POLS 6481. Research Design and Quantitative Methods II
Lecture 2. Simple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 2.1 + 2.2 + 2.3a

Outline:

Prelude – PED Use in Baseball

1. Review of the Linear Simple Regression Model
2. Summation (Review Appendix A.1, A.2)
3. Definitions (Review Appendix B.3, B.4)

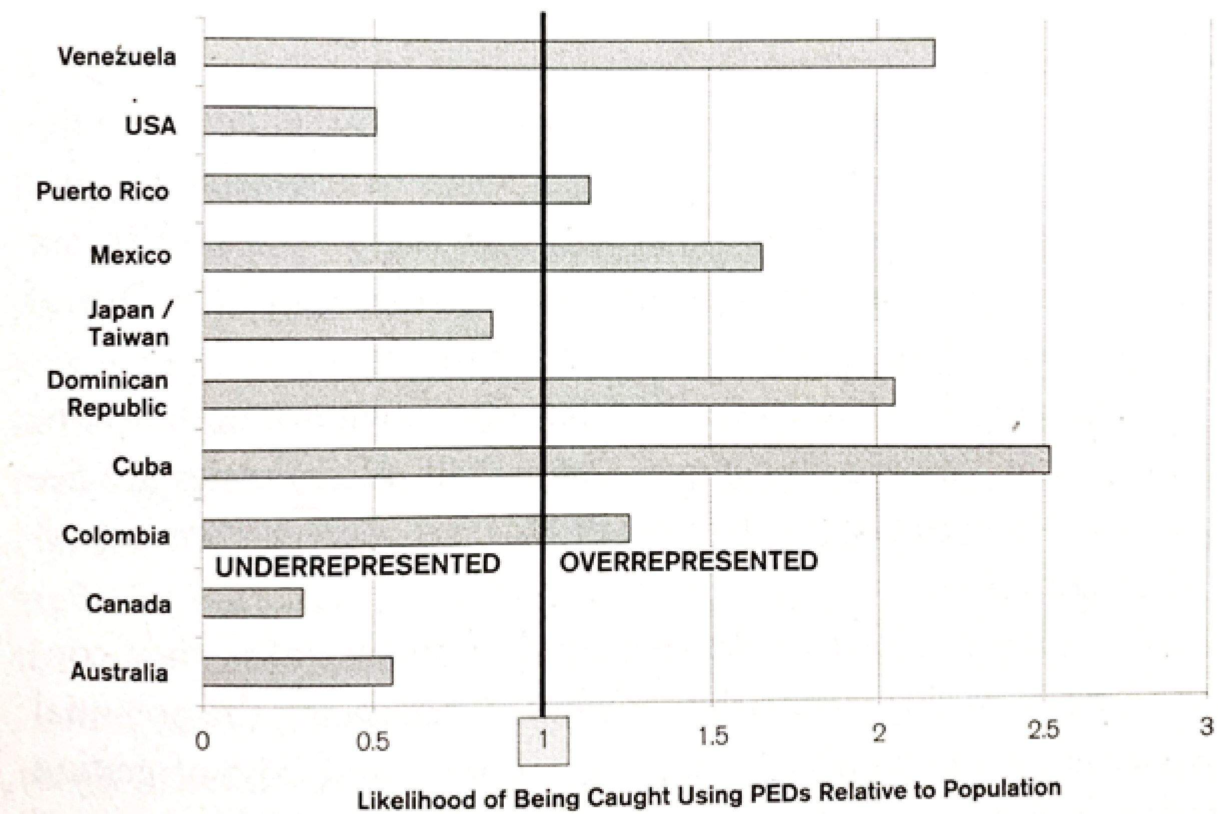
Interlude – PED Use in Baseball

4. Assumptions of the Linear Simple Regression Model

Postlude – PED Use in Baseball

POLS 6481. Research Design and Quantitative Methods II
Lecture 2. Simple Regression
Readings: Wooldridge, *Introductory Econometrics 5e,* 2.1 + 2.2 + 2.3a

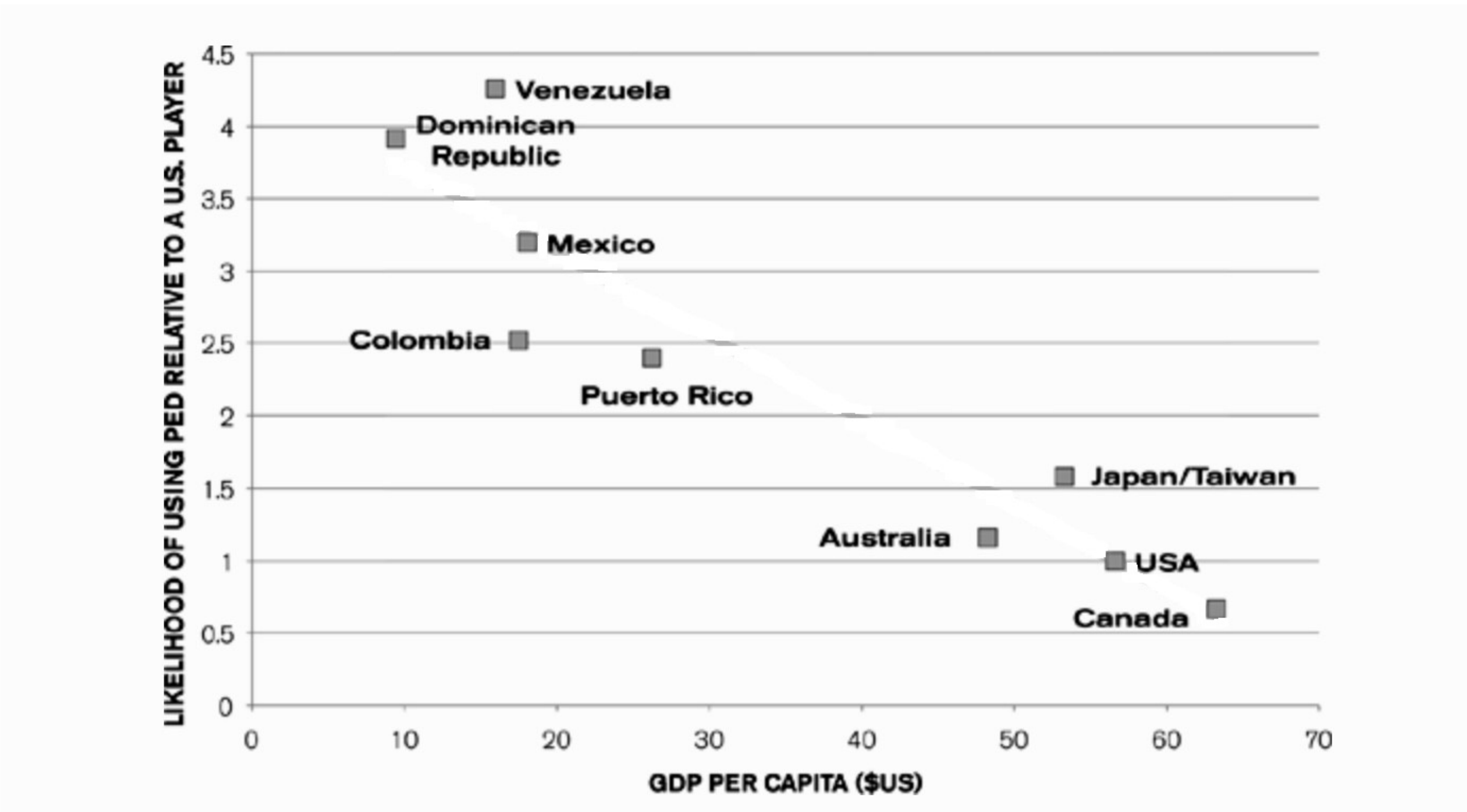
Prelude – PED Use in Baseball



Data for 1,520 U.S. Major League and 8,569 U.S. minor league players from 2005 to 2009.

POLS 6481. Research Design and Quantitative Methods II
Lecture 2. Simple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 2.1 + 2.2 + 2.3a

Prelude – PED Use in Baseball



POLS 6481. Research Design and Quantitative Methods II
Lecture 2. Simple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 2.1 + 2.2 + 2.3a

1. Review of the Linear Simple Regression Model

"explained" variable	y	a.k.a. dependent variable, *regressand*
"explanatory" variable	x	a.k.a. independent variable, *regressor*
intercept parameter	β_0	$E(y \mid x = 0)$
slope parameter	β_1	$\dfrac{\Delta E(y\|x)}{\Delta x}$
Population regression function	$E(y \mid x) = \beta_0 + \beta_1 x$	*systematic component* [2.8]
disturbance	u	*stochastic component*
Simple linear regression model	$y = \beta_0 + \beta_1 x + u$	

POLS 6481. Research Design and Quantitative Methods II
Lecture 2. Simple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 2.1 + 2.2 + 2.3a

2. Summation (Review Appendix A.1)

For the following, assume summation is over $i = 1, \ldots, n$

[A.1] $\sum x_i = x_1 + x_2 + \cdots$

[A.5] $\frac{\sum x_i}{n} = \bar{x}$

[A.6] $\sum (x_i - \bar{x}) = 0$

[A.2] $\sum c = n \cdot c$

[A.3] $\sum c \cdot x_i = c \cdot \sum x_i$

[A.7] $\sum x_i \cdot (x_i - \bar{x}) = \sum (x_i - \bar{x})^2 = \sum (x_i)^2 - \frac{(\sum x_i)^2}{n}$

$\sum x_i \cdot (y_i - \bar{y}) = \sum (x_i - \bar{x}) \cdot (y_i - \bar{y})$

3. Definitions (Review Appendix B.3, B.4)

For the following, assume summation is over $i = 1, \ldots, n$

Mean $\quad E(x) = \frac{1}{n}\sum x_i \qquad E(y) = \frac{1}{n}\sum y_i$

Variance $\quad E[(x-\mu_x)^2] = \frac{1}{n}\sum(x_i-\bar{x})^2 \qquad E\left[\left(y-\mu_y\right)^2\right] = \frac{1}{n}\sum(y_i-\bar{y})^2$ [B.23]

Covariance $\quad E\left[(x-\mu_x)\cdot\left(y-\mu_y\right)\right] = \frac{1}{n}\sum(x_i-\bar{x})\cdot(y_i-\bar{y})$ [B.26]

Correlation $\quad r_{xy} = \frac{cov(x,y)}{\sqrt{var(x)\cdot var(y)}} = \frac{\sum(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\cdot\sum(y_i-\bar{y})^2}}$ [B.29]

Regression slope $\quad \beta_1 = \frac{cov(x,y)}{var(x)} = \frac{\sum(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$ [2.19]

Regression intercept $\quad \beta_0 = \bar{y} - \beta_1\bar{x}$ [2.17]

3. Definitions, continued

Sample regression function $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} \cdot x$ [2.23]

Fitted values $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} \cdot x_i$ for $i = 1, \ldots, n$ [2.20]

Residual $\widehat{u_i} = y_i - \widehat{y_i}$

$= y_i - \widehat{\beta_0} - \widehat{\beta_1} \cdot x_i$ [2.21]

Sum of squared residuals $\sum \widehat{u_i}^2 = \sum (y_i - \widehat{y_i})^2 = \sum \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} \cdot x_i\right)^2$ [2.22]

4. Assumptions of the Linear Simple Regression Model

Assumption 1. Zero mean residual $\qquad E(\hat{u}) = 0 \qquad$ [2.10]

⟦ used to prove $\widehat{\beta_0} = \bar{y} - \beta_1 \bar{x}$; see whiteboard, p. 28 ⟧

① Assume $E(\hat{u}_i) = 0$; prove $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

define $\hat{u}_i = y_i - \hat{y} = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

then $\frac{1}{n} \sum \hat{u}_i = 0 = \frac{1}{n} \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$

$$0 = \frac{1}{n} \sum y_i - \frac{1}{n} \sum \hat{\beta}_0 - \frac{1}{n} \sum \hat{\beta}_1 x_i$$

$$0 = \bar{y} - \frac{1}{n} \cdot n \hat{\beta}_0 - \hat{\beta}_1 \bar{x}$$

def. of $\bar{y}$ — rule that $\sum_{i=1}^{n} c = nc$ — def. of $\bar{x}$ and rule that $\sum c \cdot x_i = c \sum x_i$

$$0 = \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

POLS 6481. Research Design and Quantitative Methods II
Lecture 2. Simple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 2.1 + 2.2 + 2.3a

4. Assumptions of the Linear Simple Regression Model

Assumption 2. Zero correlation of residual and regressor $E(\hat{u} \cdot x) = 0$ [2.11]

⟦ used to prove $\widehat{\beta_1} = \dfrac{cov(x,y)}{var(x)}$; see whiteboard, p. 28-29 ⟧

② Assume $E(\hat{u} \cdot x) = 0$; prove that $\hat{\beta}_1 = \dfrac{cov(x,y)}{var(x)}$ =

$$\frac{1}{n} \Sigma u_i x_i = 0 = \frac{1}{n} \Sigma x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

substitute $\bar{y} - \hat{\beta}_1 \bar{x}$ for $\hat{\beta}_0$: $\frac{1}{n} \Sigma x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)$

$$= \frac{1}{n} \Sigma x_i \left((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})\right)$$

$$= \frac{1}{n} \Sigma x_i (y_i - \bar{y}) - \hat{\beta}_1 \frac{1}{n} \Sigma x_i (x_i - \bar{x})$$

$$= \frac{1}{n} \Sigma (x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \cdot \frac{1}{n} \Sigma (x_i - \bar{x})^2$$

$$0 = cov(x,y) - \hat{\beta}_1 \, var(x)$$

$$\hat{\beta}_1 \, var(x) = cov(x,y)$$

$$\hat{\beta}_1 = \frac{cov(x,y)}{var(x)}$$

Interlude – PED Use in Baseball

Scorecasting, by T.J. Moskowitz & L.J. Wertheim (2011)

Dataset:	*PEDs in baseball.csv*
Sample:	National aggregate data (n = 9ish) from 2005 to 2009
Dependent variable:	Relative likelihood of PED use (1 = American ballplayer)
Independent variable:	Gross Domestic Product per capita (US $)

```
x <- c(48, 63, 17, 9, 53, 18, 26, 56, 16)
y <- c(1.2, 0.7, 2.5, 3.9, 1.6, 3.2, 2.4, 1.0, 4.2)
origin <- c("Australia", "Canada", "Colombia", "Dominican", "JapanTaiwan", "Mexico",
            "PuertoRico", "USA", "Venezuela")
plot(x, y, pch=16, xlim = c(0,70), ylim = c(0,4.5))
text(x, y, labels=origin, pos=1, xpd=TRUE)
```

Readings: Wooldridge, *Introductory Econometrics 5e*, 2.1 + 2.2 + 2.3a

Interlude – PED Use in Baseball

i	x_i	y_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
Australia	48	1.2					
Canada	63	0.7					
Colombia	17	2.5					
Dominican	9	3.9					
JapanTaiwan	53	1.6					
Mexico	18	3.2					
PuertoRico	26	2.4					
USA	56	1.0					
Venezuela	16	4.2					
Σ			$SST_x =$		$SST_y =$		
Σ/n	$\bar{x} =$	$\bar{y} =$	$var(x)=$		$var(y)=$		$cov(x,y)=$

Slope: $\widehat{\beta_1} = \frac{cov(x,y)}{var(x)} = \underline{\qquad\qquad} =$

Intercept: $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} =$

POLS 6481. Research Design and Quantitative Methods II
Lecture 2. Simple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 2.1 + 2.2 + 2.3a

Postlude – PED Use in Baseball

i	x_i	y_i	$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$	$\hat{u}_i = (y_i - \hat{y}_i)$	$\hat{u}_i \cdot x_i$	$\hat{u}_i^2 = (y_i - \hat{y}_i)^2$
Australia	48	1.2				
Canada	63	0.7				
Colombia	17	2.5				
Dominican	9	3.9				
JapanTaiwan	53	1.6				
Mexico	18	3.2				
PuertoRico	26	2.4				
USA	56	1.0				
Venezuela	16	4.2				
Σ			$\Sigma\hat{y}_i =$	$\Sigma\hat{u}_i =$	$\Sigma\hat{u}_i \cdot x_i =$	SSR =
Σ/n	$\bar{x} =$	$\bar{y} =$	$\bar{\hat{y}} =$			

$$sigma: \quad \sqrt{\widehat{\sigma^2}} = \sqrt{\frac{SSR}{n-2}} =$$

$$se(\widehat{\beta_1}) = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}} = - =$$

$$se(\widehat{\beta_0}) = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}} \cdot \sqrt{\frac{\sum x_i^2}{n}} =$$