Statistics 110 and 201, Fall 2008
PRACTICE MIDTERM EXAM

Open book and notes. Calculator required. There are 5 problems, with a total of 14 parts. Each part of each problem (a, b, etc) is worth 7 points, except Problem 5a, which is worth 9 points.

1. Problem 1.28 on page 37 of your textbook describes data from 84 medium-sized counties in the US. For each county, X = percentage of adults in the county having at least a high-school diploma, and Y = crime rate (crimes reported per 100,000 residents) last year. Here is Stata output from fitting a simple linear regression model to the data:

```
regress CrimeRate PctHSDiploma

      Source |       SS       df       MS              Number of obs =      84
-------------+------------------------------           F(  1,    82) =   16.83
       Model |  93462942.3     1  93462942.3           Prob > F      =  0.0001
    Residual |   455273165    82  5552111.77           R-squared     =  0.1703
-------------+------------------------------           Adj R-squared =  0.1602
       Total |   548736108    83   6611278.4           Root MSE      =  2356.3

------------------------------------------------------------------------------
   CrimeRate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
PctHSDiploma |  -170.5752   41.57433    -4.10   0.000    -253.2798   -87.87061
       _cons |    20517.6   3277.643     6.26   0.000     13997.32    27037.88
------------------------------------------------------------------------------
```

a. Write the *population* version of the regression model.

b. Write the *estimated* (sample) regression function.

c. According to the last US Census, 82.7% of Orange County adults have a high school diploma. Round this number to 83%, and obtain a point estimate for the crime rate in Orange County.

d. Interpret the slope in the context of this situation.

Problem 1, continued…

e. A 95% confidence interval for $E\{Y_h\}$ when $X_h = 70$ is 7702 to 9453. Interpret this interval in words, in the context of this situation.

f. The results indicate that counties with higher percentages of high-school graduates tend to have lower crime rates. Can we conclude from this study that having a high school diploma causes people to be less likely to commit crimes, in other words, that higher high-school graduation rates cause crime to be lower? Explain your answer.

2. Define **I** to be an $n \times n$ identity matrix, and **H** to be the usual hat matrix. A matrix that plays a useful role in regression inference is $(\mathbf{I} - \mathbf{H})$. Show using matrix algebra that $(\mathbf{I} - \mathbf{H})$ is idempotent.

3. A company offers a training course for the Math SAT. They give their students a test at the end of the course, graded from 0 to 100. They would like to use that test in the future to predict how well students will score on the Math SAT. They have scores on their test and the Math SAT for a sample of students. Thus, X = score on the company's test and Y = score on the Math SAT. They plan to use the usual simple linear regression model.

a. Would the intercept have a useful meaning in this example? Explain your answer.

b. One of the company analysts states that the intercept should be fixed at 200, because that's the lowest the SAT Math score can be. Suppose the intercept is set to 200 for this situation. Write the population model.

c. Write the full and reduced models to test whether or not it makes sense to set the intercept to be 200.

d. Write the sum that is to be minimized to get the least squares regression line, if the model you wrote in Part b is used.

4. What assumption is being examined by looking at a normal probability plot? Be specific.

5. A regression equation is to be fit for predicting Y = resting pulse rate using the predictor variables X_1 = number of minutes of exercise per week and X_2 = gender, with 1 = male and 0 = female. Here are the X values results for 6 individuals:

Exercise/week	200	10	420	50	350	140
Gender	Male	Female	Female	Male	Male	Female

a. **(9 points)** Write down the **X** matrix that would be used for this situation, filling in numerical values.

b. Explain in words what the coefficient attached to X_2 represents.