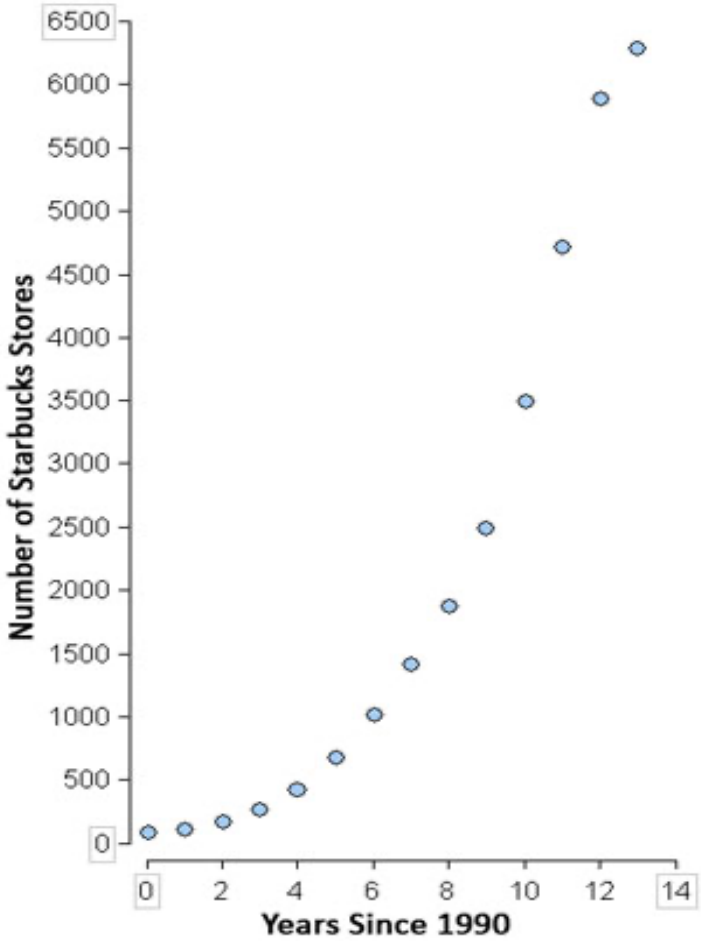
MOD 27 (PART 1) LINEAR REGRESSION

Learning Goals

- For a linear relationship, use the least squares regression line to summarize the overall pattern and to make predictions

1) These data represent the number of Starbucks stores in years since 1990.



a) Use a straight edge to draw a straight line that "best fits" the data (note: your best-fit line need not pass through any of the data points).

b) Label the point on your line that corresponds to the year 2000.

c) Use the graph of your line to predict the number of Starbucks stores in 1993.

d) Use the graph of your best-fit line to estimate the year in which there were 4500 Starbucks stores.

e) Wait for the mini-lecture on the SSE (the sum of the squared errors). How is the **least squares regression line** related to the SSE?

2) Here we take a little sidetrack to look at how statisticians write the equation of a line. Then we'll return to working on the Starbucks data.

Mathematicians often use the slope-intercept formula, $y = mx + b$, to represent the equation of a line that best fits a scatterplot. Which letter in this formula represents the slope of a line and which letter represents the y-intercept?

However, **statisticians** often write the slope-intercept formula for the equation of a line as $y = a + bx$. Which letter represents the slope and which letter represents the y-intercept in our new slope-intercept formula, $y = a + bx$?

Now we can get back to the scatterplot and our least-squares regression line.

Statisticians will often replace the variables x and y with their descriptions in the slope-intercept form of the line, $y = a + bx$. In other words, since x represents the number of years since 1990 and y represents the number of Starbucks stores, we can rewrite the slope-intercept form of the line as follows.

$$y = a + bx \quad \rightarrow \quad Number\ of\ SB\ stores = a + b(Number\ of\ years\ since\ 1990)$$

Where b is the slope and a is the y-intercept.

Use the points you labeled on the Starbuck scatterplot to find the slope b and the y-intercept a. Then substitute the values of a and b into the equation. Show your work.

$$Number\ of\ SB\ stores = a + b(Number\ of\ years\ since\ 1990)$$

3) Before we do any more work, let's all use the same best-fit line. To do this, we need to find the equation of the one true best-fit line, the **least squares regression line**.

Statisticians use the formulas below to find the slope and y-intercept of the least squares regression line.

Equation: $y = a + bx$

Slope: $b = \frac{r \cdot S_y}{S_x}$

Y-intercept: $a = \bar{y} - b\bar{x}$

Here are some descriptive statistics for the Starbucks data.

Variable	Mean	Standard Deviation	Correlation (r)
Year since 1990	6.5	4.183	0.932
Number of stores	2,067	2,192	

Use this information with the above formulas to find the equation of the least squares regression line. Show your work.

What is the slope and what does it mean?

What is the y-intercept and what does it mean?

4) Rewrite the equation of the regression line in the space below, and then use it to make the following predictions.

Regression equation:

a) In what year would we expect the number of Starbucks stores to be approximately 5700? Does your result make sense, why or why not?

b) In what year would we expect the number of Starbucks stores to be approximately 21,000? Does your answer make sense, why or why not?

c) In part a) we used **interpolation** to make a prediction, and in part b) we used **extrapolation** to make a prediction (perhaps we should define these terms together in class right now).

Interpolation:

Extrapolation:

d) Which do you think is a more reliable predictor, interpolation or extrapolation? Explain.

5) Formulas vs Technology

Here again are the formulas we used to find a regression equation.

Equation: $y = a + bx$

Slope: $b = \frac{r \cdot s_y}{s_x}$

Y-intercept: $a = \bar{y} - b\bar{x}$

We used these formulas and the given descriptive statistics to find the regression equation for the Starbucks data. But where did those descriptive statistics come from? We know how to find the means $\bar{x}$ and $\bar{y}$. And we know how to find the standard deviations s_x and s_y. Now it's time to meet the formula for the correlation coefficient r.

$$r = \frac{\sum \frac{(x-\bar{x})}{s_x} \cdot \frac{(y-\bar{y})}{s_y}}{n-1}$$

Discuss with your group-mates how you would organize your work to find the correlation coefficient for the Starbucks data.

Years (since 1990)	# of SB Stores
0	84
1	116
2	165
3	272
4	425
5	676
6	1015
7	1412
8	1886
9	2498
10	3501
11	4709
12	5886
13	6294

6) If we are given a set of data and do not have the correlation coefficient r, we can use technology to find the regression equation $y = a + bx$. In this activity we'll use our graphing calculators to find the regression equation for the Starbucks data. In the next module we'll use an applet to find regression equations, and in the lab, we'll use statistical software such as SPSS or StatCrunch.

WARNING! When you use your graphing calculator to find the equation of the least squares regression line, be sure that diagnostics are turned on so that the correlation coefficient *r* is displayed with the regression equation.

a: b:

Regression equation:

Correlation coefficient, r: Squared correlation coefficient, r^2:

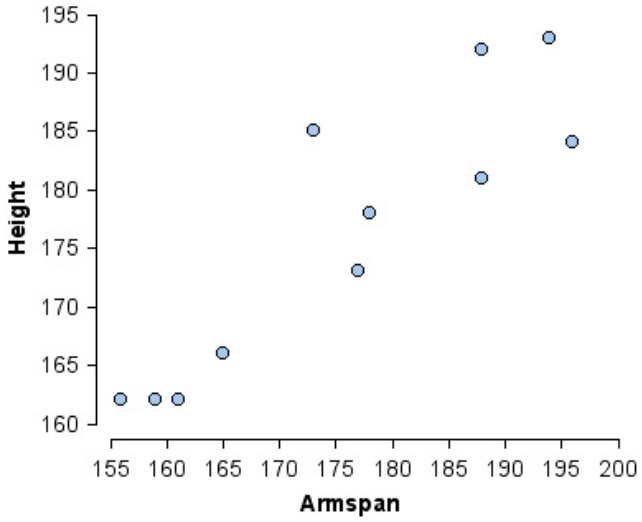
Let's compare the regression equation to the scatterplot for the Starbucks data. Graph the regression equation with the Starbucks data in the same graphing window. Does the line look like a good fit with the data?

Years (since 1990)	# of Starbucks Stores
0	84
1	116
2	165
3	272
4	425
5	676
6	1015
7	1412
8	1886
9	2498
10	3501
11	4709
12	5886
13	6294

MOD 27 – LET'S PRACTICE PART 1

1) The data in the table below gives the arm-span and height (in cm) for 11 men.

Arm Span	Observed Height
161	162
196	184
177	173
188	181
159	162
178	178
194	193
188	192
173	185
165	166
156	162



a) Use the regression capabilities of your graphing calculator to find a linear model for these data. Round to 2 decimal places.

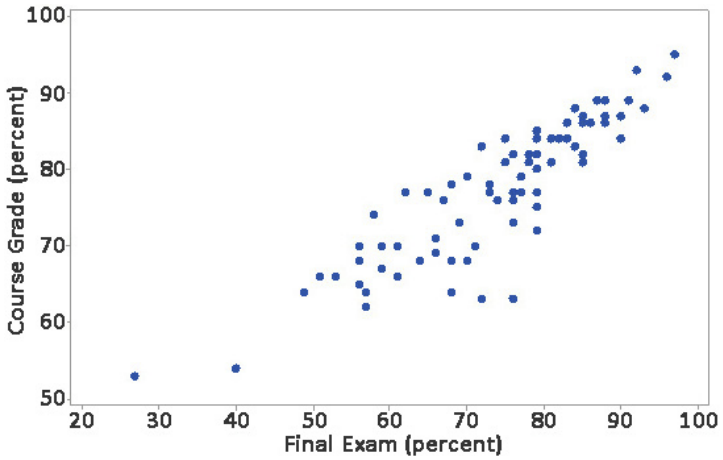
b) What is the slope and what does it mean in context?

c) What is the y-intercept and what does it mean in context?

d) Use your regression equation to predict the height for a man with an arm-span of 164 cm. Is this interpolation or extrapolation? Do you think your prediction is reliable? Explain.

e) Use your regression equation to predict the arm-span for a man who is of 198 cm tall. Is this interpolation or extrapolation? Do you think your prediction is reliable? Explain.

2) We recorded the final exam (in percent) and course grade (in percent) for 75 statistics students. Then we generated the following scatterplot of the data.



a) Which of the following can possibly be the regression line for this data?

- $y = 50.18 + 0.62x$
- $y = 31.72 - 0.62x$
- $y = 31.72 + 0.62x$
- $y = 50.18 - 0.62x$

b) What does each variable in the regression equation represent?

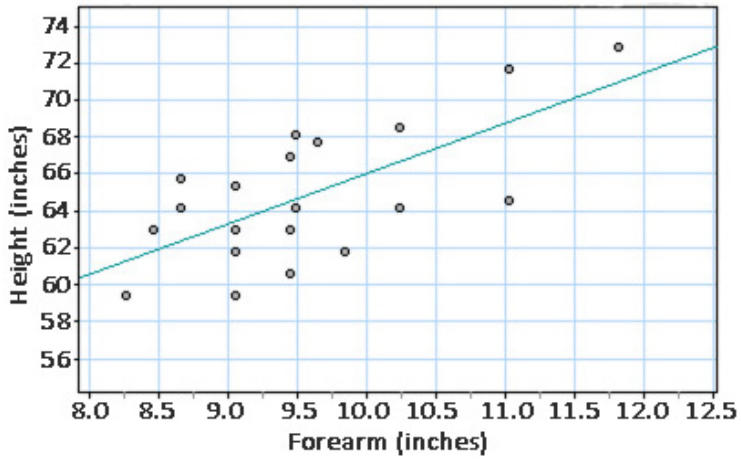
c) What is the most precise and accurate interpretation of the slope?

d) What is the most precise and accurate interpretation of the y-intercept?

e) Using the equation $y = 31.72 + 0.62x$, a teacher's assistant (TA) determined that a student who earns 45% on the final exam is predicted to earn a course grade of approximately 32%. Which one of the following is a reason the prediction is not accurate.

- The TA made a calculation error.
- The TA used the wrong regression equation.
- There is no student in the data set who earned 45% on the final exam.
- This is considered an extrapolation.

3) Below we have data from 21 college students. Forearm length in inches is the explanatory variable. Height in inches is the response variable. The line is a good summary of the linear pattern in the data.



a) Based on the graph of this line, which is the best prediction of the height of a woman with a forearm of 11 inches?

- 64.5 inches
- 68 inches
- 68.75 inches
- just under 72 inches

The equation of the line is given by the following regression equation.

$$Predicted\ height = 39 + 2.7 \cdot forearm\ length$$

b) What is the most precise and accurate interpretation of the slope?

c) What is the most precise and accurate interpretation of the y-intercept?

d) Use the equation of the line to predict the height of a woman with a 10-inch forearm. Show your work.

4) At Cuyamaca College your statistics instructor posted the following information at the end of the semester:

Variable	Class Mean	Standard Deviation	Correlation (r)
Pre-final grade (%)	72	6	0.8
Final exam grade (%)	76	14	

Individual course grades (%) for the final exam have not been posted yet. Hassan wants to predict his final exam score based on this information. He has an 82 pre-final exam average.

What does the least-squares regression line predict for Hassan's final exam score?

MOD 27 GRAPHING CALCULATOR DIRECTIONS

Turn Diagnostics On

To turn diagnostics on, we need to access the *catalog*. Press the **2nd** button, and then press the number 0. Scroll down until you find Diagnostics On. With the arrow pointing at Diagnostics On, press **ENTER**. On your screen you should see the words *Diagnostics On* following by the flashing cursor. Press **ENTER** again. You should see the word *Done* on your screen. Now when you use your calculator to find the least squares regression equation, you'll get the correlation coefficient r and the squared coefficient of determination r^2 along with the regression equation.

Enter Data into the STAT LIST Editor

To access the *list editor*, press the **STAT** button (just to the left of the arrow pad). With option 1 (EDIT) highlighted, press **ENTER**. If your lists already contain data, see *Clear Lists* at the end of this document. Enter the data for the independent variable (a.k.a. the explanatory variable) in L1 and the corresponding data for the dependent variable (a.k.a. the response variable) in L2.

Find the Least Squares Regression Line

After entering the data for the independent (explanatory) variable and the data for the dependent (response) variable in the list editor, you can use your calculator to calculate the equation of the least squares regression line. Press the **STAT** button (just to the left of the arrow pad). Use the arrow pad to scroll right to the **CALC** menu. Scroll down to the **LinReg (a + b*x*)** option. With the **LinReg (a + b*x*)** option highlighted press **ENTER**. The calculator will display **LinReg (a + b*x*)** on the home-screen followed by a blinking cursor. Assuming the independent (explanatory) variable data is stored in L1, press the **2nd** button and then press the number 1 button. Next press the comma button (in the row under the arrow pad). Assuming the dependent (response) variable data is stored in L2, press the **2nd** button and then press the number 2 button. Press **ENTER**.

Clear Lists

These instructions assume you need to clear data from L1. To access the *list editor*, press the **STAT** button (just to the left of the arrow pad). With option 1 (EDIT) highlighted, press **ENTER**. Scroll up until L1 is highlighted. Press the **CLEAR** button. Press **ENTER**. If you need to clear another list, scroll over to that list and repeat these instructions.

Creating a Scatterplot

Input the data into the STAT LIST editor (see above). Press [2nd] [Y=] to access the STAT **PLOT** editor. Press [ENTER] to edit Plot1. Scroll down and highlight the **scatter plot graph** type (first option in the first row). Press [ENTER] to select the **scatter plot graph** type. Scroll down and make sure Xlist: is set to L1 and Ylist: is set to L2. Press [GRAPH] to display the scatterplot. To get a better view of the graph, press [ZOOM][STAT] to perform a ZoomStat.

MOD 27 (PART 2) ASSESSING THE FIT OF A LINE

Learning Objectives

- Use residuals, standard error, and r^2 to assess the fit of a linear model

1) ***Example: House Price*** Suppose that we would like to predict the price of houses based on size in a particular city. A random sample of 30 houses that are for sale is selected. The researchers record the following data for each house: $x = size$ (in square feet) and $y = price$.

   a) What are some of the factors that influence the variability in house price?

   b) We need to know the percentage of the variability in house price that is explained by the linear relationship between *size* and *price*. As it turns out, the coefficient of determination, r^2, will give us that percentage.

   Suppose that you found the least squares regression line that predicts house price based on size, and the correlation coefficient is $r = 0.75$. Calculate r^2 and convert it to a percent.

   Fill in the blank to interpret r^2.

   _____ of the variability in house price is explained by *house size*."

2) Here are data from nine undergraduate public universities and colleges in the western United States (all have enrollments between 10,000 and 20,000 students). The median SAT scores and the six-year graduation rates (as percentages) are given.

Median SAT	Observed Grad Rate (as %)
1242	75.0
1114	71.5
1014	59.3
1070	56.4
920	52.4
888	48.0
970	45.8
937	42.7
871	41.1

   a) Find an equation for the least squares regression line. What is the value of the correlation coefficient r? What is the value of coefficient of determination r^2?

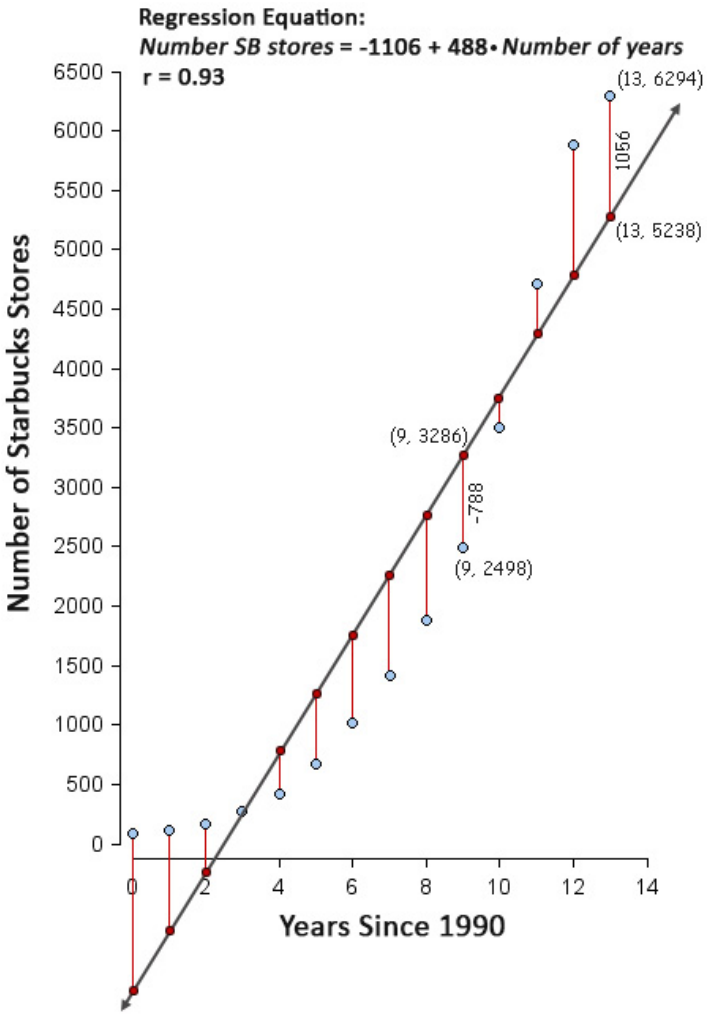
   b) Interpret r^2.

Set up for the computer lab activity

3) Use the Starbucks regression equation to complete the table. Two rows are completed for you. Label a few more signed "distances" (residuals) on the graph.

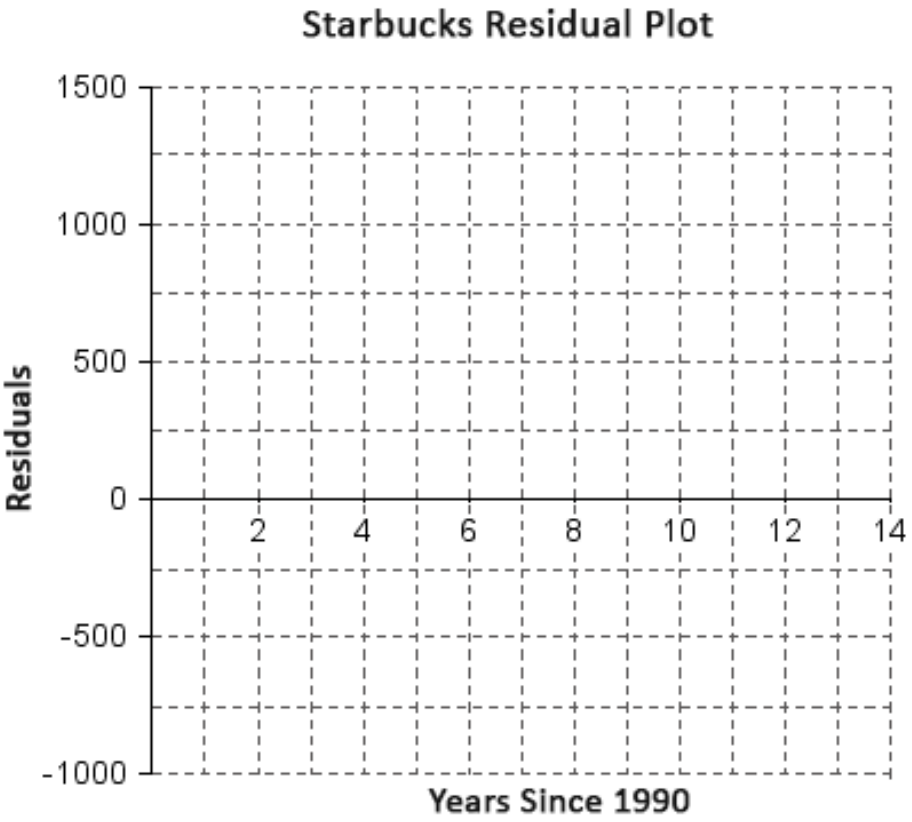
observed – predicted

Years since 1990	Observed # SB Stores	Predicted # SB Stores	Predicted Error (residual)
0	84		
1	116		
2	165		
3	272		
4	425		
5	676		
6	1015		
7	1412		
8	1886		
9	2498	3286	-788
10	3501		
11	4709		
12	5886		
13	6294	5238	1056



Note: the predicted error is also known as the residual. So the vertical lines representing the signed distances from the actual (i.e. observed) value to the predicted value are called the residual lines.

4) Use the predicted errors (residuals) to create a residual plot for the Starbucks data. In other words – create a scatterplot with the explanatory variable (years since 1990) on the horizontal axis and the residuals on the vertical axis.



When you finish the plot, wait for the class discussion before answering the following question.

What does the Starbucks residual plot tell us about our linear regression model? Explain.

5) Here again are some descriptive statistics for the Starbucks data.

Variable	Mean	Standard Deviation	Correlation (r)
Year since 1990	6.5	4.183	0.932
Number of stores	2,067	2,192	

Interpret the standard deviation for the explanatory variable.

Interpret the standard deviation for the response variable.

Fill in the blanks:
When we calculate the standard deviation for one quantitative variable, we get a rough measurement for the average _____ from the _____.
But the _____ from any observed data point to the least squares regression line involves _____ quantitative variables.
So we need a new standard deviation that gives us a rough measurement of the average _____ from the _____.

<u>Standard Deviation vs. Standard Error</u>

Standard Deviation	Standard Error
Roughly measures the average distance of the data points from the mean.	Roughly measures the average distance of the data points from the regression line.
Formula: $\text{sd} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$	Formula: $s_e = \sqrt{\frac{\text{SSE}}{n-2}}$
Where $\sum(x-\bar{x})^2$ is the sum of the squared distances from the mean.	Where SSE is the sum of the squared errors, i.e. the squared residuals.

Note: we'll use technology to find the standard error, s_e, for two quantitative variables.

THE COMPUTER LAB ACTIVITY

6) Use the directions below to find the equation of the regression line and meaningful numerical summaries for the high-school vs college GPA data set. Then answer the following questions.

- Open the *gpa* datafile in StatCrunch (StatCrunch → My Groups → Stats at Cuyamaca College → *gpa*)
- Select **Stat** → **Regression** → **Simple linear**
- Select HS GPA for the X variable.
- Select GPA for the Y variable.
- In the ***Graphs:*** section, click on **Fitted line plot**
- Click **Compute!**
- Use the arrows in the bottom right corner of the StatCrunch output window to toggle between the numerical summaries and graph.

a) Write the equation of the least squares regression line in the space below. State the slope and interpret its meaning in context. State the y-intercept and interpret its meaning in context.

Equation:

Slope:

Y-intercept:

b) State and interpret r^2.

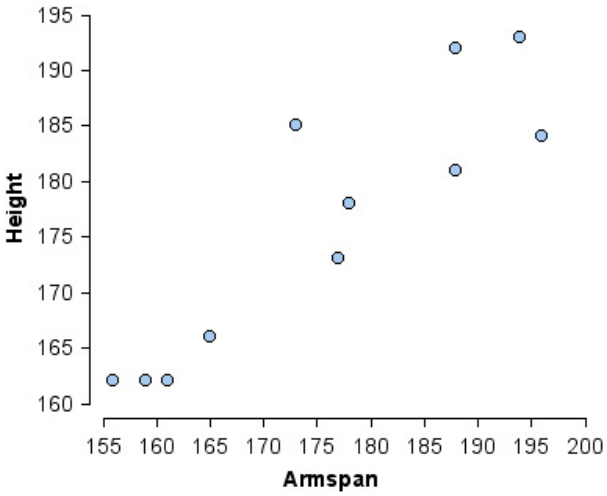
c) State and interpret the standard error, s_e.

d) Considered together, what do r^2 and s_e tell us about the data.

MOD 27 – LET'S PRACTICE PART 2

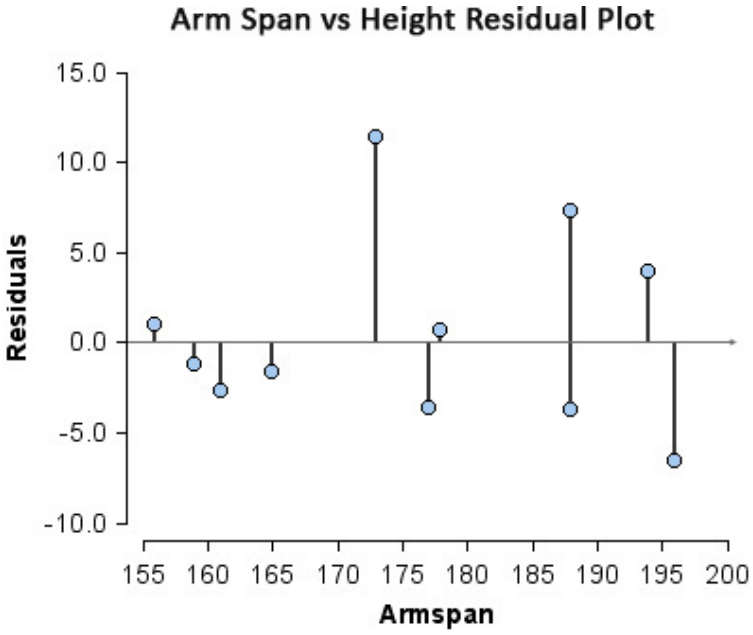
1) The data in the table below gives the arm-span and height (in cm) for 11 men.

Arm Span	Observed Height
161	162
196	184
177	173
188	181
159	162
178	178
194	193
188	192
173	185
165	166
156	162



a) Use StatCrunch to find the regression equation and descriptive statistics for this data (directions are available at the end of this assignment). You'll enter your results on the next page.

b) Here is the residual plot for the data. Explain what the residual plot indicates for your regression equation.



Continued on the next page …

c) What is the regression equation for the arm-span vs height data? (Hint: you found this in Mod 27 – Let's Practice Part 1)

d) What is the coefficient of determination r^2, and what does it mean in context?

e) The standard error, s_e, is about 5.57 cm. What does s_e mean in context?

f) Taken together, what does r^2 and s_e tell us about the regression equation?

2) Here are data for the *instructional expenditure per full-time student* (in dollars) and graduation rates (as a %) for nine universities and colleges.

$ per Full-time Student	Observed Grad Rate (as %)
6960	75.0
7274	71.5
5361	59.3
5374	56.4
5070	52.4
5226	48.0
5927	45.8
5600	42.7
5073	41.1

a) Find an equation for the least squares regression line. What is the value of the correlation coefficient r? What is the value of the coefficient of determination r^2?

b) What is the slope and what does it mean in context?

c) What is the y-intercept and what does it mean in context?

d) Interpret r^2 in context.

e) Use technology to find the standard error, s_e. Then interpret s_e in context.

f) Based on your answers to parts d) and e) above, what can you conclude about the linear relationship between instructional expenditures on full time students and graduation rates?

3) Here is a summary of the descriptive statistics for the graduation data.

$x =$ **median SAT** $y =$ **graduation rate (as a %)**
$S_e = 5.27$
$r^2 = 83.5\%$

$x =$ **\$ per full-time student** $y =$ **graduation rate (as a %)**
$S_e = 8.01$
$r^2 = 61.8\%$

Analyze this information to explain why *median SAT* is a better predictor of *graduation rate* when using linear models.

MOD 27 STATCRUNCH DIRECTIONS

To complete numbers 1 and 2 from the *Mod 27 Let's Practice* homework, you can use StatCrunch to find the least squares regression line, the correlation coefficient r, the coefficient of determination r^2, the standard error s_e, and other information. Here are the steps.

- Open StatCrunch and enter the explanatory variable data in the first column and the response variable data in the second column.
- Choose **Stat** ⟶ **Regression** ⟶ **Simple Linear**
- In the *X variable:* drop down menu to select the column for the explanatory variable.
- In the *Y variable:* drop down menu to select the column for the response variable.
- Scroll down to the *Graphs:* list and click on **Fitted line plot**.
- Click **Compute!**
- Use the arrows in the output window to toggle back and forth between the descriptive statistics (numerical summaries) and the scatterplot with the regression line.