Section 2.6

Two Quantitative Variables: Linear Regression

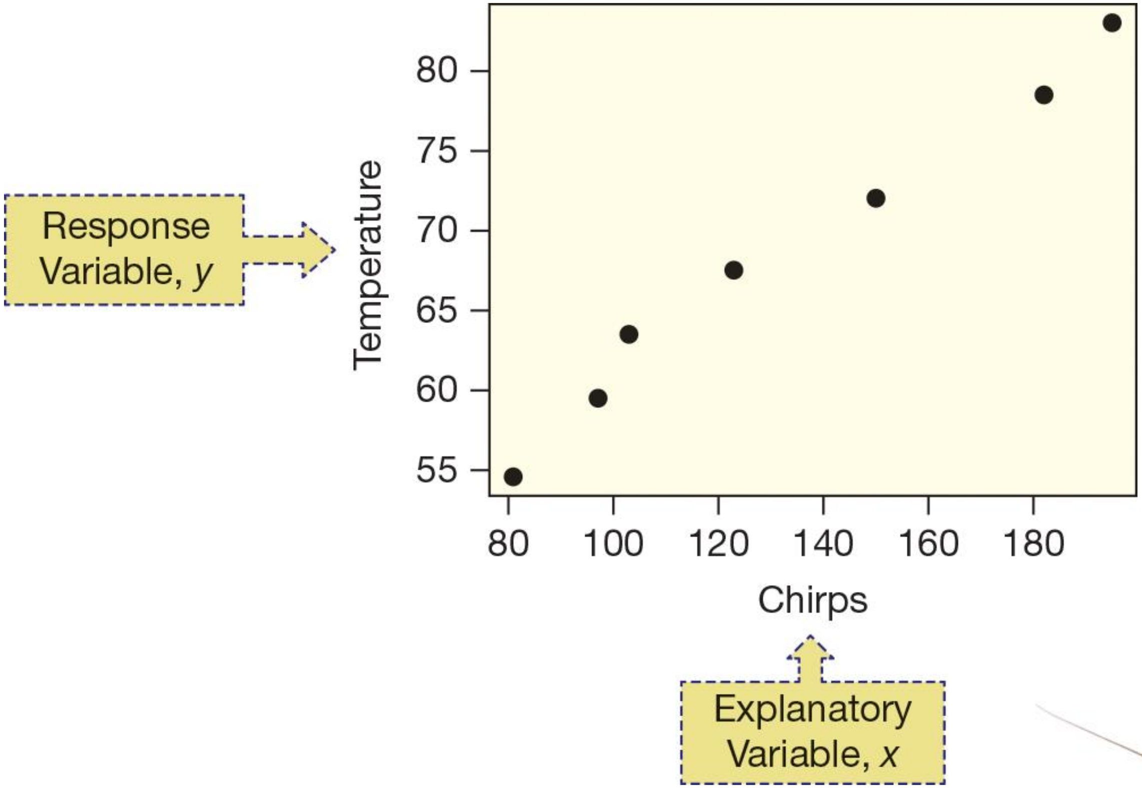
Outline

- Regression line
- Predicted values
- Residuals
- Interpreting slope and intercept
- Cautions

Crickets and Temperature (Question)

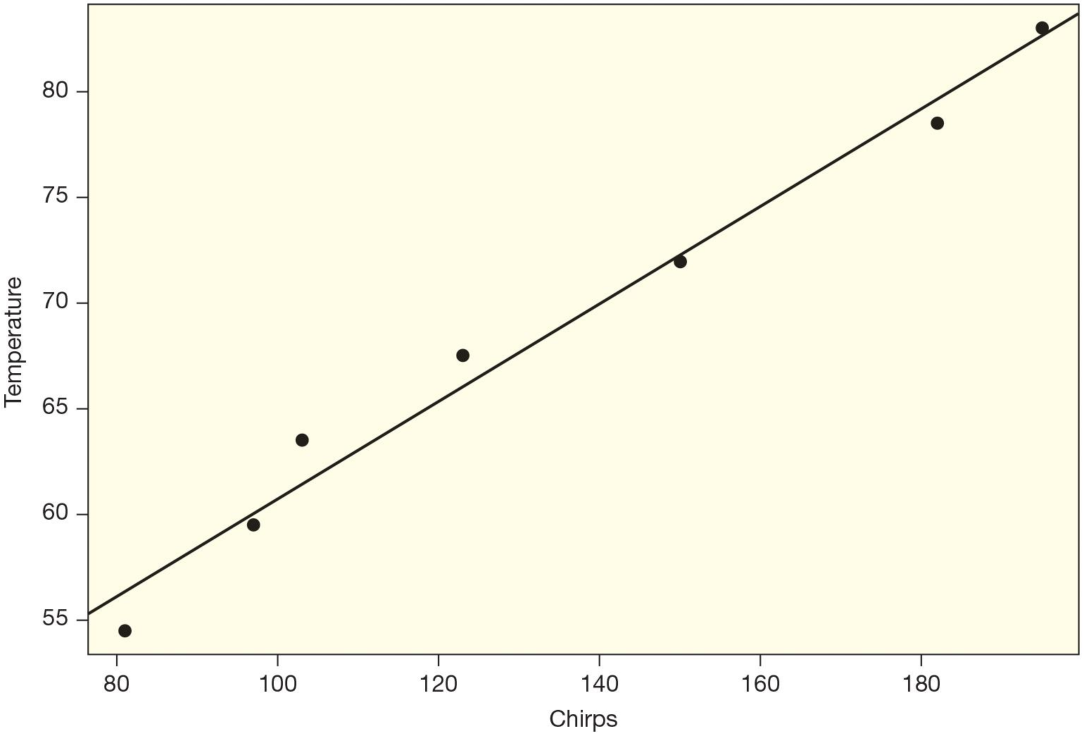
- Can you estimate the temperature on a summer evening, just by listening to crickets chirp?

Crickets and Temperature (Scatterplot)



Regression Line – What Is It?

Goal: Find a straight line that best fits the data in a scatterplot



Equation of the Line (Formula)

The estimated regression line is

$$\boxed{\hat{y} = a + bx}$$

Equation of the Line (Variables)

The estimated regression line is

$$\boxed{\hat{y} = a + bx}$$

Predicted Response

Explanatory

Prediction (Chirps and Temp)

- Type equation here.The regression equation can be used to predict *y* for a given value of *x*

$$\widehat{Temp} = 37.7 + 0.23 chirps$$

- If you listen and hear crickets chirping about 140 times per minute, your best guess at the outside temperature is

$$37.7 + 0.23 \cdot 140 = 69.9°\text{F}$$

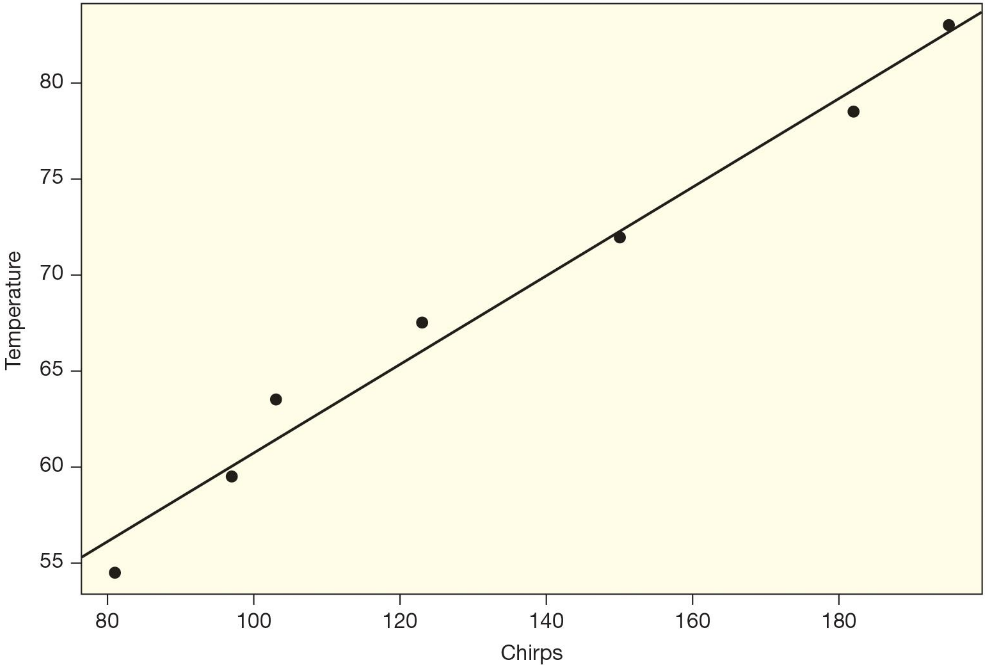
Prediction (The Temperature)

- What is the predicted temperature when the crickets do 103 chirps per minute?

$$37.7 + 0.23(103) = 61.39$$

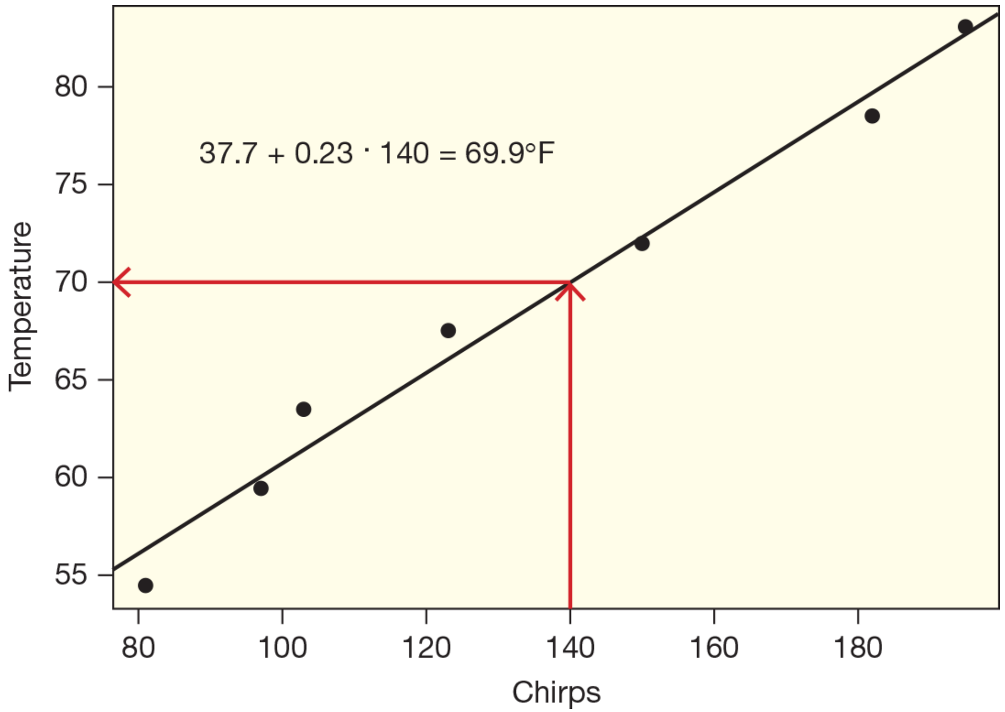
Prediction
(Where is the predicted response?)

- The predicted response is on the regression line directly above the *x* value

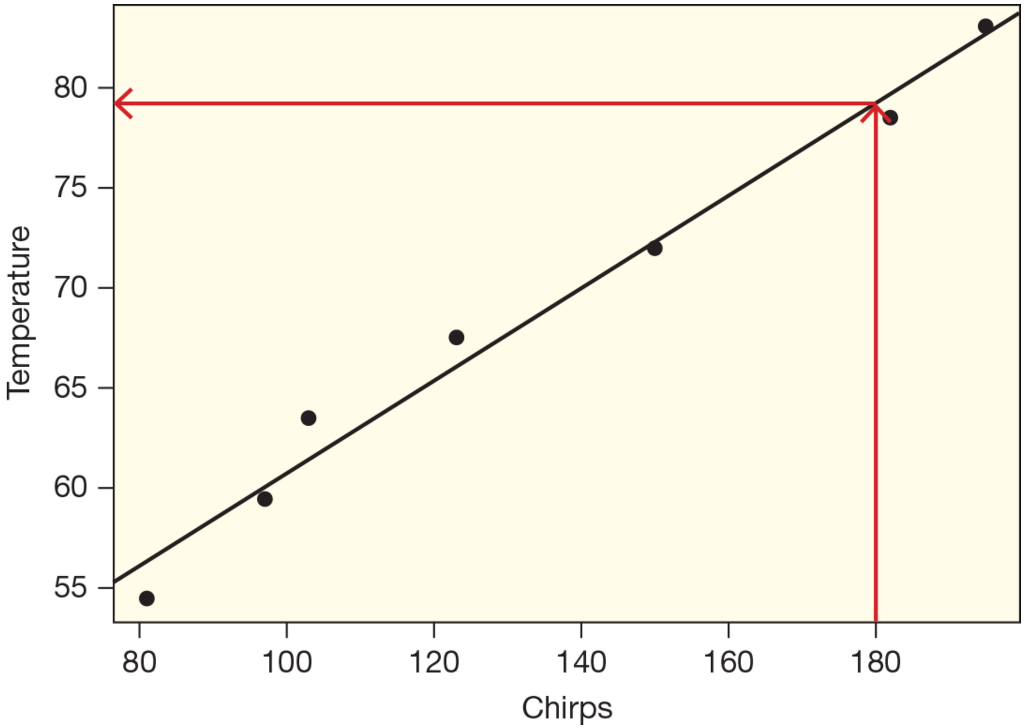


Prediction (For 140 Chirps)

- The predicted response is on the regression line directly above the *x* value



Prediction (For 180 Chirps)



If the crickets are chirping about 180 times per minute, your best guess at the temperature is

a) 60°

b) 70°

c) 80°

Prediction (Predicted vs Observed)

- One of the cases in the cricket dataset is 103 chirps per minute and 63.5°F
- How far is the predicted temperature from the observed temperature for this case?

$$37.7 + 0.23(103) = 61.39$$

$$63.5 - 61.39 = 2.11$$

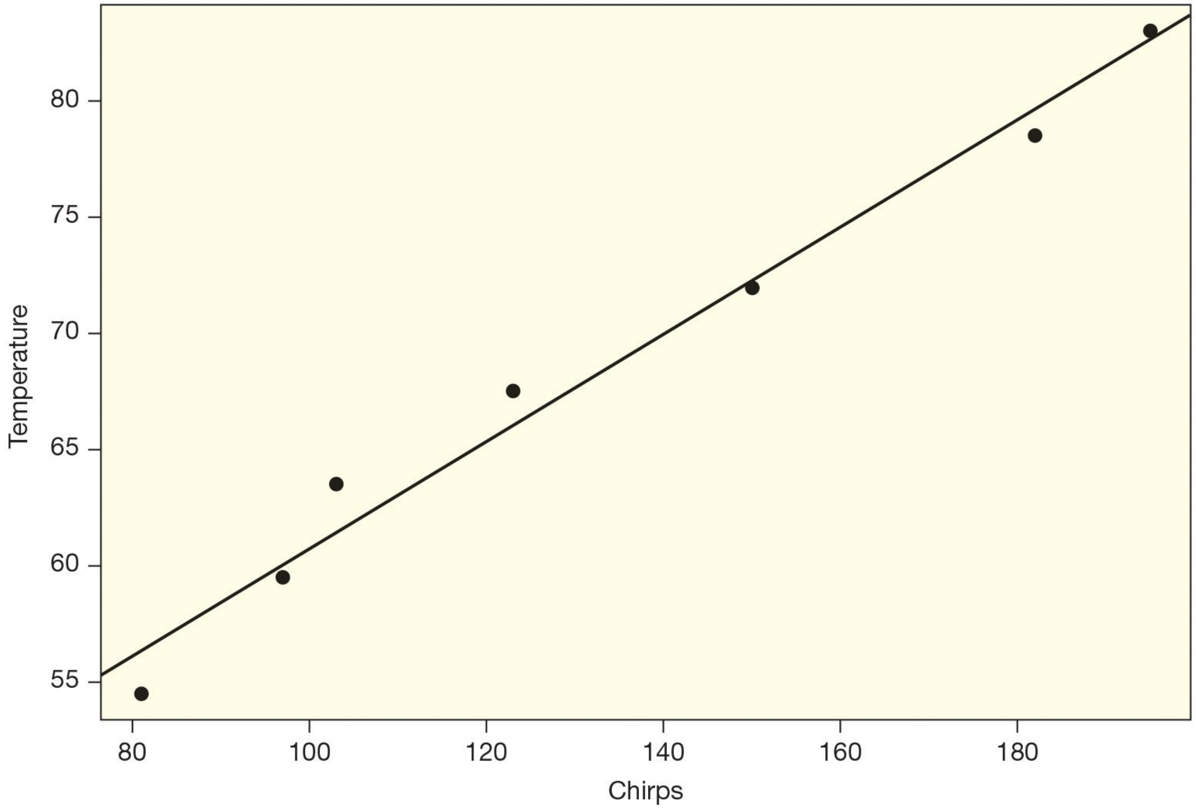
Regression Line

- How do we find the best fitting line???

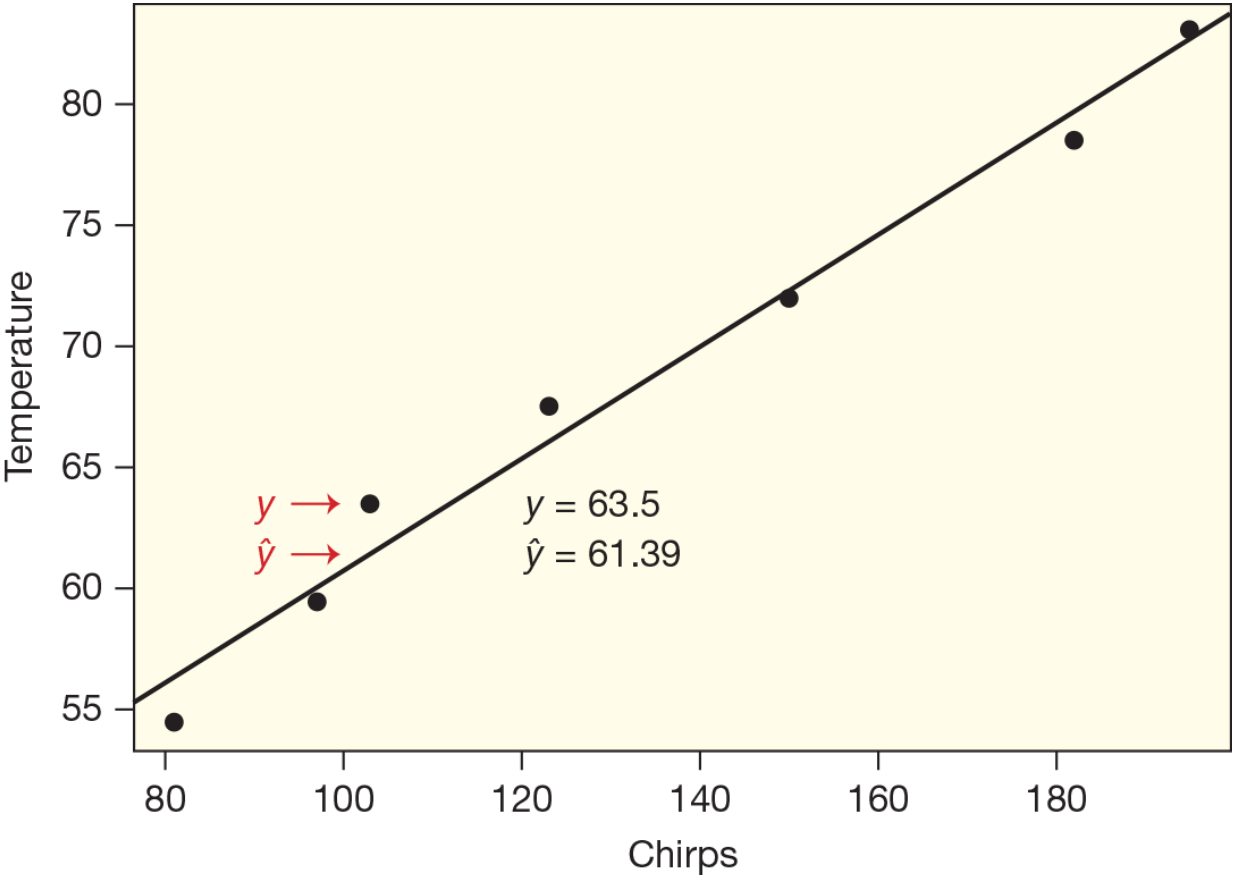
Predicted and Actual Values (Definition)

- The ***observed response value***, y, is the response value observed for a particular data point
- The ***predicted response value***, $\hat{y}$, is the response value that would be predicted for a given x value, based on a model
- The best fitting line is that which makes the predicted values closest to the actual values

Predicted and Actual Values (Chirps and Temp)



Predicted and Actual Values (A Visual)



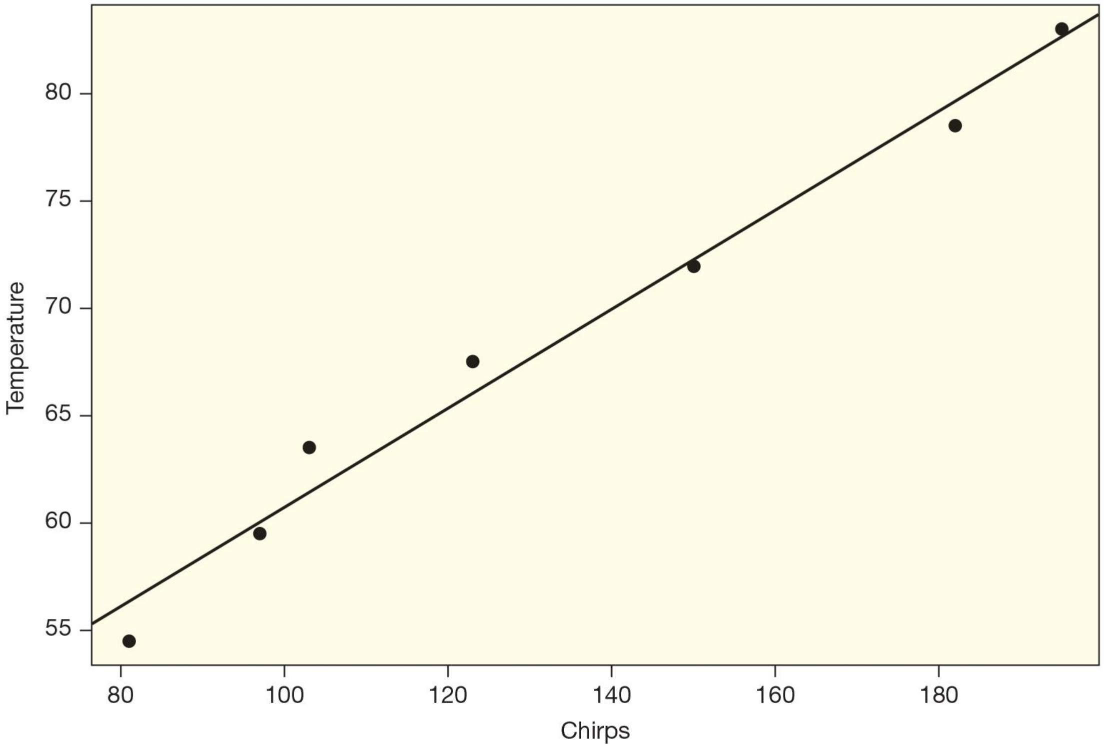
Residual (Definition)

The ***residual*** for each data point is

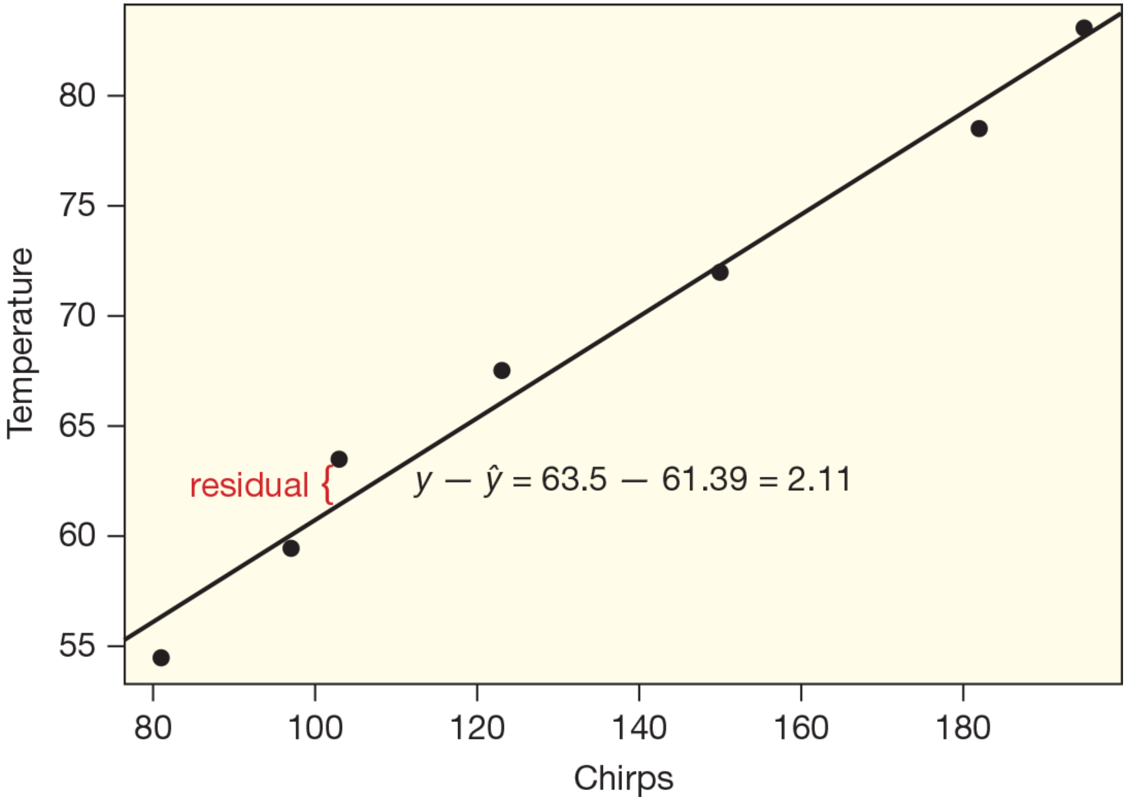
$$observed - predicted = y - \hat{y}$$

- The residual is also the vertical distance from each point to the line

Residual (Chirps and Temp)



Residual (Measurement)



- Want to make all the residuals as small as possible.
- How would you measure this?

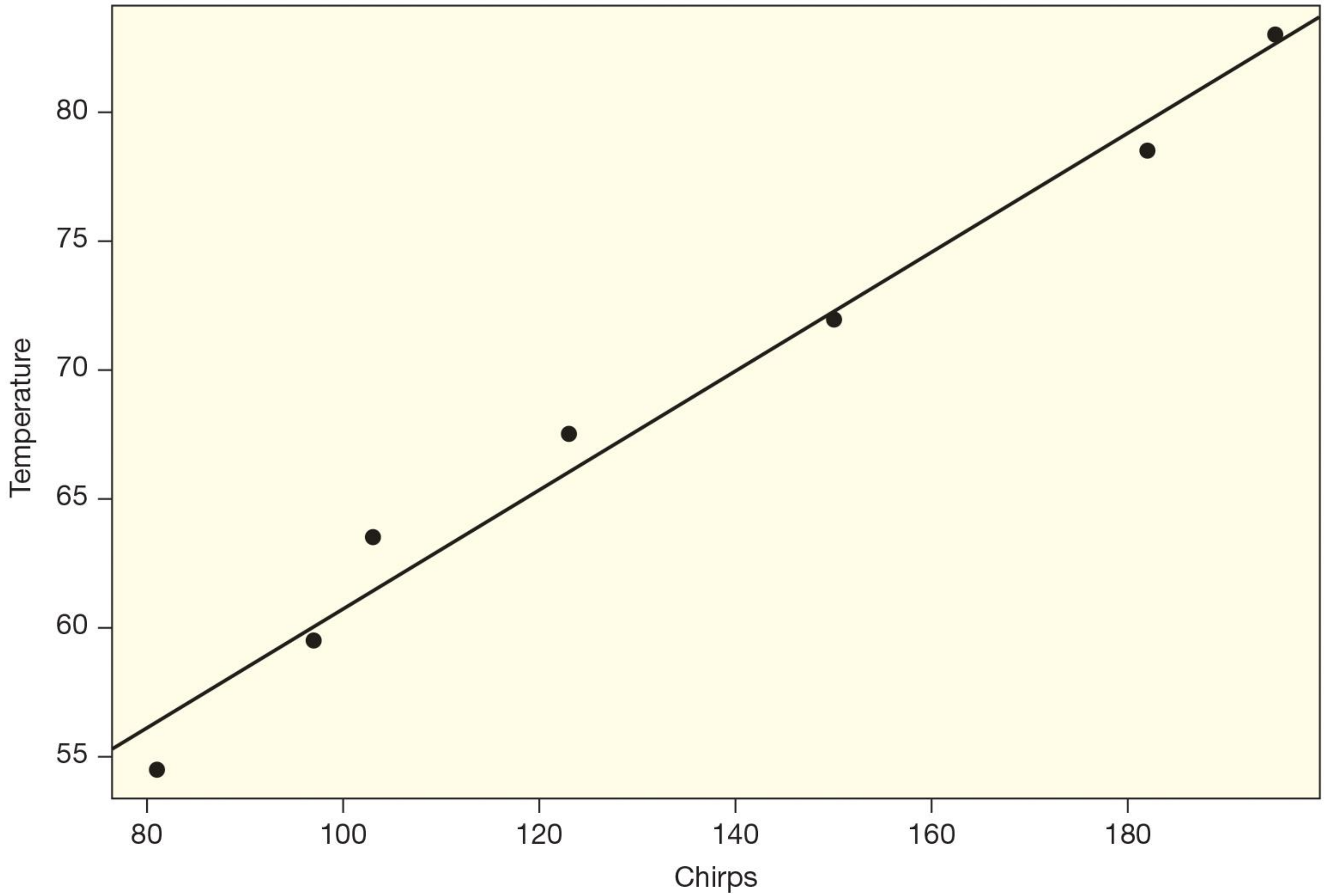
Least Squares Line

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

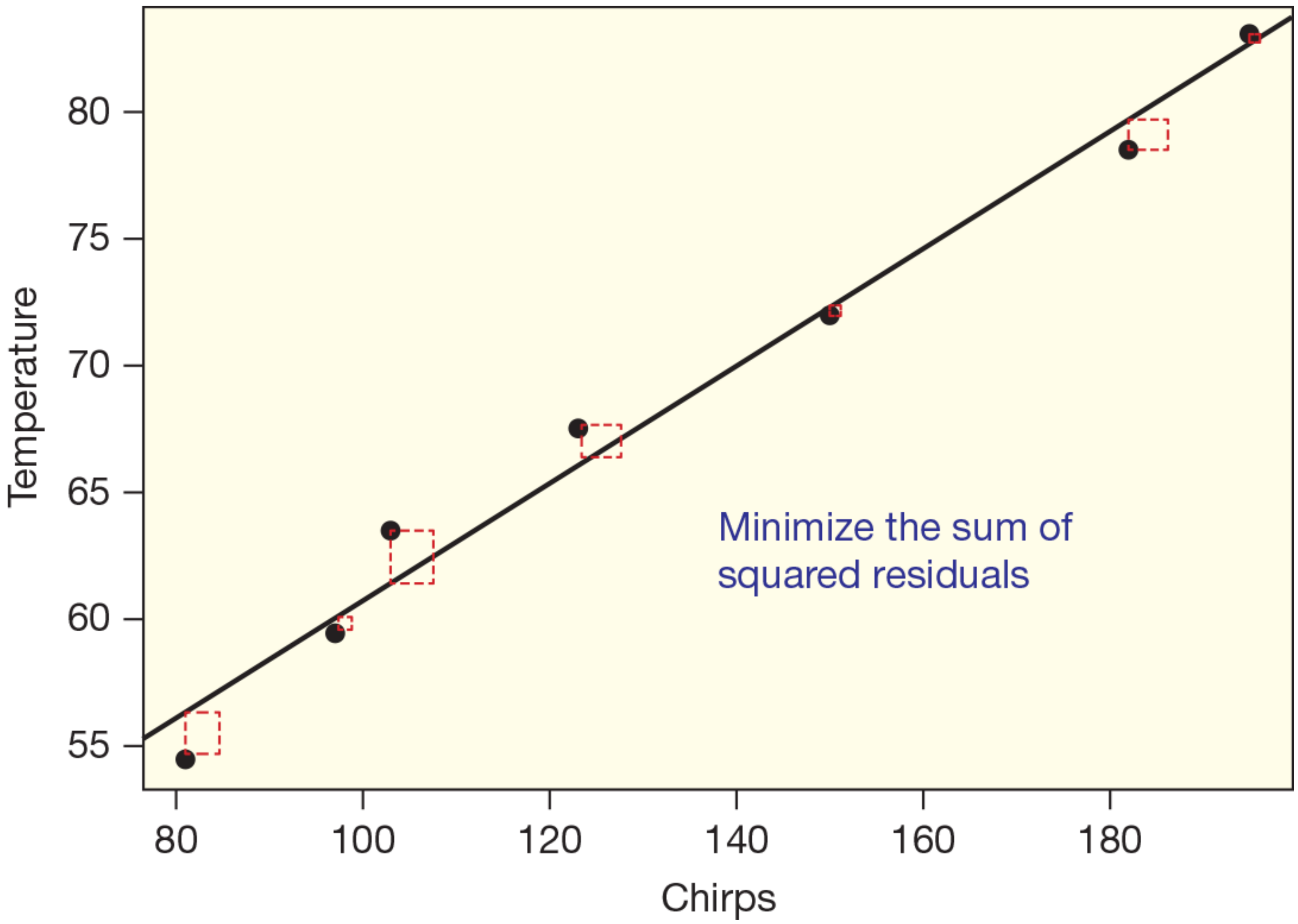
$$\text{minimize } \sum (y - \hat{y})^2$$

- Rely on technology for this finding the least square line.
- “least squares line” = “regression line”

Least Squares Regression (Chirps and Temp)



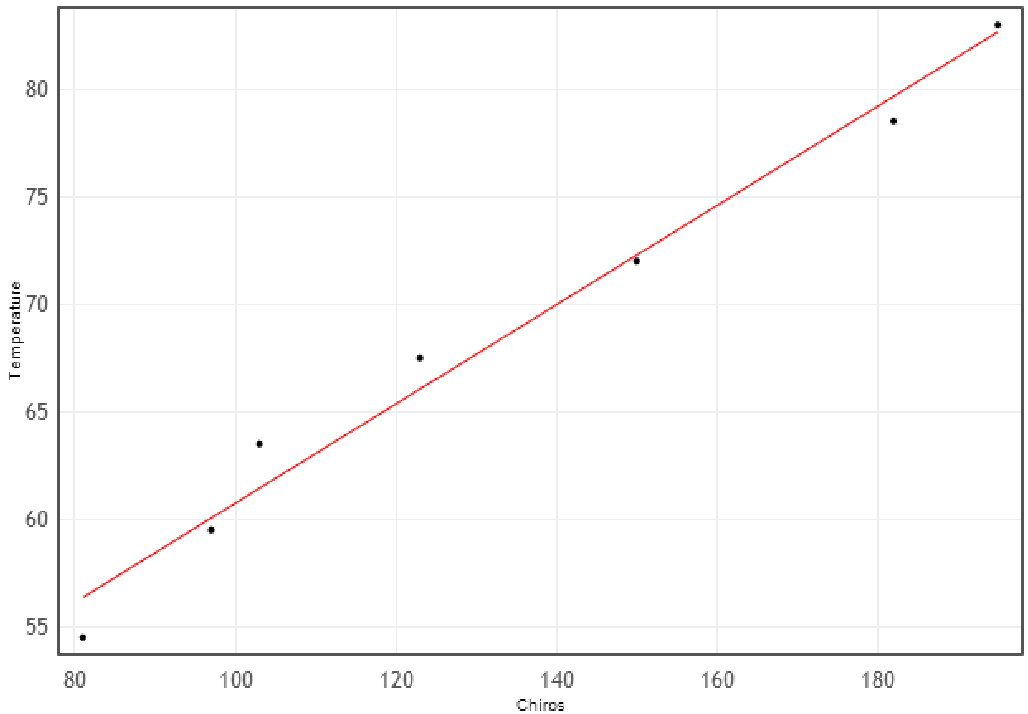
Least Squares Regression (Residuals)



Regression in StatKey

StatKey Descriptive Statistics for Two Quantitative Variables

Cricket Chirps (Temperature vs Chirps) ▾ | Show Data Table | Edit Data



Summary Statistics

Switch Variables

Statistic	Chirps	Temperature
Mean	133.0	68.4
Standard Deviation	43.8	10.2
Sample Size	7	
Correlation	0.991	
Slope	0.231	
Intercept	37.679	

Scatterplot Controls

☑ Show Regression Line

$$\widehat{Temp} = 37.7 + 0.23chirps$$

Technology Examples

```
The regression equation is
Temperature = 37.7 + 0.231 Chirps

Predictor      Coef   SE Coef      T      P
Constant     37.679     1.978  19.05  0.000
Chirps      0.23067   0.01423  16.21  0.000

S = 1.52778   R-Sq = 98.1%
```

Minitab

Model of CricketChirps Simple Regression

Response attribute (numeric): Temperature

Predictor attribute (numeric): Chirps

```
Sample count:          7
Equation:              Temperature = 0.230666 Chirps + 37.679
r:                     0.990625
r-squared:             0.98134
Slope:                 0.230666 +/- 0.0365682
SE Slope:              0.0142257
Confidence level:      95 %
```

When **Chirps** = 0 , the predicted value for a future observation of **Temperature** is **37.6786 +/- 6.42506**

Fathom

R

```
> lm(Temperature~Chirps)
Coefficients:
(Intercept)        Chirps
37.6786            0.2307
```

$$\widehat{Temp} = 37.7 + 0.23 * Chirps$$

Explanatory and Response

- Unlike correlation, for linear regression it does matter which is the explanatory variable and which is the response

$$\widehat{Temp} = 37.7 + 0.23chirps$$

$$\widehat{Chirps} = -157.8 + 4.25Temp$$

Slope and Intercept (Regression Line)

The estimated regression line is

$$\boxed{\hat{y} = a + bx}$$

Slope and Intercept (and predicted *y*)

The estimated regression line is

$$\boxed{\hat{y} = \underbrace{a}_{\text{Intercept}} + \underbrace{b}_{\text{Slope}}x}$$

- ***Slope***: increase in predicted y for every unit increase in x
- ***Intercept***: predicted y value when $x = 0$

Interpreting Slope and Intercept

$$\widehat{Temp} = 37.7 + 0.23chirps$$

Slope = 0.23:

The predicted temperature goes up by about 0.23^{o}F for every increase of one in the chirp rate.

Intercept = 37.7:

Predicted temperature when crickets stop chirping???

Predicted Grade

For a certain course, the regression line to predict grade G on the final based on number of hours studying H is

$$\hat{G} = 50 + 3 \cdot H$$

One person studied 10 hours and received a 88 on the final. The predicted grade for this person is:

A. 8

B. 10

C. 50

D. 80

E. 88

Residual Grade

For a certain course, the regression line to predict grade G on the final based on number of hours studying H is

$$\hat{G} = 50 + 3 \cdot H$$

One person studied 10 hours and received an 88 on the final. The residual for this person is:

A. 8
B. 10
C. 50
D. 80
E. 88

Slope and Predicted Grade

For a certain course, the regression line to predict grade G on the final based on number of hours studying H is

$$\hat{G} = 50 + 3 \cdot H$$

We can interpret the slope in context to mean that:

A. Predicted grade will go up by 1 point for a person who studies 3 more hours.

B. Predicted grade will go up by 3 points for a person who studies 1 more hour.

C. Three more hours of studying gives 3 more points on the final

D. The rise over the run is 3/1.

E. The response variable goes up by 1 if the explanatory variable goes up by 3.

Intercept and Predicted Grade

For a certain course, the regression line to predict grade G on the final based on number of hours studying H is

$$\hat{G} = 50 + 3 \cdot H$$

We can interpret the slope in context to mean that:

A. If a person does not study at all, the predicted grade will be 50.
B. A predicted grade of zero goes with studying 50 hours.
C. The more someone studies, the higher the predicted grade.
D. The line crosses the axis at 50.
E. The response variable is 50 if the explanatory variable is 0.

Regression Caution 1

- Do not use the regression equation or line to predict outside the range of x values available in your data (do not extrapolate!)
- If none of the x values are anywhere near 0, then the intercept is meaningless!

Units

- It is helpful to think about units when interpreting a regression equation

$$\underbrace{\hat{y}}_{y\text{ units}} = \underbrace{a}_{y\text{ units}} + \underbrace{b}_{\frac{y\text{ units}}{x\text{ units}}} \cdot \underbrace{x}_{x\text{ units}}$$

$$\underbrace{\widehat{Temp}}_{\text{degrees}} = \underbrace{37.7}_{\text{degrees}} + \underbrace{0.23}_{\text{degrees/ chirps per min}} \underbrace{Chirps}_{\text{chirps per minute}}$$

Regression Model

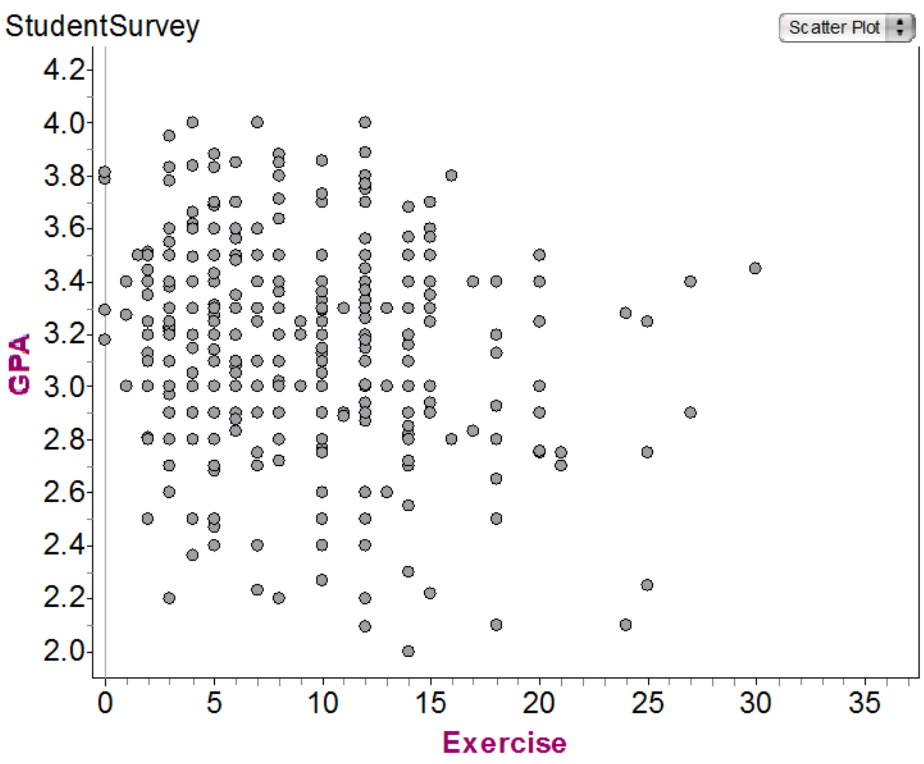
$$\widehat{Wgt} = -260.4 + 6.02 \cdot Hgt$$

Which is a correct interpretation?

a) The average subject is just over 6 feet tall.
b) For every extra 6.02 inches in height, the predicted weight goes up by one pound.
c) Predicted weight increases by 6.02 pounds for every additional inch in height.
d) A zero inch tall person is predicted to weigh about −260.4 pounds.

Exercise and GPA (Associated?)

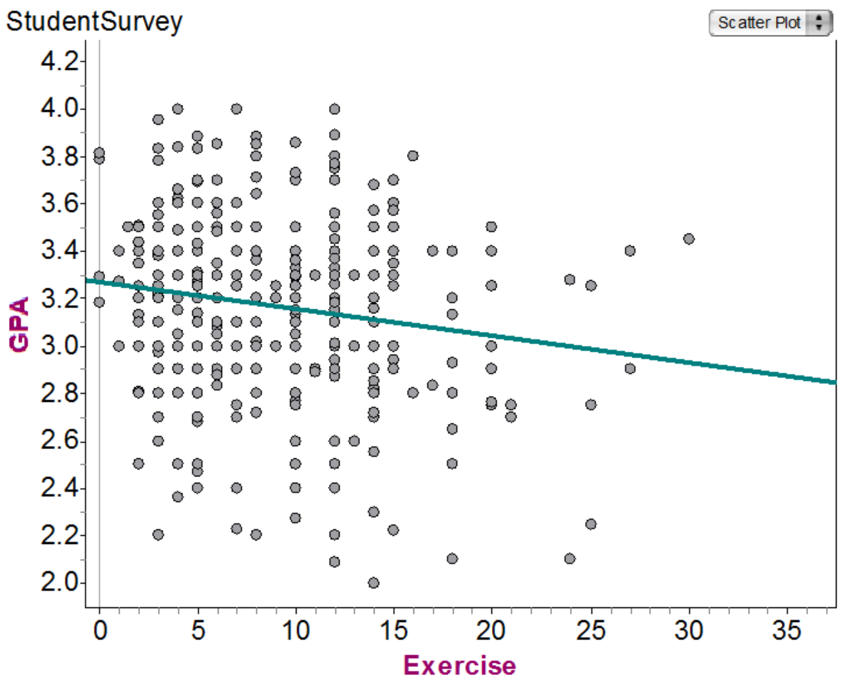
Are the hours of exercise per week and grade point average of students



a) positively associated
b) negatively associated
c) not associated
d) other

Exercise and GPA (Regression Line)

Are the hours of exercise per week and grade point average of students



$$\widehat{GPA} = 3.26 - 0.0114 \cdot Exercise$$

a) positively associated

b) negatively associated

c) not associated

d) other

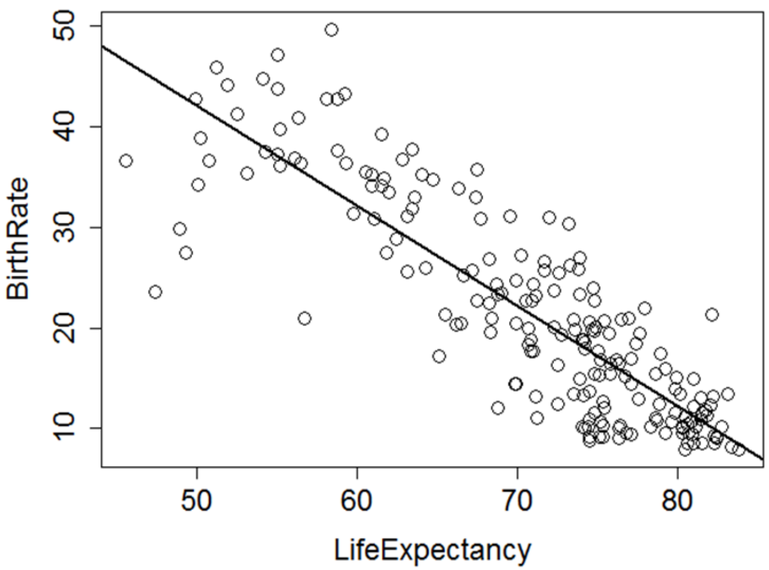
Regression Caution 2

- Computers will calculate a regression line for any two quantitative variables, even if they are not associated or if the association is not linear
- ALWAYS PLOT YOUR DATA!
- The regression line/equation should only be used if the association is approximately linear

Regression Caution 3

- Outliers (especially outliers in both variables) can be very influential on the regression line
- ALWAYS PLOT YOUR DATA!

Life Expectancy and Birth Rate




```
Coefficients:
(Intercept)   LifeExpectancy
91.8703        -0.9953
```

Which of the following interpretations is correct?

a) A decrease of 0.9953 in the birth rate corresponds to a 1 year increase in predicted life expectancy

b) Increasing life expectancy by 1 year will cause birth rate to decrease by 0.9953

c) Both

d) Neither

a) The model is predicting birth rate based on life expectancy, not the other way around

b) Can only make conclusions about causality from a randomized experiment.

Regression Caution 4

- Higher values of x may lead to higher (or lower) predicted values of y, but this does **NOT** mean that changing x will cause y to increase or decrease

- Causation can only be determined if the values of the explanatory variable were determined randomly (which is rarely the case for a continuous explanatory variable)

$r = 0$

Challenge: If the correlation between *x* and *y* is 0, what would the regression line be?

Summary: Least Squares Regression (Explanations)

- For a quantitative response *y* and quantitative predictor *x*, the least squares line is

$$\hat{y} = a + bx$$

  where the slope (*b*) and intercept (*a*) are chosen to minimize the sum of squared residuals.
- For each data case the residual is $y - \hat{y}$.
- We rely on technology to give the prediction equation (also known as the least squares fit or regression line).
- The slope is interpreted as the change in the predicted response (y) when the explanatory variable (x) increases by one.

Summary: Least Squares Regression (Cautions)

- Cautions:
  - Don’t extrapolate far beyond where the model is built.
  - Estimating a least squares line does *not* mean there is a linear trend in the data.
  - Watch out for outliers that don’t fit the pattern or can greatly influence the line.
  - Even a strong linear fit does not (necessarily) imply a cause/effect relationship.
  - ALWAYS PLOT YOUR DATA!