STATISTICS 201, FALL 2013, MIDTERM EXAM

NAME:_____KEY_____

Open book and notes, calculator required. You should have 4 pages – make sure you have them all. Each part of each question is worth **6 points**, except where indicated otherwise. Use the back of the pages if you need more space, but please *indicate that you have done so or it may be missed*.

1. A researcher plans to compare two simple linear regression models for predicting Y. One uses X as the predictor and the other one uses X^2. Can this comparison be done using the "general linear test approach" of Section 2.8, which compares full and reduced models? If you think so, write the full and reduced models. If you think not, explain why not.

No, neither model has predictors that are a subset of the other. One has X as the only explanatory variable, and the other has X^2 as the only one.

2. Suppose a 95% confidence interval for β_1 in a simple linear regression situation was (0.9 to 1.8).

a. Based only on the information in the confidence interval, would you be able to make a conclusion about whether or not there is a linear relationship between *X* and *Y*? Explain why or why not.

No. Even though the slope of the least square line is significantly different from 0, there is no way to know if a different relationship, such as a curve, would fit better without seeing a plot of the data.

b. [4 points] What would be the conclusion for a test of H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$ using $\alpha = .05$?

Because 0 is not contained in the interval, we would reject the null hypothesis and conclude that 0 is not a plausible value for the slope.

3. [6 points total] Four conditions required for using linear regression are (1)Linear relationship between X and Y, (2)Constant variance, (3)Normal distribution for the errors, (4)Independence of the Y values. Circle the number(s) corresponding to which (one or more) of these conditions can be checked using each of the following plots:

a. A plot of the residuals versus predicted values: (1) (2) (3) (4)

b. A normal probability plot of the residuals: (1) (2) (3) (4)

Accompanying this exam you will be given a page with R output for the following scenario, which applies to the remainder of the questions. Unless you write something on it that you want graded, you do not need to turn that page in.

A staff person in charge of ordering caps for graduating students has noticed that there is a relationship between head circumference and other physical variables that are easier to measure, including height and length of forearm. She decides to use data from past students to try to quantify the relationship.

The variables include:
Y = Head circumference in centimeters (called **HeadCirc**)
$\mathbf{X_1}$ = Height in inches (**Height**)
$\mathbf{X_2}$ = 1 if the student is male and 0 if female (**Male**)
$\mathbf{X_3}$ = Length of right forearm in centimeters (**RtArm**)

4. Which of the 3 models on the output do you think is best to use? Give a numerical justification for your answer.
The model with Height and Male is the best because it has the highest Adjusted R-squared. You could also say it's best because it has the lowest Residual standard error.

You could also argue that it's the best model by noting that the test for "RtArm" has a p-value of .478, indicating that it doesn't add much to the model once "height" and "male" are there. So, RtArm does not seem to be needed. On the other hand, Height and Male are both needed, as indicated by the p-values for the tests for their coefficients.

For Questions 5 and 6 use the model with Height only (Model 1, at the top of the output page).

5. [2 pts each blank] Fill in a numerical value in each blank. *If a numerical value cannot be determined from the computer output, write NA (not available).* No *extensive* computations are required.

a. b_1 = __*0.31104*_

b. The p-value for testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$ is ___*2.62×10^{-6}*_____

c. β_0 = __*NA (This is the population intercept, which is not known.)*___

d. $\sqrt{MSE}$ = _*1.771 (This is the Residual standard error.)*____

e. SSTotal = __*247.629 (This is computed as 87.674 + 159.955)*___

f. $\hat{Y}$ for a student whose height is 62 inches = _*35.6409 + 0.31104(62) = 54.92538*__

g. A 95% confidence interval for the population slope is _*0.31104*_ ± (2.008) _*0.05883*_, where 2.008 is found from the *t*-distribution with _*51*____ degrees of freedom.

6. Does the intercept of 35.64 have a useful interpretation in this situation? If so, give the interpretation. If not, explain why not.
No. It would correspond to the predicted head circumference for someone who is 0 inches tall, which is obviously impossible.

For Questions 7 to 11 use the model with Height and Male (Model 2, in the middle of the page)

7. Write the estimated (sample) regression function (plugging in numbers). Use X, Y notation rather than variable names.

$$\hat{Y}_i = 43.21 + 0.19X_{i1} + 1.42X_{i2}$$

Rounded off to 2 decimal places each; you could use more.

8. One female who was 65 inches tall had a head circumference of 57 centimeters.

a. [2 points] Write the row of the X matrix for this person.

[1 65 0] *(The first 1 is for the intercept, then her height, then a 0 because she is female.)*

b. [4 points] Calculate $\hat{Y}$ for her.

$\hat{Y}_i = 43.21 + .19X_{i1} + 1.42X_{i2} = 43.21 + (.19)(65) + 0 = 55.56$ *cm. If you did not round off, you would get 55.60156 cm.*

c. [4 points] Calculate her residual.

$Y_i - \hat{Y}_i = 57 - 55.56 = 1.44$ *(or 1.39844 if you did not round off)*

9. Interpret in words the coefficient for "male" of 1.42.

This is an estimate of the average difference in head circumference (in centimeters) for males and females of the same height.

10. For the new values $\mathbf{X_h'} = [1 \ 72 \ 1]$, one of the following is a confidence interval for $E\{Y_h\}$ and the other is a prediction interval for $Y_{h(new)}$. Circle the prediction interval and explain how you know which is which:

(57.6, 59.1) **(54.8, 61.9)**

The prediction interval is 54.8 to 61.9. We know that's the correct one because prediction intervals are always wider than confidence intervals for the mean at that same set of X values.

11. Refer to the previous question, giving a confidence interval and prediction interval. Interpret the *confidence interval* in words, in the context of this situation. Be specific, including relevant numbers.

We are 95% confident that the average head circumference for the population of males who are 72 inches tall is between 57.6 centimeters and 59.1 centimeters.

12. [8 points] The staff person is trying to determine whether it is necessary to include "Male" in the model, once Height (but not RtArm) is in the model. State the null and alternative hypotheses she is testing, provide a test statistic and p-value for the test, and make a conclusion

H_0: $\beta_2 = 0$, Ha: $\beta_2 \neq 0$, test statistic is t = .2.135, p-value = .0376.

Reject the null hypothesis. Conclude that it is necessary to include Male in the model.

13. The correlation between Head circumference and RtArm (forearm measurement) is +0.39. So the staff person was perplexed that the coefficient for RtArm in Model 3 is −0.1226, a negative value. Explain how the coefficient in this case could be negative when the correlation between the two variables is positive.

The coefficient for each variable in multiple regression is what's needed <u>after</u> accounting for the other variables in the model. In this case, RtArm and Height would clearly be correlated. So the coefficient is saying that for two people of the same height and gender, we would predict the one with the longer forearm to have a slightly smaller head circumference.