Unit 2B (Chapters 9 & 10) Review

AP Statistics - Ch 9 Summary/Review Sheet

Toolbox:

(Use this section for TI83 commands, important terms, formulas, etc)

Chapter Summary:

(Pretend you are explaining what you learned in this chapter to your grandparents and write it here)

I Can Statements:

(For each of the following statements, assess your ability level using the boxes at the right. Make sure you go back to review the "No" statements!!)

	Yes	Maybe	No
I understand that we cannot fit linear models to nonlinear relationships			
I understand that sometimes there may be subsets in the data worth exploring separately.			
I know the danger of extrapolating into the "future" (beyond the domain of x-values used to find the linear model).			
I understand that points that are outliers in different ways can have different effects on the regression line.			
I can describe how different kinds of outliers affect the regression model.			
I know to look for lurking variables that affect the variables rather than assuming a causal relationship.			
I can create a residual plot and look for patterns in the plot.			
I can explain what fanning of the residual plot means, in context			
I can explain how removing unusual points will likely affect the linear model and summary statistics			

Additional Questions (and Answers!!):

(Use this space to jot down questions that you still have concerning the material in this chapter. Then make sure you ask either your study partner or your teacher about each of your questions and record your answers here!) Use the back for more space, if needed.

AP Statistics – Ch 10 Summary/Review Sheet

Toolbox:

(Use this section for TI83 commands, important terms, formulas, etc)

Chapter Summary:

(Pretend you are explaining what you learned in this chapter to your grandparents and write it here)

I Can Statements:

(For each of the following statements, assess your ability level using the boxes at the right. Make sure you go back to review the "No" statements!!)

	Yes	Maybe	No
I can recognize when a non-linear model may be a better choice to represent the data.			
I understand why a re-expression of the data may be necessary to linearize the data.			
I can write and interpret the equation of a re-expressed scatterplot			
I can recognize when a re-expressed (linearized) model is appropriate			
I can use the re-expressed model in order to make predictions.			
I can use my TI83 to transform a list using logarithms. I will be told which variables to "log".			

Additional Questions (and Answers!!):

(Use this space to jot down questions that you still have concerning the material in this chapter. Then make sure you ask either your study partner or your teacher about each of your questions and record your answers here!) Use the back for more space, if needed.

Some Reminders: For this OTSO, everything from chapters 7-8 is fair game.

Stuff to know for Unit 2B OTSO:

- Know how to use a scatterplot and residual plot to determine if a linear model is appropriate.
- Know how to create a residual plot on the calculator
- Explain what fanning in the scatterplot/residual plot means: Prediction errors will be significantly different at different values of x. Of course, put this in context.
- Understand the effect of different types of unusual points on the linear model (slope, y-intercept, correlation coefficient):
    - Y-outliers with large residuals (over/under x-bar, y-bar)
    - High-leverage points within the trend of the other data points
    - Influential points
- Know how to talk about relationships in context – don't talk about "the data…" is straight, is negative, is positive, is good, etc. Instead "The relationship between _____ and ____ is approximately linear/negative/positive", etc.
- Know that influential points often hide in residual plots due to their high leverage
- Know that Correlation does not *necessarily* imply causation. Although one may cause the other, there may be lurking variables. Don't be too certain in either direction.
- Know the averaged values result in a stronger association
- Know how to analyze scatterplots for distinct groups, and how to determine the models to use for each group.

- Know why we re-express: we wish to linearize the association or reduce fanning.
- Know how to use your TI calculator to re-express the x and/or y-variable using log or ln.
- Know how to create a scatterplot and residual plot of the re-expressed data.
- Write the equation of the re-expressed model using the output from the calculator or computer output. *Be careful with logarithms!!!*
- Know how to make predictions using a model with logarithms and other transformations
- Know that there are many different transformations that we can use to linearize, and we use them for all the same reasons.
- You don't need to know the ladder of powers
- Know how to describe the non-linear relationship between two variables (SUDS, explain what negative/positive direction means).
- Know when a re-expressed model is appropriate (scatterplot approximately linear, residuals plot is approximately randomly scattered)
    - Know that a curve in residuals plot may be OK, as long as the R-squared value is high and the residuals are tiny.
    - ***We are interested in the most useful models, and we take the best we can get***.
    - ***We can talk about a model "***likely" being the most useful.

- Know that two variables that follow an exponential model will be linearized using a "log-linear" transformation, that is $\log(\hat{y}) = b_0 + b_1 x$.
- Know that two variables that follow a power model will be linearized using a "log-log" transformation, that is $\log(\hat{y}) = b_0 + b_1 \log(x)$.

Practice Questions:

1. True or False? Transforming variables in a non-linear fashion will always make a linear model appropriate

2. Which statement about regression analysis is true?

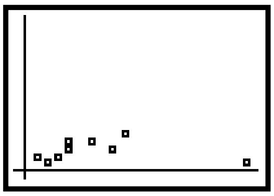
   I. Regressions based on data that are summary statistics (similar subgroups) tend to result in a higher correlation.

   II. If $r^2 = 0.95$, then the response variable increases as the explanatory variable increases.

   III. An outlier always decreases the correlation of a scatter plot.

   (A) I only
   (B) I and II only
   (C) II and III only
   (D) I, II, and III
   (E) None of the above

3. The effect of removing the right-most point (near the positive *x*-axis) in the scatter plot shown is best represented by which of the following?



   (A) The slope of the line will increase; *r* will increase
   (B) The slope of the line will decrease; *r* will increase
   (C) The slope of the line will increase; *r* will decrease
   (D) The slope of the line will decrease; *r* will decrease
   (E) Cannot be determined

4. If removing an observation from a data set would have a marked change on the slope of the regression line fit to the data, the point is called:

   (A) Influential Point
   (B) Outlier
   (C) Residual
   (D) Robust
   (E) None of the above

5. Suppose the correlation between two variables *x* and *y* is due to the fact that both are responding to changes in some unobserved third variable. What is this due to?

   (A) Lurking variable
   (B) Cause and effect between *x* and *y*
   (C) Regression to the Mean
   (D) Outlier effect
   (E) None of the above

6. Which of the following statements is true?

   (A) If the original residual plot of a dataset shows a curved pattern, then the residual plot of the transformed data will also have a curved pattern.
   (B) If the original residual plot of a dataset shows a curved pattern, then the residual plot of the transformed data will not have a curved pattern.
   (C) If you transform data by taking the natural log of both the *x* and *y* variables, the value of r^2 will always increase.
   (D) If both the residual plots of the original data and the transformed data show a curved pattern, then the residual plot with the smallest residuals is the better model.
   (E) None of the above statements is/are true

7. The model $\sqrt{s\hat{t}r}$ = 12 + 20*dia* can be used to predict the breaking strength of a rope (in pounds) from its diameter (in inches). According to this model, how much force should a rope one-half inch in diameter be able to withstand?

   (A) 4.7 lbs.
   (B) 16 lbs.
   (C) 22 lbs.
   (D) 256 lbs.
   (E) 484 lbs.

8. You are shown a scatter plot with an extremely strong positive linear association starting from the origin, steadily increasing, and ending around the point (50,50). Which of the following is not true about the point (2, 75)?

   (A) This point is an outlier.
   (B) This point is an influential point.
   (C) This point may affect the slope of the regression line.
   (D) This point is not very influential because it's rather close to $\overline{x}$
   (E) This point, if removed, will affect the correlation coefficient *r*

Use the following information for the next two questions:

Doctors studying how the human body adjusts to medication inject several patients with a lethal dose of Draino®. The doctors then laughed sheepishly as they recorded how many patients were still alive as time transpired. They hope to see if the amount of time that has elapsed can somehow predict how many patients are remaining. The results are shown below:

Time (hours)	Patients remaining
1	42
2	28
3	19
4	13
5	9
6	6
7	4

9. Which of the following is true with respects to the residual plot for these data?

   (A) This residual plot indicates that a linear model is not appropriate
   (B) This residual plot indicates that the precision of the regression line
   (C) It will always share the same sign as the correlation coefficient.
   (D) If the correlation is found to be zero, the slope of the regression line will also be zero.
   (E) All of the above are true

10. For a residual plot, which of the following indicates that a linear model is appropriate?

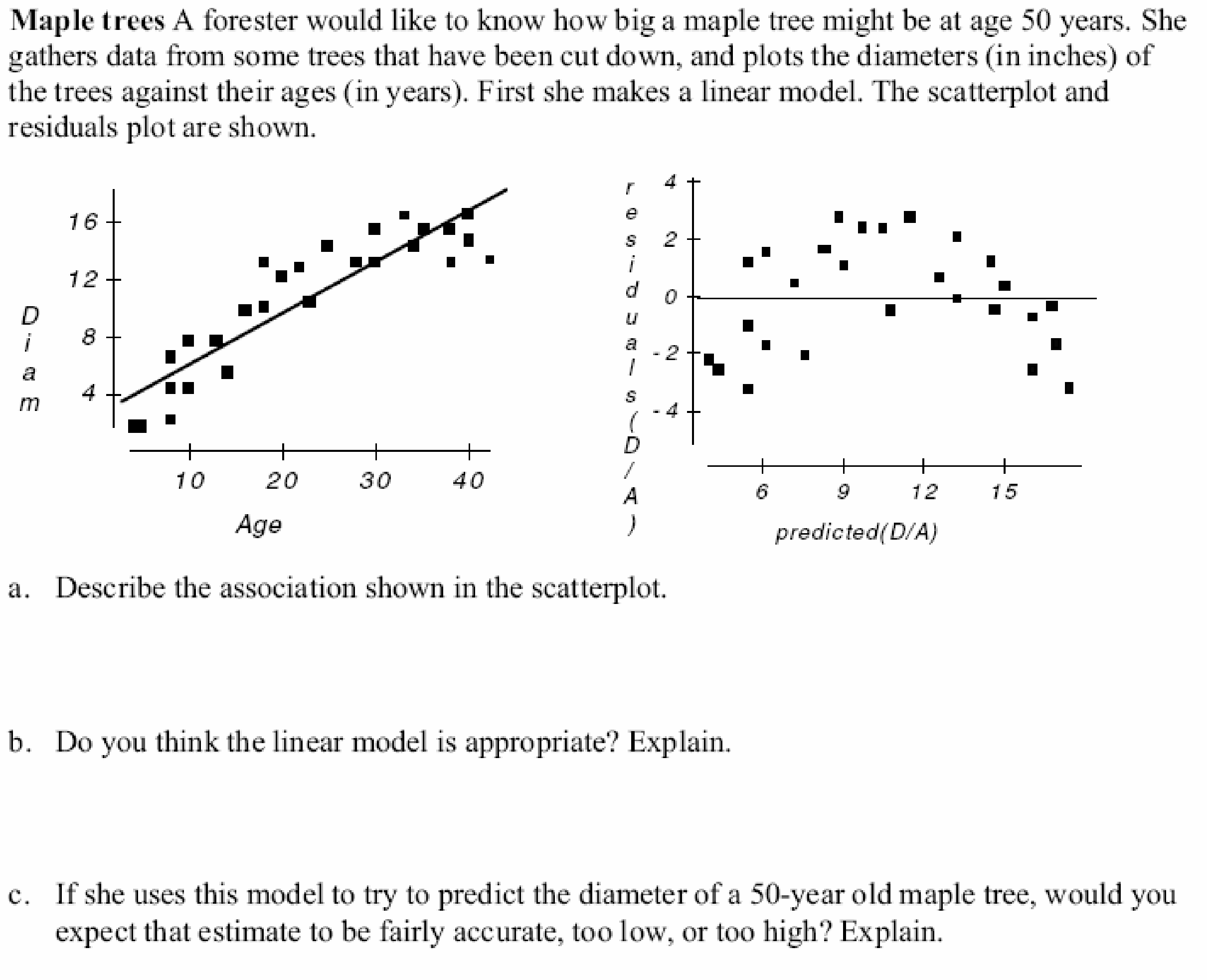
   (A) A strong correlation coefficient
   (B) A large percent of variation in the dependent variable explained by the model
   (C) A small standard deviation of the residuals
   (D) Random scatter of the residuals both above and below zero.
   (E) All of the above

11. Which of the following statements is/are true?

A. A scatterplot will always reveal if a linear model is/isn't appropriate
B. An outlier that is close to the mean of x will dramatically alter the slope of the line of best fit.
C. The units of the residuals depend upon the units of *x*.
D. If in a regression output $R^2 = 100\%$, then $s = 0$.
E. All of the above are true.

Free-response:

- Chapter 9: 5, 12, 18 (You will need to create two linear models for this problem, one for one subset of the years, and another for a different subset. Create a scatterplot, and use it to determine which data to use for each model.)
- Chapter 10: 7 (in a, linearize using log(Salary); in b, define your variables), 22,
- Review of Part II, pp 204-213: 4



Maple trees A forester would like to know how big a maple tree might be at age 50 years. She gathers data from some trees that have been cut down, and plots the diameters (in inches) of the trees against their ages (in years). First she makes a linear model. The scatterplot and residuals plot are shown.

Diam
16
12
8
4
10
20
30
40
Age
residuals(D/A)
4
2
0
-2
-4
6
9
12
15
predicted(D/A)

a. Describe the association shown in the scatterplot.

b. Do you think the linear model is appropriate? Explain.

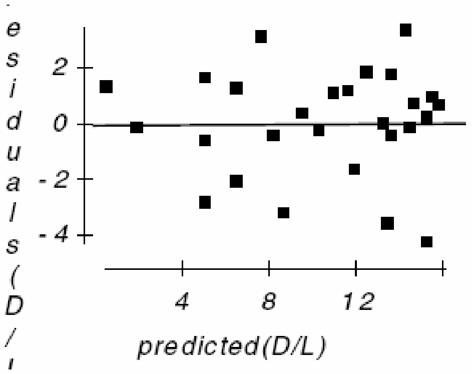
c. If she uses this model to try to predict the diameter of a 50-year old maple tree, would you expect that estimate to be fairly accurate, too low, or too high? Explain.

Now she re-expresses the data, using the logarithm of age to try to predict the diameter of the tree. Here are the regression analysis and the residuals plot.

Dependent variable is: **Diam**

R squared = 84.3%

Variable	*Coefficient*	*s.e. of Coeff*
Constant	*- 8.60770*	*1.681*
Log(Age)	*15.0701*	*1.299*



(d) Comment on whether or not this is a better model.

(e) Use the model to predict the diameter of a 50 year-old maple tree.