NAME:________**KEY_**____________________

STATISTICS 110/201, FALL 2009, MIDTERM EXAM

Open book and notes, calculator required. You should have 5 pages – make sure you have them all. Each question is worth **7 points**, except questions 8 and 12 to 15, which are worth **6 points** each. Use the back of the pages if you need more space, but please *indicate to us to turn the page over and look*.

Attached to this exam as the last page are Stata results for the following scenario, which applies to Questions 1 to 11. You may remove the last page and use it to answer 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 fore arm. She decides to use data from past students to try to quantify the relationship.

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

For Questions 1 to 5 use the model with Height and Male only (middle of last page)

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

$$\hat{Y}_i = 43.21 + .19X_{i1} + 1.42X_{i2}$$

2. One female who was 65 inches tall had a head circumference of 57 centimeters. Find her residual.

For this person, $Y_i = 57$, $X_{i1} = 65$ and $X_{i3} = 0$. So, $\hat{Y}_i = 43.21 + .19(65) = 43.21 + 12.35 = 55.56$
Residual = $\hat{Y}_i - Y_i = 57 – 55.56 = 1.44$

3. 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.

4. 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.

5. 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 males who are 72 inches tall is between 57.6 centimeters and 59.1 centimeters.

For Questions 6 and 7 use the model with Height only (bottom of last page).

6. Give the numerical values for two different test statistics that can be used to test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. You don't need to actually carry out the test.

$t = 5.29$ and $F = 27.95$

7. 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 does not. It would be the average head circumference for people who are 0 inches tall, which obviously does not make sense.

Questions 8 to 11 require use of information from some or all of the models.

8. Continuing to define X_1 = height, X_2 = right arm and X_3 = male (1 for male, 0 for female), write numerical values for the following (**2 points each**):

SSR(X_1) = **87.67** (This is the sum of squares in the “Model” row for the model with Height only.)

SSR(X_1, X_3) = **101.04** (The sum of squares in the “Model” row for the model with Height and Male.

SSR(X_2| X_1,X_3) = 102.55 – 101.04 = **1.01** (See Question 9.)

9. Explain what SSR(X_2| X_1,X_3) represents in the context of this situation.

This is the extra regression sum of squares for adding the X_2 variable (RtArm), given that X_1 and X_2 were already in the model. It’s found by subtracting SSR from the model with Height and Male only from the model with all three variables.

You could also say that it’s the reduction in the error sum of squares, or the unexplained variability, when Right Arm is added to the model. Or you could say it’s the increase in the explained variability when Right Arm is added to the model (given that Height and Male are already in the model.)

10. Using information from the results of all 3 models, which do you think is the best model? Explain how you reached your decision.

The best way to compare these models is to use Adjusted R-squared. It’s highest for the model with Height and Male, so that’s the best model.

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.

11. The staff person is trying to determine whether it is necessary to include “Male” in the model, once Height is in the model. State the null and alternative hypotheses she is testing, and provide a test statistic and p-value for the test. (Continue to number things as originally numbered, so Height is X_1 and Male is X_3.)

H_0: $\beta_3 = 0$, Ha: $\beta_3 \neq 0$, test statistic is t = .2.14, p-value = .038. These are from the model with Height and Male (the middle of the page).

The remaining questions do not use the above scenario and are worth 6 pts each.

12. 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? If you think so, write the full and reduced models. If you think not, explain why not.

The method cannot be used because one model is not a reduced version of the other model. They include different predictors – one is not a subset of the other.

THE FOLLOWING SCENARIO APPLIES TO QUESTIONS 13 TO 15:

A medical study is done in which participants with high blood pressure are randomly assigned to one of two stress reduction programs – meditation or exercise. The variables measured are:

Y_i = drop in systolic blood pressure during the 10 weeks of the program
X_{i1} = percent body fat at the beginning of the study
X_{i2} = 1 if in the meditation program, 0 if in the exercise program

The regression model to be fit (without the subscript *i*) is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

13. Here are the X values for the first 6 people in the study:

Body Fat %	30	35	28	48	25	40
Program	Meditation	Meditation	Exercise	Meditation	Exercise	Exercise

Write down the first six rows of the **X** matrix for this situation, filling in numerical values.

$$\begin{bmatrix} 1 & 30 & 1 \\ 1 & 35 & 1 \\ 1 & 28 & 0 \\ 1 & 48 & 1 \\ 1 & 25 & 0 \\ 1 & 40 & 0 \end{bmatrix}$$

14. Write in symbols the null and alternative hypotheses for testing whether there is a difference in effectiveness for the two programs in reducing blood pressure.

H_0: $\beta_2 = 0$ versus Ha: $\beta_2 \neq 0$

(The difference in effectiveness for the two programs is measured by β_2, the coefficient for X_2. If it isn't needed in the model, then there is no difference in the effect of the two programs on blood pressure.)

15. If the null hypothesis in Question 14 is rejected, can the researchers conclude that the difference in programs *caused* the difference in blood pressure results? Why or why not?

Yes. This is a randomized experiment, so a cause and effect conclusion can be made.

Stata results using all three X variables:

```
. regress HeadCirc Height RtArm Male

      Source |       SS       df       MS              Number of obs =      53
-------------+------------------------------           F(  3,    49) =   11.55
       Model |  102.554509     3  34.1848363           Prob > F      =  0.0000
    Residual |  145.074167    49  2.96069729           R-squared     =  0.4141
-------------+------------------------------           Adj R-squared =  0.3783
       Total |  247.628676    52  4.76208992           Root MSE      =  1.7207

------------------------------------------------------------------------------
    HeadCirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Height |   .2373183   .1036248     2.29   0.026     .0290765    .4455602
       RtArm |   -.122591   .1715517    -0.71   0.478    -.4673371    .2221551
        Male |   1.410817   .6686432     2.11   0.040     .0671279    2.754506
       _cons |   43.15818   5.286619     8.16   0.000     32.53432    53.78204
------------------------------------------------------------------------------
```

Stata results for the model including "Height" and "Male" but *not* "RtArm":

```
. regress HeadCirc Height Male

      Source |       SS       df       MS              Number of obs =      53
-------------+------------------------------           F(  2,    50) =   17.23
       Model |  101.042616     2  50.5213081           Prob > F      =  0.0000
    Residual |   146.58606    50   2.9317212           R-squared     =  0.4080
-------------+------------------------------           Adj R-squared =  0.3844
       Total |  247.628676    52  4.76208992           Root MSE      =  1.7122

------------------------------------------------------------------------------
    HeadCirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Height |   .1906646   .0800775     2.38   0.021     .0298242    .3515049
        Male |   1.420539   .6652255     2.14   0.038     .0843943    2.756684
       _cons |   43.20867   5.260216     8.21   0.000     32.64321    53.77412
------------------------------------------------------------------------------
```

Stata results for the model including the predictor "Height" only:

```
. regress HeadCirc Height

      Source |       SS       df       MS              Number of obs =      53
-------------+------------------------------           F(  1,    51) =   27.95
       Model |  87.6738528     1  87.6738528           Prob > F      =  0.0000
    Residual |  159.954823    51  3.13636908           R-squared     =  0.3541
-------------+------------------------------           Adj R-squared =  0.3414
       Total |  247.628676    52  4.76208992           Root MSE      =   1.771

------------------------------------------------------------------------------
    HeadCirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Height |   .3110355   .0588286     5.29   0.000     .1929321    .4291388
       _cons |   35.64059   4.020475     8.86   0.000     27.56915    43.71203
------------------------------------------------------------------------------
```