Lecture 6: Correlation and Linear Regression Basics

ECON 480 - Econometrics - Fall 2018

Ryan Safner

September 17, 2018

Covariance and Correlation

- We looked at single variables for descriptive statistics

- We looked at single variables for descriptive statistics
- Most uses of statistics in economics and business investigate relationships *between* variables

- We looked at single variables for descriptive statistics
- Most uses of statistics in economics and business investigate relationships *between* variables
    - # of police & crime rates

- We looked at single variables for descriptive statistics
- Most uses of statistics in economics and business investigate relationships *between* variables
    - # of police & crime rates
    - healthcare spending & life expectancy

- We looked at single variables for descriptive statistics
- Most uses of statistics in economics and business investigate relationships *between* variables
    - # of police & crime rates
    - healthcare spending & life expectancy
    - government spending & GDP growth

- We looked at single variables for descriptive statistics
- Most uses of statistics in economics and business investigate relationships *between* variables
    - # of police & crime rates
    - healthcare spending & life expectancy
    - government spending & GDP growth
    - carbon dioxide emissions & temperatures

- We looked at single variables for descriptive statistics
- Most uses of statistics in economics and business investigate relationships *between* variables
    - # of police & crime rates
    - healthcare spending & life expectancy
    - government spending & GDP growth
    - carbon dioxide emissions & temperatures
- We will begin with bivariate data for relationships between X and Y

Bivariate Data and Relationships

- We looked at single variables for descriptive statistics
- Most uses of statistics in economics and business investigate relationships *between* variables
    - # of police & crime rates
    - healthcare spending & life expectancy
    - government spending & GDP growth
    - carbon dioxide emissions & temperatures
- We will begin with bivariate data for relationships between X and Y
    - Immediate aim is to explore associations between variables, quantified with **correlation** and **linear regression**

Bivariate Data and Relationships

- We looked at single variables for descriptive statistics
- Most uses of statistics in economics and business investigate relationships *between* variables
    - # of police & crime rates
    - healthcare spending & life expectancy
    - government spending & GDP growth
    - carbon dioxide emissions & temperatures
- We will begin with bivariate data for relationships between X and Y
    - Immediate aim is to explore associations between variables, quantified with **correlation** and **linear regression**
    - Later we want to develop more sophisticated tools to argue for **causation**

Bivariate Data: Spreadsheets

```r
econfreedom<-read.csv("~/Dropbox/Teaching/Hood College/ECON 480 - Econometrics/D
head(econfreedom)
```

```
##   ISO.Code         Country Economic.Freedom.Summary.Index GDP.Per.Capita
## 1      AGO          Angola                           5.08       4153.146
## 2      ALB         Albania                           7.40       4543.088
## 3      ARE Unit. Arab Em.                            7.98      39313.274
## 4      ARG       Argentina                           4.81      10501.660
## 5      ARM         Armenia                           7.71       3796.517
## 6      AUS       Australia                           7.93      54688.446
```

- **Rows** are individual observations

```r
econfreedom<-read.csv("~/Dropbox/Teaching/Hood College/ECON 480 - Econometrics/D
head(econfreedom)
```

```
##   ISO.Code          Country Economic.Freedom.Summary.Index GDP.Per.Capita
## 1      AGO           Angola                           5.08       4153.146
## 2      ALB          Albania                           7.40       4543.088
## 3      ARE Unit. Arab Em.                             7.98      39313.274
## 4      ARG        Argentina                           4.81      10501.660
## 5      ARM          Armenia                           7.71       3796.517
## 6      AUS        Australia                           7.93      54688.446
```

- **Rows** are individual observations
- **Columns** are variables on all individuals

```r
econfreedom<-read.csv("~/Dropbox/Teaching/Hood College/ECON 480 - Econometrics/D
head(econfreedom)
```

```
##   ISO.Code         Country Economic.Freedom.Summary.Index GDP.Per.Capita
## 1      AGO          Angola                           5.08       4153.146
## 2      ALB         Albania                           7.40       4543.088
## 3      ARE Unit. Arab Em.                            7.98      39313.274
## 4      ARG       Argentina                           4.81      10501.660
## 5      ARM         Armenia                           7.71       3796.517
## 6      AUS       Australia                           7.93      54688.446
```

- **Rows** are individual observations
- **Columns** are variables on all individuals
- Let X be Economic Freedom and Y be GDP per capita

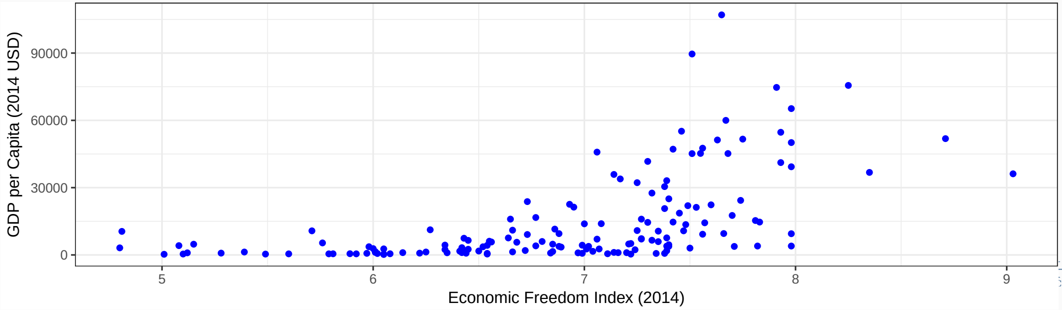
```r
str(econfreedom)
```

```
## 'data.frame':    152 obs. of  4 variables:
##  $ ISO.Code                      : Factor w/ 152 levels "AGO","ALB","ARE",..:
##  $ Country                       : Factor w/ 152 levels "Albania","Algeria",.
##  $ Economic.Freedom.Summary.Index: num  5.08 7.4 7.98 4.81 7.71 7.93 7.56 6.5
##  $ GDP.Per.Capita                : num  4153 4543 39313 10502 3797 ...
```

```r
summary(econfreedom)
```

```
##      ISO.Code           Country      Economic.Freedom.Summary.Index
##   AGO     :  1    Albania   :  1    Min.    :4.800
##   ALB     :  1    Algeria   :  1    1st Qu.:6.430
##   ARE     :  1    Angola    :  1    Median :7.050
##   ARG     :  1    Argentina:  1    Mean    :6.909
##   ARM     :  1    Armenia   :  1    3rd Qu.:7.428
##   AUS     :  1    Australia:  1    Max.    :9.030
##   (Other):146    (Other)   :146
##   GDP.Per.Capita
##   Min.    :   206.7
##   1st Qu.:  1588.3
##   Median :  5719.3
```

```r
library("ggplot2")
ggplot(econfreedom, aes(x=Economic.Freedom.Summary.Index,y=GDP.Per.Capita))+
  geom_point(color="blue")+theme_bw()+
  xlab("Economic Freedom Index (2014)")+ylab("GDP per Capita (2014 USD)")
```



- The best way to visualize an association between two variables is with a scatterplot

- Look for **association** between independent and dependent variables

- Look for **association** between independent and dependent variables
    1. *Direction*: is the trend positive or negative?

Associations

- Look for **association** between independent and dependent variables

    1. *Direction*: is the trend positive or negative?
    2. *Form*: is the trend linear, quadratic, something else, or no pattern?

- Look for **association** between independent and dependent variables

    1. *Direction*: is the trend positive or negative?
    2. *Form*: is the trend linear, quadratic, something else, or no pattern?
    3. *Strength*: is the association strong or weak?

- Look for **association** between independent and dependent variables

    1. *Direction*: is the trend positive or negative?
    2. *Form*: is the trend linear, quadratic, something else, or no pattern?
    3. *Strength*: is the association strong or weak?
    4. *Outliers*: do any observations break the trends above?

HOOD COLLEGE

- For any two variables, we can measure their sample covariance, $cov(X, Y)$ or $s_{X,Y}$ to quantify how they vary *together*[1]

$$s_{X,Y} = E\left[(X - \bar{X})(Y - \bar{Y})\right]$$

[1] Henceforth we limit to samples, for convenience. Population covariance is denoted $\sigma_{X,Y}$

- For any two variables, we can measure their sample covariance, $cov(X, Y)$ or $s_{X,Y}$ to quantify how they vary *together*[1]

$$s_{X,Y} = E\left[(X - \bar{X})(Y - \bar{Y})\right]$$

- Intuition: if X is above its mean, would we expect Y:

[1] Henceforth we limit to samples, for convenience. Population covariance is denoted $\sigma_{X,Y}$

- For any two variables, we can measure their sample covariance, $cov(X, Y)$ or $s_{X,Y}$ to quantify how they vary *together*[1]

$$s_{X,Y} = E\left[(X - \bar{X})(Y - \bar{Y})\right]$$

- Intuition: if X is above its mean, would we expect Y:
    - to be *above* its mean also (X and Y covary *positively*)

[1] Henceforth we limit to samples, for convenience. Population covariance is denoted $\sigma_{X,Y}$

- For any two variables, we can measure their sample covariance, $cov(X, Y)$ or $s_{X,Y}$ to quantify how they vary *together*[1]

$$s_{X,Y} = E\left[(X - \bar{X})(Y - \bar{Y})\right]$$

- Intuition: if X is above its mean, would we expect Y:
  - to be *above* its mean also (X and Y covary *positively*)
  - to be *below* its mean (X and Y covary *negatively*)

[1] Henceforth we limit to samples, for convenience. Population covariance is denoted $\sigma_{X,Y}$

- For any two variables, we can measure their sample covariance, $cov(X, Y)$ or $s_{X,Y}$ to quantify how they vary *together*[1]

$$s_{X,Y} = E\Big[\big(X - \bar{X}\big)\big(Y - \bar{Y}\big)\Big]$$

- Intuition: if X is above its mean, would we expect Y:
    - to be *above* its mean also (X and Y covary *positively*)
    - to be *below* its mean (X and Y covary *negatively*)
- Covariance is a common measure, but the units are meaningless, thus we rarely need to use it so **don't worry about learning the formula**

[1] Henceforth we limit to samples, for convenience. Population covariance is denoted $\sigma_{X,Y}$

- More convenient to standardize covariance into a more intuitive concept: correlation (ρ or r), normalized to be between -1 and 1

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y} = \frac{cov(X,Y)}{sd(X)sd(Y)}$$

- More convenient to standardize covariance into a more intuitive concept: correlation (ρ or r), normalized to be between -1 and 1

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y} = \frac{cov(X,Y)}{sd(X)sd(Y)}$$

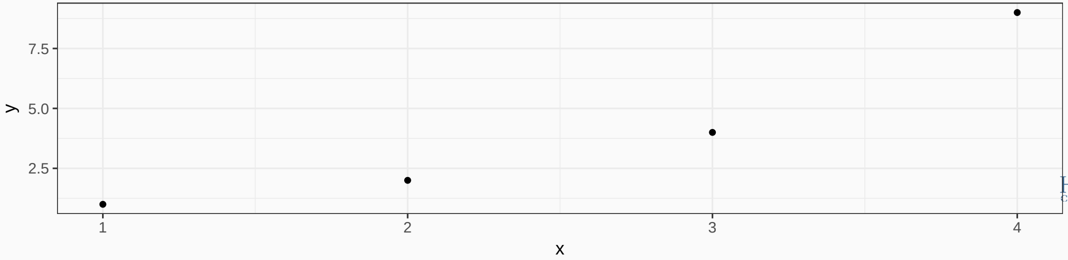
- Alternatively, sample correlation can be found by standardizing (finding the Z-score) X and Y and multiplying, for each (X,Y) pair, and then averaging (over $n-1$, due to sampling df, again):

$$\begin{aligned} r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n} Z_X Z_Y \end{aligned}$$

Example

$$(1,1),(2,2),(3,4),(4,9)$$

```
corr.example<-data.frame(x=c(1,2,3,4),
                         y=c(1,2,4,9))
ggplot(corr.example,aes(x=x,y=y))+geom_point()
```



Correlation: Example Calculation II

```r
mean(corr.example$x) #find mean of x
```

```
## [1] 2.5
```

```r
mean(corr.example$y) #find mean of y
```

```
## [1] 4
```

```r
sd(corr.example$x) #find sd of x
```

```
## [1] 1.290994
```

```r
sd(corr.example$y) #find sd of y
```

```
## [1] 3.559026
```

Correlation: Example Calculation III

```r
#take z score of x,y for each pair and multiply them
corr.example$z.product<-(((corr.example$x-2.5)/1.291)*
                                        ((corr.example$y-4)/3.559))

corr.example
```

```
##   x y z.product
## 1 1 1 0.9793959
## 2 2 2 0.2176435
## 3 3 4 0.0000000
## 4 4 9 1.6323265
```

```r
(sum(corr.example$z.product)/3) #average z products over n-1
```

```
## [1] 0.943122
```

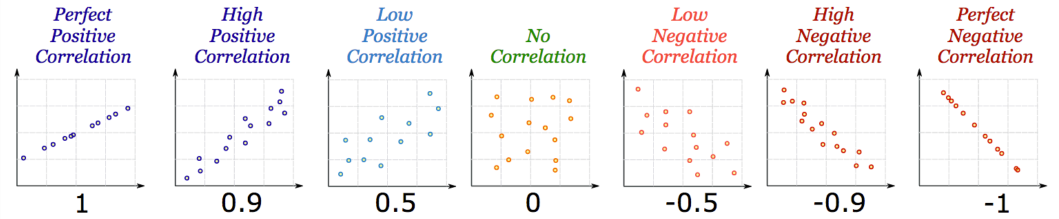
```r
cor(corr.example$x, corr.example$y) #compare our answer to cor() command
```

```
## [1] 0.9431191
```

```r
cov(corr.example$x, corr.example$y) #just for kicks - covariance
```

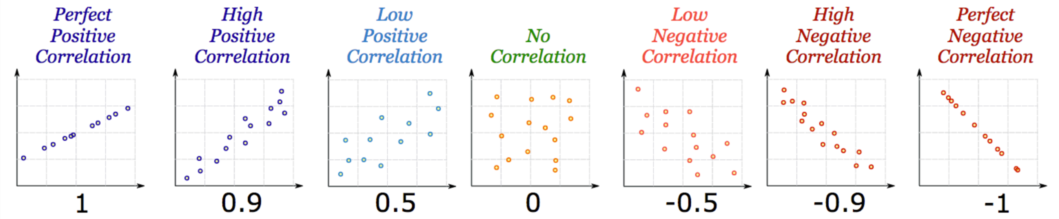
```
## [1] 4.333333
```

- Correlation is standardized to $-1 \leq r \leq 1$



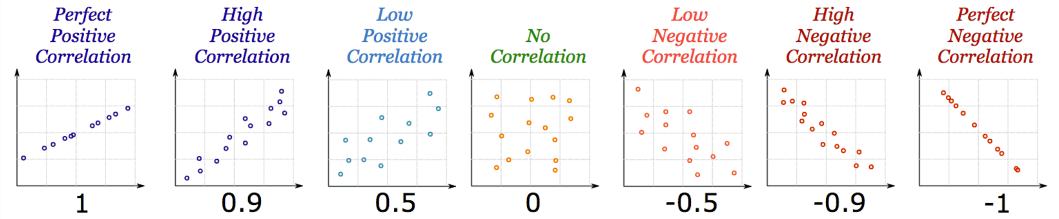
Correlation: Interpretation

- Correlation is standardized to $-1 \leq r \leq 1$
  - Negative values $\implies$ negative association

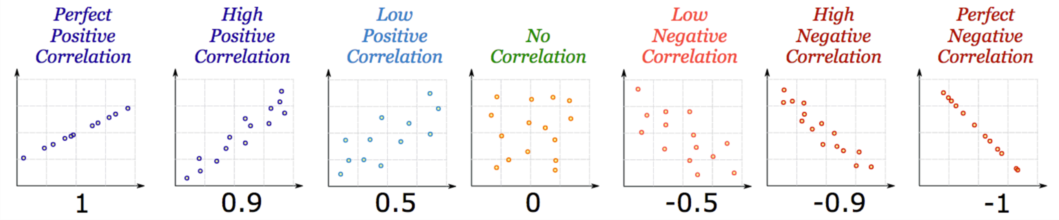


Correlation: Interpretation

- Correlation is standardized to $-1 \leq r \leq 1$
  - Negative values $\implies$ negative association
  - Positive values $\implies$ positive association

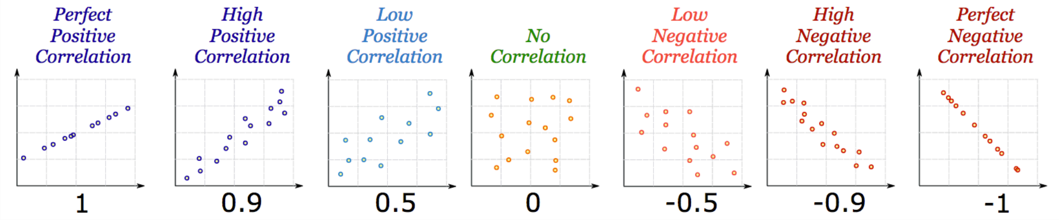


- Correlation is standardized to $-1 \leq r \leq 1$
  - Negative values $\implies$ negative association
  - Positive values $\implies$ positive association
  - Correlation of 0 $\implies$ no association

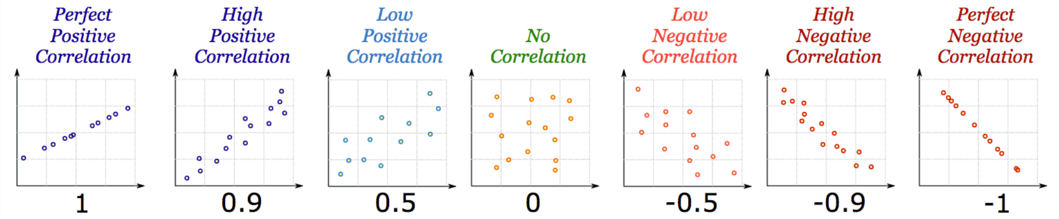


Correlation: Interpretation

- Correlation is standardized to $-1 \leq r \leq 1$
  - Negative values $\implies$ negative association
  - Positive values $\implies$ positive association
  - Correlation of 0 $\implies$ no association
  - As $|r| \to 1 \implies$ the stronger the association



- Correlation is standardized to $-1 \leq r \leq 1$
  - Negative values $\implies$ negative association
  - Positive values $\implies$ positive association
  - Correlation of 0 $\implies$ no association
  - As $|r| \to 1 \implies$ the stronger the association
  - Correlation of $|r| = 1 \implies$ a perfect linear relationship

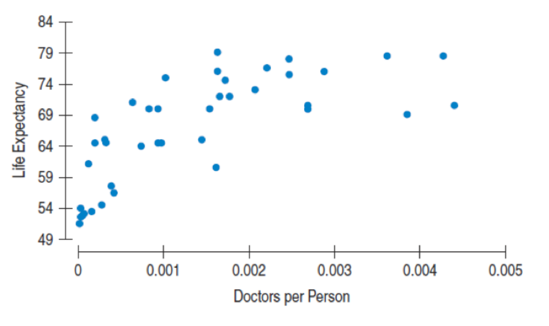


Guess The Correlation Game

- Reminder: Correlation does not imply causation!

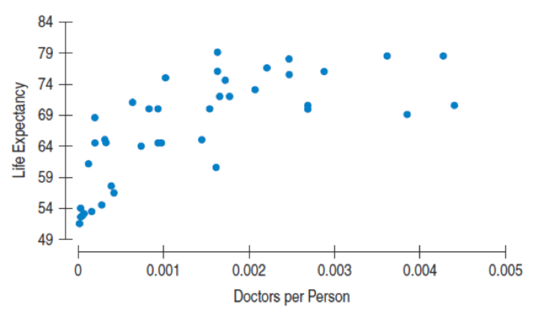
- Reminder: Correlation does not imply causation!
- See the **Handout** for more on Covariance and Correlation

Example



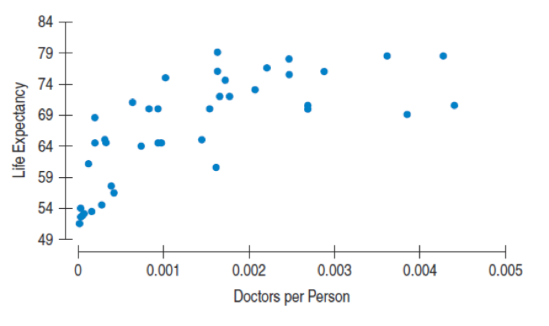
- The correlation between Life Expectancy and Doctors Per Person is 0.705.

Example

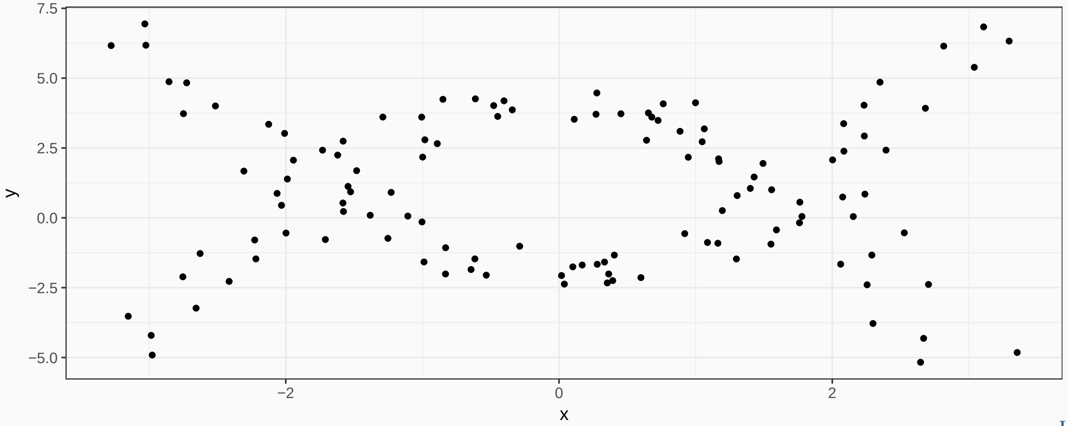


- The correlation between Life Expectancy and Doctors Per Person is 0.705.
- So should we send more doctors to developing countries to increase their life expectancy?

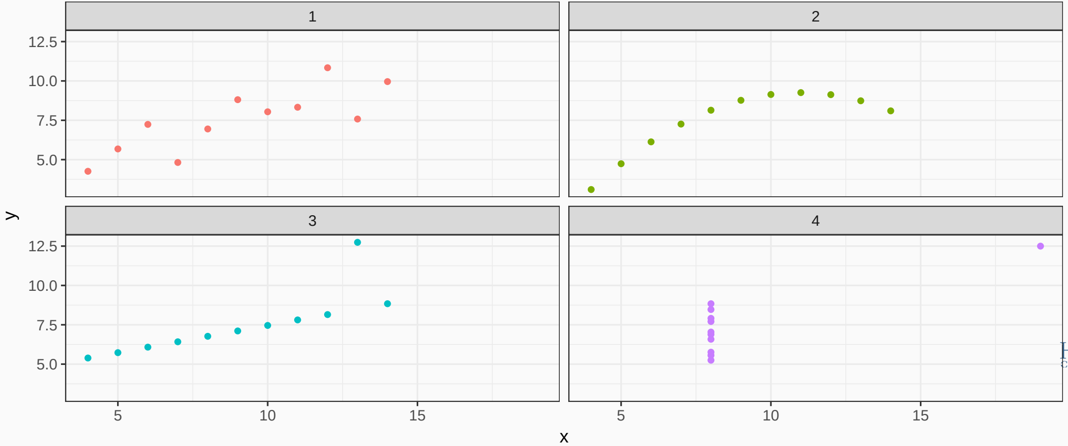
Example



- The correlation between Life Expectancy and Doctors Per Person is 0.705.
- So should we send more doctors to developing countries to increase their life expectancy?
- Properly interpreting relationships requires both statistical *and* economic intuition!


7.5
5.0
2.5
0.0
−2.5
−5.0
y
−2
0
2
x

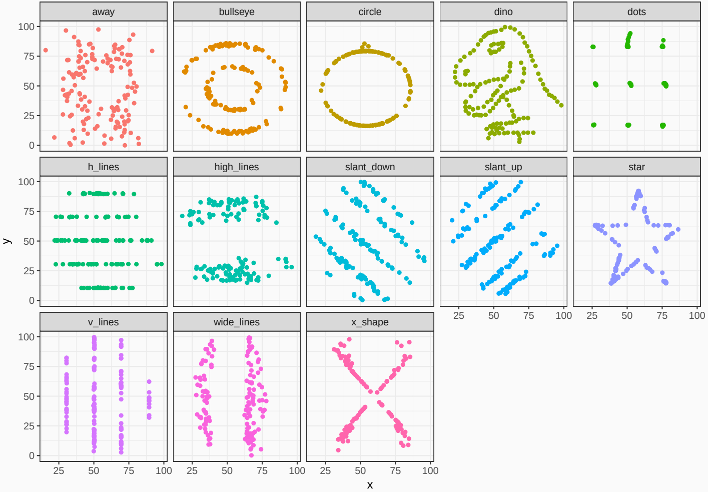
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      4.0     6.5     9.0     9.0    11.5    14.0
```



```
##    dataset                 x               y
##  Length:1846        Min.   :15.56   Min.   : 0.01512
##  Class :character   1st Qu.:41.07   1st Qu.:22.56107
##  Mode  :character   Median :52.59   Median :47.59445
##                     Mean   :54.27   Mean   :47.83510
##                     3rd Qu.:67.28   3rd Qu.:71.81078
##                     Max.   :98.29   Max.   :99.69468
```

HOOD COLLEGE

Anscombe's Quartet: A Modern Re-interpretation II



See the Datasaurus

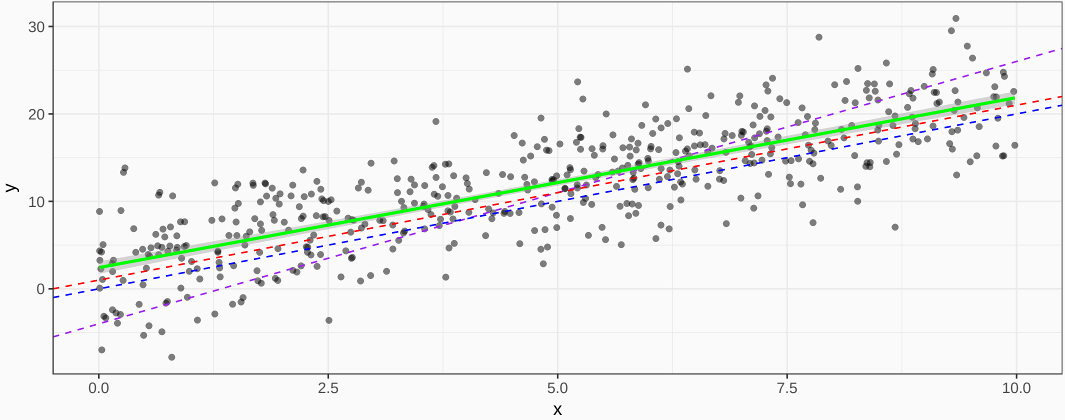
Population Linear Regression Model

- If an association appears linear, we can estimate the equation of a line that would “fit” the data

- If an association appears linear, we can estimate the equation of a line that would "fit" the data

$$Y = a + bX$$

- Recall a linear equation describing a line contains: - a: vertical intercept - b: slope - Note we will use different symbols for a and b, in line with standard econometric notation



- How do we choose the equation that best fits the data? Process is called linear regression

- Linear regression lets us estimate the slope of the population regression line between X and Y

- Linear regression lets us estimate the slope of the population regression line between X and Y
- We can make **inferences** about the population slope coefficient

- Linear regression lets us estimate the slope of the population regression line between X and Y
- We can make **inferences** about the population slope coefficient
    - eventually, a causal interpretation

Population Linear Regression Model

- Linear regression lets us estimate the slope of the population regression line between X and Y
- We can make **inferences** about the population slope coefficient
    - eventually, a causal interpretation
    - slope $= \frac{\Delta Y}{\Delta X}$: for a 1-unit change in X, how many units will this *cause* Y to change?

- Statistically, we want to use the population regression model for:

- Statistically, we want to use the population regression model for:

1. Estimation of the marginal effect of X on Y (slope of population regression line)

- Statistically, we want to use the population regression model for:

1. Estimation of the marginal effect of X on Y (slope of population regression line)
2. Hypothesis Testing of the value of the marginal effect (slope)

- Statistically, we want to use the population regression model for:

1. Estimation of the marginal effect of X on Y (slope of population regression line)
2. Hypothesis Testing of the value of the marginal effect (slope)
3. Confidence Interval construction of a range for the true effect (slope)

Example

What is the relationship between class size and educational performance?

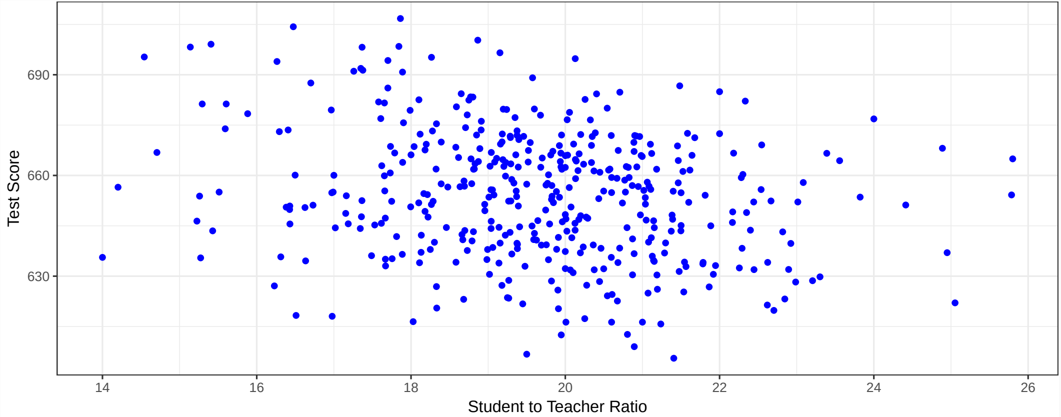
- Policy question: What is the effect of reducing class sizes by 1 student per class on test scores? 10 students?

An Extended Example: Scatterplot

```
library("foreign") #for importing .dta files
CASchool<-read.dta("~/Dropbox/Teaching/Hood College/ECON 480 - Econometrics/Data

ca.scatter<-ggplot(CASchool, aes(str,testscr))+
  geom_point(color="blue",fill="blue")+
  xlab("Student to Teacher Ratio")+
  ylab("Test Score")+theme_bw()
```

An Extended Example: Scatterplot II



- If we *change* (Δ) the class size by an amount, what would we expect the *change* in test scores to be?

$$\beta_{ClassSize} = \frac{\text{change in test score}}{\text{change in class size}} = \frac{\Delta \text{test score}}{\Delta \text{class size}}$$

- If we *change* (Δ) the class size by an amount, what would we expect the *change* in test scores to be?

$$\beta_{ClassSize} = \frac{\text{change in test score}}{\text{change in class size}} = \frac{\Delta \text{test score}}{\Delta \text{class size}}$$

- If we knew $\beta_{ClassSize}$, we could say that changing class size by 1 student will change test scores by $\beta_{ClassSize}$

An Extended Example: Slope II

- Rearranging:

$$\Delta \text{test score} = \beta_{ClassSize} \times \Delta \text{class size}$$

- Rearranging:

$$\Delta \text{test score} = \beta_{ClassSize} \times \Delta \text{class size}$$

- Suppose $\beta_{ClassSize} = -0.6$. If we shrank class size by 2 students, our model predicts:

- Rearranging:

$$\Delta\text{test score} = \beta_{ClassSize} \times \Delta\text{class size}$$

- Suppose $\beta_{ClassSize} = -0.6$. If we shrank class size by 2 students, our model predicts:

$$\Delta\text{test score} = -0.6$$

$$\Delta\text{test score} = \times -2 = 1.2$$

$$\text{test score} = \beta_0 + \beta_{ClassSize} \times \text{class size}$$

- The line relating class size and test scores has the above equation

An Extended Example: Slope III

$$\text{test score} = \beta_0 + \beta_{ClassSize} \times \text{class size}$$

- The line relating class size and test scores has the above equation
  - β_0 is the vertical-intercept, test score where class size is 0

$$\text{test score} = \beta_0 + \beta_{ClassSize} \times \text{class size}$$

- The line relating class size and test scores has the above equation
    - β_0 is the vertical-intercept, test score where class size is 0
    - $\beta_{ClassSize}$ is the slope of the regression line

An Extended Example: Slope III

$$\text{test score} = \beta_0 + \beta_{ClassSize} \times \text{class size}$$

- The line relating class size and test scores has the above equation
  - β_0 is the vertical-intercept, test score where class size is 0
  - $\beta_{ClassSize}$ is the slope of the regression line
- This relationship only holds **on average** for all districts in the population, individual districts are also affected by other factors

- To get an equation that holds for *each* district, we need to include other factors

$$\text{test score} = \beta_0 + \beta_{ClassSize} \times \text{class size} + \text{other factors}$$

- To get an equation that holds for *each* district, we need to include other factors

$$\text{test score} = \beta_0 + \beta_{ClassSize} \times \text{class size} + \text{other factors}$$

- For now, we will ignore these until the next lesson

- To get an equation that holds for *each* district, we need to include other factors

$$\text{test score} = \beta_0 + \beta_{ClassSize} \times \text{class size} + \text{other factors}$$

- For now, we will ignore these until the next lesson
- Thus, $\beta_0 + \beta_{ClassSize} \times$ class size gives the **average effect** of class sizes on scores

- To get an equation that holds for *each* district, we need to include other factors

$$\text{test score} = \beta_0 + \beta_{ClassSize} \times \text{class size} + \text{other factors}$$

- For now, we will ignore these until the next lesson
- Thus, $\beta_0 + \beta_{ClassSize} \times$ class size gives the **average effect** of class sizes on scores
- Later, we will want to estimate the **marginal effect** (**causal effect**) of each factor on an individual district's test score, holding all other factors constant

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- y is the dependent variable of interest
    - AKA "response variable," "regressand," "Left-hand side (LHS) variable"

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- y is the dependent variable of interest
    - AKA "response variable," "regressand," "Left-hand side (LHS) variable"
- x_1 and x_2 are independent variables
    - AKA "explanatory variables," "regressors," "Right-hand side (RHS) variables," "covariates," "control variables"

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- y is the dependent variable of interest
  - AKA "response variable," "regressand," "Left-hand side (LHS) variable"
- x_1 and x_2 are independent variables
  - AKA "explanatory variables," "regressors," "Right-hand side (RHS) variables," "covariates," "control variables"
- We have observed values of y, x_1, and x_2 & "regress y on x_1 and x_2"

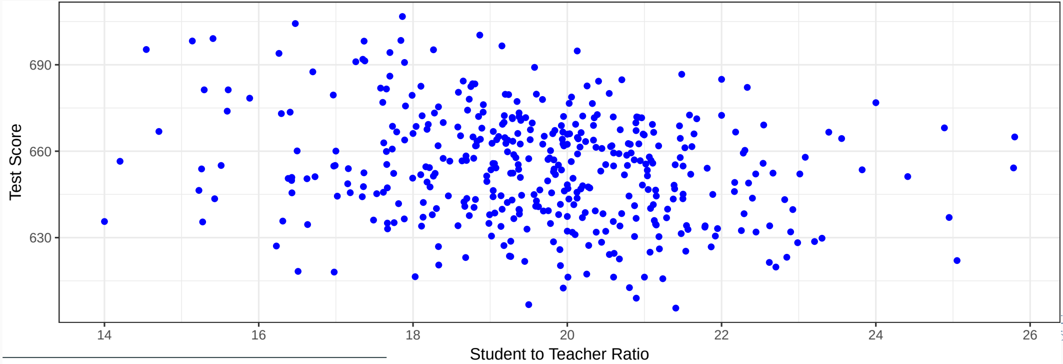
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- y is the dependent variable of interest
  - AKA “response variable,” “regressand,” “Left-hand side (LHS) variable”
- x_1 and x_2 are independent variables
  - AKA “explanatory variables,” “regressors,” “Right-hand side (RHS) variables,” “covariates,” “control variables”
- We have observed values of y, x_1, and x_2 & “regress y on x_1 and x_2”
- β_0, β_1, and β_2 are unknown parameters to *estimate*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

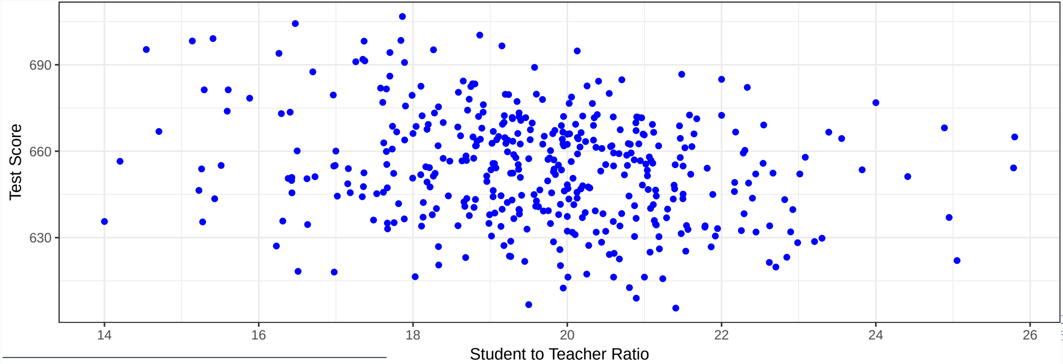
- y is the dependent variable of interest
  - AKA "response variable," "regressand," "Left-hand side (LHS) variable"
- x_1 and x_2 are independent variables
  - AKA "explanatory variables," "regressors," "Right-hand side (RHS) variables," "covariates," "control variables"
- We have observed values of y, x_1, and x_2 & "regress y on x_1 and x_2"
- β_0, β_1, and β_2 are unknown parameters to *estimate*
- ϵ is the error term
  - It is **stochastic** (random)
  - We can never measure the error term

- How do we draw a line through the scatterplot? We do not know the true $\beta_{ClassSize}$



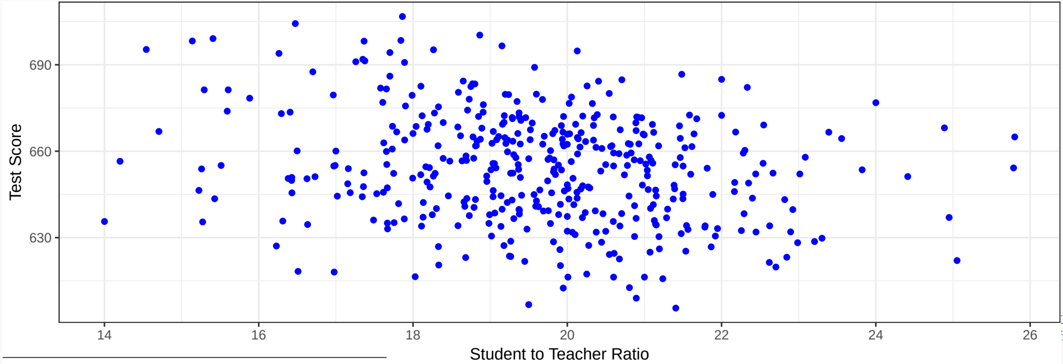
[2] Data is student-teacher-ratio and average test scores on Stanford 9 Achievement Test for 5th grade students for 420 K-6 and K-8 school districts in California in 1999, (Stock and Watson, 2015: p. 141)

- How do we draw a line through the scatterplot? We do not know the true $\beta_{ClassSize}$
- We do have data from a *sample* of class sizes and test scores[2]



[2] Data is student-teacher-ratio and average test scores on Stanford 9 Achievement Test for 5th grade students for 420 K-6 and K-8 school districts in California in 1999, (Stock and Watson, 2015: p. 141)

- How do we draw a line through the scatterplot? We do not know the true $\beta_{ClassSize}$
- We do have data from a *sample* of class sizes and test scores[2]
- So the real question is, **how can we estimate β_0 and β_1?**



[2] Data is student-teacher-ratio and average test scores on Stanford 9 Achievement Test for 5th grade students for 420 K-6 and K-8 school districts in California in 1999, (Stock and Watson, 2015: p. 141)

OLS Estimators and Sample Regression Model