Assignment 3.3 AP Statistics Name:

1. Define each of the following terms using your own words:

Outliers An observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction of a scatterplot have large regression residuals, but other outliers need not have large residuals.

Influential observations in regression: 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.

Lurking variable A variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.

2. Given each scenario, indicate the possible lurking variables.

i) A study explored the effects of listening to Mozart at a young age and the IQ of the participants when they became young teenagers. It was then donned the "Mozart effect" where listening to Mozart music makes you smarter.
The parents can be a lurking variable. Parents that get their kids to listen to Mozart could be eager to have their kids succeed. These parents also tend to be more involved in fostering their kids development.

ii) A positive correlation existed in a Miami beach where the number of ice cream sales increased with the number of shark attacks.
The weather can be a lurking variable. When the weather is sunny more ice creams will be sold. Likewise, when the weather is sunny, more people head to the beach and thus more likely to have a shark attack.

iii) A study was done to measure the number of firefighters and the amount of damage done by the fire. Is it enough to claim that the more firefighters you have, the more damage is created?
The size of the fire would be a lurking variable. When there is a big fire, more firefighters will be called to the rescue. Likewise, a bigger fire would do more damage.

iv) In WWII, bombers that attacked London tend to be less accurate with the absence of Allied resisting fighter pilots. Whereas, when fighter pilots were defending the city, the bombers were more accurate. What is the lurking variable?
Cloudy weather was the lurking variable. During WWII, bombers did not have guiding system for their bombs and relied mainly on eyesight. When it was cloudy, they wouldn't be able to see where their bombs were landing. Likewise, resistance fighter pilots wouldn't engage. On a sunny day when the bombers can see, fighter pilots would engage in defending the city.

v) Does AP exams cause Global warming? In the past few years the number of AP exams taken have been increasing steadily. Likewise, the average global temperature is also increasing steadily.

There is obviously no association between the two. Possible Lurking variables: Increase in world population, advancement in society, or time.

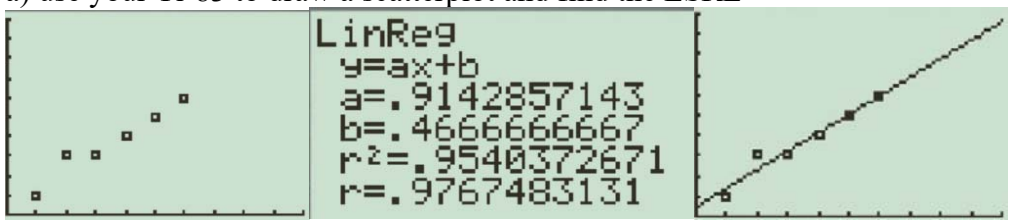
3. What are the differences between an "outlier" and an "influential observation" in regression?

- An outlier is an observation that lies outside the overall pattern of the other observations
- All influential points are outliers, but not all outliers are influential points.
- 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 directions of a scatterplot are often influential for the least-square regression line*

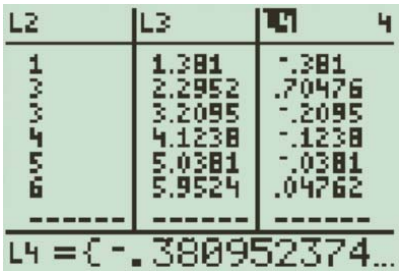
4. Given the following table of values,

x	1	2	3	4	5	6
y	1	3	3	4	5	6

a) use your Ti-83 to draw a scatterplot and find the LSRL



b) Calculate the residual for each data point

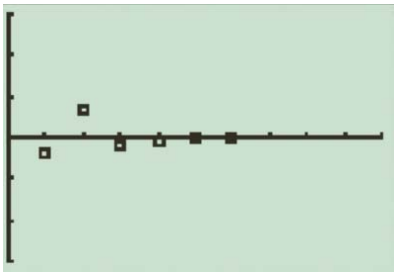


L2 : actual values "y"

L3: Predicted response values $\hat{y}$

L4: Residuals $y - \hat{y}$

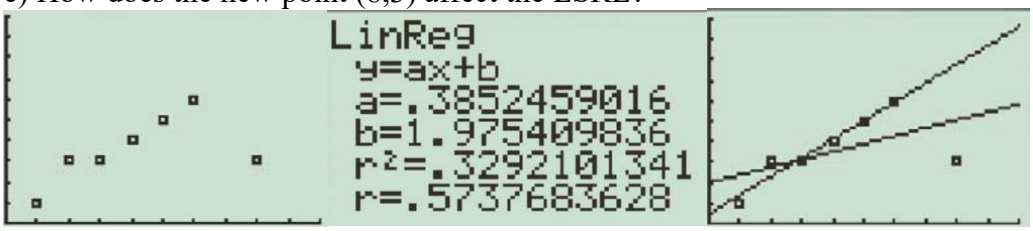
c) Draw a residual graph



The residual plot does not have any clear patterns and most points are very close to the line.

d) A new point (8,3) was collected. Is this new point an "outlier" or an "influential point" Explain
This point would be an 'outlier" because it has a larger residual point. Although this point is greater than most of other points, it would not be considered having an extremely larger "x" value. Thus, it would not be considered as an influential point.

e) How does the new point (8,3) affect the LSRL?



With the addition of the outlier the slope decreased from 0.914285714 to 0.3852459. The Y-intercept increased from 0.46666 to 1.9754098. From the graph on the right you can see how the LSRL can be changed. The slope decreased and the Y-intercept increased.

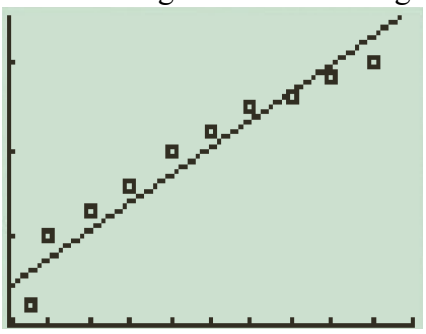
5. What are the two tools that you can use to measure if a LSRL is a good fit?

r^2, coefficient of determination and residual plot. To determine if a LSRL is a good fit, r^2 should be close to 1.

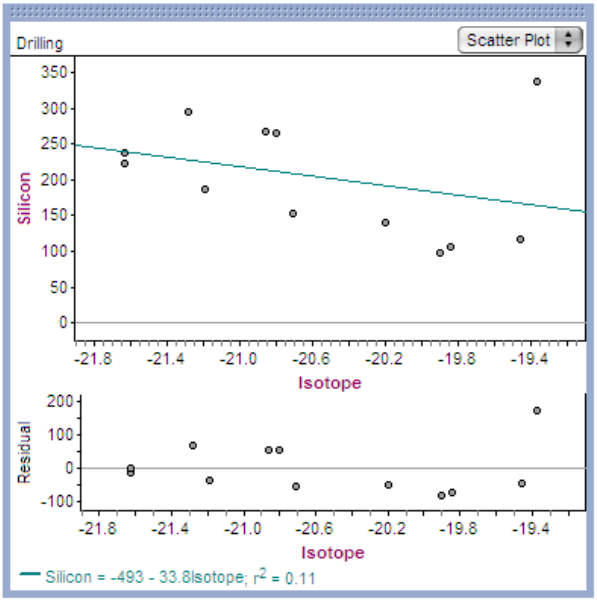
6. How do you use a residual plot to measure if a LSRL is a good fit? Explain:
A residual would be a good fit

A regression is a good model if the points in a residual plot are uniformly spread out (half/half on top and bottom/no obvious pattern, random, unstructured

7. Suppose r^2 is very close to 1 (ie: 0.96). Under what circumstances will r^2 still be a poor tool for determining if the LSRL is a good fit for the data. Provide an example.



8. Drilling down beneath a lake in Alaska yields chemical evidence of past changes in climate. Biological silicon, left by the skeletons of single-celled creatures called diatoms, is a measure of the abundance of life in the lake. A rather complex variable based on the ratio of certain isotopes relative to ocean water gives an indirect measure of moisture, mostly from snow. As we drill down, we look further into the past. Here is some computer output from a linear regression analysis of data from 2300 to 12,000 years ago:

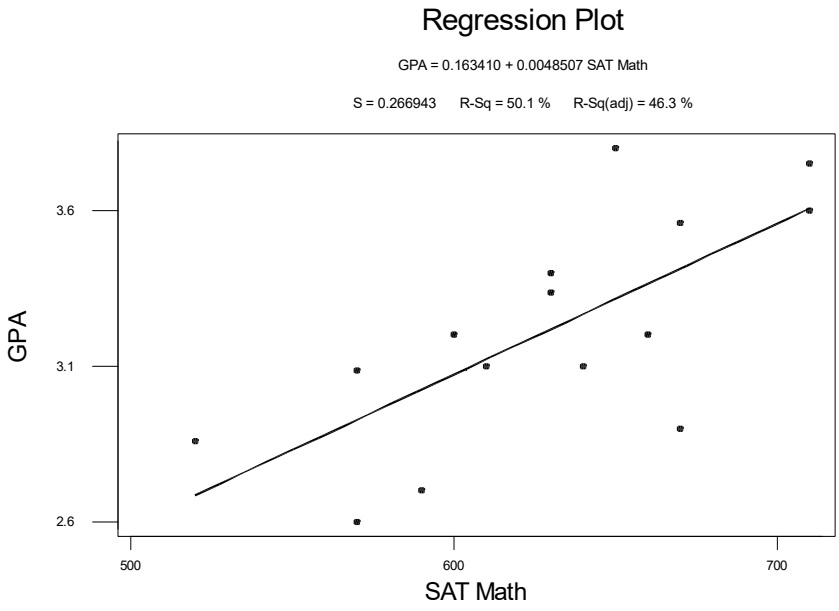


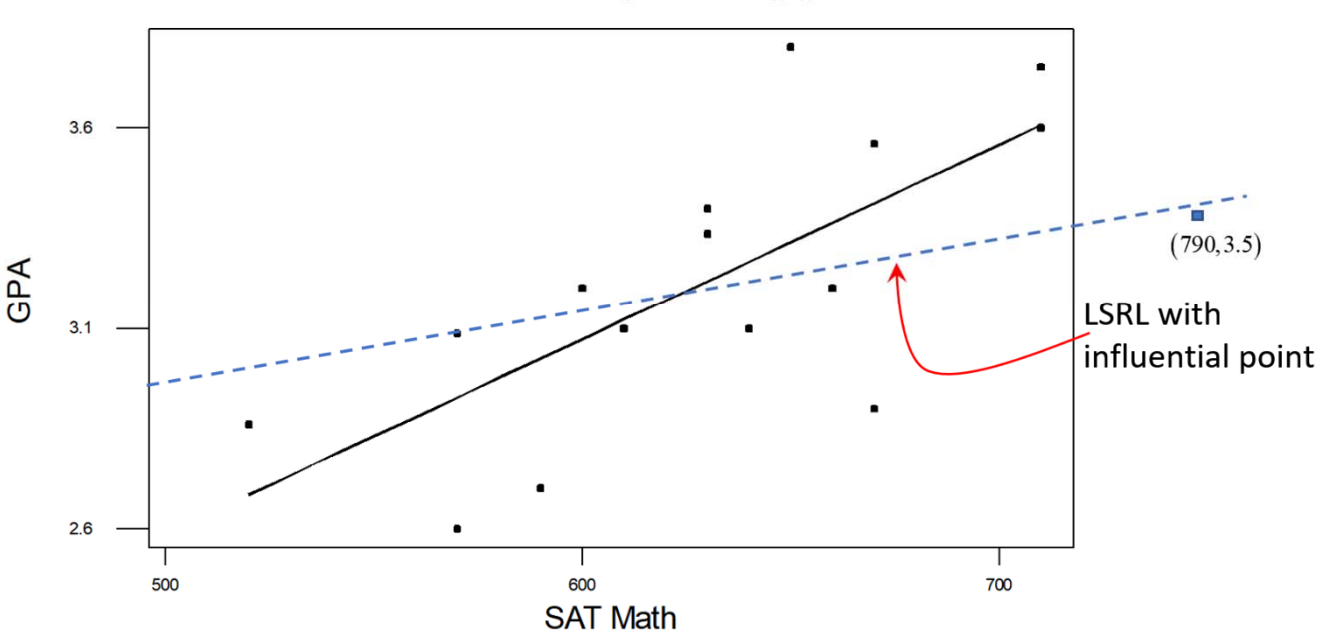
(a) Circle the unusual point in the scatterplot and the residual plot.

(b) If this point was removed, describe the effect on
i) the correlation ii) the least-squares line.

8. What explains grade inflation? Students at almost all colleges and universities get higher grades than was the case 10 or 20 years ago. Is grade inflation caused by lower grading standards? Suggest a lurking variable that might explain higher grades even if standards have remained the same. Justify your answer.
Competition, students spend more time and are better prepared than 10/20 years ago.

10. Mr. Wright believed that he might be able to use students' SAT Math scores to predict their overall grade point averages. He recorded data on a sample of 15 of his students. The scatterplot below displays the data, along with the least-squares regression line.





One student was absent that day. His SAT Math score is 790 and his grade point average is 3.5. What effect would adding this student's point to the scatterplot have on each of the following? Justify your answers.

(a) the correlation

the correlation with this point would be lower because the point is very far out in the "x" direction and well below the linear pattern of other data points. It will "pull" the LSRL towards itself reducing its residual.

(b) the slope and *y* intercept of the regression line

since new point is influential, the leverage exerted on the LSRL would result in a less steep/decreased slope. For instance, the slope would be less positive and the Y-intercept would increase

11. The effect of a lurking variable can be surprising when individuals are divided into groups. In recent years, the mean SAT score of all high school seniors has increased. But the mean SAT score has decreased for students at each level of high school grades (A, B, C, and so on). Explain how grade inflation in high school (the lurking variable) can account for this pattern.

Students who normally get A's today would probably only make A's or B's before meaning there's probably more variability (and perhaps less ability) in this group than the group of "A" students in the past. Therefore, current "A" students would probably have lower mean SAT scores than previous than previous "A" students. The same rationale can also be applied to current "B" and "C" students compared to their past counterparts. Since mean SAT scores is based on averaged data, a reason why current students have higher mean scores is perhaps due to the fact there's a larger group of "A" students factored heavily into the overall mean. Whereas past groups are much smaller. Therefore, it would be less of a factor in the overall mean (Based on GPA)