute for you something called the "F-statistic for the regression." If you have a constant term, it is usually the first one, X_1 in our terminology. The F-statistic for the regression test the null hypothesis that all the coefficients on the non-constant terms are zero,

$$H_0\colon \beta_1 = \cdots = \beta_K = 0.$$

The F-statistic for this test has $\big(K, N - (K+1)\big)$ degrees of freedom, and it only makes sense to perform this test if one of the columns is a constant term. This is a hypothesis that you probably would like to reject, so a small p-value (or large F-statistic) is a good thing. The reason this is important is that if the columns of X are nearly collinear, each individual $\hat{\boldsymbol{\beta}}_{\text{LS}k}$ is hard to estimate, that is, it will have a large standard error, so each one may turn out to be statistically insignificant different from zero, but together they could truly determine y.

25.3 ⋆ Lagrange Multiplier Theorem

Before I can explain why the test of general linear restriction is based on an F-test, or the ratio of independent χ^2 random variables, we need to characterize how to impose restrictions on a minimization problem.

25.3.1 Lagrange Multiplier Theorem *Let X be a subset of $\boldsymbol{R}^{\mathrm{n}}$, and assume that the functions $f, g_1, \dots, g_m \colon X \to \boldsymbol{R}$ are continuous. Let x^* be an interior constrained local minimizer of f subject to $g_i(x) = 0$, $i = 1, \dots, m$. Suppose f, $g_1, \dots, g_m$ are differentiable at x^*, and that $\nabla g_1(x^*), \dots, \nabla g_m(x^*)$ are linearly independent.*

Then there exist real numbers $\lambda_1^*, \dots, \lambda_m^*$, such that

$$\nabla f(x^*) - \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) = 0. \tag{2}$$

The function

$$L(x, \lambda) = f(x) - \sum_{i=1}^{n} \lambda_i g_i(x)$$

is called the Lagrangean and (2) is the set of **first order conditions** for a constrained extremum. It is also the set of FOCs for the minimization of the Lagrangean. When f is quadratic, and each g_i is linear, it turns out that the constrained minimizer of f is an unconstrained minimizer of the Lagrangean $L(x, \lambda^*)$. For a discussion and proof of the Lagrange Multiplier Theorem, see my online notes [1].

25.4 ⋆ Restricted regression

In Section 25.2 ⋆, I asserted how to test a linear restriction on the coefficients. Here's how to derive that.

Not in Larsen–Marx [4].

In the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3}$$

to test the linear restriction on $\boldsymbol{\beta}$

$$H_0 \colon \boldsymbol{a} = \boldsymbol{A}\boldsymbol{\beta},$$

where $\boldsymbol{A}$ is $q \times (K+1)$ and has full rank, we first estimate (3) by minimizing the sum of squared residuals without restrictions, this gives us our usual estimates, but let me label them

$$\hat{\boldsymbol{b}}_u = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \qquad \hat{\boldsymbol{y}}_u = \boldsymbol{X}\hat{\boldsymbol{b}}_{\boldsymbol{u}}, \qquad \boldsymbol{e}_u = \boldsymbol{y} - \hat{\boldsymbol{y}}_u.$$

Next we impose the restriction $\boldsymbol{a} = \boldsymbol{A}\boldsymbol{b}$ and minimize the sum of squared residuals (with respect to $\boldsymbol{b}$) to get the restricted estimates: $\hat{\boldsymbol{b}}_r$, $\hat{\boldsymbol{y}}_r$, and residuals $\boldsymbol{e}_r$.

To minimize the sum of squared residuals subject to the constraints, instead of projecting $\boldsymbol{y}$ onto the space $\{\boldsymbol{X}\boldsymbol{b} : \boldsymbol{b} \in \boldsymbol{R}^{(K+1)}\}$ spanned by the columns of $\boldsymbol{X}$, we have to project $\boldsymbol{y}$ onto $\{\boldsymbol{X}\boldsymbol{b} : \boldsymbol{b} \in \boldsymbol{R}^{(K+1)}$ and $\boldsymbol{A}\boldsymbol{b} = \boldsymbol{a}\}$. The point $\hat{\boldsymbol{y}}_r$ is still a linear combination of the columns of $\boldsymbol{X}$, so the unrestricted residual vector $\boldsymbol{e}_u$ is orthogonal to $\boldsymbol{y}_u - \boldsymbol{y}_r$ and the Pythagorean Theorem tells us that

$$\boldsymbol{e}_r{}'\boldsymbol{e}_r - \boldsymbol{e}_u{}'\boldsymbol{e}_u = (\hat{\boldsymbol{y}}_u - \hat{\boldsymbol{y}}_r)'(\hat{\boldsymbol{y}}_u - \hat{\boldsymbol{y}}_r).$$

See Figure 25.2.

The first question is, "What is the formula for $\hat{\boldsymbol{b}}_r$?" To answer this, note that the Lagrangean for this minimization is

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) - \boldsymbol{\lambda}'(\boldsymbol{A}\boldsymbol{b} - \boldsymbol{a})$$

where $\boldsymbol{\lambda}$ is a q-vector of Lagrange multipliers. The Lagrange Multiplier Theorem states that the first order condition for a constrained minimum is that the vector of partial derivatives with respect to $\boldsymbol{b}$ of the Lagrangean must be zero,

$$-2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}}_r - \boldsymbol{A}'\boldsymbol{\lambda}^* = \boldsymbol{0}. \tag{4}$$

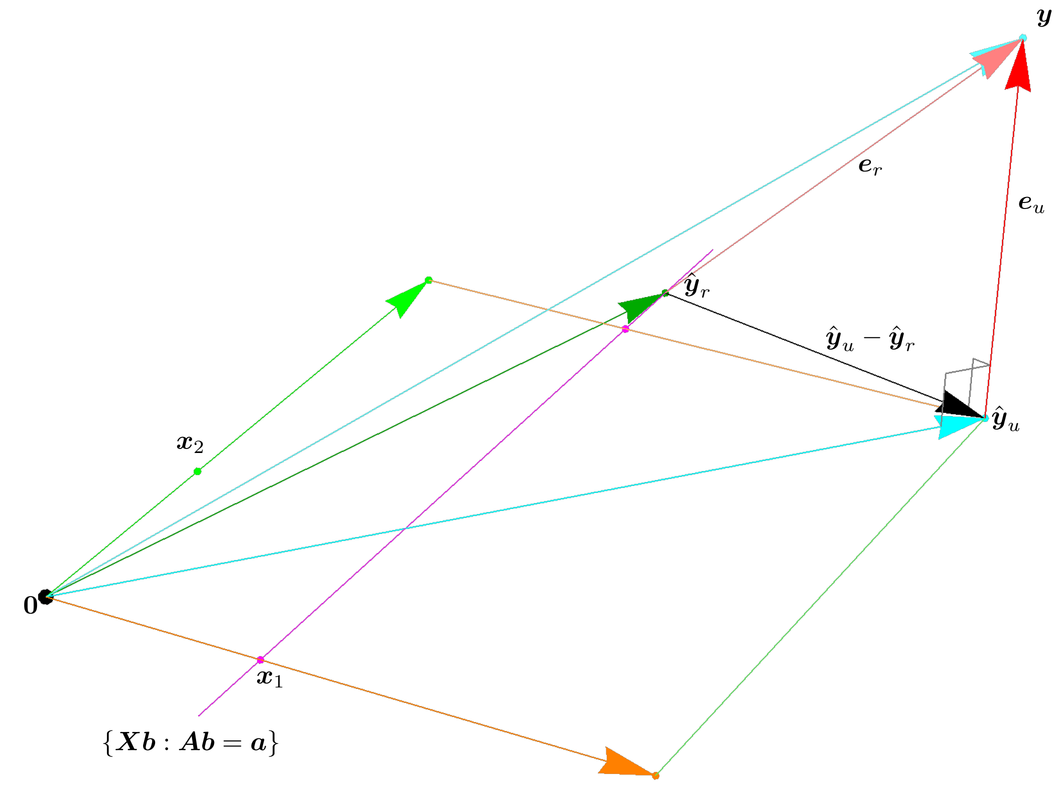


Figure 25.2. Restricted regression with restriction $\beta_1 = 1$. The points $\hat{\boldsymbol{y}}_r$, $\hat{\boldsymbol{y}}_u$, $\boldsymbol{y}$ form a right triangle with hypotenuse $\overline{\hat{\boldsymbol{y}}_r \boldsymbol{y}}$.

Premultiply by $A(\boldsymbol{X}'\boldsymbol{X})^{-1}$:

$$-2\boldsymbol{A}\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}}_{=\hat{\boldsymbol{b}}_u}+2\underbrace{\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{b}}_r}_{=\boldsymbol{A}\hat{\boldsymbol{b}}_r=\boldsymbol{a}}-\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\boldsymbol{\lambda}^*=\boldsymbol{0},$$

or

$$\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\boldsymbol{\lambda}^*=2(\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_u). \tag{5}$$

I now claim that $\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'$ is a $q \times q$ matrix of full rank,[1] so it has an inverse. This let's us solve (5) for $\boldsymbol{\lambda}^*$:

$$\boldsymbol{\lambda}^*=2\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_u].$$

Substitute this into (4) to get

$$-\boldsymbol{X}'\boldsymbol{y}+\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}}_r-\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_u]=\boldsymbol{0}.$$

Premultiply this by $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ to get

$$-\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}}_{=\hat{\boldsymbol{b}}_u}+\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}}_r}_{=\hat{\boldsymbol{b}}_r}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_u]=\boldsymbol{0}$$

so the formula for the restricted estimator is:

$$\boxed{\hat{\boldsymbol{b}}_r=\hat{\boldsymbol{b}}_u+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_u] \qquad (6)}$$

The next question is, “What is the sum of squared residuals?” Well

$$\boldsymbol{e}_r{}'\boldsymbol{e}_r=(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{b}}_r)'(\boldsymbol{y}-\hat{\boldsymbol{b}}_r)$$

and we can use (6) to write

$$\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{b}}_r=\underbrace{\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{b}}_u}_{=\boldsymbol{e}_u}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_u]=\boldsymbol{e}_u+\boldsymbol{X}\boldsymbol{C},$$

where $\boldsymbol{C}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_u]$. This means we can write

$$\boldsymbol{e}_r{}'\boldsymbol{e}_r=(\boldsymbol{e}_u+\boldsymbol{X}\boldsymbol{C})'(\boldsymbol{e}_u+\boldsymbol{X}\boldsymbol{C})=\boldsymbol{e}_u{}'\boldsymbol{e}_u+2\boldsymbol{e}_u{}'\boldsymbol{X}\boldsymbol{C}+(\boldsymbol{X}\boldsymbol{C})'(\boldsymbol{X}\boldsymbol{C}).$$

But since the unrestricted residual vector $\boldsymbol{e}_u$ is orthogonal to the columns of $\boldsymbol{X}$, we have $\boldsymbol{e}_u{}'\boldsymbol{X}=\boldsymbol{0}$. Now let's write out (recalling that the transpose of a product is the product of the transpose in reverse order, and observing that $[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ are symmetric):

$$\begin{aligned}
&(\boldsymbol{X}\boldsymbol{C})'(\boldsymbol{X}\boldsymbol{C})=\\
&\quad[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_u]'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_u]\\
&\qquad\qquad=[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_{\boldsymbol{u}}]'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_{\boldsymbol{u}}]
\end{aligned}$$

Thus

$$\boxed{\boldsymbol{e}_r{}'\boldsymbol{e}_r=\boldsymbol{e}_u{}'\boldsymbol{e}_u+[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_u]'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}A']^{-1}[\boldsymbol{a}-\boldsymbol{A}\hat{\boldsymbol{b}}_u]. \qquad (7)}$$

[1] To see that the $q \times q$ matrix $\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'$ has full rank, suppose

$$\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\boldsymbol{b}=\boldsymbol{0}$$

for some q-vector $\boldsymbol{b}$. We want to show that this implies that $\boldsymbol{b}=\boldsymbol{0}$.

Since $\boldsymbol{A}$ has rank q, so does $\boldsymbol{A}'\boldsymbol{A}$, so $(\boldsymbol{A}'\boldsymbol{A})^{-1}$ exists. Premultiply the equation above by $(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ to get

$$(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{A}'\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\boldsymbol{b}=\boldsymbol{A}'\boldsymbol{b}=\boldsymbol{0}.$$

Since $\boldsymbol{A}$ has rank q, this implies $\boldsymbol{b}=\boldsymbol{0}$. q.e.d.

A test of H_0 amounts to a test of the null hypothesis

$$H_0\colon \quad \frac{\boldsymbol{e}'_r\boldsymbol{e}_r}{\boldsymbol{e}_u{}'\boldsymbol{e}_u} = 1$$

versus the alternative hypothesis

$$H_1\colon \quad \frac{\boldsymbol{e}'_r\boldsymbol{e}_r}{\boldsymbol{e}_u{}'\boldsymbol{e}_u} = \frac{\boldsymbol{e}_u{}'\boldsymbol{e}_u + [\boldsymbol{a} - \boldsymbol{A}\hat{\boldsymbol{b}}_u]'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}[\boldsymbol{a} - \boldsymbol{A}\hat{\boldsymbol{b}}_u]}{\boldsymbol{e}_u{}'\boldsymbol{e}_u} > 1.$$

Now recalling that $s^2 = \boldsymbol{e}_u{}'\boldsymbol{e}_u/(N-(K+1))$ is the unbiased estimate of the variance we can rewrite the ratio as

$$1 + \frac{(\boldsymbol{a} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_{\text{LS}})'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{a} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_{\text{LS}})}{\big(N-(K+1)\big)s^2}.$$

So form the test statistic

$$F = \frac{[\boldsymbol{a} - \boldsymbol{A}\hat{\boldsymbol{b}}_u]'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}[\boldsymbol{a} - \boldsymbol{A}\hat{\boldsymbol{b}}_u]/q}{\boldsymbol{e}_u{}'\boldsymbol{e}_u/\big(N-(K+1)\big)}$$

Using a little tedious manipulation, you can show that if $\boldsymbol{a} = \boldsymbol{A}\hat{\boldsymbol{b}}_u$, that is, if the null hypothesis is true, then the numerator is distributed $\chi^2(q)$. The denominator is distributed as $\chi^2\big(N-(K+1)\big)$ and you can show that it is independent of the numerator. Thus, under the null hypothesis, the test statistic has an F-distribution with $\big(q, N-(K+1)\big)$ degrees of freedom. The null hypothesis should be rejected if $F \geqslant F_{1-\alpha,q,N-(K+1)}$.

For the missing details see, e.g., Theil [6], Section 1.8, pages 42–45, and the discussion on pages 141–144.

25.5 The $\boldsymbol{t}$-test vs. the $\boldsymbol{F}$-test

We have now seen to ways to test the null hypothesis $H_0\colon \beta_j = 0$ against the alternative $H_1\colon \beta_j \neq 0$. One way is to use the t-test from Section 24.10 based the test statistic

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

The other is to treat it as the linear restriction $\boldsymbol{e}^{j\prime}\boldsymbol{\beta} = 0$, and use the F-statistic, which for this case ($\boldsymbol{A} = \boldsymbol{e}^{j\prime}$, $\boldsymbol{a} = \boldsymbol{0}$, $q = 1$) becomes

$$F = \frac{\hat{\boldsymbol{\beta}}_{\text{LS}j}[(\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj}]^{-1}\hat{\boldsymbol{\beta}}_{\text{LS}j}}{s^2}.$$

Note that $F = t^2$, so either test (with the same significance level α) leads to the same decision.

25.6 Two $\boldsymbol{t}$-tests vs. a single $\boldsymbol{F}$-test

While the t-test and the F-test for a single $\hat{\boldsymbol{\beta}}_{\text{LS}j}$ agree, there can be disagreement between a joint test and two separate tests. For instance, consider the joint null hypothesis $H_0 : (\hat\beta_1, \hat\beta_2) = (0,0)$. This will be rejected at the $\alpha = 0.05\%$ level if the vector $(\hat\beta_1, \hat\beta_2)$ lies outside the ellipse shown in Figure 25.3. The point A is such a point, but B is not. Now consider the two hypotheses $H'_0\colon \hat\beta_1 = 0$ and $H''_0\colon \hat\beta_2 = 0$. The first will be rejected if L09$(\hat\beta_1, \hat\beta_2)$ lies outside the vertical lines, and the second will be rejected if $(\hat\beta_1, \hat\beta_2)$ lies outside the horizontal lines. Thus A fails the joint test, but passes both separate tests; and B passes the joint test, fails the each of the

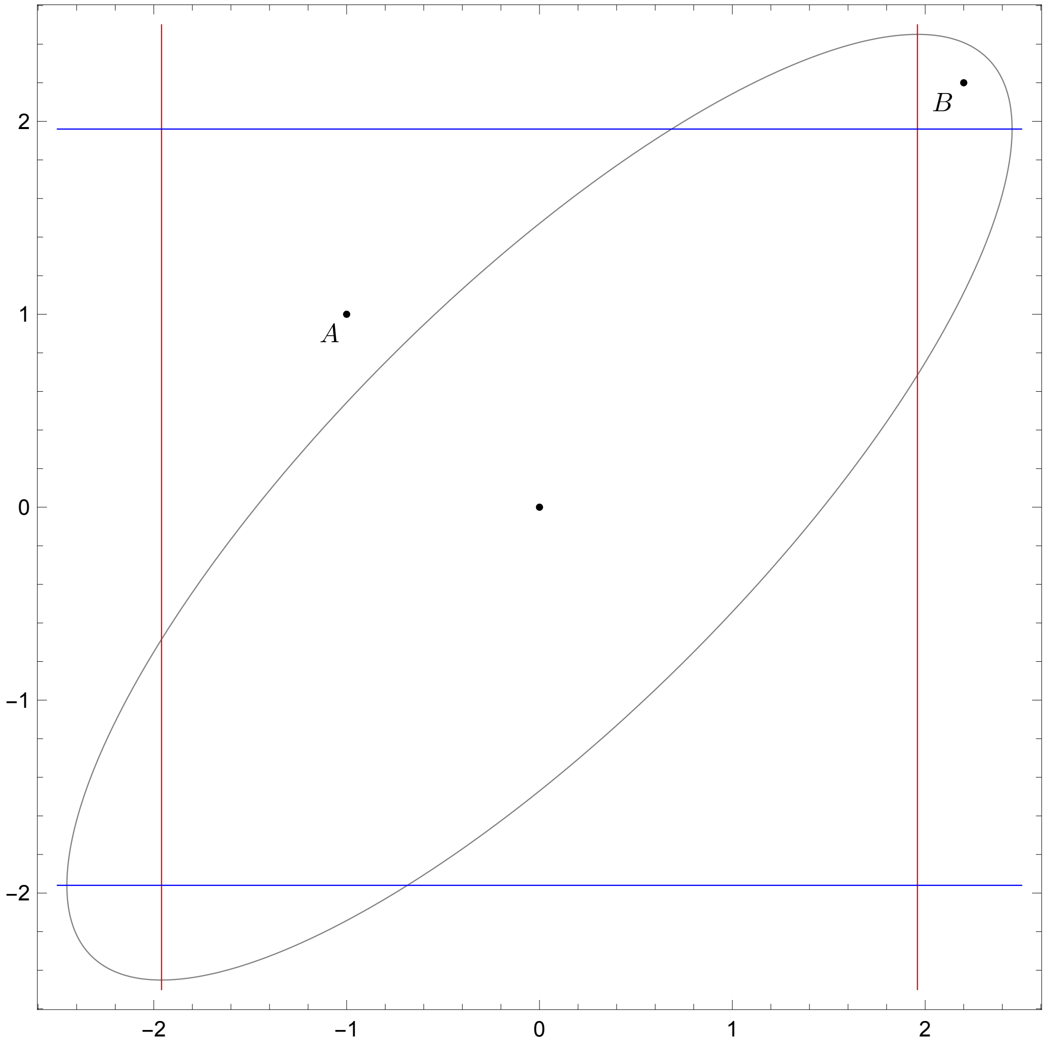


Figure 25.3. A joint test vs. two separate tests.

separate tests. This need not always happen, but it might especially when the matrix X is nearly singular. You as a researcher have to decide what is the appropriate test.

(By the way the figure is drawn assuming that $(\hat\beta_1, \hat\beta_2)$ is bivariate normal with mean zero and covariance matrix $\begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix}$. The correlation 0f 0.8 is important in making the ellipse the right size and shape for the example. The fact that I use the multivariate normal instead of a multivariate t is secondary.)

25.7 The Coefficient of Determination, R^2

Given the standard linear model

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}} + \boldsymbol{e} \\ &= \hat{\boldsymbol{y}} + \boldsymbol{e} \end{aligned}$$

the Pythagorean theorem tells us that

$$\boldsymbol{y}'\boldsymbol{y} = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}} + \boldsymbol{e}'\boldsymbol{e} + 2\hat{\boldsymbol{\beta}}_{\mathrm{LS}}'\underbrace{\boldsymbol{X}'\boldsymbol{e}}_{=0} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \boldsymbol{e}'\boldsymbol{e} \tag{8}$$

as $X'e = 0$.

Let $\bar y = \sum_t y_t/N$ be the sample average value of $\boldsymbol{y}$, and let $\tilde{\boldsymbol{y}}$ be the vector of centered values of $\boldsymbol{y}$, that is, $\tilde y_t = y_t - \bar y$. When the set of regressors includes a constant term, that is, the model is

$$Y_t = 1 \cdot \beta_0 + X_{t1}\beta_1 + \cdots + X_{tK}\beta_K + \varepsilon_t,$$

then regressing $\boldsymbol{y}$ on $\boldsymbol{x}$ and regressing $\tilde{\boldsymbol{y}}$ on $\boldsymbol{x}$ will give the same estimates of $\hat\beta_i$, $i = 1, \dots, K-1$ and exactly the same residuals e_t, $t = 1, \dots, N$. But the estimated coefficient $\hat\beta_0$ on the constant term shifts down by $\bar y$. In fact, the scatterplot of the data and the regression line just shift down by $\bar y$.

Now $\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}} = \sum_t (y_t - \bar y)^2$ is a measure of the total variation of y in the sample. There are two reasons the values y_t are not all the same and equal to $\bar y$. One is that the error terms ε_t are not all the same. The other is that the values x_{tk} of the regressors are not all the same. How much of the variation is due to randomness (ε_t's) and how much is "explained" by the regressors?

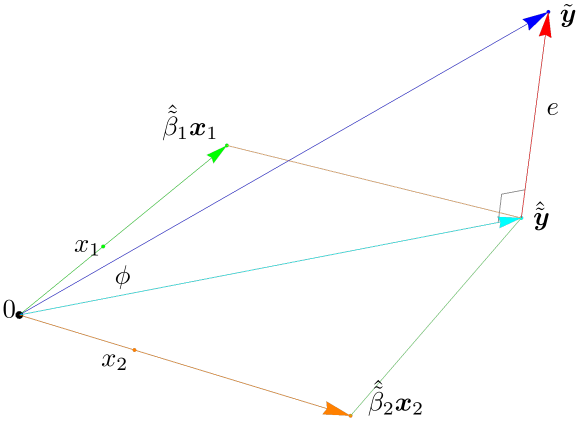
The **coefficient of determination**, which is denoted either by R^2 or r^2, is defined to be

$$R^2 = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}}},$$

and is often used as a measure of the fraction of the variation in $\boldsymbol{y}$ "explained" by the regressors. One of the consequences of (8) is that $\boldsymbol{y}'\boldsymbol{y} \geqslant \boldsymbol{e}'\boldsymbol{e}$ and $\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}} \geqslant \boldsymbol{e}'\boldsymbol{e}$, so it is always the case that

$$0 \leqslant R^2 \leqslant 1.$$

The geometric interpretation of R^2 is that $R = \sqrt{R^2}$ is the cosine of the angle between $\tilde{\boldsymbol{y}}$ and $\hat{\tilde{\boldsymbol{y}}}$. (Since $\boldsymbol{e}$ is orthogonal to $\hat{\tilde{\boldsymbol{y}}} = \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}}_{\mathrm{LS}}$, the three points $\boldsymbol{0}$, $\hat{\tilde{\boldsymbol{y}}}$, and $\boldsymbol{y}$ form a right triangle in $\boldsymbol{R}^{\mathrm{N}}$ with $\overline{\boldsymbol{0y}}$ as the hypotenuse. Thus R is the ratio $\|\hat{\tilde{\boldsymbol{y}}}\| \,/\, \|\tilde{\boldsymbol{y}}\|$, which is the cosine of the angle between them. An R near one corresponds to an angle near zero. The quantity R^2 is usually referred to simply as "R squared." Incidentally, almost no one talks about the value R for a regression. Instead they talk about R^2, and most software reports R^2.


$\tilde{\boldsymbol{y}}$
e
$\hat{\tilde{\beta}}_1 \boldsymbol{x}_1$
$\hat{\tilde{\boldsymbol{y}}}$
x_1
ϕ
0
x_2
$\hat{\tilde{\beta}}_2 \boldsymbol{x}_2$

Note that increasing the number of right-hand side variates can only decrease the sum of squared residuals, so it is desirable to penalize the measure of "fit." One way to do this is with the **adjusted $\boldsymbol{R^2}$**.

The **adjusted $\boldsymbol{\bar{R}^2}$** is defined by:

$$(1-\bar{R}^2) = \frac{\frac{1}{N-(K+1)}\boldsymbol{e}'\boldsymbol{e}}{\frac{1}{N-1}\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}}} = \frac{N-1}{N-(K+1)}(1-R^2)$$

or

$$\bar{R}^2 = \frac{1-(K+1)}{N-(K+1)} + \frac{N-1}{N-(K+1)}R^2.$$

It is possible for the adjusted R^2 to be negative.

I am embarrassed to say that if you read the textbooks in econometrics that I used as an undergraduate, you will find an entirely different definition of R^2, which is sometimes referred to as the **coefficient of multiple correlation**. In this definition, the vector $\tilde{\boldsymbol{y}}$ is replaced by $\boldsymbol{y}$ and R^2 is defined by

$$R^2 = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}'\boldsymbol{y}} = \frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{y}}.$$

(See e.g., Johnston [2, p. 36], Theil [6, p. 164].)

You might ask why do we divide $\boldsymbol{e}'\boldsymbol{e}$ by $\boldsymbol{y}'\boldsymbol{y}$ instead of $\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}}$? Theil [6, p. 164n] claims that it is for notational simplicity. But if the regression model includes a constant term (one of the columns of X is a vector of ones), then the mean residual is zero, so $\boldsymbol{e}'\boldsymbol{e}/\big(N-(K+1)\big)$ is the unbiased estimate of σ^2. Moreover the regression plane passes through the sample mean; $\bar{\boldsymbol{y}} = \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{\text{LS}}$, where $\bar{\boldsymbol{x}}$ is the vector of regressor sample means. So if we drop the constant term and express all variates as deviations from their sample mean, we still get the same estimate of $\boldsymbol{\beta}$.

What is a "good" value of $\boldsymbol{R^2}$

A natural question is, "Is this a good R^2?" Unfortunately there is no simple answer to that question. Sometimes you might hope to get a low value for the R^2 in a regression. For instance, in HW 6, you were given Problem 7.4.9 in Larsen–Marx [4, pp. 399–400], which concerned the age at which scientists did their best work. You were asked, "Plot the age versus date for these twelve discoveries. (Put the date on the abscissa.) Does the variability of the y_i's appear to be random with respect to time?" In HW 8 you are asked to regress the age on the date. In each instance, you are hoping that there is no significant time trend, so that the data from different centuries are comparable. In this case, you would hope for an insignificant coefficient and a low R^2. Other times you may hope to get a high R^2. For instance, if you are testing Hubble's Law using the data from [4, p. 543], you might be disappointed if your R^2 is less than 0.95.

There is no necessary connection between the R^2 of a regression and the significance of a particular coefficient. I use significance here in the usual sense that an estimated coefficient $\hat{\beta}_i$ is significant at the α level if the null hypothesis $H_0\colon \beta_i = 0$ is rejected at the α level. It could well be that the value of X has a nonzero effect on the value of Y, but that a regression of Y on X and a constant still has a low value of R^2 because the variance of the error term ε is large. On the other hand, I could have a zero coefficient on one regressor and the R^2 could still be close to one if the other regressor determined Y.

25.8 Prediction intervals in the linear model

We often want to "predict" the value of Y for values of $X_1, \dots, X_K$ that we have not observed. (Recall the *Road & Track* example.) Let $\boldsymbol{x}_* = (1, x_1, \dots, x_K)$ be a vector of values for the independent variables (with a constant term). Our model predicts

$$y_* = \boldsymbol{x}_*'\boldsymbol{\beta} + \varepsilon_*$$

so we form the prediction

$$\hat{y}_* = x_*'\hat{\boldsymbol{\beta}}_{\text{LS}}.$$

What is the confidence interval for y_*? Now

$$\hat{y}_* - y_* = \boldsymbol{x}_*'\hat{\boldsymbol{\beta}}_{\text{LS}} - x_*'\boldsymbol{\beta} - \varepsilon_* = \boldsymbol{x}_*'(\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta}) - \varepsilon_*.$$

Therefore

$$\begin{aligned}\boldsymbol{Var}(\hat{y}_* - y_*) &= \boldsymbol{Var}\Big(\boldsymbol{x}_*'(\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta}) - \varepsilon_*\Big)\\ &= \boldsymbol{Var}\Big(\boldsymbol{x}_*'(\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta})\Big) + \boldsymbol{Var}(\varepsilon_*)\\ &= \sigma^2(\boldsymbol{x}_*'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_* + 1).\end{aligned}$$

This tells us that there are two components to the variance of the predicted $\hat{y}_*$ and the realization y_*. In addition to the variance due to the error term ε_*, there is additional variance due to the fact that we are predicting based on our estimate $\hat{\boldsymbol{\beta}}_{\text{LS}}$, which is a random variable with its own variance.

Therefore

$$\frac{\boldsymbol{x}_*'\hat{\boldsymbol{\beta}}_{\text{LS}} - y_*}{\sqrt{\sigma^2(\boldsymbol{x}_*'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_* + 1)}} \sim N(0,1). \tag{9}$$

Recall that $s^2 = \boldsymbol{e}'\boldsymbol{e}/(N - (K+1))$, and that

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

and note that this is independent of the ratio (9), because s^2 and $\hat{\boldsymbol{\beta}}_{\text{LS}}$ are independent and ε_* is independent of ε from the sample. Therefore the ratio

$$\frac{\frac{\boldsymbol{x}_*'\hat{\boldsymbol{\beta}}_{\text{LS}} - y_*}{\sqrt{\sigma^2(\boldsymbol{x}_*'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_* + 1)}}}{\frac{(N-(K+1))s^2}{\sigma^2}} = \frac{\boldsymbol{x}_*'\hat{\boldsymbol{\beta}}_{\text{LS}} - y_*}{s\sqrt{\boldsymbol{x}_*'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_* + 1}} \sim t(N - (K+1)).$$

Thus a $(1-\alpha)$ confidence interval of y_* is

$$\Big[\hat{y}_* - t_{\alpha/2,N-(K+1)}\, s\,\sqrt{\boldsymbol{x}'_*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_* + 1},\ \ \hat{y}_* + t_{\alpha/2,N-(K+1)}\, s\,\sqrt{\boldsymbol{x}'_*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_* + 1}\Big].$$

25.9 ⋆ Measurement error

See Johnston [2, pp. 281–289]

When the variables $X_1, \dots, X_K$ are measured with error, then the previous results need to be modified. It is no longer true that $\hat{\boldsymbol{\beta}}_{\text{LS}}$ is consistent and unbiased.

Suppose that the observed data $\tilde{X}$ is actually

$$\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{V},$$

where $\boldsymbol{V}$ is a matrix of measurement errors. We estimate the model

$$\boldsymbol{y} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{u}, \tag{10}$$

when in fact the correct model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

or

$$\boldsymbol{y} = (\tilde{\boldsymbol{X}} - \boldsymbol{V})\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + (\boldsymbol{\varepsilon} - \boldsymbol{V}\boldsymbol{\beta}).$$

The LS estimate derived from (10) is

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{\varepsilon} - \boldsymbol{V}\boldsymbol{\beta})$$

The expectation is

$$\boldsymbol{E}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{E}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}\boldsymbol{\beta}$$

One way to correct for this problem is to find a vector of random variables $\boldsymbol{Z}$, called **instrumental variables**, which is uncorrelated with the errors $\boldsymbol{V}$ and with $\boldsymbol{\varepsilon}$, then we can use these to estimate X and then use the estimated X's to estimate $\boldsymbol{\beta}$. See, e.g., Johnston [2, pp. 282–283]

25.10 ⋆ ANOVA is Regression in disguise

Larsen–Marx [4]: Chapter 12

According to Wikipedia, the “law of the instrument” was formulated by Abraham Kaplan [3, p. 28]:

> Give a small boy a hammer, and he will find that everything he encounters needs pounding.

This was reformulated by Abraham Maslow [5, p. 15]:

> ... it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.

Perhaps that is why I think of **ANalysis Of VAriance**, or **ANOVA** for short, as a special case of multiple linear regression (which it is). However, the language of those who use ANOVA tends to be different from that of those who use linear regression, and part of what we will do today is learn that language.

25.11 Some Terminology

According to Larsen–Marx [4, pp. 431–432],

> The word *factor* is used to denote any treatment or therapy “applied to” the subjects being measured or any relevant feature (age, sex, ethnicity, etc.) “characteristic” of those subjects. Different versions, extents, or aspects, of a factor are referred to as *levels*. ... Sometimes subjects or environments share certain characteristics that affect the way levels of a factor respond, yet those characteristics are of no intrinsic interest to the experimenter. Any such set of conditions or subjects is called a *block*.

I would probably call a factor an explanatory variable, and the level its value. Blocks would correspond to additional variables. The thing actually being measured is called the **response**, or what I would call the dependent, or left-hand side, variable.

Analysis of variance is designed to address data from a completely randomized single factor design with two or more levels.

25.12 The model

In terms of model equations (which is the most transparent way to look at models), we have

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \qquad (i = 1, \dots, n_j;\ j = 1, \dots, k) \tag{11}$$

where there are k different factor levels, we have n_j observations of the response Y at level j, and Y_{ij} is the i^{th} measurement of the response at level j, $j = 1, \dots, k$. Let $n = n_1 + \cdots + n_k$ be the total number of observations

The random variables ε_{ij} are assumed to be independent, have common mean zero and common variance σ^2. Thus μ_j is just the expected value of the response at level j.

The object of the analysis is estimate the vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)$ and test hypotheses regarding it. The most common hypothesis is

$$H_0 \colon \mu_1 = \cdots = \mu_k,$$

against the alternative

$$H_1 \colon \text{not all the } \mu_j\text{'s are equal.}$$

To see why this framework is subsumed by regression, stack the vectors of observations $\boldsymbol{y}$ and errors ε_{ij}, let X_j be a **dummy variable** or **indicator** for the j^{th} level, and create an $n \times k$ matrix $\boldsymbol{X}$ and rewrite the system (11) as

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{n_1 1} \\ y_{12} \\ \vdots \\ y_{n_2 2} \\ \vdots \\ \vdots \\ y_{1k} \\ \vdots \\ y_{n_k k} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2 2} \\ \vdots \\ \vdots \\ \varepsilon_{1k} \\ \vdots \\ \varepsilon_{n_k k} \end{bmatrix}.$$

This is a regression model with no constant term. If were to add a constant term it would be the sum of the of the other columns so $\boldsymbol{X}'\boldsymbol{X}$ would be singular. As it is, $\boldsymbol{X}'\boldsymbol{X}$ is the diagonal matrix with n_j on the diagonal (and therefore nonsingular) and it is easy to see that the LS estimated coefficients are the sample means for each level:

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & & \vdots & & \ddots & & 0 & \cdots & 0 \\ \vdots & & \vdots & & & \ddots & 0 & & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & 0 \\ 0 & & & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & & & 0 & 1 \end{bmatrix} = \begin{bmatrix} n_1 & 0 & \cdots & \cdots & 0 \\ 0 & n_2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \\ 0 & \cdots & \cdots & 0 & n_k \end{bmatrix}$$

and

$$\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & & \vdots & & \ddots & & 0 & \cdots & 0 \\ \vdots & & \vdots & & & \ddots & 0 & & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_{11} \\ \vdots \\ y_{n_1 1} \\ y_{12} \\ \vdots \\ y_{n_2 2} \\ \vdots \\ \vdots \\ y_{1k} \\ \vdots \\ y_{n_k k} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n_1} y_{i1} \\ \sum_{i=1}^{n_2} y_{i2} \\ \vdots \\ \sum_{i=1}^{n_k} y_{ik} \end{bmatrix}$$

so

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \dfrac{\sum_{i=1}^{n_1} y_{i1}}{n_1} \\ \dfrac{\sum_{i=1}^{n_2} y_{i2}}{n_2} \\ \vdots \\ \dfrac{\sum_{i=1}^{n_k} y_{ik}}{n_k} \end{bmatrix}$$

In Lecture 25 we saw how to use an F-statistic to test linear restrictions on the coefficient vector $(\hat{\mu}_1, \dots, \hat{\mu}_k)$. So we could stop here.

One reason for not stopping here is that the F-statistic is more transparent and intuitive in the traditional analysis. The other reason is that ANOVA analysis is usually presented in particular format, so you should learn to understand that. Historically, the traditional analysis may have developed because it is less computationally intensive than LS. (The general matrix inversion to obtain $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is replaced by k division operations, since $\boldsymbol{X}'\boldsymbol{X}$ is a diagonal matrix.)

25.13 ANOVA terminology

Time for some more terminology:

- y_{ij} is the response of the i^{th} observation at level j.
- $T_{\bullet j} = \sum_{i=1}^{n_j} y_{ij}$ is the **response total** at level j.
- $\bar{Y}_{\bullet j} = \dfrac{T_{\bullet j}}{n_j}$ is the sample mean at level j.
- $T_{\bullet\bullet} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij} = \sum_{j=1}^{n} T_{\bullet j}$ is the sample overall total response.
- $\bar{Y}_{\bullet\bullet} = \dfrac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij} = \dfrac{1}{n}\sum_{j=1}^{k} T_{\bullet j}$ is the sample overall average response.
- The **treatment sum of squares** SSTR is defined to be

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2$$

- $\mu = \sum_{j=1}^{k} \dfrac{n_j}{n}\mu_j$ is the overall average of the (unobserved) μ_j's.

Now the MLE of $\mu_j = \bar{Y}_{\bullet j}$, so the MLE of μ is $\bar{Y}_{\bullet\bullet}$.
It is not hard to show that (see Larsen–Marx [4, Theorem 12.2.1, p. 598–599])

$$\boldsymbol{E}(\text{SSTR}) = (k-1)\sigma^2 + \sum_{j=1}^{k} n_j(\mu_j - \mu)^2. \tag{12}$$

Now σ^2 is not generally known, so we must estimate it. Start by defining

$$s_j^2 = \frac{\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_{\bullet j})^2}{n_j - 1},$$

and aggregating

$$\text{SSE} = \sum_{j=1}^{k}(n_j-1)s_j^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{\bullet j})^2, \tag{13}$$

which is called the **error sum of squares**. The important fact about these is:

25.13.1 Theorem *(Larsen–Marx [4, Theorem 12.2.3, p. 600])*

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-k)$$

and SSE and SSTR are stochastically independent.

There is one last sum of squares of interest. If we ignore the levels (treatments) the variance about μ can be estimated by the **total sum of squares** SSTOT:

$$\text{SSTOT} = \sum_{j=1}^{k}(n_j-1)s_j^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{\bullet\bullet})^2$$

Tedious arithmetic (Larsen–Marx [4, Theorem 12.2.4, pp. 600–601]) shows that

$$\text{SSTOT} = \text{SSTR} + \text{SSE}$$

25.14 Testing equality of means

Note that if the null hypothesis

$$H_0\colon \mu_1 = \cdots = \mu_k$$

is true, then

$$\boldsymbol{E}\,\text{SSTR} = (k-1)\sigma^2 (\text{by (12)}) \qquad \text{and} \qquad \boldsymbol{E}\,\text{SSE} = (n-k)\sigma^2 (\text{by (13)}).$$

Therefore we should expect

$$F = \frac{\text{SSTR}/(k-1)}{\text{SSE}/(n-k)}$$

to be close to one. Otherwise it will be significantly greater than one. So the appropriate test is a one-sided F-test,

> Reject the null hypothesis H_0 if $F \geqslant F_{1-\alpha,k-1,n-k}$.

25.15 ANOVA tables

The traditional way to present ANOVA data is in the form of a table like this:

Source	df	SS	MS	F	P
Treatment	k-1	SSTR	$\dfrac{\text{SSTR}}{k-1}$	$\dfrac{\text{SSTR}/(k-1)}{\text{SSE}/(n-k)}$	$F \leqslant F_{k-1,n-k}$
Error	n-k	SSE	$\dfrac{\text{SSE}}{n-k}$		
Total	n-1	SSTOT			

Two more terms: the **mean square for treatments** is

$$\text{MSTR} = \frac{\text{SSTR}}{k-1}$$

the **mean square for errors** is

$$\text{MSE} = \frac{\text{SSE}}{n-k}.$$

25.16 Contrasts

A linear combination of the form

$$C = \boldsymbol{w}'\boldsymbol{\mu},$$

where $\mathbf{1}'\boldsymbol{w} = 0$ is called a **contrast**. A typical contrast uses a vector of the form

$$\boldsymbol{w} = (0,\dots,0,\underset{j}{1},0,\dots,0,\underset{j'}{-1},0,\dots,0,$$

so

$$C = \boldsymbol{w}'\boldsymbol{\mu} = \mu_j - \mu_{j'}.$$

Then the hypothesis $H_0\colon C = 0$ amounts to $H_0\colon \mu_j = \mu_{j'}$. This is why it is called a contrast.

To test a hypothesis that $C = 0$, we weight the sample means

$$\hat{C} = \sum_{j=1}^{k} w_j \bar{y}_{\bullet j}.$$

Then

$$\boldsymbol{E}\,\hat{C} = C \qquad \boldsymbol{Var}\,\hat{C} = \sigma^2 \sum_{j=1}^{k} \frac{w_j^2}{n_j}.$$

Define

$$\mathrm{SS_C} = \frac{\hat{C}^2}{\sum_{j=1}^{k} \frac{w_j^2}{n_j}}.$$

25.16.1 Theorem *Larsen–Marx [4, Theorem 12.4.1, p.614] The test statistic*

$$F = \frac{SS_C}{SSE/(n-k)}$$

has an F-distribution with $(1, n-k)$ degrees of freedom. The null hypothesis

$$H_0 \colon \boldsymbol{w}'\boldsymbol{\mu} = 0$$

should be rejected if $F \geqslant F_{1-\alpha,1,n-k}$.

Bibliography

[1] K. C. Border. Notes on maximization with more than one variable. http://www.its.caltech.edu/~kcborder/Notes/Max2.pdf

[2] J. Johnston. 1972. *Econometric methods*, 2d. ed. New York: McGraw–Hill.

[3] A. Kaplan. 1964. *The conduct of inquiry: Methodology for behavioral science*. San Francisco: Chandler Publishing Co.

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

[5] A. H. Maslow. 1966. *The psychology of science: A reconnaissance*. NY: Harper & Row.

[6] H. Theil. 1971. *Principles of econometrics*. New York: Wiley.