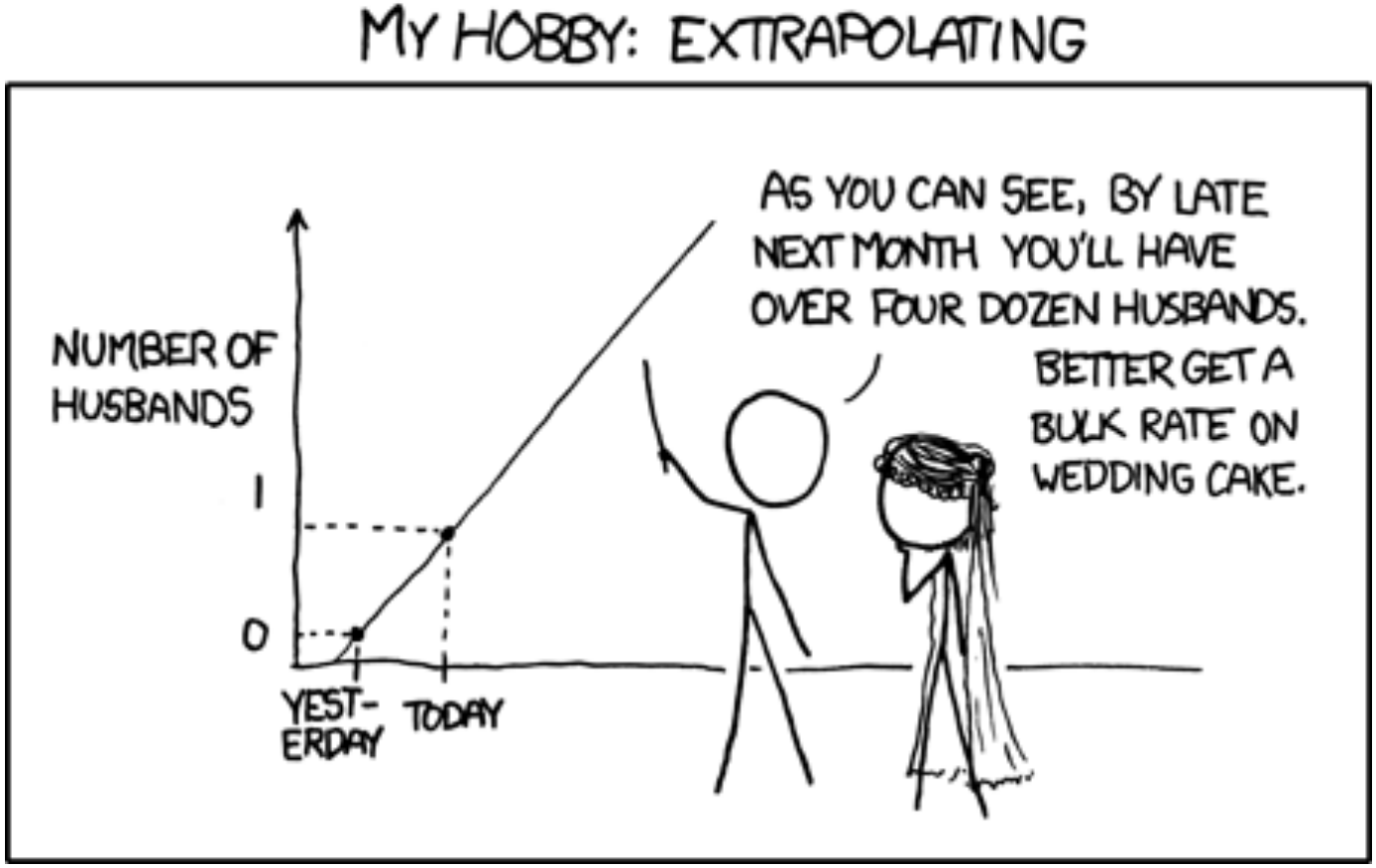
3.2: Least Squares Regressions



Section 3.2
Least-Squares Regression

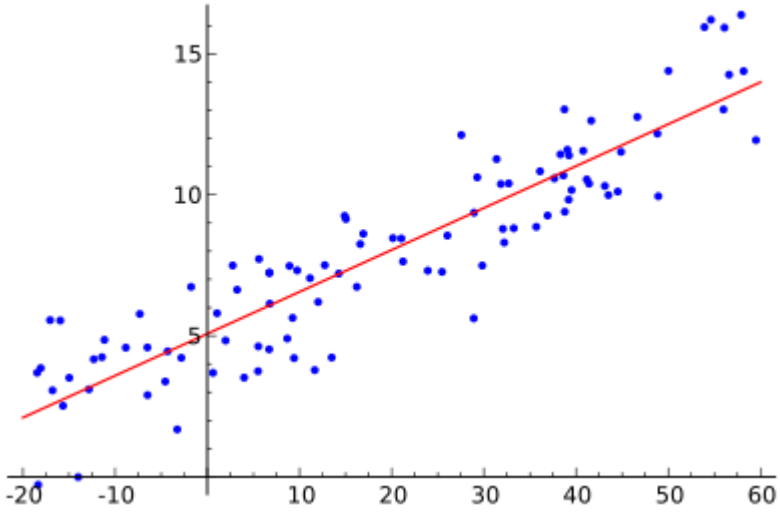
After this section, you should be able to…

- ✓ INTERPRET a regression line
- ✓ CALCULATE the equation of the least-squares regression line
- ✓ CALCULATE residuals
- ✓ CONSTRUCT and INTERPRET residual plots
- ✓ DETERMINE how well a line fits observed data
- ✓ INTERPRET computer regression output

Regression Lines

A **regression line** summarizes the relationship between two variables, but only in settings where one of the variables helps **explain** or **predict** the other.

A **regression line** is a line that describes how a response variable *y* changes as an explanatory variable *x* changes.
We often use a regression line to predict the value of *y* for a given value of *x*.



Regression Lines

Regression lines are used to conduct analysis.

- Colleges use student's SAT and GPAs to predict college success
- Professional sports teams use player's vital stats (40 yard dash, height, weight) to predict success
- Macy's uses shipping, sales and inventory data predict future sales.
- MDCPS uses student data to evaluate teachers using the VAM model

Regression Line Equation

Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis).
A regression line relating y to x has an equation of the form:

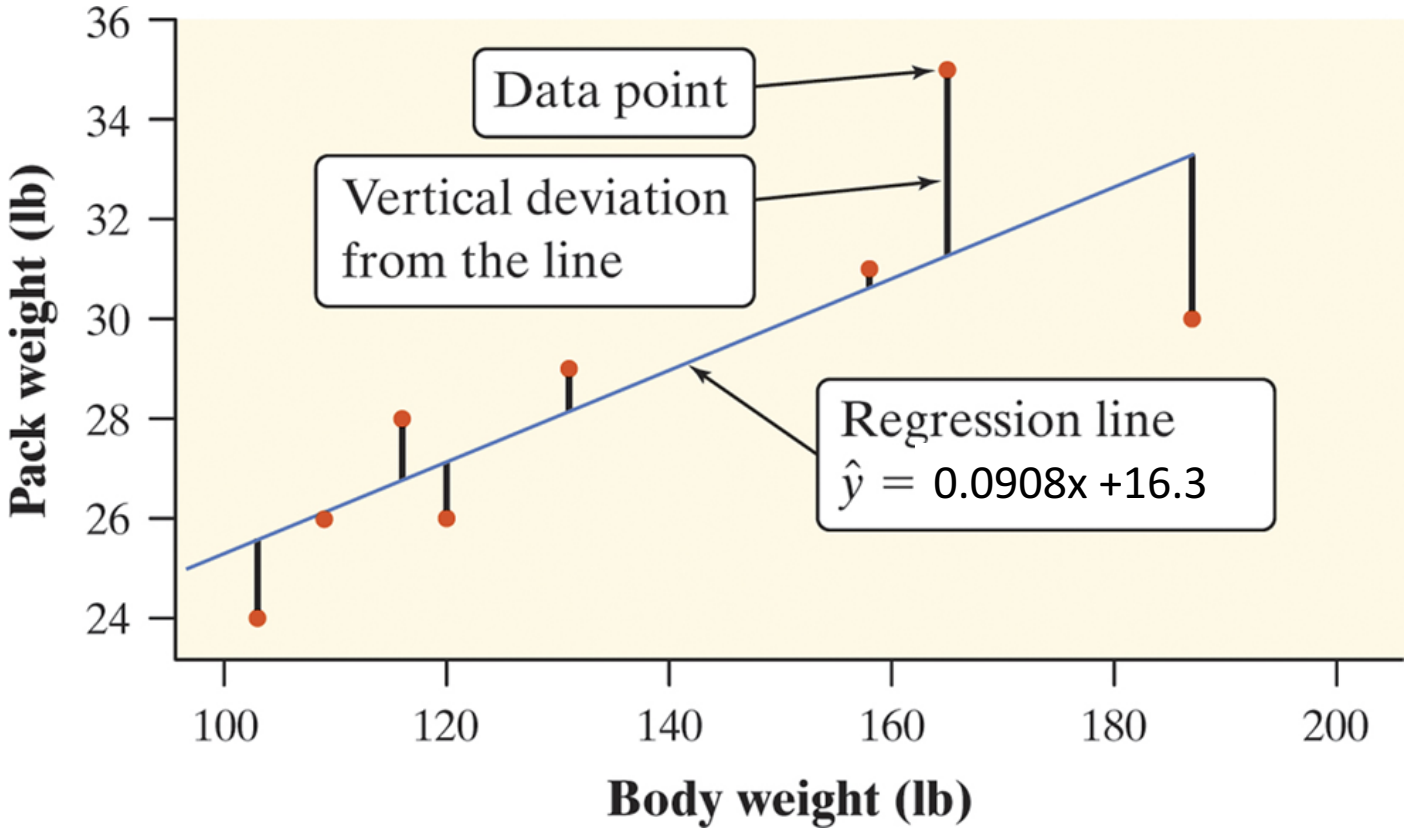
$$\hat{y} = ax + b$$

In this equation,

- $\hat{y}$ (read "y hat") is the predicted value of the response variable y for a given value of the explanatory variable x.
- a is the slope, the amount by which y is predicted to change when x increases by one unit.
- b is the y intercept, the predicted value of y when x = 0.

Regression Line Equation

Body weight (lb):	120	187	109	103	131	165	158	116
Backpack weight (lb):	26	30	26	24	29	35	31	28



Format of Regression Lines

Format 1:

$\hat{y}$ = 0.0908x + 16.3

$\hat{y}$ = predicted back pack weight

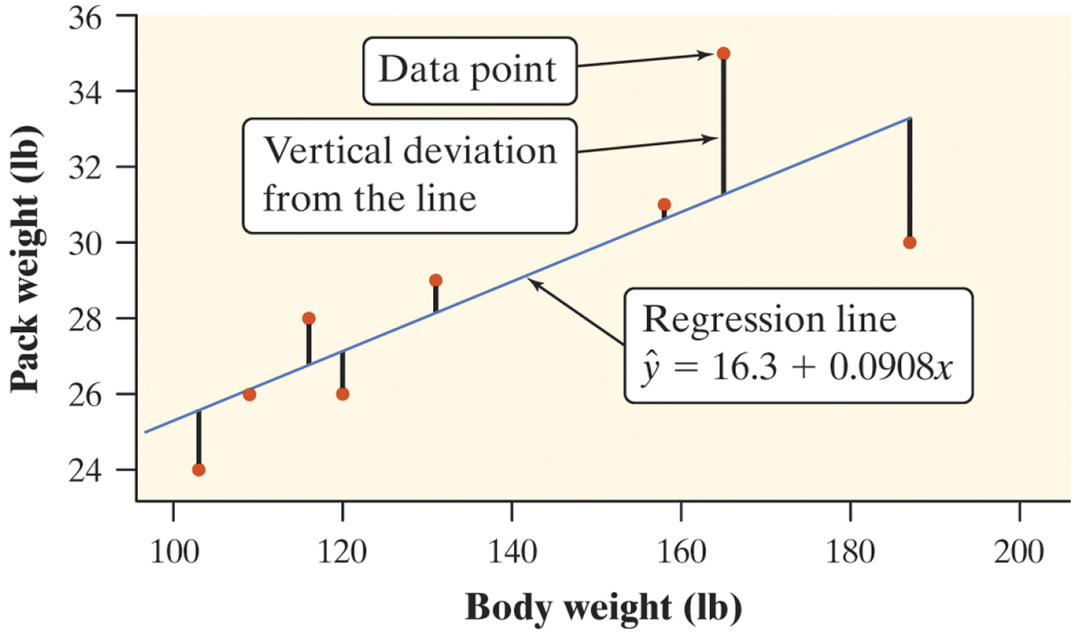
x = student's weight

Format 2:

Predicted back pack weight= 16.3 + 0.0908(student's weight)

Interpreting Linear Regression

- **Y-intercept:** A student weighing zero pounds is predicted to have a backpack weight of 16.3 pounds (no practical interpretation).
- **Slope:** For each additional pound that the student weighs, it is predicted that their backpack will weigh an additional 0.0908 pounds more, on average.



Interpreting Linear Regression

Interpret the y-intercept and slope values in context. Is there any practical interpretation?

$$\hat{y} = 37x + 270$$

x= Hours Studied for the SAT

$\hat{y}$ =Predicted SAT Math Score

Interpreting Linear Regression

$$\hat{y} = 37x + 270$$

Slope: For each additional hour the student studies, his/her score is predicted to increase 37 points, on average.

This makes sense

OR this does not make sense; it is unreasonable for scores to increase by 37 points for JUST one hour of studying.

Interpreting Linear Regression

$$\hat{y} = 37x + 270$$

Y-intercept: If a student studies for zero hours, then the student's predicted SAT score is 270 points.

This makes sense

OR This does not make sense because a SAT scores of 270 is very low regardless of study.

Predicted Value

What is the predicted SAT Math score for a student who studies 12 hours?

$$\hat{y} = 37x + 270$$

Hours Studied for the SAT (x)

Predicted SAT Math Score (y)

Predicted Value

What is the predicted SAT Math score for a student who studies 12 hours?

$$\hat{y} = 37x + 270$$

Hours Studied for the SAT (x)

Predicted SAT Math Score (y)

$$\hat{y} = 37(12) + 270$$

Predicted Score: 714 points

Self Check Quiz!

Self Check Quiz: Calculate the Regression Equation

A crazy professor believes that a child with IQ 100 should have a reading test score of 50, and that reading score should increase by 1 point for every additional point of IQ. What is the equation of the professor's regression line for predicting reading score from IQ? Be sure to identify all variables used.

Self Check Quiz: Calculate the Regression Equation

A crazy professor believes that a child with IQ 100 should have a reading test score of 50, and that reading score should increase by 1 point for every additional point of IQ. What is the equation of the professor's regression line for predicting reading score from IQ? Be sure to identify all variables used.

Answer:

$\hat{y}$ = 50 + x

$\hat{y}$ = predicted reading score

x = number of IQ points above 100

Self Check Quiz: Interpreting Regression Lines & Predicted Value

Data on the IQ test scores and reading test scores for a group of fifth-grade children resulted in the following regression line:

predicted reading score = −33.4 + 0.882(IQ score)

(a) What's the slope of this line? Interpret this value in context.

(b) What's the y-intercept? Explain why the value of the intercept is not statistically meaningful.

(c) Find the predicted reading scores for two children with IQ scores of 90 and 130, respectively.

predicted reading score = −33.4 + 0.882(IQ score)

(a) Slope = 0.882. For each 1 point increase of IQ score, the reading score is predicted to increase 0.882 points, on average.

(b) Y-intercept= -33.4. If the student has an IQ of zero, which is essential impossible (would not be able to hold a pencil to take the exam), the score would be -33.4. This has no practical interpretation.

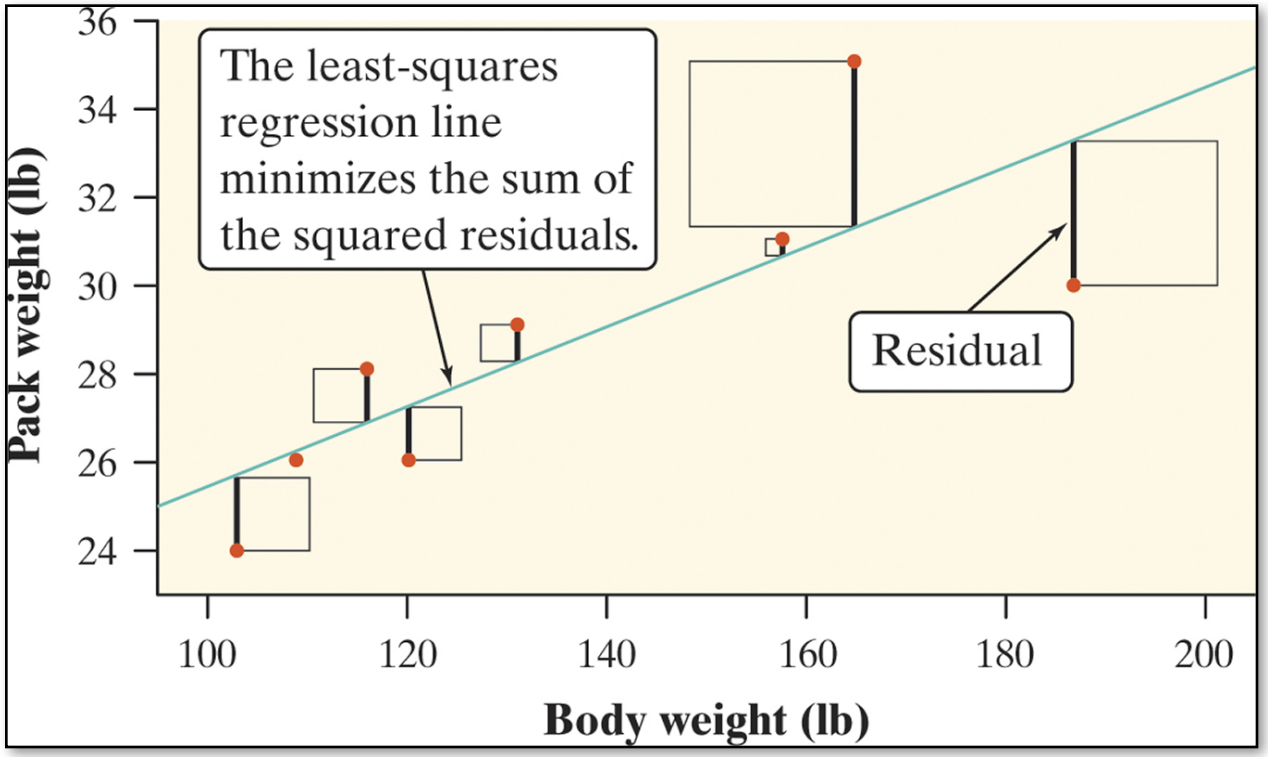
(c) Predicted Value: 90: -33.4 + 0.882(90) = 45.98

130: -33.4 + 0.882(130) = 81.26 points.

Least-Squares Regression Line

Different regression lines produce different residuals. The regression line we use in AP Stats is Least-Squares Regression.

The least-squares regression line of y on x is the line that makes the sum of the squared residuals as small as possible.

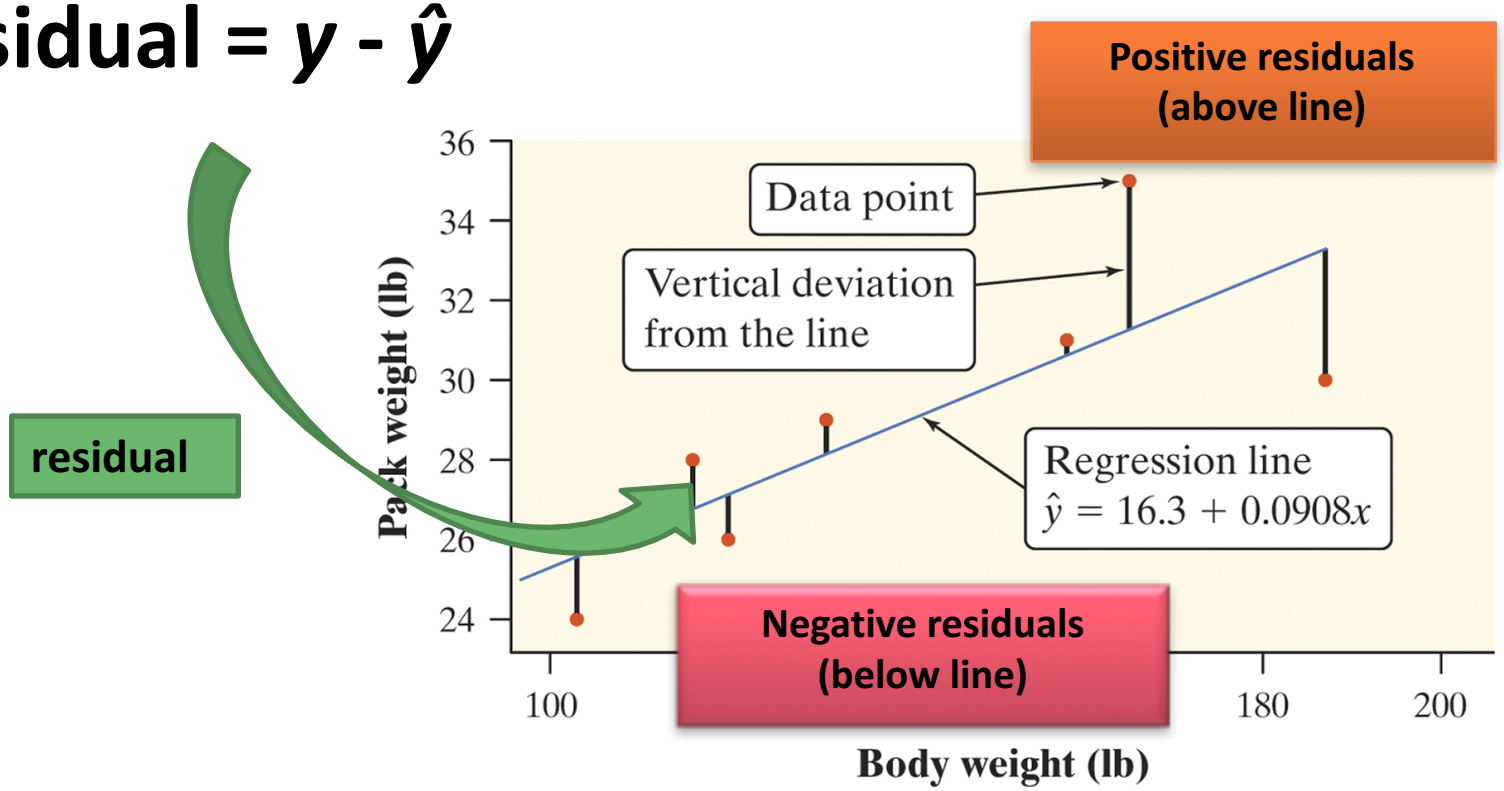


Residuals

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

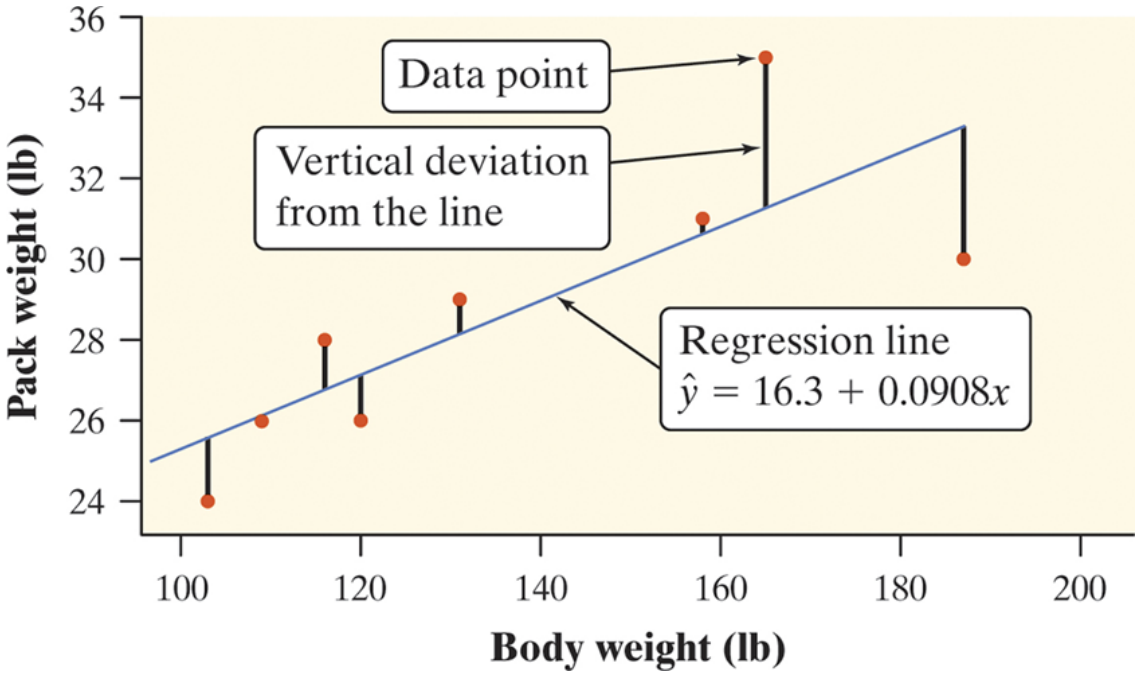
residual = actual *y* – predicted *y* **(remember AP)**

residual = $y - \hat{y}$



How to Calculate the Residual

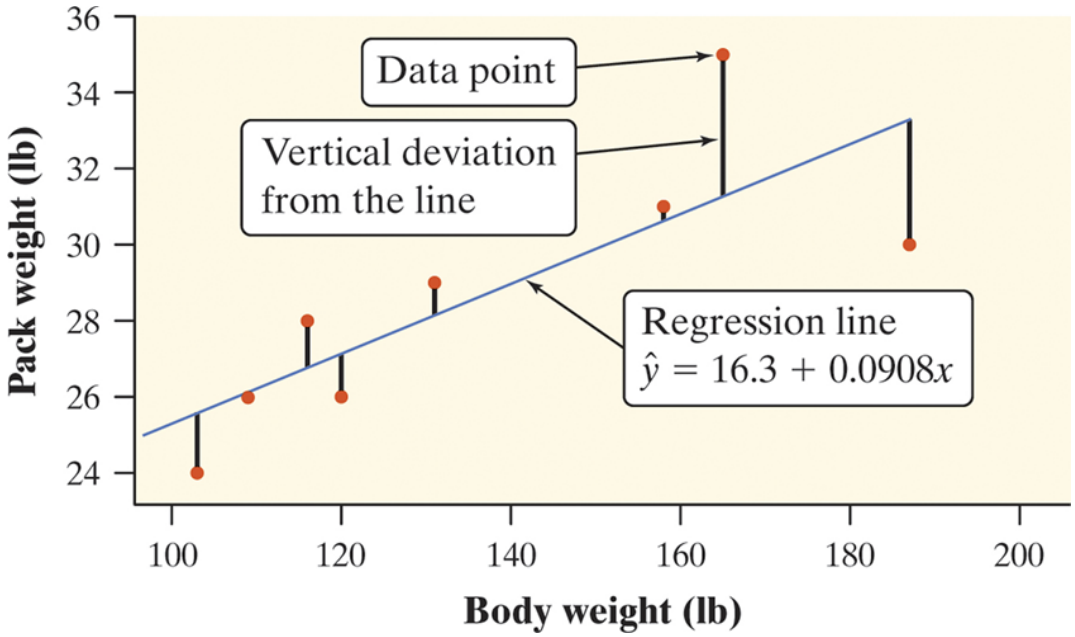
1. Calculate the predicted value, by plugging in x to the LSRE.
2. Determine the observed/actual value.
3. Subtract.



Calculate the Residual

1. If a student weighs 170 pounds and their backpack weighs 35 pounds, what is the value of the residual?

2. If a student weighs 105 pounds and their backpack weighs 24 pounds, what is the value of the residual?



Calculate the Residual

1. If a student weighs 170 pounds and their backpack weighs 35 pounds, what is the value of the residual?

Predicted: $\hat{y} = 16.3 + 0.0908\ (170) = 31.736$

Observed: 35

Residual: 35 - 31.736 = **3.264 pounds**

The student's backpack weighs 3.264 pounds more than predicted.

Calculate the Residual

2. If a student weighs 105 pounds and their backpack weighs 24 pounds, what is the value of the residual?

Predicted: $\hat{y} = 16.3 + 0.0908\ (105) = 25.834$

Observed: 24

Residual: $24 - 25.834 = -1.834$

The student's backpack weighs 1.834 pounds less than predicted

Check Your Understanding

Some data were collected on the weight of a male white laboratory rat for the first 25 weeks after its birth. A scatterplot of *y* = weight (in grams) and *x* = time since birth (in weeks) shows a fairly strong, positive linear relationship. The regression equation $\hat{y} = \mathbf{100} + \mathbf{40}\boldsymbol{x}$ models the data well.

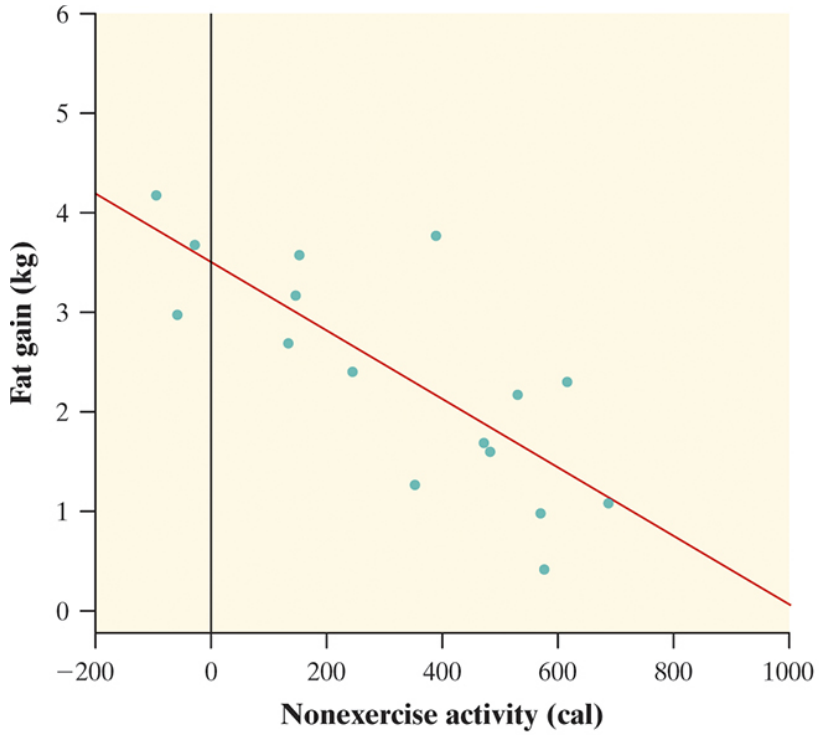
A. Predict the rat's weight at 16 weeks old.

B. Calculate and interpret the residual if the rat weighed 700 grams at 16 weeks old

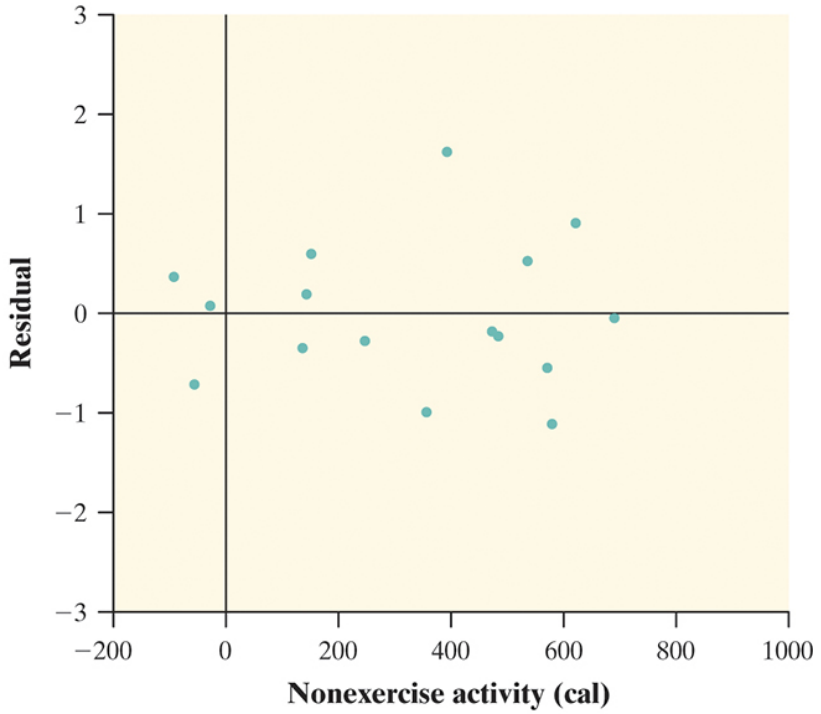
C. Should you use this line to predict the rat's weight at 2 years old?

Residual Plots

A **residual plot** is a scatterplot of the residuals against the explanatory variable. Residual plots help us assess how well a regression line fits the data.



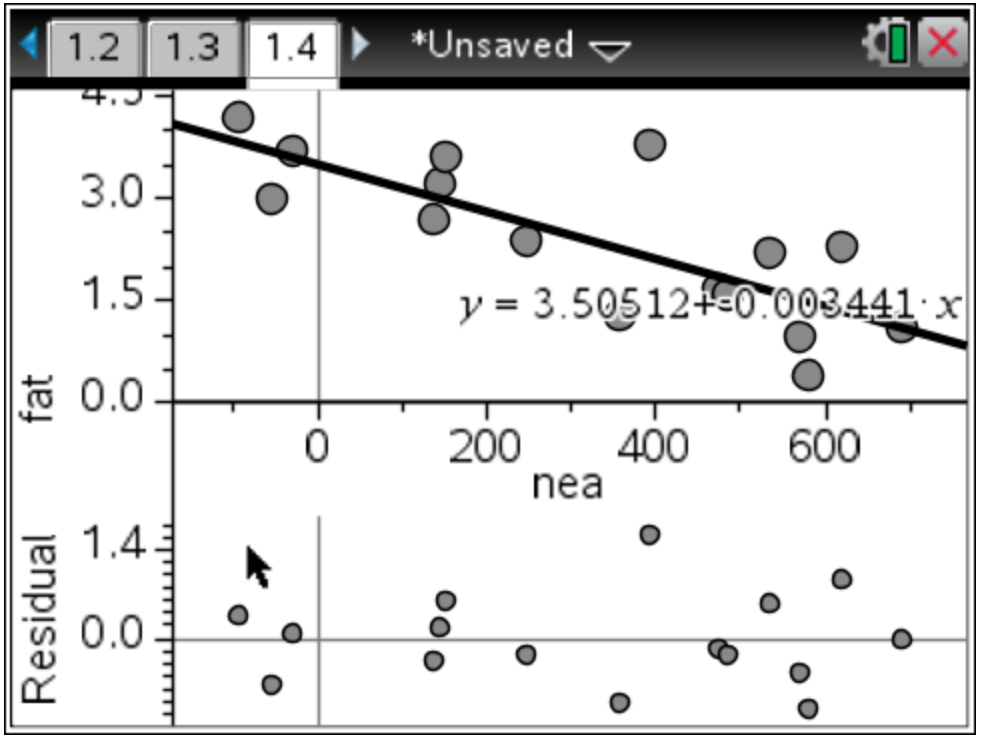
(a)



(b)

TI-NSpire: Residual Plots

1. Press MENU, 4: Analyze
2. Option 6: Residual, Option 2: Show Residual Plot

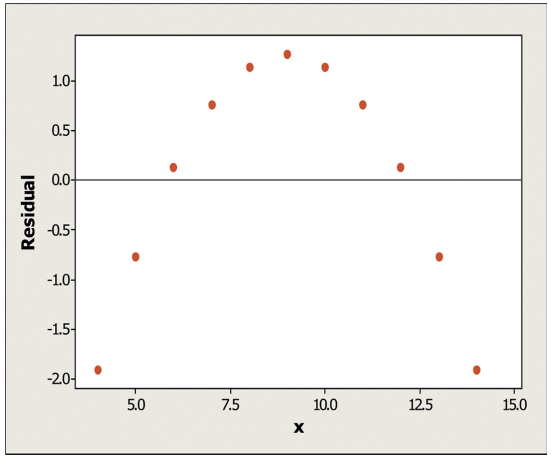


Interpreting Residual Plots

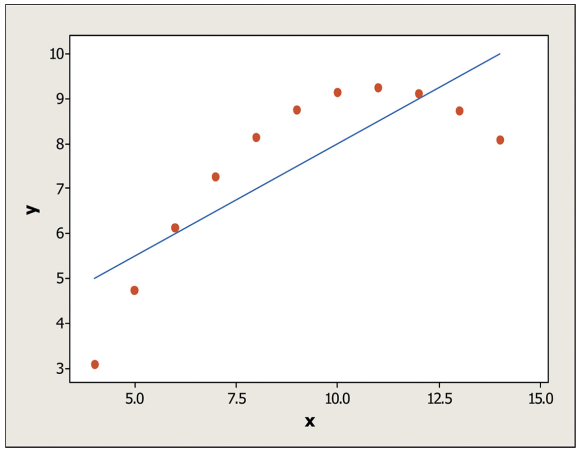
A residual plot magnifies the deviations of the points from the line, making it easier to see unusual observations and patterns.

1) The residual plot should show no obvious patterns
2) The residuals should be relatively small in size.

A valid residual plot should look like the "night sky" with approximately equal amounts of positive and negative residuals.

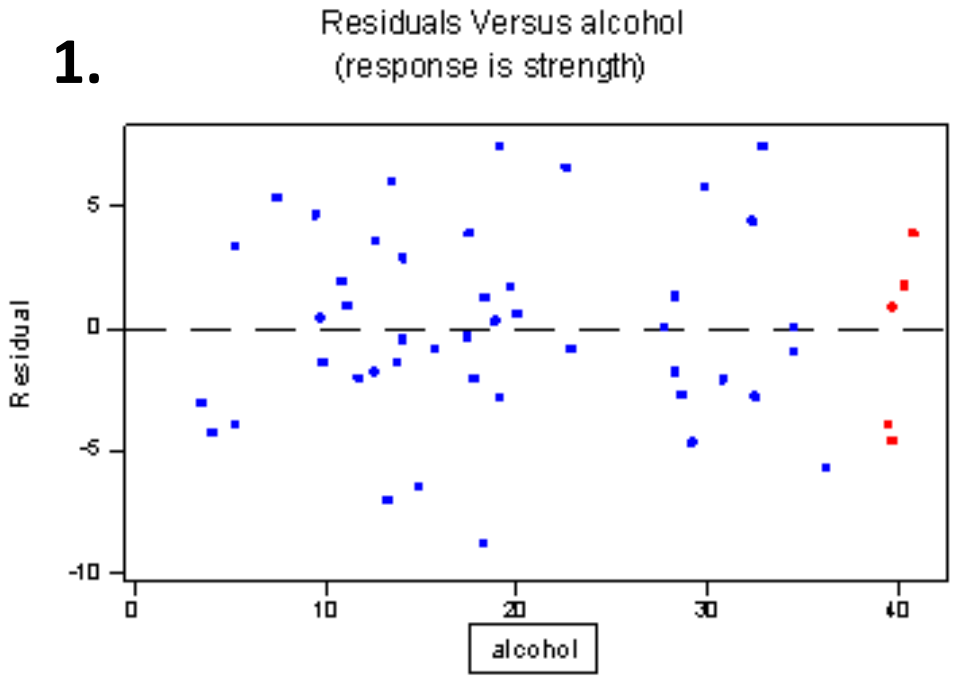


Pattern in residuals
Linear model not appropriate

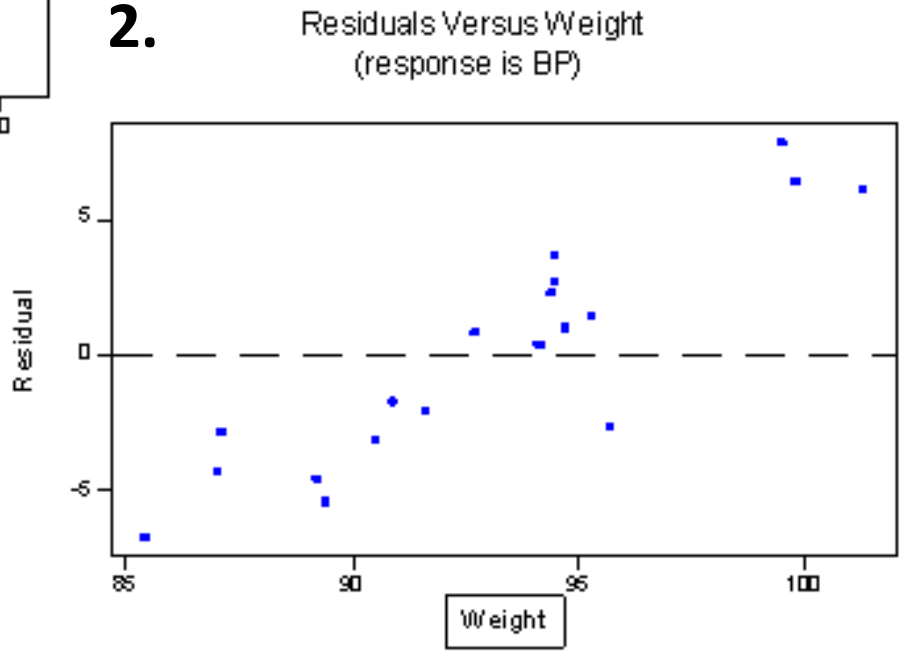


Should You Use LSRL?

1.

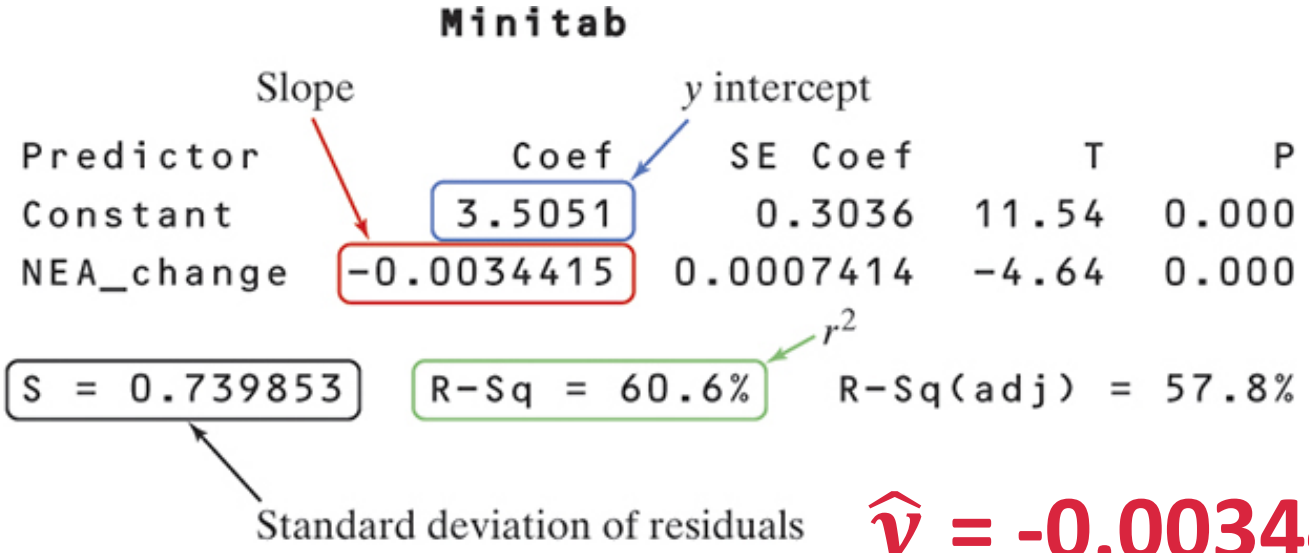


2.



Interpreting Computer Regression Output

Be sure you can locate: the slope, the *y* intercept and determine the equation of the LSRL.



$\hat{y}$ = -0.0034415x + 3.5051

$\hat{y}$ = predicted....

x = explanatory variable

Determine is the equation of the LSRL.

Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001
S = 200.01		R-Sq = 73.33%	R-Sq (adj) = 70.37%	

Determine is the equation of the LSRL.

Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001
S = 200.01		R-Sq = 73.33%	R-Sq (adj) = 70.37%	

$\hat{y}$ = 174.40x + 72.95

x = customers in line

$\hat{y}$ = predicted seconds it takes to check out.

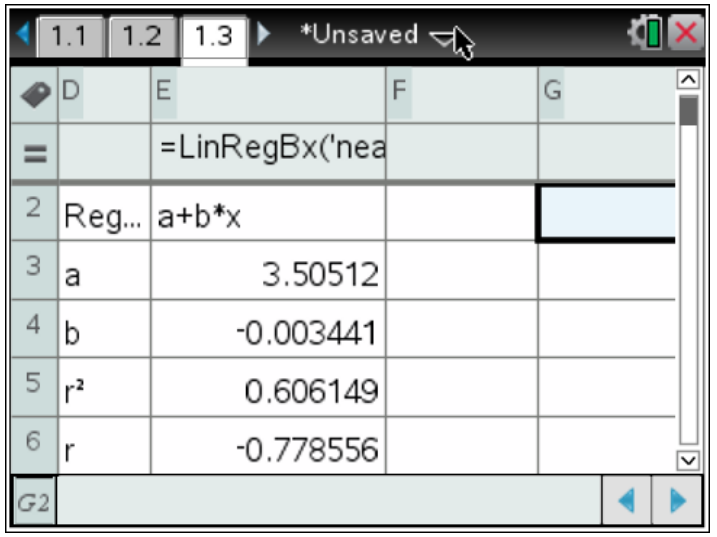
r^2: Coefficient of Determination

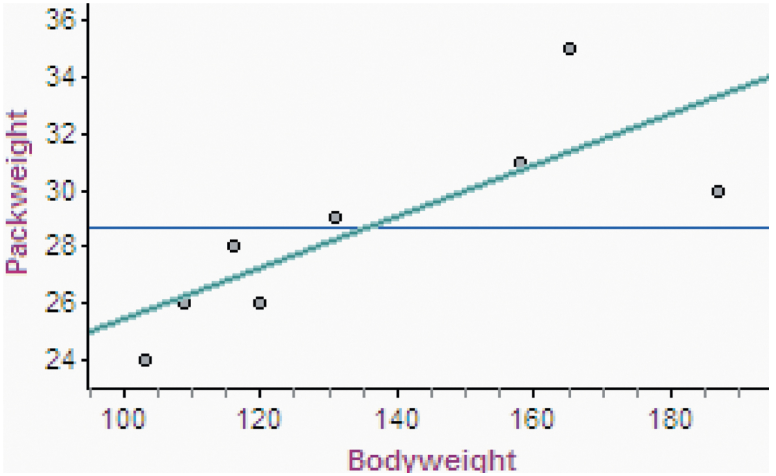
r^2 tells us how much better the LSRL does at predicting values of *y* than simply guessing the mean *y* for each value in the dataset.

In this example, r^2 equals 60.6%.

60.6% of the variation in pack weight is explained by the linear relationship with bodyweight.

(Insert r^2)% of the variation in **y** is explained by the linear relationship with **x**.





Interpret r^2

Interpret in a sentence (how much variation is accounted for?)

1. $r^2 = 0.875$, x= hours studied, y= SAT score
2. $r^2 = 0.523$, x= hours slept, y= alertness score

Interpret r^2

Answers:

1. 87.5% of the variation in SAT score is explained by the linear relationship with the number of hours studied.

2. 52.3% of the variation in alertness score is explained by the linear relationship with the number of hours slept.

S: Standard Deviation of the Residuals

If we use a least-squares regression line to predict the values of a response variable *y* from an explanatory variable *x*, the **standard deviation of the residuals (s)** is given by

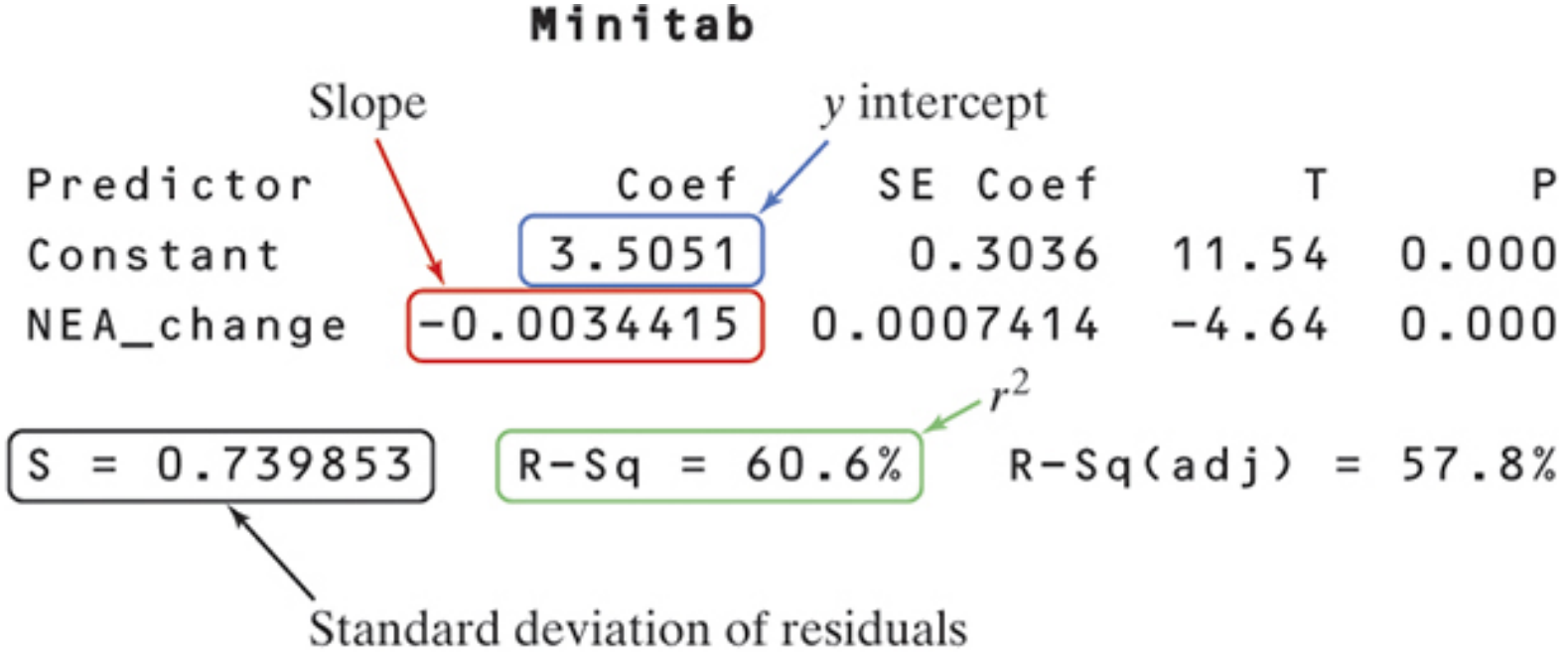
$$s = \sqrt{\frac{residuals^2}{n \quad 2}} = \sqrt{\frac{(y_i \quad \hat{y})^2}{n \quad 2}}$$

S represents the typical or average ERROR (residual).

Positive = UNDER predicts

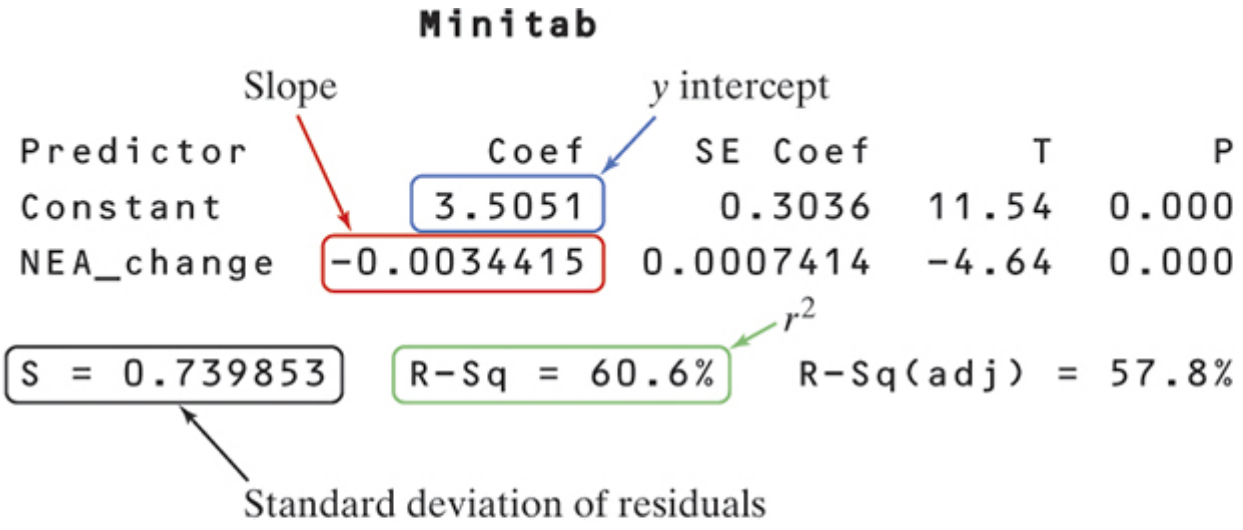
Negative = OVER predicts

S: Standard Deviation of the Residuals



1. Identify and interpret the standard deviation of the residual.

S: Standard Deviation of the Residuals



Answer:
S= 0.740

Interpretation: On average, the model under predicts fat gain by 0.740 kilograms using the least-squares regression line.

Self Check Quiz!

The data is a random sample of 10 trains comparing number of cars on the train and fuel consumption in pounds of coal.

- What is the regression equation? Be sure to define all variables.
- What is r^2 telling you?
- Define and interpret the slope in context. Does it have a practical interpretation?
- Define and interpret the y-intercept in context.
- What is s telling you?

Predictor	Coef	StDev	T	P
Constant	10.677	5.157	2.07	0.072
Railcar	2.1495	0.1396	15.40	0.000

S = 4.361 R-Sq = 96.7% R-Sq(adj) = 96.3%

1. $\hat{y} = 2.1495x + 10.667$

$\hat{y}$ = predicted fuel consumption in pounds of coal

x = number of rail cars

2. 96.7 % of the varation is fuel consumption is explained by the linear relationship with the number of rail cars.

3. Slope = 2.1495. With each additional car, the fuel consuption increased by 2.1495 pounds of coal, on average. This makes practical sense.

4. Y-interpect = 10.667. When there are no cars attached to the train the fuel consuption is 10.667 pounds of coal. This has no practical intrepretation beacuse there is always at least one car, the engine.

5. S= 4.361. On average, the model under predicts fuel consumption by 4.361 pounds of coal using the least-squares regression line.

Extrapolation

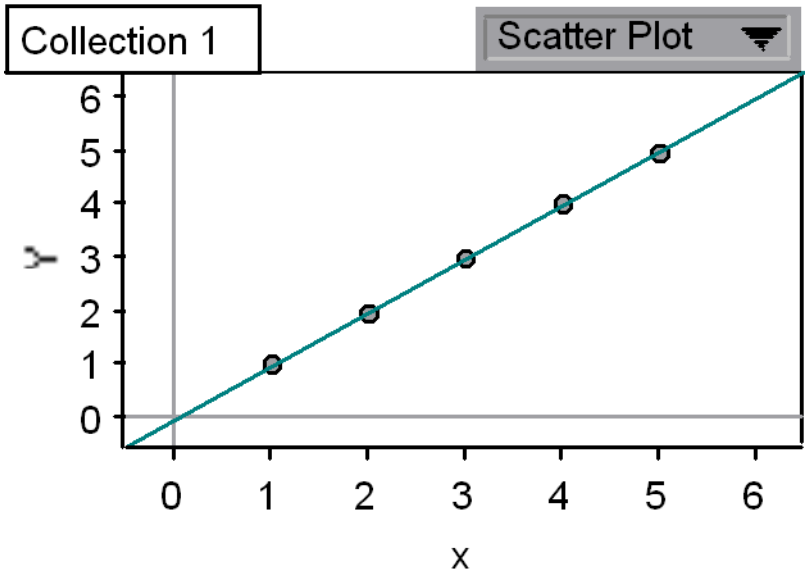
We can use a regression line to predict the response $\hat{y}$ for a specific value of the explanatory variable x. The accuracy of the prediction depends on how much the data scatter about the line. **Exercise caution in making predictions outside the observed values of *x*.**

Extrapolation is the use of a regression line for prediction far outside the interval of values of the explanatory variable *x* used to obtain the line. Such predictions are often not accurate.

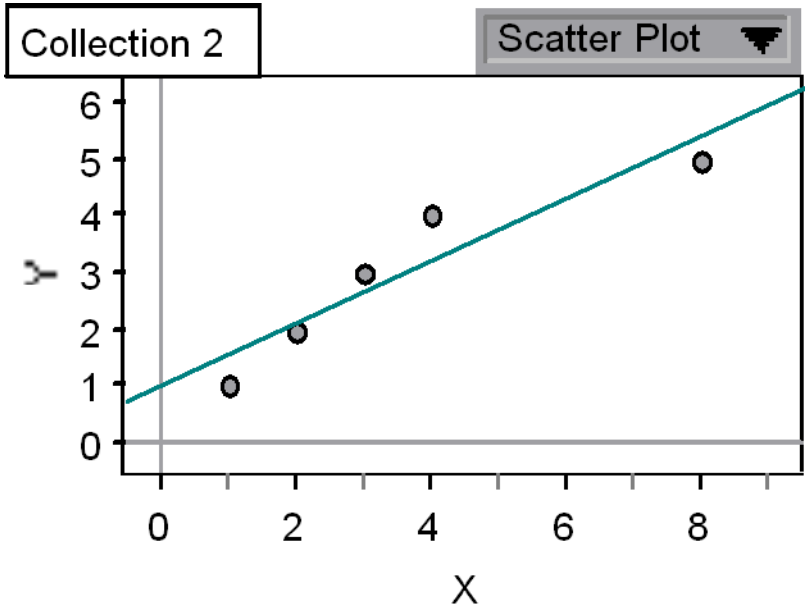
Outliers and Influential Points

- An outlier is an observation that lies outside the overall pattern of the other observations.
- An observation is influential for a statistical calculation if removing it would markedly change the result of the calculation.
- Points that are outliers in the *x* direction of a scatterplot are often influential for the least-squares regression line.
- **Note**: Not all influential points are outliers, nor are all outliers influential points.

Outliers and Influential Points



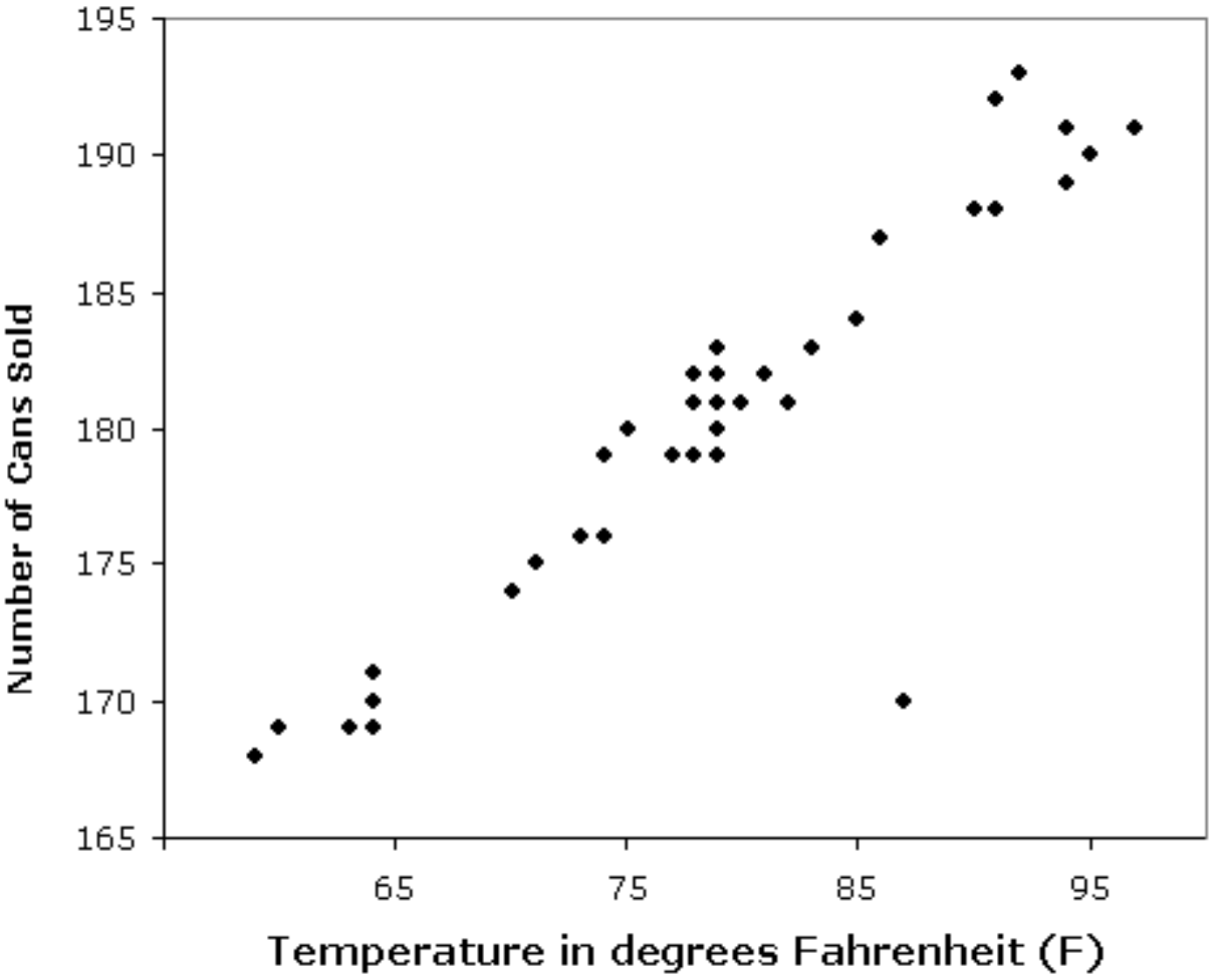
Y = x; r^2 = 1.0



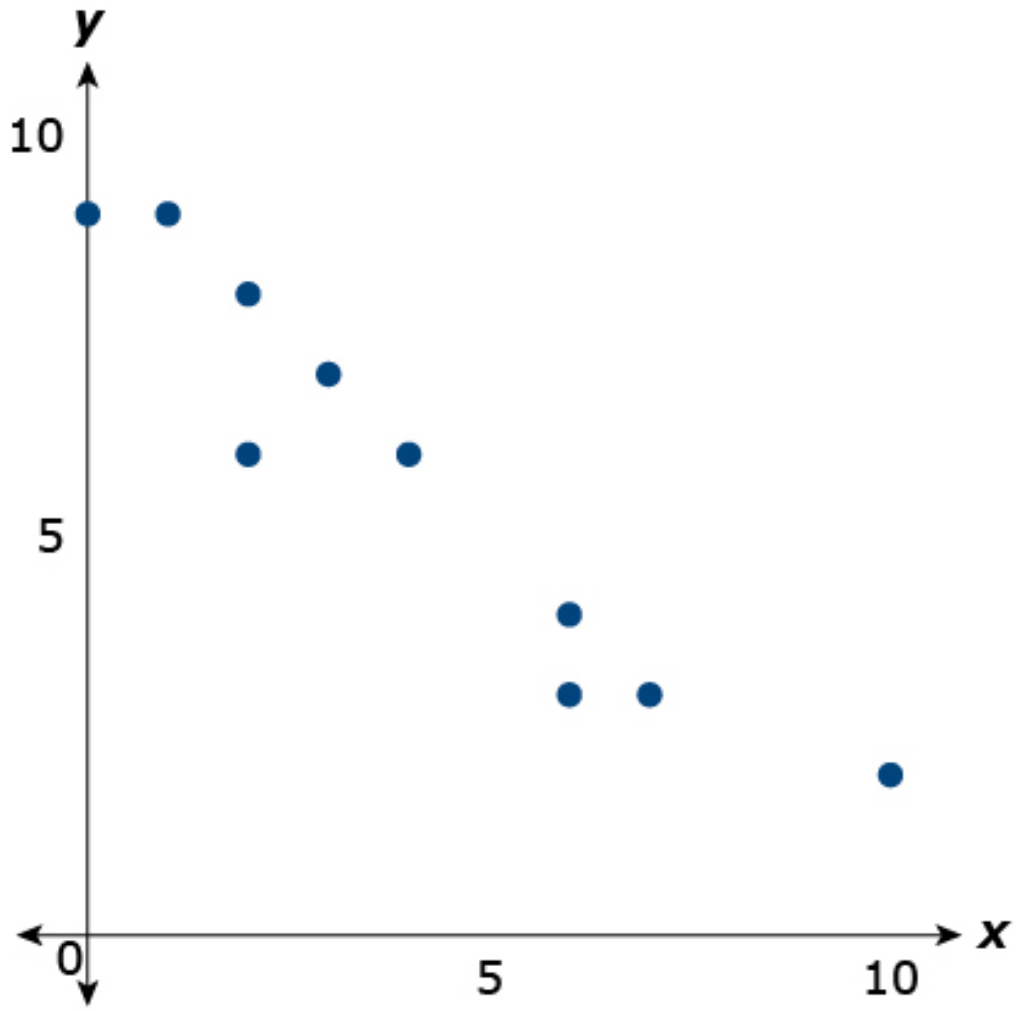
Y = 0.548X + 1.0; r^2 = 0.88

The left graph is perfectly linear. In the right graph, the last value was changed from (5, 5) to (8, 5)…clearly influential, because it changed the graph significantly. However, the residual is very small.

Identify the Outlier…

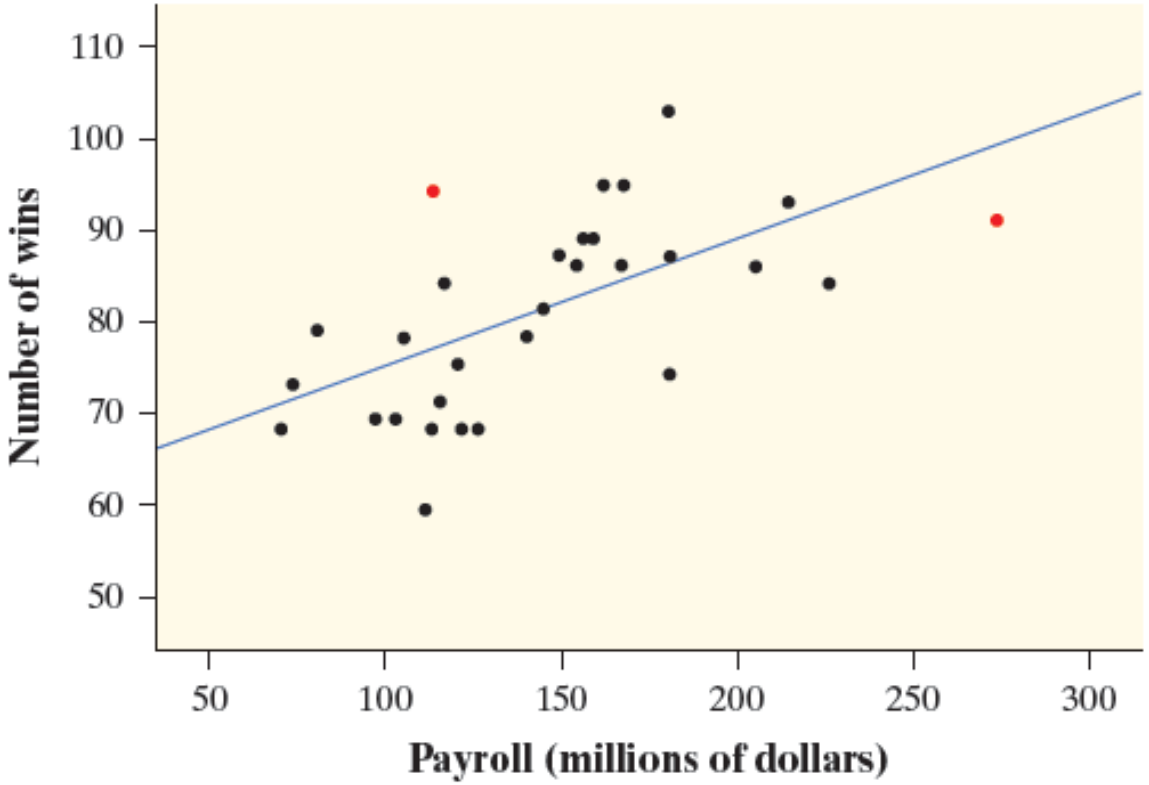


Identify the Outlier…



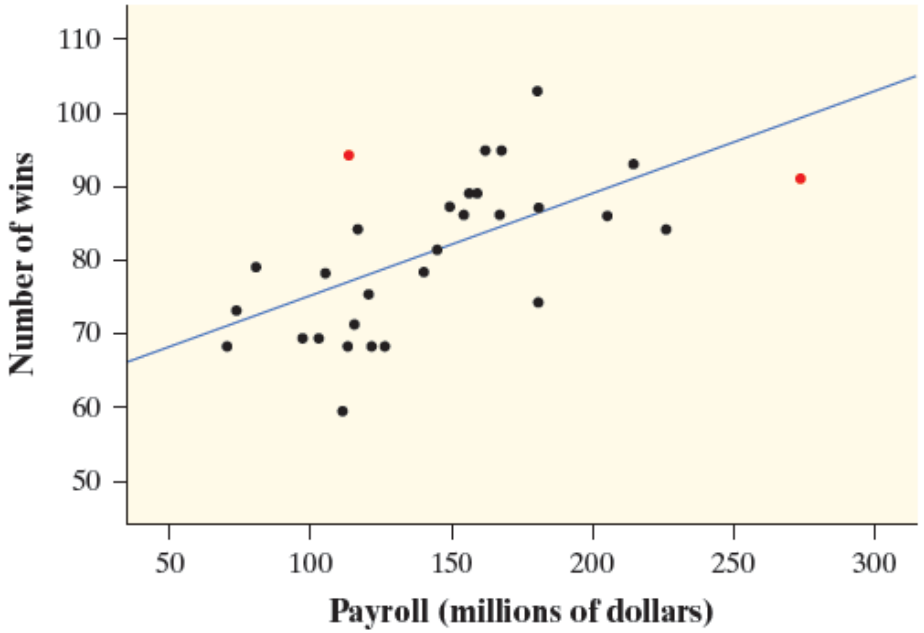
Check Your Understanding

The scatterplot shows the payroll (in millions of dollars) and number of wins for Major League Baseball teams in 2016, along with the least-squares regression line. The points highlighted in red represent the Los Angeles Dodgers (far right) and the Cleveland Indians (upper left).



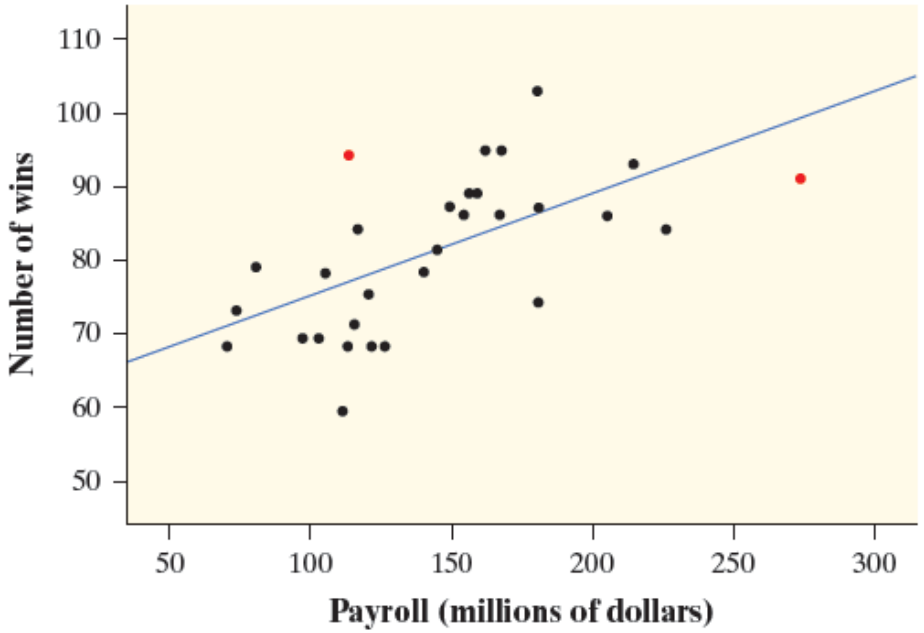
Check Your Understanding

A. Describe what influence the point representing the Los Angeles Dodgers has on the **equation of the least-squares** regression line. Explain your reasoning.



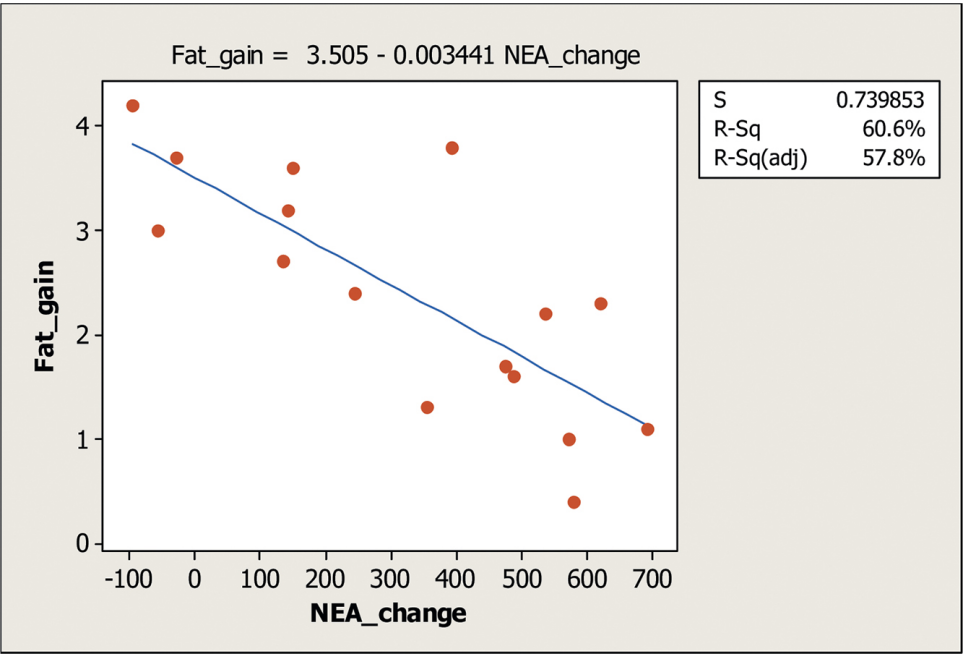
Check Your Understanding

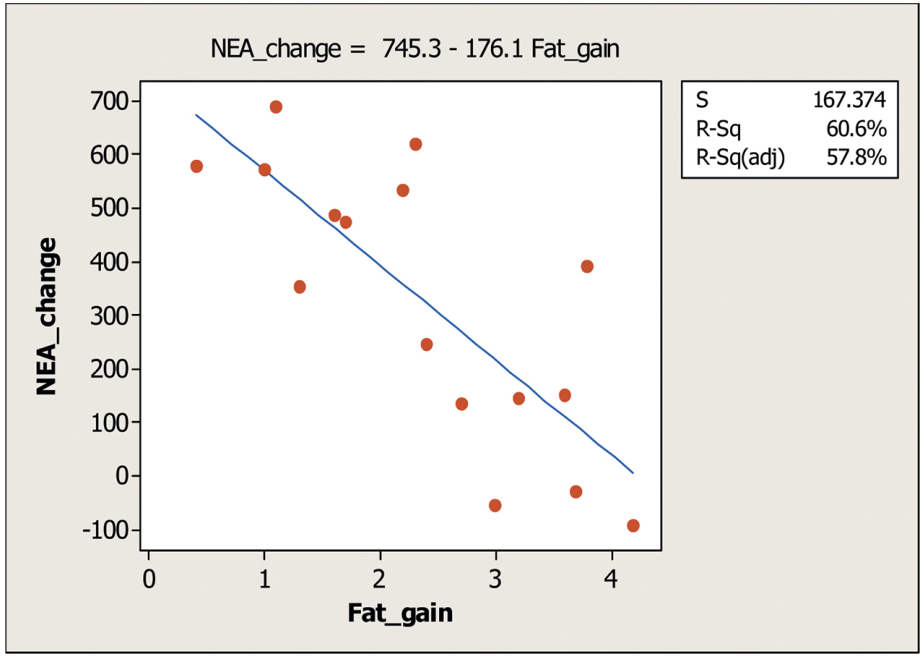
B. Describe what influence the point representing the Cleveland Indians has on the **standard deviation** of the residuals and r^2. Explain your reasoning.



Correlation and Regression Limitations

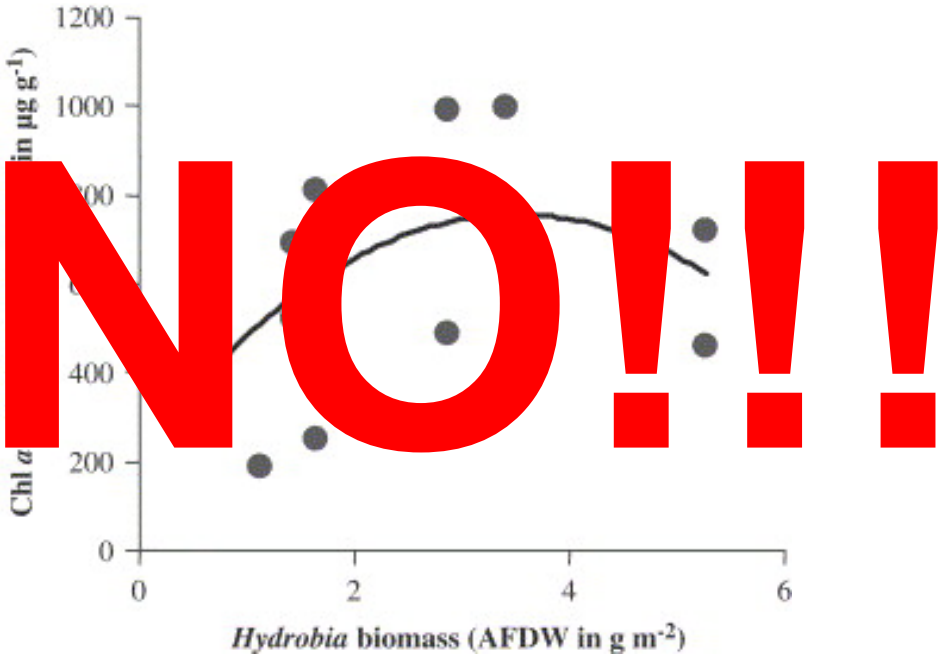
The distinction between explanatory and response variables is important in regression.





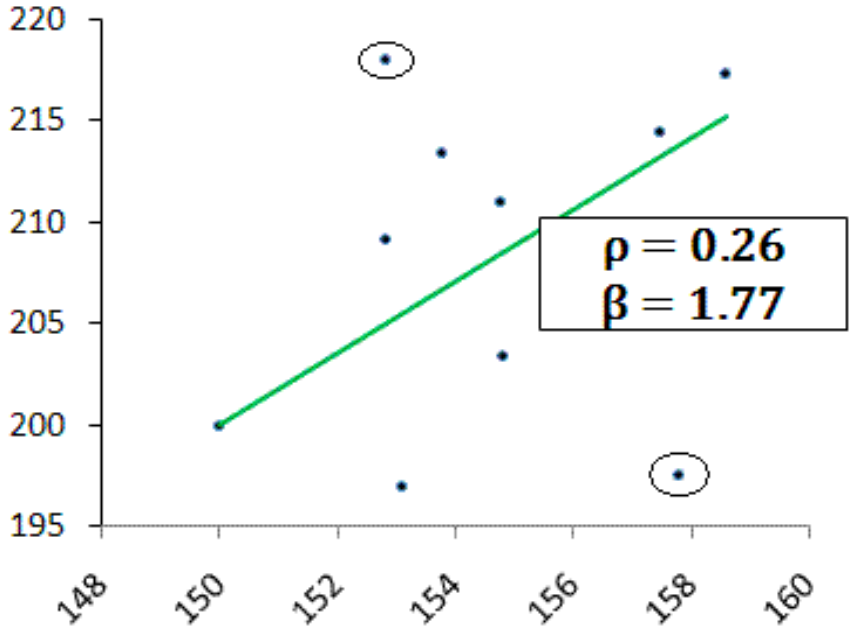
Correlation and Regression Limitations

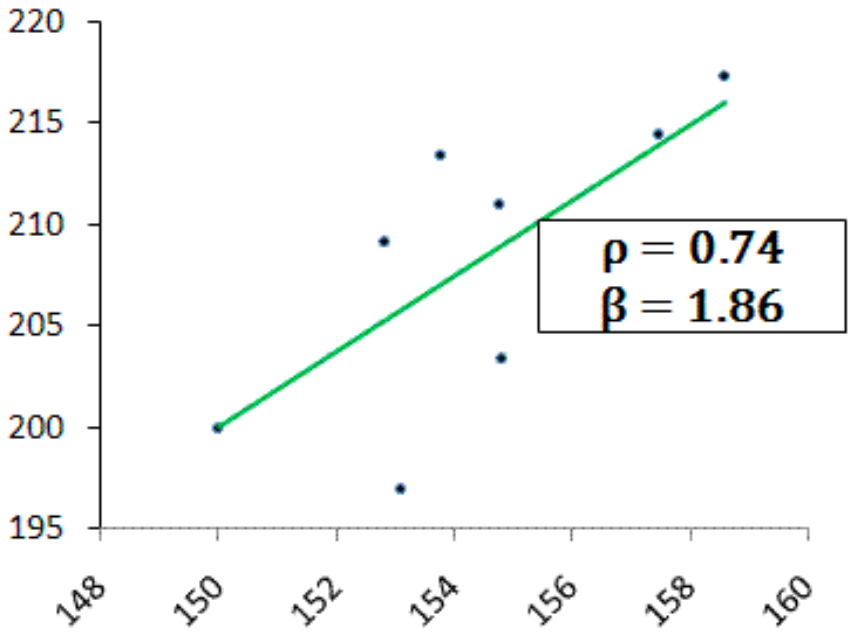
Correlation and regression lines describe only linear relationships.

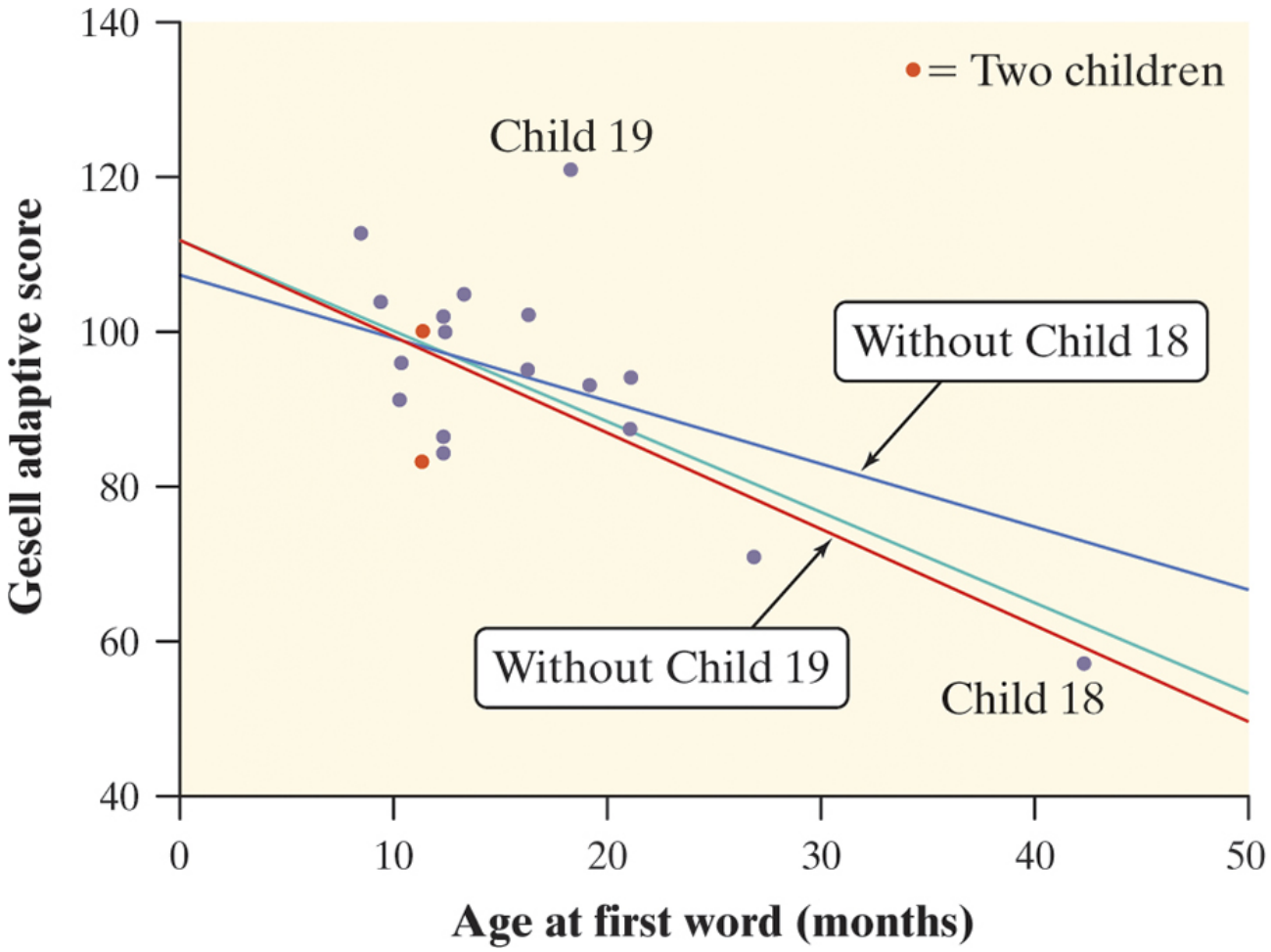


Correlation and Regression Limitations

Correlation and least-squares regression lines are not resistant.







With all 19 children:

$r = -0.64$

$\hat{y} = 109.874 - 1.127x$

Without Child 19:

$r = -0.76$

$\hat{y} = 109.305 - 1.193x$

Without Child 18:

$r = -0.33$

$\hat{y} = 105.630 - 0.779x$

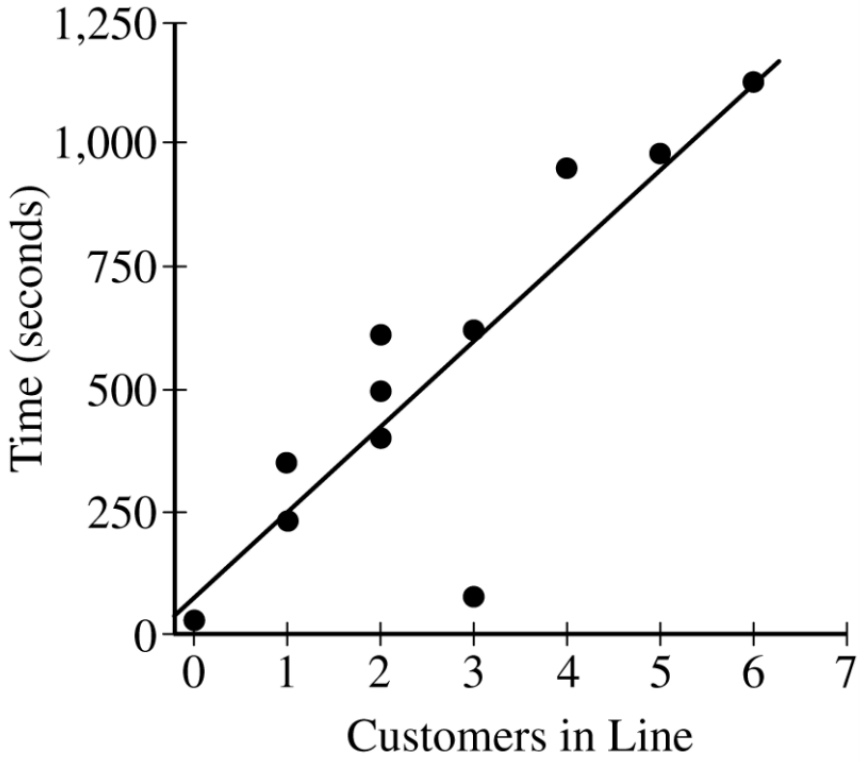
Correlation and Regression Wisdom

Association Does Not Imply Causation

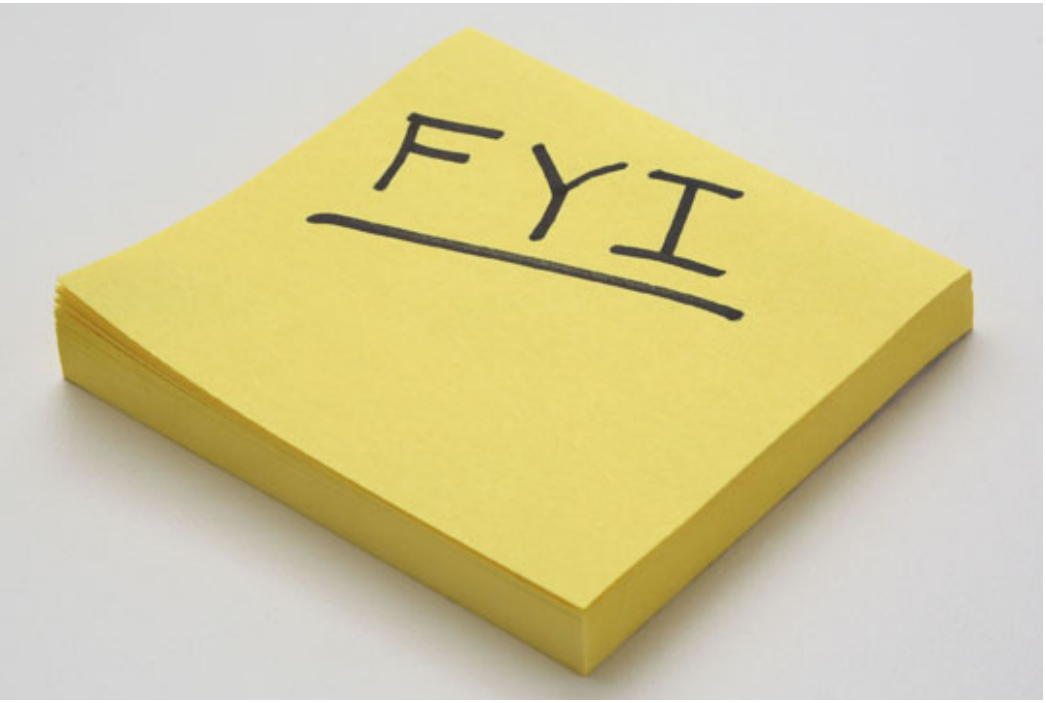
An association between an explanatory variable *x* and a response variable *y*, even if it is very strong, is not by itself good evidence that changes in *x* actually cause changes in *y*.

A serious study once found that people with two cars live longer than people who only own one car. Owning three cars is even better, and so on. There is a substantial positive correlation between number of cars *x* and length of life *y*. **Why?**

FRQ 2018 #1



Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001
S = 200.01		R-Sq = 73.33%	R-Sq (adj) = 70.37%	



Additional Calculations & Proofs

Least-Squares Regression Line

We can use technology to find the equation of the least-squares regression line. We can also write it in terms of the means and standard deviations of the two variables and their correlation.

Equation of the least-squares regression line

We have data on an explanatory variable *x* and a response variable *y* for *n* individuals. From the data, calculate the means and standard deviations of the two variables and their correlation. The least squares regression line is the line $\hat{y} = a + bx$ with

slope $b = r\frac{s_y}{s_x}$ and ***y* intercept** $a = \bar{y} \quad b\bar{x}$

Calculate the Least Squares Regression Line

Some people think that the behavior of the stock market in January predicts its behavior for the rest of the year. Take the explanatory variable *x* to be the percent change in a stock market index in January and the response variable *y* to be the change in the index for the entire year. We expect a positive correlation between *x* and *y* because the change during January contributes to the full year's change. Calculation from data for an 18-year period gives

Mean x =1.75 % S_x= 5.36% Mean y = 9.07%

S_y = 15.35% r = 0.596

Find the equation of the least-squares line for predicting full-year change from January change. Show your work.

The Role of r^2 in Regression

The standard deviation of the residuals gives us a numerical estimate of the average size of our prediction errors.

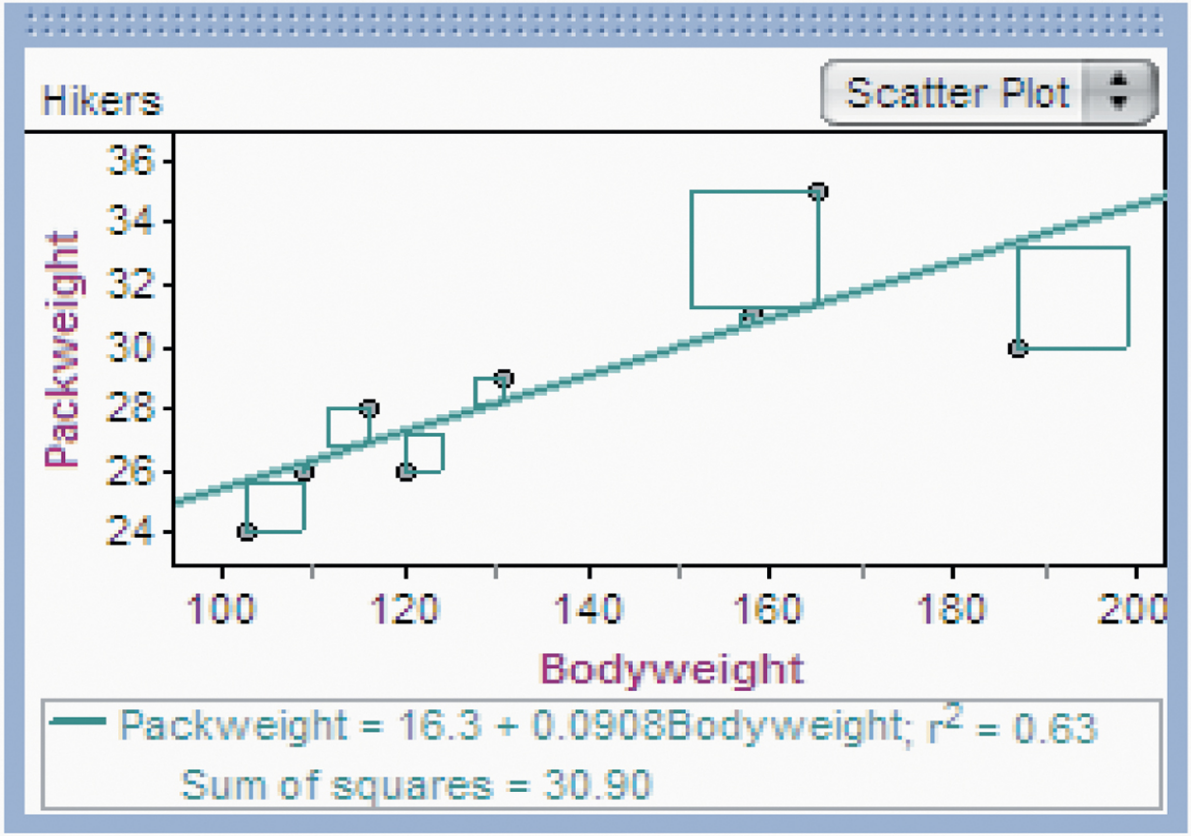
The **coefficient of determination r^2** is the fraction of the variation in the values of *y* that is accounted for by the least-squares regression line of *y* on *x*. We can calculate r^2 using the following formula:

$$r^2 = 1 \quad \frac{\text{SSE}}{\text{SST}}$$

$$SSE = \quad \text{residual}^2 \qquad\qquad SST = \quad (y_i \quad \bar{y})^2$$

In practicality, just square the correlation r.

Accounted for Error



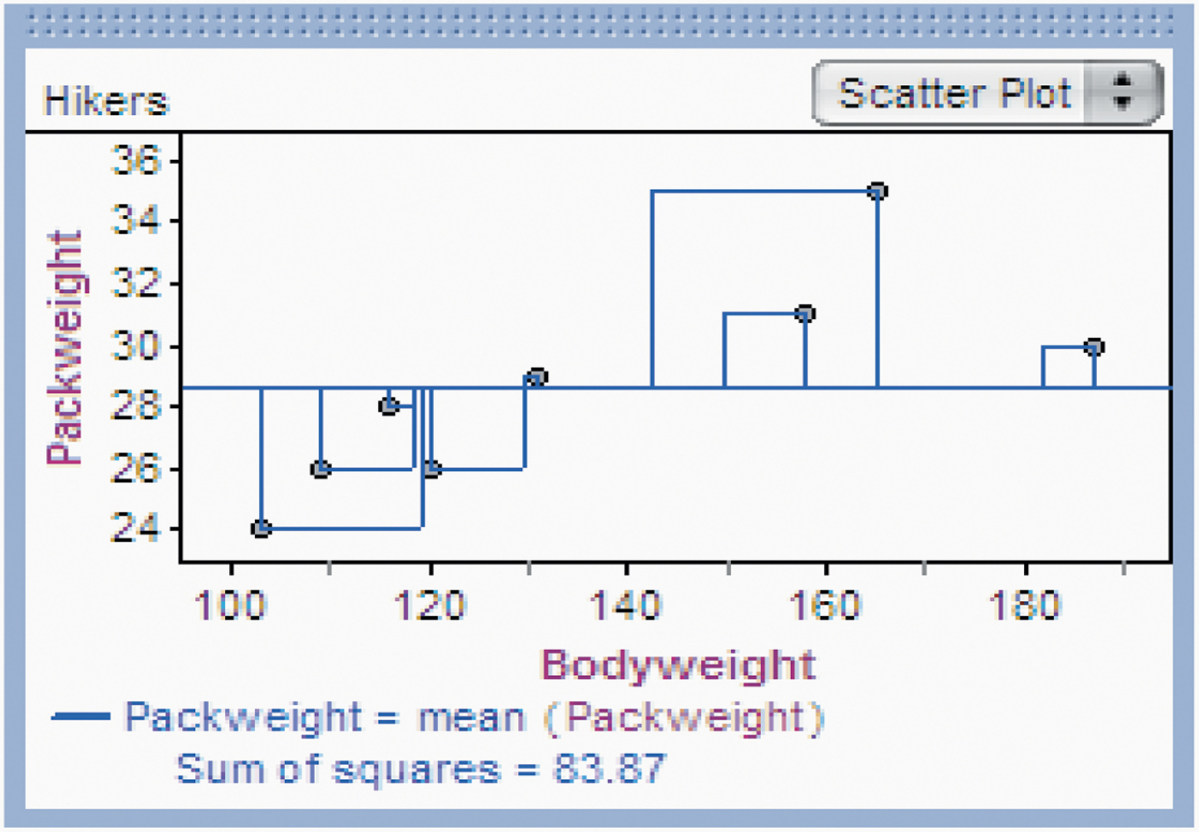
If we use the LSRL to make our predictions, the sum of the squared residuals is 30.90.
SSE = 30.90

1 – SSE/SST = 1 – 30.97/83.87

r^2 = 0.632

63.2 % of the variation in backpack weight is accounted for by the linear model relating pack weight to body weight.

Unaccounted for Error



If we use the mean backpack weight as our prediction, the sum of the squared residuals is 83.87.

SST = 83.87

SSE/SST = 30.97/83.87

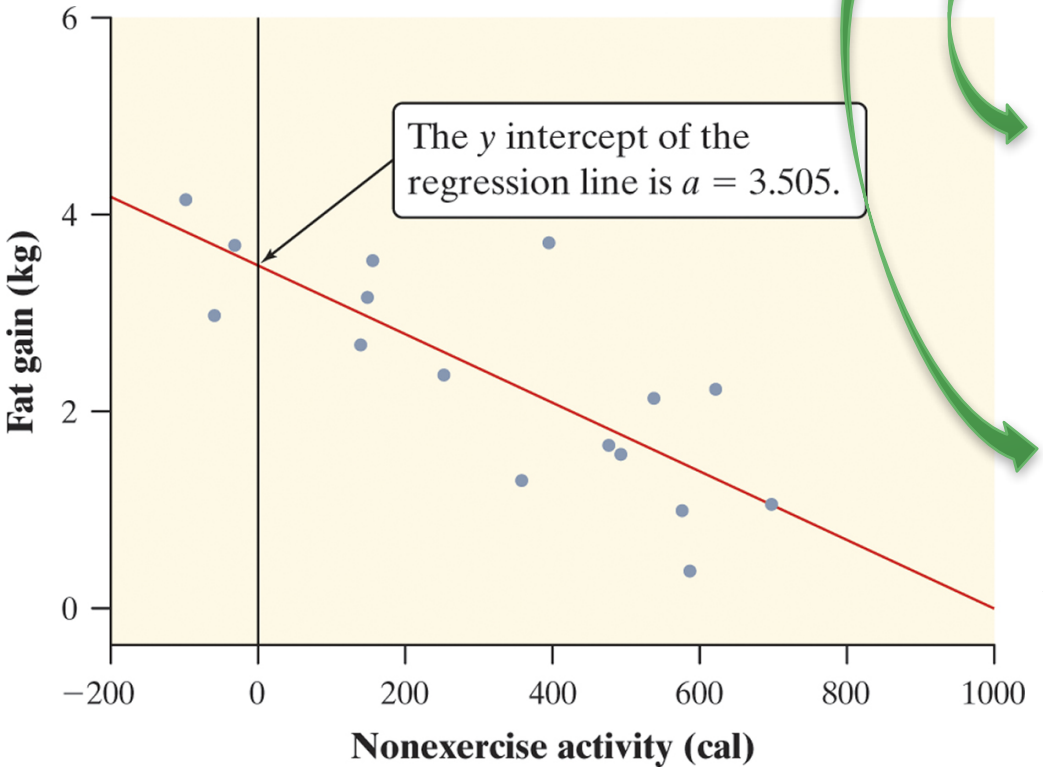
SSE/SST = 0.368

Therefore, 36.8% of the variation in pack weight is *unaccounted for* by the least-squares regression line.

Interpreting a Regression Line

Consider the regression line from the example (pg. 164) "Does Fidgeting Keep You Slim?" Identify the slope and *y*-intercept and interpret each value in context.

$$\widehat{fatgain} = 3.505 - 0.00344(NEA\ change)$$



The slope $b = -0.00344$ tells us that the amount of fat gained is predicted to go down by 0.00344 kg for each added calorie of NEA.

The *y*-intercept $a = 3.505$ kg is the fat gain estimated by this model if NEA does not change when a person overeats.