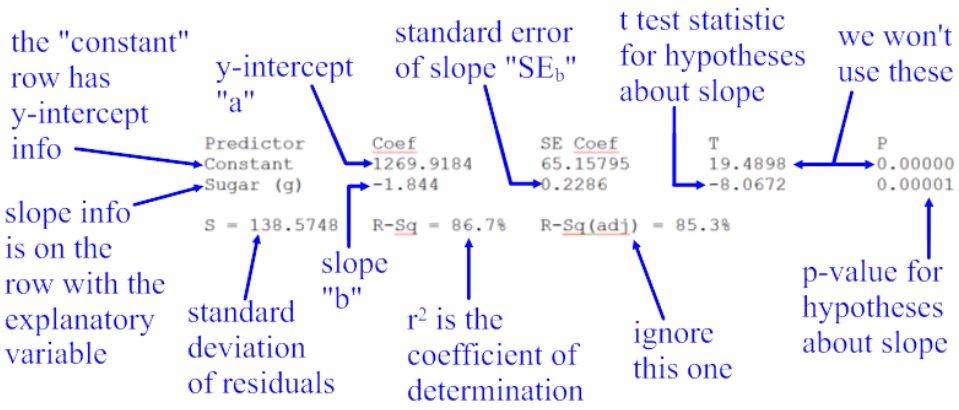
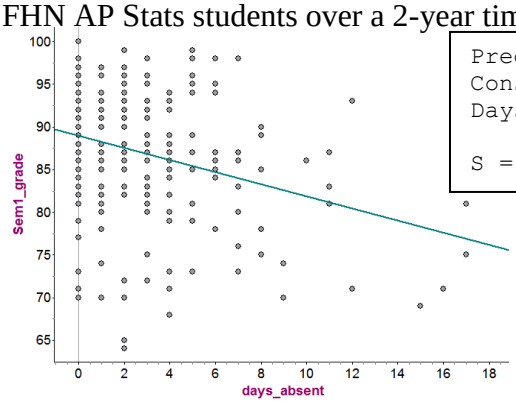
Name _____

Chapter 12 Learning Objectives	**Section**	**Related Example on Page(s)**	**Relevant Chapter Review Exercise(s)**	**Can I do this?**
Check the conditions for performing inference about the slope β of the population (true) regression line.	12.1	745	R12.2, R12.3, R12.4	
Interpret the values of a, b, s, SE_b, and r^2 in context, and determine these values from computer output.	12.1	748, 754	R12.1	
Construct and interpret a confidence interval for the slope β of the population (true) regression line.	12.1	749	R12.3	
Perform a significance test about the slope β of the population (true) regression line.	12.1	754	R12.2	
Use transformations involving powers and roots to find a power model that describes the relationship between two variables, and use the model to make predictions.	12.2	768, 770	R12.5	
Use transformations involving logarithms to find a power model or an exponential model that describes the relationship between two variables, and use the model to make predictions.	12.2	772, 773, 776	R12.6	
Determine which of several transformations does a better job of producing a linear relationship.	12.2	779	R12.6	

Chapter 12 Introduction



Does better attendance *cause* higher achievement, or do better students simply tend to also have good attendance? We can't do an experiment and randomly assign some students to attend more than others, so we have to rely on observational studies. Let's focus just on AP Stats students at FHN. These data come from 204 FHN AP Stats students over a 2-year time period.



Predictor	Coef	SE Coef	T	P
Constant	88.801	7.77	11.429	<0.00001
Days_absent	-0.7099	0.1597	-4.445	0.000007

S = 7.4343 R-Sq = 8.91%

1. Identify the explanatory variable and the response variable.

2. Describe the association shown in the scatterplot.

3. Using the computer output, determine the equation of the least-squares regression line (LSRL).

4. Calculate the value of the correlation, r.

Predictor	Coef	SE Coef	T	P
Constant	88.801	7.77	11.429	<0.00001
Days_absent	-0.7099	0.1597	-4.445	0.000007

S = 7.4343 R-Sq = 8.91%

5. Calculate and interpret the residual for the student who missed 10 days and had a grade of 86%.

6. Interpret the slope, b, of the least-squares regression line.

7. Interpret the standard deviation of the residuals, s.

8. Interpret the value of the coefficient of determination, r^2.

HW #38: page 669 (AP3.1–AP3.35)
12.1 Sampling Distribution of *b*
Read 739–747

The difference between…

The sample regression line	The population (true) regression line
Based on a random sample of individuals	Based on the whole population
$\hat{y} = a + bx$	$\mu_y = \alpha + \beta x$
We use the y-intercept, a, to estimate…	…alpha, α, the y-intercept in the population.
We use the slope, b, to estimate…	…beta, β, the slope in the population.

What is the sampling distribution of *b*?
Take a sample of n, things from a population and find the slope, b, for the LSRL relating two variables x and y for those things. Do that repeatedly, recording the value of the slope, b, for each sample. This process is how we begin finding the sampling distribution of b. That is, **the sampling distribution of b** is the distribution of the slopes of the lines that relate 2 variables for all possible samples of the same size from some population.

What shape, center, and spread does the sampling distribution of b have?

Choose an SRS of n observations (x, y) from a population of size N with least-squares regression line

$$\text{predicted } y = \alpha + \beta x$$

Let b be the slope of the sample regression line. Then:

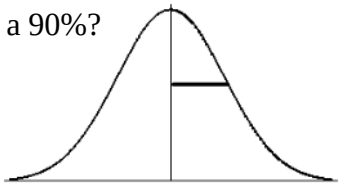
- The **mean** of the sampling distribution of b is $\mu_b = \beta$.
- The **standard deviation** of the sampling distribution of b is

$$\sigma_b = \frac{\sigma}{\sigma_x \sqrt{n}}$$

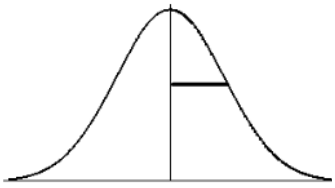
as long as the *10% condition* is satisfied: $n \leq \frac{1}{10} N$.

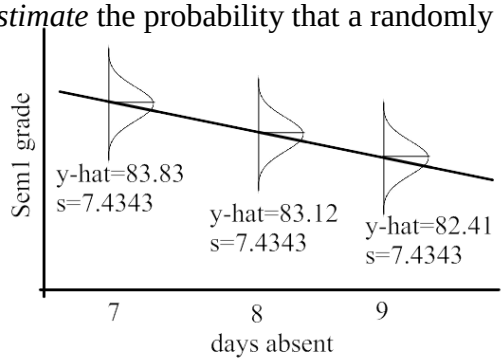
- The sampling distribution of b will be **approximately Normal** if the values of the response variable y follow a Normal distribution for each value of the explanatory variable x (the *Normal condition*).

Suppose that the *true* regression line for all FHN AP Stats students is $\mu_y = 89 - 1x$ with $\sigma = 7$. What is the *actual* probability that a randomly selected student missing 8 days will get at least a 90%?



Use the regression model we found, $\hat{y} = 88.801 - 0.7099x$, to *estimate* the probability that a randomly selected student missing 8 days will get at least a 90%.





What are the conditions for regression inference?

Suppose we have n observations on an explanatory variable x and a response variable y. Our goal is to study or predict the behavior of y for given values of x.

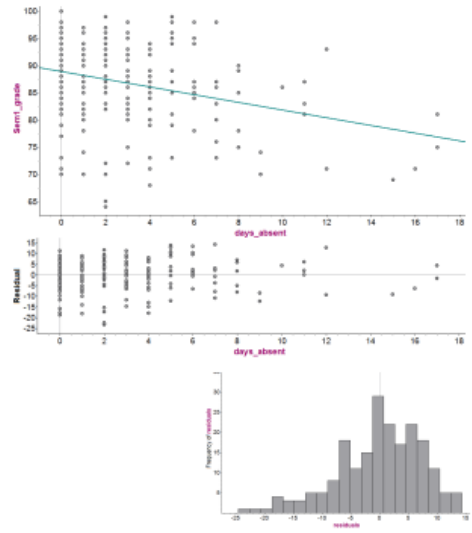
- **Linear:** The actual relationship between x and y is linear. For any fixed value of x, the mean response μ_y falls on the population (true) regression line $\mu_y = \alpha + \beta x$.
- **Independent:** Individual observations are independent of each other. When sampling without replacement, check the *10% condition*.
- **Normal:** For any fixed value of x, the response y varies according to a Normal distribution.
- **Equal SD:** The standard deviation of y (call it σ) is the same for all values of x.
- **Random:** The data come from a well-designed random sample or randomized experiment.

How do we check them?

Linear	Is the scatterplot linear and does the residual plot lack any curved pattern?
Independent	If sampling without replacement, is $n \leq \frac{1}{10}N$?
Normal	Make a stemplot, histogram, dotplot, or NPP of the residuals. Does it lack any clear departures from Normality?
Equal SD	In the residual plot, is the vertical spread roughly the same for all values of x?
Random	Is the data from a random sample or a randomized experiment?

Check the conditions for inference are satisfied for the absence/AP Stats grade data. Which one(s) lead you to hesitate?

Linear	Is the scatterplot linear and does the residual plot lack any curved pattern?
Independent	If sampling without replacement, is $n \leq \frac{1}{10}N$?
Normal	Make a stemplot, histogram, dotplot, or NPP of the residuals. Does it lack any clear departures from Normality?
Equal SD	In the residual plot, is the vertical spread roughly the same for all values of x?
Random	Is the data from a random sample or a randomized experiment?



What are the three parameters in a regression model? What statistics do we use to estimate them?

The 3 parameters in our regression model:	The 3 statistics we use to estimate the parameters:
β, the slope	b, the slope from our data
α, the y intercept	a, the y-intercept from our data
σ, the standard deviation describing spread around the population (true) regression line	s, the standard deviation of the residuals

HW #42: page 759 (1, 3, 5, 29, 30)

Confidence Intervals for β

Read 747–751

The **standard deviation of the slope**, σ_b, is the typical difference between slopes of sample regression lines and the slope of the population regression line.

$$\sigma_b = \frac{\sigma}{\sigma_x \sqrt{n}}$$

The **standard error of the slope**, SE_b, is our *estimate* of the typical difference between slopes of sample regression lines and the slope of the population regression line.

$$SE_b = \frac{s}{s_x \sqrt{n-1}}$$

Here's the computer output again for the absence & AP Stats grade data.
Identify and interpret the standard error of the slope.

```
Predictor      Coef          SE Coef       T             P
Constant       88.801        7.77          11.429        <0.00001
Days_absent    -0.7099       0.1597        -4.445        0.000007

S = 7.4343           R-Sq = 8.91%
```

This formula for constructing a confidence interval for a slope appears in words on the formula packet.

$$b \pm t^* SE_b$$

We get the value of t* from a t-table or invT.

For the kind of regression we do, the degrees of freedom = n-2.

For his science fair project, John decided to investigate the effect of sugar on the volume of bread. He knew that yeast consumes sugar, producing carbon dioxide, but also had read that too much sugar in a recipe could inhibit yeast growth. He used the same recipe and the same oven to make 12 loaves of bread, but varied the amount of sugar.
Here are the data and computer output.

```
Predictor      Coef          SE Coef       T             P
Constant       1269.9184     65.15795      19.4898       0.00000
Sugar (g)      -1.844        0.2286        -8.0672       0.00001

S = 138.5748   R-Sq = 86.7%   R-Sq(adj) = 85.3%
```

Sugar (g)	Volume (cm^3)
50	1296
50	1188
50	1296
100	1188
100	1080
100	1080
250	672
250	576
250	588
500	432
500	504
500	360

Construct and interpret a 99% confidence interval for the slope of the true regression line. Don't do all 4 steps, just the "State", "Do", and "Conclude".

HW #43 page 759 (2, 4, 6, 7, 9, 11)

12.1 Significance Tests for β Read 753–757

What is the standardized test statistic for a significance test for the slope? Is this formula on the formula sheet?

Suppose the conditions for inference are met. To test the hypothesis $H_0: \beta = \beta_0$, compute the test statistic

$$t = \frac{b - \beta_0}{SE_b}$$

Find the P-value by calculating the probability of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the t distribution with df $= n - 2$.

Notice again that the degrees of freedom is n-2. *Note: Computer printout always gives 2-sided P-values.*

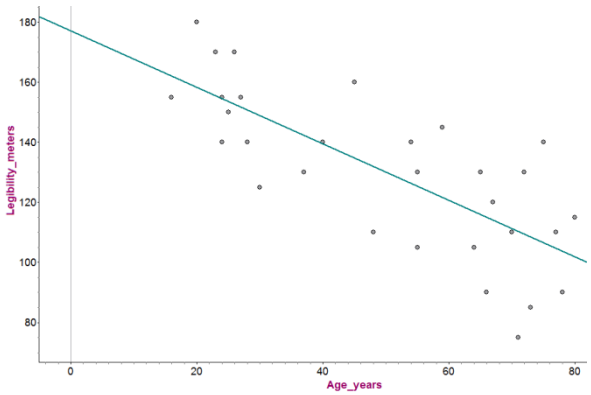
Ooops! My printer messed up. Find the missing t statistic and then use tcdf on the calculator to find the p-value. Assume df=17.

Model	Coef	SE Coef	T	P
Tons mined	4.25	1.25	[redacted]	[redacted]
Intercept	23.35	2.57	9.086	<0.00001

What are two explanations for the negative association seen between absences and AP Stats semester 1 grades?

Does age affect the ability to read road signs at a distance? A college statistics student randomly selected 30 individuals leaving the DMV over the course of a week, asking their age (in years) and for them to read several signs at various distances (in meters) that she had set up around the area, recording the greatest distance the individuals could read accurately. Do these data provide convincing evidence that older individuals tend to have shorter legibility distances?

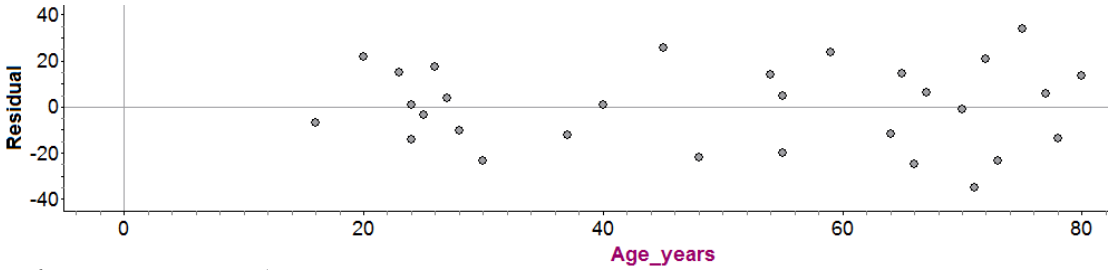
(a) Here is a scatterplot of the data with the least-squares regression line added. Describe what this graph tells you about the relationship between the two variables, Age (years) and Legibility Distance (meters).

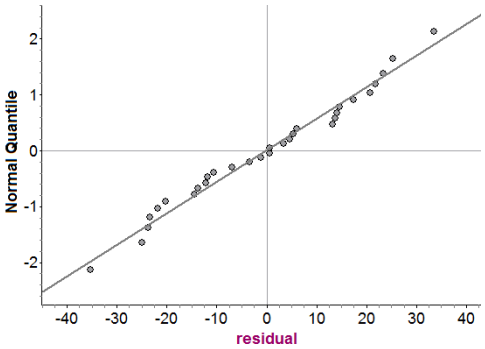


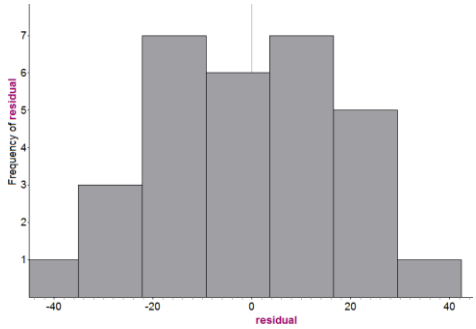
More computer output from a linear regression analysis on these data is shown below. The graphs are a residual plot, a normal probability plot of the residuals, and a histogram of the residuals.

Predictor	Coef	SE Coef	T	P
Age_years	-0.939645	0.158013	-5.947	0.00000569
Constant	176.79	8.05857	21.938	0.00000000

S = 17.7388 R-Sq = 55.8% R-Sq(adj) = 57.9%







Predictor	Coef	SE Coef	T	P
Age_years	-0.939645	0.158013	-5.947	0.00000569
Constant	176.79	8.05857	21.938	0.00000000

S = 17.7388 R-Sq = 55.8% R-Sq(adj) = 57.9%

(b) What is the equation of the least-squares regression line for predicting the maximum distance a driver can legibly read by knowing his/her age? Define any variables you use.

(c) Interpret the slope of the least-squares regression line in context.

(d) Even though it doesn't make sense to do so (newborns don't drive), interpret the y intercept of the least-squares regression line in context. This will serve as an example of how to interpret the y-intercept.

(e) Carry out an appropriate test to answer the student's question.
Note: Computer printout always gives 2-sided P-values.

The calculator can find a lot of the things that the computer output provides, but only when you have data to enter into lists.

HW #44: page 761 (13, 15, 17, 25–28)

12.2 Transformations to Achieve Linearity

Read 765–771 *(focus on fish example primarily)*

When associations are non-linear, we can use powers and roots to model the associations. In order to achieve a linear relationship from a scatterplot that appears to be nonlinear, we will transform one or both variables by using logarithms.

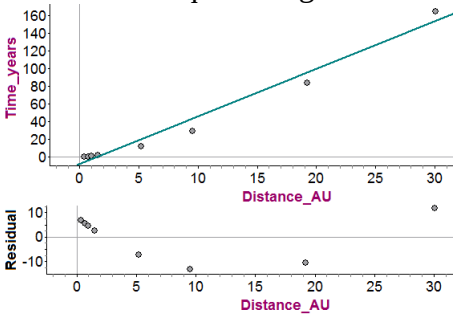
Power Models (made linear by taking log of both variables)
Exponential Models (made linear by taking log of one variables)

12.2 Transformations to Achieve Linearity—Power Models

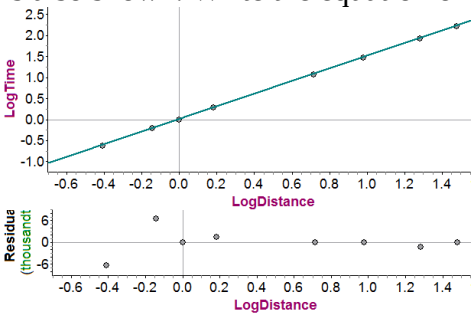
Kepler's Third Law of Planetary Motion predicts the time it takes a planet to go around the sun, based on its distance from the sun. Here are the data for the 8 planetary objects closest to the sun.

Distance	Time
0.39	0.24
0.72	0.62
1	1
1.52	1.88
5.20	11.86
9.54	29.46
19.22	84.01
30.06	164.80

(a) Based on this scatterplot and residual plot, describe why it would not be appropriate to use the least-squares regression line shown.



(b) We take the logarithm (base 10) of each variable to make the association linear. The resulting scatterplot is shown, along with a portion of the computer output from the analysis of the transformed data. The residual plot is also shown. Write the equation of the resulting linear equation.



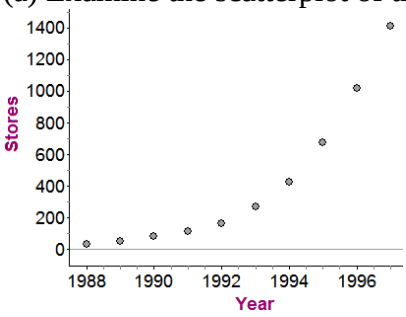

```
Predictor           Coef          SE Coef
Constant            -0.000019     0.0012709
LogDistance         1.50005       0.0020336
```

(c) Predict the time (in years) for the dwarf planet Ceres, located in asteroid belt, to travel around the sun. Ceres is located 2.7675 AU from the sun.

Read 771–774
(d) The dwarf planet Pluto, located beyond Neptune, has an actual orbital period of 248.00 years. Even though this is extrapolating, use the equation to predict the time (in years) for t to travel around the sun. Pluto is located 39.54 AU from the sun.

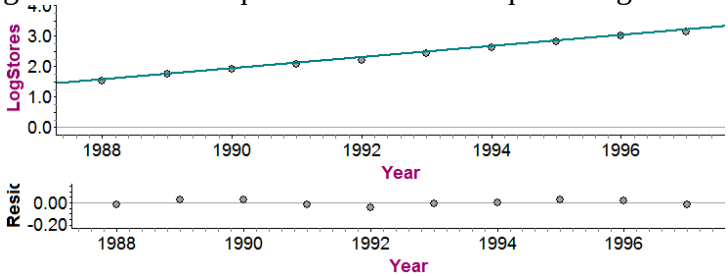
HW #45: page 763 (19–24), page 785 (31–37 odd)
12.2 Transformations to Achieve Linearity—Exponential Models
Read 774–782

The table shows the number of Starbucks coffee shops in existence for several years during a period of rapid growth for the company.
(a) Examine the scatterplot of the Starbucks data. Does the relationship look linear?



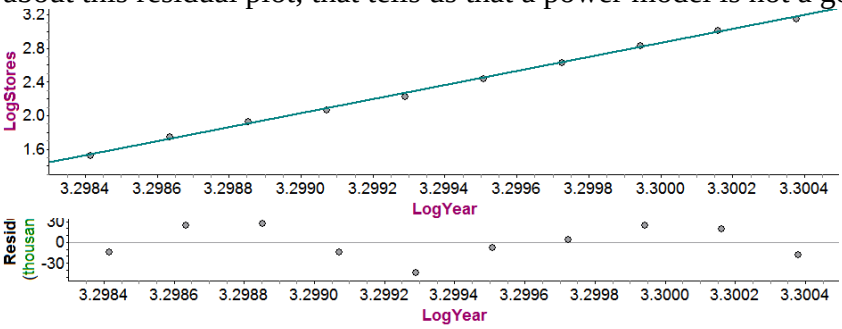
Year	Stores
1988	33
1989	55
1990	84
1991	116
1992	165
1993	272
1994	425
1995	676
1996	1015
1997	1412

(b) This is a scatterplot of log(Stores) vs. year with the LSRL and residual plot as well. It appears that an exponential model would be a good model for this data. A portion of the corresponding computer output is also given. Write the equation of the least-squares regression line.



Predictor	Coef	SE Coef
Constant	-359.70	6.3574897
Year	0.181708	0.0027673

(c) This scatterplot of log(stores) vs. log(year) looks fairly linear. However, look at the residual plot. What is it about this residual plot, that tells us that a power model is not a good fit for this data?



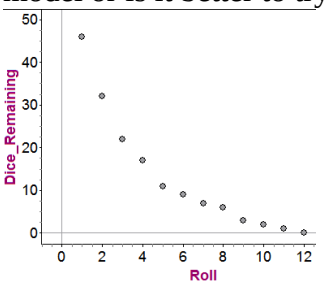
(d) Use the model from part (b) to predict the number of Starbucks locations in 1998.

Predictor	Coef	SE Coef
Constant	-359.70	6.3574897
Year	0.181708	0.0027673

A student rolled some four-sided dice and removed all the ones that landed with the "1" displayed. When he finished, he picked up the dice and repeated the same process over and over until all the dice were removed. Here is a table showing the number of dice remaining at the end of each "roll".

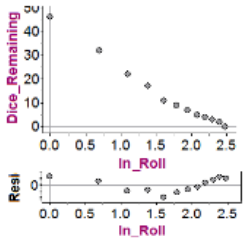
Roll	Dice remaining
1	46
2	32
3	22
4	17
5	11
6	9
7	7
8	6
9	3
10	2
11	1
12	0

(a) Here's a scatterplot of this data. Does the relationship look like it follows a linear model or is it better to try a transformation to make it linear?

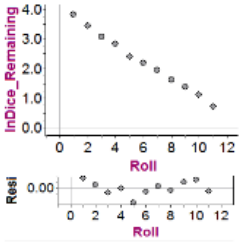


(b) Three different combinations of transformations are done. Examine each scatterplot and residual plot. On two of these, the last observation in the table is not included since ln(0) is undefined. Which one appears to do the best job of making the relationship linear? Explain.

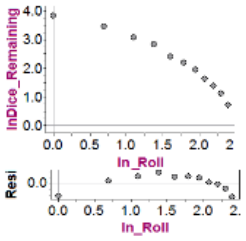
lnRoll



lnDice



lnRoll & lnDice



(c) Computer output from a linear regression analysis on the 11 transformed data points is shown below. This output uses the transformation we just selected. Give the equation of the least-squares regression line defining any variables you use.

Predictor	Coef	SE Coef	T	P
Constant	4.0403	0.04352	92.84	0.000
Roll	-0.301452	0.006413	-47.01	0.000

S = 0.0638 R-Sq = 99.6% R-Sq(adj) = 97.6%

(d) Use your model from part (c) to predict the original number of dice the student used.

(e) Calculate a 95% confidence interval for the slope of the least-squares regression line using the transformed data. Assume all the conditions for inference have been met.

HW #44: page 787 (39–45 odd, 47–50)
Chapter 12 Review / FRAPPY!
FRAPPY from page 793
HW #45: page 795 Chapter 12 Review Exercises
Chapter 12 Review
HW #46: page 797 Chapter 12 AP® Statistics Practice Test
Chapter 12 Test
HW #47: Page 800 (1, 4, 5, 9, 10, 11, 13, 14, 15, 18, 19, 26, 27, 28, 37, 38, 39, 41, 42)

Review for Final / Picking the correct inference procedure

The table below lists the 13 different inference procedures you should know for the AP exam. In each of the scenarios below, choose the correct inference procedure.

8	One-sample *z* interval for *p*	One-sample *z* test for *p*	9
8	One-sample *t* interval for μ, including paired data	One-sample *t* test for μ, including paired data	9
10	Two-sample *z* interval for $p_1 - p_2$	Two-sample *z* test for $p_1 - p_2$	10
10	Two-sample *t* interval for $\mu_1 - \mu_2$	Two-sample *t* test for $\mu_1 - \mu_2$	10
12	*t* interval for the slope of a least-squares regression line	*t* test for the slope of a least-squares regression line	12
		Chi-square test for goodness-of-fit	11
		Chi-square test for homogeneity	11
		Chi-square test for association/independence	11

1. Which kind of light bulbs last longer—60W or 40W?

2. According to a recent survey, a typical teenager has 247 "friends" on Facebook. Is this true at your school?

3. What percent of students at your school have a Twitter ID?

4. Is there a relationship between the age of a car and the mileage reading on the odometer in our county?

5. Is there a relationship between students' favorite academic subject and preferred pizza toppings at a large high school?

6. Who is more likely to own an iPhone—middle school girls or middle school boys?

7. How long do teens typically spend on social media each day?

8. Are the colors equally distributed in Skittles?

9. Which brand of nail polish resists chipping longer? To answer this question, researchers recruited 25 women to have nails on one hand painted with enamel A and the other hand with enamel B.

10. How much more effective is exercise and drug treatment than drug treatment alone at reducing the incidence of stroke among men aged 65 and older?

For more problems like these, search online for: "Larry Green Practice Classifying Statistics Problems"

HW #50 page 800 (2, 3, 6–8, 12, 16, 17, 20–25, 29–36, 40, 43–46)

Preparing for the AP Statistics Exam

The Multiple Choice Section:

- Worth 50% of your overall score. (Each question is 1.25% of your overall score.)
- 40 questions in 90 minutes (there is usually enough time for this section).
- There is no penalty for wrong answers, so ANSWER EVERY QUESTION!
- Generally the questions get harder as you go.
- Skip tough questions and return to them later.

The Free Response Section:

- Worth 50% of your overall score. (Questions 1-5 are each 7.5% of your overall score and question 6 is 12.5% of your overall score.)
- 6 questions in 90 minutes (students might feel a bit more rushed on this section).
- The first 5 questions are shorter and should take 10-15 minutes each.
- The 6th and final question is called the investigative task. It is worth 25% of the free response portion and usually takes 25-30 minutes. The question usually has a "flow" (meaning the parts are connected) and almost always asks the students to do something somewhat new, *just beyond* what you've done in class. Don't save it until the end of the exam, you will be too tired and rushed to think creatively.
- A good strategy is to do 1 question out of 1-5, then question 6, then the remaining 4 questions. Read each question first so you can get the big picture and prioritize your time.
- Communication is very important. Make sure the grader knows what you are doing and why. Don't use statistical vocabulary unless you use it correctly. Define all symbols, draw pictures, etc. Never just give a numerical answer. My best last-minute advice is to communicate clearly.
- Don't just rely on calculator commands. If you use calculator commands, clearly label each number.
- Explain your reasoning. When asked to choose between several options, give reasons for your choice AND reasons why you did not choose the others.
- When you are asked to compare two distributions, use explicit comparison phrases such as "higher than" or "approximately the same as." Lists of characteristics do not count as a comparison.
- Don't waste time erasing. Cross out wrong answers and draw arrows to help the reader follow your work.
- Don't give 2 different solutions to a problem. Both will be read and then you get the lower of the 2 scores.
- Answer all questions in the context of the problem.
- If the question asks you to use results from previous parts of the question, make sure you explicitly refer to them in your answer.
- If you cannot get an answer for an early part of a question but need it for a later part, *make up* a value or carefully explain what you would do if you knew the answer.
- Space on the exam is not suggestive of the desired length of an answer. The best answers are usually quite succinct. There is no need for "extra fluff" on an AP Stats exam.
- Don't automatically enter data into your calculator. In most cases, you will not need to.
- Use words like "approximately" liberally, especially with the word "Normal."

Other Stuff:

- Bring a watch to help pace yourself. Bring an extra calculator, or at least extra batteries and an extra pencil.
- You will be provided formulas and tables (normal, t, chi-square) on both sections (usually, formulas are at the front of the section and tables are at the back).
- Do NOT bring a cell phone (or any other communication device).
- You may not use rulers, white-out, or highlighters.
- You may not discuss the multiple choice questions (ever) and may not discuss the free response questions until they are released on AP Central (not all FR questions will be released).
- Be sure to review computer output and the formula sheets.
- The AP Exam is harder than a normal classroom test. Scoring at least 40% will almost guarantee a 3 or higher on the exam. Don't panic if you cannot answer a question or two.
- You may not have any programs on your calculator except those which upgrade its capabilities to match newer calculators. For example, you may have a program to do inverse-t but not one that lists conditions.