Ma 3/103: Lecture 25
Linear Regression II:
Hypothesis Testing and ANOVA

March 6, 2017

1. OLS estimator
2. Restricted regression
3. Errors in variables
4. ANOVA
5. The F test in an ANOVA framework
6. Contrasts

Standard Linear Model

$$y = X\beta + \varepsilon,$$

where

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

OLS estimation

With N observations on $X_1, \ldots, X_K, Y$, let X be the $N \times K$ matrix of regressors, and y be the $N \times 1$ vector of observations on the response Y. Then if X has rank K, the OLS estimator $\hat{\beta}_{\text{OLS}}$ of the parameter vector β is given by

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

It is obtained by orthogonally projecting y onto the column space of X.

Regression and Correlation

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

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

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

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

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

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

Given pairs (x_t, y_t), $t = 1, \dots, N$, of observations, define the sample **correlation coefficient** r by

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

which is the sample analog of the correlation. It is also known as the **Pearson product-moment correlation coefficient**.

Consider the **centered variables**

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

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

But by construction,

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

Now look at the formula for the correlation coefficient. It can be rewritten as

$$r = \frac{\sum_{t=1}^{N} \tilde{x}_t \tilde{y}_t}{s_x s_y} = \hat{\beta}_1 \frac{s_x}{s_y},$$

where $s_x = \sqrt{\sum_t (x_t - \bar{x})^2} = \sqrt{\sum_t \tilde{x}_t^2}$ and $s_y = \sqrt{\sum_t \tilde{y}_t^2}$. 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 the slope is not identifiable.)

Testing for serial correlation

Regress e_t on e_{t-1}:

$$e_t = \beta_0 + \beta_1 e_{t-1}.$$

Test $\hat{\beta}_1 = 0$.

Testing linear restrictions on β

To test q simultaneous restrictions, let

$$H_0 \colon a = A\beta,$$

where A is a $q \times K$ matrix with rank q.

Theorem

Under the null hypothesis, the test statistic

$$F = \frac{1}{qs^2}(a - A\hat{\beta}_{OLS})'\left[A(X'X)^{-1}A'\right](a - A\hat{\beta}_{OLS})$$

has an F-distribution with $(q, N - K)$ degrees of freedom.

The F-test of the regression

Many software packages, including R, compute for you something called the "F-statistic for the regression."

The F-statistic for the regression tests the null hypothesis that all the coefficients on the non-constant terms are zero,

$$H_0 \colon \beta_2 = \beta_3 = \cdots = \beta_K = 0.$$

(If you have a constant term, it is usually X_1 in our terminology.)

Coefficient of Multiple Correlation R

$$\|y\|^2 = y'y = \hat{\beta}'_{\text{OLS}} X'X \hat{\beta}_{\text{OLS}} + e'e + 2\hat{\beta}'_{\text{OLS}} \underbrace{X'e}_{=0}$$

The **coefficient of multiple correlation** R is a measure of the fraction of $y'y$ "explained" by the regressors. Specifically,

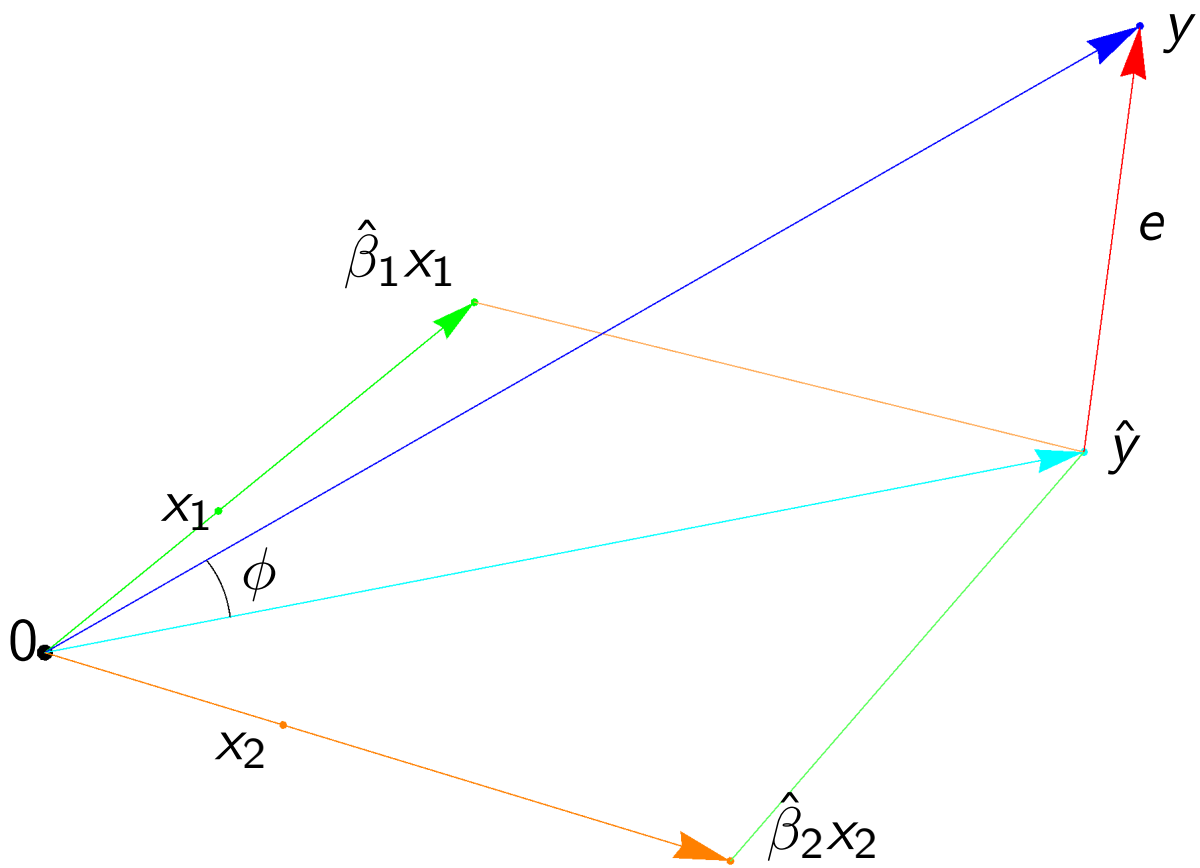
$$1 - R^2 = \frac{e'e}{y'y}, \quad \text{or } R^2 = \frac{\hat{y}'\hat{y}}{y'y} = \frac{\hat{\beta}'_{\text{OLS}} X'X \hat{\beta}_{\text{OLS}}}{y'y}.$$

The Pythagorean Theorem implies $y'y = e'e + \hat{y}'\hat{y}$, so

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

Geometry of R^2

$R = \sqrt{R^2}$ is the cosine of the angle ϕ between y and $\hat{y} = X\hat{\beta}_{\text{OLS}}$.



Adjusted R^2

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

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

or

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

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

What is a "good" value for R^2?

Prediction intervals

Let

$$y_* = x_*'\beta + \varepsilon_*, \qquad \hat{y}_* = x_*'\hat{\beta}_{\text{OLS}}.$$

But what is the confidence interval for y_*?

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

Therefore

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

Confidence intervals

Under the normality hypothesis,

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

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

$$\left[\hat{y}_* - t_{\frac{\alpha}{2},N-K}\, s\sqrt{x_*'(X'X)^{-1}x_*+1},\ \hat{y}_* + t_{\frac{\alpha}{2},N-K}\, s\sqrt{x_*'(X'X)^{-1}x_*+1}\right].$$

The Lagrange Multiplier Theorem

If

$$x^* \text{ minimizes } f(x) \quad \text{subject to } g_i(x) = 0 \ (i = 1, \dots, m),$$

and if

$$\nabla g_1(x^*), \dots, \nabla g_m(x^*) \text{ are linearly independent,}$$

then there exist **Lagrange multipliers** λ_i^*, $i = 1, \dots, m$ such that

$$\nabla f(x^*) + \lambda_1^* \nabla g_1(x^*) + \cdots + \lambda_m^* \nabla g_m(x^*) = 0.$$

Restricted OLS

To minimize $(y - Xb)'(y - Xb)$ subject to the constraint $Ab = a$ (where A is $q \times k$), the LMT tells us to form the Lagrangean

$$(y - Xb)'(y - Xb) + \boldsymbol{\lambda}'(Ab - a)$$

and solve the FOC

$$-2X'y + 2X'Xb + A'\lambda = 0. \tag{2}$$

Solving the FOC

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

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

so, solving for $\boldsymbol{\lambda}$

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

Substitute this into (2) to get

$$-X'y + X'Xb - A'[A(X'X)^{-1}A']^{-1}[a - A(X'X)^{-1}X'y]$$

which after premultiplying by $(X'X)^{-1}$, with some work simplifies to

$$b^* = \hat{\beta}_{\text{OLS}} + (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(a - A\hat{\beta}_{\text{OLS}}).$$

Restricted residuals

Let

$$e_r = y - Xb^*.$$

be the vector of residuals from the restricted regression.
It can be shown that

$$e_r{}'e_r = e_u{}'e_u + (a - A\hat{\beta}_{\text{OLS}})'[A(X'X)^{-1}A']^{-1}(a - A\hat{\beta}_{\text{OLS}}),$$

where $e_u{}'e_u$ is the sum of squares from the unrestricted OLS regression.
Thus

$$e_r{}'e_r - e_u{}'e_u = (a - A\hat{\beta}_{\text{OLS}})'[A(X'X)^{-1}A']^{-1}(a - A\hat{\beta}_{\text{OLS}})$$

is a quadratic form in the q variables $a - A\hat{\beta}_{\text{OLS}}$.

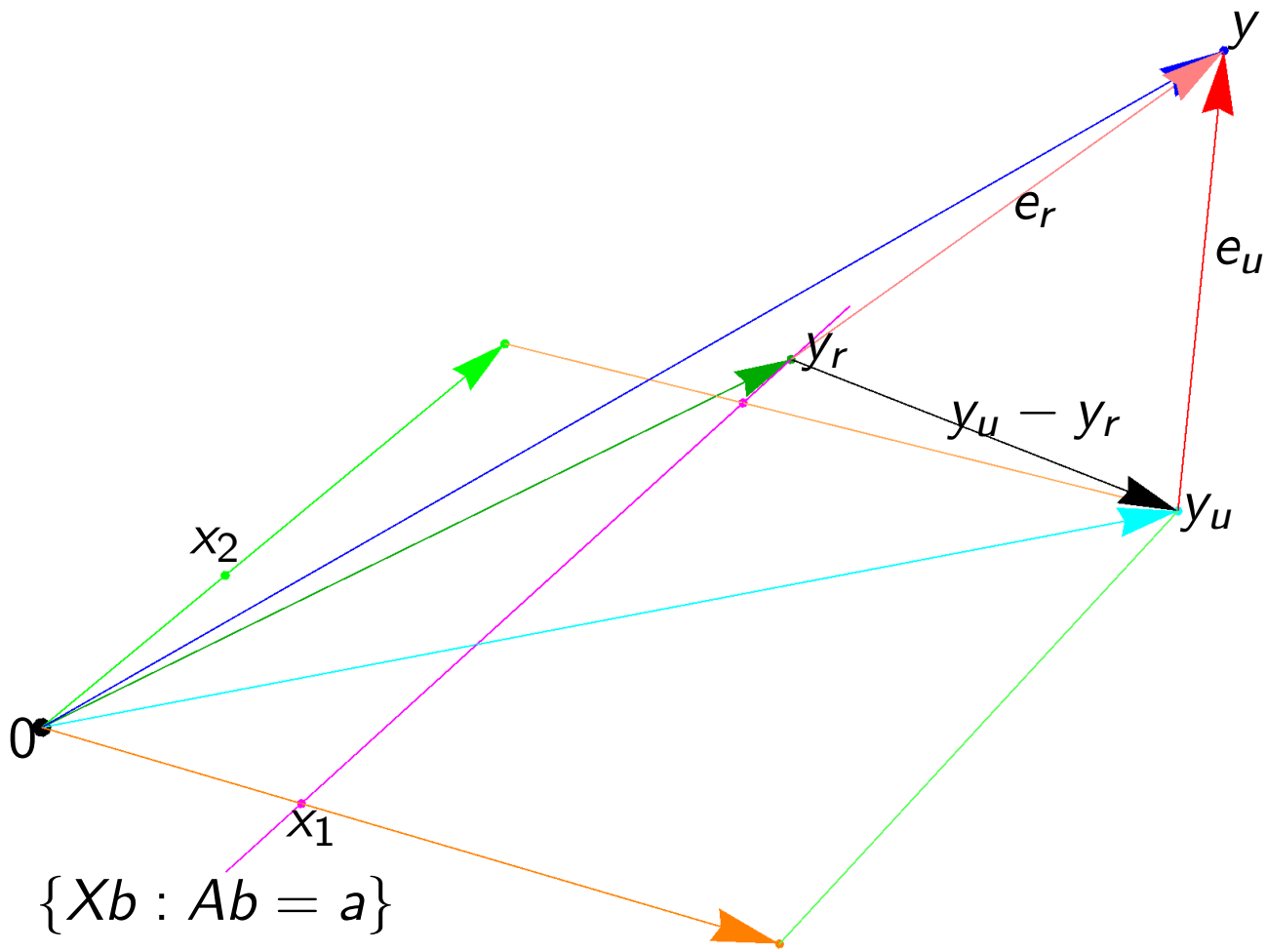
Testing a linear restriction

$$H_0 \colon A\beta = a$$

Let e_u and e_r be the vector of residuals from the unrestricted and restricted regressions. Then under the null hypothesis,

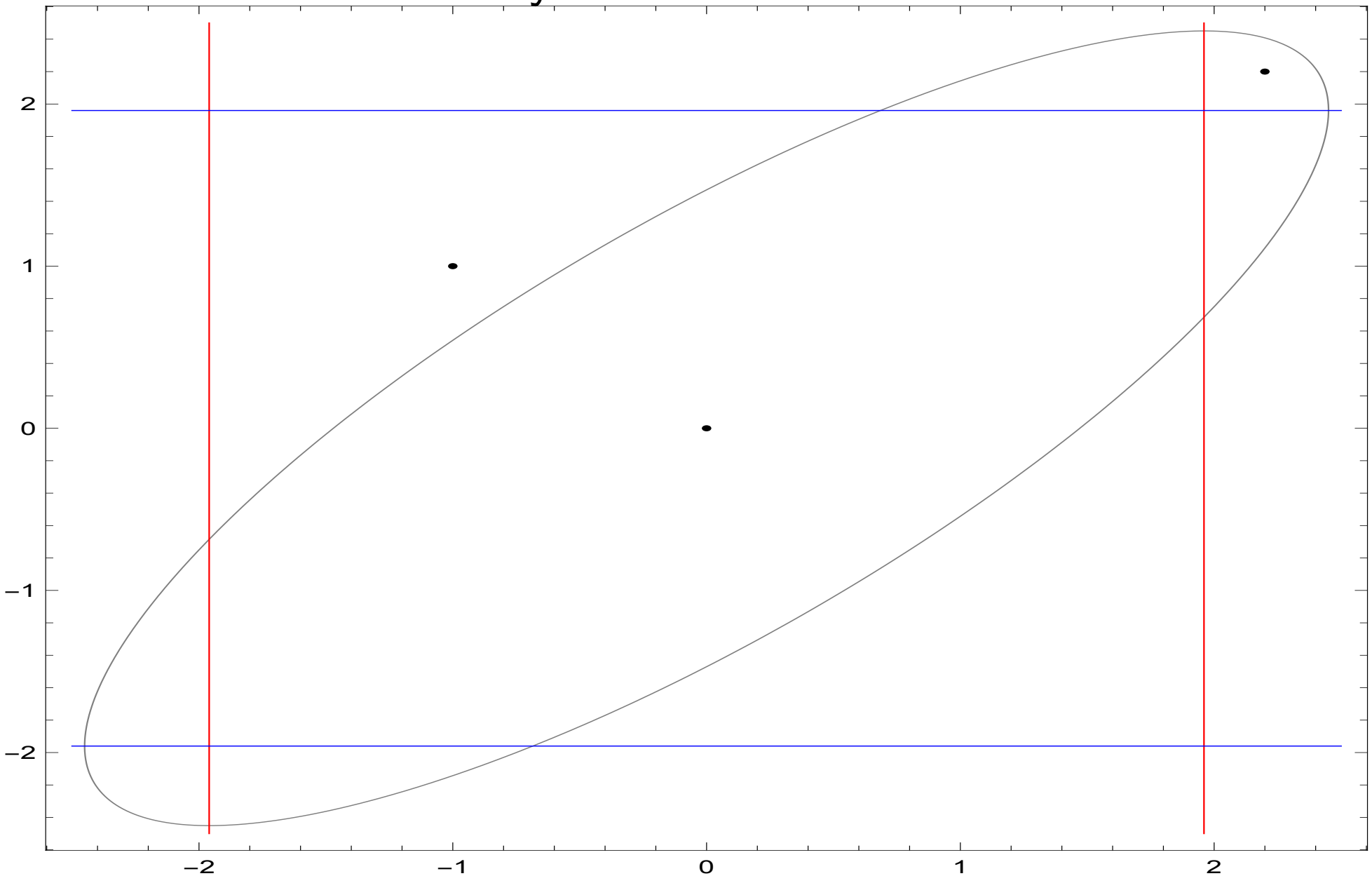
$$F = \frac{\frac{e_r{}' e_r - e_u{}' e_u}{q}}{\frac{e_u{}' e_u}{N-K}}$$

has an F-distribution with $(q, N-K)$ degrees of freedom.
The null hypothesis should be rejected if $F \geqslant F_{1-\alpha,q,N-K}$.



Restricted regression with restriction $\beta_1 = 1$. The points y_r, y_u, y form a right triangle with hypotenuse $\overline{y_r y}$.

F-tests and t-tests may seem to conflict!



Measurement error

True model:

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

but observe

$$X = \tilde{X} + V.$$

So the estimated model is

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

The OLS estimate derived from (3) is

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

The expectation is

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

which is not, in general, unbiased, nor consistent.

Some jargon

According to Larsen and Marx, 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*.*

ANOVA

ANOVA is an acronym for **AN**alysis **O**f **VA**riance

Model equations

One factor with k levels.
Y_{ij} is the i^{th} measurement of the response at factor level j.
n_j observations at level j.

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

$n = n_1 + \cdots + n_k$ is the total number of observations.
ε_{ij} are assumed to be independent, have common mean zero and common variance σ^2.
μ_j is just the expected value of the response at level j.

ANOVA is a special case of the Standard Linear Model

X_j is a **dummy variable** or **indicator** for the j^{th} level.

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

OLS and ANOVA

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

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

$$X'y = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & & \vdots & & \ddots & & 0 & \cdots & 0 \\ \vdots & & \vdots & & & \ddots & 0 & & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_{11} \\ \vdots \\ y_{n_1 1} \\ y_{12} \\ \vdots \\ y_{n_2 2} \\ \vdots \\ \vdots \\ y_{1k} \\ \vdots \\ y_{n_k k} \end{bmatrix}$$

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

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

Hypothesis testing in ANOVA

The most common hypothesis is

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

against the alternative

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

A little more jargon

- 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 fundamental identity

For any list $x_1, \ldots, x_n$,

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n\bar{x}^2$$

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \bar{x})^2 &= \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2.
\end{aligned}$$

- The **treatment sum of squares** SSTR is defined to be

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

It is not hard to show using the fundamental identity and other tricks that (see L&M Theorem 12.2.1, p. 598–599)

$$\boldsymbol{E}(\text{SSTR}) = (k-1)\sigma^2 + \sum_{j=1}^{k} n_j(\mu_j - \mu)^2, \tag{4}$$

where $\mu = \sum_{j=1}^{k} \frac{n_j}{n} \mu_j$ is the overall average of the (unobserved) μ_js.

That is, a large value of SSTR relative to $(k-1)\sigma^2$ indicates that the null hypothesis $H_0 \colon \mu_1 = \cdots = \mu_k = \mu$ should be rejected.

Estimating σ^2

Start by defining

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

and aggregating

$$\text{SSE} = \sum_{j=1}^{k}(n_j - 1)s_j^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{\bullet j})^2, \tag{5}$$

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

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

and SSE and SSTR are stochastically independent. (L&M, Theorem 12.2.3, p. 600)

Under the null $H_0\colon \mu_1 = \cdots = \mu_k = \mu$,

$$\frac{\text{SSTR}}{\sigma^2} \sim \chi^2(k-1)$$

Therefore, under the null,

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

The F-test

At the α-level of significance, reject $H_0\colon \mu_1 = \cdots = \mu_k$ if

$$\frac{\text{SSTR}/(k-1)}{\text{SSE}/(n-k)} \geqslant F_{1-\alpha,k-1,n-k}.$$

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

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

Contrasts

A linear combination of the form

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

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

$$w = (0, \ldots, 0, \underset{j}{1}, 0, \ldots, 0, \underset{j'}{-1}, 0 \ldots, 0),$$

so

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

Then the hypothesis $H_0\colon C = 0$ amounts to $H_0\colon \mu_j = \mu_{j'}$. This is probably 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

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

F test of a contrast

The test statistic

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

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

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

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