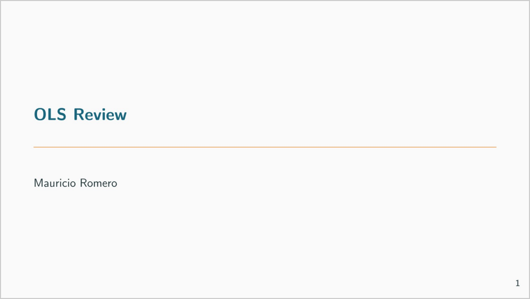
Lecture 7 - OLS review

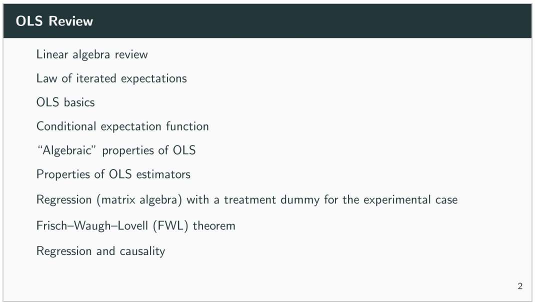
Thursday, September 09, 2021 3:58 PM

Lecture 7 - OLS review



OLS Review

Mauricio Romero



OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

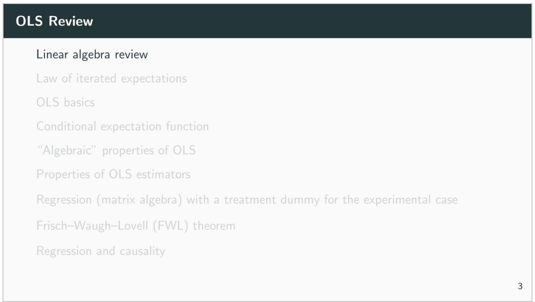
"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality



OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

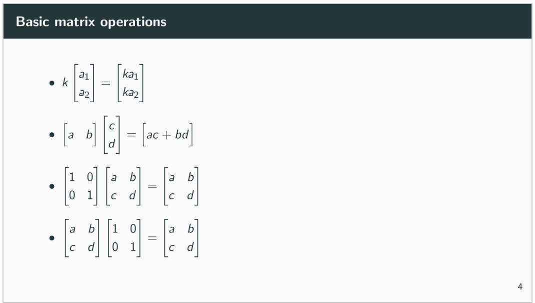
"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

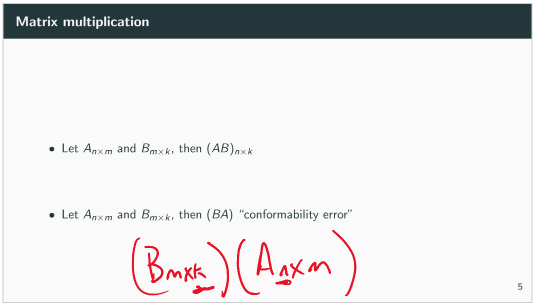
Frisch–Waugh–Lovell (FWL) theorem

Regression and causality



Basic matrix operations

- $k \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} ka_1 \\ ka_2 \end{bmatrix}$
- $\begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} ac + bd \end{bmatrix}$
- $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$
- $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$



Matrix multiplication

- Let $A_{n\times m}$ and $B_{m\times k}$, then $(AB)_{n\times k}$
- Let $A_{n\times m}$ and $B_{m\times k}$, then (BA) "conformability error"

$(B_{m\times \underline{k}})(A_{\underline{n}\times m})$

Transpose and inverse of a matrix

$(A_{a\times m}\ B_{m\times k})' = (AB_{a\times k})' = [k\times n] \longrightarrow B'_{k\times m} A'_{m\times n}$ ✓

- Transpose of Product $(AB)' = B'A'$ and $(ABC)' = C'B'A'$
- Inverse of Product $(AB)^{-1} = B^{-1}A^{-1}$ and $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$
- Transpose of an inverse equals inverse of a transpose $(D^{-1})' = (D')^{-1}$

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

Law of Iterated Expectations (LIE): A useful trick

- **Formally**: The unconditional expectation of a random variable is equal to the expectation of the conditional expectation of the random variable conditional on some other random variable

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}[Y|X])$$

- **Informally**: the weighted average of the conditional averages is the unconditional average

Example of LIE

- Say want average wage but only know average wage by education level
- LIE says we get the former by taking conditional expectations by education level and combining them (properly weighted)

$$\begin{aligned}\mathbb{E}[Wage] &= \mathbb{E}(\mathbb{E}[Wage|Education]) \\ &= \sum_{Education_i} Pr(Education_i) \cdot E[Wage|Education_i]\end{aligned}$$

Person	Gender	IQ
1	M	120
2	M	115
3	M	110
4	F	130
5	F	125
6	F	120

- E[IQ] = 120
- E[IQ | Male] = 115; E[IQ | Female] = 125
- LIE: E (E [IQ | Sex]) = (0.5)×115 + (0.5)×125 = 120

LIE: Proof for the discrete case

$$
\begin{aligned}
\mathbb{E}(\mathbb{E}[Y|X]) &= \sum_x \mathbb{E}[Y|X=x]p(x) \\
&= \sum_x \left(\sum_y yp(y|x)\right) p(x) \\
&= \sum_x \sum_y yp(x,y) \\
&= \sum_y y \sum_x p(x,y) \\
&= \sum_y yp(y) \\
&= \mathbb{E}(Y)
\end{aligned}
$$

$x \; y$

$$\rightarrow \sum_x \sum_Y y \underbrace{P(y|x)\, P(x)}_{P(x,y)} \rightarrow \text{Regla Bayes}$$

$$\rightarrow \sum_y \sum_x y P(x,y)$$

LIE: Proof for the continuous case

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}(Y|X)] &= \int \mathbb{E}(Y|X=u)g_x(u)du \\
&= \int \left[\int tf_{y|x}(t|X=u)dt\right] g_x(u)du \\
&= \int \int tf_{y|x}(t|X=u)g_x(u)dudt \\
&= \int t \left[\int f_{y|x}(t|X=u)g_x(u)du\right] dt \\
&= \int t\,[f_{x,y}du]\,dt \\
&= \int tg_y(t)dt \\
&= \mathbb{E}(y)
\end{aligned}
$$

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

OLS - As minimizing residuals

- Data with n observations and two variables: $(x_1, ...x_n)$ and $(y_1, ..., y_n)$
- Find the line $(\widehat{\beta}_0 + \widehat{\beta}_1 x)$ that best fits the data
- $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ is the fitted value for i
- The residual is $\widehat{u}_i = y_i - \widehat{y}_i$
- Goal: minimize residuals or distance from the line (fitted values) to the data

OLS - As minimizing residuals

- We don't care if the residual $\widehat{u}_i$ is positive or negative, we want it to be small
- So we square it: ${\widehat{u}_i}^2$
- Why not the absolute value? Good statistical reasons + harder to work with $|\cdot|$
- We want all the mistakes to be small, so we really want to minimize $\sum_{i=1}^n {\widehat{u}_i}^2$

OLS - As minimizing residuals

$$\min_{\widehat{\beta}_0,\widehat{\beta}_1} \sum_{i=1}^{n} \widehat{u}_i^{\,2} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \boxed{\sum_{i=1}^{n} \left(y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right)^2}$$

$$\frac{\partial}{\partial \widehat{\beta}_0} = 0$$

$$\frac{\partial}{\partial \widehat{\beta}_1} = 0$$

- Using calculus (deriving with respect to $\widehat{\beta}_0, \widehat{\beta}_1$ and equating to zero):

$$\boxed{\widehat{\beta}_1^{\,*} = \frac{\sum_{i=1}^{n}(x_i - \overline{x_i})(y_i - \overline{y_i})}{\sum_{i=1}^{n}(x_i - \overline{x_i})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_i})(y_i - \overline{y_i})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_i})^2} = \frac{\text{Sample covariance (x,y)}}{\text{Sample variance (x)}}$$

$$\boxed{\widehat{\beta}_0^{\,*} = \overline{y_i} - \widehat{\beta}_1 \overline{x_i}}$$

Visual tour of OLS

- https://ryansafner.shinyapps.io/ols_estimation_by_min_sse/
- https://seeing-theory.brown.edu/regression-analysis/index.html#section1
- https://setosa.io/ev/ordinary-least-squares-regression/
- https://mgimond.github.io/Stats-in-R/regression.html

OLS as an estimator

- There is a population with two random variables x and y
- We take a **random sample** of size n: $(x_1, x_2, ..x_n)$ and $(y_1, y_2, ..., y_n)$
- We would like to see how y varies with changes in x
  - What if y is affected by factors other than x?
  - What is the functional form connecting these two variables?
  - If interested in causal effect of x on y, how to distinguish from mere correlation?

OLS as an estimator of the DGP parameters

- Assume the data generating proces (DGP)s is:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- That is, this model holds in the **population**
- Not only x_i affects y_i, u_i (called the error term) also does
- Do not confuse u_i with $\widehat{u_i}$
- We assume there is a linear relationship between y_i and x_i
- We never observe β_0 and β_1

Inference

- Goal: Estimate unknown parameters
- To approximate parameters, we use an estimator, which is a function of the data
- Thus, estimator is a random variable (it is a function of a random variable)
- Infer something about the parameters from the distribution of the estimator

Important notation

Based on this tweet: https://twitter.com/nickchk/status/1272993322395557888

- Greek letters (e.g., μ) are the truth (i.e., parameters of the true DGP)
- Greek letters with hats (e.g., $\widehat{\mu}$) are estimates (i.e., what we *think* the truth is)
- Non-Greek letters (e.g., X) denote sample/data
- Non-Greek letters with lines on top (e.g., $\overline{X}$) denote calculations from the data
- We want to estimate the truth, with some calculation from the data ($\widehat{\mu} = \overline{X}$)
- Data $\longrightarrow$ Calculations $\longrightarrow$ Estimate $\underbrace{\longrightarrow}_{\text{Hopefully}}$ Truth
- Example: X $\longrightarrow \overline{X} \longrightarrow \widehat{\mu} \underbrace{\longrightarrow}_{\text{Hopefully}} \mu$

OLS as an estimator of the DGP parameters

- Assume the data generating process is:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- Also assume $\mathbb{E}u_i = 0$
  - Without loss of generality
  - We can just change the intercept to force $\mathbb{E}u_i = 0$
  - For example if $\mathbb{E}u_i = \alpha_0$
  - Redefine model to $y_i = \underbrace{\beta_0 + \alpha_0}_{\text{new intercept}} + \beta_1 x_i + \underbrace{u_i - \alpha_0}_{\text{new error term}}$
- Assume mean independence $\mathbb{E}(u_i|x_i) = \mathbb{E}(u_i)$ for all values x
  - ~~This is a non-trivial assumption, but let's take it for granted for now~~
  - Implies that $\mathbb{E}(u_i|x) = \mathbb{E}(u_i) = 0$
  - Implies that $\mathbb{E}(u_i x_i) = \mathbb{E}(\mathbb{E}(u_i|x_i)) = 0$

OLS as an estimator of the DGP parameters

- $\mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_i$ → $E(y_i|x_i) = E(\beta_0|x_i) + E(\beta_1 x_i|x_i) + \cancel{E(u_i|x_i)} \to 0$

$E(\beta_0 + \beta_1 x_i + u_i|x_i)$

- $\mathbb{E}(y_i|x_i)$: population regression function or conditional expectation function
- By our assumptions:
  - $\mathbb{E}(u_i|x_i) = \mathbb{E}(y_i - \beta_0 - \beta_1 x_i) = 0$
  - $\mathbb{E}(u_i x_i) = \mathbb{E}(x(y_i - \beta_0 - \beta_1 x_i)) = 0$
- These two conditions determine β_0 and β_1

OLS as an estimator of the DGP parameters

First equation

$$\begin{aligned}
\mathbb{E}(y_i - \beta_0 - \beta_1 x_i) &= 0\\
\mathbb{E}y_i - \beta_0 - \beta_1 \mathbb{E}x_i &= 0\\
\boxed{\mathbb{E}y_i - \beta_1 \mathbb{E}x_i = \beta_0}
\end{aligned}$$

OLS as an estimator of the DGP parameters

Second equation

→β_0

$$\begin{aligned}
\mathbb{E}x_i(y_i - \beta_0 - \beta_1 x_i) &= 0\\
\mathbb{E}x_i(y_i - (\mathbb{E}y_i - \beta_1 \mathbb{E}x_i) - \beta_1 x_i) &= 0\\
\mathbb{E}x_i(y_i - \mathbb{E}y_i - \beta_1(x_i - \mathbb{E}x_i)) &= 0\\
\boxed{\mathbb{E}x_i(y_i - \mathbb{E}y_i)} &= \mathbb{E}x_i\beta_1(x_i - \mathbb{E}x_i)\\
\cancel{\mathbb{E}(x_i - \mathbb{E}x_i)(y_i - \mathbb{E}y_i)} &= \beta_1\mathbb{E}(x_i - \mathbb{E}x_i)(x_i - \mathbb{E}x_i)\\
\boxed{\frac{\mathbb{E}(x_i - \mathbb{E}x_i)(y_i - \mathbb{E}y_i)}{\mathbb{E}(x_i - \mathbb{E}x_i)(x_i - \mathbb{E}x_i)}} &= \beta_1\\
\frac{\text{Population covariance (x,y)}}{\text{Population variance (x)}} &= \beta_1
\end{aligned}$$

→ ✓

$\beta_1 \, E(x_i - E(x_i))(x_i - E(x_i))$

$\beta_1(E(x_i^2 - 2E(x_i)x_i + E^2(x_i))$

$\beta_1(E(x_i^2 - 2E^2(x_i) + E^2(x_i))$

$\beta_1(E(x_i^2 - E^2(x_i))$

OLS as an estimator of the DGP parameters

- But we don't have x and y, nor do we know $\mathbb{E}y_i$ or $\mathbb{E}x_i$
- We only have a **random sample** of size n: $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$
- The sample analogs:
  - $\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$
  - $\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$

OLS as an estimator of the DGP parameters

First equation

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) &= 0\\
\frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{\beta}_0 - \widehat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i &= 0\\
\overline{y_i} - \widehat{\beta}_0 - \widehat{\beta}_1 \overline{x_i} &= 0\\
\boxed{\overline{y_i} - \widehat{\beta}_1 \overline{x_i} = \widehat{\beta}_0}
\end{aligned}$$

OLS as an estimator of the DGP parameters

Second equation

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) &= 0\\
\frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - \left(\overline{y_i} - \widehat{\beta}_1\overline{x_i}\right) - \widehat{\beta}_1 x_i\right) &= 0\\
\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \overline{y_i} + \widehat{\beta}_1(\overline{x_i} - x_i)) &= 0\\
\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \overline{y_i}) &= \frac{1}{n}\sum_{i=1}^{n} x_i(\widehat{\beta}_1(x_i - \overline{x_i}))\\
\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x_i})(y_i - \overline{y_i}) &= \widehat{\beta}_1\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_i})(x_i - \overline{x_i})\\
\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_i})(y_i - \overline{y_i})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_i})^2} &= \widehat{\beta}_1\\
\boxed{\frac{\text{Sample covariance (x,y)}}{\text{Sample variance (x)}} = \widehat{\beta}_1}
\end{aligned}$$

OLS as an estimator of the DGP parameters

- Formulas are the same as "minimizing residuals"
- Show the OLS coefficients as estimator of the population parameters (β_0 and β_1)
- Some remarks:
  - Can only estimate if the sample variance of x_i is not zero
  - In other words, if x_i is not constant across all values of i
  - Intuitively, the variation in x is what permits us to identify its impact in y

Multiple regression – notation

- Consider the multiple linear regression model

$$y_i = x_i'\beta + u_i$$

  where $\beta = (\beta_0, \beta_1, ..., \beta_K)'$ and $x_i = (1, ..., x_K)'$
  - β is of size $(k \times 1)$
  - x_i is of size $(k \times 1)$
  - $x_i'\beta$ is of size $(1 \times k)(k \times 1) = 1 \times 1$
- Equivalent

$$y = X\beta + u$$

  where $\beta = (\beta_0, \beta_1, ..., \beta_K)'$
  - β is of size $(k \times 1)$
  - X is of size $(n \times k)$
  - $X\beta$ is of size $(n \times k)(k \times 1) = n \times 1$

Multiple regression

- Consider the multiple linear regression model

$$\boxed{y_i = x_i'\beta + u_i} \rightarrow \mathbb{E}(Y_i|X_i) = \mathbb{E}(X_i'\beta|X_i) + \mathbb{E}(\cancel{u_i|X_i})$$

- $x_i'\beta = \beta_0 + \sum_{k=1}^{K} x_{ik}\beta_k$ is the conditional expectation function ($\mathbb{E}(y_i|x_i)$)
- The population regression β coefficients solve

$$\boxed{\beta = \mathbb{E}[x_i xi']^{-1}\mathbb{E}[x_i y_i]} = \arg\min_b \mathbb{E}\left[(y_i - x_i'b)^2\right]$$

- The sample equivalent is

$$\boxed{\widehat{\beta}} = \frac{\cancel{\frac{1}{n}}\sum_{i=1}^{n}(x_i y_i)}{\cancel{\frac{1}{n}}\sum_{i=1}^{n}(x_i x_i')} = \boxed{\frac{\sum_{i=1}^{n}(x_i y_i)}{\sum_{i=1}^{n}(x_i x_i')}} = \boxed{(X'X)^{-1}X'y}$$

$$X_i' = (X_{1i}, ..., X_{Ki})$$

$$X_i'\beta = (X_{0i}, ..., X_{Ki})\begin{pmatrix}\beta_0\\ \vdots\\ \beta_K\end{pmatrix}$$

$$\sum_{j=0}^{K} X_{ji}\beta_j$$

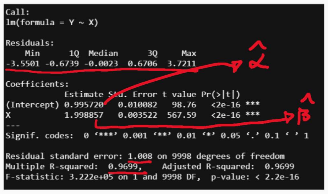
Solving for $\widehat{\beta}$

- We let the computer do the calculations, which are tedious even for small n
- Good to know what's going on behind the scenes
- But I honestly do not care if you know how invert a matrix

Simulations!

```
alpha=1 #intercept
beta=2 #slope
Nobs=10000 #how many observations?
X=runif(Nobs,-5,5)
#use the DGP to generate data
Y=alpha+beta*X+rnorm(Nobs)
OLS=lm(Y~X)
summary(OLS)
```

Our estimate of the coefficient are pretty close to the truth



```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5501 -0.6739 -0.0023  0.6706  3.7211 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.995720   0.010082   98.76   <2e-16 ***
X           1.998857   0.003522  567.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.008 on 9998 degrees of freedom
Multiple R-squared:  0.9699,	Adjusted R-squared:  0.9699 
F-statistic: 3.222e+05 on 1 and 9998 DF,  p-value: < 2.2e-16
```

$\hat{\alpha}$

$\hat{\beta}$

But how close on average?

```
#Now lets repeat the process and see how close our estimates are
Reps=1000
alpha_estimate=NULL
beta_estimate=NULL
for(r in 1:Reps){
  X=runif(Nobs,-5,5)
  #use the DGP to generate data
  Y=alpha+beta*X+rnorm(Nobs)
  OLS=lm(Y~X)
  Estimates=summary(OLS)$coef[,"Estimate"]
  alpha_estimate=c(alpha_estimate,Estimates[1])
  beta_estimate=c(beta_estimate,Estimates[2])
}
hist(beta_estimate,freq=F,breaks=30,
     main="",las=1,xlab="Estimate of beta")
abline(v=beta,col='red',lwd=2,lty=2)
```

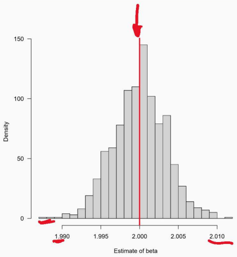
Vector $\hat{\alpha}$

Vector $\hat{\beta}$

$$\underline{\hat{\beta}} \sim N(\beta, \overset{\Sigma}{\boxed{\sigma_\beta}})$$

ERROR ESTANDAR DE $\hat{\beta}$

Distribution of estimate of $\widehat{\beta}$



OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

“Algebraic” properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

“Algebraic” properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

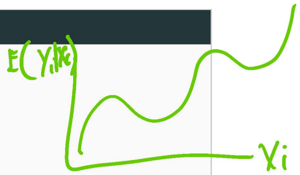
Conditional expectation function (CEF)

- Assume we are interested in the returns to schooling
- Summarize the effect of schooling on wages with the CEF ($\mathbb{E}(y_i|x_i)$)
- The CEF is the expectation (i.e, population average) of y_i with x_i held constant
- $\mathbb{E}(y_i|x_i)$ provides a reasonable representation of how y changes with x
- Because x_i is random, $\mathbb{E}[y_i \mid x_i]$ is random
- Sometimes work with a particular value of the CEF (e.g., $\mathbb{E}[y_i \mid x_i = 12]$)

Property 1: CEF Decomposition Property

- $y_i = \mathbb{E}(y_i|x_i) + u_i$ where
  1. u_i is mean independent of x_i; that is $\mathbb{E}(u_i|x_i) = 0$
  2. u_i is uncorrelated with any function of x_i
- In words: any random variable, y_i, can be decomposed into two parts: the part that can be explained by x_i and the part left over that cannot be explained by x_i
- Proof is in Angrist and Pischke (ch. 3)

Property 2: CEF Prediction Property



- Let $m(x_i)$ be any function of x_i
- $\mathbb{E}(y_i|x_i) = \arg\min_{m(x_i)} \mathbb{E}[(y_i - m(x_i))^2]$
- In words: The CEF is the minimum mean squared error predictor of y_i given x_i
- Proof is in Angrist and Pischke (ch. 3)

Property 3: Best linear approximation

- The population regression is the best linear approximation to the true nonlinear CEF in a mean squared error sense:

$$\beta = \mathbb{E}[x_i xi']^{-1}\mathbb{E}[x_i y_i] = \arg\min_{b} \mathbb{E}[(\mathbb{E}[y_i \mid x_i] - x_i' b)^2]$$

- In words: even if the true CEF is nonlinear (for example, $E[y_i \mid x_i] = \log(x_i)$), regression is still a good approximation to the truth

Why linear regression may be of interest (summary)

- If the CEF is linear. Then the population regression is it
  - Then it makes the most sense to use linear regression to estimate it
- Linear regression may be interesting even if the underlying CEF is not linear
  - $\mathbb{E}(y_i|x_i)$, is the minimum mean squared error predictor of y_i given x_i in the class of all functions of x_i
  - The population regression function is the best we can do in the class of all linear functions to approximate $\mathbb{E}(y_i|x_i)$

Big picture

1. Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable
2. If we prefer to think of approximating $\mathbb{E}(y_i|x_i)$ as opposed to predicting y_i, even if the CEF is nonlinear, regression provides the best linear approximation to it

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

“Algebraic” properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

OLS Review

- Linear algebra review
- Law of iterated expectations
- OLS basics
- Conditional expectation function
- "Algebraic" properties of OLS
- Properties of OLS estimators
- Regression (matrix algebra) with a treatment dummy for the experimental case
- Frisch–Waugh–Lovell (FWL) theorem
- Regression and causality

Residuals add up to zero

- Remembering how the **first moment** condition allows us to obtain $\widehat{\beta}_0$ and $\widehat{\beta}_1$:

$E(U_i)=0$

$$\boxed{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0} = \frac{1}{n}\sum \hat{U_i} = 0$$

$\frac{1}{n}\sum \hat{U_i} = Y_i - \beta_0 - \beta_1 X_i = 0$

- This means the OLS residuals *always* add up to zero, by *construction*,

$$\boxed{\frac{1}{n}}\sum_{i=1}^{n}\widehat{u}_i = 0$$

$$\boxed{\sum_{i=1}^{n}\widehat{u}_i = 0}$$

The mean of the fitted values is the mean of the data

Because $\underline{y_i} = \underline{\widehat{y}_i} + \underline{\widehat{u}_i}$ by definition,

$$\boxed{\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\widehat{y}_i + \boxed{\cancel{\sum_{i=1}^{n}\widehat{u}_i}}} = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i$$

$$\underline{\overline{y}} = \overline{\widehat{y}}$$

$Y_i = \mathbb{E}(Y_i|X_i) + \underline{\underline{U_i}}$

$U_i \neq \hat{U_i}$

$$\boxed{\hat{U_i} = Y_i - \hat{Y_i}}$$

Sample correlation between x_i and residuals is zero

Similarly the way we obtained our estimates,

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0} \qquad (1)$$

$\boxed{\mathbb{E}(X_i U_i) = 0}$

$\frac{1}{n}\sum X_i \hat{U_i} = 0$

The sample covariance (and therefore the sample correlation) between the explanatory variables and the residuals is always zero:

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} x_i \widehat{u}_i = 0} \qquad (2)$$

$\mathbb{E}(X_i U_i) \neq 0$

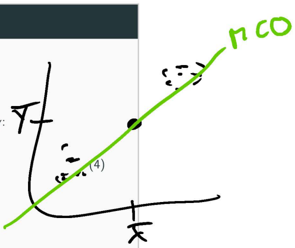
Bringing things together

Because the $\widehat{y}_i$ are linear functions of the x_i, the fitted values and residuals are uncorrelated, too:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i\widehat{u}_i = 0 \qquad (3)$$

The point $(\overline{x}, \overline{y})$ is always on the OLS regression line

If we plug in the average for x, we predict the sample average for y:

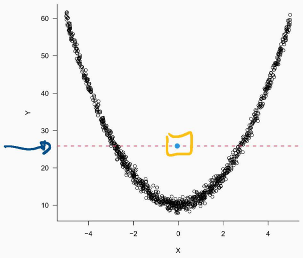


$$\overline{y} = \underline{\widehat{\beta}_0 + \widehat{\beta}_1\overline{x}} \qquad (4)$$

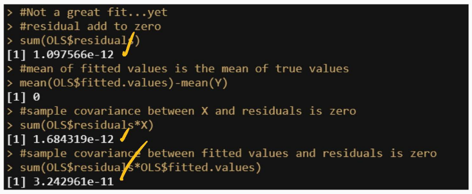
(see formula for $\widehat{\beta}_0$)

```
Nobs=1000 #how many observations?
X=runif(Nobs,-5,5)
#use the DGP to generate data
Y=10+2*X^2+rnorm(Nobs)
OLS=lm(Y~X)
summary(OLS)
plot(X,Y,bty="L")
abline(OLS,col=2,lwd=2,lty=2)
points(mean(X),mean(Y),pch=19,col=4,cex=1.5)
#Not a great fit...yet
#residual add to zero
sum(OLS$residuals)
#mean of fitted values is the mean of true values
mean(OLS$fitted.values)-mean(Y)
#sample covariance between X and residuals is zero
sum(OLS$residuals*X)
#sample covariance between fitted values and residuals is zero
sum(OLS$residuals*OLS$fitted.values)
```

y - Y



Algebraic properties



```
> #Not a great fit...yet
> #residual add to zero
> sum(OLS$residuals)
[1] 1.097566e-12
> #mean of fitted values is the mean of true values
> mean(OLS$fitted.values)-mean(Y)
[1] 0
> #sample covariance between X and residuals is zero
> sum(OLS$residuals*X)
[1] 1.684319e-12
> #sample covariance between fitted values and residuals is zero
> sum(OLS$residuals*OLS$fitted.values)
[1] 3.242961e-11
```

Big picture

Don't let anyone tell you the model is good because any of the following happens

1. Residuals add to zero
2. Fitted values mean is equal to data mean
3. Residuals are uncorrelated with x
4. If we plug in the average for *x*, we predict the sample average for *y*

These results are mechanical: Unrelated to how appropriate the model is or “causality”

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

“Algebraic” properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

“Algebraic” properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

Expected Value of OLS

- Mathematical statistics: How do our estimators behave across different samples of data? On average, would we get the right answer if we could repeatedly sample?
- Find the expected value of the OLS estimators – the average outcome across all possible random samples – and determine if we are right on average
- Leads to the notion of **unbiasedness**, a "desirable" characteristic for estimators.

$$\mathbb{E}(\widehat{\beta}) = \beta \tag{5}$$

Don't forget why we're here

- The **population** parameter that describes the relationship between y and x is β
- Goal: estimate β with a sample of data
- $\widehat{\beta}$ is an **estimator** obtained with a *specific* sample from the population

Uncertainty and sampling variance

- Different samples will generate different estimates ($\widehat{\beta}$) for the "true" β
- Thus, $\widehat{\beta}$ a random variable (depends on random samples)
- Unbiasedness is the idea that if we could take as many random samples on y as we want from the population, and compute an estimate each time, the average of these estimates would be equal to β
- But, this also implies that $\widehat{\beta}$ has spread and therefore variance

Assumption 1 (Linear in Parameters)

- The population model can be written as

$$y = X\beta + u \tag{6}$$

  where β are the (unknown) population parameters
- We view X and u as outcomes of random variables; thus, y is random
- Our goal is to estimate β
- u is the unobserved error. It is not the residual that we compute from the data!

Assumption 2 (Random Sampling)

- We have a random sample of size n, $\{(x_i, y_i) : i = 1, ..., n\}$
- We know how to use this data to estimate β by OLS

Assumption 3 (Zero Conditional Mean)

$$\underset{\text{COVID}}{Y} = \underset{\underset{\text{EDUCACION}}{\downarrow}}{X}\beta + U$$

- In the population, the error term has zero mean given any value of X:

$$\underline{\mathbb{E}(u|X)} = \mathbb{E}(u) = 0. \tag{7}$$

- This is the key assumption for showing that OLS is unbiased, with the zero value not being important once we assume $\mathbb{E}(u|X)$ does not change with X

Assumption 1-3

- We can compute the OLS estimates whether or not these assumption hold
- But we might not get a "good" estimate

Assumption 4 (Sample Variation in the Explanatory Variable)

- The sample outcomes on x_i are not all the same value
- Same as saying the sample variance of $\{x_i : i = 1, ..., n\}$ is not zero
- If the x_i are all the same value, we cannot learn how x affects y

Showing OLS is unbiased

- How do we show $\widehat{\beta}$ is unbiased for β?
- We know $\boxed{\widehat{\beta} = (X'X)^{-1}X'y}$
- And that $\underline{y = X\beta + u \text{ (by assumption 1)}}$

$$\to \underline{(X'X)^{-1}X'X}\beta + (X'X)^{-1}X'u$$

- Therefore: $\widehat{\beta} = (X'X)^{-1}X'(X\beta + u) = \beta + (X'X)^{-1}X'u$
- $\mathbb{E}(\widehat{\beta} \mid X) = \beta + (X'X)^{-1}X' \boxed{\underbrace{\mathbb{E}(u \mid X)}_{=0 \text{ by assumption 3}}}$
- $\mathbb{E}(\widehat{\beta} \mid X) = \beta$

$$\mathbb{E}(\mathbb{E}(\hat{\beta} \mid X)) = \beta$$

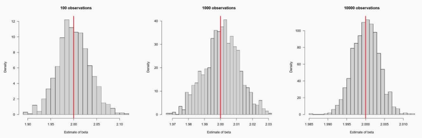
$$\mathbb{E}(\hat{\beta}) = \beta$$

- Each sample leads to a different estimate, $\widehat{\beta}$
- Some will be very close to the true values β
- Some **could** be very far from those values
- If we repeat the experiment and average the estimates → very close to β
- But in a single sample, we can never know whether we are close to β
- Next: measure of dispersion (spread) in the distribution of the estimators

Repeat our simulations with different N

```
alpha=1 #intercept
beta=2 #slope
Reps=1000
for(Nobs in c(100,1000,10000)){
  alpha_estimate=NULL
  beta_estimate=NULL
  for(r in 1:Reps){
    X=runif(Nobs,-5,5)
    Y=alpha+beta*X+rnorm(Nobs)
    OLS=lm(Y~X)
    Estimates=summary(OLS)$coef[,"Estimate"]
    alpha_estimate=c(alpha_estimate,Estimates[1])
    beta_estimate=c(beta_estimate,Estimates[2])
  }
  hist(beta_estimate,freq=F,breaks=30,main="",las=1)
  abline(v=beta,col='red',lwd=3,lty=1)
}
```

Repeat our simulations with different N — Look at the x-axis scale



Reminder

- **Errors** are the vertical distances between observations and the **unknown** Conditional Expectation Function. Therefore, they are unknown.
- **Residuals** are the vertical distances between observations and the **estimated** regression function. Therefore, they are known.

Variance of OLS estimators

The correct variance estimation procedure is given by the structure of the data

- It is very unlikely that all observations in a dataset are unrelated, but drawn from identical distributions (**homoskedasticity**)
- For instance, the variance of income is often greater in families belonging to top deciles than among poorer families (**heteroskedasticity**)
- Some phenomena do not affect observations individually, but they do affect groups of observations uniformly within each group (**clustered data**)

Assumption 5 (Homoskedasticity, or Constant Variance)

The error has the same variance given any value of the explanatory variable x:

$$Var(u|X) = \sigma^2 > 0 \tag{8}$$

where σ^2 is (virtually always) unknown.

Because $\mathbb{E}(u|x) = 0$ we can also write

$$\mathbb{E}(u^2|x) = \sigma^2 = \mathbb{E}(u^2) \tag{9}$$

Assumption 5 (Homoskedasticity, or Constant Variance)

Under the our assumptions

$$\begin{aligned} y &= X\beta + u \\ \underline{\mathbb{E}(y|x)} &\underline{= X\beta} \\ Var(y|x) &= \sigma^2 \end{aligned}$$

$$V(y) = \underline{V(X\beta)} + V(u)$$

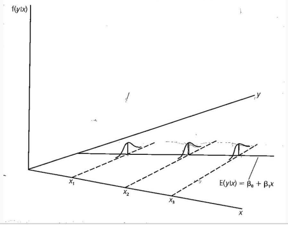
$$V(y|x) = V(u|x) = \sigma^2$$

The average or expected value of y is allowed to change with x, but the variance does not change with x

Assumption 5 (Homoskedasticity, or Constant Variance)

Figure 2.8

The simple regression model under homoskedasticity.



Variance of OLS estimators In matrix form the property that $\underline{V(aW)} = \underline{\boxed{a^2} V(W)}$ where a is constant and W is a random variable is written as:

$$\underline{V(\boxed{A}W)} = \overset{\downarrow}{A}V(W)\overset{\downarrow}{A'}$$

where A is a constant matrix and W is a random variable

Variance of OLS estimators

$$((X'X)^{-1})^t = ((X'X)^t)^{-1}$$
$$= (X^tX)^{-1}$$

- We know $\widehat{\beta} = (X'X)^{-1}X'y$
- And that $y = X\beta + u$ (by assumption 1)
- Therefore: $\widehat{\beta} = (X'X)^{-1}X'(X\beta + u) = \beta + (X'X)^{-1}X'u$
- $V(\widehat{\beta} \mid X) = \underbrace{V(\beta \mid X)}_{=0 \text{since it's constant}} + (X'X)^{-1}X' \underbrace{V(u \mid X)}_{=\sigma^2 \text{ by assumption 3}} \boxed{X(X'X)^{-1}}$
- $V(\widehat{\beta} \mid X) = (X'X)^{-1}X'\sigma^2 X(X'X)^{-1} = \sigma^2(X'X)^{-1}$

$$\sigma^2(X'X)^{-1}X'X(X'X)^{-1}$$

Estimating the Error Variance

- In the formula
$$V(\widehat{\beta} \mid X) = (X'X)^{-1}X'\sigma^2 X(X'X)^{-1} = \boxed{\sigma^2 \boxed{(X'X)^{-1}}}$$
we can compute $(X'X)^{-1}$ but we need to estimate σ^2
- Recall that
$$\sigma^2 = \mathbb{E}(u^2)$$

Estimating the Error Variance

- If we could observe the errors (u_i) an unbiased estimator of σ^2 would be
$$\boxed{\frac{1}{n}\sum_{i=1}^{n} u_i^2} \tag{10}$$
- But this not a *feasible* estimator because the u_i are unobserved
- How about replacing each u_i with its "estimate", the OLS residual $\widehat{u}_i$?

$$\boxed{u_i = y_i - x_i'\beta}$$
$$\boxed{\widehat{u}_i = y_i - x_i'\widehat{\beta}}$$

Estimating the Error Variance

$\widehat{u}_i$ can be computed from the data, but $\widehat{u}_i \neq u_i$ for any i:

$$\begin{aligned} \widehat{u}_i &= y_i - x_i'\widehat{\beta} = x_i'\beta + u_i - x_i'\widehat{\beta} \\ &= u_i - (\widehat{\beta} - \beta)x_i \end{aligned}$$

$\mathbb{E}(\widehat{\beta}) = \beta$ but the estimators differ from the population values in a given sample

Estimating the Error Variance

- Now, what about this as an estimator of σ^2?
$$\boxed{\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i^2} \tag{11}$$
- It is a *feasible* estimator and easily computed from the data after OLS
- As it turns out, this estimator is slightly biased

Estimating the Error Variance

The estimator does not account for the restrictions on the residuals, used to obtain $\widehat{\beta}$

$$\left.\begin{aligned} &\boxed{\sum_{i=1}^{n} \widehat{u}_i = 0} \\ &\sum_{i=1}^{n} x_{1i}\widehat{u}_i = 0 \\ &\qquad \vdots \\ &\sum_{i=1}^{n} x_{ki}\widehat{u}_i = 0 \end{aligned}\right\} \text{k RESTRICCIONES}$$

There is no such restriction on the unobserved errors

Estimating the Error Variance

The unbiased estimator of σ^2 uses a **degrees-of-freedom** adjustment The residuals have only $n-k$ degrees-of-freedom (minus the k restrictions), not n

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} \widehat{u}_i^2}{\underline{(n-k)}}$$

THEOREM: Unbiased Estimator of σ^2

Under Assumptions 1-5,

$$\mathbb{E}(\widehat{\sigma}^2) = \sigma^2$$

- Given $\widehat{\sigma}$, we can now estimate $\underline{V(\widehat{\beta})}$
- $\underline{V(\widehat{\beta})}$ is a variance-covariance matrix (size $k \times k$)
- The diagonal elements of $V(\widehat{\beta})$ give us the variance of the estimators $\widehat{\beta}$
- $\widehat{\sigma_{\widehat{\beta}}}$: The square root of the diagonal elements of the estimator of $V(\widehat{\beta})$ is usually called the **standard errors** (i.e., estimate of the standard deviation of the estimator)

ERROR STD

$\widehat{\sigma}_{\widehat{\beta}}$

DESV. ESTIMADOR

$\sigma_{\widehat{\beta}}$

Bringing the central limit theorem to play

$V(aX) = \underline{a}^2 V(X)$

- By some version of the central limit theorem:

$$\begin{aligned} \frac{\widehat{\beta}-\beta}{\boxed{\sigma_{\widehat{\beta}}}} &\rightarrow_d N(0,1) \\ \widehat{\beta} &\rightarrow_d \underline{\sigma_{\widehat{\beta}}} N(0,1) + \beta \\ \widehat{\beta} &\rightarrow_d N(\beta, \sigma_{\widehat{\beta}}^2) \end{aligned}$$

- $\sigma_{\beta} = \sigma^2 (X'X)^{-1}$
- Since we do not know $\underline{\sigma^2}$, we estimate it
- $\boxed{\sigma_{\widehat{\beta}} = \widehat{\sigma}^2 (X'X)^{-1}}$
- By some version of the central limit theorem + some statistical properties

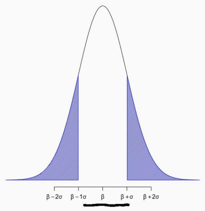
$$\begin{aligned} \frac{\widehat{\beta}-\beta}{\widehat{\sigma}_{\widehat{\beta}}} &\rightarrow_d t_{n-k} \\ \widehat{\beta} &\rightarrow_d \widehat{\sigma_{\widehat{\beta}}} t_{n-k} + \beta \end{aligned}$$

To keep things simple

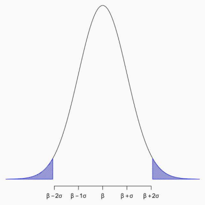
- $\underline{t_{n-k}} \rightarrow \underline{N(0,1)}$ as $(n-k) \rightarrow \infty$
- So as long as your sample is large, we can keep thinking of normal distributions

$$\widehat{\beta} \approx N(\beta, \widehat{\sigma}_{\widehat{\beta}}^2)$$

31.7% of estimates will be more than $\underline{\widehat{\sigma_{\widehat{\beta}}}}$ away from β



4.55% of estimates will be more than $2\widehat{\sigma_{\widehat{\beta}}}$ away from β

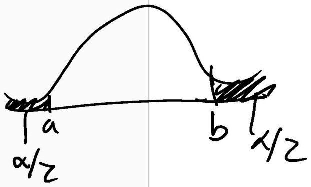


We can know learn something about the true β

- We know $\widehat{\beta} \sim N(\beta, \widehat{\sigma_{\beta}})$
- We want to find some interval on which β is likely to live:

$$P\left(\underline{a} \leq \beta \leq \underline{b}\right) = \underline{1-\alpha}$$

- $P\left(\underline{-a \geq -\beta \geq b}\right) = 1-\alpha$
- $P\left(\frac{\widehat{\beta}-a}{\widehat{\sigma_{\widehat{\beta}}}} \geq \underbrace{\boxed{\frac{\widehat{\beta}-\beta}{\widehat{\sigma_{\widehat{\beta}}}}}}_{\substack{\text{standard normal} \\ (t_{n-k}\text{to be exact})}} \geq \boxed{\frac{\widehat{\beta}-b}{\widehat{\sigma_{\widehat{\beta}}}}}\right) = 1-\alpha$
- Assuming we want symmetry (so $\frac{\alpha}{2}$ on each side), then:
  - $\boxed{\Phi}\left(\frac{\widehat{\beta}-a}{\widehat{\sigma_{\widehat{\beta}}}}\right) = 1-\frac{\alpha}{2}$
  - $\boxed{\Phi}\left(\frac{\widehat{\beta}-b}{\widehat{\sigma_{\widehat{\beta}}}}\right) = \frac{\alpha}{2}$



Confidence interval

- Thus:
  - $\Phi^{-1}\left(\frac{\alpha}{2}\right) = \boxed{\frac{\widehat{\beta}-b}{\widehat{\sigma_{\widehat{\beta}}}}}$
  - $\underline{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)} = \frac{\widehat{\beta}-a}{\widehat{\sigma_{\widehat{\beta}}}}$
  - $b = \widehat{\beta} - \Phi^{-1}\left(\frac{\alpha}{2}\right)\widehat{\sigma_{\widehat{\beta}}}$
  - $a = \widehat{\beta} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\widehat{\sigma_{\widehat{\beta}}}$
- β is between $\widehat{\beta} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\widehat{\sigma_{\widehat{\beta}}}$ and $\widehat{\beta} - \Phi^{-1}\left(\frac{\alpha}{2}\right)\widehat{\sigma_{\widehat{\beta}}}$ with probability $\underline{1-\alpha}$

Confidence interval

- Say $\alpha = 5\%$, then $\underline{\Phi^{-1}\left(\frac{\alpha}{2}\right)} = \underline{-1.96}$ and $\underline{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)} = \underline{1.96}$
- Then we know β is between with probability 95%:
  - $\widehat{\beta} - 1.96\widehat{\sigma_{\beta}}$
  - $\widehat{\beta} + 1.96\widehat{\sigma_{\beta}}$
- Generally speaking, confidence intervals are wider, the smaller α is

QNORM(0.025)
QNorm(1-0.025)

Simulations!

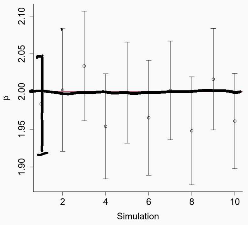
```
beta0=1 #intercept
beta1=2 #slope
#Now lets repeat the process and see how close our estimates are
Reps=1000
Nobs=100 #number of obs
beta0_estimate=NULL #vector to store estimates of beta0
beta1_estimate=NULL #vector to store estimates of beta1
EstimateInCI=NULL #vector to store whether estimate is in CI
LowerCI=NULL #vector to store lower bound CI
UpperCI=NULL #vector to store upper bound CI
Confidence_level=0.05 #alpha
for(r in 1:Reps){
  X=runif(Nobs,-5,5)
  Y=beta0+beta1*X+rnorm(Nobs)   #use the DGP to generate data
  OLS=lm(Y~X)
  Estimates=summary(OLS)$coef[,"Estimate"] #estimate
  SE=summary(OLS)$coef[,"Std. Error"] #estimate sigma_beta
  beta0_estimate=c(beta0_estimate,Estimates[1])
  beta1_estimate=c(beta1_estimate,Estimates[2])
  CI_Beta_lowerbound=Estimates[2]+qnorm(Confidence_level/2)*SE[2]   #Confidence intervals
  LowerCI=c(LowerCI,CI_Beta_lowerbound)
  CI_Beta_upperbound=Estimates[2]+qnorm(1-Confidence_level/2)*SE[2]
  UpperCI=c(UpperCI,CI_Beta_upperbound)
  DummyInCI=(CI_Beta_lowerbound<beta1 & CI_Beta_upperbound>beta1)   #Is the true value in CI?
  EstimateInCI=c(EstimateInCI,DummyInCI)
}
mean(EstimateInCI)
```

$\hat{\beta}_0$ $\hat{\beta}_1$

$\Phi^{-1}\left(\frac{\alpha}{2}\right)$

$\hat{\sigma}_{\hat{\beta}}$

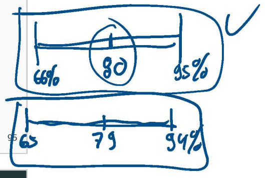
First ten simulations (red line is true β)



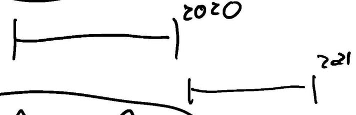
Test hypothesis

- Is $\beta \neq \beta_0$?
- Usually posed as testing $H_0 : \beta = \beta_0$ vs $H_a : \beta \neq \beta_0$
- Different way to look at this: is β_0 is in the confidence interval of β?
- Confidence interval depends on our choice of α
- Pick largest α for which β_0 is not in the confidence interval
- This is called the $p-value$
- Largest probability of obtaining results at least as extreme as those actually observed, under the assumption that the null hypothesis is correct

$H_0: \beta = 0$
P-Valor 3%



$$V(\hat{\beta}) = V\begin{pmatrix}\hat{\beta}_{2020} \\ \hat{\beta}_{2021}\end{pmatrix} = \hat{\sigma}^2_{\hat{\beta}}(X'X)^{-1}$$



$$\boxed{\hat{D} = \hat{\beta}_{2020} - \hat{\beta}_{2021}}$$

$$V(\hat{D}) = \hat{\sigma}^2_{\hat{\beta}_{2020}} + \sigma^2_{\hat{\beta}_{2021}} - \underline{2Cov(\hat{\beta}_{2020}, \hat{\beta}_{2021})}$$

Simulations!

observed, under the assumption that the null hypothesis is correct

65 79 94%

$$V(\hat{D}) = \hat{\sigma}^2_{\hat{\beta}2020} + \sigma^2_{\hat{\beta}2021} - 2\,Cov(\hat{\beta}2020, \hat{\beta}2021)$$

Simulations!

```
#### p−values
beta0=1 #intercept
beta1=2 #slope
#Now lets repeat the process and see how close our estimates are
Reps=1000
Nobs=1000 #number of obs
pvalue_beta0=NULL #vector to store whether  0 is in CI
pvalue_beta2=NULL #vector to store whether  2 is in CI
Confidence_level=0.05 #alpha
for(r in 1:Reps){
  X=runif(Nobs,−5,5)
  #use the DGP to generate data
  Y=beta0+beta1*X+rnorm(Nobs)
  OLS=lm(Y~X)
  Estimates=summary(OLS)$coef[,"Estimate"] #estimate
  SE=summary(OLS)$coef[,"Std. Error"] #estimate sigma_beta
  beta0_estimate=c(beta0_estimate,Estimates[1])
  beta1_estimate=c(beta1_estimate,Estimates[2])
  #Confidence intervals
  CI_Beta_lowerbound=Estimates[2]+qnorm(Confidence_level/2)*SE[2]
  CI_Beta_upperbound=Estimates[2]+qnorm(1−Confidence_level/2)*SE[2]
  #Is 0 in CI?
  pvalue_beta0=c(pvalue_beta0,(CI_Beta_lowerbound<0 & CI_Beta_upperbound>0))
  #Is 2 in CI?
  pvalue_beta2=c(pvalue_beta2,(CI_Beta_lowerbound<2 & CI_Beta_upperbound>2))
}
mean(pvalue_beta0)
mean(pvalue_beta2)
```

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

OLS

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

What's going on behind the scenes?

Simple case

- Relationship between outcome Y_i and treatment indicator T_i
- Regress the outcome on the treatment indicator, and a constant
- $X_i = \begin{pmatrix} 1 & T_i \end{pmatrix}$
- Assume first N_T units are treated ($N_C = N - N_T$ units are untreated)
- $X = \begin{pmatrix} 1 & T_1 \\ 1 & T_2 \\ \vdots & \\ 1 & T_{N_T} \\ 1 & T_{N_T+1} \\ \vdots & \\ 1 & T_N \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} N_T \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} N_c \end{matrix}$

4

Simple case

- $(X'X) = \underbrace{\begin{pmatrix} 1 & 1\cdots1 \\ 1 & 1\cdots1 \end{pmatrix}}_{NT}\underbrace{\begin{pmatrix} 1 & \cdots1 \\ 0 & \cdots0 \end{pmatrix}}_{N_C} \overset{X}{\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\1\\1\end{matrix}}\right\} N_T \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} \end{matrix} = \boxed{\begin{pmatrix} N & N_T \\ N_T & N_T \end{pmatrix}}$
- $(X'X)^{-1} = \frac{1}{N_T(N-N_T)}\begin{pmatrix} N_T & -N_T \\ -N_T & N \end{pmatrix} = \frac{1}{N_C}\begin{pmatrix} 1 & -1 \\ -1 & \frac{N}{N_T} \end{pmatrix}$

$$\frac{N_T}{N_T(N-N_T)} = \frac{1}{N_C} \qquad -\frac{1}{N_C}$$

$$\frac{-N_T}{N_T(N-N_T)} = -\frac{1}{N_C} \qquad \frac{N}{N_T(N-N_T)} = \frac{N}{N_T N_C}$$

Simple case

- $X'y = \begin{pmatrix} 1 & 1\cdots1 & 1 & \cdots1 \\ 1 & 1\cdots1 & 0 & \cdots0 \end{pmatrix}\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{N_T} \\ Y_{N_T+1} \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N_T} Y_i \end{pmatrix}$

Simple case

$$(X'X)^{-1}X'y = \underbrace{\frac{1}{N_C}\begin{pmatrix} 1 & -1 \\ -1 & \frac{N}{N_T} \end{pmatrix}}\underbrace{\begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N_T} Y_i \end{pmatrix}}$$

Simple case

$$\begin{aligned} (X'X)^{-1}X'y &= \frac{1}{N_C}\begin{pmatrix} 1 & -1 \\ -1 & \frac{N}{N_T} \end{pmatrix}\begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N_T} Y_i \end{pmatrix} \\ &= \frac{1}{N_C}\begin{pmatrix} \sum_{i=1}^{N} Y_i - \sum_{i=1}^{N_T} Y_i \\ \frac{N}{N_T}\sum_{i=1}^{N_T} Y_i - \sum_{i=1}^{N} Y_i \end{pmatrix} \rightarrow \end{aligned}$$

Simple case

$$\begin{aligned} (X'X)^{-1}X'y &= \frac{1}{N_C}\begin{pmatrix} 1 & -1 \\ -1 & \frac{N}{N_T} \end{pmatrix}\begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N_T} Y_i \end{pmatrix} \\ &= \frac{1}{N_C}\begin{pmatrix} \sum_{i=1}^{N} Y_i - \sum_{i=1}^{N_T} Y_i \\ \frac{N}{N_T}\sum_{i=1}^{N_T} Y_i - \sum_{i=1}^{N} Y_i \end{pmatrix} \\ &= \frac{1}{N_C}\begin{pmatrix} \sum_C Y_i \\ \frac{N}{N_T}\sum_T Y_i - \sum_T Y_i - \sum_C Y_i \end{pmatrix} \\ &= \frac{1}{N_C}\begin{pmatrix} \sum_C Y_i \\ \frac{N-N_T}{N_T}(\sum_T Y_i) - \sum_C Y_i \end{pmatrix} \end{aligned}$$

Simple case

$$\begin{aligned} (X'X)^{-1}X'y &= \frac{1}{N_C}\begin{pmatrix} 1 & -1 \\ -1 & \frac{N}{N_T} \end{pmatrix}\begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N_T} Y_i \end{pmatrix} \\ &= \frac{1}{N_C}\begin{pmatrix} \sum_{i=1}^{N} Y_i - \sum_{i=1}^{N_T} Y_i \\ \frac{N}{N_T}\sum_{i=1}^{N_T} Y_i - \sum_{i=1}^{N} Y_i \end{pmatrix} \\ &= \frac{1}{N_C}\begin{pmatrix} \sum_C Y_i \\ \frac{N}{N_T}\sum_T Y_i - \sum_T Y_i - \sum_C Y_i \end{pmatrix} \rightarrow \frac{1}{N_C}\Sigma_C Y_i = \overline{Y_C} \\ &= \frac{1}{N_C}\begin{pmatrix} \sum_C Y_i \\ \frac{N-N_T}{N_T}(\sum_T Y_i) - \sum_C Y_i \end{pmatrix} \rightarrow \frac{\cancel{N_C}}{NT\cancel{N_C}}\Sigma_T Y - \frac{1}{N_C}\Sigma_C Y_i \\ \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} &= \boxed{\begin{pmatrix} \overline{Y_C} \\ \overline{Y_T} - \overline{Y_C} \end{pmatrix}} \qquad \overline{Y_T} - \overline{Y_C} \end{aligned}$$

Simple case

Simple case

$$\begin{aligned}
(X'X)^{-1}X'y &= \frac{1}{N_C}\begin{pmatrix} 1 & -1 \\ -1 & \frac{N}{N_T}\end{pmatrix}\begin{pmatrix}\sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N_T} Y_i\end{pmatrix} \\
&= \frac{1}{N_C}\begin{pmatrix}\sum_{i=1}^{N} Y_i - \sum_{i=1}^{N_T} Y_i \\ \frac{N}{N_T}\sum_{i=1}^{N_T} Y_i - \sum_{i=1}^{N} Y_i\end{pmatrix} \\
&= \frac{1}{N_C}\begin{pmatrix}\sum_C Y_i \\ \frac{N}{N_T}\sum_T Y_i - \sum_T Y_i - \sum_C Y_i\end{pmatrix} \\
&= \frac{1}{N_C}\begin{pmatrix}\sum_C Y_i \\ \frac{N-N_T}{N_T}(\sum_T Y_i) - \sum_C Y_i\end{pmatrix} \\
\hat{\beta} &= \begin{pmatrix}\overline{Y_C} \\ \overline{Y_T} - \overline{Y_C}\end{pmatrix}
\end{aligned}$$

Simple case

$$(X'X)^{-1}X'y = \begin{pmatrix}\overline{Y_C} \\ \overline{Y_T} - \overline{Y_C}\end{pmatrix}$$

Simple case

$$(X'X)^{-1}X'y = \begin{pmatrix}\boxed{\overline{Y_C}} \\ \overline{Y_T} - \overline{Y_C}\end{pmatrix}$$

- The OLS estimate of the intercept is $\overline{Y_C}$
- The coefficient of the treatment dummy is $\overline{Y_T} - \overline{Y_C}$

$\overline{Y_C} = \boxed{80\%}$

$\overline{Y_T} - \overline{Y_C} = \underline{-0.5} = \hat{\beta}$

How precise are these estimates?

- What is the variance of $\widehat{\beta} = (X'X)^{-1}X'y$
- Recall $Y = X\beta + u$
- $\widehat{\beta} = (X'X)^{-1}X'(X\beta + u)$
- $\widehat{\beta} = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u$
- $\widehat{\beta} = \beta + (X'X)^{-1}X'u$
- If $\mathbb{E}(uX) = 0$
- $V(\widehat{\beta}) = (X'X)^{-1}X'V(u)X(X'X)^{-1}$ [matrix version of $V(b + aY) = a^2Y$]
- If $V(\varepsilon) = \sigma^2 I$ [Homoskedasticity] then
- $V(\widehat{\beta}) = (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} = \sigma^2\cancel{(X'X)^{-1}X'X(X'X)^{-1}} = \boxed{\sigma^2(X'X)^{-1}}$
- $\underline{V(\widehat{\beta})} = \sigma^2\frac{1}{N_T(N-N_T)}\begin{pmatrix} N_T & -N_T \\ -N_T & N\end{pmatrix} = \boxed{\frac{\sigma^2}{N_C}\begin{pmatrix} 1 & -1 \\ -1 & \frac{N}{N_T}\end{pmatrix}}$

How precise are these estimates?

- $\boxed{V(\widehat{\beta}) = \sigma^2\begin{pmatrix}\boxed{\frac{1}{N_C}} & -\frac{1}{N_C} \\ -\frac{1}{N_C} & \boxed{\frac{N}{N_T N_C}}\end{pmatrix}} \longrightarrow$
- Let $N_T = \kappa N$ and $N_C = (1-\kappa)N$
- Since we don't know σ^2, use $\frac{1}{N-1}(Y - \widehat{Y})^2 = \frac{1}{N-1}(\widehat{u})^2$ as an estimator

$V(\hat{\beta}_T) = \left(\frac{N}{N_T N_C}\right)\sigma^2$

$\boxed{\left(\frac{N_T + N_C}{N_T\ N_C}\right)\sigma^2}$

$N_T + N_C = N$

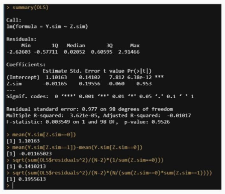
$N_T = N_C = \frac{1}{2}N$

Simulations!

```
N=100 #Number of individuals
mu0=1
s.sq=1
beta=0.2
#Let's create potential outcomes
Y0 <-  rnorm(n=N, mean=mu0, sd=s.sq)  # control potential outcome
Y1 <- Y0 + beta   # treatment potential outcome
#Lets randomly assign people to treatment
Z.sim <- rbinom(n=N, size=1, prob=.5) # Do a random assignment
Y.sim <- Y1*Z.sim + Y0*(1-Z.sim) # Reveal outcomes according to assignment

OLS=lm(Y.sim~Z.sim)
summary(OLS)
```

OLS estimator



```
> summary(OLS)

Call:
lm(formula = Y.sim ~ Z.sim)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.62603 -0.57711  0.02052  0.60595  2.91466 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.10163    0.14102   7.812 6.38e-12 ***
Z.sim       -0.01165    0.19556  -0.060    0.953    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.977 on 98 degrees of freedom
Multiple R-squared:  3.621e-05, Adjusted R-squared:  -0.01017 
F-statistic: 0.003549 on 1 and 98 DF,  p-value: 0.9526

> mean(Y.sim[Z.sim==0])
[1] 1.10163
> mean(Y.sim[Z.sim==1])-mean(Y.sim[Z.sim==0])
[1] -0.01165023
> sqrt(sum(OLS$residuals^2)/(N-2)*(1/sum(Z.sim==0)))
[1] 0.1410213
> sqrt(sum(OLS$residuals^2)/(N-2)*(N/(sum(Z.sim==0)*sum(Z.sim==1))))
[1] 0.1955613
> 
```

How precise are these estimates?

- Standard error of the intercept is: $\boxed{\sqrt{\hat{\sigma}\frac{1}{N_C}}}$
- Standard error of the slope is: $\sqrt{\hat{\sigma}\frac{N}{N_T N_C}}$
- This should tell us how much our estimates vary on different samples

```
sqrt(sum(OLS$residuals^2)/(N-2)*(1/sum(Z.sim==0)))
sqrt(sum(OLS$residuals^2)/(N-2)*(N/(sum(Z.sim==0)*sum(Z.sim==1))))
```

N_C N_T

Simulations!

```
N=100 #Number of individuals
mu0=1
s.sq=1
beta=0.2
Reps=1000
estimate_vector=NULL
for(r in 1:Reps){
  Y0 <-  rnorm(n=N, mean=mu0, sd=s.sq) # control potential outcome
  Y1 <- Y0 + beta # treatment potential outcome
  Z.sim <- rbinom(n=N, size=1, prob=.5) # Do a random assignment
  Y.sim <- Y1*Z.sim + Y0*(1-Z.sim) # Reveal outcomes according to assignmen
  OLS=lm(Y.sim~Z.sim)
  beta_estimate=summary(OLS)$coef[2,1]
  estimate_vector=c(estimate_vector,beta_estimate)
}
sd(estimate_vector)
```

Big picture

- We let the computer do the calculations, which are tedious even for small n
- Good to know what's going on behind the scenes
- But I honestly do not care if you know how invert a matrix
- Important things in life to understand:
    - What $\hat{\beta}$ is (an estimator of a parameter we do not observe)
    - What the standard error is (the standard deviation of the estimator)
    - What a confidence interval is (an interval where we know with some probability the true estimate lives)
    - What a p-value is (largest probability of obtaining results at least as extreme as those actually observed, under the assumption that the null hypothesis is correct)

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

Regression Anatomy Theorem – Frisch–Waugh–Lovell (FWL) theorem

- Assume your main multiple regression model of interest:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i$$

- An auxiliary regression in which the variable x_{1i} is regressed on all the remaining independent variables

$$\boxed{x_{1i}} = \underline{\gamma_0 + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \cdots + \gamma_K x_{Ki} + f_i}$$

- $\underline{\tilde{x}_{1i}} = \underline{x_{1i}} - \underline{\widehat{x}_{1i}}$ is the residual from the auxiliary regression
- The parameter β_1 can be rewritten as

$$\rightarrow \underline{\beta_1} = \frac{Cov(y_i, x_{1i})}{Var(x_{1i})} = \frac{Cov(y_i, \tilde{x}_{1i})}{Var(\underline{\tilde{x}_{1i}})}$$

- $\widehat{\beta}_1$ is a scaled covariance with the actual data $x$or with the $\tilde{x}_1$ residual

REG Y_i Vs $\underline{X_2, \ldots, X_K}$ → RESID $\tilde{Y}$

REG X_1 Vs $X_2, \ldots, X_K$ → RESID $\tilde{X}_1$

REG $\tilde{Y}$ vs $\tilde{X}_1$

→ COEF $\underline{\hat{\tilde{\beta}}_1}$ = COEF $\underline{\underline{\hat{\beta}_1}}$ = COEF $\hat{\hat{\hat{\beta}}}_1$

→ REG Y_1 vs $\boxed{X_1}, \ldots, X_K$

↓ REG Y_1 vs $\tilde{X}_1$

Regression Anatomy Theorem – Frisch–Waugh–Lovell (FWL) theorem II

- Assume your main multiple regression model of interest:

$$y_i = \beta_0 + \boxed{\beta_1} x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i \rightarrow \hat{\beta}_1$$

- Two auxiliary regressions
  - x_{1i} is regressed on all the remaining independent variables

$$x_{1i} = \underline{\gamma_0 + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \cdots + \gamma_K x_{Ki}} + f_i$$

  - $y_i i$ is regressed on all the remaining independent variables

$$y_i = \underline{\alpha_0 + \alpha_{k-1} x_{k-1i} + \alpha_{k+1} x_{k+1i} + \cdots + \alpha_K x_{Ki}} + g_i$$

- $\underline{\tilde{x}_{1i} = x_{1i} - \widehat{x}_{1i}}$ and $\underline{\tilde{y}_i = y_i - \widehat{y}_i}$ residuals from auxiliary regressions
- The parameter β_1 can be rewritten as

$$\beta_1 = \frac{Cov(y_i, x_{1i})}{Var(x_{1i})} = \frac{Cov(\tilde{y}, \tilde{x}_{1i})}{Var(\tilde{x}_{1i})}$$

- $\widehat{\beta}_1$ is a scaled covariance with the actual data or with the residuals

Big picture

- Regression anatomy theorem helps us interpret a single slope coefficient in a multiple regression model by the aforementioned decomposition
- Also, help us understand "OLS" as a "matching estimator" (try to compare observations that are alike in the Xs)

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

OLS Review

Linear algebra review

Law of iterated expectations

OLS basics

Conditional expectation function

"Algebraic" properties of OLS

Properties of OLS estimators

Regression (matrix algebra) with a treatment dummy for the experimental case

Frisch–Waugh–Lovell (FWL) theorem

Regression and causality

Regression and causality

- When is regression causal? Whenever the CEF that regression approximates (or equals if the truth is linear) is causal
- Next: discuss one assumption under which the CEF has a causal interpretation

Potential outcomes - reminder

- A treatment (T) induces two "potential outcomes" for individual i
  - The untreated outcome Y_{0i}
  - The treated outcome Y_{1i}

Potential outcomes - reminder

- A treatment (T) induces two "potential outcomes" for individual i
  - The untreated outcome Y_{0i}
  - The treated outcome Y_{1i}
- The observed outcome

$$\begin{aligned} Y_i &= \begin{cases} Y_{1i} & \text{if } T_i = 1 \\ Y_{0i} & \text{if } T_i = 0 \end{cases} \\ &= Y_{0i} + (Y_{1i} - Y_{0i})T_i \end{aligned}$$

Potential outcomes - reminder

- A treatment (T) induces two "potential outcomes" for individual i
  - The untreated outcome Y_{0i}
  - The treated outcome Y_{1i}
- The observed outcome

$$\begin{aligned} Y_i &= \begin{cases} Y_{1i} & \text{if } T_i = 1 \\ Y_{0i} & \text{if } T_i = 0 \end{cases} \\ &= Y_{0i} + (Y_{1i} - Y_{0i})T_i \end{aligned}$$

- The impact for any individual is $\delta_i = Y_{1i} - Y_{0i}$

Potential outcomes - reminder

- A treatment (T) induces two "potential outcomes" for individual i
  - The untreated outcome Y_{0i}
  - The treated outcome Y_{1i}
- The observed outcome

$$\begin{aligned} Y_i &= \begin{cases} Y_{1i} & \text{if } T_i = 1 \\ Y_{0i} & \text{if } T_i = 0 \end{cases} \\ &= Y_{0i} + (Y_{1i} - Y_{0i})T_i \end{aligned}$$

- The impact for any individual is $\delta_i = Y_{1i} - Y_{0i}$
- Fundamental problem: **Never observe both potential outcomes for the same individual**

We can't just compared treated/untreated individuals

- We observe $Y_i = Y_{0i} + \underbrace{(Y_{1i} - Y_{0i})}_{\delta_i = \text{impact}} T_i$
- If we compare the outcomes of treated and untreated individuals:

$$\underbrace{\mathbb{E}(Y_i|T_i = 1) - \mathbb{E}(Y_i|T_i = 0)}_{\text{Observed difference}} =$$

We can't just compared treated/untreated individuals

- We observe $Y_i = Y_{0i} + \underbrace{(Y_{1i} - Y_{0i})}_{\delta_i = \text{impact}} T_i$
- If we compare the outcomes of treated and untreated individuals:

$$\begin{aligned} \underbrace{\mathbb{E}(Y_i|T_i = 1) - \mathbb{E}(Y_i|T_i = 0)}_{\text{Observed difference}} &= \mathbb{E}(Y_{1i}|T_i = 1) - \mathbb{E}(Y_{0i}|T_i = 1) + \\ & \quad \mathbb{E}(Y_{0i}|T_i = 1) - \mathbb{E}(Y_{0i}|T_i = 0) \end{aligned}$$

We can't just compared treated/untreated individuals

- We observe $Y_i = Y_{0i} + \underbrace{(Y_{1i} - Y_{0i})}_{\delta_i = \text{impact}} T_i$
- If we compare the outcomes of treated and untreated individuals:

$$\begin{aligned}\underbrace{\mathbb{E}(Y_i|T_i=1) - \mathbb{E}(Y_i|T_i=0)}_{\text{Observed difference}} &= \mathbb{E}(Y_{1i}|T_i=1) - \mathbb{E}(Y_{0i}|T_i=1) + \\ &\quad \mathbb{E}(Y_{0i}|T_i=1) - \mathbb{E}(Y_{0i}|T_i=0) \\ &= \underbrace{\mathbb{E}(Y_{1i}|T_i=1) - \mathbb{E}(Y_{0i}|T_i=1)}_{\text{average treatment effect on the treated}} + \\ &\quad \underbrace{\mathbb{E}(Y_{0i}|T_i=1) - \mathbb{E}(Y_{0i}|T_i=0)}_{\text{selection bias}}\end{aligned}$$

Unconfoundedness

Assumption (Unconfoundedness)
$(Y_{1i}, Y_{0i}) \coprod T_i \mid X_i$

In words:

1. Once we condition on observable characteristics X_i, the treatment T_i is as good as randomly assigned
2. Put differently, within the group of individuals with the same characteristics x_i, we have a randomized experiment
3. Yet another way of saying it is that conditional on x_i, the selection bias disappears

Uncounfoundedness is **fundamentally untestable** and should always be discussed!

Overlap

- In order to exploit the unconfoundedness assumption, for all values of x_i we need to have both treated and untreated units
- Otherwise, either no treatment or no control group for some values of x_i
- *Propensity score*, which gives us the probability of $T_i = 1$ given $X_i = x$

$$p(x) = P(T_i = 1 \mid x_i = x)$$

- $p(x) = 1$ means that there are no control units (everyone is treated)
- $p(x) = 0$ means that there are no treated units (no one is treated)

Assumption (Overlap)
$0 < p(x) < 1$ *for all* x

- In contrast to unconfoundedness, overlap is testable since we can compute $p(x)$ from the data

Identification under unconfoundedness

- How can we identify treatment effects under unconfoundedness?
- Define the conditional mean difference as

$$\boxed{\delta_x} = E[Y_i \mid T_i = 1, x_i = x] - E[Y_i \mid T_i = 0, x_i = x]$$

- Conditional on $x_i = x$, we can use the same arguments as the experimental case:

$$\begin{aligned}\boxed{\delta_x} &= E[Y_i \mid T_i = 1, x_i = x] - E[Y_i \mid T_i = 0, x_i = x] \\ &= E[Y_{1i} \mid T_i = 1, x_i = x] - E[Y_{0i} \mid T_i = 0, x_i = x] \\ &= E[Y_{1i} \mid x_i = x] - E[Y_{0i} \mid x_i = x] \\ &= E[Y_{1i} - Y_{0i} \mid x_i = x]\end{aligned}$$

- The second equality is well-defined for every x by the overlap assumption
- The third equality is by unconfoundedness

Identification under unconfoundedness

- δ_x is the ATE for individuals with charateristics $x_i = x$
- We can get the (unconditional) ATE as

$$\begin{aligned}\delta = E[\delta_{x_i}] &= E[E[Y_{1i} - Y_{0i} \mid x_i]] \\ &= E[Y_{1i} - Y_{0i}]\end{aligned}$$

Discrete covariates

- The results so far are rather abstract.
- It is easier to understand the results with discrete covariates x_i
- In this case,

$$\text{ATE} = E[Y_{1i} - Y_{0i}] = \sum_x \delta_x P(x_i = x) = E(\delta_x)$$

- Suppose x_i is binary. In this case the formula becomes:

$$\begin{aligned}E[Y_{1i} - Y_{0i}] = &\quad \underbrace{\boxed{\delta_{x_i=1}}}_{\text{mean diff. in group with } x_i = 1} \cdot \underbrace{\boxed{P(x_i = 1)}}_{\text{fract. with } x_i = 1} \\ &+ \underbrace{\boxed{\delta_{x_i=0}}}_{\text{mean diff. in group with } x_i = 0} \cdot \underbrace{\boxed{P(x_i = 0)}}_{\text{fract. with } x_i = 0}\end{aligned}$$

An example: causal effect of gender on admissions

Major	Admissions	Admit	Deny	Total
A	Men	400	200	600
B	Men	100	300	400
A	Women	50	50	100
B	Women	300	100	400

- T_i is gender ($T_i = 1$ if male and $T_i = 0$ if female)
- $x_i = M_i$ is choice of major
- Unconfoundedness: gender is independent of admission outcomes conditional on major

An example: causal effect of gender on admissions

50% — 75% → Δ 25pp

δ 50% +

$$\boxed{\delta_A} = 400/600 - 50/100 = 0.166$$
$$\boxed{\delta_B} = 100/400 - 300/400 = -0.5$$
$$P(M_i = A) = (600 + 100)/(1500) = 0.466$$
$$P(M_i = B) = (400 + 400)/1500 = 0.533$$
$$\delta = ATE = 0.167 \cdot 0.47 + (-0.5) \cdot 0.533 = -0.19$$

Regression and causality

- Under unconfoundedness, the CEF $\overset{\delta x}{E[Y_i \mid T_i, X_i]}$ has a causal interpretation
- Thus, a linear regression model has an (approximate) causal interpretation under unconfoundedness.

Regression and causality

- If the population regression model is:

$$Y_i = \theta T_i + X_i'\beta + u_i$$

- Then $\boxed{\delta_x = \theta}$ is constant across x and thus $\underset{\hat{\theta}}{ATE = \theta}$

What if δ_x is not constant?

Major	Admissions	Admit	Deny	Total
A	Men	400	200	600
B	Men	100	300	400
A	Women	50	50	100
B	Women	300	100	400

Simulations!

```
#Lets create the data
#600 men in A,  400 men in B, 100 women in A, 400 women in B
#code major A as 1 and B as zeroq()
Data=data.frame(AdminExact=c(rep(1,400),rep(0,200),
                             rep(1,100),rep(0,300),
                             rep(1,50),rep(0,50),
                             rep(1,300),rep(0,100)),
                Major=c(rep(1,600),rep(0,400),rep(1,100),rep(0,400)),
                Treatment=c(rep(1,600),rep(1,400),rep(0,100),rep(0,400)))


summary(lm(AdminExact~Treatment+Major,subset=Major==1,data=Data))
summary(lm(AdminExact~Treatment+Major,subset=Major==0,data=Data))
summary(lm(AdminExact~Treatment+Major,data=Data))
```

What if δ_x is not constant?

- Estimate

$$Y_i = \theta T_i + X_i'\beta + u_i$$

- Regression yields $\widehat{\theta} = -0.3 \neq ATE = -0.19$
- In general, we get the following weighted average

$$\boxed{\theta = \frac{\mathbb{E}(\sigma^2_{T_i|X_i}\delta_x)}{\mathbb{E}(\sigma^2_{T_i|X_i})}}$$

- Regression produces a treatment-variance weighted average of δ_x (proof in Angrist and Pischke MHE 3.3.1)
- In our case $\sigma^2_{T_i|X_i} = \boxed{P(T_i = 1|X_i)(1 - P(T_i = 1|X_i))}$
  - $\sigma^2_{T_i|Major=A} = \frac{600}{700}\frac{100}{700}$
  - $\sigma^2_{T_i|Major=B} = \frac{400}{800}\frac{400}{800}$

What if δ_x is not constant?

- Therefore:

$$\begin{aligned}\widehat{\theta} &= \frac{\overbrace{\delta_A}^{0.166}\overbrace{\sigma^2_{T_i|Major=A}}^{\frac{600}{700}\frac{100}{700}}\overbrace{P(Major=A)}^{\frac{700}{1500}} + \overbrace{\delta_B}^{-0.5}\overbrace{\sigma^2_{T_i|Major=B}}^{\frac{400}{800}\frac{400}{800}}\overbrace{P(Major=B)}^{\frac{800}{1500}}}{\boxed{\sigma^2_{T_i|Major=A}P(Major=A) + \sigma^2_{T_i|Major=B}P(Major=B)}}\\ &= -0.3\end{aligned}$$

Big picture

- Beware of what OLS gives you
- Still causal interpretation, even if δ_x is not constant
- Weighted average of different δ_x
- Weights depend on the variance!

Beyond regression

- Regression is only one method to obtain causal effects under uncounfoundedness
- Other popular methods are: matching and inverse probability weighting
  - Assumption are the same, they generally yield similar results (but implicit weights are different)
- A great review is: *Recent Developments in the Econometrics of Program Evaluation* by Imbens and Wooldrige (2009)
- Check this out: `http://www.nber.org/minicourse3.html`

Some important remarks (based on Cyrus Samii's lecture notes)

> *For most researchers, the math obscures the assumptions. Without an experiment, a natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive. (Sekhon, 2009, p. 503)*

- At the end of the day, OLS (and other matching/weighting estimators) "mop up" imbalances that makes CIA plausible
- Thought experiment necessary to test CIA:
  - How could it be that two units that are identical with respect to all meaningful background factors nonetheless receive different treatment?
- Your answer to this question is your source of identification