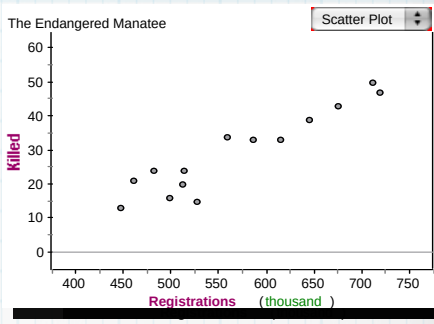
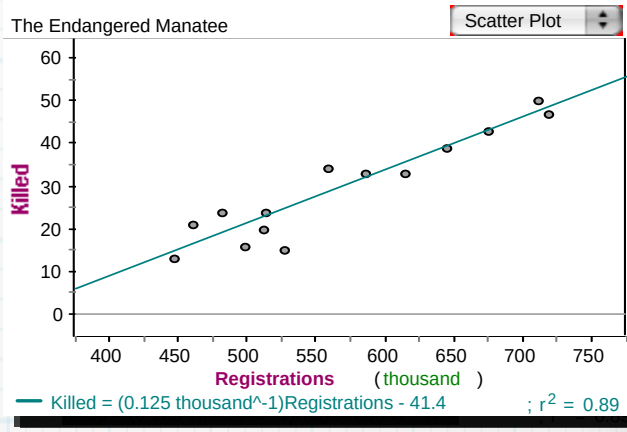
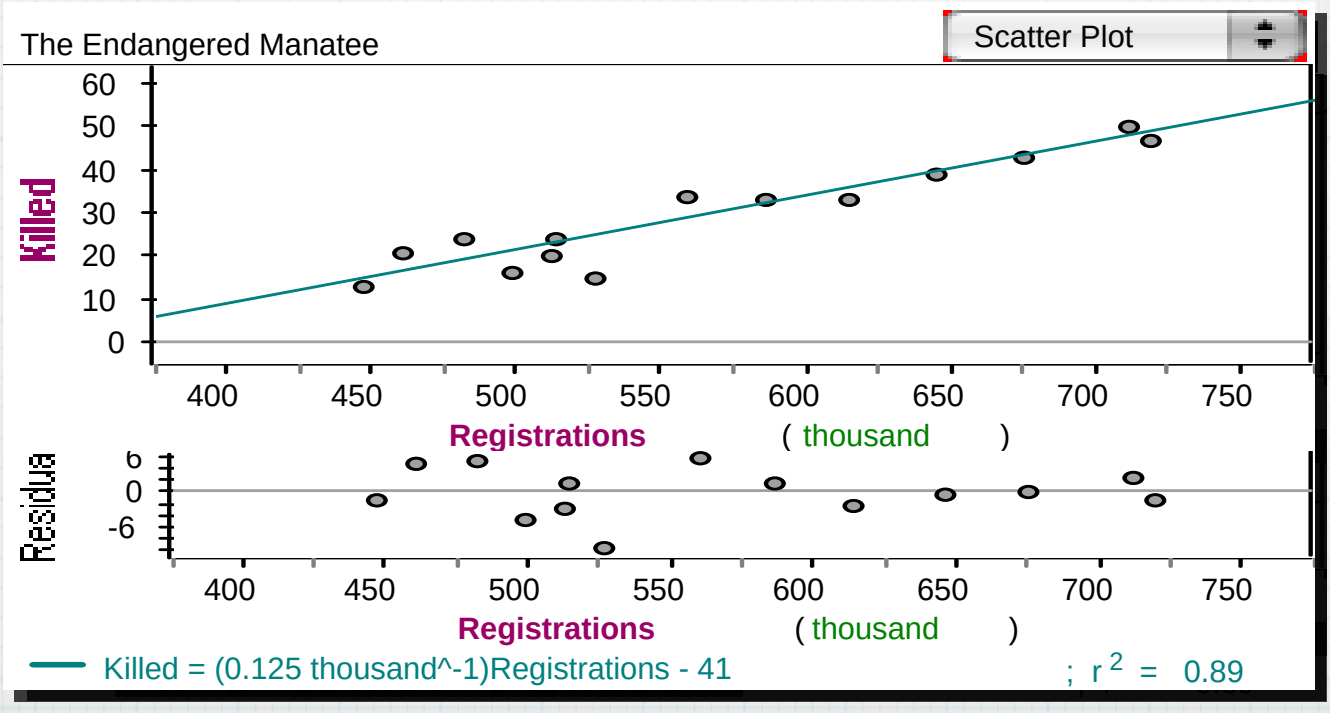
Review of Regression Basics

- When describing a Bivariate Relationship:
  - Make a Scatterplot
    - Strength, Direction, Form
  - Model: y-hat=a+bx
    - Interpret slope in context
  - Make Predictions
    - Residual = Observed-Predicted
  - Assess the Model
    - Interpret “r”
    - Residual Plot







Reading Minitab Output

Regression Analysis: Fat gain versus NEA

The regression equation is

FatGain = **** + ******(NEA)**

Predictor	Coef	SE Coef	T	P
Constant	3.5051	0.3036	11.54	0.000
NEA	-0.0034415	0.000[illegible]14		

S=0.739853 **R-Sq = 60.6%**

The Intercept is also known as a "constant"

a=3.5051

The Slope and Intercept are "coefficients" in the LSRL.

The Slope is the coefficient of the explanatory variable.

b=-0.0034415

Regression equations aren't always as easy to spot as they are on your TI-84. Can you find the slope and intercept above?

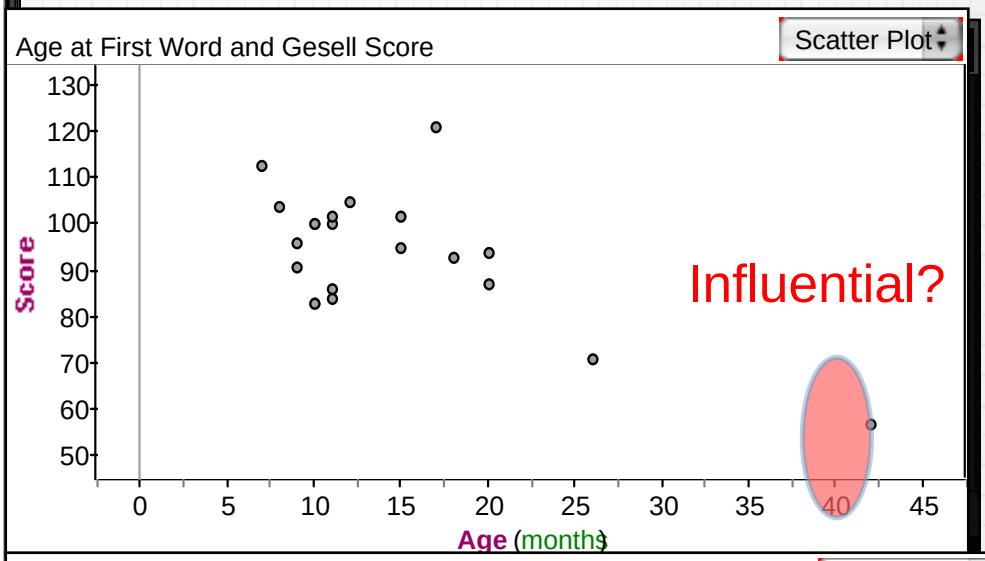
FatGain = 3.5051 - 0.0034415(NEA)

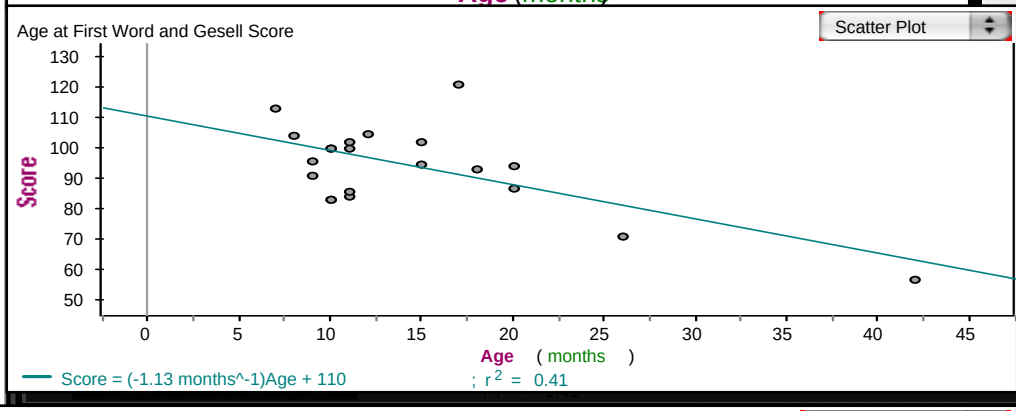
Outliers/Influential Points

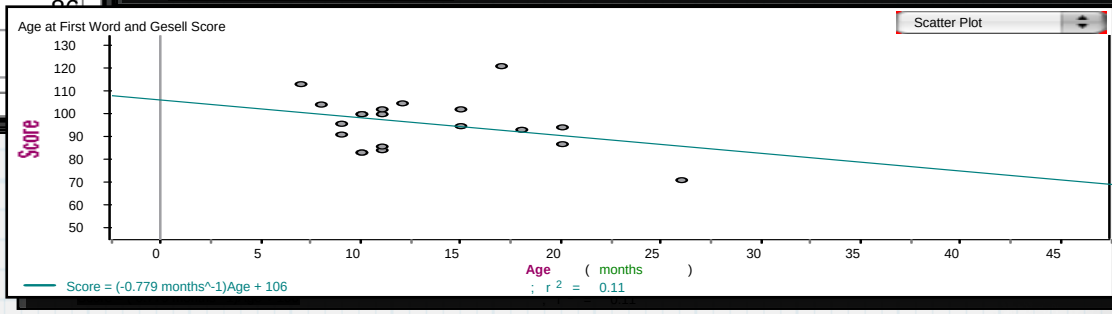
Does the age of a child's first word predict his/her mental ability? Consider the following data on (age of first word, Gesell Adaptive Score) for 21 children.

Age at First Word and Gesell Score

	Child	Age	Score
1	1	15 months	95
2	2	26 months	71
3	3	10 months	83
4	4	9 months	91
5	5	15 months	102
6	6	20 months	87
7	7	18 months	93
8	8	11 months	100
9	9	8 months	104
10	10	20 months	94
11	11	7 months	113
12	12	9 months	96
13	13	10 months	83
14	14	11 months	84
15	15	11 months	102
16	16	10 months	100
17	17	12 months	105
18	18	42 months	57
19	19	17 months	121
20	20	11 months	96
21	21	10 months	94





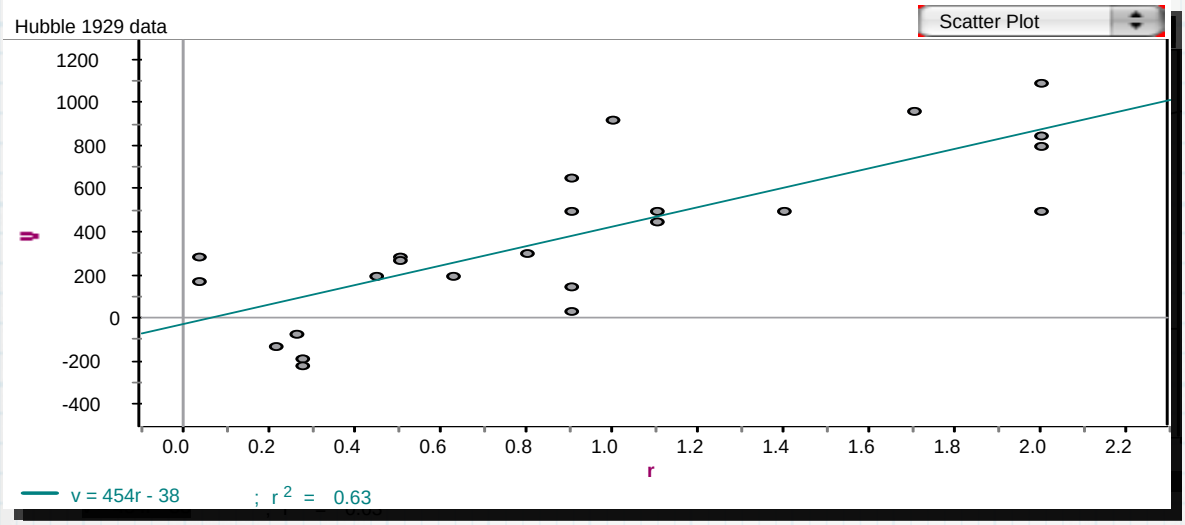


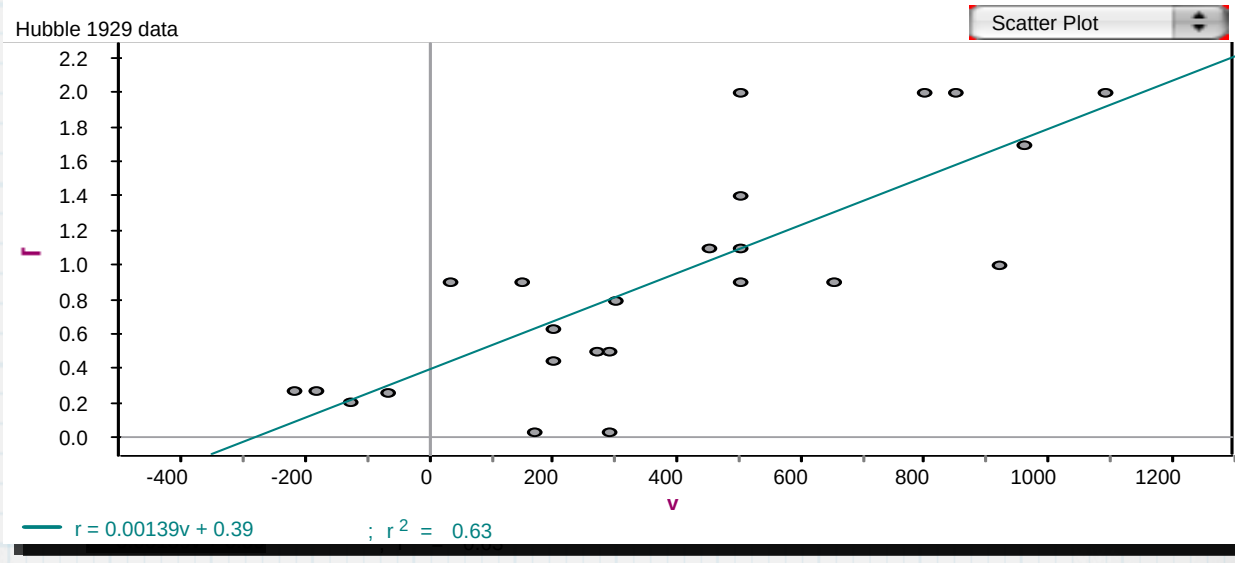
Does the highlighted point markedly affect the equation of the LSRL? If so, it is "influential".

Test by removing the point and finding the new LSRL.

Explanatory vs. Response

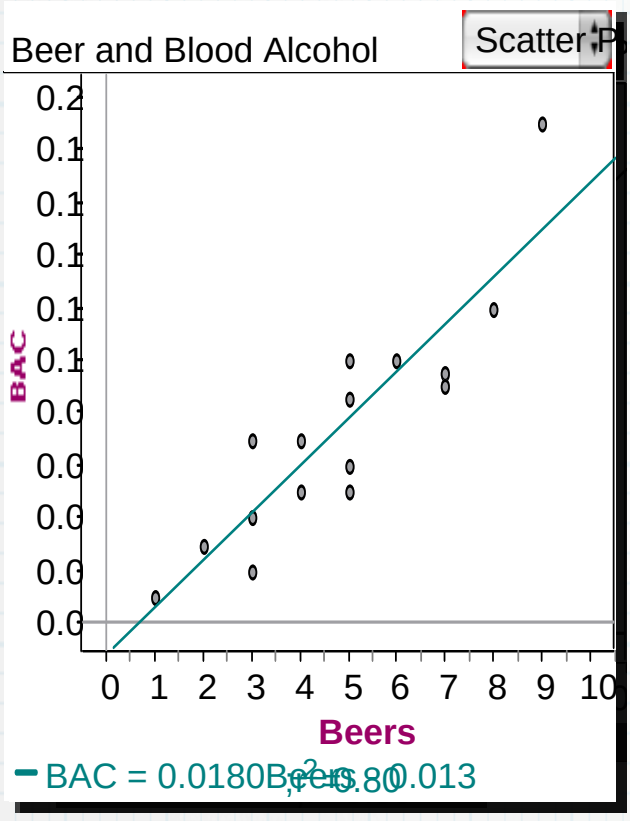
- **The Distinction Between Explanatory and Response variables is essential in regression.**
  - **Switching the distinction results in a different least-squares regression line.**





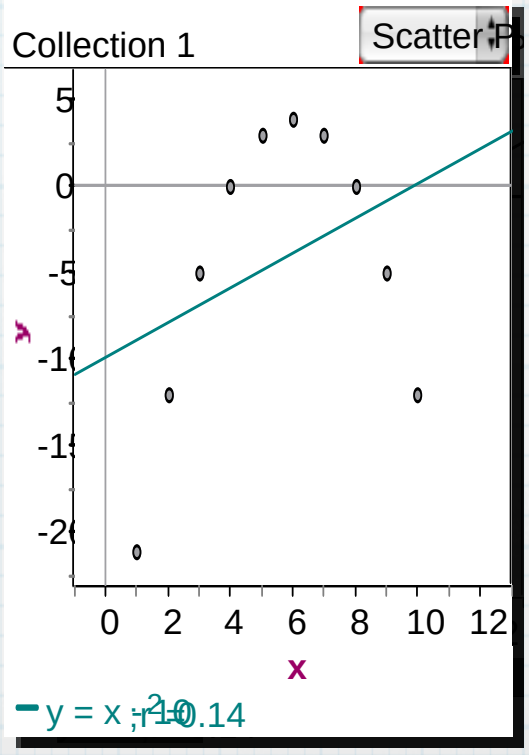
- Note: The correlation value, r, does NOT depend on the distinction between Explanatory and Response.

Correlation



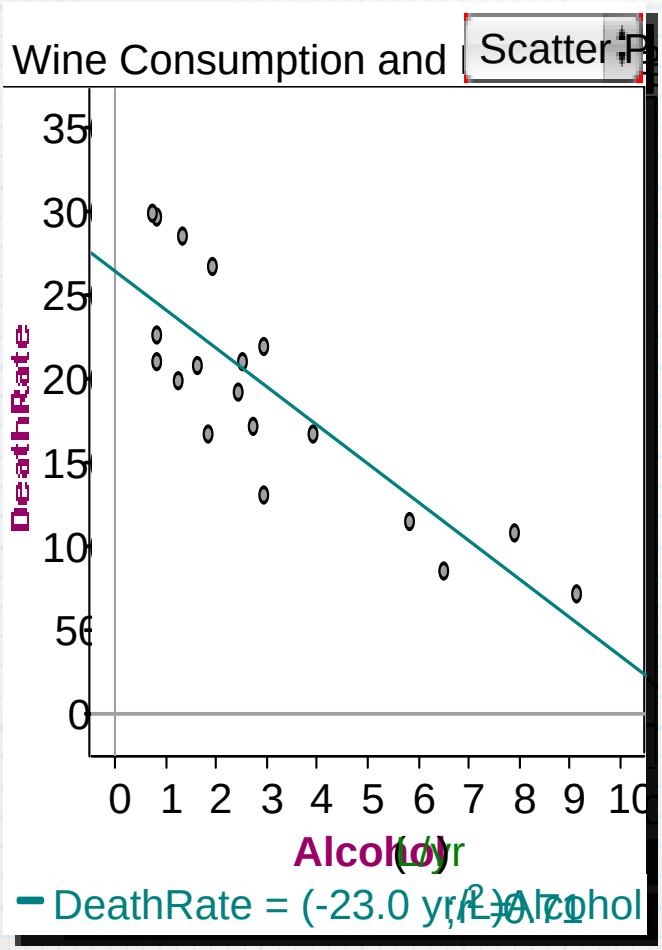
- The correlation, r, describes the strength of the straight-line relationship between x and y.
  - Ex: There is a strong, positive, LINEAR relationship between # of beers and BAC.

- There is a weak, positive, linear relationship between x and y. However, there is a strong nonlinear relationship.
- r measures the strength of linearity...



Coefficient of Determination

- The coefficient of determination, r^2, describes the percent of variability in y that is explained by the linear regression on x.



- 71% of the variability in death rates due to heart disease can be explained by the LSRL on alcohol consumption.
- That is, alcohol consumption provides us with a fairly good prediction of death rate due to heart disease, but other factors contribute to this rate, so our prediction will be off somewhat.

Cautions

- Correlation and Regression are NOT RESISTANT to outliers and Influential Points!

- Correlations based on “averaged data” tend to be higher than correlations based on all raw data.

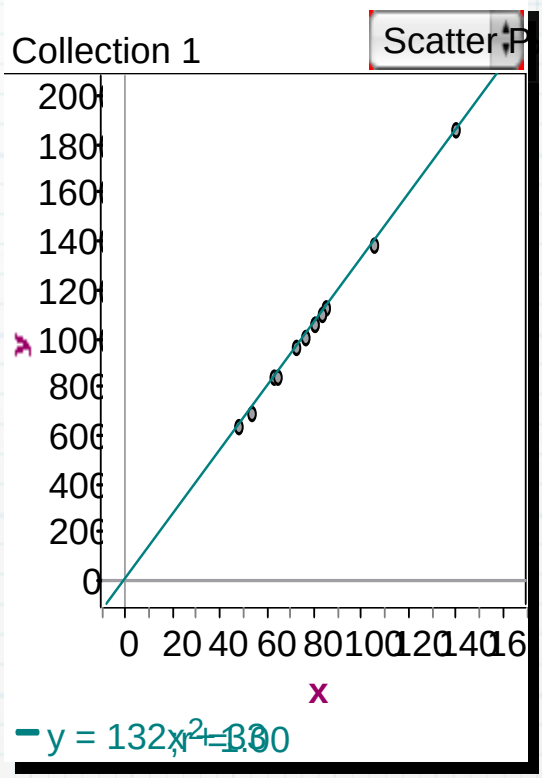
- Extrapolating beyond the observed data can result in predictions that are unreliable.

Correlation vs. Causation

- Consider the following historical data:

Collection 1

	Year		
1	1860	63	837
2	1865	48	640
3	1870	53	700
4	1875	64	848
5	1880	72	959
6	1885	80	106
7	1890	85	112
8	1895	76	100
9	1900	80	105
10	1905	83	110
11	1910	105	138
12	1915	140	185



- There is an almost perfect linear relationship between x and y. (r=0.999997)

- x = # Methodist Ministers in New England
- y = # of Barrels of Rum Imported to Boston
- CORRELATION DOES NOT IMPLY CAUSATION!

Summary

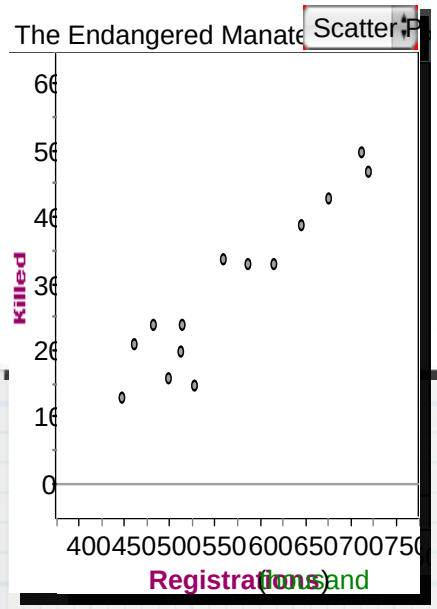
Plot your data.
Scatterplot

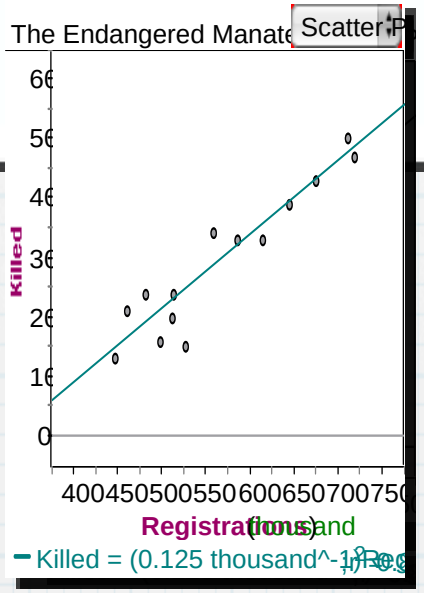
Interpret what you see:
direction, form, strength, outliers

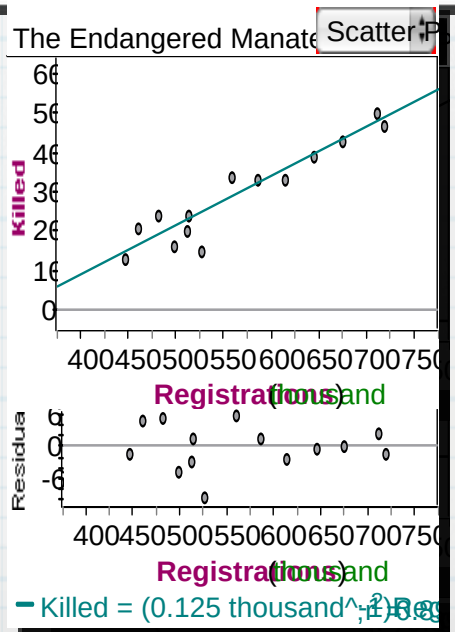
Numerical summary?
$\bar{x}, \bar{y}, s_x, s_y$, and r?

Mathematical model?
Regression line?

How well does it fit?
Residuals and r^2



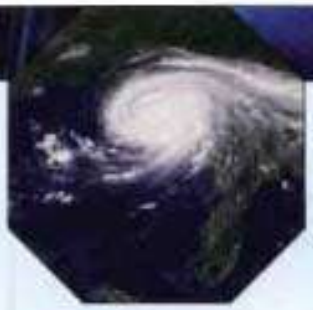




CONNECTIONS

Scatterplots are the basic tool for examining the relationship between two quantitative variables. We start with a picture when we want to understand the distribution of a single variable, and we always make a scatterplot to begin to understand the relationship between two quantitative variables.

We used z-scores as a way to measure the statistical distance of data values from their means. Now we've seen the z-scores of x and y working together to build the correlation coefficient. Correlation is a summary statistic like the mean and standard deviation—only it summarizes the strength of a linear relationship. And we interpret it as we did z-scores, using the standard deviations as our rulers in both x and y.



WHAT HAVE WE LEARNED?

In recent chapters we learned how to listen to the story told by data from a single variable. Now we've turned our attention to the more complicated (and more interesting) story we can discover in the association between two quantitative variables.

We've learned to begin our investigation by looking at a scatterplot. We're interested in the *direction* of the association, the *form* it takes, and its *strength*.

We've learned that, although not every relationship is linear, when the scatterplot is straight enough, the *correlation coefficient* is a useful numerical summary.

- The sign of the correlation tells us the direction of the association.
- The magnitude of the correlation tells us the *strength* of a linear association. Strong associations have correlations near −1 or +1 and very weak associations near 0.
- Correlation has no units, so shifting or scaling the data, standardizing, or even swapping the variables has no effect on the numerical value.

Once again we've learned that doing Statistics right means we have to *Think* about whether our choice of methods is appropriate.

- The correlation coefficient is appropriate only if the underlying relationship is linear.
- We'll check the **Straight Enough Condition** by looking at a scatterplot.
- And, as always, we'll watch out for outliers!

Finally, we've learned not to make the mistake of assuming that a high correlation or strong association is evidence of a cause-and-effect relationship. Beware of lurking variables!

A S ***Simulation:* Correlation, Center, and Scale.** If you have any lingering doubts that shifting and rescaling the data won't change the correlation, watch nothing happen right before your eyes!

Terms

Scatterplots

147. A scatterplot shows the relationship between two quantitative variables measured on the same cases.

Association

- 147. **Direction:** A positive direction or association means that, in general, as one variable increases, so does the other. When increases in one variable generally correspond to decreases in the other, the association is negative.
- 147. **Form:** The form we care about most is straight, but you should certainly describe other patterns you see in scatterplots.
- 148. **Strength:** A scatterplot is said to show a strong association if there is little scatter around the underlying relationship.

Outlier

148. A point that does not fit the overall pattern seen in the scatterplot.

Response variable, Explanatory variable, *x*-variable, *y*-variable

149. In a scatterplot, you must choose a role for each variable. Assign to the *y*-axis the response variable that you hope to predict or explain. Assign to the *x*-axis the explanatory or predictor variable that accounts for, explains, predicts, or is otherwise responsible for the *y*-variable.

Correlation Coefficient

152. The correlation coefficient is a numerical measure of the direction and strength of a linear association.

$$r = \frac{\sum z_x z_y}{n-1}$$

Lurking variable

157. A variable other than *x* and *y* that simultaneously affects both variables, accounting for the correlation between the two.

Skills

THINK

- Recognize when interest in the pattern of a possible relationship between two quantitative variables suggests making a scatterplot.
- Know how to identify the roles of the variables and that you should place the response variable on the *y*-axis and the explanatory variable on the *x*-axis.
- Know the conditions for correlation and how to check them.
- Know that correlations are between −1 and +1, and that each extreme indicates a perfect linear association.
- Understand how the magnitude of the correlation reflects the strength of a linear association as viewed in a scatterplot.
- Know that correlation has no units.
- Know that the correlation coefficient is not changed by changing the center or scale of either variable.
- Understand that causation cannot be demonstrated by a scatterplot or correlation.

SHOW

- Know how to make a scatterplot by hand (for a small set of data) or with technology.
- Know how to compute the correlation of two variables.
- Know how to read a correlation table produced by a statistics program.

TELL

- Be able to describe the direction, form, and strength of a scatterplot.
- Be prepared to identify and describe points that deviate from the overall pattern.
- Be able to use correlation as part of the description of a scatterplot.
- Be alert to misinterpretations of correlation.
- Understand that finding a correlation between two variables does not indicate a causal relationship between them. Beware the dangers of suggesting causal relationships when describing correlations.

SCATTERPLOTS AND CORRELATION ON THE COMPUTER

Statistics packages generally make it easy to look at a scatterplot to check whether the correlation is appropriate. Some packages make this easier than others.

Many packages allow you to modify or enhance a scatterplot, altering the axis labels, the axis numbering, the plot symbols, or the colors used. Some options, such as color and symbol choice, can be used to display additional information on the scatterplot.

WHAT HAVE WE LEARNED?

We've learned that when the relationship between quantitative variables is fairly straight, a linear model can help summarize that relationship and give us insights about it:

- The regression (best fit) line doesn't pass through all the points, but it is the best compromise in the sense that the sum of squares of the residuals is the smallest possible.

We've learned several things the correlation, r, tells us about the regression:

- The slope of the line is based on the correlation, adjusted for the units of x and y:

$$b_1 = \frac{r s_y}{s_x}$$

We've learned to interpret that slope in context:

- For each SD of x that we are away from the x mean, we expect to be r SDs of y away from the y mean.
- Because r is always between -1 and $+1$, each predicted y is fewer SDs away from its mean than the corresponding x was, a phenomenon called regression to the mean.
- The square of the correlation coefficient, R^2, gives us the fraction of the variation of the response accounted for by the regression model. The remaining $1 - R^2$ of the variation is left in the residuals.

The residuals also reveal how well the model works:

- If a plot of residuals against predicted values shows a pattern, we should re-examine the data to see why.
- The standard deviation of the residuals, s_e, quantifies the amount of scatter around the line.

Of course, the linear model makes no sense unless the **Linearity Assumption** is satisfied. We check the **Straight Enough Condition** and **Outlier Condition** with a scatterplot, as we did for correlation, and also with a plot of residuals against either the x or the predicted values. For the standard deviation of the residuals to make sense as a summary, we have to make the **Equal Variance Assumption**. We check it by looking at both the original scatterplot and the residual plot for the **Does the Plot Thicken? Condition.**

Terms

Model 172. An equation or formula that simplifies and represents reality.

Linear model 172. A linear model is an equation of a line. To interpret a linear model, we need to know the variables (along with their W's) and their units.

Predicted value 172. The value of $\hat{y}$ found for a given x-value in the data. A predicted value is found by substituting the x-value in the regression equation. The predicted values are the values on the fitted line; the points $(x, \hat{y})$ all lie exactly on the fitted line.

Residuals 172. Residuals are the differences between data values and the corresponding values predicted by the regression model—or, more generally, values predicted by any model.

$$\text{Residual} = \text{observed value} - \text{predicted value} = e = y - \hat{y}$$

Least squares 172. The least squares criterion specifies the unique line that minimizes the variance of the residuals or, equivalently, the sum of the squared residuals.

Regression to the mean 174. Because the correlation is always less than 1.0 in magnitude, each predicted $\hat{y}$ tends to be fewer standard deviations from its mean than its corresponding x was from its mean. This is called regression to the mean.

Regression line
Line of best fit 174. The particular linear equation

$$\hat{y} = b_0 + b_1 x$$

that satisfies the least squares criterion is called the least squares regression line. Casually, we often just call it the regression line, or the line of best fit.

Slope 176. The slope, b_1, gives a value in "y-units *per* x-unit." Changes of one unit in x are associated with changes of b_1 units in predicted values of y. The slope can be found by

$$b_1 = \frac{rs_y}{s_x}.$$

Intercept 176. The intercept, b_0, gives a starting value in y-units. It's the $\hat{y}$-value when x is 0. You can find it from $b_0 = \bar{y} - b_1\bar{x}$.

s_e 181. The standard deviation of the residuals is found by $s_e = \sqrt{\frac{\Sigma e^2}{n-2}}$. When the assumptions and conditions are met, the residuals can be well described by using this standard deviation and the 68–95–99.7 Rule.

R^2
- 182. R^2 is the square of the correlation between y and x.
- R^2 gives the fraction of the variability of y accounted for by the least squares linear regression on x.
- R^2 is an overall measure of how successful the regression is in linearly relating y to x.

Skills

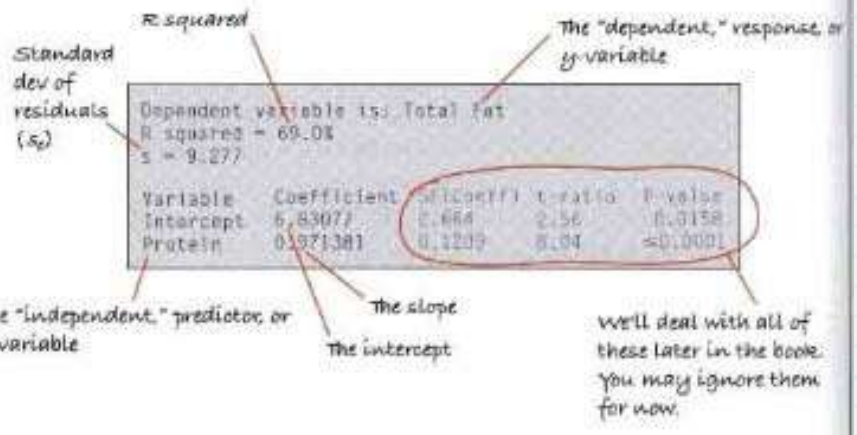
THINK
- Be able to identify response (y) and explanatory (x) variables in context.
- Understand how a linear equation summarizes the relationship between two variables.
- Recognize when a regression should be used to summarize a linear relationship between two quantitative variables.
- Be able to judge whether the slope of a regression makes sense.
- Know how to examine your data for violations of the **Straight Enough Condition** that would make it inappropriate to compute a regression.
- Understand that the least squares slope is easily affected by extreme values.
- Know that residuals are the differences between the data values and the corresponding values predicted by the line and that the *least squares criterion* finds the line that minimizes the sum of the squared residuals.
- Know how to use a plot of residuals against predicted values to check the **Straight Enough Condition**, the **Does the Plot Thicken? Condition**, and the **Outlier Condition**.
- Understand that the standard deviation of the residuals, s_e, measures variability around the line. A large s_e means the points are widely scattered; a small s_e means they lie close to the line.

SHOW
- Know how to find a regression equation from the summary statistics for each variable and the correlation between the variables.
- Know how to find a regression equation using your statistics software and how to find the slope and intercept values in the regression output table.
- Know how to use regression to predict a value of y for a given x.
- Know how to compute the residual for each data value and how to display the residuals.

TELL
- Be able to write a sentence explaining what a linear equation says about the relationship between y and x, basing it on the fact that the slope is given in *y-units per x-unit*.
- Understand how the correlation coefficient and the regression slope are related. Know how R^2 describes how much of the variation in y is accounted for by its linear relationship with x.
- Be able to describe a prediction made from a regression equation, relating the predicted value to the specified x-value.
- Be able to write a sentence interpreting s_e as representing typical errors in predictions—the amounts by which actual y-values differ from the $\hat{y}$'s estimated by the model.

REGRESSION ON THE COMPUTER

All statistics packages make a table of results for a regression. These tables may differ slightly from one package to another, but all are essentially the same—and all include much more than we need to know for now. Every computer regression table includes a section that looks something like this:



AS Finding Least Squares Lines. We almost always use technology to find regressions. Practice now—just in time for the exercises.

The slope and intercept coefficient are given in a table such as this one. Usually the slope is labeled with the name of the x-variable, and the intercept is labeled "Intercept" or "Constant." So the regression equation shown here is

$$\widehat{\text{Fat}} = 6.83077 + 0.397138\,\text{Protein}$$

It is not unusual for statistics packages to give many more digits of the estimated slope and intercept than could possibly be estimated from the data. (The original data were reported to the nearest gram.) Ordinarily, you should round most of the reported numbers to one digit more than the precision of the data, and the slope to two. We will learn about the other numbers in the regression table later in the book. For now, all you need to be able to do is find the coefficients, the s_e, and the R^2 value.