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)

Solution:
95% CI: $\hat{\beta}_1 \pm t_{n-p}^{(0.975)} \text{SE}\left(\hat{\beta}_1\right)$

$n - p = 60 - 4 = 56 \implies t_{n-p}^{(0.975)} = 2.00$

$$\begin{aligned} \hat{\beta}_1 \pm t_{n-p}^{(0.975)} \text{SE}\left(\hat{\beta}_1\right) &= 6.4 \pm 2.00(0.8) \\ &= (4.8, 8) \end{aligned}$$

(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** $\log_2(2) = 1$

Solution: For guinea pigs who received the orange juice supplement, we estimate, that a doubling in dose, results in a 6.4 unit increase in mean tooth length.

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

Solution: t-test of $H_0 : \beta_3 = 0$. No additional information required.

(Half credit: F-test, need RSS for model without interaction)

(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)

Solution: $VC_i = 0$ and $\text{dose}_i = 1 \implies \log_2(\text{dose}_i) = 0$:

$$\widehat{\text{length}}_i = \hat{\beta}_0 = 20.7$$

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

Solution: None, use SE $\left(\hat{\beta}_0\right)$.

(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)

Solution:

$$\text{RSS}_R = \hat{\sigma}^2 * \text{d.f.}_R = 3.72^2(56) = 774.95$$

$$\begin{aligned} F &= \frac{(\text{RSS}_R - \text{RSS}_F)/(\text{d.f.}_R - \text{d.f.}_F)}{\text{RSS}_F/\text{d.f.}_F} \\ &= \frac{(774.95 - 712.11)/(56 - 54)}{712.11/54} \\ &= \frac{31.42}{13.19} \\ &= 2.38 \end{aligned}$$

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

Solution: There is no evidence of lack of fit. Or, the model that treats length as a linear function of log dose is adequate.

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

Solution:

$$Y = X\beta + \epsilon$$

where

$$\epsilon \sim N(0, \sigma^2 I)$$

.

Or, in words:

- Linearity, systematic form of model is correct
- Constant variance, errors all have same variance
- Indpendence, errors are all independent
- Normality, errors are Normally distributed

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

Solution:

- Systematic form of the model, $\mathrm{E}\left(Y\right) = X\beta$. If violated, the parameters in the model may be meaningless, estimates may be biased.
- Independence of errors, ϵ_i independent of ϵ_j for all i and j. If violated, estimates are still unbiased, but standard errors are generally inappropriate.
- Constant variance, $\mathrm{Var}\left(\epsilon_i\right) = \sigma^2$ for all i. If violated, variance in predictions may not be properly quantified.
- Normality, $\epsilon \sim N()$. Can rely on CLT for large samples. If violated, prediction intervals are probably innappropriate.

Question 2 continues on the next page...

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)

Solution: Linearity, constant spread, and Normality.

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

Solution:

- High leverage observations are unusual in their combination of explanatory values and have the potential to be influential.
- Outliers don’t fit the model well (their combination of response and explanatories is unusual according to the model)
- Influential observations substantially change the model when included/excluded. We don’t want out conclusions to rely heavily on a few influential observations. Generally are also one of high leverage and/or outliers.

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

Solution: They don’t pick up groups or clusters of unusual points.

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)

Solution: No, the climate covariates are likely to be very correlated with each other, e.g. a location with low daytime temperatures in Jan is likely to have low nighttime temperatures in Jan too.
Hence, this is probably a situation of ”multicollinearity”. The variance inflation of the individual standard errors, leads lot’s of uncertainty in the point estimates (hence large p-values), but the variables as a group might be explaining a large proprtion of the variation in the response.

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

Solution: Proceed without any changes.

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)

Solution: Do an F-test comparing this model to one without any of the climate variables.

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

Solution: Variable selection refers to the process of selecting a subset of variables for inclusion in the regression model from some larger set.

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

Solution: Start with full model. Drop the variable that has the highest p-value above some critical level, α_{crit}. Repeat until all variables in the model have p-values below α_{crit}.

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

Solution: The do a more complete search of the model space. They let us compare non-nested models.

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

Solution: You can't do the usual inference on the final model.

(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.