Section 3.1 – Scatterplots and Correlation (pp. 141-164)

Most statistical studies examine data on more than one variable. We will continue to use tools we have already learned as well as adding others to assist us in analysis.

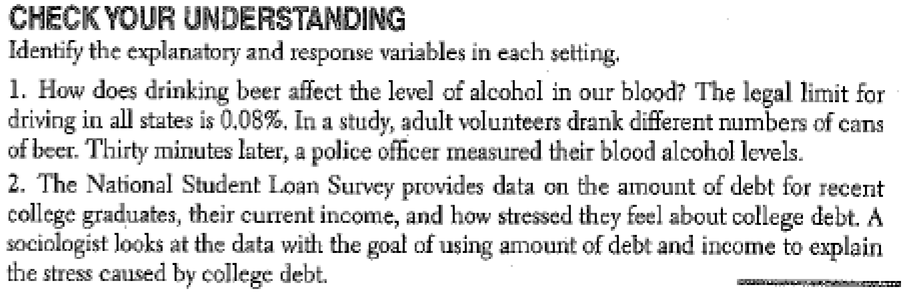
- Plot the data, add numerical summaries
- Look for overall patterns and deviations from those patterns
- If there is a regular pattern, use a simplified model to describe it

1. Explanatory and Response Variables

Definition: A **response variable** measures the outcome of a study. An **explanatory variable** *may* help explain or influence changes in a response variable.

This means that the *explanatory variable* "accounts for" or "predicts" changes in the response variable.

Examples:



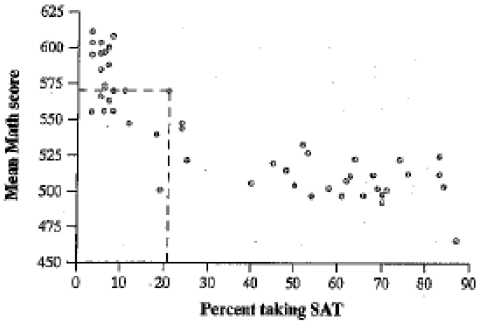
CHECK YOUR UNDERSTANDING

Identify the explanatory and response variables in each setting.

1. How does drinking beer affect the level of alcohol in our blood? The legal limit for driving in all states is 0.08%. In a study, adult volunteers drank different numbers of cans of beer. Thirty minutes later, a police officer measured their blood alcohol levels.

2. The National Student Loan Survey provides data on the amount of debt for recent college graduates, their current income, and how stressed they feel about college debt. A sociologist looks at the data with the goal of using amount of debt and income to explain the stress caused by college debt.

2. Displaying Relationships: Scatterplots

<table>
<tr><td colspan="2"><b>Definition:</b> A <b>scatterplot</b> shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis and the values of the other variable appear on the vertical axis. Each individual in the data set appears as a point on the graph.</td></tr>
<tr><td>
</td><td>If there is an explanatory variable, it is plotted on the x-axis and the response variable is on the y-axis.

If there is no explanatory-response distinction, either variable can go on the x-axis.</td></tr>
</table>

How to make a scatterplot:
1. Decide which variable should go on which axis.
2. Label and scale your axes
3. Plot individual data values

(Common error on AP Exam – failing to label axes.)

Calculator:

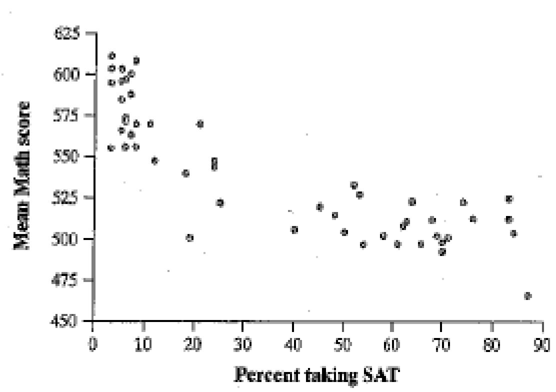
3. Interpreting Scatterplots

How to examine a scatterplot

Look for *overall pattern* and for striking *departures* from that pattern

- Overall pattern is described by the **direction**, **form**, and **strength** of the relationship.
- An important type of departure is an **outlier**, an individual pattern that falls outside the overall pattern of the relationship.

DOFS



D:

F:

S:

O:

Definition:

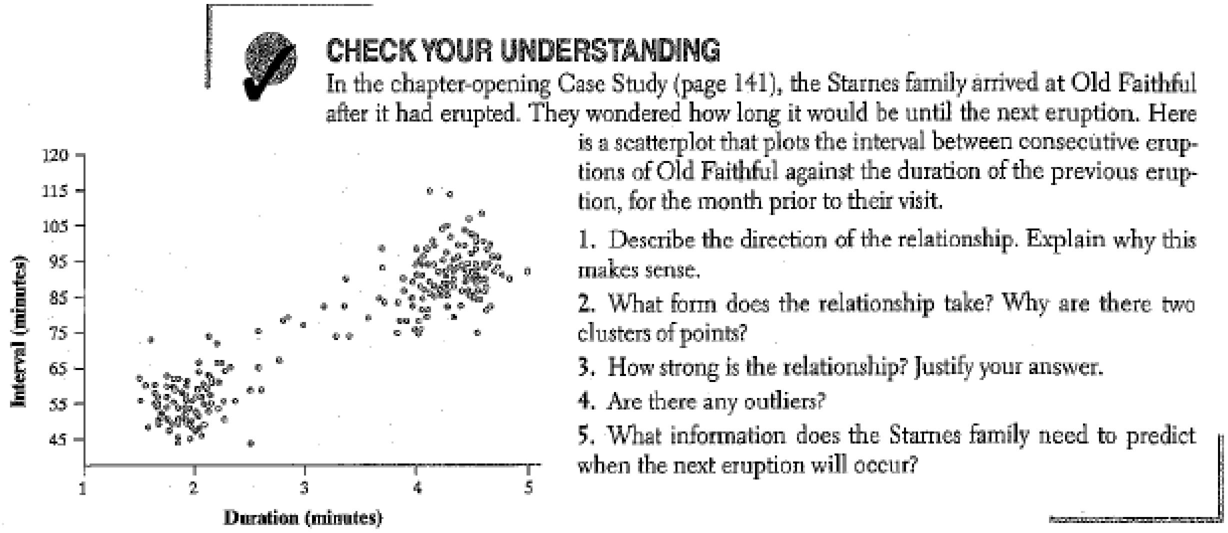
Two variables have a **positive association** when the above average values of one tend to accompany above average values of the other and when below average values also tend to occur together.

Two variables have a **negative association** when the above average values of one tend to accompany below average values of the other.

******Causation and Association******

Association does not imply causation!!!!!

Examples:



CHECK YOUR UNDERSTANDING

In the chapter-opening Case Study (page 141), the Starnes family arrived at Old Faithful after it had erupted. They wondered how long it would be until the next eruption. Here is a scatterplot that plots the interval between consecutive eruptions of Old Faithful against the duration of the previous eruption, for the month prior to their visit.

1. Describe the direction of the relationship. Explain why this makes sense.
2. What form does the relationship take? Why are there two clusters of points?
3. How strong is the relationship? Justify your answer.
4. Are there any outliers?
5. What information does the Starnes family need to predict when the next eruption will occur?

4. Measuring Linear Association: Correlation

A linear relationship may appear in a scatterplot. The linear relationship is strong if the points lie close to a straight line and weak if they are widely scattered about a line. We are going to use a statistic called **correlation** to measure linearity in a scatterplot. **Correlation *r*** measures the *direction* and *strength* of the linear relationship between two quantitative variables.

The correlation r is always a number between -1 and 1. The sign indicates the direction of the association. Values close to 0 indicate a weak linear relationship. As r approaches -1 or 1, the strength of the relationship increases. -1 and 1 only occur if the values lie *exactly* on a straight line.

Team work: The following data give the weight in pounds and cost in dollars of a sample of 11 stand mixers.

Wt	23	28	19	17	25	26	21	32	16	17	8
Price	180	250	300	150	300	370	400	350	200	150	30

1. Scatterplot your data and sketch the scatterplot below. Be sure to scale and label it properly.

2. Calculate the correlation.

3. The last mixer in the table is from Walmart. What happens to the correlation when you remove this point?

4. What happens to the correlation if the Walmart mixer weighs 25 pounds instead of 8 pounds? Add the point (25, 30) and recalculate the correlation.

5. Suppose a new titanium mixer was introduced that weighed 8 points, but the cost was $500. Remove the point (25, 30) and add the point (8, 500). Recalculate the correlation.

6. Summarize what you learned about the effect of a single point on the correlation.

How to calculate r

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

What does this mean?

Note: A value of r close to 1 or -1 *does not guarantee a linear relationship between two variables*. A scatterplot with a clear curved form can have a correlation that is near -1 or 1. **Always plot your data!**

5. Facts about Correlation

1. Correlation makes no distinction between explanatory and response variables.

2. Because r uses the standardized values of the observations, r does not change when we change the units of measurement of x, y, or both.

3. The correlation r itself has no unit of measurement.

4. Correlation requires that both variables be quantitative.

5. Correlation measures the strength of only the linear relationship between tow variables. It does not describe curved relationships between variables.

6. The correlation is not *resistant*: it is strongly affected by a few outlying observations.

7. Correlation is not a complete summary of two-variable data. You should always give means and standard deviations of both x and y along with the correlation.

Section 3.2 - Least-Squares Regression

In the previous section we examined scatterplots for linear relationships. Correlation measures the direction and strength of these relationships. When the plot shows a linear relationship, we would like to summarize the overall pattern by drawing a line on the scatterplot. This is called a **Regression Line**. In order to do this we must have an *explanatory* and a *response variable*.

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.

1. Interpreting a Regression Line – A regression line is a *model* for the data, much like the density curves we considered in Chapter 2. It gives a compact mathematical description of the relationship between the response variable y and the explanatory variable x.

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

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 is the slope (rate of change), the amount by which y is *predicted* to change when x increases by one unit.

a is the y-intercept, the *predicted* value of y when x = 0.

Note: on the AP Exam formula sheet the regression equation is written $\hat{y} = b_0 + b_1 x$. (Regardless of notation, the coefficient with x is the slope.)

Example: Suppose there is a strong negative relationship between the miles driven and the advertised price of used Ford F-150 SuperCrew 4 x 4 trucks and that the regression equation is $\hat{y} = 38257 - 0.1629x$ where $\hat{y}$ is in dollars and x is in thousands of miles driven. Identify the slope and y-intercept of the regression line and interpret each in context.

2. Prediction – The regression line can be used to predict the response variable $\hat{y}$ for a specific value of the explanatory variable x.

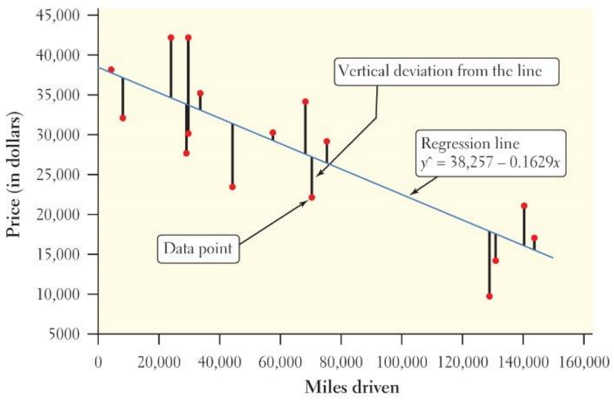
Example: Predict the price of a used F-150 with 100,000 miles on it.

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**.

3. Residuals and the Least-Squares Regression Line – In most cases, no line will pass exactly through all the points in a scatterplot. The predicted values (y-hat) will not be the actual values of the response variable y. *A good regression line makes the vertical distance between the actual points from the line as small as possible.*

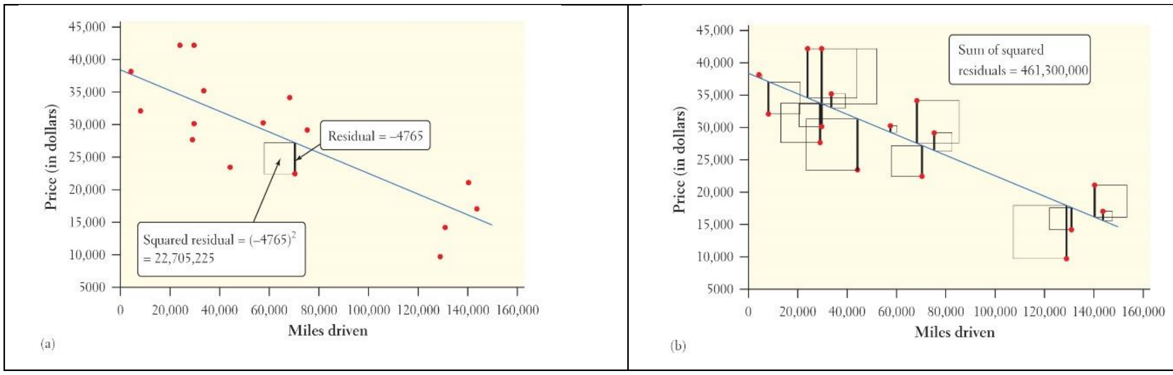
Definition: A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is

Residual = Observed y – Predicted y = y - $\hat{y}$



Example: Find and interpret the residual for the F-150 that had 70,583 miles driven and a price of $21,994.

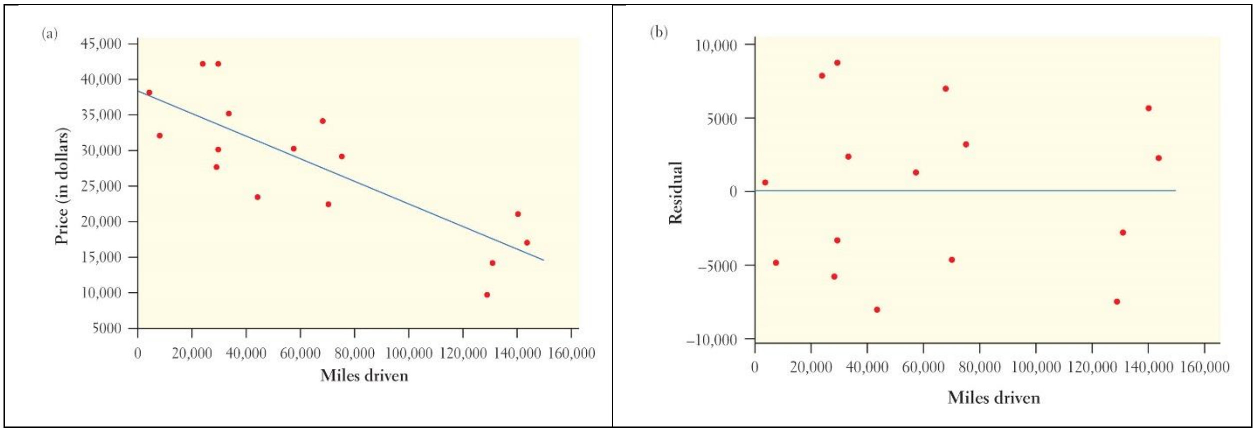
In order to create the **Least-Squares regression line** we will choose the line that makes the *sum of the squared residuals as small as possible*.



Technology – The least-squares regression line can be found by using your graphing calculator. Details are listed on p. 171 of the text as well as p. 32 of NTA.

4. 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*.

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.



Examining Residual Plots

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.

Technology - Using the calculator to graph residuals is covered on p. 178 of the text and p. 33 of NTA. To find the standard deviation of the residuals, divide the sum of the squared residuals by n-2 and take the square root.

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

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:

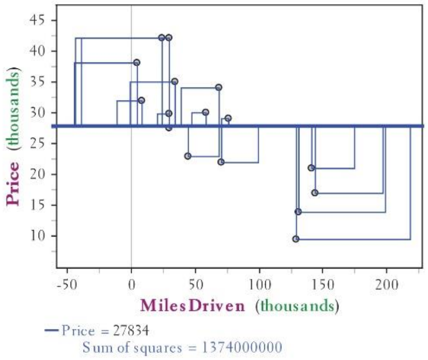
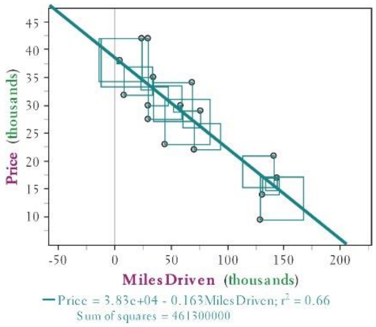
$$\boldsymbol{s = \sqrt{\frac{\sum residuals^2}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y})^2}{n-2}}}$$

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

Example. For the used F-150s, the standard deviation of the residuals is $5740. So when we use the number of miles to estimate the price, we will be off by an average of $5740.

Coefficient of Determination. While 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 regression line predicts values of the response variable, y. It is called the **coefficient of determination, r^2**.

Suppose we want to estimate the advertised price of a used F-150 from CarMax but do not know the number of miles. The mean price of the other used F-150s would be a reasonable guess.

<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>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 <b>total sum of squares (SST)</b>.</td>
<td>The second scatterplot shows the least-squares regression line.

The sum of the squared prediction errors when using the least-squares regression line is called the <b>sum of squared errors (SSE).</b></td>
</tr>
</table>

We can use the SST and SSE to find the variation in asking price that is *unaccounted* for by the least-squares regression line.

Using this we can find the variability in advertised price that *is* accounted for by the least-squares regression line.

Definition: The **coefficient of determination, r²** 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 **r²** using

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

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

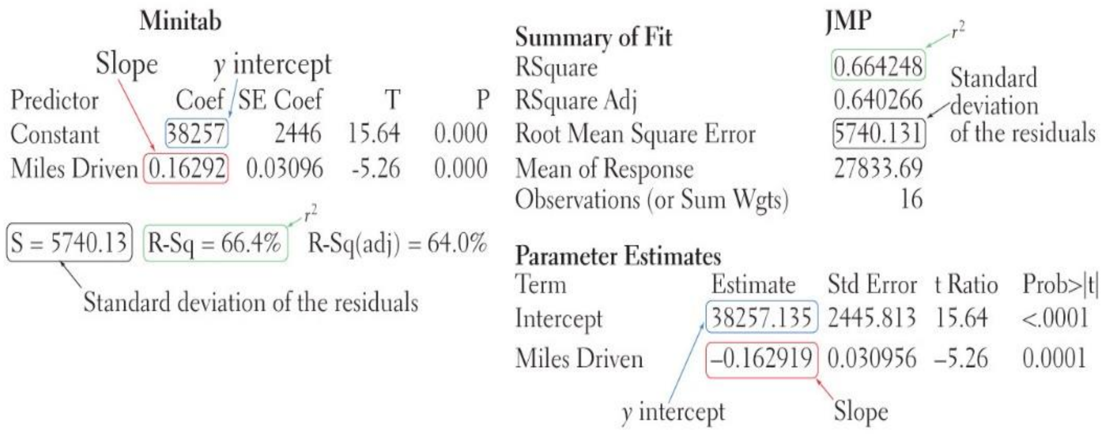
****

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

****

---

6. Interpreting Computer Output - When looking at computer output, always look for *slope, y-intercept,* and the values of *s* and *r²*.



Minitab

```
Predictor      Coef   SE Coef      T       P
Constant      38257      2446  15.64   0.000
Miles Driven 0.16292  0.03096  -5.26   0.000

S = 5740.13   R-Sq = 66.4%   R-Sq(adj) = 64.0%
```

JMP

```
Summary of Fit
RSquare                          0.664248
RSquare Adj                      0.640266
Root Mean Square Error           5740.131
Mean of Response                 27833.69
Observations (or Sum Wgts)             16

Parameter Estimates
Term           Estimate    Std Error  t Ratio  Prob>|t|
Intercept      38257.135   2445.813   15.64    <.0001
Miles Driven   -0.162919   0.030956   -5.26    0.0001
```

Slope
y intercept
r²
Standard deviation of the residuals
r²
Standard deviation of the residuals
y intercept
Slope

7. Regression to the Mean

It is possible to calculate the equation of the least-squares regression line using only the means and standard deviations of the two variables and their correlations.

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

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

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

9. 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.
- *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.