Gov 50: 25. Inference for Linear Regression

Matthew Blackwell

Harvard University

Roadmap

1. Inference for linear regression
2. Presenting OLS regressions
3. Wrapping up the class

1/ Inference for linear regression

Data

- Do political institutions promote economic development?

Data

- Do political institutions promote economic development?
    - Famous paper on this: Acemoglu, Johnson, and Robinson (2001)

Data

- Do political institutions promote economic development?
    - Famous paper on this: Acemoglu, Johnson, and Robinson (2001)
    - Relationship between strength of property rights in a country and GDP.

Data

- Do political institutions promote economic development?
  - Famous paper on this: Acemoglu, Johnson, and Robinson (2001)
  - Relationship between strength of property rights in a country and GDP.
- Data:

Data

- Do political institutions promote economic development?
    - Famous paper on this: Acemoglu, Johnson, and Robinson (2001)
    - Relationship between strength of property rights in a country and GDP.
- Data:

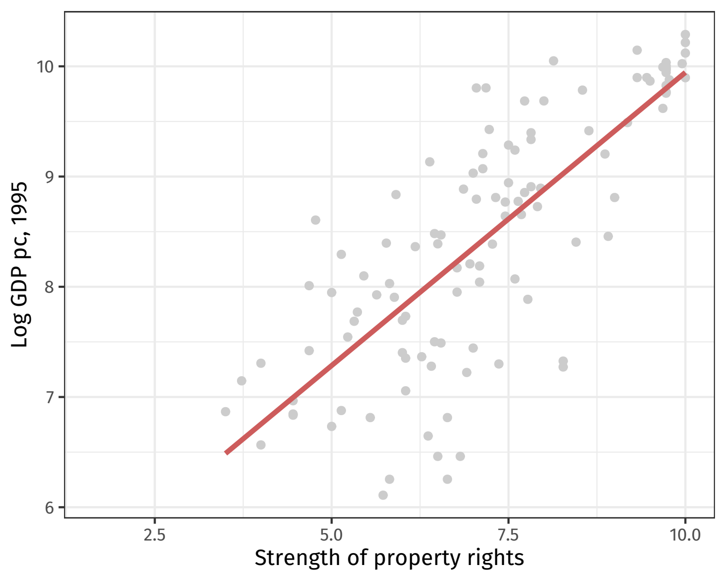
Name	Description
`shortnam`	three-letter country code
`africa`	indicator for if the country is in Africa
`asia`	indicator for if country is in Asia
`avexpr`	strength of property rights (protection against expropriation)
`logpgp95`	log GDP per capita

Loading the data

```r
library(gov50data)
head(ajr)
```

```
## # A tibble: 6 x 15
##   short~1 africa lat_a~2 malfa~3 avexpr logpg~4 logem4  asia
##   <chr>    <dbl>   <dbl>   <dbl>  <dbl>   <dbl>  <dbl> <dbl>
## 1 AFG          0   0.367 0.00372 NA      NA       4.54     1
## 2 AGO          1   0.137 0.950     5.36    7.77   5.63     0
## 3 ARE          0   0.267 0.0123    7.18    9.80  NA        1
## 4 ARG          0   0.378 0         6.39    9.13   4.23     0
## 5 ARM          0   0.444 0       NA        7.68  NA        1
## 6 AUS          0   0.300 0         9.32    9.90   2.15     0
## # ... with 7 more variables: yellow <dbl>, baseco <dbl>,
## #   leb95 <dbl>, imr95 <dbl>, meantemp <dbl>,
## #   lt100km <dbl>, latabs <dbl>, and abbreviated variable
## #   names 1: shortnam, 2: lat_abst, 3: malfal94,
## #   4: logpgp95
```

AJR scatterplot



Simple linear regression model

- We are going to assume a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Simple linear regression model

- We are going to assume a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Data:

Simple linear regression model

- We are going to assume a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Data:
    - Dependent variable: Y_i

Simple linear regression model

- We are going to assume a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Data:
    - Dependent variable: Y_i
    - Independent variable: X_i

Simple linear regression model

- We are going to assume a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Data:
    - Dependent variable: Y_i
    - Independent variable: X_i
- Population parameters:

Simple linear regression model

- We are going to assume a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Data:
  - Dependent variable: Y_i
  - Independent variable: X_i
- Population parameters:
  - Population intercept: β_0

Simple linear regression model

- We are going to assume a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Data:
  - Dependent variable: Y_i
  - Independent variable: X_i
- Population parameters:
  - Population intercept: β_0
  - Population slope: β_1

Simple linear regression model

- We are going to assume a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Data:
    - Dependent variable: Y_i
    - Independent variable: X_i
- Population parameters:
    - Population intercept: β_0
    - Population slope: β_1
- Error/disturbance: ϵ_i

Simple linear regression model

- We are going to assume a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Data:
  - Dependent variable: Y_i
  - Independent variable: X_i
- Population parameters:
  - Population intercept: β_0
  - Population slope: β_1
- Error/disturbance: ϵ_i
  - Represents all unobserved error factors influencing Y_i other than X_i.

Least squares

- How do we figure out the best line to draw?

Least squares

- How do we figure out the best line to draw?
    - Alt question: how do we figure out β_0 and β_1?

Least squares

- How do we figure out the best line to draw?
  - Alt question: how do we figure out β_0 and β_1?
  - $(\hat{\beta}_0, \hat{\beta}_1)$: estimated coefficients.

Least squares

- How do we figure out the best line to draw?
  - Alt question: how do we figure out β_0 and β_1?
  - $(\hat{\beta}_0, \hat{\beta}_1)$: estimated coefficients.
  - $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$: predicted/fitted value.

Least squares

- How do we figure out the best line to draw?
    - Alt question: how do we figure out β_0 and β_1?
    - $(\hat{\beta}_0, \hat{\beta}_1)$: estimated coefficients.
    - $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$: predicted/fitted value.
    - $\hat{\epsilon}_i = Y_i - \widehat{Y}$: residual.

Least squares

- How do we figure out the best line to draw?
    - Alt question: how do we figure out β_0 and β_1?
    - $(\hat{\beta}_0, \hat{\beta}_1)$: estimated coefficients.
    - $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$: predicted/fitted value.
    - $\hat{\epsilon}_i = Y_i - \widehat{Y}$: residual.
- Get these estimates by the **least squares method**.

Least squares

- How do we figure out the best line to draw?
  - Alt question: how do we figure out β_0 and β_1?
  - $(\widehat{\beta}_0, \widehat{\beta}_1)$: estimated coefficients.
  - $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$: predicted/fitted value.
  - $\widehat{\epsilon}_i = Y_i - \widehat{Y}$: residual.
- Get these estimates by the **least squares method**.
- Minimize the **sum of the squared residuals** (SSR):

$$\text{SSR} = \sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

Estimators

- Least squares is an **estimator**

Estimators

- Least squares is an **estimator**
    - it's a machine that we plug data into and we get out estimates.

Estimators

- Least squares is an **estimator**
  - it's a machine that we plug data into and we get out estimates.

$$\boxed{\text{OLS}}$$

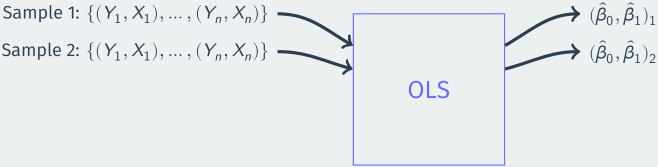
Estimators

- Least squares is an **estimator**
    - it's a machine that we plug data into and we get out estimates.

Sample 1: $\{(Y_1, X_1), \ldots, (Y_n, X_n)\} \rightarrow \boxed{\text{OLS}} \rightarrow (\hat{\beta}_0, \hat{\beta}_1)_1$

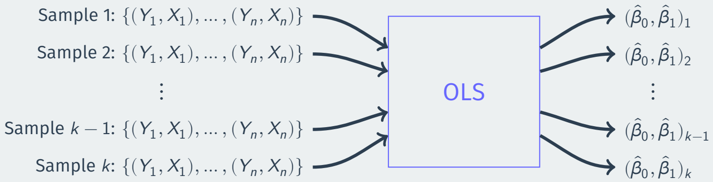
Estimators

- Least squares is an **estimator**
    - it's a machine that we plug data into and we get out estimates.



Estimators

- Least squares is an **estimator**
    - it's a machine that we plug data into and we get out estimates.



Estimators

- Least squares is an **estimator**
  - it's a machine that we plug data into and we get out estimates.

Sample 1: $\{(Y_1, X_1), \dots, (Y_n, X_n)\}$ → OLS → $(\hat{\beta}_0, \hat{\beta}_1)_1$

Sample 2: $\{(Y_1, X_1), \dots, (Y_n, X_n)\}$ → OLS → $(\hat{\beta}_0, \hat{\beta}_1)_2$

$\vdots$

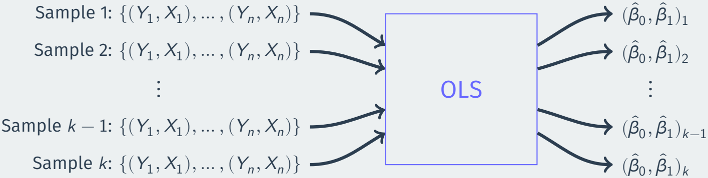
Sample $k-1$: $\{(Y_1, X_1), \dots, (Y_n, X_n)\}$ → OLS → $(\hat{\beta}_0, \hat{\beta}_1)_{k-1}$

Sample k: $\{(Y_1, X_1), \dots, (Y_n, X_n)\}$ → OLS → $(\hat{\beta}_0, \hat{\beta}_1)_k$

- Just like the sample mean or difference in sample means

Estimators

- Least squares is an **estimator**
  - it's a machine that we plug data into and we get out estimates.



- Just like the sample mean or difference in sample means
- $\rightsquigarrow$ sampling distribution with a standard error, etc.

Simulation procedure

- Let's take a simulation approach to demonstrate:

Simulation procedure

- Let's take a simulation approach to demonstrate:
    - Pretend that the AJR data represents the population of interest

Simulation procedure

- Let's take a simulation approach to demonstrate:
  - Pretend that the AJR data represents the population of interest
  - See how the line varies from sample to sample

Simulation procedure

- Let's take a simulation approach to demonstrate:
  - Pretend that the AJR data represents the population of interest
  - See how the line varies from sample to sample

1. Randomly sample $n = 30$ countries w/ replacement using `sample()`

Simulation procedure

- Let's take a simulation approach to demonstrate:
  - Pretend that the AJR data represents the population of interest
  - See how the line varies from sample to sample

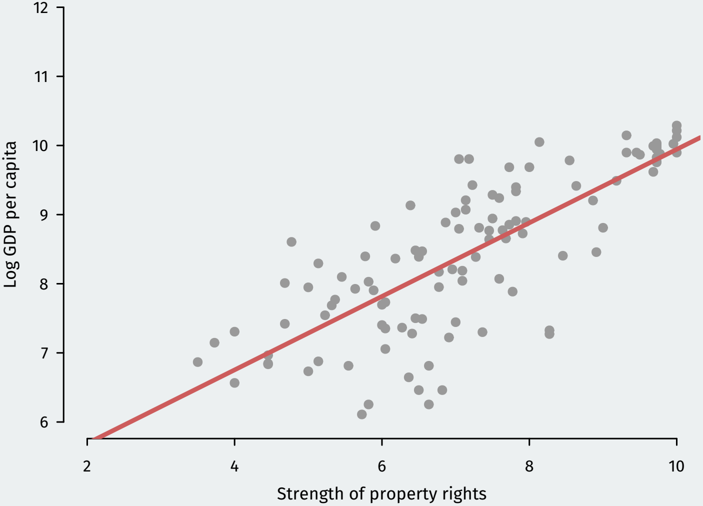
1. Randomly sample $n = 30$ countries w/ replacement using `sample()`
2. Use `lm()` to calculate the OLS estimates of the slope and intercept

Simulation procedure

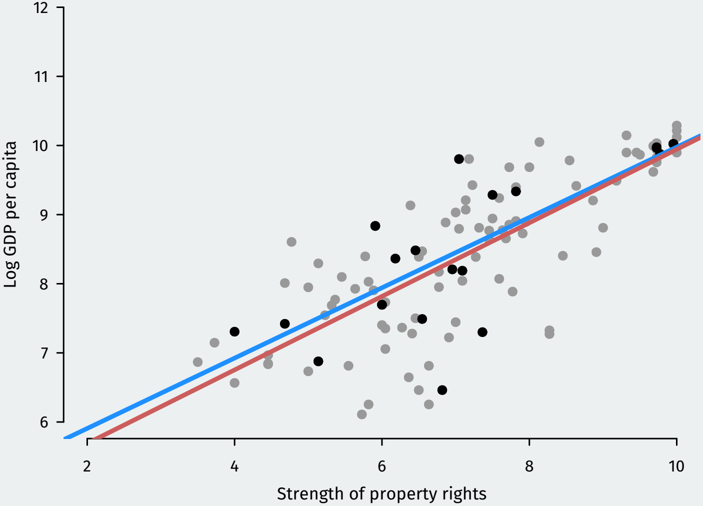
- Let's take a simulation approach to demonstrate:
  - Pretend that the AJR data represents the population of interest
  - See how the line varies from sample to sample

1. Randomly sample $n = 30$ countries w/ replacement using `sample()`
2. Use `lm()` to calculate the OLS estimates of the slope and intercept
3. Plot the estimated regression line

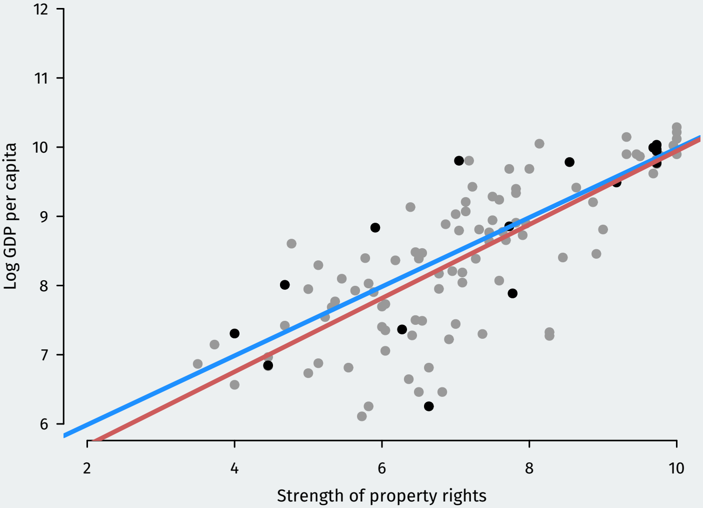
Population regression



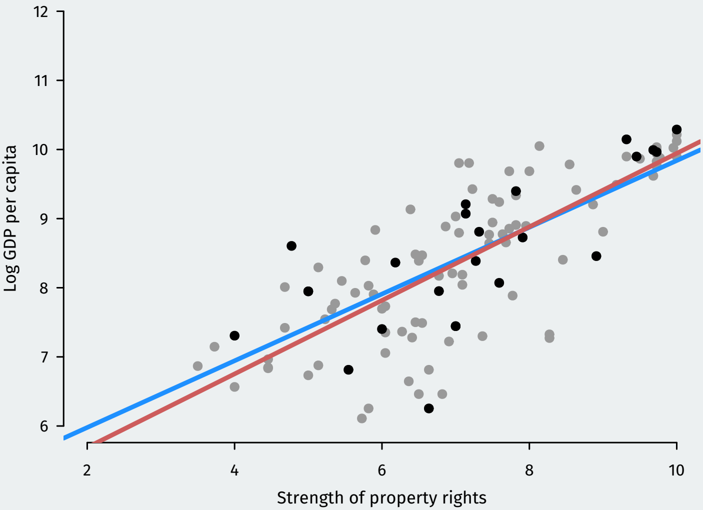
Randomly sample from AJR



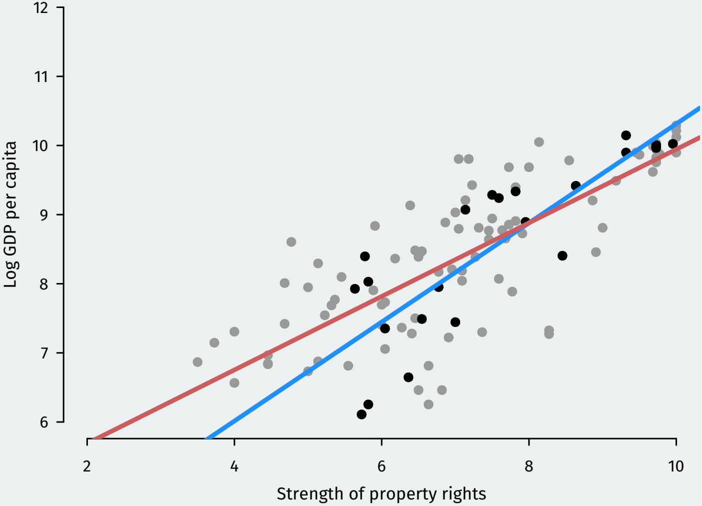
Randomly sample from AJR



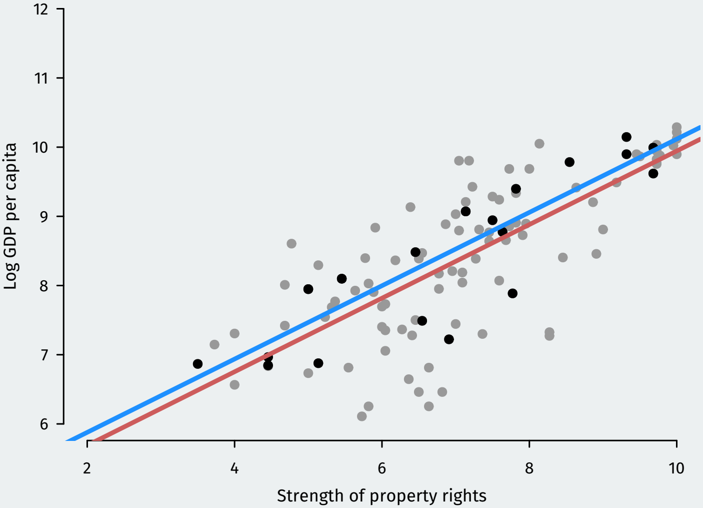
Randomly sample from AJR



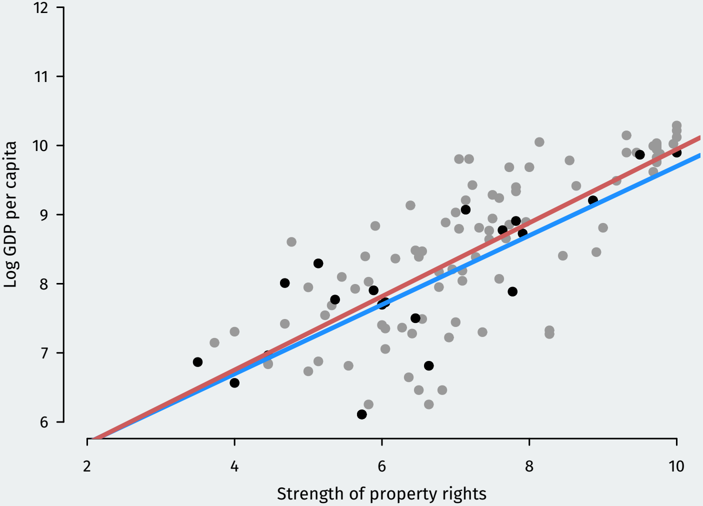
Randomly sample from AJR



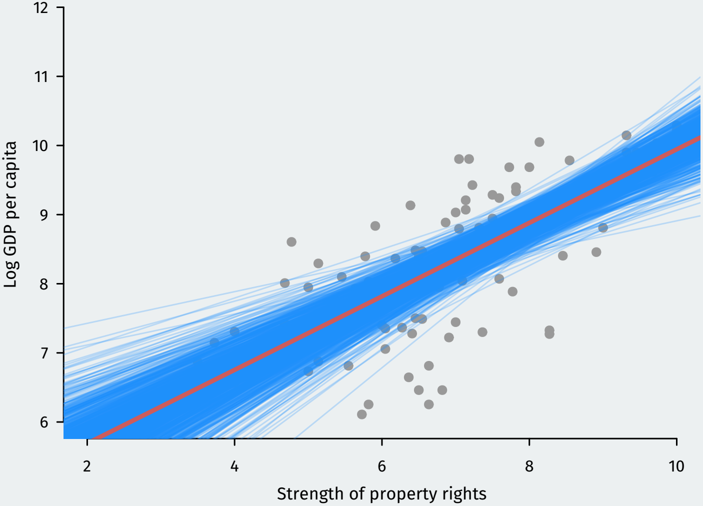
Randomly sample from AJR



Randomly sample from AJR

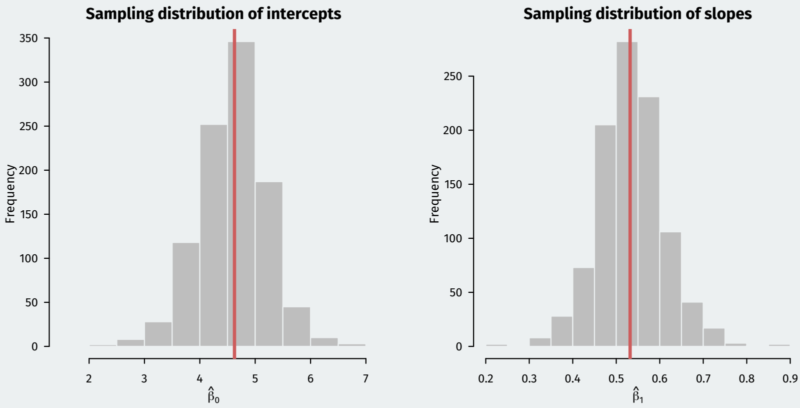


Randomly sample from AJR



Sampling distribution of OLS

- Estimated slope and intercept vary between samples, centered on truth.



Properties of OLS

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables

Properties of OLS

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables
    - Are they on average equal to the true values (bias)?

Properties of OLS

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables
    - Are they on average equal to the true values (bias)?
    - How spread out are they around their center (variance)?

Properties of OLS

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables
    - Are they on average equal to the true values (bias)?
    - How spread out are they around their center (variance)?
- Under minimal conditions, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for the population line of best fit, but...

Properties of OLS

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables
  - Are they on average equal to the true values (bias)?
  - How spread out are they around their center (variance)?
- Under minimal conditions, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for the population line of best fit, but...
  - This might be misleading if the true relationship is nonlinear.

Properties of OLS

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables
    - Are they on average equal to the true values (bias)?
    - How spread out are they around their center (variance)?
- Under minimal conditions, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for the population line of best fit, but...
    - This might be misleading if the true relationship is nonlinear.
    - May not represent a causal effect unless causal assumptions hold.

Standard errors of OLS

R will also calculate an estimate of the standard error: $\widehat{\text{SE}}(\hat{\beta}_1)$

Standard errors of OLS

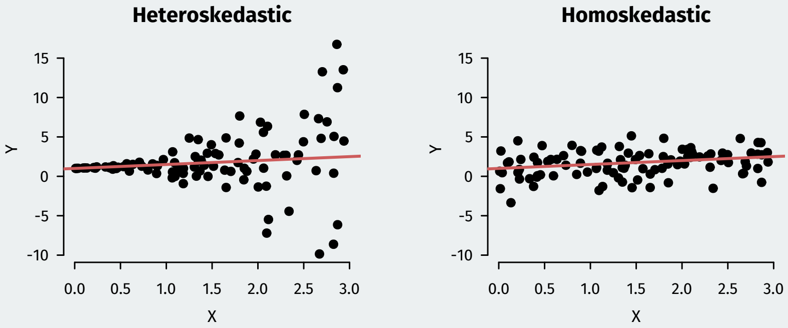
R will also calculate an estimate of the standard error: $\widehat{\text{SE}}(\hat{\beta}_1)$

Default estimators for the SEs assume **homoskedasticity** or that the spread around the regression line is the same for all values of the independent variables.

Standard errors of OLS

R will also calculate an estimate of the standard error: $\widehat{\text{SE}}(\hat{\beta}_1)$

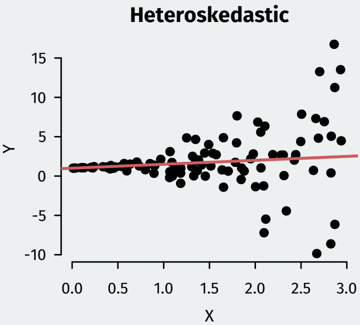
Default estimators for the SEs assume **homoskedasticity** or that the spread around the regression line is the same for all values of the independent variables.

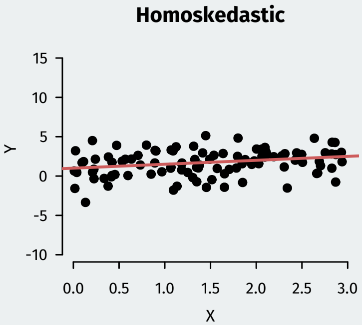


Standard errors of OLS

R will also calculate an estimate of the standard error: $\widehat{\text{SE}}(\hat{\beta}_1)$

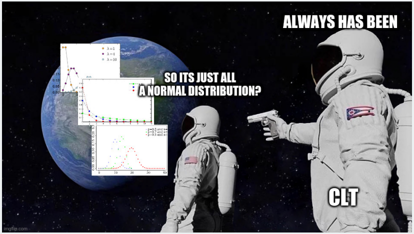
Default estimators for the SEs assume **homoskedasticity** or that the spread around the regression line is the same for all values of the independent variables.





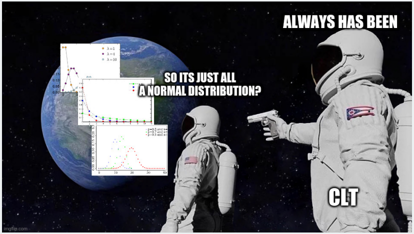
Relatively easy fixes exist, but beyond the scope of this class.

Tests and CIs for regression



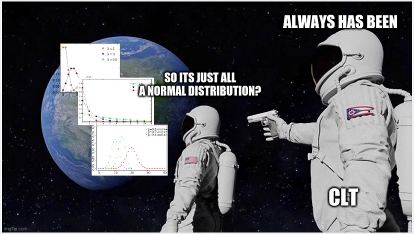
- $(\hat{\beta}_0, \hat{\beta}_1)$ can be written as weighted averages of the outcome...

Tests and CIs for regression



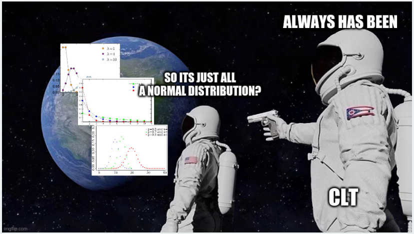
- $(\hat{\beta}_0, \hat{\beta}_1)$ can be written as weighted averages of the outcome…
  - Which means they follow the Central Limit Theorem!

Tests and CIs for regression



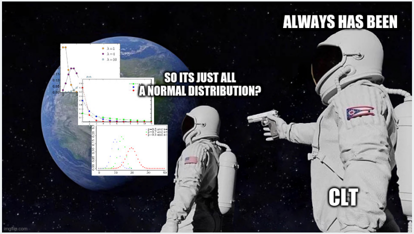
- $(\hat{\beta}_0, \hat{\beta}_1)$ can be written as weighted averages of the outcome...
    - Which means they follow the Central Limit Theorem!
- BAM! 95% confidence intervals: $\hat{\beta}_1 \pm 1.96 \times \widehat{\text{SE}}(\hat{\beta}_1)$

Tests and CIs for regression



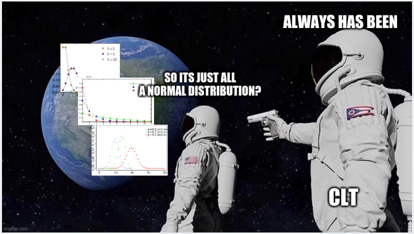
- $(\hat{\beta}_0, \hat{\beta}_1)$ can be written as weighted averages of the outcome…
    - Which means they follow the Central Limit Theorem!
- BAM! 95% confidence intervals: $\hat{\beta}_1 \pm 1.96 \times \widehat{\text{SE}}(\hat{\beta}_1)$
- BOOM! Hypothesis tests:

Tests and CIs for regression



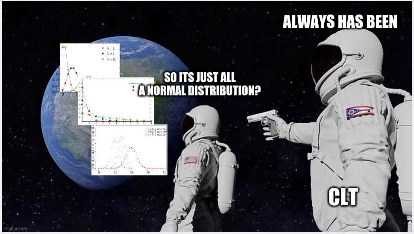
- $(\hat{\beta}_0, \hat{\beta}_1)$ can be written as weighted averages of the outcome...
    - Which means they follow the Central Limit Theorem!
- BAM! 95% confidence intervals: $\hat{\beta}_1 \pm 1.96 \times \widehat{\text{SE}}(\hat{\beta}_1)$
- BOOM! Hypothesis tests:
    - Null hypothesis: $H_0 : \beta_1 = \beta_1^*$

Tests and CIs for regression



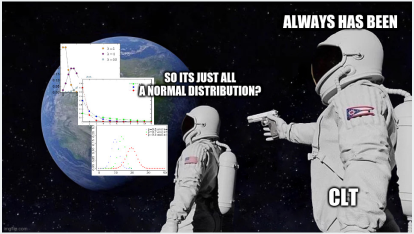
- $(\hat{\beta}_0, \hat{\beta}_1)$ can be written as weighted averages of the outcome...
    - Which means they follow the Central Limit Theorem!
- BAM! 95% confidence intervals: $\hat{\beta}_1 \pm 1.96 \times \widehat{\text{SE}}(\hat{\beta}_1)$
- BOOM! Hypothesis tests:
    - Null hypothesis: $H_0 : \beta_1 = \beta_1^*$
    - Test statistic: $\frac{\hat{\beta}_1 - \beta_1^*}{\widehat{\text{SE}}(\hat{\beta}_1)} \sim N(0, 1)$

Tests and CIs for regression



- $(\hat{\beta}_0, \hat{\beta}_1)$ can be written as weighted averages of the outcome...
  - Which means they follow the Central Limit Theorem!
- BAM! 95% confidence intervals: $\hat{\beta}_1 \pm 1.96 \times \widehat{\text{SE}}(\hat{\beta}_1)$
- BOOM! Hypothesis tests:
  - Null hypothesis: $H_0 : \beta_1 = \beta_1^*$
  - Test statistic: $\frac{\hat{\beta}_1 - \beta_1^*}{\widehat{\text{SE}}(\hat{\beta}_1)} \sim N(0, 1)$
  - Usual test is of $\beta_1 = 0$.

Tests and CIs for regression



- $(\hat{\beta}_0, \hat{\beta}_1)$ can be written as weighted averages of the outcome...
    - Which means they follow the Central Limit Theorem!
- BAM! 95% confidence intervals: $\hat{\beta}_1 \pm 1.96 \times \widehat{\text{SE}}(\hat{\beta}_1)$
- BOOM! Hypothesis tests:
    - Null hypothesis: $H_0 : \beta_1 = \beta_1^*$
    - Test statistic: $\frac{\hat{\beta}_1 - \beta_1^*}{\widehat{\text{SE}}(\hat{\beta}_1)} \sim N(0, 1)$
    - Usual test is of $\beta_1 = 0$.
    - $\hat{\beta}_1$ is **statistically significant** if its p-value from this test is below some threshold (usually 0.05)

```r
ajr.reg <- lm(logpgp95 ~ avexpr, data = ajr)
summary(ajr.reg)
```

```
##
## Call:
## lm(formula = logpgp95 ~ avexpr, data = ajr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.902 -0.316  0.138  0.422  1.441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.6261     0.3006    15.4   <2e-16 ***
## avexpr        0.5319     0.0406    13.1   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.718 on 109 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.611,  Adjusted R-squared:  0.608
## F-statistic:  171 on 1 and 109 DF,  p-value: <2e-16
```

Using broom with regression

```r
library(broom)
tidy(ajr.reg)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    4.63     0.301       15.4 4.28e-29
## 2 avexpr         0.532    0.0406      13.1 4.16e-24
```

Multiple regression

- Correlation doesn’t imply causation

Multiple regression

- Correlation doesn't imply causation
- Omitted variables $\rightsquigarrow$ violation of exogeneity

Multiple regression

- Correlation doesn't imply causation
- Omitted variables $\rightsquigarrow$ violation of exogeneity
- You can adjust for multiple confounding variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

Multiple regression

- Correlation doesn't imply causation
- Omitted variables $\rightsquigarrow$ violation of exogeneity
- You can adjust for multiple confounding variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

- Interpretation of β_j: an increase in the outcome associated with a one-unit increase in X_{ij} when other variables don't change their values

Multiple regression

- Correlation doesn't imply causation
- Omitted variables $\rightsquigarrow$ violation of exogeneity
- You can adjust for multiple confounding variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

- Interpretation of β_j: an increase in the outcome associated with a one-unit increase in X_{ij} when other variables don't change their values
- Inference:

Multiple regression

- Correlation doesn't imply causation
- Omitted variables $\rightsquigarrow$ violation of exogeneity
- You can adjust for multiple confounding variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

- Interpretation of β_j: an increase in the outcome associated with a one-unit increase in X_{ij} when other variables don't change their values
- Inference:
  - Confidence intervals constructed exactly the same for $\hat{\beta}_j$

Multiple regression

- Correlation doesn't imply causation
- Omitted variables $\rightsquigarrow$ violation of exogeneity
- You can adjust for multiple confounding variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

- Interpretation of β_j: an increase in the outcome associated with a one-unit increase in X_{ij} when other variables don't change their values
- Inference:
  - Confidence intervals constructed exactly the same for $\hat{\beta}_j$
  - Hypothesis tests done exactly the same for $\hat{\beta}_j$

Multiple regression

- Correlation doesn't imply causation
- Omitted variables $\rightsquigarrow$ violation of exogeneity
- You can adjust for multiple confounding variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

- Interpretation of β_j: an increase in the outcome associated with a one-unit increase in X_{ij} when other variables don't change their values
- Inference:
  - Confidence intervals constructed exactly the same for $\hat{\beta}_j$
  - Hypothesis tests done exactly the same for $\hat{\beta}_j$
  - $\rightsquigarrow$ interpret p-values the same as before.

Using knitr::kable to produce tables

```
ajr.multreg <- lm(logpgp95 ~ avexpr + lat_abst + asia + africa, data = ajr)
tidy(ajr.multreg) |>
  knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	5.840	0.339	17.239	0.000
avexpr	0.394	0.050	7.843	0.000
lat_abst	0.312	0.444	0.703	0.484
asia	-0.170	0.153	-1.108	0.270
africa	-0.930	0.165	-5.628	0.000

2/ Presenting OLS regressions

Regression tables

- In papers, you'll often find regression tables that have several models.

Regression tables

- In papers, you'll often find regression tables that have several models.
- Each column is a different regression:

Regression tables

- In papers, you'll often find regression tables that have several models.
- Each column is a different regression:
  - Might differ by independent variables, dependent variables, sample, etc.

Regression tables

- In papers, you'll often find regression tables that have several models.
- Each column is a different regression:
    - Might differ by independent variables, dependent variables, sample, etc.
- Standard errors, p-values, sample size, and R^2 may be reported as well.

AJR regression table

TABLE 2—OLS REGRESSIONS

	Whole world (1)	Base sample (2)	Whole world (3)	Whole world (4)	Base sample (5)	Base sample (6)	Whole world (7)	Base sample (8)
	Dependent variable is log GDP per capita in 1995						Dependent variable is log output per worker in 1988	
Average protection against expropriation risk, 1985–1995	0.54 (0.04)	0.52 (0.06)	0.47 (0.06)	0.43 (0.05)	0.47 (0.06)	0.41 (0.06)	0.45 (0.04)	0.46 (0.06)
Latitude			0.89 (0.49)	0.37 (0.51)	1.60 (0.70)	0.92 (0.63)		
Asia dummy				−0.62 (0.19)		−0.60 (0.23)		
Africa dummy				−1.00 (0.15)		−0.90 (0.17)		
"Other" continent dummy				−0.25 (0.20)		−0.04 (0.32)		
R^2	0.62	0.54	0.63	0.73	0.56	0.69	0.55	0.49
Number of observations	110	64	110	110	64	64	108	61

modelsummary() to produce tables

We can use modelsummary() to produce a table. It takes a list of outputs from lm and aligns them in the correct way.

```
modelsummary::modelsummary(list(ajr.reg, ajr.multreg))
```

Output

```r
modelsummary::modelsummary(list(ajr.reg, ajr.multreg))
```

	Model 1	Model 2
(Intercept)	4.626	5.840
	(0.301)	(0.339)
avexpr	0.532	0.394
	(0.041)	(0.050)
lat_abst		0.312
		(0.444)
asia		−0.170
		(0.153)
africa		−0.930
		(0.165)
Num.Obs.	111	111
R2	0.611	0.713
R2 Adj.	0.608	0.703
AIC	245.4	217.6
BIC	253.5	233.8
Log.Lik.	−119.709	−102.795
RMSE	0.71	0.61

Cleaning up the goodness of fit statistics

```
modelsummary::modelsummary(
  list(ajr.reg, ajr.multreg),
  gof_map = c("nobs", "r.squared", "adj.r.squared"))
```

	Model 1	Model 2
(Intercept)	4.626	5.840
	(0.301)	(0.339)
avexpr	0.532	0.394
	(0.041)	(0.050)
lat_abst		0.312
		(0.444)
asia		−0.170
		(0.153)
africa		−0.930
		(0.165)
Num.Obs.	111	111
R2	0.611	0.713
R2 Adj.	0.608	0.703

Cleaning up the variable names

We can also map the variable names to more readable names using the `coef_map` argument. But first, we should do the mapping in a vector. Any term omitted from this vector will be omitted from the table

```
var_labels <- c(
  "avexpr" = "Avg. Expropriation Risk",
  "lat_abst" = "Abs. Value of Latitude",
  "asia" = "Asian country",
  "africa" = "African country"
)
var_labels
```

```
##                    avexpr                  lat_abst
## "Avg. Expropriation Risk"  "Abs. Value of Latitude"
##                      asia                    africa
##           "Asian country"         "African country"
```

Nice table

```
modelsummary::modelsummary(
  list(ajr.reg, ajr.multreg),
  coef_map = var_labels,
  gof_map = c("nobs", "r.squared", "adj.r.squared"))
```

	Model 1	Model 2
Avg. Expropriation Risk	0.532	0.394
	(0.041)	(0.050)
Abs. Value of Latitude		0.312
		(0.444)
Asian country		−0.170
		(0.153)
African country		−0.930
		(0.165)
Num.Obs.	111	111
R2	0.611	0.713
R2 Adj.	0.608	0.703

3/ Wrapping up the class

Big takeaways

Important takeaways from the course:

1. Data wrangling and data visualizations are really important skills that you now have!

Big takeaways

Important takeaways from the course:

1. Data wrangling and data visualizations are really important skills that you now have!
2. Causality is hugely important in the world but difficult to establish.

Big takeaways

Important takeaways from the course:

1. Data wrangling and data visualizations are really important skills that you now have!
2. Causality is hugely important in the world but difficult to establish.
3. Really important to understand and assess statistical uncertainty when working with data.

I’m really proud of you!

You’ve come a long way! Hopefully the tools you learned in this course will help you throughout your life and career!

What next?

- Gov 51 with Naijia Liu:

What next?

- Gov 51 with Naijia Liu:
    - A more in-depth review of some ideas from Gov 50 including causality and regression plus new models (maybe some machine learning).

What next?

- Gov 51 with Naijia Liu:
  - A more in-depth review of some ideas from Gov 50 including causality and regression plus new models (maybe some machine learning).
  - Really helpful for students looking to write senior theses.

What next?

- Gov 51 with Naijia Liu:
    - A more in-depth review of some ideas from Gov 50 including causality and regression plus new models (maybe some machine learning).
    - Really helpful for students looking to write senior theses.
- Only need 3 more classes to finish the data science track in Gov!

What next?

- Gov 51 with Naijia Liu:
  - A more in-depth review of some ideas from Gov 50 including causality and regression plus new models (maybe some machine learning).
  - Really helpful for students looking to write senior theses.
- Only need 3 more classes to finish the data science track in Gov!
- More theoretical stats side: Stat 110/111

What next?

- Gov 51 with Naijia Liu:
    - A more in-depth review of some ideas from Gov 50 including causality and regression plus new models (maybe some machine learning).
    - Really helpful for students looking to write senior theses.
- Only need 3 more classes to finish the data science track in Gov!
- More theoretical stats side: Stat 110/111
- More CS approach to data science: CS109 (Data Science 1)

Thanks!

Fill out your evaluations!