Stat 2470, 4/9 Discussion Questions Name ______________________________

Instructions: Attempt to answer these questions by reading the textbook or with online resources before coming to class on the date above.

1. What are some terms that refer to the input variable in an equation?

   independent, predictor, explanatory variable

2. What are some terms that refer to the output variable in an equation?

   dependent, response variable

3. What is a scatterplot?

   a display of (x_i, y_i) points from the data
   (not connected by lines)

4. What does a simple linear regression equation do that just picking two points from the dataset cannot do?

   it produces a line which has the smallest sum of squares from the points to the line.

5. How is the error term in a regression equation distributed?

   normally

6. How can we tell from the scatterplot that a linear regression equation is reasonable?

   the data should appear to be approximately linear. If the data appears too nonlinear a different kind of regression equation would be more appropriate.

7. What does the notation x^* refer to?

the value of x at which we want to predict a value of $\hat{y}$.
it may be in the original list or it may not be.

8. What is the principle of least squares?

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 - b_i x_i)]^2 \text{ is minimized}$$

$$\text{i.e. } \frac{\partial f}{\partial b_0} = 0 \text{ and } \frac{\partial f}{\partial b_1} = 0$$

9. What are the formulas for finding b_1 (the estimate for the slope parameter) and b_0 (the estimate for the intercept parameter)?

$$b_1 = \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$b_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} = \hat{\beta}_0$$

10. What is the problem with extrapolating beyond the data range input into our model?

we do not know if the behaviour of the data continues in the same fashion or for how long.

11. How are the residuals for a model calculated?

$$y_i - \hat{y}_i$$

12. Why is the degrees of freedom for a linear regression model $n - 2$ rather than $n - 1$?

predicting 2 variables $(\hat{\beta}_0, \hat{\beta}_1)$ and not just one

13. What is the coefficient of determination and how is it calculated?

$$SSE = \sum (y_i - \hat{y}_i)^2 \qquad SST = S_{yy} = \sum y_i^2 - \frac{1}{n}\left[\sum y_i\right]^2$$

$$r^2 = 1 - \frac{SSE}{SST}$$

14. What does the coefficient of determination tell us about our model and its relationship to the data?

the proportion of the variation explained by the linear relationship between the variables

15. What is the formula for the standard deviation for the slope parameter β_1? How is it calculated from the data?

$$S_{\hat{\beta}_1} = \frac{S}{\sqrt{S_{xx}}}$$

S is the error on The residuals

16. Why do we generally use a T-test statistic for the linear regression model rather than a Z?

Sample sizes are generally small and more difficult to check normality assumption of errors. also, we are using the estimate S rather than known σ

17. What is the confidence interval for the slope of the linear regression line?

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \cdot S_{\hat{\beta}_1}$$

18. How do we find the confidence interval for the slope in the calculator?

Stat → Tests → LinRegTInt

19. We can also conduct a hypothesis test on the slope parameter. How can we perform this test by hand and in the calculator? What is the typical H_a for this test?

in calc: Stat → Tests → LinRegTTest

however, working from data we can only compare slope to zero; if comparing to a non-zero slope, we must do it by hand

by hand: $T = \dfrac{\hat{\beta}_1 - \beta_{10}}{\frac{S}{\sqrt{S_{xx}}}} = \dfrac{\hat{\beta}_1 - \beta_{10}}{S_{\hat{\beta}_1}}$ w/ $df = n-2$

we can also use the confidence interval. if β_{10} inside the interval, fail to reject. If not, reject.