CHAPTER 6: INTRODUCTION TO LINEAR REGRESSION

math 189 : data analysis and inference : winter 2019

Jelena Bradic
http://www.math.ucsd.edu/~jbradic/
Associate Professor, Department of Mathematics, University of California, San Diego
jbradic@ucsd.edu

Line fitting, residuals, and correlation

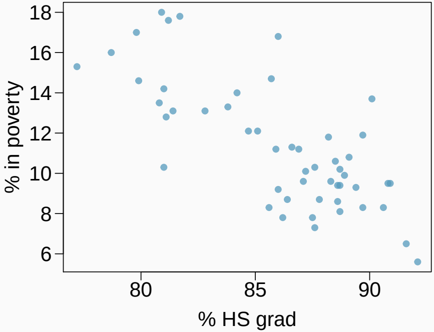
Fitting a line by least squares regression

Types of outliers in linear regression

Inference for linear regression

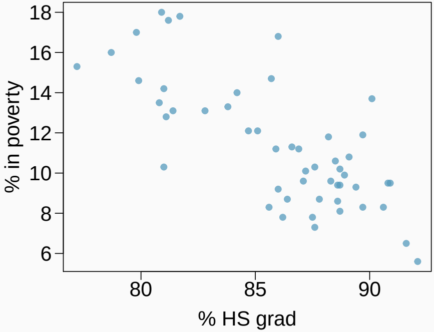
In this unit we will learn to quantify the relationship between two numerical variables, as well as modeling numerical response variables using a numerical or categorical explanatory variable.

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).

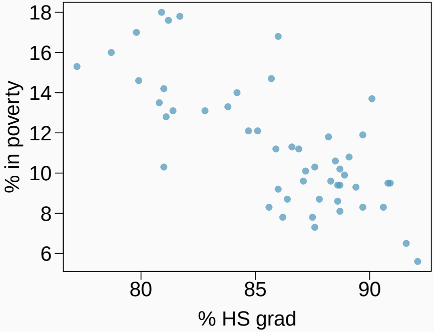


Response variable?

% in poverty

Poverty vs. HS graduate rate

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



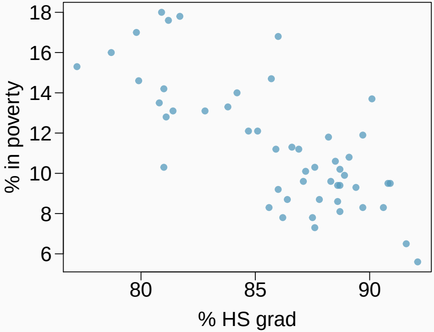
Response variable?

% in poverty

Explanatory variable?

Poverty vs. HS graduate rate

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

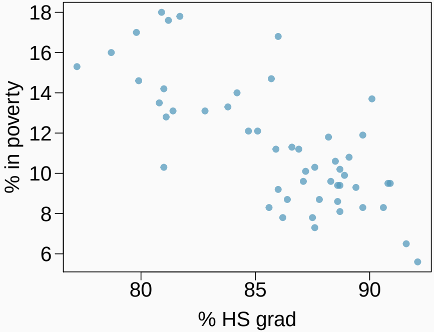
% in poverty

Explanatory variable?

% HS grad

Poverty vs. HS graduate rate

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

% in poverty

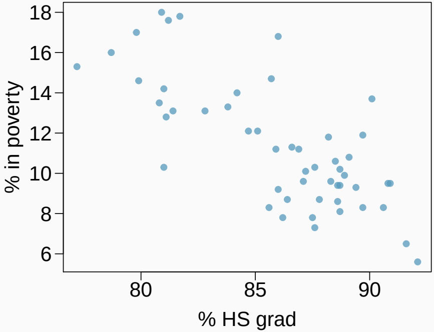
Explanatory variable?

% HS grad

Relationship?

Poverty vs. HS graduate rate

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

% in poverty

Explanatory variable?

% HS grad

Relationship?

linear, negative, moderately strong

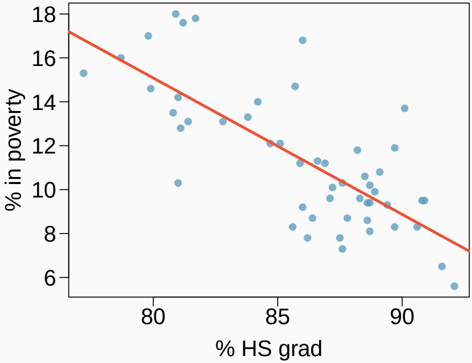
* Correlation describes the strength of the linear association between two variables.

* Correlation describes the strength of the linear association between two variables.
* It takes values between -1 (perfect negative) and +1 (perfect positive).

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

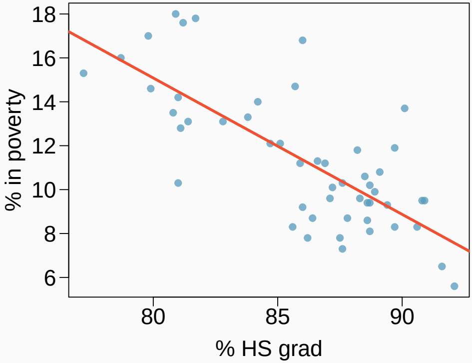
Which of the following is the best guess for the correlation between % in poverty and % HS grad?

* 0.6
* -0.75
* -0.1
* 0.02
* -1.5



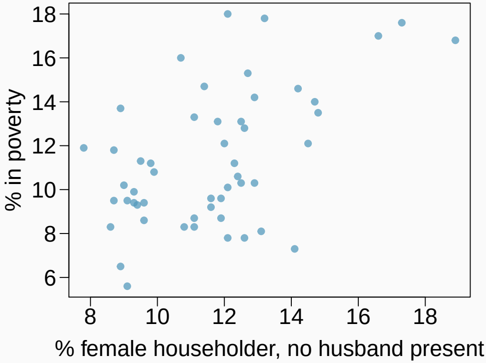
Which of the following is the best guess for the correlation between % in poverty and % HS grad?

* 0.6
* -0.75
* -0.1
* 0.02
* -1.5



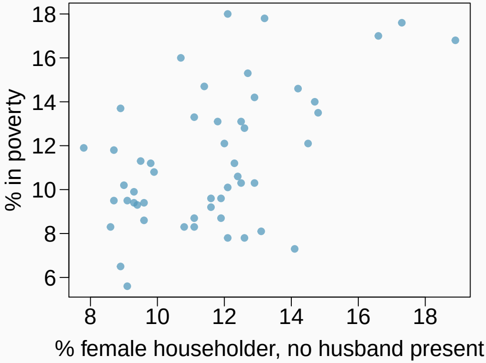
Which of the following is the best guess for the correlation between % in poverty and % HS grad?

* 0.1
* -0.6
* -0.4
* 0.9
* 0.5

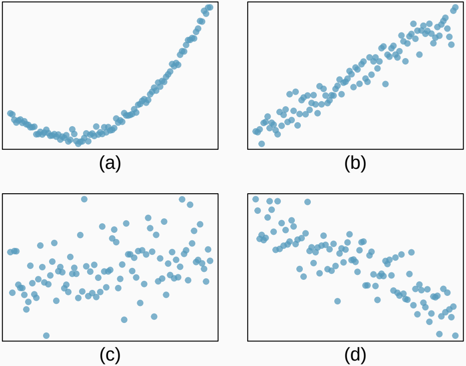


Which of the following is the best guess for the correlation between % in poverty and % HS grad?

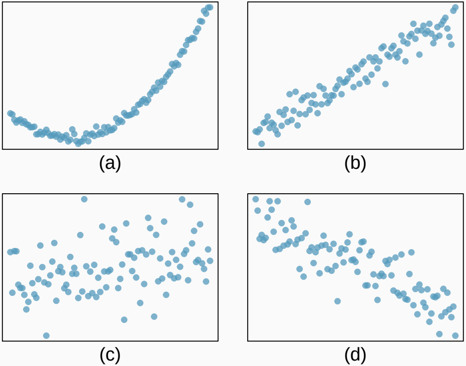
* 0.1
* -0.6
* -0.4
* 0.9
* 0.5



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

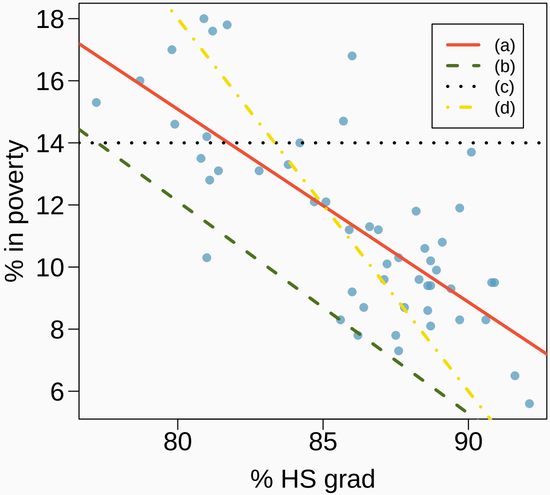


Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



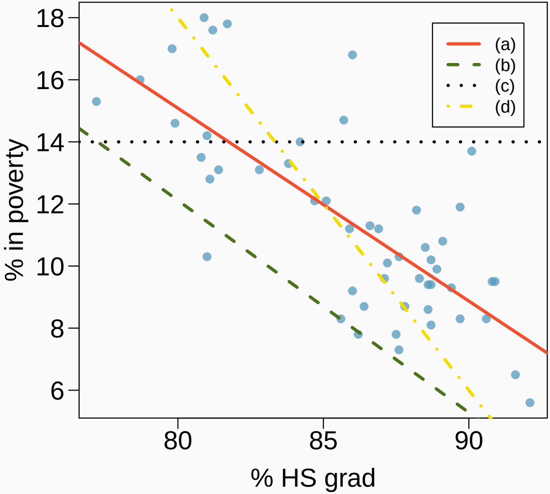
(b) → correlation means <u>linear</u> association

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

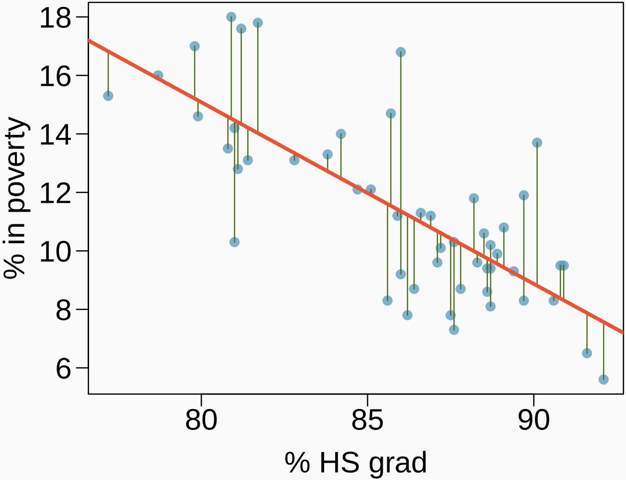


Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one.

(a)



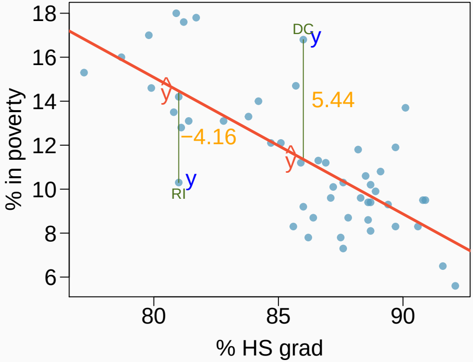
Residuals are the leftovers from the model fit: Data = Fit + Residual



Residual

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

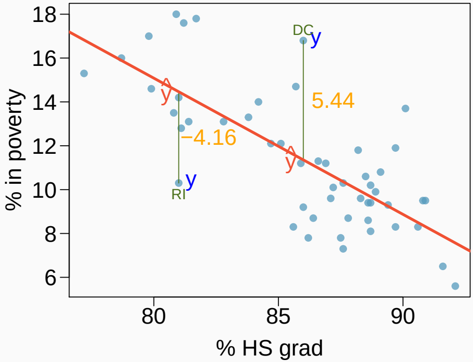
$$e_i = y_i - \hat{y}_i$$



Residual

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

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

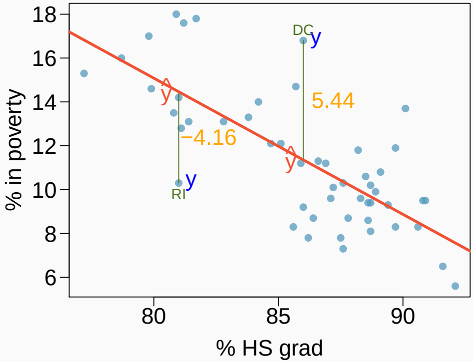


- % living in poverty in DC is 5.44% more than predicted.

Residual

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

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



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

* We want a line that has small residuals:

* We want a line that has small residuals:
  * Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

* We want a line that has small residuals:
  * Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

  * Option 2: Minimize the sum of squared residuals – least squares

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

* We want a line that has small residuals:
  * Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

  * Option 2: Minimize the sum of squared residuals – least squares

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

* Why least squares?

A measure for the best line

* We want a line that has small residuals:
  * Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

  * Option 2: Minimize the sum of squared residuals – least squares

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

* Why least squares?
  * Most commonly used

A measure for the best line

* We want a line that has small residuals:
    * Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

    * Option 2: Minimize the sum of squared residuals – least squares

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

* Why least squares?
    * Most commonly used
    * Easier to compute by hand and using software

A measure for the best line

* We want a line that has small residuals:
  * Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

  * Option 2: Minimize the sum of squared residuals – least squares

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

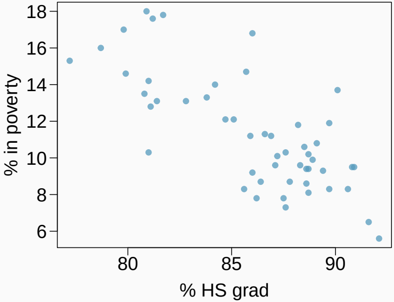
* Why least squares?
  * Most commonly used
  * Easier to compute by hand and using software
  * In many applications, a residual twice as large as another is usually more than twice as bad

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

- predicted y
- intercept
- slope
- explanatory variable

Notation:

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



	% 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

The slope of the regression can be calculated as

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

Slope

The slope of the regression can be calculated as

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

In context...

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

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 additional % point in HS graduate rate, we would expect the % living in poverty to be lower on average by 0.62% points.

Intercept

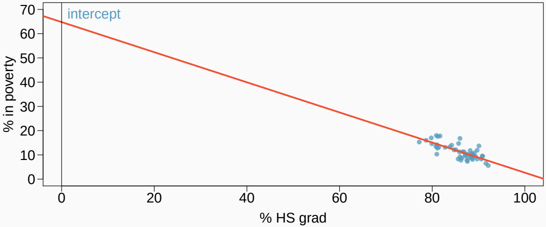
The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

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

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 $(\bar{x}, \bar{y})$.

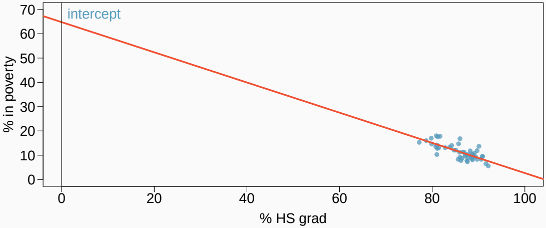
$$b_0 = \bar{y} - b_1\bar{x}$$



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 $(\bar{x}, \bar{y})$.

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



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

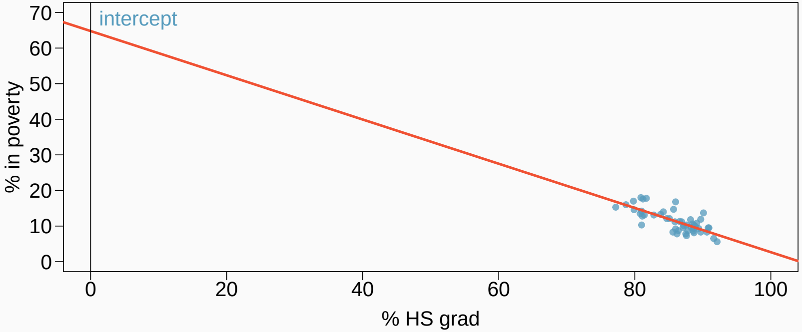
Which of the following is the correct interpretation of the intercept?

* For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.
* For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.
* Having no HS graduates leads to 64.68% of residents living below the poverty line.
* States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.
* In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

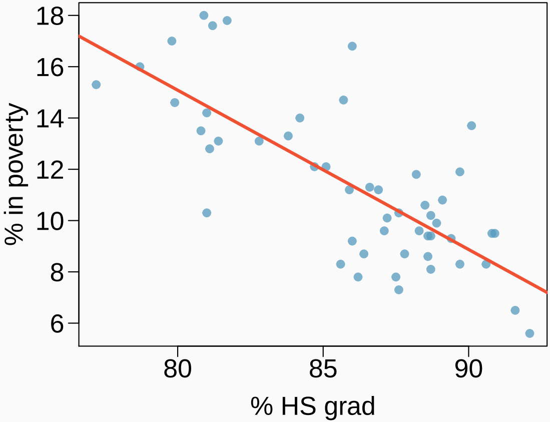
Which of the following is the correct interpretation of the intercept?

* For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.
* For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.
* Having no HS graduates leads to 64.68% of residents living below the poverty line.
* States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.
* In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

Since there are no states in the dataset with no HS graduates, the intercept is of no interest, not very useful, and also not reliable since the predicted value of the intercept is so far from the bulk of the data.

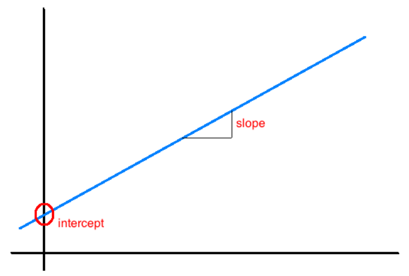


$$\widehat{\%\ \text{in poverty}} = 64.68 - 0.62\ \%\ \text{HS grad}$$



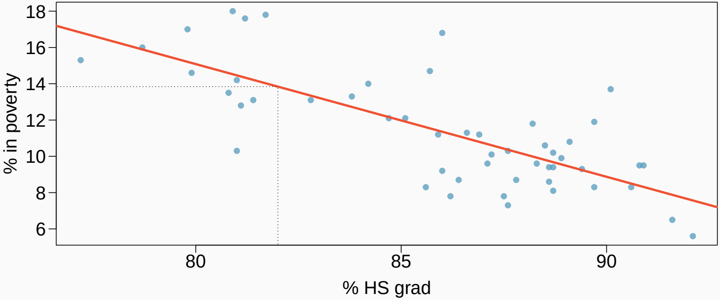
* Intercept: When $x = 0$, y is expected to equal the intercept.

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

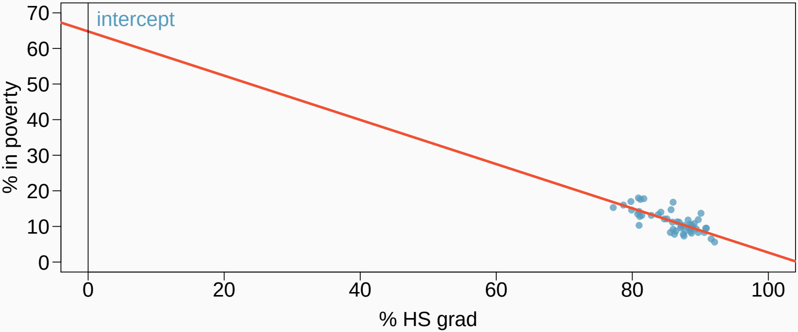


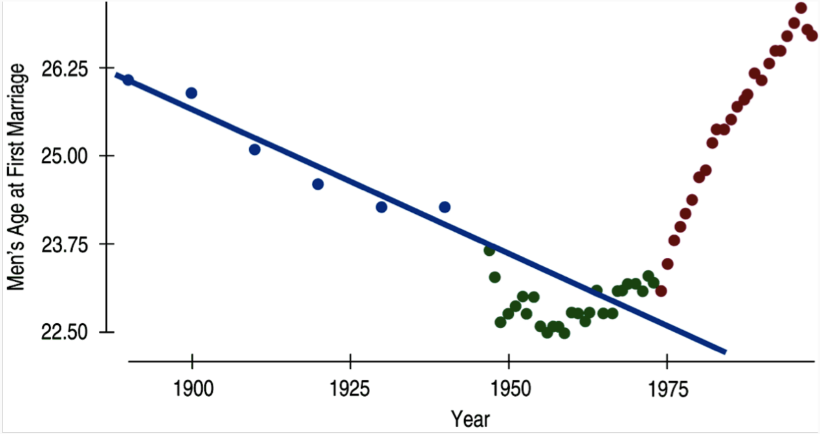
Note: These statements are not causal, unless the study is a randomized controlled experiment.

* Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called prediction, simply by plugging in the value of x in the linear model equation.
* There will be some uncertainty associated with the predicted value.



* Applying a model estimate to values outside of the realm of the original data is called extrapolation.
* Sometimes the intercept might be an extrapolation.




Men's Age at First Marriage
26.25
25.00
23.75
22.50
1900
1925
1950
1975
Year

Examples of extrapolation

BBC NEWS

Watch One-Minute World News

News Front Page

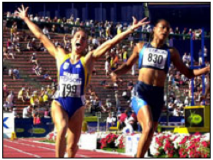
- Africa
- Americas
- Asia-Pacific
- Europe
- Middle East
- South Asia
- UK
  - England
  - Northern Ireland
  - Scotland
  - Wales
  - UK Politics
  - Education
  - Magazine
- Business
- Health
- Science & Environment
- Technology
- Entertainment
- Also in the news

Last Updated: Thursday, 30 September, 2004, 04:04 GMT 05:04 UK

E-mail this to a friend   Printable version

Women 'may outsprint men by 2156'

Women sprinters may be outrunning men in the 2156 Olympics if they continue to close the gap at the rate they are doing, according to scientists.



Women are set to become the dominant sprinters

An Oxford University study found that women are running faster than they have ever done over 100m.

At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.

The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.

However, former British Olympic sprinter Derek Redmond told the BBC: "I find it difficult to believe.

"I can see the gap closing between men and women but I can't necessarily see it being overtaken because mens' times are also going to improve."

Examples of extrapolation

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

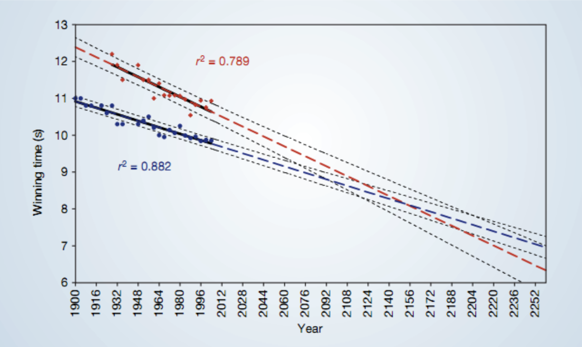


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

* Linearity

Conditions for the least squares line

* Linearity
* Nearly normal residuals

Conditions for the least squares line

* Linearity
* Nearly normal residuals
* Constant variability

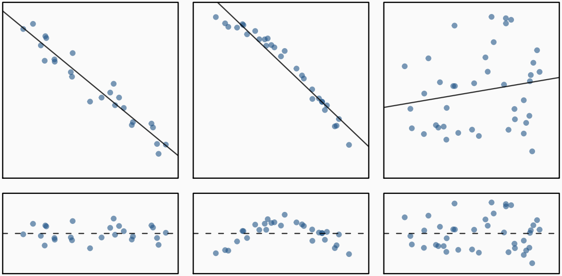
Conditions: (1) Linearity

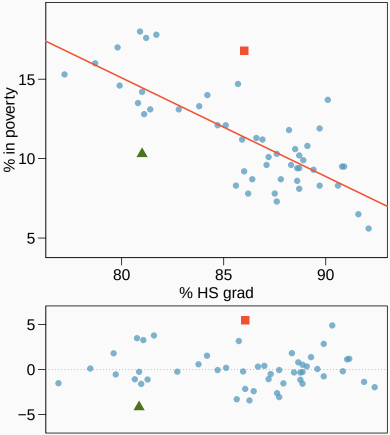
* The relationship between the explanatory and the response variable should be linear.

Conditions: (1) Linearity

* The relationship between the explanatory and the response variable should be linear.
* Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class. If this topic is of interest, an Online Extra is available on openintro.org covering new techniques.

* The relationship between the explanatory and the response variable should be linear.
* Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class. If this topic is of interest, an Online Extra is available on openintro.org covering new techniques.
* Check using a scatterplot of the data, or a residuals plot.





▲ RI:

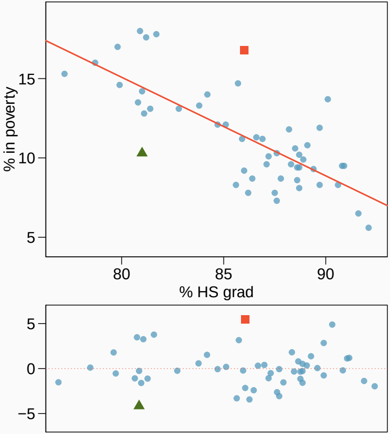
$$\% \text{ HS grad} = 81 \qquad \% \text{ in poverty} = 10.3$$

$$\widehat{\% \text{ in poverty}} = 64.68 - 0.62 * 81 = 14.46$$

$$e = \% \text{ in poverty} - \widehat{\% \text{ in poverty}}$$

$$= 10.3 - 14.46 = -4.16$$

Anatomy of a residuals plot



▲ RI:

$$\% \text{ HS grad} = 81 \qquad \% \text{ in poverty} = 10.3$$

$$\widehat{\% \text{ in poverty}} = 64.68 - 0.62 * 81 = 14.46$$

$$e = \% \text{ in poverty} - \widehat{\% \text{ in poverty}}$$

$$= 10.3 - 14.46 = -4.16$$

■ DC:

$$\% \text{ HS grad} = 86 \qquad \% \text{ in poverty} = 16.8$$

$$\widehat{\% \text{ in poverty}} = 64.68 - 0.62 * 86 = 11.36$$

$$e = \% \text{ in poverty} - \widehat{\% \text{ in poverty}}$$

$$= 16.8 - 11.36 = 5.44$$

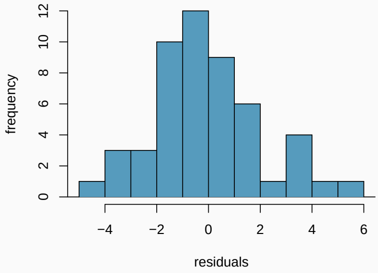
Conditions: (2) Nearly normal residuals

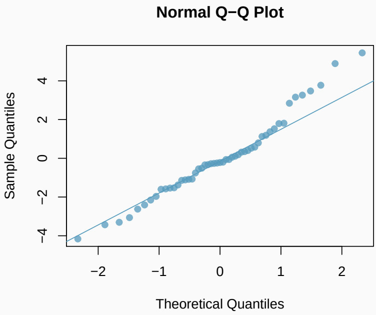
* The residuals should be nearly normal.

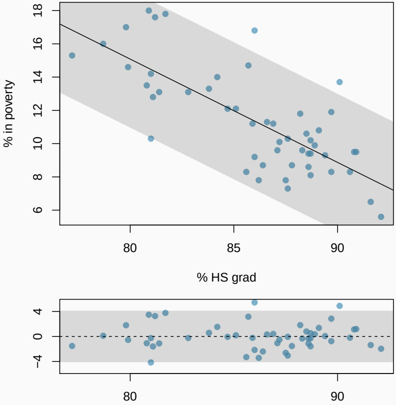
Conditions: (2) Nearly normal residuals

* The residuals should be nearly normal.
* This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.

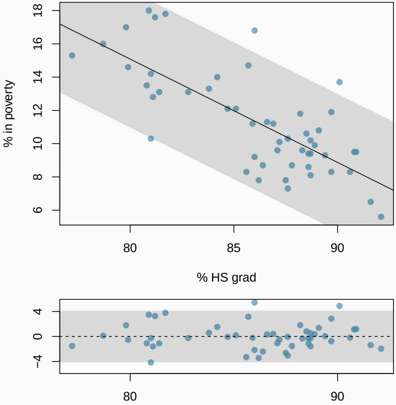
* The residuals should be nearly normal.
* This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
* Check using a histogram or normal probability plot of residuals.



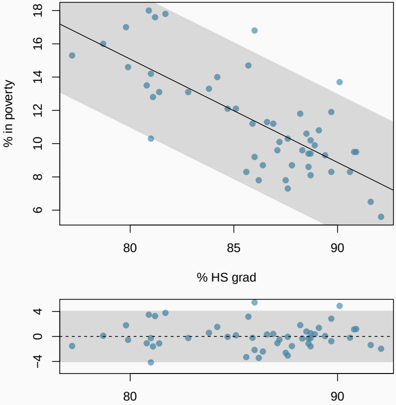




* The variability of points around the least squares line should be roughly constant.

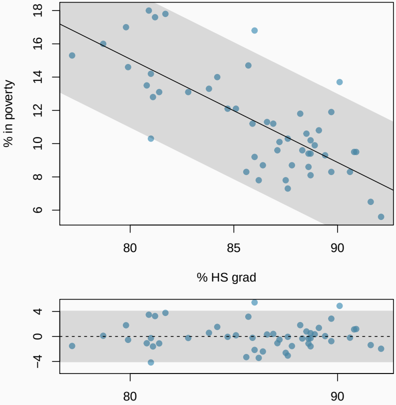


- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.



* The variability of points around the least squares line should be roughly constant.
* This implies that the variability of residuals around the 0 line should be roughly constant as well.
* Also called homoscedasticity.

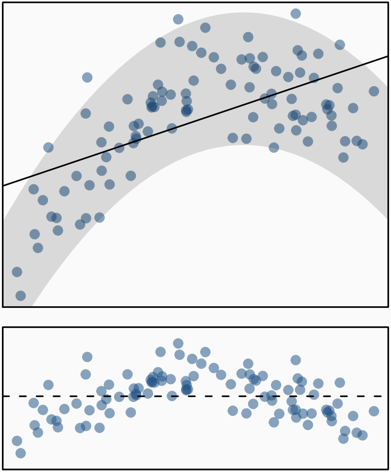
Conditions: (3) Constant variability



* The variability of points around the least squares line should be roughly constant.
* This implies that the variability of residuals around the 0 line should be roughly constant as well.
* Also called homoscedasticity.
* Check using a histogram or normal probability plot of residuals.

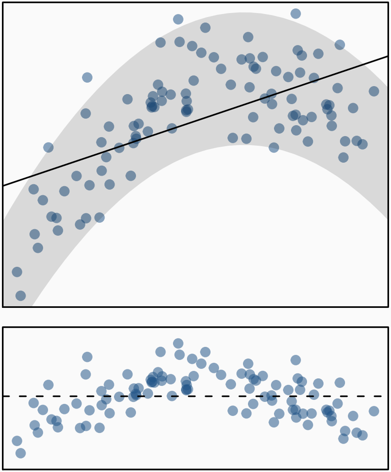
What condition is this linear model obviously violating?

* Constant variability
* Linear relationship
* Normal residuals
* No extreme outliers



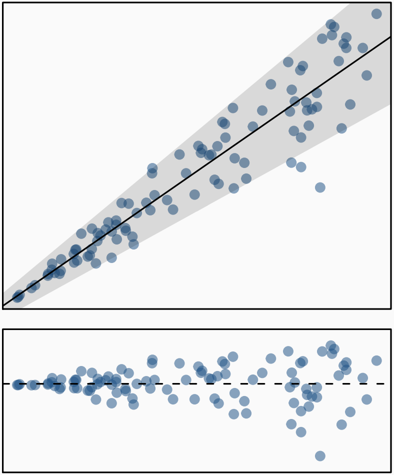
What condition is this linear model obviously violating?

* Constant variability
* Linear relationship
* Normal residuals
* No extreme outliers



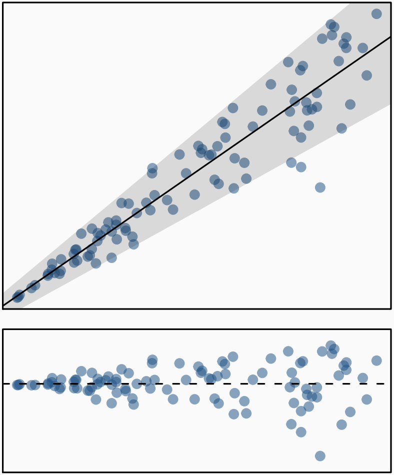
What condition is this linear model obviously violating?

* Constant variability
* Linear relationship
* Normal residuals
* No extreme outliers



What condition is this linear model obviously violating?

* Constant variability
* Linear relationship
* Normal residuals
* No extreme outliers



R^2

* The strength of the fit of a linear model is most commonly evaluated using R^2.

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.

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.

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 or by inherent randomness in the data.

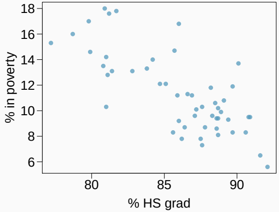
R^2

* 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 or by inherent randomness in the data.
* For the model we've been working with, $R^2 = -0.62^2 = 0.38$.

Interpretation of R^2

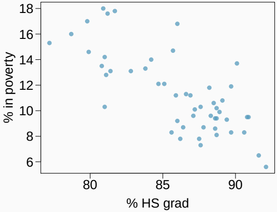
Which of the below is the correct interpretation of $R = -0.62$, $R^2 = 0.38$?

* 38% of the variability in the % of HG graduates among the 51 states is explained by the model.
* 38% of the variability in the % of residents living in poverty among the 51 states is explained by the model.
* 38% of the time % HS graduates predict % living in poverty correctly.
* 62% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



Which of the below is the correct interpretation of $R = -0.62$, $R^2 = 0.38$?

* 38% of the variability in the % of HG graduates among the 51 states is explained by the model.
* 38% of the variability in the % of residents living in poverty among the 51 states is explained by the model.
* 38% of the time % HS graduates predict % living in poverty correctly.
* 62% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



Poverty vs. region (east, west)

$$\widehat{\text{poverty}} = 11.17 + 0.38 \times \text{west}$$

* Explanatory variable: region, reference level: east
* Intercept: The estimated average poverty percentage in eastern states is 11.17%

$$\widehat{\text{poverty}} = 11.17 + 0.38 \times \text{west}$$

* Explanatory variable: region, reference level: east
* Intercept: The estimated average poverty percentage in eastern states is 11.17%
  * This is the value we get if we plug in 0 for the explanatory variable

Poverty vs. region (east, west)

$$\widehat{\text{poverty}} = 11.17 + 0.38 \times \text{west}$$

* Explanatory variable: region, reference level: east
* Intercept: The estimated average poverty percentage in eastern states is 11.17%
    * This is the value we get if we plug in 0 for the explanatory variable
* Slope: The estimated average poverty percentage in western states is 0.38% higher than eastern states.

Poverty vs. region (east, west)

$$\widehat{\text{poverty}} = 11.17 + 0.38 \times \text{west}$$

* Explanatory variable: region, reference level: east
* Intercept: The estimated average poverty percentage in eastern states is 11.17%
  * This is the value we get if we plug in 0 for the explanatory variable
* Slope: The estimated average poverty percentage in western states is 0.38% higher than eastern states.
  * Then, the estimated average poverty percentage in western states is 11.17 + 0.38 = 11.55%.

Poverty vs. region (east, west)

$$\widehat{\text{poverty}} = 11.17 + 0.38 \times \text{west}$$

* Explanatory variable: region, reference level: east
* Intercept: The estimated average poverty percentage in eastern states is 11.17%
  * This is the value we get if we plug in 0 for the explanatory variable
* Slope: The estimated average poverty percentage in western states is 0.38% higher than eastern states.
  * Then, the estimated average poverty percentage in western states is 11.17 + 0.38 = 11.55%.
  * This is the value we get if we plug in 1 for the explanatory variable

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) is the reference level?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

* northeast
* midwest
* west
* south
* cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) is the reference level?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

* northeast
* midwest
* west
* south
* cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

* northeast
* midwest
* west
* south
* cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

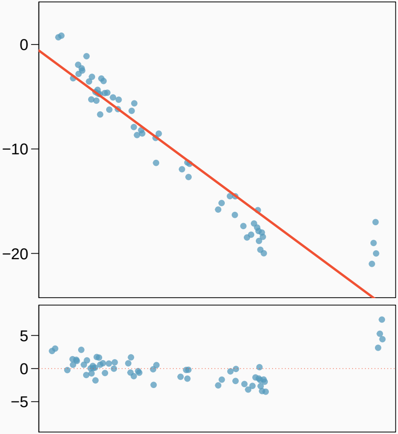
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

* northeast
* midwest
* west
* south
* cannot tell

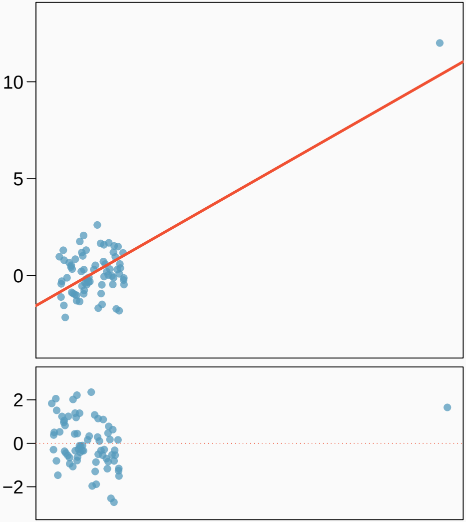
Types of outliers

How do outliers influence the least squares line in this plot?

To answer this question think of where the regression line would be with and without the outlier(s). Without the outliers the regression line would be steeper, and lie closer to the larger group of observations. With the outliers the line is pulled up and away from some of the observations in the larger group.

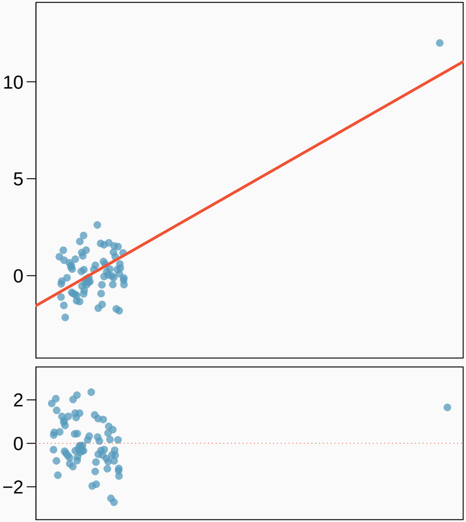


How do outliers influence the least squares line in this plot?



How do outliers influence the least squares line in this plot?

Without the outlier there is no evident relationship between x and y.



Some terminology

* Outliers are points that lie away from the cloud of points.

Some terminology

* Outliers are points that lie away from the cloud of points.
* Outliers that lie horizontally away from the center of the cloud are called high leverage points.

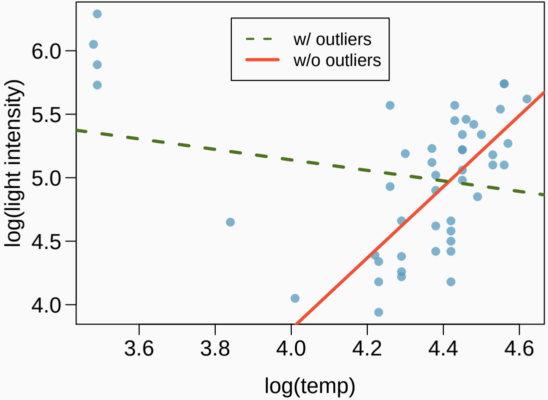
Some terminology

* Outliers are points that lie away from the cloud of points.
* Outliers that lie horizontally away from the center of the cloud are called high leverage points.
* High leverage points that actually influence the slope of the regression line are called influential points.

Some terminology

* Outliers are points that lie away from the cloud of points.
* Outliers that lie horizontally away from the center of the cloud are called high leverage points.
* High leverage points that actually influence the slope of the regression line are called influential points.
* In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential. If not, then itnot an influential point.

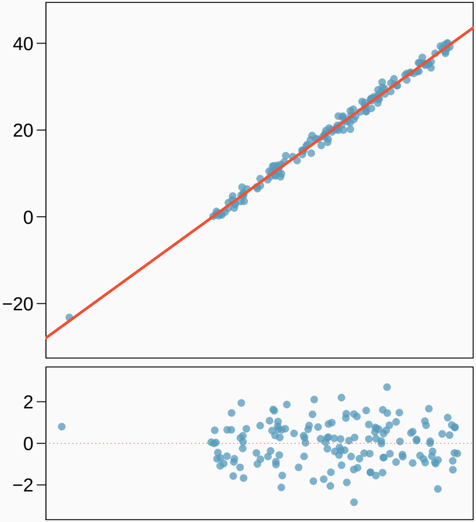
Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.



Types of outliers

Which of the below best describes the outlier?

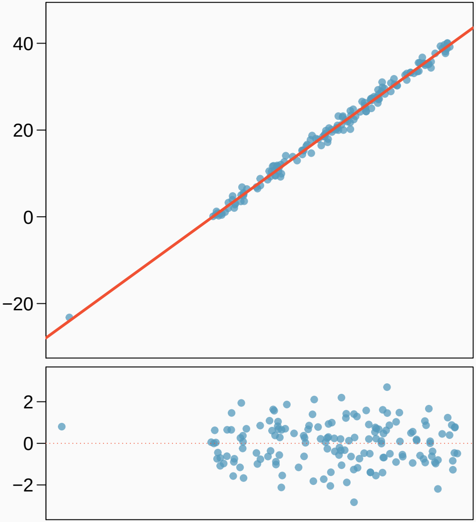
* influential
* high leverage
* none of the above
* there are no outliers



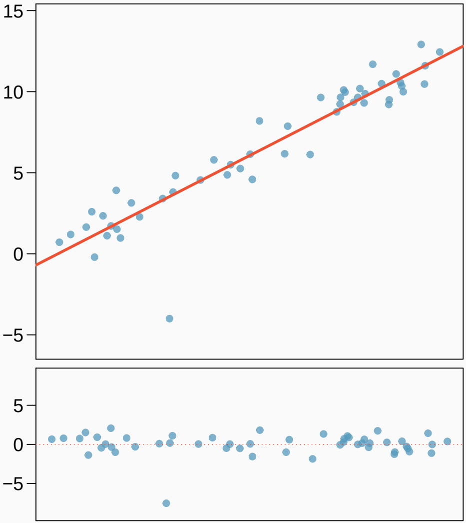
Types of outliers

Which of the below best describes the outlier?

* influential
* high leverage
* none of the above
* there are no outliers

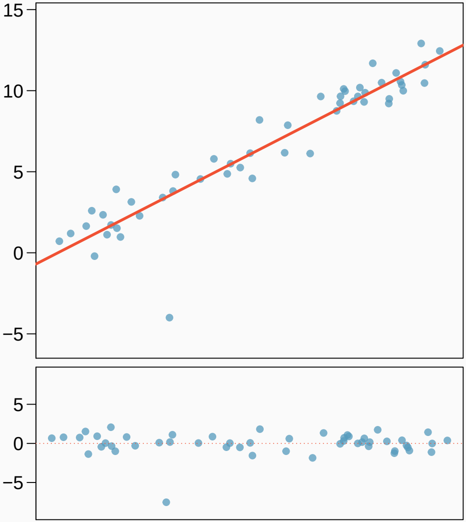


Does this outlier influence the slope of the regression line?



Does this outlier influence the slope of the regression line?

Not much...



Which of following is true?

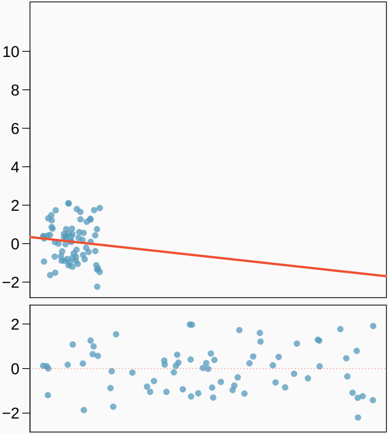
* Influential points always change the intercept of the regression line.
* Influential points always reduce R^2.
* It is much more likely for a low leverage point to be influential, than a high leverage point.
* When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.
* None of the above.

Recap

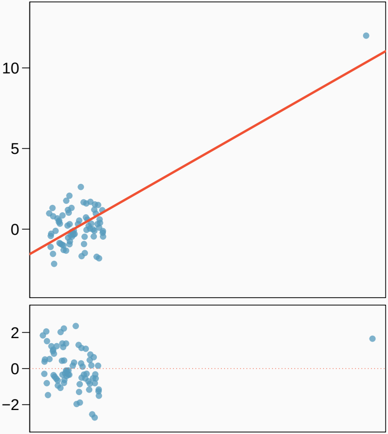
Which of following is true?

* Influential points always change the intercept of the regression line.
* Influential points always reduce R^2.
* It is much more likely for a low leverage point to be influential, than a high leverage point.
* When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.
* None of the above.

$R = 0.08, R^2 = 0.0064$



$R = 0.79, R^2 = 0.6241$



Line fitting, residuals, and correlation

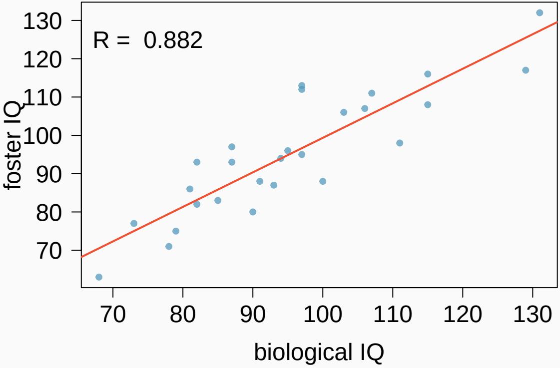
Fitting a line by least squares regression

Types of outliers in linear regression

Inference for linear regression

- Understanding regression output from software
- HT for the slope
- CI for the slope

In 1966 Cyril Burt published a paper called “The genetic determination of differences in intelligence: A study of monozygotic twins reared apart?” The data consist of IQ scores for [an assumed random sample of] 27 identical twins, one raised by foster parents, the other by the biological parents.



Which of the following is false?

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        9.20760    9.29990   0.990    0.332
bioIQ              0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
Multiple R-squared: 0.7779,Adjusted R-squared: 0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

* Additional 10 points in the biological twin's IQ is associated with additional 9 points in the foster twin's IQ, on average.
* Roughly 78% of the foster twins' IQs can be accurately predicted by the model.
* The linear model is $\widehat{\text{fosterIQ}} = 9.2 + 0.9 \times \text{bioIQ}$.
* Foster twins with IQs higher than average IQs tend to have biological twins with higher than average IQs as well.

Which of the following is false?

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.20760    9.29990   0.990    0.332
bioIQ             0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
Multiple R-squared: 0.7779,Adjusted R-squared: 0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

* Additional 10 points in the biological twin's IQ is associated with additional 9 points in the foster twin's IQ, on average.
* Roughly 78% of the foster twins' IQs can be accurately predicted by the model.
* The linear model is $\widehat{\text{fosterIQ}} = 9.2 + 0.9 \times \text{bioIQ}$.
* Foster twins with IQs higher than average IQs tend to have biological twins with higher than average IQs as well.

Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin. What are the appropriate hypotheses?

* $H_0 : b_0 = 0$; $H_A : b_0 \neq 0$
* $H_0 : \beta_0 = 0$; $H_A : \beta_0 \neq 0$
* $H_0 : b_1 = 0$; $H_A : b_1 \neq 0$
* $H_0 : \beta_1 = 0$; $H_A : \beta_1 \neq 0$

Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin. What are the appropriate hypotheses?

* $H_0 : b_0 = 0$; $H_A : b_0 \neq 0$
* $H_0 : \beta_0 = 0$; $H_A : \beta_0 \neq 0$
* $H_0 : b_1 = 0$; $H_A : b_1 \neq 0$
* $H_0 : \beta_1 = 0$; $H_A : \beta_1 \neq 0$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

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

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

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

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

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

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

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

* Point estimate = b_1 is the observed slope.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

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

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

* Point estimate = b_1 is the observed slope.
* SE_{b_1} is the standard error associated with the slope.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

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

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

* Point estimate = b_1 is the observed slope.
* SE_{b_1} is the standard error associated with the slope.
* 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)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

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

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

* Point estimate = b_1 is the observed slope.
* SE_{b_1} is the standard error associated with the slope.
* Degrees of freedom associated with the slope is $df = n - 2$, where n is the sample size.

  Remember: We lose 1 degree of freedom for each parameter we estimate, and in simple linear regression we estimate 2 parameters, β_0 and β_1.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$$T = \frac{0.9014 - 0}{0.0963} = 9.36$$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

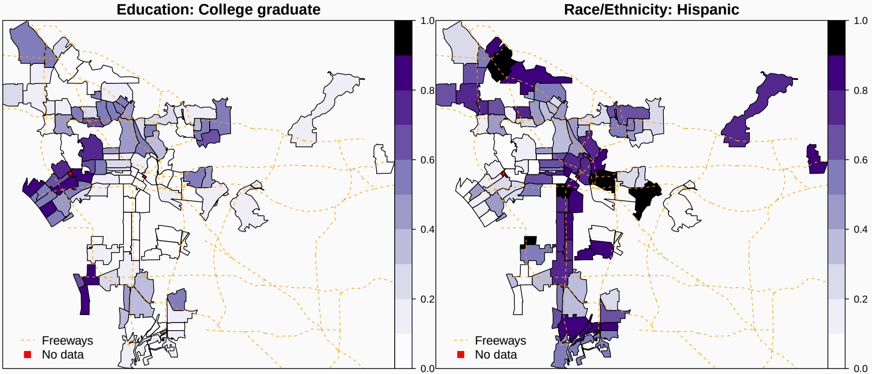
$$
\begin{aligned}
T &= \frac{0.9014 - 0}{0.0963} = 9.36 \\
df &= 27 - 2 = 25
\end{aligned}
$$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

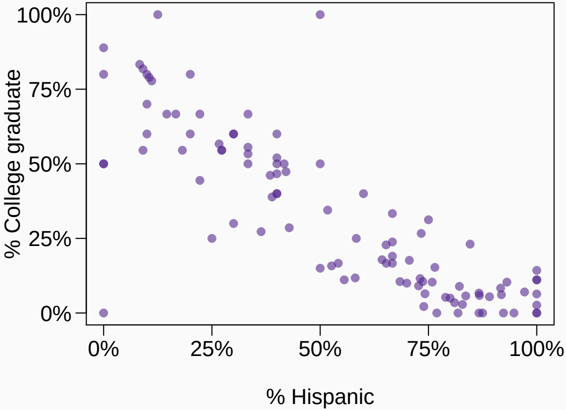
$$
\begin{aligned}
T &= \frac{0.9014 - 0}{0.0963} = 9.36 \\
df &= 27 - 2 = 25 \\
p - value &= P(|T| > 9.36) < 0.01
\end{aligned}
$$

What can you say about the relationship between % 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

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 1% increase in Hispanic residents in a zip code area in LA is associated with a 75% decrease in % of college grads.
- A 1% increase in Hispanic residents in a zip code area in LA is associated with a 0.75% decrease in % of college grads.
- An additional 1% of Hispanic residents decreases the % of college graduates in a zip code area in LA by 0.75%.
- In zip code areas with no Hispanic residents, % of college graduates is expected to be 75%.

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

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 1% increase in Hispanic residents in a zip code area in LA is associated with a 75% decrease in % of college grads.
* A 1% increase in Hispanic residents in a zip code area in LA is associated with a 0.75% decrease in % of college grads.
* An additional 1% of Hispanic residents decreases the % of college graduates in a zip code area in LA by 0.75%.
* 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

How reliable is this p-value if these zip code areas are not randomly selected?

% 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.

How reliable is this p-value if these zip code areas are not randomly selected?

% 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.

How reliable is this p-value if these zip code areas are not randomly selected?

Not very...

Confidence interval for the slope

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 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

* $9.2076 \pm 1.65 \times 9.2999$
* $0.9014 \pm 2.06 \times 0.0963$
* $0.9014 \pm 1.96 \times 0.0963$
* $9.2076 \pm 1.96 \times 0.0963$

Remember that a confidence interval is calculated as point estimate ± 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 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$n = 27$ $df = 27 - 2 = 25$

* $9.2076 \pm 1.65 \times 9.2999$
* $0.9014 \pm 2.06 \times 0.0963$
* $0.9014 \pm 1.96 \times 0.0963$
* $9.2076 \pm 1.96 \times 0.0963$

Confidence interval for the slope

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 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

* $9.2076 \pm 1.65 \times 9.2999$
* $0.9014 \pm 2.06 \times 0.0963$
* $0.9014 \pm 1.96 \times 0.0963$
* $9.2076 \pm 1.96 \times 0.0963$

$$n = 27 \qquad df = 27 - 2 = 25$$

$$95\% : t^{\star}_{25} = 2.06$$

Remember that a confidence interval is calculated as point estimate ± 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 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

* $9.2076 \pm 1.65 \times 9.2999$
* $0.9014 \pm 2.06 \times 0.0963$
* $0.9014 \pm 1.96 \times 0.0963$
* $9.2076 \pm 1.96 \times 0.0963$

$$n = 27 \qquad df = 27 - 2 = 25$$

$$95\% : \; t^\star_{25} = 2.06$$

$$0.9014 \pm 2.06 \times 0.0963$$

Confidence interval for the slope

Remember that a confidence interval is calculated as point estimate ± 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 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

* $9.2076 \pm 1.65 \times 9.2999$
* $0.9014 \pm 2.06 \times 0.0963$
* $0.9014 \pm 1.96 \times 0.0963$
* $9.2076 \pm 1.96 \times 0.0963$

$$n = 27 \qquad df = 27 - 2 = 25$$

$$95\% : t^\star_{25} = 2.06$$

$$0.9014 \pm 2.06 \times 0.0963$$

$$(0.7 \; , \; 1.1)$$

* Inference for the slope for a single-predictor linear regression model:

Recap

* Inference for the slope for a single-predictor linear regression model:
  * Hypothesis test:

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

Recap

* Inference for the slope for a single-predictor linear regression model:
  * Hypothesis test:

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

  * Confidence interval:

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

Recap

* Inference for the slope for a single-predictor linear regression model:
  * Hypothesis test:

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

  * 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.

Recap

* Inference for the slope for a single-predictor linear regression model:
  * Hypothesis test:
  $$T = \frac{b_1 - \text{null value}}{SE_{b_1}} \qquad df = n - 2$$
  * 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.

Recap

* Inference for the slope for a single-predictor linear regression model:
  * Hypothesis test:
  $$T = \frac{b_1 - \text{null value}}{SE_{b_1}} \qquad df = n - 2$$
  * 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.

Caution

* Always be aware of the type of data you're working with: random sample, non-random sample, or population.

Caution

* Always be aware of the type of data you're working with: random sample, non-random sample, or population.
* Statistical inference, and the resulting p-values, are meaningless when you already have population data.

Caution

* Always be aware of the type of data you're working with: random sample, non-random sample, or population.
* Statistical inference, and the resulting p-values, are meaningless when you already have population data.
* If you have a sample that is non-random (biased), inference on the results will be unreliable.

Caution

* Always be aware of the type of data you're working with: random sample, non-random sample, or population.
* Statistical inference, and the resulting p-values, are meaningless when you already have population data.
* If you have a sample that is non-random (biased), inference on the results will be unreliable.
* The ultimate goal is to have independent observations.