Content Specific Tips & Traps

1. DESCRIBING: When describing a scatterplot:
* Comment on Form (linear), Outliers, Association, Direction and Strength of the relationship (FOADS)
* Look for patterns in the data and then deviations from those patterns
1. RELATIONSHIP: A correlation near 0 doesn’t necessarily mean there are no meaningful relationships between the two variables. Consider the following data points:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Y | 6 | 30 | 8 | 50 | 10 | 70 | 12 | 90 | 14 | 110 | 16 |

In this case, r=0.38, indicating fairly weak correlation, but a scatterplot displays something quite interesting. Moral of the story: ALWAYS PLOT YOUR DATA.

1. CORRELATION (and SLOPE): Don’t confuse correlation coefficient and slope of the LSRL.
* A slope close to 1 or -1 doesn’t mean strong correlation.
* An r value close to 1 or -1 doesn’t mean the slope is of the LSRL is close to 1 or -1.
* The relationship between b (slope of the regression line) and r (correlation coefficient) is $b=r \left(\frac{S\_{y}}{S\_{x}}\right)$. This is on the formula sheet provided on the exam.
* Remember that r2 > 0 doesn’t mean r > 0. For instance, if r2 = 0.81, then r = 0.9 OR r = -0.9.
1. 4. Residual= Observed – Predicted: $y- \hat{y}$
2. Know the difference between a scatterplot and a residual plot.
3. For a residual plot, be sure to comment on:
* The balance of positive and negative residuals
* The size of the residuals relative to corresponding y values
* Whether the residuals appear to be randomly patterned.
1. Given a LSRL, you should be able to correctly interpret the slope and the y-intercept in the context of the problem. The LSRL should always be stated IN CONTEXT…define variables!
2. The LSRL contains the point ($\overbar{x}, \overbar{y}$ ), where $\overbar{x}$ is the mean of the x-values and $\overbar{y}$ is the mean of the y-values.
3. The LSRL minimizes the sum of the squared residuals (vertical deviations from the LSRL).
4. Logarithmic transformations can be practical and useful. Taking logs cuts down the magnitude of the numbers. Also, if there is an exponential relationship between x and y (i.e. y = abX ), then a scatterplot of the points (x, log y) has a linear pattern.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|

|  |  |  |
| --- | --- | --- |
| X | Y | Log y |
| 1 | 24 | 1.3802 |
| 2 | 192 | 2.2833 |
| 3 | 1,536 | 3.1864 |
| 4 | 12,188 | 4.0859 |
| 7 | 6,290,000 | 6.7987 |
| 8 | 49,900,000 | 7.6981 |

 | An exponential fit to (x,y) yields . When , this model predicts A linear fit to (x, log y) yields , with . If , then Hence  |

1. If the relationship between x and y is described by a power function ( ), then a scatterplot of (log x, logy) will have a linear pattern.

Example:

|  |  |
| --- | --- |
|  |  |

A power fit to (x,y) on the calculator yields . When , this model predicts .

A linear fit to (log x, log y) yields with r=1. When , this model predicts Hence .

1. 12. Remember the correct interpretation of , the coefficient of determination, is: of the variation in y can be explained by the linear relationship with x. Stated IN CONTEXT!
2. 13. ,

Calculator Tips:

Know how to find the LSRL: Stat, Calc, Linreg xlist,ylist.

Know that the residuals are store in a list called RESID when a regression is performed.

Know how to create a scatterplot and residual plot on the calculator.

Multiple Choice Questions

1. Which of the following would not be a correct interpretation of a correlation coefficient of r = -0.30?
2. The variables are inversely related.
3. The coefficient of determination is .09.
4. 30% of the variation between the variables is linear.
5. There exists a weak variation between the variables.
6. All of these are correct.
7. If the correlation between (x,y) is r, then which of the following is true?
8. The variables x and y are linearly related.
9. The correlation of the set (y,x) is also r.
10. The correlation of the set (x,ax) is a\*r.
11. The correlation of the set (ax,ay) is a\*r.
12. None of these.
13. A coefficient of determination is found to be 0.81. Which of the following is true?
14. 81% of the variation between the variables is accounted for in the linear relationship.
15. 81% of the data points lie on a line.
16. The correlation coefficient is approximately ±0.9.
17. 19% of the variation between the variables is accounted for by the linear relationship.
18. All of these is true.
19. Given a set of ordered pairs (x,y) so that , What is the slope of the LSRL?
20. 1.82
21. 1.17
22. 2.18
23. 0.26
24. 0.78
25. A simple random sample of 50 families has produced the following statistics:

Number of children in family: $\overbar{x}=2.1, S\_{x}=1.4$

Annual Gross Income: $\overbar{y}=34,250, S\_{y}=10,540, r=0.75$

The linear regression equation relating these variables, based on data of this study is

1. income = 5646(number of children) + 22392
2. income = 34250 + 0.0001(number of children)
3. income = .0001(number of children) - 1.312
4. number of children = 5646(income) + 22392
5. equation cannot be determined from the given information.
6. You are given the regression equation: $\hat{temperature}=30.4-0.72 (distance)$ where ***temperature*** is the temperature displayed on a sensor in degrees Celsius and ***distance*** is the distance in centimeters from the sensor to the heat source.

Which of the following is not a reasonable conclusion?

1. the temperature of the heat source is 30.4 degrees C.
2. the temperature decreases approximately 0.72 degrees C for each centimeter the sensor is moved away from the heat source.
3. we can predict that the sensor displays a temperature of 21.76 degrees C when the sensor is 12 centimeters away from the heat source.
4. the correlation coefficient between temperature and distance indicates a negative relationship.
5. all of these are reasonable.
6. A study utilizing a simple random sample of 40 college students studied their hours of part-time work per week and GPA and found that the correlation coefficient between these variables was -.43. If the resulting LSRL is: $\hat{GPA }= 3.75 - .05 (number of hours),$ Which of the following is not a correct statement?
7. the average GPA of those who don’t work is approximately 3.75
8. if the correlation coefficient were -.60, the slope of the regression equation would be approximately -.07.
9. students who work 40 hours per week have a mean GPA of approximately 1.75.
10. the value of the correlation coefficient and steepness of the regression line are not related.
11. all of these statements are correct.
12. Which of the following could be the regression line for the plot below?



1. $\hat{y}=384.66-93.72x$
2. $\hat{y}=93.72-384.66x$
3. $\hat{y}=-384.66+93.72x$
4. $\hat{y}=384.66+93.72x$
5. none of these is reasonable.
6. Which of the following residual plots indicates a reasonable fit for the data?



1. Which of the following is a correct conclusion based on the residual plot displayed?



1. the line overestimates the data.
2. the line underestimates the data.
3. it is not appropriate to fit a line to these data since there is clearly no correlation between the variables.
4. the data are not related
5. none of these is correct

Free Response

1. Complete a regression analysis for the following age and income data as indicated.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age (yrs) | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 |
| Income ($1,000) | 18.5 | 23.6 | 29.8 | 38.5 | 49.0 | 64.1 | 78.5 | 102.0 | 130.8 |

a. Construct and label a scatterplot of the data.

b. Perform a linear regression on the data; plot the line on the scatterplot.

c. Discuss the goodness of fit of the linear regression referencing the correlation coefficient and residual plot.

d. Perform the following transformations: exponential and power.

e. Perform the linear regression on both sets of transformed data.

f. Discuss the goodness of fit of these linear regressions referencing the correlation coefficients and each their residual plots.

g. Transform the linear models into the exponential and power models and plot each one on the original scatterplot.

g. Comment on which of the three regression models fits the data the best. Explain your answer.

Multiple Choice KEY

|  |  |  |
| --- | --- | --- |
| 1 | **C** |  |
| 2 | **B** |  |
| 3 | **C** |  |
| 4 | **D** |  |
| 5 | **A** |  |
| 6 | **E** |  |
| 7 | **D** |  |
| 8 | **D** |  |
| 9 | **D** |  |
| 10 | **B** |  |

Free Response KEY

|  |  |
| --- | --- |
| **1.** | * 1. .
 |
| **2.** | 1. ).
 |

|  |  |
| --- | --- |
| **3.** |  |