First Midterm Exam
Economics 401
Tues., March 3, 2009

Show All Work. Only partial credit will be given for correct answers if we can not figure out how they were derived.

Points:

Problem 1:	25
Problem 2:	25
Problem 3:	25
Problem 4:	25
Total:	100

Problem 1: Consider the following output from stata:

```
. reg unem inf_1 unem_1 year

      Source |       SS       df       MS              Number of obs =      55
-------------+------------------------------           F(  3,    51) =   39.16
       Model |  84.9148875     3  28.3049625           Prob > F      =  0.0000
    Residual |  ??????????    51  .722866721           R-squared     =  ??????
-------------+------------------------------           Adj R-squared =  0.6795
       Total |   121.78109    54  2.25520538           Root MSE      =  .85022

------------------------------------------------------------------------------
        unem |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       inf_1 |   .1818597   .0391665     4.64   0.000     .1032296    .2604898
      unem_1 |    .639145   .0824823     7.75   0.000     .4735548    .8047351
        year |   .0031555   .0076661     0.41   0.682    -.0122348    .0185457
       _cons |  -4.876141   15.00256    -0.33   0.746    -34.99503    25.24275
------------------------------------------------------------------------------
```

This is a regression we looked at in class. We are predicting unemployment rate using lagged unemployment (unem_1) lagged inflation (inf_1) and the year.

a) What is the R^2 in this model (which is missing)?

$$\begin{aligned} R^2 &= \frac{SSE}{SST} \\ &= \frac{84.91}{121.78} \\ &= 0.6972 \end{aligned}$$

b) What is the residual sum of squares up to two decimal places (which is missing)?

$$\begin{aligned} SSR &= SST - SSE \\ &= 121.78 - 84.91 \\ &= 36.87 \end{aligned}$$

c) In 2002 the inflation rate was 1.6 and the unemployment rate was 5.8. Come up with a forecast for the unemployment rate in 2003.

$$\begin{aligned} \widehat{U}_{2003} &= \widehat{\beta}_0 + \widehat{\beta}_1 U\_1 + \widehat{\beta}_2 U\_1 + \widehat{\beta}_3 year \\ &= -4.87 + 0.18 \times 1.6 + 0.64 \times 5.8 + 0.0032 \times 2003 \\ &= 5.54 \end{aligned}$$

Problem 2: Suppose you have data on people's exercise habits ($exer_i$ measured in average minutes per day) and their age at death (age_i measured in years). Suppose now that you run a regression of age at death on exercise and find that

$$age_i = 65 + 0.1 exer_i + \hat{u}_i.$$

a) Thinking about the model in terms of a descriptive manner, explain what it means that the slope coefficient is 0.1.

Suppose I randomly grabbed two people from the data and it turned out that one exercised 1 minute per day more than the other. I would expect the person that exersised more to live for an extra one tenth of a year.

b) Thinking about the model in terms of a causal manner, explain what it means that the slope coefficient is 0.1.

If I exercise an extra minute per day longer, my life expectancy will go up by a tenth of a year.

c) What is the crucial assumption justifying the causal claim? Can you think of a potential omitted variable that would bias the result?

The crucial assumption is that $E(u_i \mid exer_i) = 0$. This means that there is no correlation between the amount I exercise and other factors related to the length of my life. This seems unreasonable as people who exercise more probably do other things related to living a healthy life style. One example of an omitted variable would be consumption of Cheeseburgers.

c) Do you think the omitted variable bias is negative or positive in this case (and please explain why)? If you got data on that variable and included that variable into the model, what would that do to the coefficient on exercise?

I know that the omitted variable bias is $\beta_2\delta_1$ where β_2 picks up the effect of cheeseburgers on length of life and δ_1 picks up the relationship between cheeseburgers and exersise. I would expect that the effect of cheeseburgers on length of life is negative which means that $\beta_2 < 0$. I expect that people who exercise more are probably more diet concious and eat fewer cheeseburgers so that $\delta_1 < 0$. Thus I would expect $\beta_2\delta_1$ to be positive meaning I expect the bias to be positive. This means that if I included Cheeseburgers in the regression, the bias should go away meaning that the effect should fall.

Problem 3: You have data from different cities on the following variables:

- p_i: price of housing in the city
- s_i: the size of the city (population)
- z_i: the zip code in the center of the city
- w_i: the average wealth level in the city

Consider the following 6 regressions:

$$p_i = a_0 + a_1 s_i + u_{ai} \tag{1}$$
$$p_i = b_0 + b_1 z_i + u_{bi} \tag{2}$$
$$p_i = c_0 + c_1 s_i + c_2 z_i + u_{ci} \tag{3}$$
$$p_i = d_0 + d_1 s_i + d_2 w_i + u_{di} \tag{4}$$
$$p_i = e_0 + e_1 z_i + e_2 w_i + u_{ei} \tag{5}$$
$$p_i = f_0 + f_1 s_i + f_2 z_i + f_3 w_i + u_{fi} \tag{6}$$

Think about the comparison between the R^2 in the 6 regressions. What can you say <u>for sure</u> about the relative values of R^2 in the different regressions? (You don't need to tell me the ones you can't tell apart, but please tell me about all the ones you can.)

We know that when we add a variable to a regression, the R^2 can not fall. However, we can not compare two regressions in which the bigger one does not nest the smaller one. Thus we know that

$$\begin{aligned}
(1) &\leq (3) \\
(1) &\leq (4) \\
(1) &\leq (6) \\
(2) &\leq (3) \\
(2) &\leq (5) \\
(3) &\leq (6) \\
(4) &\leq (6) \\
(5) &\leq (6)
\end{aligned}$$

Problem 4: Assume that the causal regression model takes the form:

$$Y_i = \beta_0^{\beta_1} + \beta_1 X_{1i} + (\beta_2 - \beta_0)X_{2i} + X_{1i}u_i.$$

Assume further that

$$E(u_i \mid X_{1i}, X_{2i}) = 0$$

I want you to explain how you would estimate β_0, β_1 and β_2. That is first come up with three population equations you could use. Then translate these to three equations to sample analogues that depend only on the data and the three unknown parameters.

We can just use the same population equations we have used all along:

$$\begin{aligned} E(u_i) &= 0 \\ E(x_{1i}u_i) &= 0 \\ E(x_{2i}u_i) &= 0 \end{aligned}$$

We write the sample regression function as

$$Y_i = \hat{\beta}_0^{\hat{\beta}_1} + \hat{\beta}_1 X_{1i} + (\hat{\beta}_2 - \hat{\beta}_0)X_{2i} + X_{1i}\hat{u}_i.$$

or

$$\hat{u}_i = \frac{Y_i - \hat{\beta}_0^{\hat{\beta}_1} - \hat{\beta}_1 X_{1i} - (\hat{\beta}_2 - \hat{\beta}_0)X_{2i}}{X_{1i}}.$$

Then we can write the sample analogues of the population equations as:

$$\begin{aligned} 0 &= \frac{1}{N}\sum_{i=1}^{N} \hat{u}_i \\ 0 &= \frac{1}{N}\sum_{i=1}^{N} X_{1i}\hat{u}_i \\ 0 &= \frac{1}{N}\sum_{i=1}^{N} X_{2i}\hat{u}_i \end{aligned}$$

which we can write as

$$\begin{aligned} 0 &= \frac{1}{N}\sum_{i=1}^{N} \frac{Y_i - \hat{\beta}_0^{\hat{\beta}_1} - \hat{\beta}_1 X_{1i} - (\hat{\beta}_2 - \hat{\beta}_0)X_{2i}}{X_{1i}} \\ 0 &= \frac{1}{N}\sum_{i=1}^{N} Y_i - \hat{\beta}_0^{\hat{\beta}_1} - \hat{\beta}_1 X_{1i} - (\hat{\beta}_2 - \hat{\beta}_0)X_{2i} \\ 0 &= \frac{1}{N}\sum_{i=1}^{N} X_{2i}\frac{\left(Y_i - \hat{\beta}_0^{\hat{\beta}_1} - \hat{\beta}_1 X_{1i} - (\hat{\beta}_2 - \hat{\beta}_0)X_{2i}\right)}{X_{1i}} \end{aligned}$$