NAME:_____

STATISTICS 110/201, FALL 2008, MIDTERM EXAM

Open book and notes, calculator required. You should have 4 pages – make sure you have them all. Each part of each problem is worth **7 points**, except where indicated.

The following lead-in applies to Questions 1 to 8:
The dataset for these questions was collected on the first day of this course, from most of you:
Y = hand span (in centimeters)
X_1 = height (in inches)
X_2 = a variable called "male" which is 1 for males and 0 for females
There were n = 30 usable observations.

Here are the Stata results for the full model (Y, X_1 and X_2 as defined above):

```
. regress handspan height male

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  2,    27) =   15.24
       Model |  79.1672586     2  39.5836293           Prob > F      =  0.0000
    Residual |   70.144734    27  2.59795311           R-squared     =  0.5302
-------------+------------------------------           Adj R-squared =  0.4954
       Total |  149.311993    29   5.1486894           Root MSE      =  1.6118

------------------------------------------------------------------------------
    handspan |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .2511856   .0860979     2.92   0.007     .0745274    .4278438
        male |   1.258293   .7838402     1.61   0.120     -.350014    2.866601
       _cons |   3.705192   5.651104     0.66   0.518    -7.889916     15.3003
------------------------------------------------------------------------------
```

1. Write the *population* version of the model. (Use X and Y notation rather than the variable names.)

$$E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \qquad \text{OR} \qquad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

2. Write the numerical equation (sample regression equation) that would be used to predict hand span for a *female* with height X_h.

$$\hat{Y}_i = 3.705 + .251 X_h$$ [Because it's a female, there is no X_2 term.]

3. Interpret in words the coefficient for "male" of 1.258.

The coefficient means that on average, for a male and female of the same height the male's handspan will be 1.258 cm more than the female's handspan. You could also say that the average handspan for males of a certain height is about 1.258 cm higher than the average handspan for females of that same height.

4. The first three heights in the sample were 66, 69.5 and 61 inches, and they were for a female, male, and female, respectively. Write the first 3 rows of the **X** matrix used for this analysis.

$$\begin{bmatrix} 1 & 66 & 0 \\ 1 & 69.5 & 1 \\ 1 & 61 & 0 \end{bmatrix}$$

5. Consider *males* who are 70 inches tall.

a. Find $\hat{Y}_h$ and show your work.

$$\hat{Y}_h = 3.705 + .251\ (70) + 1.258 = 22.533 \text{ centimeters}$$

b. [**8 points**] Give *two* interpretations for $\hat{Y}_h$, i.e. two ways in which it is used.

First, it is an estimate of the population mean hand span for all males who are 70 inches tall.

Second, it is the best single number prediction for the hand span for one male who is 70 inches tall.

6. Does the intercept have a useful interpretation in this situation? If so, give the interpretation. If not, explain why not.

The intercept is the predicted hand span measurement for a female with height of 0 inches. Since a height of 0 inches is not possible, and in fact there are no data anywhere near that value, the intercept does not have a useful interpretation in this situation.

The following information is to be used for Questions 7 and 8 , in conjunction with the Stata output provided on page 1.
Below are the Stata results from fitting the model with a single explanatory variable, X_1 = height. (The response variable is still hand span.) In other words, the indicator variable "male" was removed from the equation.

```
. regress handspan height

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =   26.41
       Model |  72.4724222     1  72.4724222           Prob > F      =  0.0000
    Residual |  76.8395704    28  2.74427037           R-squared     =  0.4854
-------------+------------------------------           Adj R-squared =  0.4670
       Total |  149.311993    29   5.1486894           Root MSE      =  1.6566

------------------------------------------------------------------------------
    handspan |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .3422062   .0665909     5.14   0.000     .2058009    .4786115
       _cons |  -1.921275   4.555989    -0.42   0.676     -11.2538    7.411246
------------------------------------------------------------------------------
```

7. [8 points] Using information for this model *only* (not from the model on page 1), give the numerical value for two different test statistics that can be used to test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. Identify where you found them on the Stata output. You don't need to actually carry out the test.

t = 5.14 and F* = 26.41. They are in bold in the output above.*

8. a. Continuing to define X_1 = height and X_2 = male (1 for male, 0 for female), write numerical values for the following, (using the output from the model above and the full model on page 1):
[**3 points each**]

$SSR(X_1)$ = __72.472__ $SSR(X_1, X_2)$ = __79.167____ $SSR(X_2| X_1)$ = __6.695__

(The last one is just a difference of the first two.)

b. Using the output from both models, construct the test statistic F* for the general linear test of H_0: $\beta_2 = 0$ versus H_a: $\beta_2 \neq 0$, and find the numerical value of the test statistic.

$$F^* = \frac{SSR(X_2 \mid X_1)}{MSE(X_1, X_2)} = \frac{6.695}{2.598} = 2.577$$

(Note that this gives the same information as the t test in the analysis on page 1;*
$\sqrt{F^*} = \sqrt{2.577}$ = 1.6053, which is the t statistic of 1.61 given for the "male" coefficient.*

c. [6 points]Give the numerator and denominator degrees of freedom for the test statistics in part b.
Numerator df = 1 and denominator df = n – p = 30 – 3 = 27.

9. For each of the following regression situations, state which would be more useful – a confidence interval for $E\{Y_h\}$ or a prediction interval for $Y_{h(new)}$. You don't need to explain your answer, unless you think it would help in a situation where you aren't sure. [**2 points each**]

a. For children from ages 3 to 15, a regression relationship is found between Y = child's height, and the two variables X_1 = child's age in months and X_2 = child's sex (1 for males and 0 for females). A doctor would like to know if a boy who is 72 months old is abnormally tall.

Prediction interval (interest is in a single individual).

b. A bank loans money to farmers who grow a certain crop. A regression relationship has been found between Y = crop yield for the summer and a set of X variables related to the size and health of the farm. The bank is trying to decide whether to make a loan to a particular farmer, and has access to the X values for that farm.

Prediction interval (interest is in that particular farm).

c. An insurance company has found the regression relationship between Y = age at death and a set of X variables related to lifestyle, education, etc. The amount they will pay out for life insurance policies depends on the policy holders' ages at death. They would like to estimate the average payout for a particular combination of the X variables.

Confidence interval (interest is in the average for the population with that combination of X variables).

10. (Fictitious data) The variables Y = resting pulse rate and X = water consumed during one day (in ounces) were measured for a random sample of 7 adults. A plot of the original data and residuals from a regression showed that a 1/X transformation was needed. The resulting regression equation is $\hat{Y} = 48.2 + 646\ X'$ where $X' = 1/X$.

a. Find the predicted resting pulse rate for someone who consumes 50 ounces of water in a day.

$\hat{Y} = 48.2 + 646\ X' = 48.2 + 646(1/50) = 61.12$, *or about 61 beats per minute.*

b. The test for $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ had a p-value of 0.000. Note that *higher* water consumption X corresponds to *lower* 1/X, and therefore *lower* predicted resting pulse. From this can we conclude that drinking more water causes one's resting pulse to go down? Explain.

No, because this was an observational study, not a randomized experiment. There are probably many other related variables that can explain why people with higher water intake have lower resting pulse rates. The most obvious is amount of exercise – more exercise would lead to needing more water, and also to lower resting pulse rate.