Ma 3/103: Lecture 24
Linear Regression I: Estimation

March 3, 2017

Regression analysis

Estimate and test

$$\boldsymbol{E}(Y \mid \boldsymbol{X}) = f(\boldsymbol{X}).$$

f is the **regression function**; components of $\boldsymbol{X} = (X_1, \ldots, X_K)$ are **regressors**.

The standard linear model

$$Y = X\beta + \varepsilon$$

or

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

The linear model is more general than you might think

- 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.

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

- Newton's Law of Gravity:

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

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

- Polynomials:

$$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_K x^K$$

- Geometric means:

$$y = b_0 x_1^{b_1} x_2^{b_2} \cdots x_K^{b_K} \quad \Longleftrightarrow \quad \ln y = \ln b_0 + b_1 \ln x_1 + \cdots + b_K \ln x_K$$

- "Dummy variables," or indicators: e.g.,

$$X_1 = \begin{cases} 1 & \text{Honda} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{Kawasaki} \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$X_\ell = \begin{cases} 1 & \text{Ducati} \\ 0 & \text{otherwise} \end{cases}$$

"Variates"

The variates X_k may be fixed constants chosen by an experimenter or they may be random variables themselves. They are called **regressors**.

Almost always a constant "variate" is included.

Data

N **observations** of the values $x_1, \ldots, x_K$ and y.

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

where the ε_ts are unobserved errors. In matrix form:

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

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

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

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} \text{ is a } K \times 1 \text{ column vector,}$$

and

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

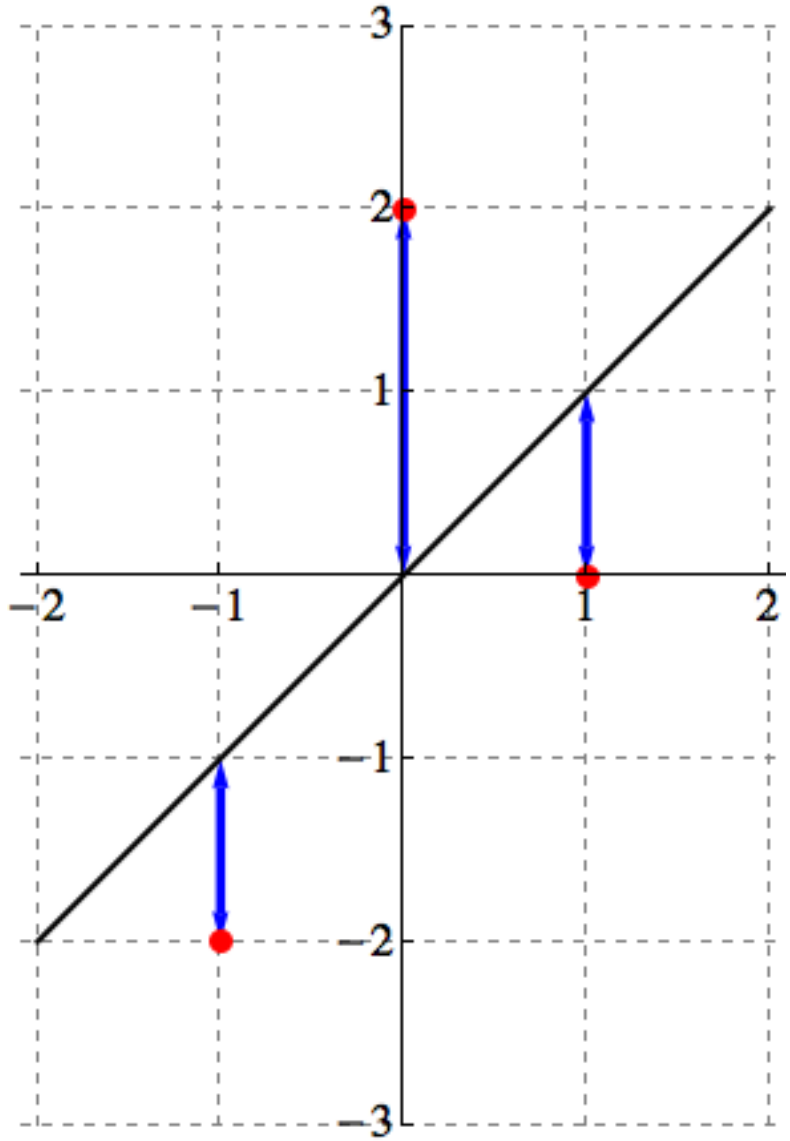
The estimation problem

The problem is to estimate $(\beta_1, \ldots, \beta_K)$.
Statistical assumptions of the standard model:

$$\boldsymbol{E}(\varepsilon|X) = 0,$$
$$\boldsymbol{Var}(\varepsilon|X) = \boldsymbol{E}(\varepsilon\varepsilon' \mid X) = \sigma^2 I_{N\times N}.$$

This last assumption is known as **homoskedasticity**.

The Least Squares approach



Sum of squared residuals

Vector of **residuals** as a function of b is

$$y - Xb$$

The **sum of squared residuals (SSR)** is

$$(y - Xb)'(y - Xb).$$

Expanding yields

$$\text{SSR}(b) = y'y - 2y'Xb + b'X'Xb.$$

which is a convex quadratic function in the components of b.

Minimizing the sum of squared residuals

By convexity, the minimum occurs whenever the gradient equals zero. The gradient of this function is

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

Thus the minimizer $\hat{\beta}_{\mathrm{OLS}}$ satisfies the first-order condition $\nabla \mathrm{SSR}(\hat{\beta}_{\mathrm{OLS}}) = 0$:

$$X'y = X'X\hat{\beta}_{\mathrm{OLS}}.$$

This matrix equation is known as the **normal equation** for $\hat{\beta}_{\mathrm{OLS}}$.

Least Squares Estimator

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

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

minimizes the sum of squared residuals.

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

The singular case

What if $X'X$ is singular? Then

$$a_1X^1 + \cdots + a_KX^K = 0,$$

where not all a_k are zero.
Then

$$y = \beta_1X^1 + \cdots + \beta_KX^K + \varepsilon + c\underbrace{(a_1X^1 + \cdots + a_KX^K)}_{=0}$$

$$= (\beta_1 + ca_1)X^1 + \cdots + (\beta_K + ca_k)X^K + \varepsilon$$

for any value of c.
Whenever a_k is nonzero, the coefficient on X^k 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*.

Properties

- $\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$.
- This is a random vector.
- Set $e = y - X\hat{\beta}_{\text{OLS}}$,
- the vector e of **residuals** is orthogonal to each k^{th} column vector of the values of the **regressor** X_k.
- $X'e = 0$, since

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

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

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

Proof:

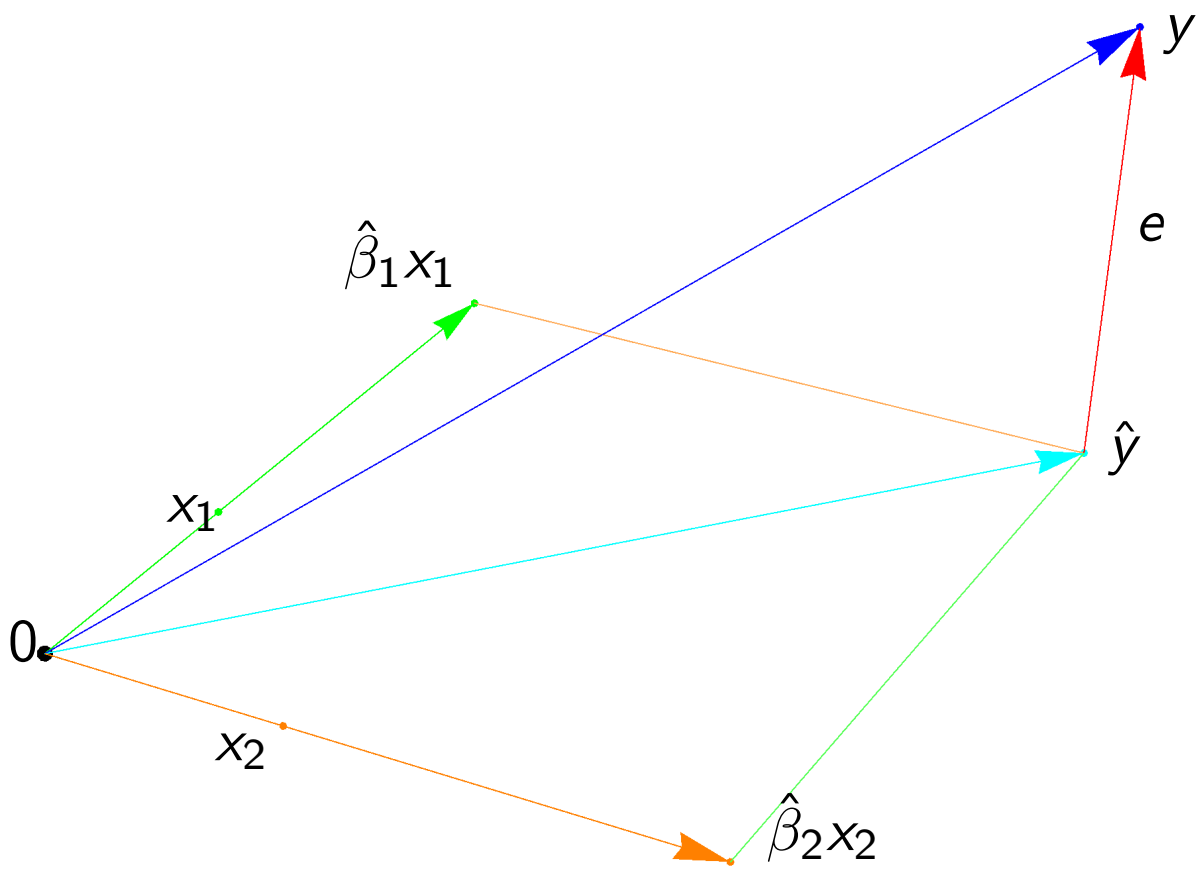
$$y = X\hat{\beta}_{\text{OLS}} + e,$$

so

$$\mathbf{1}'y = \mathbf{1}'X\hat{\beta}_{\text{OLS}} + \mathbf{1}'e,$$

where $\mathbf{1}$ is a N-vector of ones. Since it is one of the regressors, $\mathbf{1}'e = 0$. Dividing by N gives $\bar{y} = \bar{x}_1\hat{\beta}_1 + \cdots + \bar{x}_K\hat{\beta}_K$.

The Geometry of LSE



OLS and MLE

When the error vector ε has a multivariate normal distribution $N(0, \sigma^2 I)$ distribution,
then the OLS estimator of β is also the Maximum Likelihood Estimator.

MLE of β

The density of $\varepsilon = y - X\beta$ is the multivariate normal density $N(0, \sigma^2 I)$

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

$$\left(\frac{1}{\sqrt{2\pi}}\right)^N \left(\frac{1}{(\sigma^2)^N}\right)^{\frac{1}{2}}\; e^{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)}$$

Taking logs, we find the log likelihood function is

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

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

MLE of σ^2

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

$$-\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2}(y - X\beta)'(y - X\beta)\frac{1}{(\sigma^2)^2} = 0.$$

Then multiply by $2(\sigma^2)^2$ to get

$$-N\sigma^2 + (y - X\beta)'(y - X\beta) = 0,$$

so

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

where

$$e = y - X\hat{\beta}.$$

$\hat{\beta}_{\text{OLS}}$ is unbiased

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

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

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

$\hat{\beta}_{\text{OLS}}$ is **unbiased**,

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

Variance-covariance matrix $\hat{\beta}_{\text{OLS}}$

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

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

Gauss–Markov Theorem

In the standard linear model, if X has rank K, then the OLS estimator $\hat{\beta}_{\text{OLS}}$ is the **Best Linear Unbiased Estimate (BLUE)** of β in the following sense.

Given any other estimator b of β which is linear in y and which satisfies $\boldsymbol{E}\, b = \beta$ for any possible value of β, then

$$\boldsymbol{Var}\, b = \boldsymbol{Var}\, \hat{\beta}_{\text{OLS}} + P, \qquad \text{where } P \text{ is positive semidefinite.}$$

This implies that for any vector w of weights

$$\boldsymbol{Var}\, w'b \geqslant \boldsymbol{Var}\, w'\hat{\beta}_{\text{OLS}}.$$

Proof of Gauss–Markov

Let $b = Ay$. Define

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

Then

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

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

So in expectation,

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

Proof of Gauss–Markov, continued

Now b is unbiased if and only if $DX\beta = 0$ for all β. Therefore

$$DX = 0,$$

so (1) becomes

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

Proof of Gauss–Markov, continued

So for an unbiased linear estimator b,

$$\begin{aligned}
\boldsymbol{Var}\, b &= \boldsymbol{E}(b-\beta)(b-\beta)' \\
&= \left(D + (X'X)^{-1}X'\right)\boldsymbol{E}(\varepsilon\varepsilon')\left(D + (X'X)^{-1}X'\right)' \\
&= \sigma^2\left(D + (X'X)^{-1}X'\right)\left(D' + X(X'X)^{-1}\right) \\
&= \sigma^2\Big(DD' + \underbrace{DX}_{=0}(X'X)^{-1} + (X'X)^{-1}\underbrace{X'D'}_{=0}\Big) + (X'X)^{-1} \\
&= \sigma^2 DD' + \boldsymbol{Var}\,\hat{\beta}_{\text{OLS}}.
\end{aligned}$$

But $P = \sigma^2 DD'$ is positive semidefinite as

$$w'DD'w = (D'w)'(D'w) \geqslant 0.$$

q.e.d.

Estimating σ^2

$$e = My = M\varepsilon,$$

where

$$M = I - X(X'X)^{-1}X'.$$

$$e'e = \varepsilon' M' M\varepsilon = \varepsilon' M\varepsilon.$$

Since $\varepsilon' M\varepsilon$ is 1×1, 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 by the magic of linear algebra,

$$\begin{aligned} \boldsymbol{E}(e'e) &= \boldsymbol{E}(\varepsilon' M\varepsilon) \\ &= (N - K)\sigma^2 \end{aligned}$$

Estimating σ^2, continued

Define

$$s^2 = \frac{e'e}{N-K}, \qquad s = \sqrt{\frac{e'e}{N-K}}.$$

Theorem

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

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

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

Test statistics

If ε is jointly Normal, then for any K-vector w of weights,

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

so

$$\frac{w'(\hat{\beta}_{\text{OLS}} - \beta)}{s\sqrt{w'(X'X)^{-1}w}} \sim t(N-K). \tag{2}$$

Standard error of $\hat{\beta}_{k\text{OLS}}$

Special case, w is the k^{th} unit coordinate vector:

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

Since $\sigma^2(X'X)^{-1}_{kk} = \boldsymbol{Var}\,\hat{\beta}_{k\text{OLS}}$, we have that $s\sqrt{(X'X)^{-1}_{kk}}$ is the estimated standard deviation of $\hat{\beta}_{k\text{OLS}}$, and is called the **standard error** of $\hat{\beta}_{k\text{OLS}}$.

Confidence intervals for β_k

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

$$\left(\hat{\beta}_k - t_{\frac{\alpha}{2},N-K}\, s\sqrt{(X'X)^{-1}_{kk}},\ \hat{\beta}_k + t_{1-\frac{\alpha}{2},N-K}\, s\sqrt{(X'X)^{-1}_{kk}}\right)$$

Testing β_k

To test

$$H_0 \colon \beta_k = \beta_k^0 \quad \text{versus} \quad H_1 \colon \beta_k \neq \beta_k^0$$

Compute

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

We reject the null hypothesis if $|t| > t_{\frac{\alpha}{2},N-K}$.
For the null hypothesis

$$H_0 \colon \hat{\beta}_k = 0, \quad \text{we have} \quad t = \frac{\hat{\beta}_{k\text{OLS}}}{s\sqrt{(X'X)^{-1}_{kk}}}.$$

It is this value of t that statistical software reports as the "t-value" for β_k.