Gov 50: 22. Uncertainty in Regressions

Matthew Blackwell

Harvard University

Fall 2018

1/ Today's agenda

Where are we? Where are we going?

- Learned about uncertainty for sample means and sample difference-in-means.
- What about our regression estimates?

- Final project:
  - Analyses due tonight.
  - Template Rmd file uploaded to Canvas.
  - Final write-up due 12/10
  - We'll learn some key concepts for interpreting regression coefficients today.

2/ Regression review

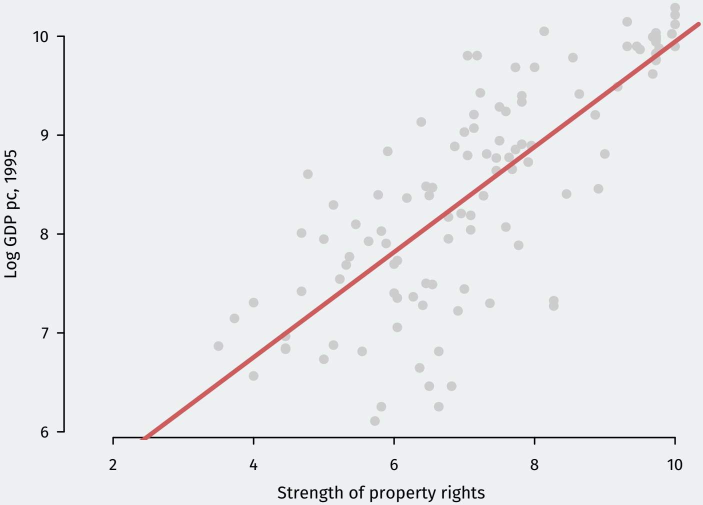
Data

- Do political institutions promote economic development?
- Acemoglu, Johnson, and Robinson (2001) look at the relationship between strength of property rights in a country and GDP.
- Data:

```r
ajr <- foreign::read.dta("data/ajr.dta")
```

Name	Description
`shortnam`	three-letter country code
`africa`	indicator for if the country is in Africa
`avexpr`	strength of property rights (protection against expropriation)
`logpgp95`	log GDP per capita
`imr95`	infant mortality rate

AJR scatterplot



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

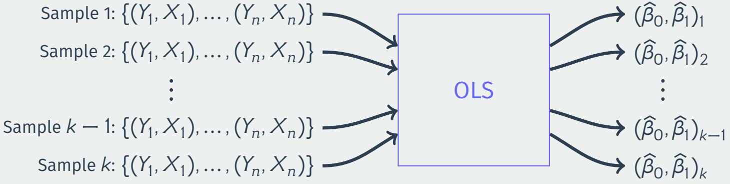
- How do we figure out the best line to draw?
  - alt question: how do we figure out β_0 and β_1?
  - $(\widehat{\beta}_0, \widehat{\beta}_1)$: estimated coefficients.
  - $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$: predicted/fitted value.
  - $\hat{\epsilon}_i = Y_i - \widehat{Y}$: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

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

3/ OLS as a estimator

Estimators

- Remember: least squares is an estimator—it's a machine that we plug data into and we get out estimates.



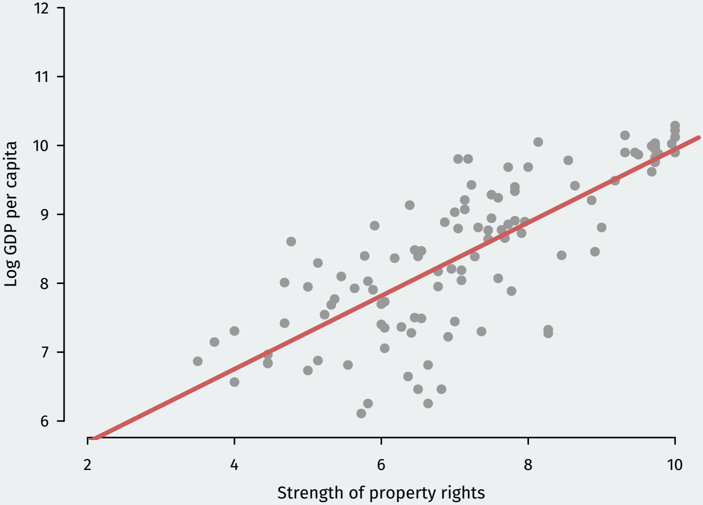
- Just like the sample mean, sample difference in means, or the sample variance
- It has a sampling distribution, with a sampling variance/standard error, etc.

Simulation procedure

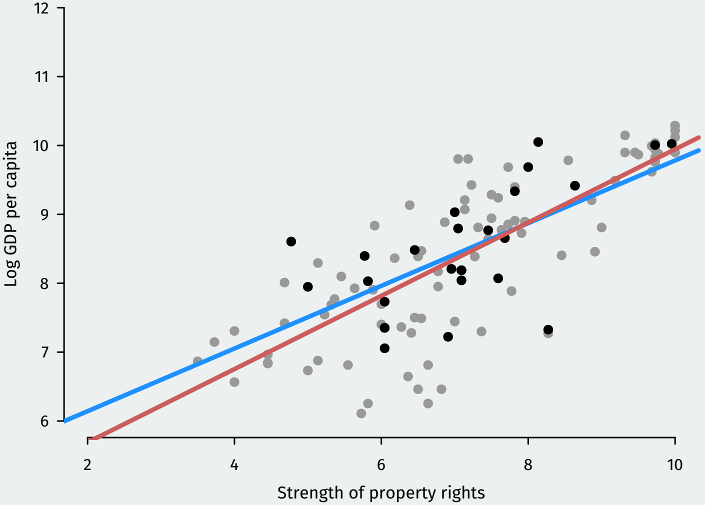
- Let's take a simulation approach to demonstrate:
  - Pretend that the AJR data represents the population of interest
  - See how the line varies from sample to sample

1. Draw a random sample of size $n = 30$ with replacement using `sample()`
2. Use `lm()` to calculate the OLS estimates of the slope and intercept
3. Plot the estimated regression line

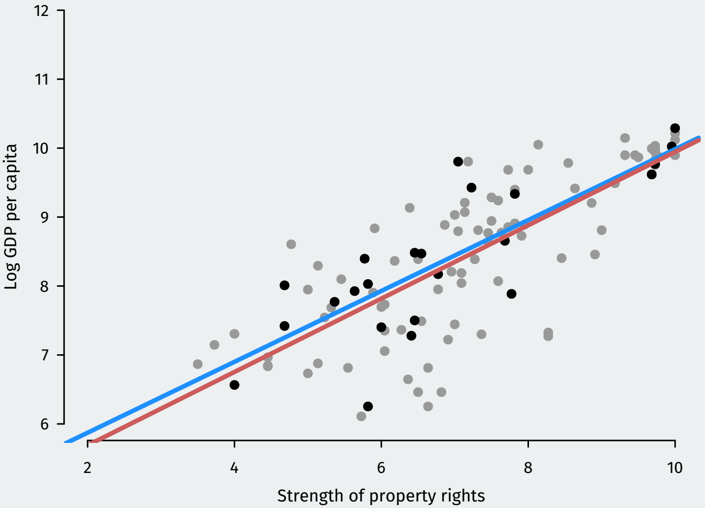
Population regression



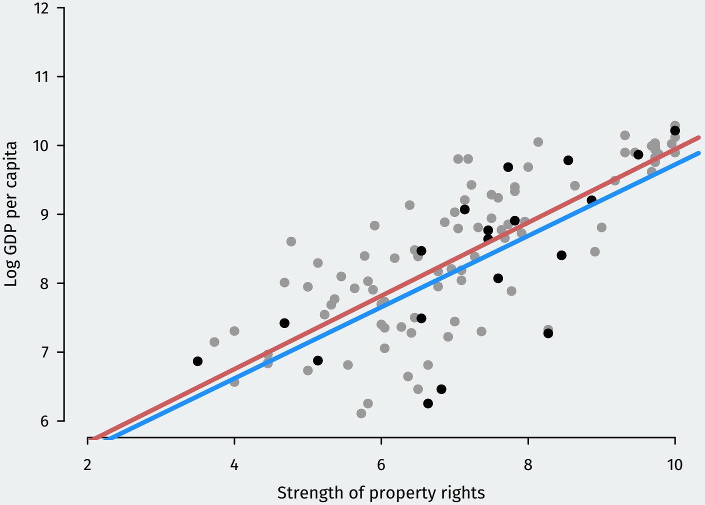
Randomly sample from AJR



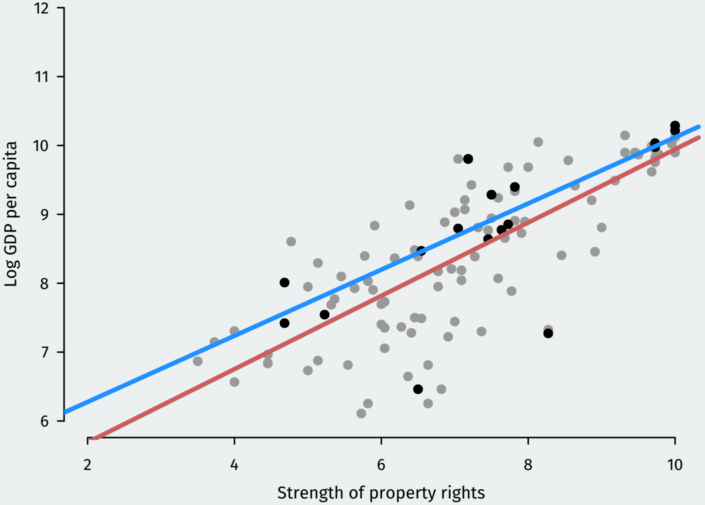
Randomly sample from AJR



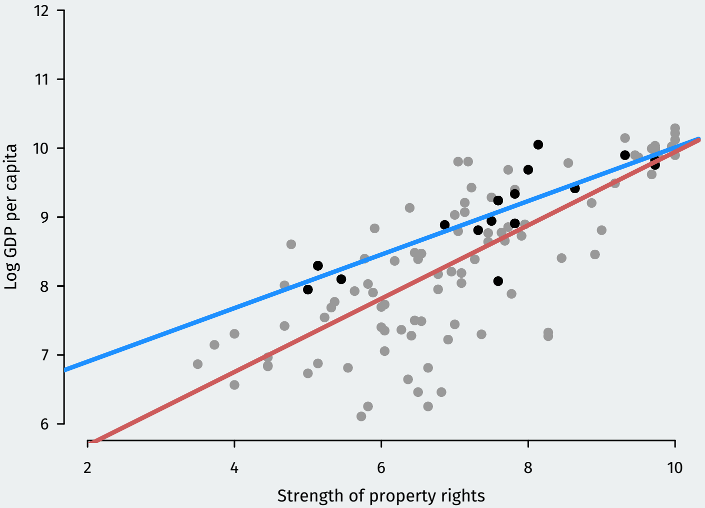
Randomly sample from AJR



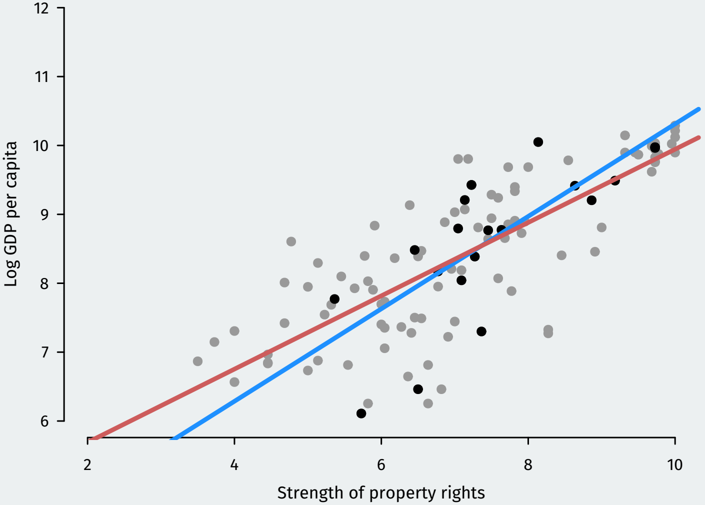
Randomly sample from AJR



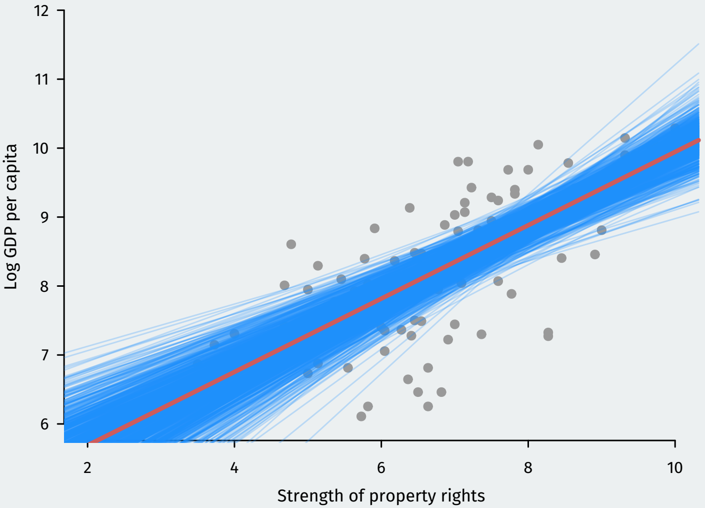
Randomly sample from AJR



Randomly sample from AJR

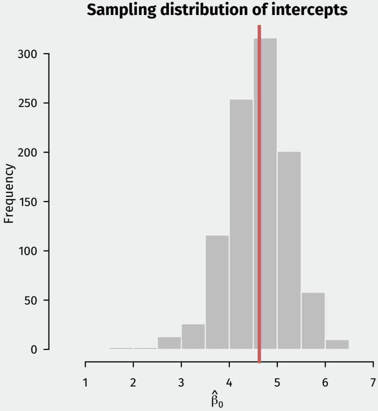


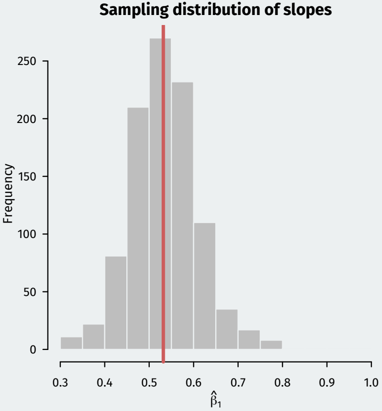
Randomly sample from AJR



Sampling distribution of OLS

- You can see that the estimated slopes and intercepts vary from sample to sample, but that the “average” of the lines looks about right.





Assumptions

- Key assumptions of regression:
  1. **Exogeneity**: mean of ϵ_i does not depend on X_i:

$$\mathbb{E}(\epsilon_i|X_i) = \mathbb{E}(\epsilon_i) = 0$$

  2. **Homoskedasticity**: variance of ϵ_i does not depend on X_i:

$$\mathbb{V}(\epsilon_i|X_i) = \mathbb{V}(\epsilon_i) = \sigma^2$$

- Exogeneity violated if there are unmeasured confounders between Y_i and X_i.
  - i.e., things in ϵ_i that are related to X_i
- Homoskedasticity violated when spead of Y_i depends on X_i.
  - easy fix for this, but beyond the scope of this class.

Properties of OLS

- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are random variables
  - Are they on average equal to the true values (bias)?
  - How spread out are they around their center (variance)?
- We can also estimate their standard error: $\widehat{\mathrm{SE}}(\widehat{\beta}_1)$
  - Our best guess at the spread of the estimator
- Under exogeneity and homoskedasticity,
  - $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased
  - Estimated standard errors are unbiased

Tests and CIs for regression

- 95% confidence intervals:
  - $\widehat{\beta}_0 \pm 1.96 \times \widehat{\text{SE}}(\widehat{\beta}_0)$
  - $\widehat{\beta}_1 \pm 1.96 \times \widehat{\text{SE}}(\widehat{\beta}_1)$
- Hypothesis tests:
  - Null hypothesis: $H_0 : \beta_1 = \beta_1^*$
  - Test statistic: $\frac{\widehat{\beta}_1 - \beta_1^*}{\widehat{\text{SE}}(\widehat{\beta}_1)} \sim N(0, 1)$
  - Usual test is of $\beta_1 = 0$.
  - $\widehat{\beta}_1$ 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
```

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 $\widehat{\beta}_j$
  - Hypothesis tests done exactly the same for $\widehat{\beta}_j$
  - $\rightsquigarrow$ interpret p-values the same as before.

```r
ajr.reg <- lm(logpgp95 ~ avexpr + africa + imr95, data = ajr)
summary(ajr.reg)
```

```
##
## Call:
## lm(formula = logpgp95 ~ avexpr + africa + imr95, data = ajr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3928 -0.2708  0.0865  0.2749  1.1652
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  7.01362    0.40445   17.34    < 2e-16 ***
## avexpr       0.28872    0.05046    5.72 0.00000043 ***
## africa      -0.02069    0.18622   -0.11       0.91
## imr95       -0.01549    0.00271   -5.71 0.00000045 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.492 on 56 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  0.778,  Adjusted R-squared:  0.766
## F-statistic: 65.4 on 3 and 56 DF,  p-value: <2e-16
```

Regression tables

- In papers, you'll often find regression tables that have several models.
- Each column is a different regression with different predictors or different samples.
- 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	0.52	0.47	0.43	0.47	0.41	0.45	0.46
	(0.04)	(0.06)	(0.06)	(0.05)	(0.06)	(0.06)	(0.04)	(0.06)
Latitude			0.89	0.37	1.60	0.92		
			(0.49)	(0.51)	(0.70)	(0.63)		
Asia dummy				−0.62		−0.60		
				(0.19)		(0.23)		
Africa dummy				−1.00		−0.90		
				(0.15)		(0.17)		
"Other" continent dummy				−0.25		−0.04		
				(0.20)		(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

4/ Wrapping up

Wrapping up

- Main goal of statistical methods: learn about what we don't know (population parameters) from what we do know (data).
- Messages to keep in mind moving forward:
  - A particular sample or result could be due to random chance $\rightsquigarrow$ use hypothesis tests and confidence intervals to assess
  - Be skeptical of causal claims unless groups are really comparable.
  - Think carefully about sampling biases when people make claims.

Want more?

- More Gov classes in quantitative methods:
  - Gov 61: more advanced methods for thesis writers
  - Gov 1000/2000: first methods class for PhD students.
  - Gov 1005 (Data)/Gov 1006 (Models): lots of tools for data science.
  - Classes by Prof. James Snyder have data analysis components.
- Outside Gov:
  - Stat 110/111: deeper into statistical theory.
  - Data Science 1/2: more focus on computation and prediction.
- Outside classes:
  - Work with faculty on research projects!

Thanks!

Thanks for a really fun and engaging semester! Good luck with your final projects!