Unit 6: Introduction to linear regression
1. Introduction to regression

Sta 101 - Spring 2019

Duke University, Department of Statistical Science

Dr. Ellison Slides posted at *https://www2.stat.duke.edu/courses/Spring19/sta101.001/*

Outline

1. Housekeeping

2. Main ideas:
    1. **Quantify** the Linear Relationship Between Two Numerical Variables: ***Correlation coefficient***
        - Correlation coefficient describes the strength and direction of the linear association between two numerical variables
    2. **Model** the Linear Relationship Between Two Numerical Variables: ***Simple Linear Regression Model***
        - Calculating the Model: Least squares line minimizes squared residuals
        - Interpreting the Model
        - Using the Model: Predict, but don't extrapolate

3. Summary

Coming up…

- Lab Assignment 8 due Friday 4/5 11:55pm *(extension)*
- Peer Evaluation 2 due Tuesday 4/9 11:55 pm
- Problem Set 6 due Wednesday 4/10 11:55 pm
- Readiness Assessment 7 Wednesday 4/10
- Don't forget Project Stage 2… due in ~2 weeks

Outline

1. Housekeeping

2. Main ideas:
    1. **Quantify** the Linear Relationship Between Two Numerical Variables: ***Correlation coefficient***
        - Correlation coefficient describes the strength and direction of the linear association between two numerical variables
    2. **Model** the Linear Relationship Between Two Numerical Variables: ***Simple Linear Regression Model***
        - Calculating the Model: Least squares line minimizes squared residuals
        - Interpreting the Model
        - Using the Model: Predict, but don't extrapolate

3. Summary

What's new in Unit 6?

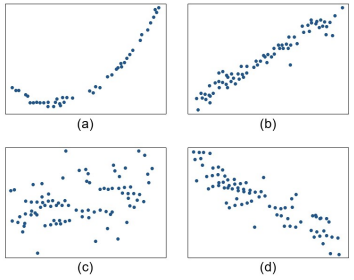
		Analyses and Ways to Conduct Them (that we know so far)	
Types of Variable(s) Involved	**Population Parameter**	**Confidence Interval for the Population Parameter**	**Hypothesis Test for the Population Parameter** *Ho: Pop. Param = # Ha: Pop. Param (≠ or < or >) #*
Single **Numerical** Variable	**μ**	CLT Confidence Interval (Unit 3+4) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 3+4) Bootstrap Hypothesis Test (Unit 4)
	μdiff	CLT Confidence Interval (Unit 4) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 4) Bootstrap Hypothesis Test (Unit 4)
	Median	Bootstrap Confidence Interval (Unit 4)	Bootstrap Hypothesis Test (Unit 4)
Single **Categorical** Variable **(2 levels)**	**p**	CLT Confidence Interval (Unit 5) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 5) Bootstrap Hypothesis Test (Unit 4) Randomization Testing (Unit 5-Selecting ball/chips out of bag, rolling dice)
Numerical Response Variable **Categorical Explanatory** Variable **(2 levels)**	**μ1-μ2**	CLT Confidence Interval (Unit 4) Bootstrap Confidence Interval (Unit 5)	CLT Hypothesis Test (Unit 4) Randomization Testing (Unit 1-Shuffling Cards)
	Median1-Median2	Bootstrap Confidence Interval (Unit 5)	Randomization Testing (Unit 1-Shuffling Cards)
Categorical Response Variable **Categorical Explanatory** Variable **(both have 2 levels)**	**p1-p2**	CLT Confidence Interval (Unit 5) Bootstrap Confidence Interval (Unit 5)	CLT Hypothesis Test (Unit 5) Randomization Testing (Unit 1-Shuffling Cards)
Numerical Response Variable **Numerical and/or Categorical Explanatory** Variable(s)	**βi (i=0,1,...,k)**	Regression Coefficient Confidence Interval (Units 6 +7)	Regression Coefficient Hypothesis Test (Units 6 +7)

1. Housekeeping

2. Main ideas:
   1. **Quantify** the *Linear* Relationship Between Two Numerical Variables: ***Correlation coefficient***
      - Correlation coefficient describes the strength and direction of the *linear association* between two numerical variables
   2. **Model** the Linear Relationship Between Two Numerical Variables: ***Simple Linear Regression Model***
      - Calculating the Model: Least squares line minimizes squared residuals
      - Interpreting the Model
      - Using the Model: Predict, but don't extrapolate

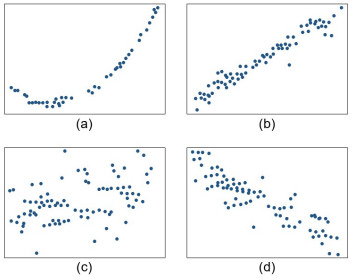
3. Summary

What are four things we can discuss about the relationship between two numerical variables?



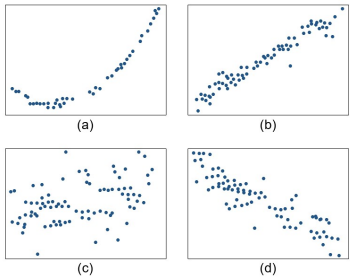
What are four things we can discuss about the relationship between two numerical variables?

1. Strength
2. Direction
3. Linearity
4. Any outliers



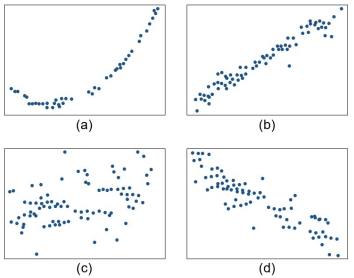
*double check your Lab 8

What metric can we use to *quantify* the strength and direction of a linear relationship?

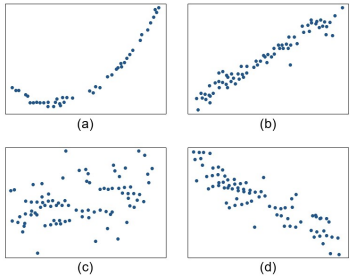


What metric can we use to *quantify* the strength and direction of a linear relationship?

Correlation Coefficient (*R*)



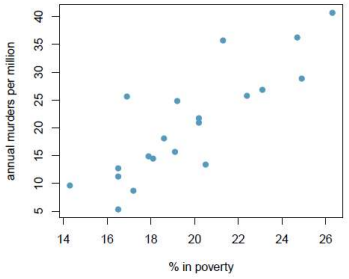
Guessing the Correlation Coefficient by Looking at a Scatter Plot



Clicker question

Which of the following is the best guess for the correlation between annual murders per million and percentage living in poverty?

(a) -1.52
(b) -0.63
(c) -0.12
(d) 0.02
(e) 0.84



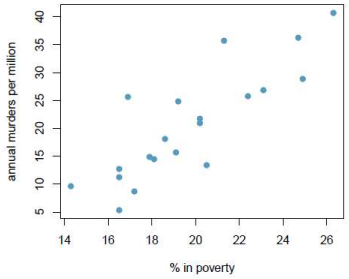
Guessing the correlation

Clicker question

Which of the following is the best guess for the correlation between annual murders per million and percentage living in poverty?

~~(a) -1.52~~
(b) -0.63
(c) -0.12
(d) 0.02
(e) 0.84

- $-1 \leq R \leq 1$



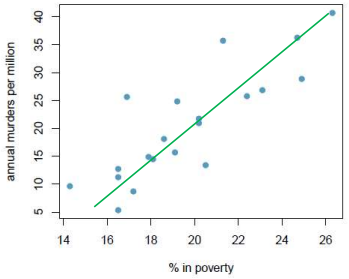
Guessing the correlation

Clicker question

Which of the following is the best guess for the correlation between annual murders per million and percentage living in poverty?

~~(a) -1.52~~
~~(b) -0.63~~
~~(c) -0.12~~
(d) 0.02
(e) 0.84

- $-1 \leq R \leq 1$
- Upwards Trending Relationship → **Positive R**



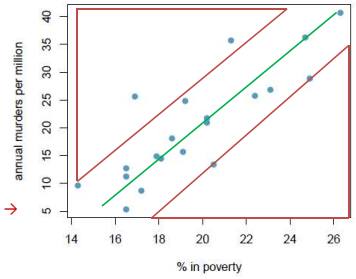
Guessing the correlation

Clicker question

Which of the following is the best guess for the correlation between annual murders per million and percentage living in poverty?

~~(a) -1.52~~
~~(b) -0.63~~
~~(c) -0.12~~
~~(d) 0.02~~
(e) 0.84

- $-1 \leq R \leq 1$
- Upwards Trending Relationship → **Positive R**
- About ~84% of the screen is roughly point-free, points pretty close to line → **$|R| \sim 0.84$**

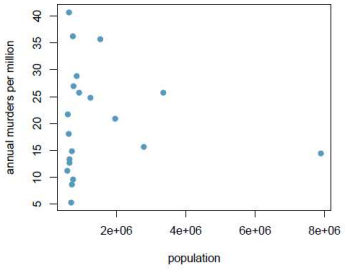


Guessing the correlation

Clicker question

Which of the following is the best guess for the correlation between annual murders per million and population size?

(a) -0.97
(b) -0.61
(c) -0.06
(d) 0.55
(e) 0.97



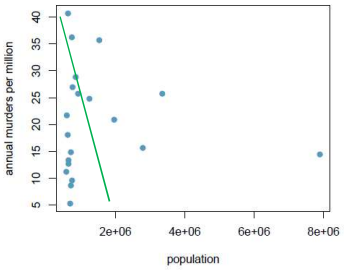
Guessing the correlation

Clicker question

Which of the following is the best guess for the correlation between annual murders per million and population size?

(a) -0.97
(b) -0.61
(c) -0.06
~~(d) 0.55~~
~~(e) 0.97~~

- Downwards Trending Relationship → **Negtive R**



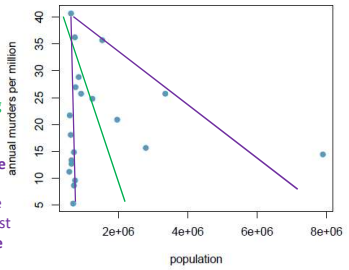
Guessing the correlation

Clicker question

Which of the following is the best guess for the correlation between annual murders per million and population size?

~~(a) -0.97~~
~~(b) -0.61~~
(c) -0.06
~~(d) 0.55~~
~~(e) 0.97~~

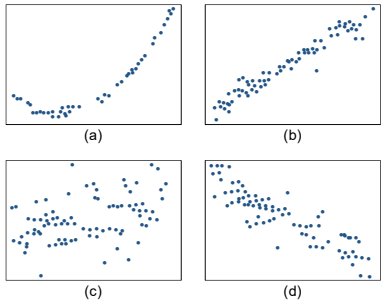
- Downwards Trending Relationship → **Negtive R**
- **Doesn't seem to have a linear relationship AND** many points are far away from the best fit-line. → **|R| will be low!**



Assessing the correlation

Clicker question

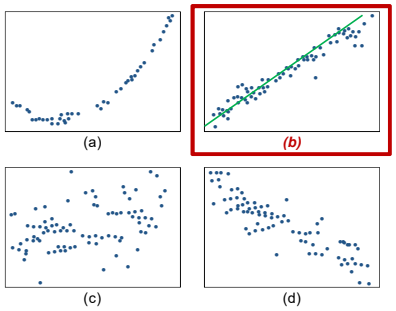
Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



Assessing the correlation

Clicker question

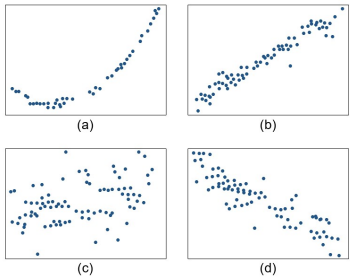
Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



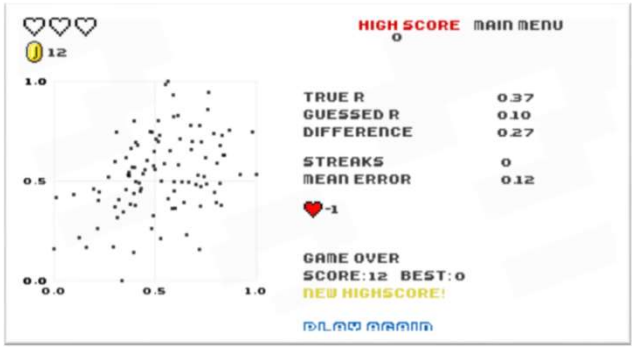
- (a) seems to have the strongest relationship **BUT**
- (b) seems to have the strongest LINEAR relationship…
  Correlation Coefficient (R) *measures the* ***strength of the LINEAR relationship.***

Outline

Need More Practice Guessing R? Extra Credit?

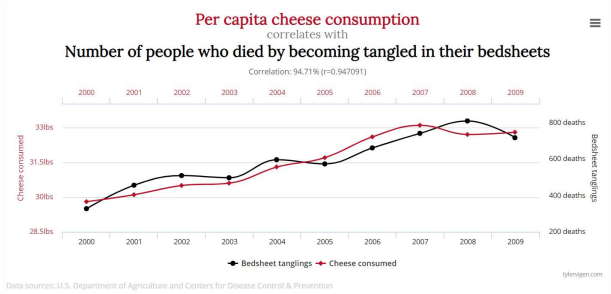


Upload a screen shot with your PS 6 (EC - Correlation Game) for extra credit PS 6 (1 pt on the problem set).
http://guessthecorrelation.com/



Spurious correlations

Remember: correlation does not always imply causation!
http://www.tylervigen.com/



Outline

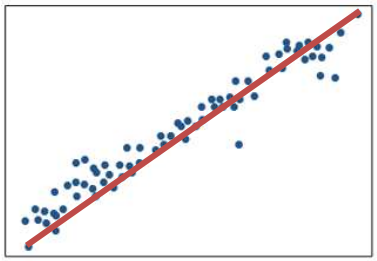
1. Housekeeping

2. Main ideas:
   1. **Quantify** the *Linear* Relationship Between Two Numerical Variables: ***Correlation coefficient***
      - Correlation coefficient describes the strength and direction of the *linear association* between two numerical variables
   2. **Model** the Linear Relationship Between Two Numerical Variables: ***Simple Linear Regression Model***
      - Calculating the Model: Least squares line minimizes squared residuals
      - Interpreting the Model
      - Using the Model: Predict, but don't extrapolate

3. Summary

Outline

What ways can we find a "best fit" line for a set of (x,y) data? How do we represent it?

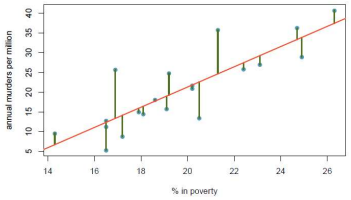


(2) Least squares line minimizes squared residuals

- The **least squares line** minimizes squared residuals.

Models a **Sample** of (x,y) Data

$\hat{y} = b_0 + b_1 x$

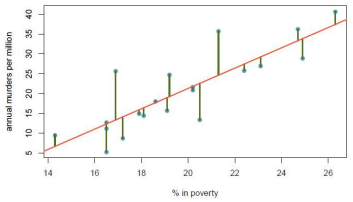


(2) Least squares line minimizes squared residuals

- The **least squares line** minimizes squared residuals.

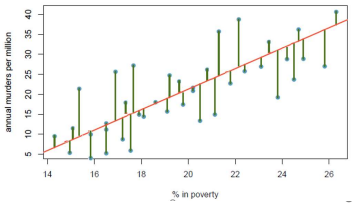
Models a **Sample** of (x,y) Data

$\hat{y} = b_0 + b_1 x$



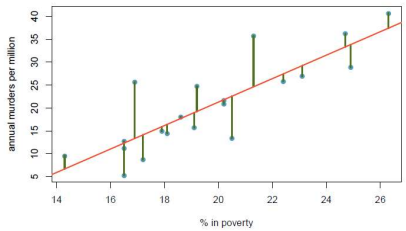
Models the **Population** of (x,y) Data

$\hat{y} = \beta_0 + \beta_1 x$



Outline

How do we assess how well our least squares line "fit" (predicted) an *individual* explanatory variable(s) value (ie x-value)?

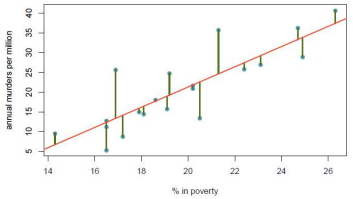


(2) Least squares line minimizes squared residuals

- The **least squares line** minimizes squared residuals.

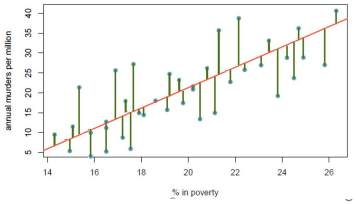
Models a **Sample** of (x,y) Data

$\hat{y} = b_0 + b_1 x$



Models the **Population** of (x,y) Data

$\hat{y} = \beta_0 + \beta_1 x$



- **Residuals** are the leftovers from the model fit, and calculated as the difference between the observed and predicted y-values, for a given x-value (explanatory var(s) value)

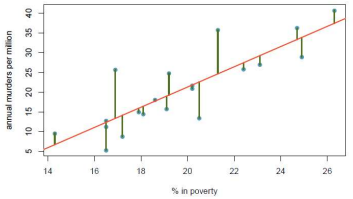
$e_i = y_i - \hat{y}_i$

(2) Least squares line minimizes squared residuals

- The **least squares line** minimizes squared residuals.

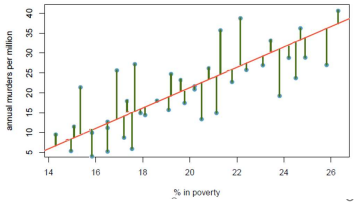
Models a **Sample** of (x,y) Data

$\hat{y} = b_0 + b_1 x$



Models the **Population** of (x,y) Data

$\hat{y} = \beta_0 + \beta_1 x$



- **Residuals** are the leftovers from the model fit, and calculated as the difference between the observed and predicted y-values, for a given x-value (explanatory var(s) value)

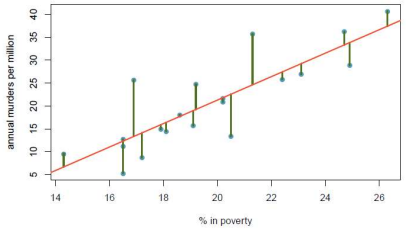
observed y for x_i

$e_i = y_i - \hat{y}_i$ ← predicted y for x_i

Outline

How do we <u>calculate</u> our least squares line?

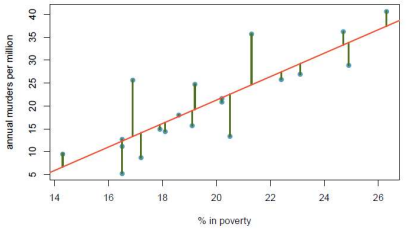
$\hat{y} = b_0 + b_1 x$



(2) Least squares line minimizes squared residuals

$\min \sum_{i=1}^{n} {e_i}^2$

Least <u>Squares</u> Regression

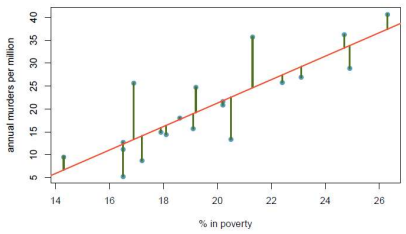


(2) Least squares line minimizes squared residuals

Least Squares Regression

$$\min \sum_{i=1}^{n} {e_i}^2$$

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



(2) Least squares line minimizes squared residuals

Least Squares Regression

$$\min \sum_{i=1}^{n} {e_i}^2$$

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\min \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

What values of b_0 and b_1 give the minimum value of this function?

Some Calculus…
The (b_0, b_1) is the critical point of this function.

(2) Least squares line minimizes squared residuals

Least Squares Regression

$$\min \sum_{i=1}^{n} {e_i}^2$$

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\min \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

What values of b_0 and b_1 give the minimum value of this function?

Slope

$$b_1 = \frac{s_y}{s_x} R$$

Intercept

$$b_0 = \bar{y} - b_1 \bar{x}$$

(2) Least squares line minimizes squared residuals

Least Squares Regression

$$\min \sum_{i=1}^{n} {e_i}^2$$

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\min \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

What values of b_0 and b_1 give the minimum value of this function?

Slope

$$b_1 = \frac{s_y}{s_x} R$$

Sample std dev. of y-values

Sample std dev. of x-values

Intercept

$$b_0 = \bar{y} - b_1 \bar{x}$$

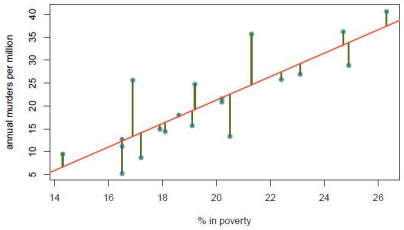
(2) Least squares line minimizes squared residuals

$$\min \sum_{i=1}^{n} |e_i|$$

$$\min \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$\min \sum_{i=1}^{n} |(y_i - (b_0 + b_1 x_i)|$$

Why not use absolute value instead?



(2) Least squares line minimizes squared residuals

$$\min \sum_{i=1}^{n} |e_i|$$

$$\min \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$\min \sum_{i=1}^{n} |(y_i - (b_0 + b_1 x_i)|$$

Why not use absolute value instead?

- Least squares easier for computation.
- Least squares more sensitive to outliers.

Outline

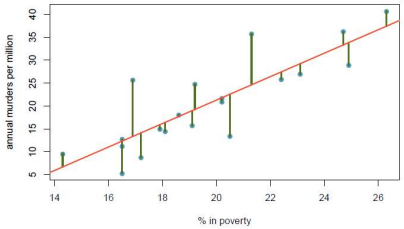
1. Housekeeping

2. Main ideas:
   1. **Quantify** the *Linear* Relationship Between Two Numerical Variables: ***Correlation coefficient***
      - Correlation coefficient describes the strength and direction of the *linear association* between two numerical variables
   2. **Model** the Linear Relationship Between Two Numerical Variables: ***Simple Linear Regression Model***
      - Calculating the Model: Least squares line minimizes squared residuals
      - Interpreting the Model
      - Using the Model: Predict, but don't extrapolate

3. Summary

Outline

How do we interpret our least squares line?



(3) Interpreting the last squares line

$$\hat{y} = b_0 + b_1 x$$

When x is numerical

- *Slope:* For each unit increase in x, y is expected to behigher/lower on average by the slope.

$$b_1 = \frac{s_y}{s_x} R$$

- *Intercept:* When $x = 0$, y is expected to equal the intercept.

$$b_0 = \bar{y} - b_1 \bar{x}$$

  - The calculation of the intercept uses the fact the a regression line **always** passes through $(\bar{x}, \bar{y})$.

(3) Interpreting the last squares line

$$\hat{y} = b_0 + b_1 x$$

When x is numerical

- *Slope:* For each unit increase in x, y is expected to behigher/lower on average by the slope.

$$b_1 = \frac{s_y}{s_x} R$$

- *Intercept:* When $x = 0$, y is expected to equal the intercept.

$$b_0 = \bar{y} - b_1 \bar{x}$$

  - The calculation of the intercept uses the fact the a regression line **always** passes through $(\bar{x}, \bar{y})$.

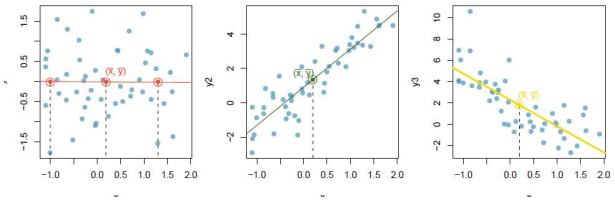
*Important: Make sure your least squares line interpretation does not imply causality.

*double check your Lab 8

Why does the regression line **always** pass through $(\bar{x}, \bar{y})$?

Why does the regression line **always** pass through $(\bar{x}, \bar{y})$?

- If there is no relationship between x and y ($b_1 = 0$), the best guess for $\hat{y}$ for any value of x is $\bar{y}$.
- Even when there is a relationship between x and y ($b_1 \neq 0$), the best guess for $\hat{y}$ when $x = \bar{x}$ is still $\bar{y}$.



Application exercise: 6.1 Linear model

See course website for details

Clicker question

What is the interpretation of the slope?

(a) Each additional percentage in those living in poverty increases number of annual murders per million by 2.56.

(b) For each percentage increase in those living in poverty, the number of annual murders per million is expected to be higher by 2.56 on average.

(c) For each percentage increase in those living in poverty, the number of annual murders per million is expected to be lower by 29.91 on average.

(d) For each percentage increase annual murders per million, the percentage of those living in poverty is expected to be higher by 2.56 on average.

Clicker question

What is the interpretation of the slope?

(a) Each additional percentage in those living in poverty increases number of annual murders per million by2.56.

(b) For each percentage increase in those living in poverty, the number of annual murders per million is expected to be higher by 2.56 on average.

(c) For each percentage increase in those living in poverty, the number of annual murders per million is expected to be lower by 29.91 on average.

(d) For each percentage increase annual murders per million, the percentage of those living in poverty is expected to be higher by 2.56 on average.

**The language in (a) implies a causal relationship. We don't want to say this, as this regression line is calculated from data from a observational study.*

Outline

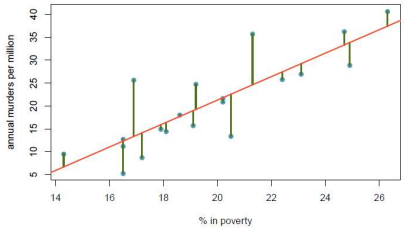
1. Housekeeping

2. Main ideas:
   1. **Quantify** the *Linear* Relationship Between Two Numerical Variables: ***Correlation coefficient***
      - Correlation coefficient describes the strength and direction of the *linear association* between two numerical variables
   2. **Model** the Linear Relationship Between Two Numerical Variables: ***Simple Linear Regression Model***
      - Calculating the Model: Least squares line minimizes squared residuals
      - Interpreting the Model
      - Using the Model: Predict, but don't extrapolate

3. Summary

Outline

How *should* we use our least squares model to make predictions?

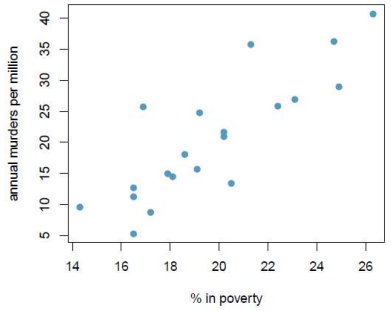


Clicker question

Suppose you want to predict annual murder count (per million) for a series of districts that were not included in the dataset. For which of the following districts would you be most comfortable with your prediction?

A district where % in poverty =

(a) 5%
(b) 15%
(c) 20%
(d) 26%
(e) 40%

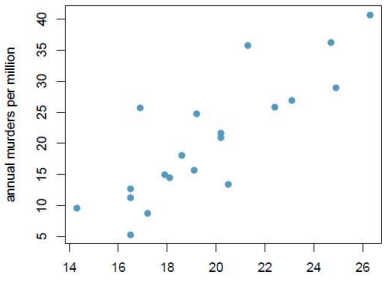


Clicker question

Suppose you want to predict annual murder count (per million) for a series of districts that were not included in the dataset. For which of the following districts would you be most comfortable with your prediction?

A district where % in poverty =

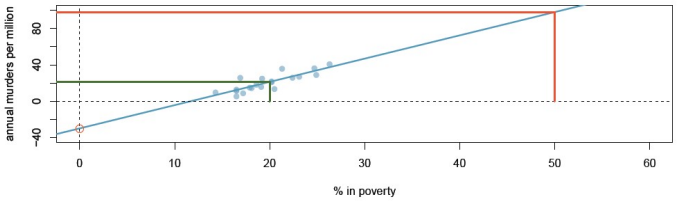
(a) 5%→ extrapolation
(b) 15%
(c) 20% → most center
(d) 26%
(e) 40%→ extrapolation



Extrapolation → Predicting with x-values outside the range of input data.

A note about the intercept

Sometimes the intercept might be an extrapolation: useful for adjusting the height of the line, but meaningless in the context of the data.



Calculating predicted values

By hand: $\widehat{murder} = -29.91 + 2.56\,poverty$

The predicted number of murders per million per year for a county with 20% poverty rate is:

Calculating predicted values

By hand: $\widehat{murder} = -29.91 + 2.56\,poverty$

The predicted number of murders per million per year for a county with 20% poverty rate is:

$$\widehat{murder} = -29.91 + 2.56 \times 20 = 21.29$$

Why did we plug in "20 percent" and not "0.20"?

Calculating predicted values

By hand: $\widehat{murder} = -29.91 + 2.56\,poverty$

The predicted number of murders per million per year for a county with 20% poverty rate is:

$$\widehat{murder} = -29.91 + 2.56 \times 20 = 21.29$$

Why did we plug in "20 percent" and not "0.20"?

We used data in the form of percentages [0,100] (not decimals [0,1]) to calculate b_0 and b_1… **be consistent!**

Calculating predicted values

By hand: $\widehat{murder} = -29.91 + 2.56\,poverty$

The predicted number of murders per million per year for a county with 20% poverty rate is:

$$\widehat{murder} = -29.91 + 2.56 \times 20 = 21.29$$

In R:

```
# load data
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
# fit model
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
# create new data
newdata <- data.frame(perc_pov = 20)
# predict
predict(m_mur_pov, newdata)
```

Calculating predicted values

By hand: $\widehat{murder} = -29.91 + 2.56\, poverty$

The predicted number of murders per million per year for a county with 20% poverty rate is:

$$\widehat{murder} = -29.91 + 2.56 \times 20 = 21.29$$

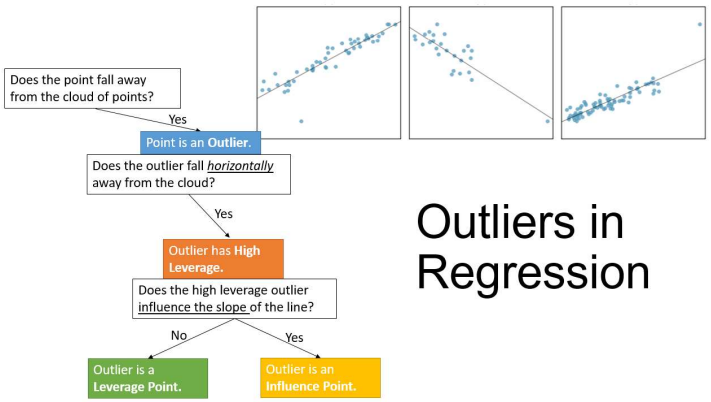
In R:

```
# load data
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
# fit model
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
# create new data
newdata <- data.frame(perc_pov = 20)
# predict
predict(m_mur_pov, newdata)
```

```
       1
21.28663
```

Summary of main ideas

1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables
2. Least squares line minimizes squared residuals
3. Interpreting the least squares line
4. Predict, but don't extrapolate



Outliers in Regression

Using Correlation Coefficient

What we might know	Possible Outcomes
Linear Relationship (from scatter plot) and High |R|	Strong linear association.
Linear Relationship (from scatter plot) and Low |R|	Weak linear association.
High |R|	Could be a strong linear association OR Could be some nonlinear relationship.
Low |R|	Could be a weak linear association OR Could be some nonlinear relationship.
Nonlinear Relationship (from scatter plot)	Could have high |R| OR Could have low R|.