Quiz 1: Chapter 8, Clarifying the Concepts

1. Indicate whether the following statements are true or false. If the statement is false, alter the statement to make it true.

a. The least-squares line is the line that minimizes the sum of the residuals. F

The Least Squares Regression Line is the line that minimizes the sum of the residuals squared. (I think this can be true)

b. If all the residuals are 0, then SST = SSR. T, SST = SSR + SSE. SSE = sum of square error. If the residuals are 0 (meaning the model is a perfect fit) then there is no square error.

c. The value of the correlation coefficient can be calculated given the value of the coefficient of determination, r^2, alone. True, to find r from r^2 you would square root r^2.

2. Describe the difference between the estimated regression line and the true regression line.

The estimated regression equation shows the equation for y hat (predicted y/target variable) while, the regression model shows the equation for the actual y (target variables) in the data.

3. Where would a data point be situated that has the smallest possible leverage?

If a point has very little leverage/no leverage, it is spot on the least squared regression line.

4. Explain the difference between a confidence interval and a prediction interval. Which interval is always wider? Why? Which interval is probably more useful to the data miner? Why?

The difference is that a confidence interval can have different levels of confidence is a prediction for the mean of the regression model. This means that it has a narrower interval because it's for the mean of the data. A prediction looks at one person or data value and tries to make a prediction for that person so the interval will be a lot wider. A data miner would want to use a confidence interval to talk more generally about a topic.

5. A colleague would like to use linear regression to predict whether customers will make a purchase, based on some predictor variable. What would you explain to your colleague?

I would tell my friend that a linear regression would only work based on if the predictor variables are continuous. regardless of the quality of the relation- ship between them, but this does not guarantee that the regression will therefore be useful. I would continue to tell him that he needs to figure out what the r^2 of his regression model is because it measures the goodness of fit of the regression. R^2 measures how well the linear approximation produced by the least-squares regression line fits the observed data.

6. Match each of the following regression terms with its definition.

Regression Term	Definition	
a. Influential observation	Measures the typical difference between the predicted response value and the actual response value.	J
b. SSE	Represents the total variability in the values of the response variable alone, without reference to the predictor.	S
c. r^2	An observation that has a very large standardized residual in absolute value.	I
d. Residual	Measures the strength of the linear relationship between two quantitative variables, with values ranging from -1 to 1.	G
e. s	An observation that significantly alters the regression parameters based on its presence or absence in the data set.	A
f. High leverage point	Measures the level of influence of an observation, by taking into account both the size of the residual and the amount of leverage for that observation.	K
g. r	Represents an overall measure of the error in prediction resulting from the use of the estimated regression equation.	B
h. SST	An observation that is extreme in the predictor space, without reference to the response variable.	F
i. Outlier	Measures the overall improvement in prediction accuracy when using the regression as opposed to ignoring the predictor information.	H
j. SSR	The vertical distance between the predicted response and the actual response.	D
k. Cook's distance	The proportion of the variability in the response that is explained by the linear relationship between the predictor and response variables.	C