Chapter 5: Linear Regression

Section 5.2

Least-Squares Regression Line (LSRL)

Warm-UP

1. Given a sample of test scores that are normally distributed with a $\bar{x} = 82$ and a $s_x = 4.2$. Determine the difference between the two scores for individual A and individual B, knowing the following:

Individual A: $z\ score = -1.5$

Individual B: $z\ score = 2.0$

2. What is the Pearson sample correlation coefficient?

Sept 22/23 Warm-UP

1. Given a sample of test scores that are normally distributed with a $\bar{x} = 82$ and a $s_x = 4.2$. Determine the difference between the two scores for individual A and individual B, knowing the following:

A: $z\ score = -1.5$

$$-1.5 = \frac{x - 82}{4.2}$$
$$-6.3 = x - 82$$
$$x = 75.7$$

B: $z\ score = 2.0$

$$2.0 = \frac{x - 82}{4.2}$$
$$8.4 = x - 82$$
$$x = 90.4$$

The difference in the scores is 14.7 points

Warm-UP

2. What is the Pearson sample correlation coefficient?

The correlation coefficient is a numerical assessment of the strength of the relationship between 2 quantitative variables (*x* and *y*) in a data set that consist of (*x, y*) pairs.

Chapter 5: Summarizing Bivariate Data

- **3.4 & 5.1** Scatterplots and Correlation
- **5.2 Least-Squares Regression Line (LSRL)**
- **5.3 Assessing the Fit of a Line (LSRL)**

Section 3.2
Least-Squares Regression

Learning Objectives

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 Line

Linear (straight-line) relationships between two quantitative variables are common and easy to understand. A **regression line** summarizes the relationship between two variables, but only in settings where one of the variables helps explain or predict the other.

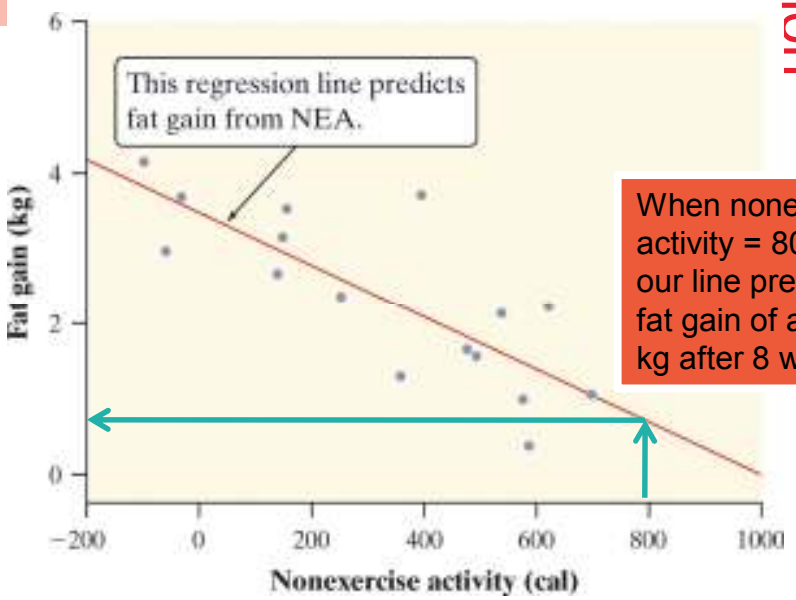
Definition:

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*.

"Does Fidgeting Keep You Slim?"

This example is a scatterplot of the change in nonexercise activity (cal) and measured fat gain (kg) after 8 weeks for 16 healthy young adults.

- ✓ The plot shows a moderately strong, negative, linear association between NEA change and fat gain with no outliers.
- ✓ The regression line predicts fat gain from change in NEA.



When nonexercise activity = 800 cal, our line predicts a fat gain of about 0.8 kg after 8 weeks.

Interpreting a Regression Line

A regression line is a *model* for the data, much like density curves. The equation of a regression line gives a compact mathematical description of what this model tells us about the relationship between the response variable *y* and the explanatory variable *x.*

Definition:

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 form:

$$\hat{y} = a + bx \quad or \quad \hat{y} = b_0 + b_1 x$$

In this equation,

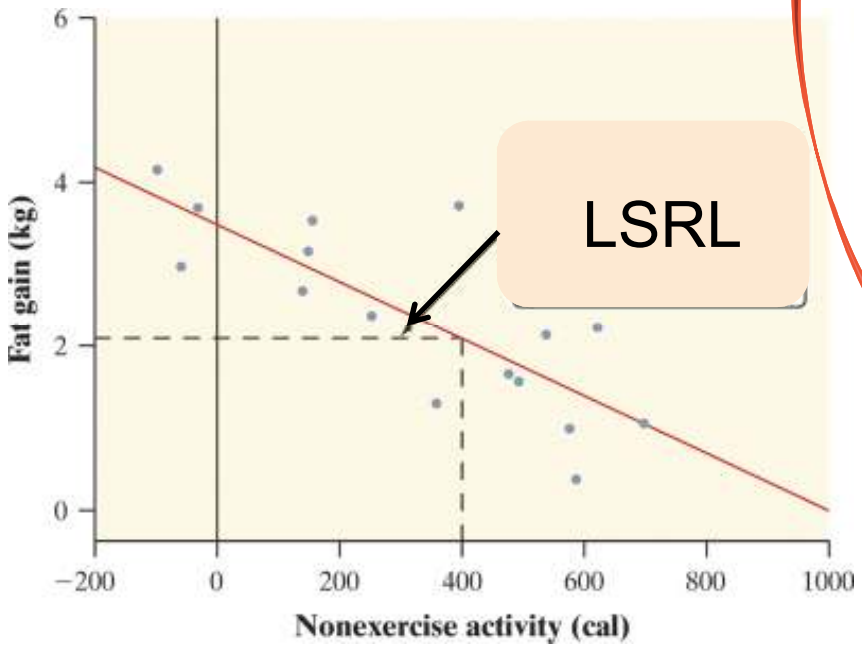
- $\hat{y}$ (read "y hat") is the **predicted value** of the response variable *y* for a given value of the explanatory variable x.
- b or b_1 is the **slope**, the amount by which *y* is predicted to change when *x* increases by one unit.
- a or b_0 is the **y intercept**, the predicted value of *y* when *x* = 0.

Interpreting a Regression Line

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

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

$$\hat{y} = 3.505 - 0.00344x$$



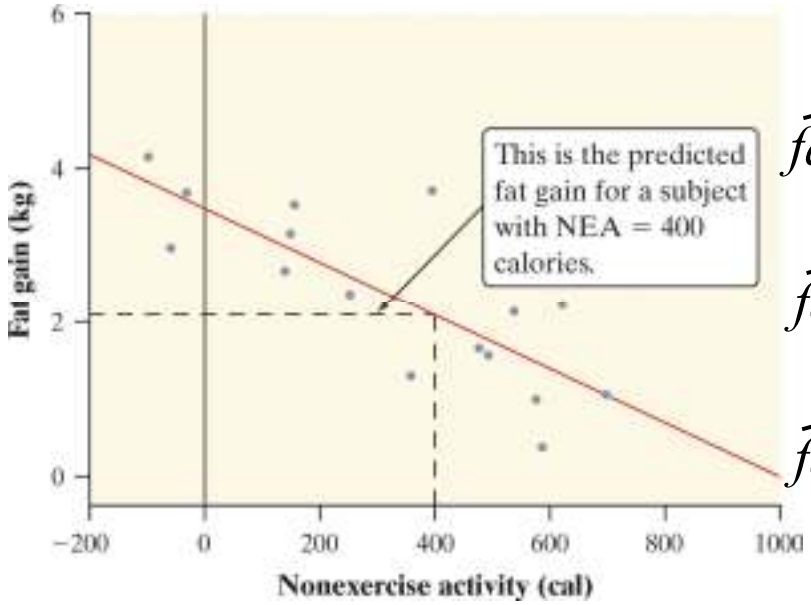
The slope *b* = -0.00344 tells us that the amount of fat gained is predicted to go down by 0.00344 kg for each added calorie of NEA.

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

Prediction

We can use a regression line to predict the response $\hat{y}$ for a specific value of the explanatory variable x.

Use the NEA and fat gain regression line to predict the fat gain for a person whose NEA increases by 400 cal when she overeats.



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

$$\widehat{fatgain} = 3.505 - 0.00344(400)$$

$$\widehat{fatgain} = 2.13$$

We predict a fat gain of 2.13 kg when a person with NEA = 400 calories.

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.

While we can substitute any value of x into the equation of the regression line, we must exercise caution in making predictions outside the observed values of x.

Definition:

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.***

Don't make predictions using values of x that are much larger or much smaller than those that actually appear in your data.

Sept 24/25
Warm-UP

- Problem #5.24 (p.224)
- Put each list into your calculator
- We'll complete the problem together

Residuals

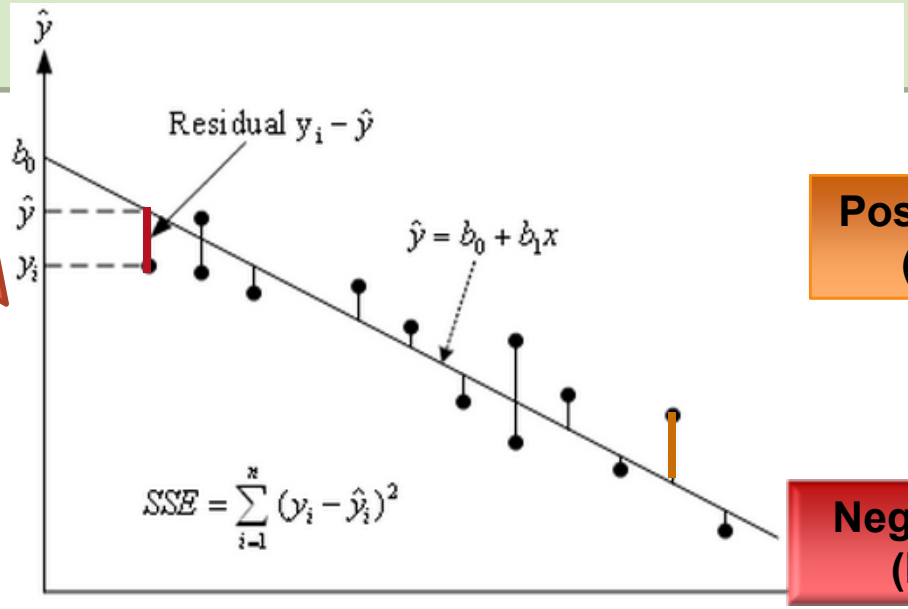
In most cases, no line will pass exactly through all the points in a scatterplot. A good regression line makes the vertical distances of the points from the line as small as possible.

Definition:

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

residual = observed *y* – predicted *y*

residual = $y - \hat{y}$



residual

Positive residuals (above line)

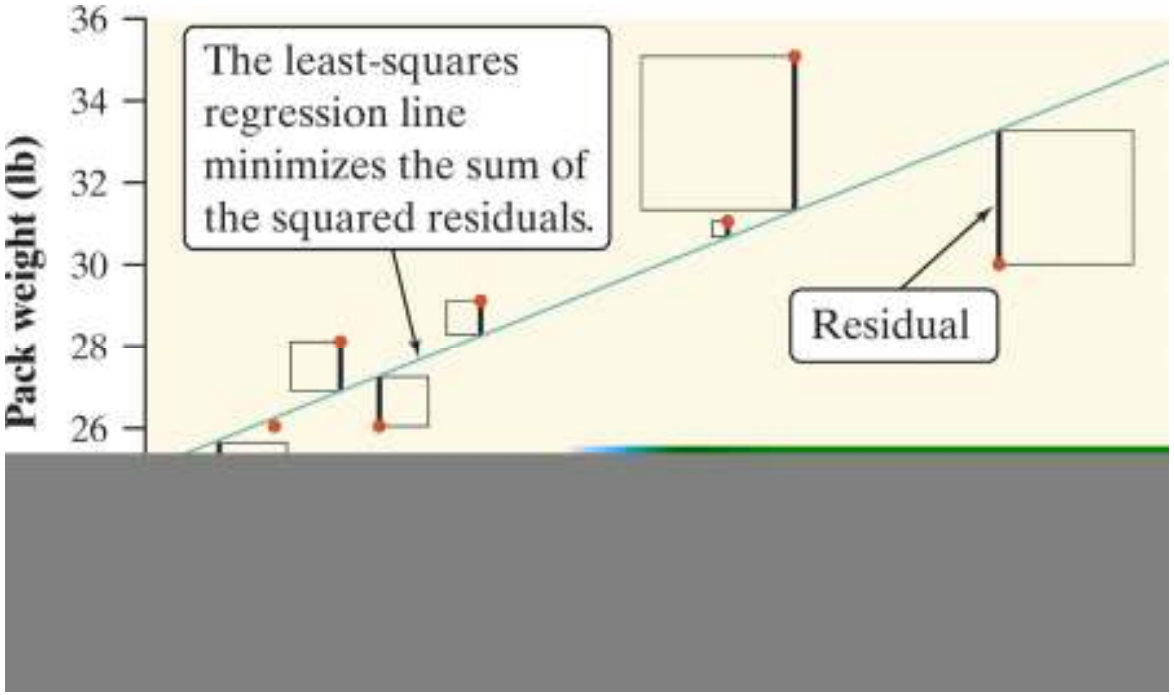
Negative residuals (below line)

Least-Squares Regression Line

Different regression lines produce different residuals. The regression line we want is the one that minimizes the sum of the squared residuals.

Definition:

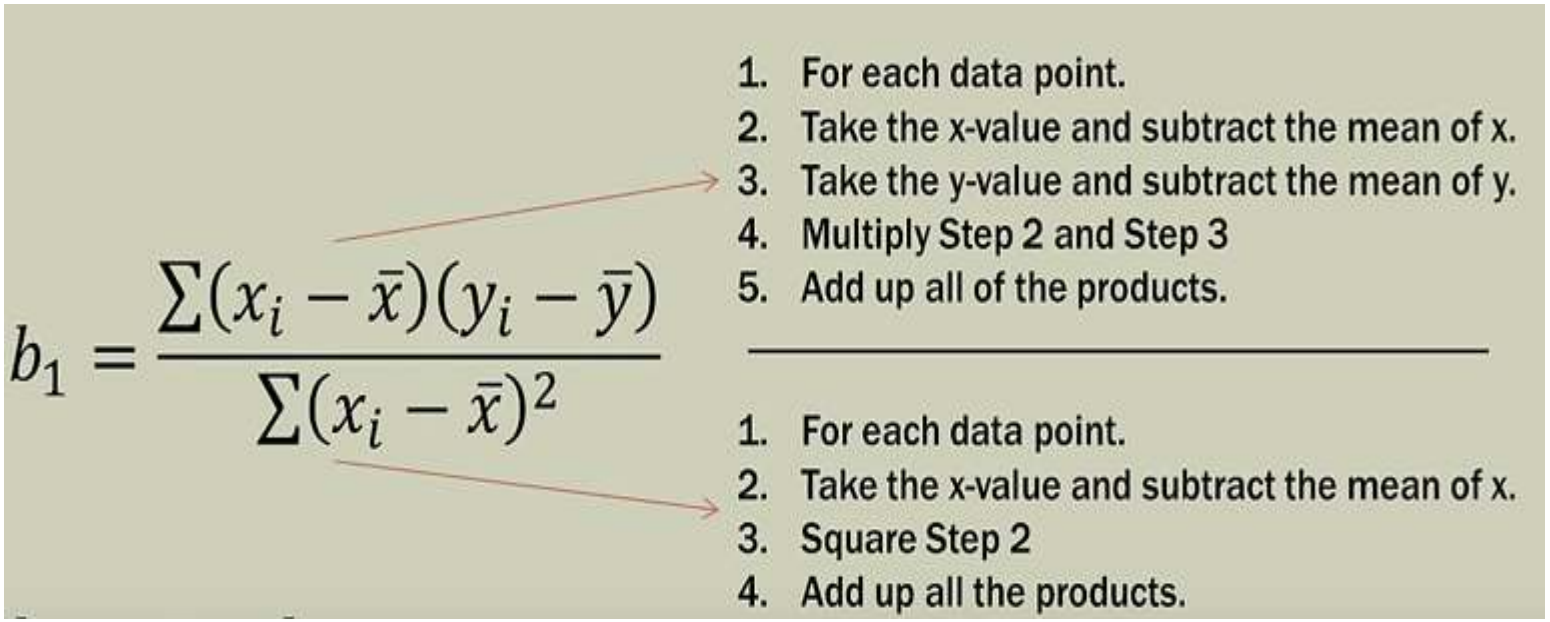
The **least-squares regression line** of *y* on *x* is the line that makes the sum of the squared residuals as small as possible.



Least-Squares Regression Line (LSRL)

- To determine any line, you need to determine the slope and the y-intercept.
- LSRL : $\hat{y} = a + bx$ or $\hat{y} = b_0 + b_1 x$
- Slope: $b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$ or $b_1 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$
- y intercept: $a = \bar{y} - b\bar{x}$ or $b_0 = \bar{y} - b_1\bar{x}$
- **Note**: Centroid $(\bar{x}, \bar{y})$ will always be on the LSRL

How the calculator calculates the slope:



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.

Definition: 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

Another formula for slope

$$b = r\frac{s_y}{s_x}$$

and ***y* intercept**

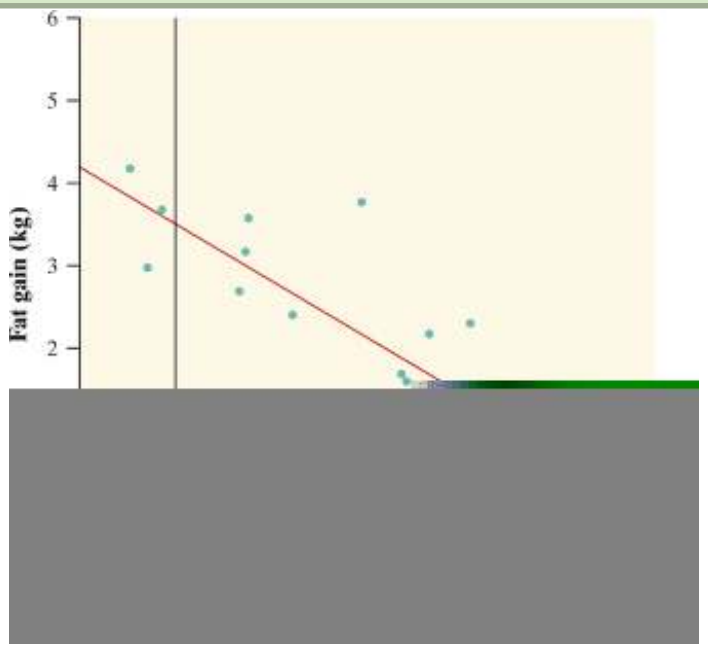
$$a = \bar{y} - b\bar{x}$$

Residual Plots

One of the first principles of data analysis is to look for an overall pattern and for striking departures from the pattern. A regression line describes the overall pattern of a linear relationship between two variables. We look for departures from this pattern by looking at the residuals.

Definition:

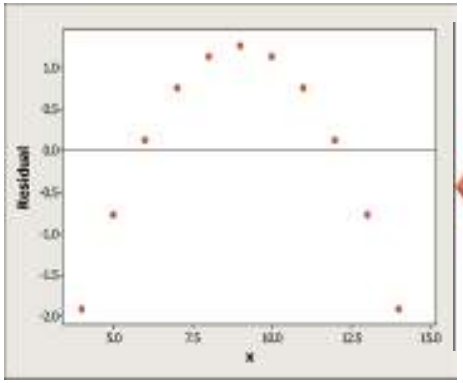
A **residual plot** is a scatterplot of the residuals against the explanatory variable; ordered pairs of (x, residual). Residual plots help us assess how well a regression line fits the data.



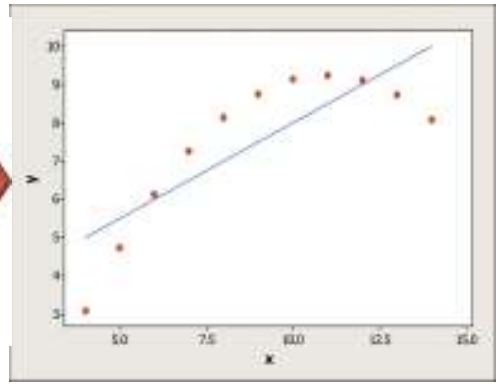
Interpreting Residual Plots

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.



Pattern in residuals
Linear model not appropriate



Definition:

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 about the LSRL (s_e)** is given by

$$s_e = \sqrt{\frac{SSResid}{n-2}} \quad or \quad s = \sqrt{\frac{\sum residuals^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

The Coefficient of determination: r^2

The standard deviation of the residuals gives us a numerical estimate of the average size of our prediction errors. There is another numerical quantity that tells us how well the least-squares regression line predicts values of the response *y*.

Definition:

The **coefficient of determination r^2** is the proportion 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:

$$r^2 = 1 - \frac{SSResid}{SSTo} \quad \text{or} \quad r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

Where *SSResid* or $SSE = \sum \text{residual}^2$

and *SSTo* $SST = \sum (y_i - \bar{y})^2$

The Role of r^2 in Regression:

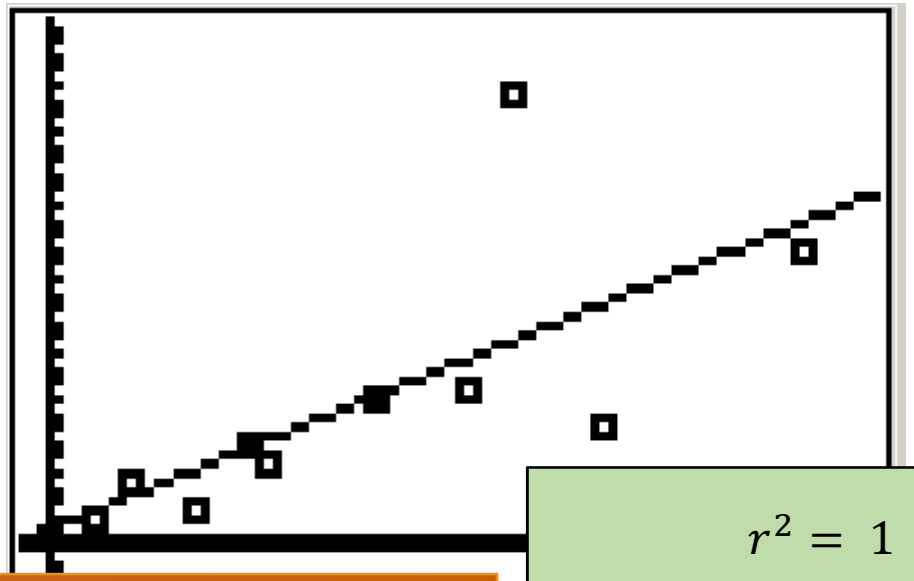
r^2 tells us how much better the LSRL does at predicting values of *y* than simply using the mean *y* for each value in the dataset.

- $r^2 = 1 - \frac{SSResid}{SSTo}$
- Find SSResid: $\sum(y - \hat{y})^2$ $= 359.54$
- Find SSTotal: $\sum(y - \bar{y})^2$ $= 707.39$
- Subtract the ratio ($\frac{SSResid}{SSTo}$) from 1 to obtain r^2 (CoD)

$$r^2 = 1 - \frac{359.54}{707.39} \approx 0.49$$

The Role of r^2 in Regression

r^2 is the proportion of the variation in the values of the *response variable* that is accounted for by the least-squares regression line of ***y* on *x***. Using the example for sales price, If we needed to predict a building price for another property, we could use the average sales price as our prediction.



$$\frac{SSResid}{SSTo} = \frac{SSE}{SST} = \frac{359.54}{707.39} \approx 0.508$$

Therefore, 50.8% of the variation in sales price is *unaccounted for* by the least-squares regression line.

$$r^2 = 1 - \frac{SSResid}{SSTo} \quad or$$

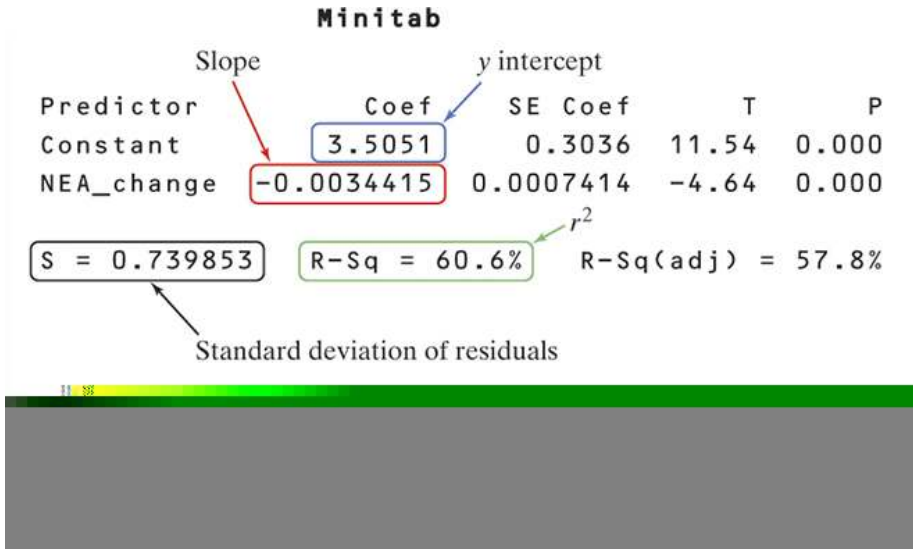
$$r^2 = 1 - \frac{359.54}{707.39} \approx 0.492$$

49.2 % of the variation in sales price is accounted for by the linear model relating sales price to size of building.

Interpreting Computer Regression Output

A number of statistical software packages produce similar regression output. Be sure you can locate

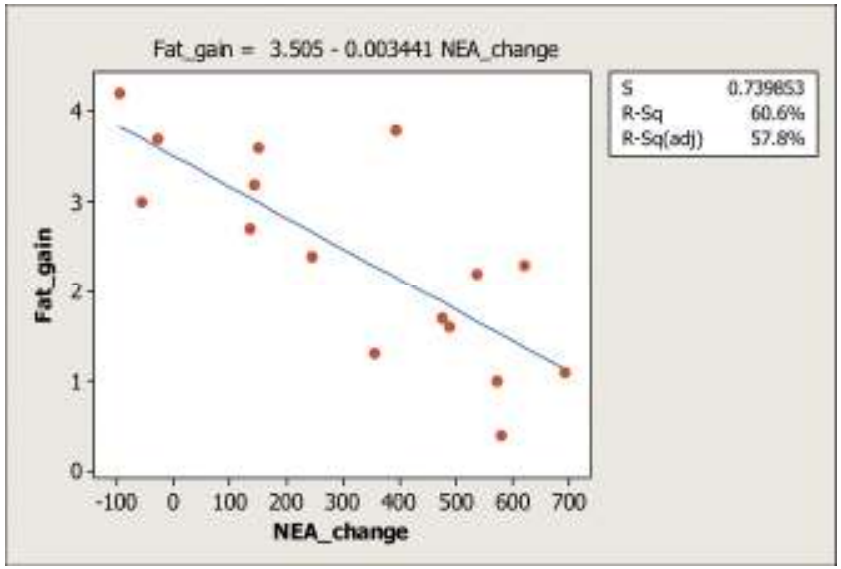
- the slope *b*,
- the *y* intercept *a*,
- and the values of *s* and r^2.

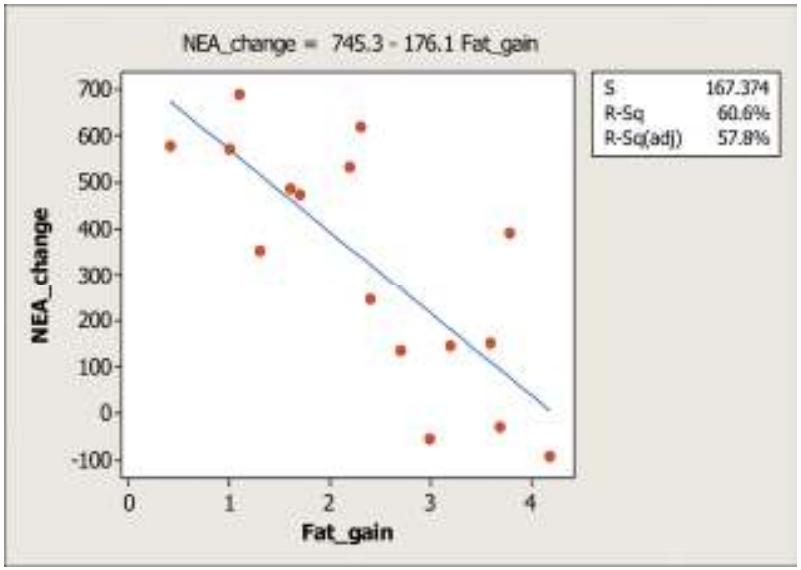


Correlation and Regression Wisdom

Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, be aware of their limitations

1. The distinction between explanatory and response variables is important in regression, because the slope and intercept are dependent on each, but r and r^2 remain the same regardless.





Correlation and Regression Wisdom

2. Correlation and regression lines describe only linear relationships.

3. Correlation and least-squares regression lines are not resistant.

Definition:

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the *y* direction but not the *x* direction of a scatterplot have large residuals. Other outliers may not have large residuals.

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.

Correlation and Regression Wisdom

4. Association does not imply causation.

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?

Section 5.2
Least-Squares Regression

Summary

In this section, we learned that…

- ✓ A **regression line** is a straight line that describes how a response variable *y* changes as an explanatory variable *x* changes. We can use a regression line to **predict** the value of *y* for any value of *x*.
- ✓ The **slope *b*** of a regression line is the rate at which the predicted response $\hat{y}$ changes along the line as the explanatory variable *x* changes. *b* is the *predicted* change in *y* when *x* increases by 1 unit.
- ✓ The ***y* intercept *a*** of a regression line is the predicted response for $\hat{y}$ when the explanatory variable $x = 0$.
- ✓ Avoid **extrapolation**, predicting values outside the range of data from which the line was calculated.

Section 5.3
Least-Squares Regression

Summary

In this section, we learned that…

- ✓ The **least-squares regression line** is the straight line $\hat{y} = a + bx$ that minimizes the sum of the squares of the vertical distances of the observed points from the line.
- ✓ You can examine the fit of a regression line by studying the **residuals** (observed y – predicted y). Be on the lookout for points with unusually large residuals and also for nonlinear patterns and uneven variation in the **residual plot.**
- ✓ The **standard deviation of the residuals *s*** measures the average size of the prediction errors (residuals) when using the regression line.

Section 5.3
Least-Squares Regression

Summary

In this section, we learned that…

- ✓ The **coefficient of determination** r^2 is the fraction of the variation in one variable that is accounted for by least-squares regression on the other variable.
- ✓ Correlation and regression must be interpreted with caution. Plot the data to be sure the relationship is roughly linear and to detect **outliers** and **influential points**.
- ✓ Be careful not to conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated.

Looking Ahead…

In the next Section…

HW Notebook Check will only go to Section 5.3. We will save Sec 5.4 for the next HW check.

NEXT, we'll learn how to complete transformations on non-linear relationships to apply regression techniques.