Chapter 7: Simple linear regression

Yingbo Li

Southern Methodist University

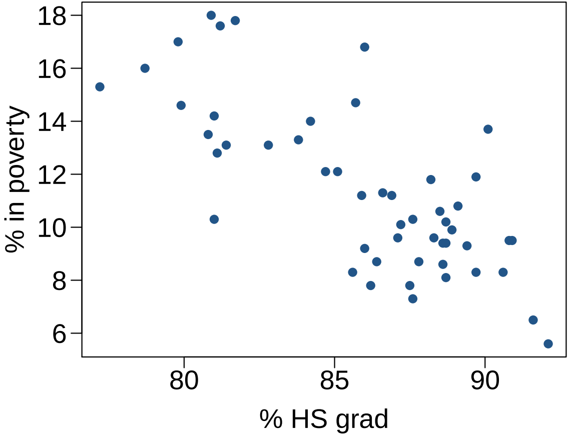
STAT 2331

Outline

1. Introduction to linear regression
2. Fitting a line by least squares regression
3. Inference for linear regression

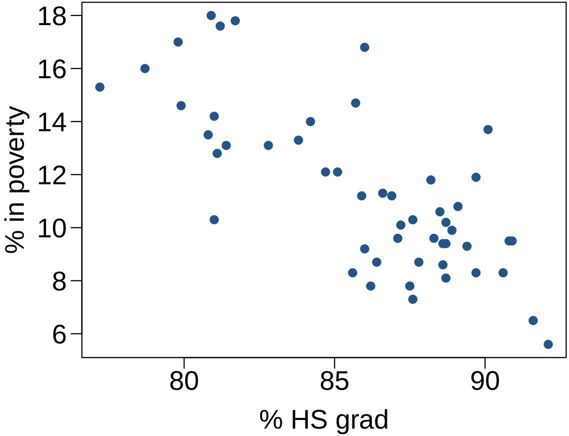
Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in the 51 states in the US (including DC) and the % of residents who live below the poverty line (income below $22,350 for a family of 4).



Response vs. explanatory

Response variable is on the y-axis, and *explanatory* variable is on the x-axis.

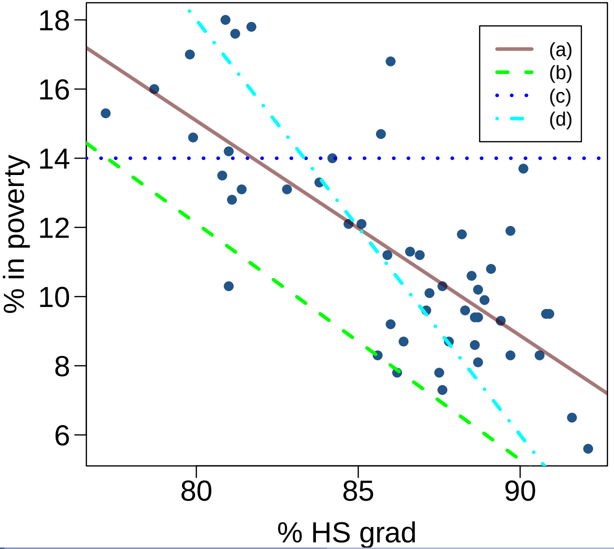


Eyeballing the line

Question

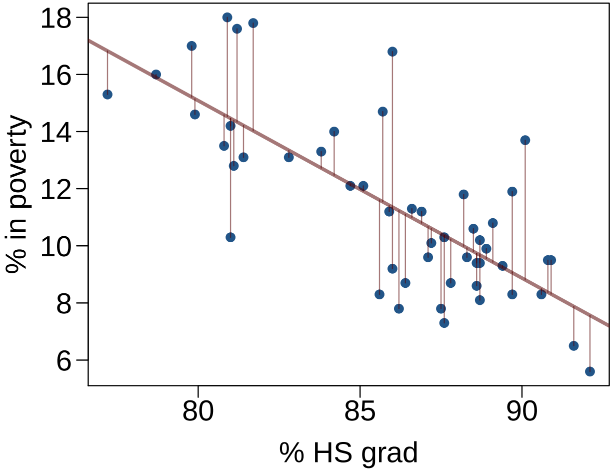
1. Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one.

(a)



Residuals

Residuals are the leftovers from the model fit: Data = Fit + Residual

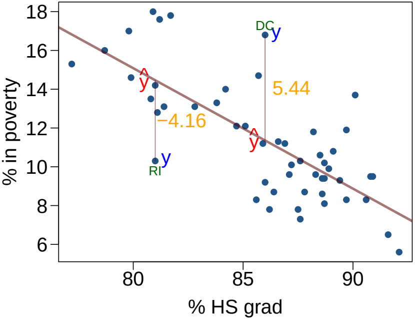


Residuals (cont.)

Residual

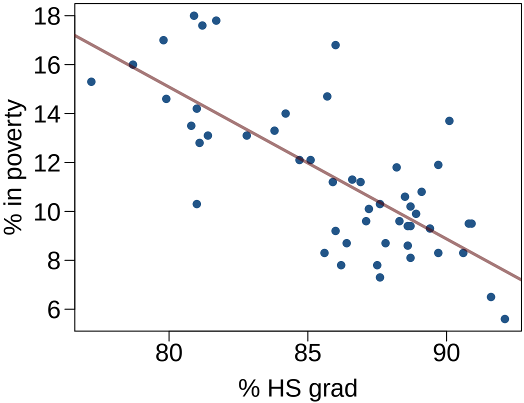
Residual is the difference between the observed and predicted y.

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



- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

Describing the relationship



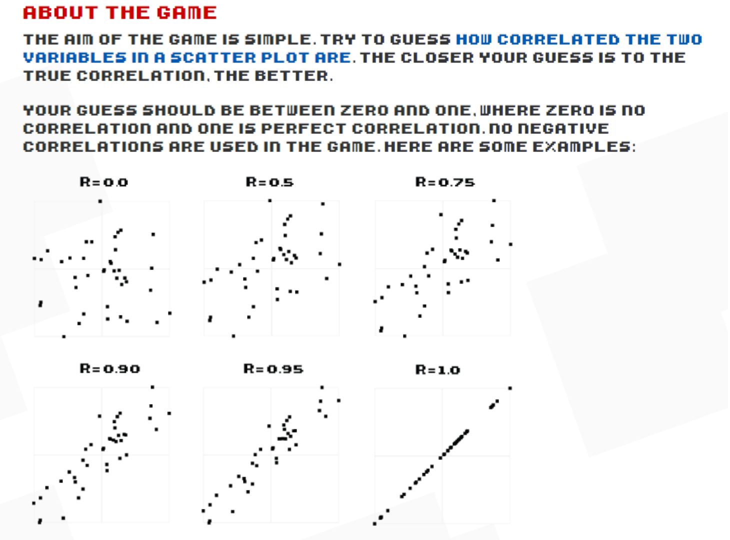
The relationship between % in poverty and % HS grad is

- linear
- negative
- somewhat strong - not a huge amount of scatter around the line

Quantifying the relationship

- *Correlation* describes the strength of the *linear* relationship between two variables.
- It takes values between -1 (perfect negative relationship) and $+1$ (perfect positive relationship).
- A value of 0 indicates no relationship.

Play the game! *http://guessthecorrelation.com/*



ABOUT THE GAME

THE AIM OF THE GAME IS SIMPLE. TRY TO GUESS HOW CORRELATED THE TWO VARIABLES IN A SCATTER PLOT ARE. THE CLOSER YOUR GUESS IS TO THE TRUE CORRELATION, THE BETTER.

YOUR GUESS SHOULD BE BETWEEN ZERO AND ONE, WHERE ZERO IS NO CORRELATION AND ONE IS PERFECT CORRELATION. NO NEGATIVE CORRELATIONS ARE USED IN THE GAME. HERE ARE SOME EXAMPLES:

Guessing the correlation

Question

2. Which of the following is the best guess for the correlation between % in poverty and % HS grad?

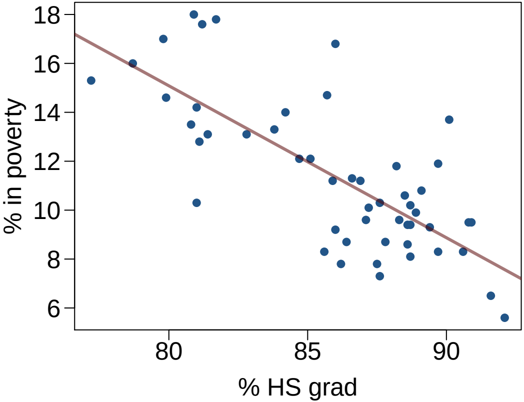
(a) 0.6

(b) *-0.75*

(c) -0.1

(d) 0.02

(e) -1.5



Calculating the correlation

- Using a formula:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

If we change the measurement units, will it affect correlation?

Note: You won't be asked you to calculate the correlation coefficient by hand, because nobody does it by hand. But you might be given a scatterplot and asked to guess the correlation.

Guessing the correlation

Question

3. Which of the following is the best guess for the correlation between % in poverty and % female householder?

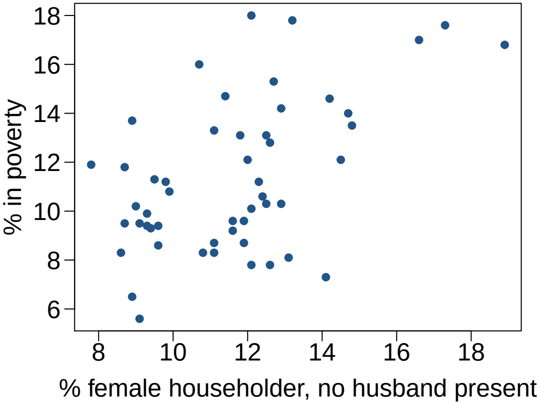
(a) 0.1

(b) -0.6

(c) -0.4

(d) 0.9

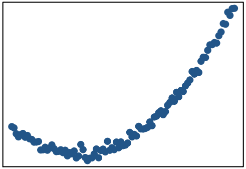
(e) *0.5*



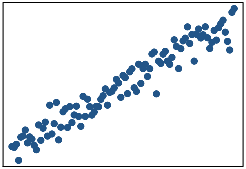
Assessing the correlation

Question

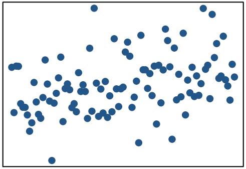
4. Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



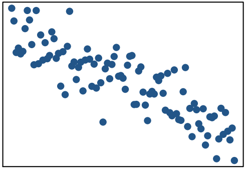
(a)



(b)



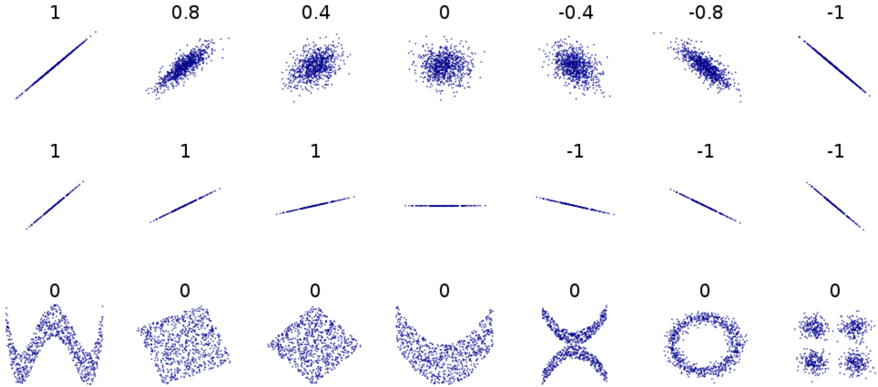
(c)



(d)

(b) → correlation means linear association

Correlations



A measure for the best line

- We want a line that has small residuals
- One option: Minimize the sum of magnitudes (absolute values) of residuals

$$|e_1| + |e_2| + \cdots + |e_n|$$

- Another option: Minimize the sum of squared residuals

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

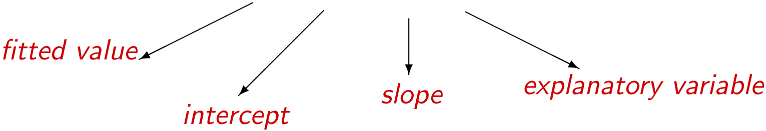
- The line that minimizes the sum of squared residuals is the *least squares line*

Why minimize squares?

1. Most commonly used
2. Easier to compute by hand and using software
3. In many applications, a residual twice as large as another is more than twice as bad

The least squares line

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



fitted value

intercept

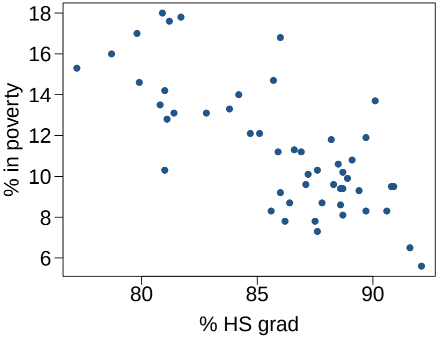
slope

explanatory variable

Notation:

- Intercept:
  - Parameter: β_0
  - Point estimate: b_0
- Slope:
  - Parameter: β_1
  - Point estimate: b_1

Given...



	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

Slope

Slope

The slope of the regression can be calculated as

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

In context...

$$b_1 = \frac{3.1}{3.73} \times (-0.75) = -0.62$$

Interpretation

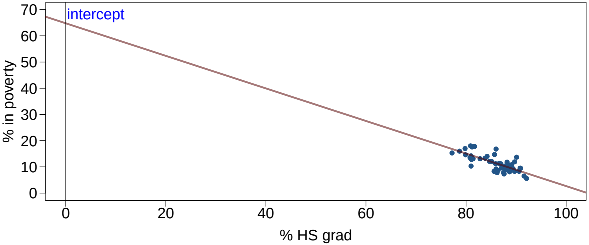
For each % point increase in HS graduate rate, we would *expect* the % living in poverty to decrease *on average* by 0.62% points.

Intercept

Intercept

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

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



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

$$\begin{aligned} b_0 &= 11.35 - (-0.62) \times 86.01 \\ &= 64.68 \end{aligned}$$

Question

5. Which of the following is the correct interpretation of the intercept?

(a) For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(b) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

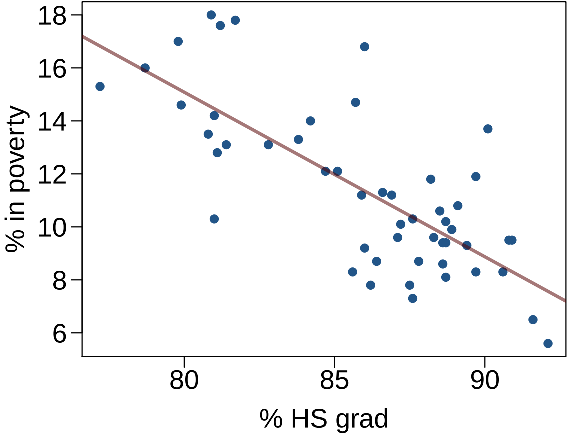
(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

(d) *States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.*

(e) In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

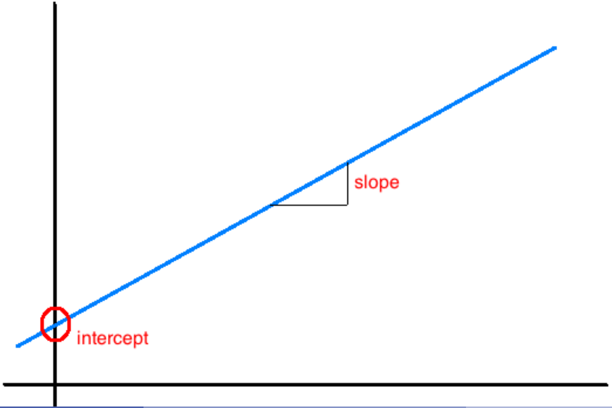
Regression line: $\widehat{y} = b_0 + b_1 x$

$$\widehat{\% \; in \; poverty} = 64.68 - 0.62 \; \% \; HS \; grad$$



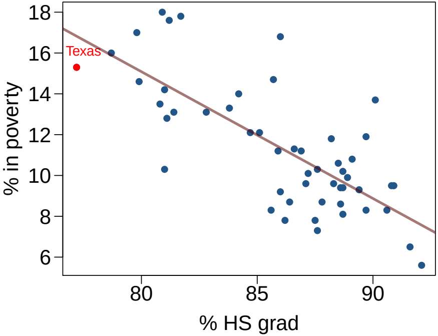
Interpretation of slope and intercept

- *Intercept:* When $\underline{x = 0}$, $\underline{y}$ is expected to equal the intercept.
- *Slope:* For each unit increase in $\underline{x}$, $\underline{y}$ is expected to increase/decrease on average by the slope.



Fitted values and residuals

Texas: % in poverty $= 15.3$, % HS grad = 77.2



$$x = 77.2, \qquad y = 15.3$$

$$\widehat{y} = 64.68 - 0.62 \times 77.2 = 16.82$$

$$e = y - \widehat{y} = 15.3 - 16.82 = -1.52$$

R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.
- R^2 is calculated as the square of the correlation coefficient.
- It tells us *what percent of variability in the response variable is explained by the model.*
- The remainder of the variability is explained by variables not included in the model.
- For the model we've been working with, $R^2 = (-0.75)^2 = 0.56$.

Interpretation of R^2

Question

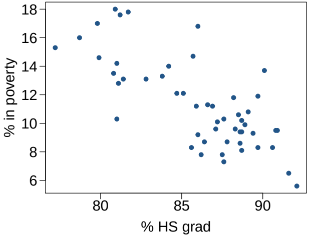
6. Which of the below is the correct interpretation of $R = -0.75$, $R^2 = 0.56$?

(a) 56% of the variability in the % of HG graduates among the 51 states is explained by the model.

(b) *56% of the variability in the % of residents living in poverty among the 51 states is explained by the model.*

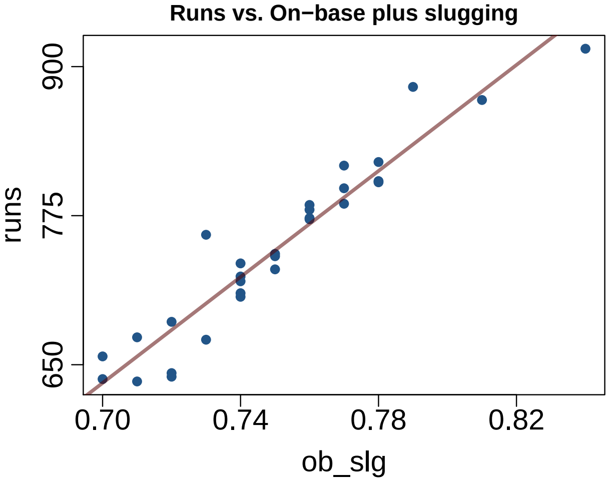
(c) 56% of the time % HS graduates predict % living in poverty correctly.

(d) 44% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



Major league baseball

We can use the 2009 Major League Baseball (MLB) data to predict runs:



Major league baseball

Question

7. R^2 for the regression line for predicting runs from on-base plus slugging is 91.31%. Which of the below is the correct interpretation of this value?

91.31% of

(a) runs can be accurately predicted by on-base plus slugging.

(b) variability in predictions of runs is explained by on-base plus slugging.

(c) variability in predictions of on-base plus slugging is explained by runs.

(d) *variability in runs is explained by on-base plus slugging.*

(e) variability in on-base plus slugging is explained by runs.

Recap

1. Correlation
   - between 2 numerical variables
   - linear association
   - between -1 to 1
2. Use a straight line to fit data points
   - Residual: $e_i = y_i - \hat{y}_i$
   - Find the line: minimizing sum of squared residuals
   - Formula: $\hat{y} = b_0 + b_1 x$
   - Interpret parameters:
     (1) intercept b_0: when x = 0, y is expected to equal the intercept.
     (2) slope b_1: for each unit increase in x, y is expected to increase/decrease on average by the slope.
   - R^2: percent of variability in the response variable that is explained by the model.

Testing for the slope

Question

8. Assuming that the 2009 season is representative of all MLB seasons, we would like to test if these data provide convincing evidence that the slope of the regression line for predicting runs from on-base plus slugging is different than 0. What are the appropriate hypotheses?

(a) $H_0 : b_0 = 0; H_A : b_0 \neq 0$

(b) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$

(c) $H_0 : b_1 = 0; H_A : b_1 \neq 0$

(d) $H_0 : \beta_0 = 0; H_A : \beta_0 \neq 0$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-921	97.38	-9.46	0.0000
ob_slg	2223	129.61	17.15	0.0000

- We always use a t-test in inference for regression

  Remember: Test statistic, $T = \frac{point\ estimate - null\ value}{SE}$

- Point estimate $= b_1$ is the observed slope, and is given in the regression output
- SE_{b_1} is the standard error associated with the slope, and can be calculated as

  $$SE_{b_1} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2/(n-2)}{\sum(x_i - \bar{x}_i)^2}}$$

  is also given in the regression output (and it's silly to try to calculate it by hand, just know that it's doable and why the formula works the way it does)

- Degrees of freedom associated with the slope is $df = n - 2$, where n is the sample size

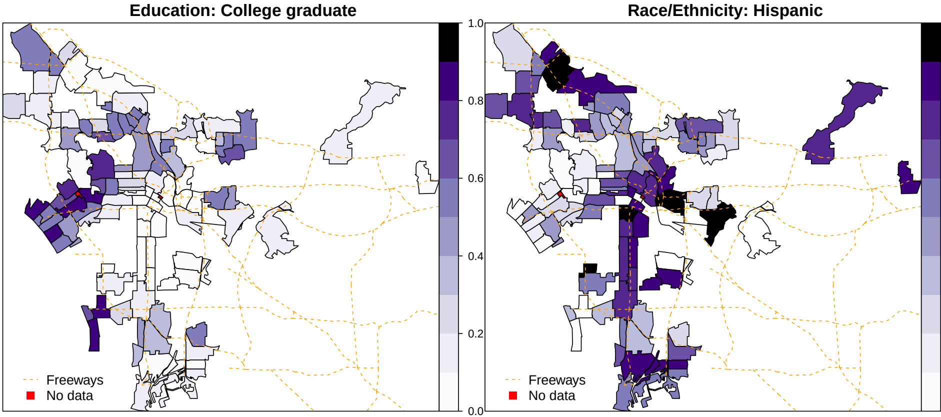
Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-921	97.38	-9.46	0.0000
ob_slg	*2223*	*129.61*	*17.15*	0.0000

$$\begin{aligned} T &= \frac{2223 - 0}{129.6116} = 17.15 \\ df &= 30 - 2 = 28 \\ \text{p-value} &= P(|T| > 17.15) < 0.01 \end{aligned}$$

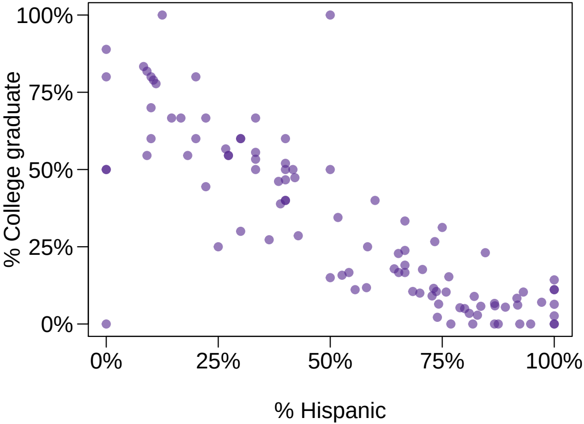
% College graduate vs. % Hispanic in LA

What can you say about the relationship between of % college graduate and % Hispanic in a sample of 100 zip code areas in LA?



% College educated vs. % Hispanic in LA - another look

What can you say about the relationship between of % college graduate and % Hispanic in a sample of 100 zip code areas in LA?



% College educated vs. % Hispanic in LA - linear model

Question

9. Which of the below is the best interpretation of the slope?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7290	0.0308	23.68	0.0000
%Hispanic	-0.7527	0.0501	-15.01	0.0000

(a) A 1% increase in Hispanic residents in a zip code area in LA is associated with a 75% decrease in % of college grads.

(b) *A 1% increase in Hispanic residents in a zip code area in LA is associated with a 0.75% decrease in % of college grads.*

(c) An additional 1% of Hispanic residents decreases the % of college graduates in a zip code area in LA by 0.75%.

(d) In zip code areas with no Hispanic residents, % of college graduates is expected to be 75%.

% College educated vs. % Hispanic in LA - linear model

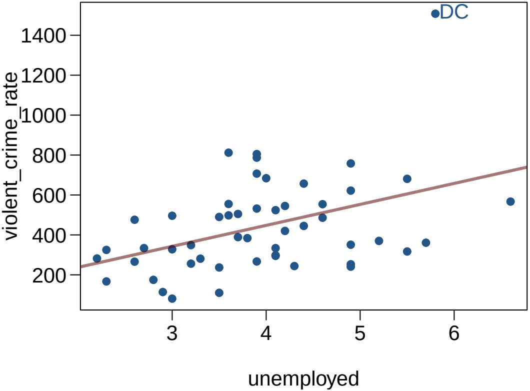
Do these data provide convincing evidence that there is a statistically significant relationship between % Hispanic and % college graduates in zip code areas in LA?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7290	0.0308	23.68	0.0000
hispanic	-0.7527	0.0501	-15.01	0.0000

Yes, the p-value for % Hispanic is low, indicating that the data provide convincing evidence that the slope parameter is different than 0.

Violent crime rate vs. unemployment

Relationship between violent crime rate (annual number of violent crimes per 100,000 population) and unemployment rate (% of work eligible population not working) in 51 US States (including DC):



Note: The data are from the 2003 Statistical Abstract of the US. A 2012 version is available online, if looking for data on states for your project, it's a good resource.

Violent crime rate vs. unemployment

Question

10. Which of the below is the correct set of hypotheses and the p-value for testing if the slope of the relationship between violent crime rate and unemployment is positive?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.68	130.00	0.21	0.8323
unemployed	105.03	32.04	3.28	0.0019

(a) $H_0 : b_1 = 0$ $H_A : b_1 \neq 0$ p-value $= 0.0019$

(b) $H_0 : \beta_1 = 0$ $H_A : \beta_1 > 0$ $p\text{-}value = 0.0019/2 = 0.00095$

(c) $H_0 : \beta_1 = 0$ $H_A : \beta_1 \neq 0$ p-value $= 0.0019/2 = 0.00095$

(d) $H_0 : b_1 = 0$ $H_A : b_1 > 0$ p-value $= 0.0019/2 = 0.00095$

(e) $H_0 : \beta_1 = 0$ $H_A : \beta_1 \neq 0$ p-value $= 0.8323$

Confidence interval for the slope

Question

11. Remember that a confidence interval is calculated as $point\ estimate \pm ME$ and the degrees of freedom associated with the slope in a simple linear regression is $n-2$. Which of the below is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 51 states.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.68	130.00	0.21	0.8323
unemployed	105.03	32.04	3.28	0.0019

(a) $27.68 \pm 1.65 \times 32.04$

(b) $105.03 \pm 2.01 \times 32.04$

(c) $105.03 \pm 1.96 \times 32.04$

(d) $27.68 \pm 1.96 \times 32.04$

$$\begin{aligned} n &= 51 \qquad df = 51-2 = 49 \\ 95\%:\ t^{\star}_{49} &= 2.01 \\ 105.03 &\pm 2.01 \times 32.04 \\ (40.63 &,\ 169.43) \end{aligned}$$

Recap

- Inference for the slope for a SLR model (only one explanatory variable):
  - Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

  - Confidence interval:

$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$

- The null value is often 0 since we are usually checking for *any* relationship between the explanatory and the response variable
- The regression output gives b_1, SE_{b_1}, and *two-tailed* p-value for the t-test for the slope where the null value is 0
- We rarely do inference on the intercept, so we'll be focusing on the estimates and inference for the slope