Stat 2470, 4/16 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 does it mean for a model to be "intrinsically linear"?

it means that after a suitable transformation of the variable(s) then the model is linear

2. What are the four types of intrinsically linear models and what transformations of variables are done in each case?

exponential : log y

power : log both x & y

log : log x

reciprocal : 1/x

3. Which of these intrinsically linear models have special functions to find the regression equations in the calculator? Which one(s) would have to be done manually?

exponential, power and ln regression all in calculator

reciprocals have to be done "by hand".

4. For each model, how does the error term behave? Is the error term ϵ normally distributed, or a transformation of it?

normally in the linear form, but

$Y = \alpha e^{\beta x} \cdot \varepsilon$ in exponential, $Y = \alpha x^{\beta} \cdot \varepsilon$ in power (multiplied)

log & reciprical models remain as $+\varepsilon$

5. What is logistic regression? How are the variables transformed to be fit to this model?

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad \text{where } p(x) \in [0,1]$$

or $$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}$$

(odds)

6. What kind of behaviour can be modeled logistically?

probability data most straightforward (categorical data) but generally exponential shape on both ends, & sharp transition in The middle where curve flips over

7. What is the general polynomial regression model?

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$$

8. Which polynomial regression models are included in our calculator?

Quadratic, Cubic & Quartic

9. What is the coefficient of multiple determination?

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

10. What is the formula for the adjusted coefficient of multiple determination? Why should we prefer the adjusted R^2 value rather than the unadjusted R^2 value?

$$\text{adjusted } R^2 = 1 - \frac{n-1}{n-(k+1)} \cdot \frac{SSE}{SST} = \frac{(n-1)R^2 - k}{n-1-k}$$

11. To test model utility, we generally want good correlation with as few variable terms as possible (fewer terms, fewer parameters). How can we test hypotheses on the coefficients in our model to see if they are needed? What is the usual H_0 for this situation?

$$t = \frac{\hat{\beta}_i - \beta_{i0}}{s_{\hat{\beta}_i}}$$

$s_{\hat{\beta}_i}$ is so complicated, it's not given in The book

12. What are the formulas for confidence intervals and prediction intervals for our regression equation?

$$CI = \hat{y} \pm t_{\alpha/2,\, n-(k+1)} \cdot s_{\hat{y}}$$

$$PI = \hat{y} \pm t_{\alpha/2,\, n-(k+1)} \cdot \sqrt{s^2 + s_{\hat{y}}^2}$$

13. What is one technique that can be used to possibly reduce the number of parameters needed in a model?

transform variables by a horizontal transformation

14. What is the general additive multiple regression model?

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$

15. Models can be expanded in multiple variables to include interaction terms, and quadratic terms. What are some terms that refer to these different model types (and give examples of each)?

w/interaction $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_1 x_2$, etc

quadratic $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$ etc.

full quadratic (w/interaction) $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$ etc.

16. How can we include categorical variables in multiple regression models?

use indicator variables that take only 0 or 1

17. Why should categorical variables only be models with 0 and 1 and not {0, 1, 2} or other sets of numerical variables?

Categorical data typically not "ranked" in this fashion So each value (minus one) should have its own indicator (the last one is the default if all others are zero)

18. How can we interpret β_i values in the various model types? Give some specific examples.

fix other variables and then it is like interpreting the slope: as x_i increase y_i increases by β_i amount per unit.

19. As with polynomial models, the coefficient of multiple determination come in adjusted and unadjusted versions. When determining whether or not to include another variable in the model, what are we looking for in the R^2 terms?

We are looking for the adjusted R^2 term to increase. if it does not, then we should not include the last variable we added.

20. What is the formula for the model utility test? What other kind of test is similar to this test that also uses the F test? What is the null and alternative hypotheses?

$$f = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$$

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_a: at least one β_i not zero (like ANOVA)

21. What are the degrees of freedom we need for the F test?

$n-(k+1)$

22. What are the formulas for the confidence interval and prediction interval for $\hat{y}$?

CI: $\hat{y} \pm t_{\alpha/2,\, n-(k+1)} \cdot s_{\hat{y}}$

PI: $\hat{y} \pm t_{\alpha/2,\, n-(k+1)} \cdot \sqrt{s^2 + s_{\hat{y}}^2}$

23. How can we look at residuals with a multiple regression model?

plot residuals against each variable separately

24. In section 13.5, there is a discussion of variable selection in multiple regression. What are some key points made there?

Start w/ least complicated model, add complication only as needed (i.e. interaction, quadratic, etc.)

may want to transform variables → horizontal shift or log

identify influential observations

dependence of x_i's on each other