Final

ST 552

March 18th 2016

Answer the questions in the spaces provided on this exam.

Name: _____

- You have 110 minutes to complete the exam.
- There are 3 questions. Answer all of the questions.
- Please
  - do not look at the exam until I tell you and
  - stop writing when I announce that the exam is over.
- There is one page of statistical tables at the end of the exam. You may remove the page of tables if you desire.

Question	Points	Score
1	15	
2	14	
3	16	
Total:	45	

1. Researchers conduct an experiment to investigate the effect of vitamin C on the tooth growth of guinea pigs.

   Each animal received one of three *dose* levels of vitamin C: 0.5, 1, or 2 mg/day, administered by one of two *supplement* methods: orange juice (OJ) or ascorbic acid (VC).

   The response is the *length* of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs.

   The following model was fit:

$$\text{length}_i = \beta_0 + \beta_1 \log_2(\text{dose}_i) + \beta_2 \text{VC}_i + \beta_3 (\log_2(\text{dose}_i) \times \text{VC}_i) + \epsilon_i$$

   where VC is an indicator variable for the ascorbic acid supplement method.

   The resulting estimates and standard errors are:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 20.7 \\ 6.4 \\ -3.7 \\ 2.7 \end{pmatrix}, \quad \text{SE}\left(\hat{\beta}\right) = \begin{pmatrix} 0.7 \\ 0.8 \\ 1.0 \\ 1.2 \end{pmatrix}, \quad \hat{\sigma} = 3.72$$

   (a) Construct a 95% confidence interval for the parameter, β_1. (2)

(b) Interpret the point estimate for β_1, in the context of the study. (3)

Hints:

1. $\log_2(x) + 1 = \log_2(2x)$
2. Restrict your answer to only guinea pigs who received the orange juice supplement, i.e. $VC_i = 0$
3. **Errata** $\log_2(2) = 1$

(c) How would you test if there is indeed an interaction between $\log_2$ (dose) and supplement? Is any additional information required? (You do not need to do the test.) (2)

(d) Estimate the mean tooth length for a guinea pig that receives the OJ supplement at a dose of 1. ($\log_2(1) = 0$) (1)

(e) What additional information is required to construct a confidence interval on the estimate in part (d)? (1)

(f) A more complicated model that treats *dose* as a categorical variable is also fit (including interactions with *supplement*), with a resulting residual sum of squares (RSS) of 712.11 on 54 degrees of freedom.

i. Find the F-statistic for the lack-of-fit F-test. (4)

ii. The p-value corresponding to the F-test above is 0.102. What would you conclude? (2)

2. (a) i. State the assumptions required for making inferences in regression. (2)

ii. For each assumption, describe the consequences of a violation of the assumption. (4)

iii. Which assumptions can be assessed by examining residual plots? For each assumption that can be checked using residual plots, sketch an example residual plot that illustrates what a violation might look like. (4)

(b) i. Describe three ways a point may be considered "unusual". (3)

ii. What is a limitation of case influence statistics? (1)

3. (a) Researchers are interested in the relationship between home heating costs and the method of heating (e.g. oil, gas, electric, or wood burning) across the USA. Since they know people in colder climates will spend more on heating they want to account for climate.

They survey a random sample of households and collect their annual heating expenditure, their heating method and their location. Based on their location they calculate the following four climate variables:

- Average daytime temperature in January
- Average nighttime temperature in January
- Average annual number of days below freezing
- Average annual number of days below 68F

They fit a regression model of annual heating cost against method of heating and the four climate variables.

The researchers are surprised to find that none of the coefficients on the climate variables are statistically significantly different from zero.

i. Should the researchers be surprised? What could explain this outcome? (2)

ii. If the researchers are primarily interested in using this model for prediction, how would you suggest they proceed? (1)

iii. If the researchers are primarily interested in whether any of the climate variables have an effect on heating costs, how would you suggest they proceed? (1)

 Question 3 continues on the next page...

(b) i. In one sentence, describe what is meant by **variable selection** in the context of multiple linear regression. (1)

ii Describe the process of **backward elimination**. (3)

iii. What advantage do **criterion methods** have over stepwise procedures? (1)

iv. What disadvantage do both **stepwise methods** and **criterion methods** share? (1)

(c) Some extensions to multiple linear regression include: (6)

- Robust regression
- Generalized least squares
- Regularized regression
- Logistic regression
- Non-linear regression

Pick **two** methods from the above list and describe how they differ from the usual case of linear regression.