Precept 5: Simple OLS

Soc 500: Applied Social Statistics

Simone Zhang[1]

Princeton University

October 2016

[1] This draws material from Matt Blackwell.

Today's Agenda

- Basic matrix operations
- Review matrix notation for linear regression
  - Notation
  - OLS estimation
  - Variance-covariance matrix
  - R-square
- F-test
- Bootstrap

Matrix Notation

- $\mathbf{X}$ is the $n \times (K+1)$ design matrix of independent variables
- β be the $(K+1) \times 1$ column vector of coefficients.
- $\mathbf{X}\beta$ will be $n \times 1$:

- We can compactly write the linear model as the following:

$$\underset{(n\times1)}{\mathbf{y}} = \underset{(n\times1)}{\mathbf{X}\beta} + \underset{(n\times1)}{\mathbf{u}}$$

$$\underset{(n\times1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \underset{(n\times(K+1))}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \quad \underset{((K+1)\times1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

OLS Estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- What's the intuition here?
- "Numerator" $\mathbf{X}'\mathbf{y}$: is roughly composed of the covariances between the columns of $\mathbf{X}$ and $\mathbf{y}$
- "Denominator" $\mathbf{X}'\mathbf{X}$ is roughly composed of the sample variances and covariances of variables within $\mathbf{X}$
- Thus, we have something like:

$$\widehat{\boldsymbol{\beta}} \approx (\text{variance of } \mathbf{X})^{-1}(\text{covariance of } \mathbf{X} \text{ \& } \mathbf{y})$$

- This is a rough sketch and isn't strictly true, but it can provide intuition.

Variance-Covariance Matrix

- The homoskedasticity assumption is different: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- In order to investigate this, we need to know what the variance of a vector is.
- The variance of a vector is actually a matrix:

$$\text{var}[\mathbf{u}] = \Sigma_u = \begin{bmatrix} \text{var}(u_1) & \text{cov}(u_1, u_2) & \dots & \text{cov}(u_1, u_n) \\ \text{cov}(u_2, u_1) & \text{var}(u_2) & \dots & \text{cov}(u_2, u_n) \\ \vdots & & \ddots & \\ \text{cov}(u_n, u_1) & \text{cov}(u_n, u_2) & \dots & \text{var}(u_n) \end{bmatrix}$$

- This matrix is **symmetric** since $\text{cov}(u_i, u_j) = \text{cov}(u_j, u_i)$

Matrix Version of Homoskedasticity

- Once again: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- $\mathbf{I}_n$ is the $n \times n$ identity matrix
- Visually:

$$\text{var}[\mathbf{u}] = \sigma_u^2 \mathbf{I}_n = \begin{bmatrix} \sigma_u^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_u^2 & 0 & \ldots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \ldots & \sigma_u^2 \end{bmatrix}$$

- In less matrix notation:
  - $\text{var}(u_i) = \sigma_u^2$ for all i (constant variance)
  - $\text{cov}(u_i, u_j) = 0$ for all $i \neq j$ (implied by iid)

Sampling Variance for OLS Estimates

- Under assumptions 1-5, the sampling variance of the OLS estimator can be written in matrix form as the following:

$$\text{var}[\widehat{\beta}] = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$$

- This matrix looks like this:

	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\cdots$	$\widehat{\beta}_K$
$\widehat{\beta}_0$	$\text{var}[\widehat{\beta}_0]$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_1]$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_2]$	$\cdots$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_K]$
$\widehat{\beta}_1$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_1]$	$\text{var}[\widehat{\beta}_1]$	$\text{cov}[\widehat{\beta}_1, \widehat{\beta}_2]$	$\cdots$	$\text{cov}[\widehat{\beta}_1, \widehat{\beta}_K]$
$\widehat{\beta}_2$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_2]$	$\text{cov}[\widehat{\beta}_1, \widehat{\beta}_2]$	$\text{var}[\widehat{\beta}_2]$	$\cdots$	$\text{cov}[\widehat{\beta}_2, \widehat{\beta}_K]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\widehat{\beta}_K$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_K]$	$\text{cov}[\widehat{\beta}_K, \widehat{\beta}_1]$	$\text{cov}[\widehat{\beta}_K, \widehat{\beta}_2]$	$\cdots$	$\text{var}[\widehat{\beta}_K]$

Estimating Error Variance

Note that we never observe the true error variance, σ_u^2. We can estimate it with the following:

$$\widehat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$$

where n-(K+1) = residual degrees of freedom and

$$\widehat{\mathbf{u}}'\widehat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\widehat{\beta})'(\mathbf{y} - \mathbf{X}\widehat{\beta})$$

Prediction error

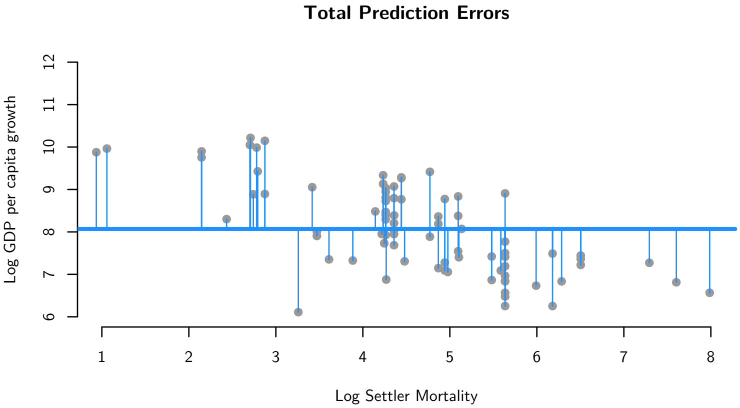
- Prediction errors without $\mathbf{X}$: best prediction is the mean, so our squared errors, or the **total sum of squares** (SS_{tot}) would be:

$$SS_{tot} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = (\mathbf{y} - \overline{y})'(\mathbf{y} - \overline{y})$$

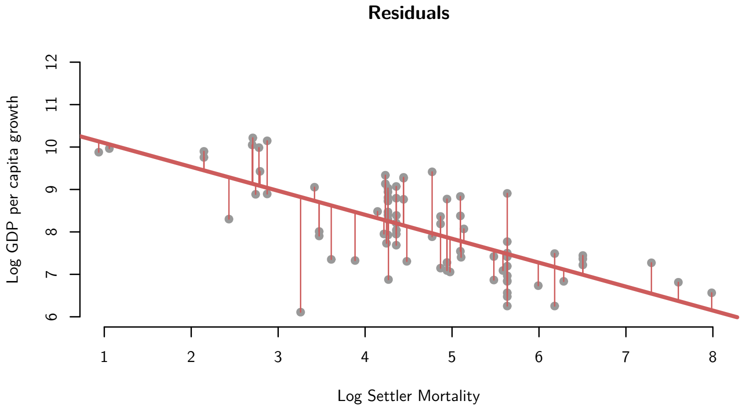
- Once we have estimated our model, we have new prediction errors, which are just the sum of the squared residuals or SS_{res}:

$$SS_{res} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \widehat{\mathbf{u}}'\widehat{\mathbf{u}}$$

Sum of Squares



Sum of Squares



R-square

- **Coefficient of determination** or R^2:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- This is the fraction of the total prediction error eliminated by providing information in $\mathbf{X}$.

F Test Procedure

The **F statistic** can be calculated by the following procedure:

1. Fit the **Unrestricted Model (UR)** which *does not* impose H_0
2. Fit the **Restricted Model (R)** which *does* impose H_0
3. From the two results, compute the **F Statistic:**

$$F_0 = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

where **SSR**=sum of squared residuals, **q**=number of restrictions, k=number of predictors in the unrestricted model, and n= # of observations.

Intuition:

$$\frac{\text{increase in prediction error}}{\text{original prediction error}}$$

The Bootstrap

We see a single sample that is a draw from a population:

- There's a true mean loan amount; we only observe one sample

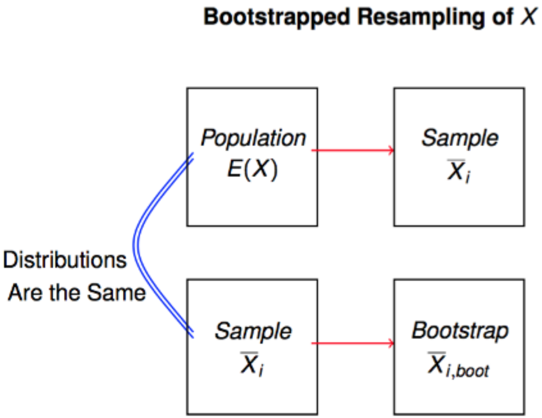
Since we cannot resample from the population, we resample from the sample!

Idea: Within a loop, generate a bootstrapped sample:

1. Sample from $\{1, 2, \ldots, N\}$ with replacement
2. Re-calculate the quantity of interest on each bootstrapped sample
3. Resampling from the sample *approximates* sampling again from the full population (giving us a sense of the sampling distribution)

(Thanks to Ted Enamorado for sharing slides on bootstrapping)

Bootstrap: Intuition



Simple Example with Sample Means

Let $X_i = \{3, 7, 9, 11, 150\}$

Bootstrapped Samples:

						$\bar{X}_{\text{boot}}$
$\text{X}_{\text{boot},1}$	3	3	9	11	3	5.8
$\text{X}_{\text{boot},1}$	7	150	11	7	11	37.2
$\text{X}_{\text{boot},1}$	11	9	9	7	3	7.8
$\vdots$						

Bootstrapped Standard Error

- Bootstrapped Standard Error
  sd($\bar{X}_{\mathrm{boot}}$)
- Bootstrapped Confidence Interval:
  Take the 2.5% and 97.5% quantiles of $\bar{X}_{\mathrm{boot}}$

Questions?