Lecture Notes & Examples 3.2 Part C and Part D

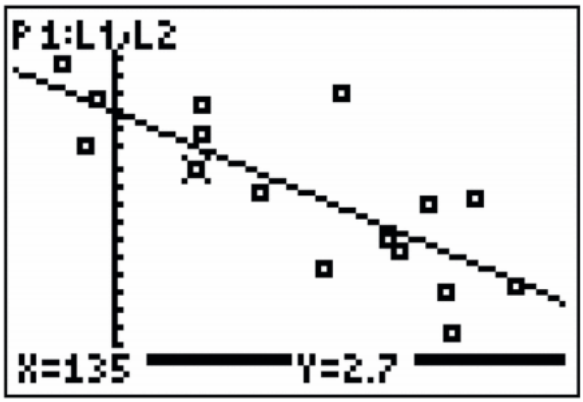
5. How Well the Line Fits the Data: Residual Plots

Because the residuals show how far the data fall from our regression line, examining the residuals helps us assess how well the line describes the data. It should be noted that the *mean of the least-squares residuals is always zero*.

Example – Does Fidgeting Keep You Slim?

Examining Residuals

Let's return to the fat gain and NEA study involving 16 young people who volunteered to overeat for 8 weeks. Those whose NEA rose substantially gained less fat than others. We confirmed that the least-squares regression line for these data is $\widehat{fat\ gain} = 3.505 - 0.00344\ (NEA\ change)$. The calculator screen shot to the right shows a scatterplot of the data with the least-squares line added.



One subject's NEA rose by 135 ca. The subject gained 2.7 kg of fat. (This point is marked in the screen shot with an X.) The predicted fat gain for 135 cal is: $\hat{y} = 3.505 - 0.00344\ (135) = 3.04\ kg.$.

The residual for this subject is therefore:

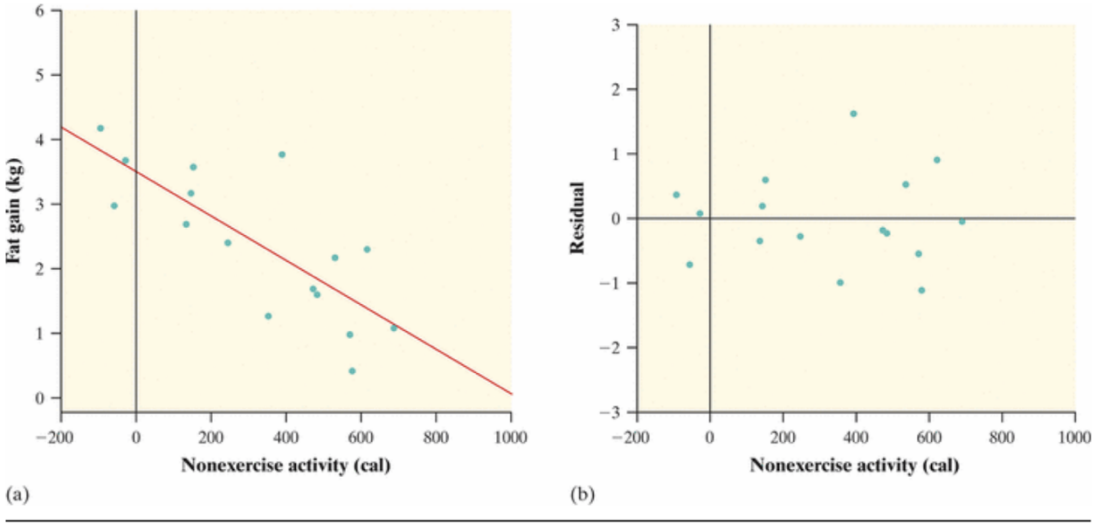
$$\text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y} = 2.7 - 3.04 = -0.34 \text{ kg}$$

This residual is negative because the data point lies below the line. The 16 data points used in calculating the least-squares line produce 16 residuals. Rounded to two decimal places, they are

0.37	−0.70	0.10	−0.34	0.19	0.61	−0.26	−0.98
1.64	−0.18	−0.23	0.54	−0.54	−1.11	0.93	−0.03

Because the residuals show how far the data fall from our regression line, examining the residuals helps assess how well the line describes the data. Although residuals can be calculated from any model that is fitted to the data, the residuals from the least-squares line have a special property: the mean of the least-squares residuals is always zero. You can check that the sum of the residuals in the above example is 0.01. The sum is not exactly 0 because we rounded to two decimal places.

You can see the residuals in the scatterplot of **(a)** by looking at the vertical deviations of the points from the line. The **residual plot** in **(b)** makes it easier to study the residuals by plotting them against the explanatory variable, change in NEA. Because the mean of the residuals is always zero, the horizontal line at zero in **(b)** helps orient us. This "residual = 0" line corresponds to the regression line in **(a)**.



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

CHECK YOUR UNDERSTANDING

Refer to the data below:

NEA change (cal):	−94	−57	−29	135	143	151	245	355
Fat gain (kg):	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA change (cal):	392	473	486	535	571	580	620	690
Fat gain (kg):	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

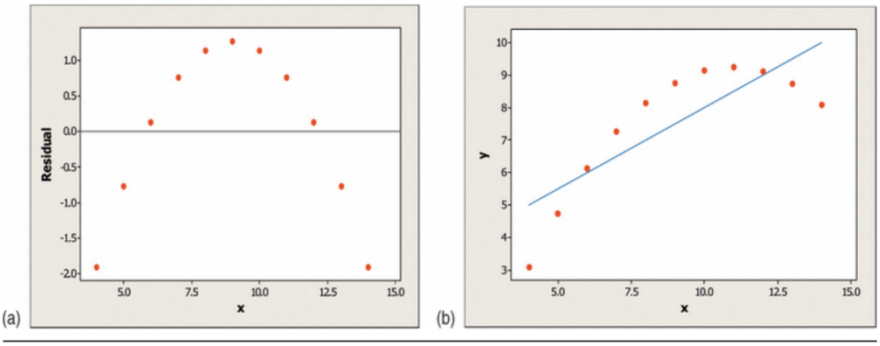
1. Find the residual for the subject who increased NEA by 620 calories. Show your work.

2. Interpret the value of this subject's residual in context.

3. For which subject did the regression line overpredict fat gain by the most? Justify your answer.

Examining Residual Plots

A residual plot in effect turns the regression line horizontal. It magnifies the deviations of the points from the line, making it easier to see unusual observations and patterns. If the regression line captures the overall pattern of the data, there should be no pattern in the residuals. **Figure (a)** shows a residual plot with a clear curved pattern. A straight line is not an appropriate model for these data, as **Figure (b)** confirms.



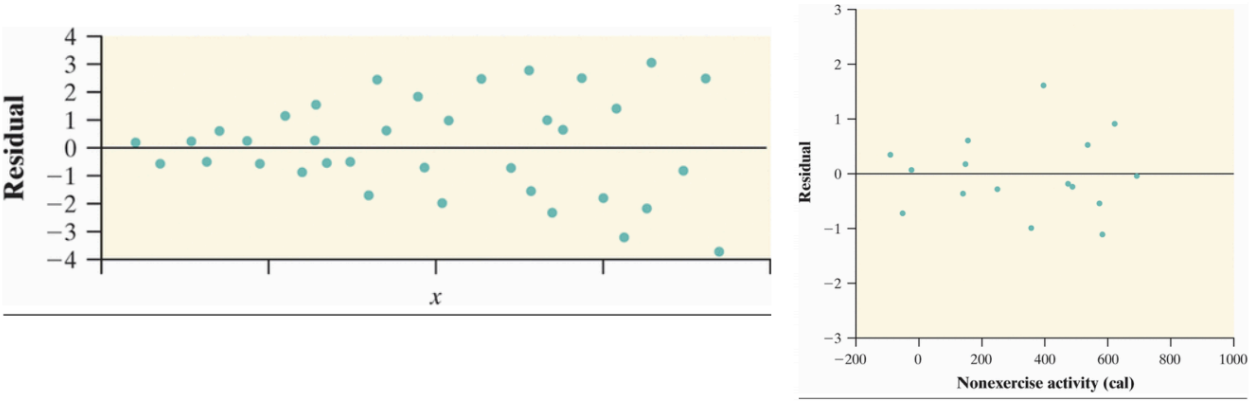
Here are two important things to look for when you examine a residual plot.

1. The residual plot should show *no obvious pattern*.

- A curved pattern shows that the relationship is not linear.
- A pattern that gets increasing larger says that the regression line will not be accurate for larger values of x.

2. The residuals should be *relatively small in size*.

To decide what "small" means, consider the size of the typical error with respect to the data points.



In the figure to the right above, for example, most of the residuals are between −0.7 and 0.7. For these individuals, the predicted fat gain from the least-squares line is within 0.7 kilogram (kg) of their actual fat gain during the study. That sounds pretty good. But the subjects gained only between 0.4 and 4.2 kg, so a prediction error of 0.7 kg is relatively large compared with the actual fat gain for an individual. The largest residual, 1.64, corresponds to a prediction error of 1.64 kg. This subject's actual fat gain was 3.8 kg, but the regression line predicted a fat gain of only 2.16 kg. That's a pretty large error, especially from the subject's perspective!

Standard deviation of the residuals We have already seen that the average prediction error (that is, the mean of the residuals) is 0 whenever we use a least-squares regression line. That's because the positive and negative residuals "balance out." But that doesn't tell us how far off the predictions are, on average. Instead, we use the standard deviation of the residuals:

$$s = \sqrt{\frac{\sum \text{residuals}^2}{n-2}}$$

For the NEA and fat gain data, the sum of the squared residuals is 7.663. So the standard deviation of the residuals is:

$$s = \sqrt{\frac{7.663}{14}} = 0.740 \text{ kg}$$

Standard Deviation of the Residuals (s) - To find out how far off the predictions are using the residuals, we can compute the Standard Deviation of the Residuals:

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

This value gives us the **approximate size of a "typical" or "average" predicted error (residual).**

Interpret the standard deviation For the NEA and fat gain data.

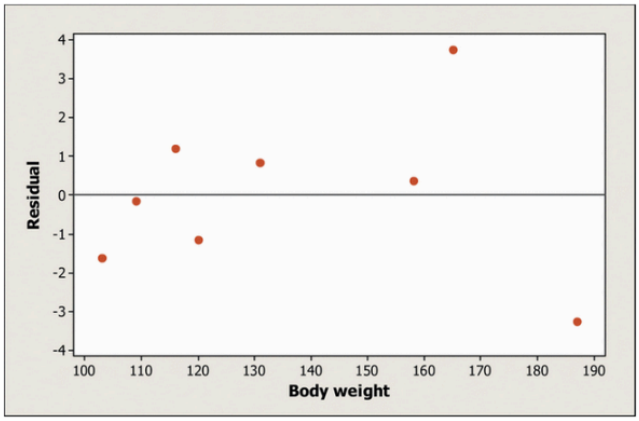
$$s = \sqrt{\frac{7.663}{14}} = 0.740 \text{ kg}$$

- The average error(residual) in prediction fat-gain is 0.740 kg using the least-squares regression line (LLSR)

Technology - Using the calculator to graph residuals is covered on p. 178 of the text. To find the standard deviation of the residuals, divide the sum of the squared residuals by n-2 and take the square root.

CHECK YOUR UNDERSTANDING

The graph shown is a residual plot for the least-squares regression of pack weight on body weight for the 8 hikers.



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

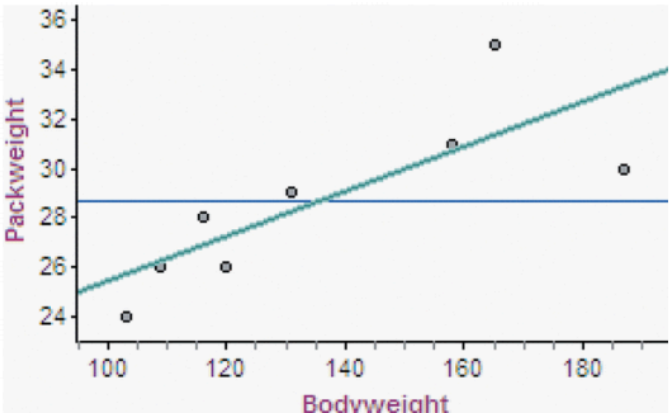
1. The residual plot does not show a random scatter. Describe the pattern you see.

2. For this regression, s = 2.27. Interpret this value in context.

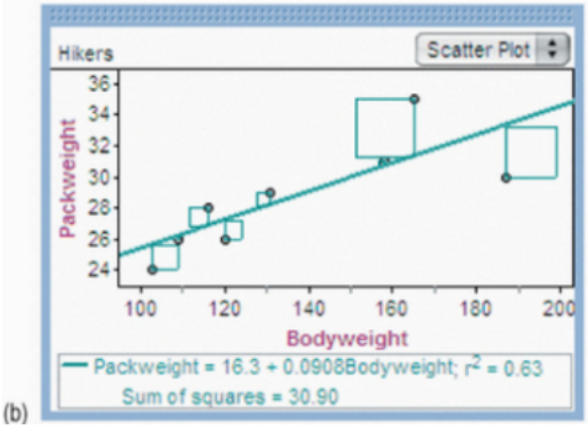
3.2.6 How Well the Line Fits the Data: The Role of r^2 in Regression

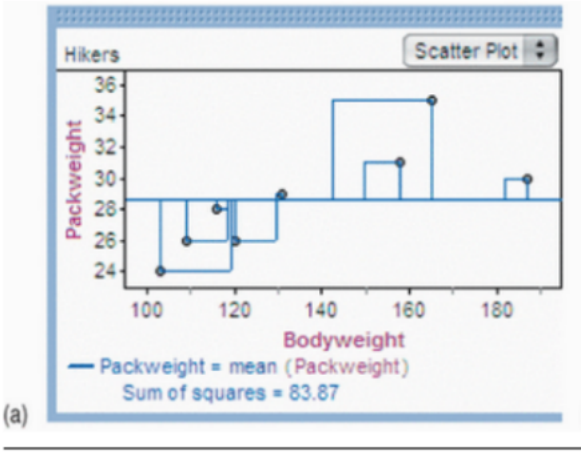
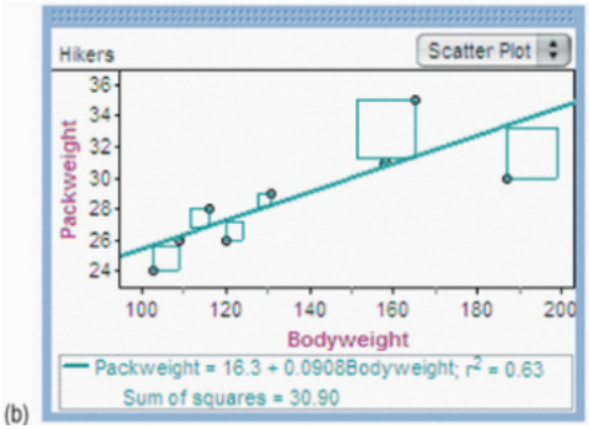
Coefficient of Determination. A residual plot is a graphical tool for evaluating how well a regression line fits the data. The standard deviation of the residuals, s, gives us a numerical estimate of the average size of our prediction errors from the regression line**. There is another numerical quantity that tells us how well the least-squares line predicts values of the response variable y. It is r^2, the coefficient of determination**. Some computer packages call it "R-sq." You may have noticed this value in some of the calculator and computer regression output that we showed earlier. Although it's true that r^2 is equal to the square of r, there is much more to this story.

Example – Pack weight and body weight
How can we predict y if we don't know x?



Suppose a new student is assigned at the last minute to our group of 8 hikers. What would we predict for his pack weight? The figure above shows a scatterplot of the hiker data that we have studied throughout this chapter. The least-squares line is drawn on the plot in green. Another line has been added in blue: a horizontal line at the mean y-value, $\bar{y}$ = 28.62. If we don't know this new student's body weight, then we can't use the regression line to make a prediction. What should we do? Our best strategy is to use the mean pack weight of the other 8 hikers as our prediction.

	
The first scatterplot shows the mean price line. The sum of the squared prediction errors when using the mean price y-bar is called the **total sum of squares (SST)**.	The second scatterplot shows the least-squares regression line. (LSRL) The sum of the squared prediction errors when using the least-squares regression line is called the **sum of squared errors (SSE).**

	
The first scatterplot shows the mean price line. The sum of the squared prediction errors when using the mean price y-bar is called the **total sum of squares (SST)**.	The second scatterplot shows the least-squares regression line. (LSRL) The sum of the squared prediction errors when using the least-squares regression line is called the **sum of squared errors (SSE).**

The figure above **(a)** shows the prediction errors if we use the average pack weight $\bar{y}$ as our prediction for the original group of 8 hikers. We can see that the sum of the squared residuals for this line is SST = $\sum(y_i - \bar{y})^2$ =83.87. SST measures the total variation in the y-values.

If we learn our new hiker's body weight, then we could use the least-squares line to predict his pack weight. How much better does the regression line do at predicting pack weights than simply using the average pack weight y of all 8 hikers? Figure **(b)** reminds us that the sum of squared residuals for the least-squares line is Σ residual^2 = 30.90. We'll call this SSE, for sum of squared errors. The ratio SSE/SST tells us what proportion of the total variation in y still remains after using the regression line to predict the values of the response variable. In this case,

$$\frac{SSE}{SST} = \frac{30.90}{83.87} = 0.368$$

This means that 36.8% of the variation in pack weight is unaccounted for by the least-squares regression line. Taking this one step further, the proportion of the total variation in y that is accounted for by the regression line is

$$1 - \frac{SSE}{SST} = 1 - 0.368 = 0.632$$

We interpret this by saying that "63.2% of the variation in backpack weight is accounted for by the linear model relating pack weight to body weight." For this reason, we define

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

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

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

Where $SSE = \sum residual^2$ and $SST = \sum(y_i - \bar{y})^2$

It seems pretty remarkable that the coefficient of determination is actually the correlation squared. This fact provides an important connection between correlation and regression. When you report a regression, give r^2 as a measure of how successful the regression was in explaining the response. When you see a correlation, square it to get a better feel for the strength of the linear relationship

AP EXAM TIP Students often have a hard time interpreting the value of r^2 on AP exam questions. They frequently leave out key words in the definition.

Our advice: Treat this as a fill-in-the-blank exercise. Write

"_____ % of the variation in the [response variable name] is accounted for by the regression line."

****

CHECK YOUR UNDERSTANDING

1. For the least-squares regression of fat gain on NEA, $r^2 = 0.606$. Which of the following gives a correct interpretation of this value in context?

(a) 60.6% of the points lie on the least-squares regression line.
(b) 60.6% of the fat gain values are accounted for by the least-squares line.
(c) 60.6% of the variation in fat gain is accounted for by the least-squares line.
(d) 77.8% of the variation in fat gain is accounted for by the least-squares line.

2. A recent study discovered that the correlation between the age at which an infant first speaks and the child's score on an IQ test given upon entering elementary school is −0.68. A scatterplot of the data shows a linear form. Which of the following statements about this finding is correct?

(a) Infants who speak at very early ages will have higher IQ scores by the beginning of elementary school than those who begin to speak later.

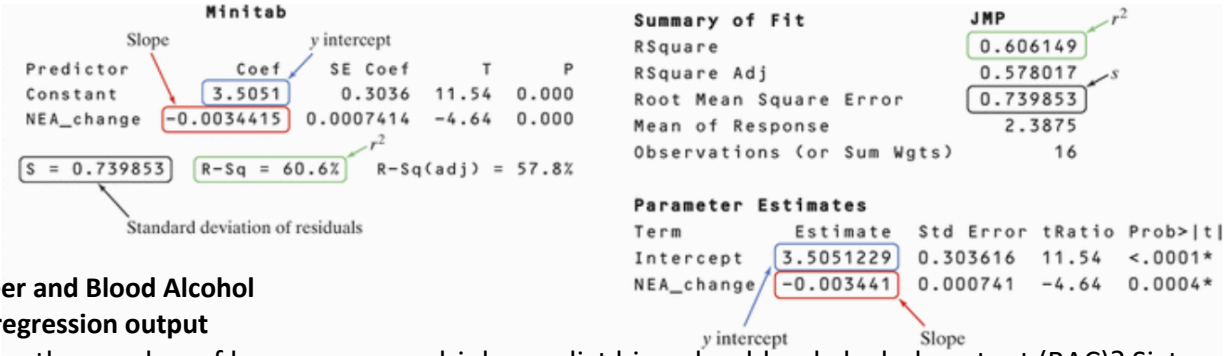
(b) 68% of the variation in IQ test scores is explained by the least-squares regression of age at first spoken word and IQ score.

(c) Encouraging infants to speak before they are ready can have a detrimental effect later in life, as evidenced by their lower IQ scores.

d) There is a moderately strong, negative linear relationship between age at first spoken word and later IQ test score for the individuals in this study.

3.2.7. Interpreting Computer Output

When looking at computer output, always look for *slope, y-intercept,* and the values of *s* and r^2.



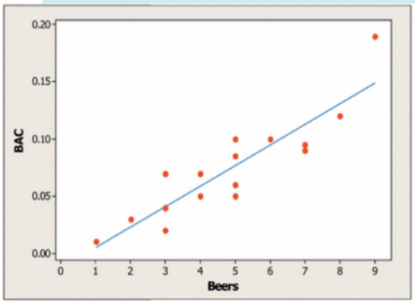
Example – Beer and Blood Alcohol

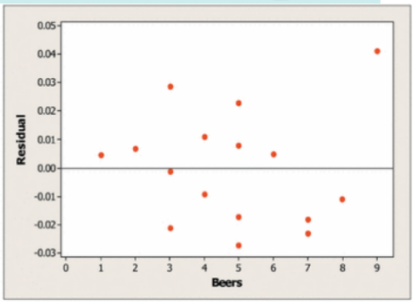
Interpreting regression output

How well does the number of beers a person drinks predict his or her blood alcohol content (BAC)? Sixteen volunteers with an initial BAC of 0 drank a randomly assigned number of cans of beer. Thirty minutes later, a police officer measured their BAC. Least-squares regression was performed on the data. A scatterplot with the regression line added, a residual plot, and some computer output from the regression are shown below.

Dependent variable is: **BAC**
No Selector
R squared = 80%
R squared (adjusted) = 78.6%
s = 0.0204 with
16 - 2 = 14 degrees of freedom

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-0.012701	0.0126	-1.00	0.3320
Beers	0.017964	0.0024	7.48	≤0.0001





(a) What is the equation of the least-squares regression line that describes the relationship between beers consumed and blood alcohol content? Define any variables you use.

(b) Interpret the slope of the regression line in context.

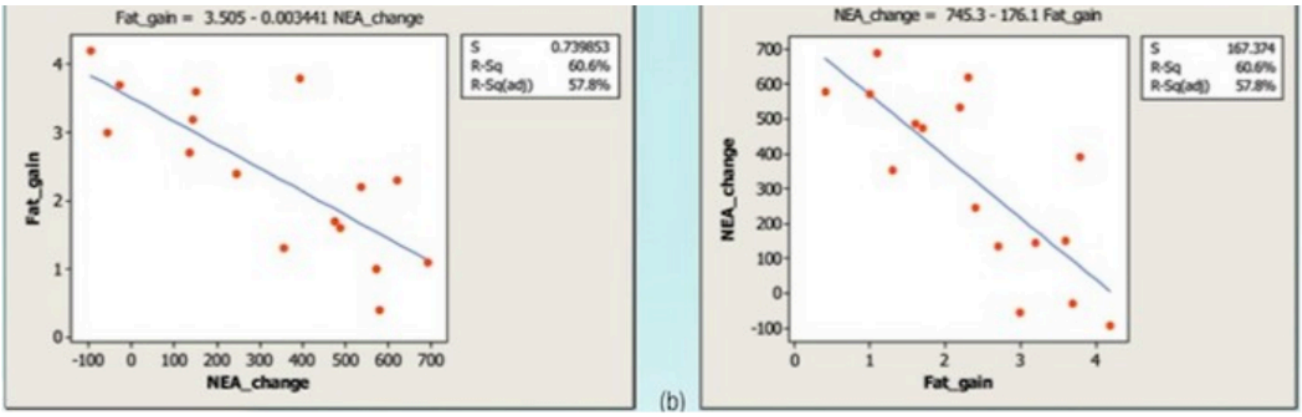
(c) Find the correlation.

(d) Is a line an appropriate model to use for these data? What information tells you this?

(e) What was the BAC reading for the person who consumed 9 beers? Show your work.

3.2.8 Correlation and Regression Wisdom

- *The distinction between explanatory and response variables is important in regression.*
    - This is not true for correlation. Switching x and y will not affect the value of r.
    - Switching x and y will give a different regression line.



The two regression lines are very different. However, no matter which variable we put on the x axis, $r^2 = 0.606$ and the correlation is $r = -0.778$.

- *Correlation and regression lines describe only linear relationships.*
    - You can calculate correlation and the regression line for any relationship between quantitative variables, but the results are only useful if the scatterplot shows a linear relationship.
    - **ALWAYS PLOT YOUR DATA!**
- *Correlation and least-squares regression lines are not resistant.*
    - One unusual point can change the correlation, r.
    - Least-squares regression makes the sum of the squares of the vertical distances to the points from the line as small as possible. A point that is extreme in the x direction with no other points near it pulls the line toward itself. This type of point is called ***influential***.

> **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.

- *Association does not imply causation*.
    - A strong association between two variables is not enough to draw conclusions about cause and effect.
    - We will learn how to establish causation in Chapter 4.

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