Economics 3950
Spring 2008
Dr. Richard Mueller

Assignment #2

Instructions: These questions should be answered using a text editor or a word processor where you can cut and paste output from your statistical program (where necessary). Please mark question numbers clearly. This assignment is **due on Friday, February 29, 2008 by 12:00 in D-552.**

1. (40 points total) Exercise 3.38, pp. 132. If the model with POP and YEAR are similar in explaining the variation of GDP, explain why? Now regress GDP on both POP and YEAR. Comment on the standard errors on their respective estimated coefficients and explain why this has happened.

2. (30 points total) Exercise 4.9, p. 190.

3. (30 points total) Exercise 4.24, p. 204.

Grand Total: 100 points

Answer Key

1. The following is the estimated model, along with standard errors, etc.

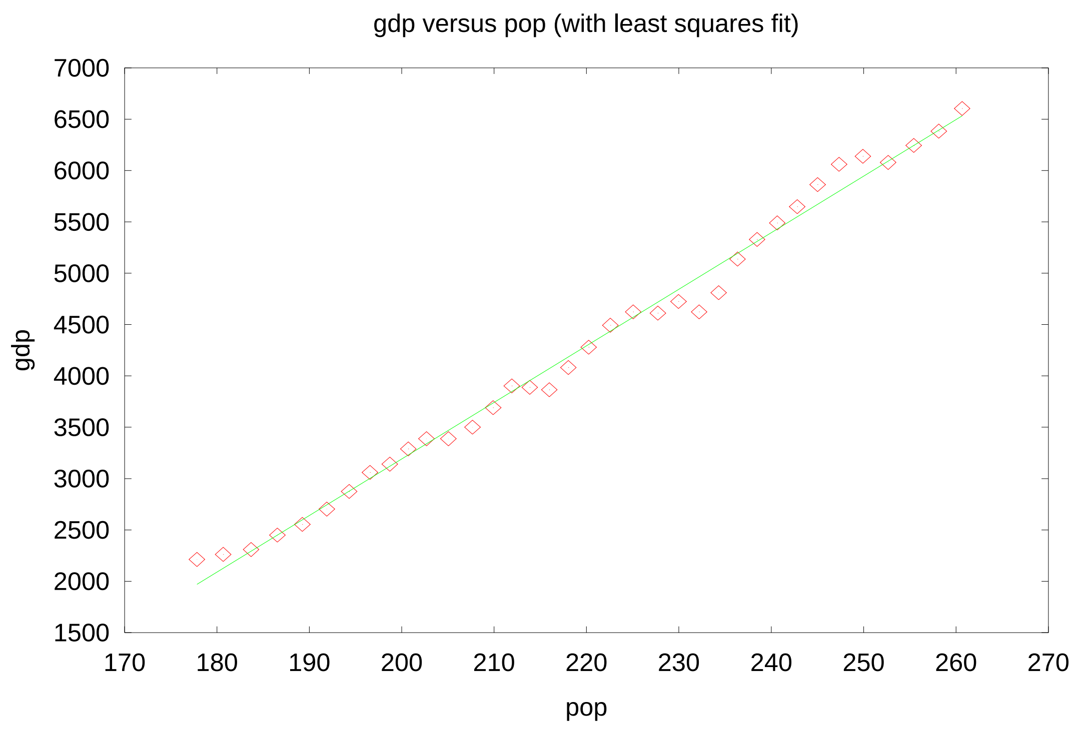
```
MODEL 1: OLS estimates using the 36 observations 1959-1994
Dependent variable: gdp

      VARIABLE      COEFFICIENT       STDERROR       T STAT    2Prob(t > |T|)

   0)    const       -7827.1909       199.4214      -39.250       0.000000 ***
   3)      pop          55.0917         0.9030       61.011       0.000000 ***

Mean of dep. var.           4269.528  S.D. of dep. variable         1330.370
Error Sum of Sq (ESS)         560686  Std Err of Resid. (sgmahat)   128.4164
Unadjusted R-squared           0.991  Adjusted R-squared               0.991
F-statistic (1, 34)           3722.4  p-value for F()               0.000000
Durbin-Watson stat.            0.591  First-order autocorr. coeff      0.654
```

The plot of GDP against POP is shown below. It looks like linear regression model should prove to be a pretty good fit for these data.



The following shows the second model (note that the R-squared is higher, but not by much, implying that this model does a slightly better job than the first):

```
MODEL 2: OLS estimates using the 36 observations 1959-1994
Dependent variable: gdp

      VARIABLE      COEFFICIENT      STDERROR       T STAT    2Prob(t > |T|)

   0)    const          -244328     3791.2313      -64.446       0.000000 ***
   1)     year         125.7769        1.9181       65.573       0.000000 ***

Mean of dep. var.           4269.528  S.D. of dep. variable          1330.370
Error Sum of Sq (ESS)         485987  Std Err of Resid. (sgmahat)    119.5564
Unadjusted R-squared           0.992  Adjusted R-squared                0.992
F-statistic (1, 34)          4299.78  p-value for F()                0.000000
Durbin-Watson stat.            0.607  First-order autocorr. coeff       0.682
```

The variables POP and TIME have a correlation coefficient close to 1. Thus, it's not surprising that they would both perform well in the above models.

If we multiplied GDP by 1000 (to put it into millions of dollars) and called it GDP2, we would get the following:

```
MODEL 3: OLS estimates using the 36 observations 1959-1994
Dependent variable: gdp2

      VARIABLE      COEFFICIENT      STDERROR       T STAT    2Prob(t > |T|)

   0)    const       -7.827e+006        199421      -39.250       0.000000 ***
   3)      pop       55091.6789      902.9723       61.011       0.000000 ***

Mean of dep. var.        4269527.778  S.D. of dep. variable       1330370.323
Error Sum of Sq (ESS)    5.6069e+011  Std Err of Resid. (sgmahat)      128416
Unadjusted R-squared           0.991  Adjusted R-squared                0.991
F-statistic (1, 34)           3722.4  p-value for F()                0.000000
Durbin-Watson stat.            0.591  First-order autocorr. coeff       0.654
```

Notice that the coefficient values and standard errors have also increased by 1000, but nothing else has changed.

If we include both POP and YEAR as regressors, notice what happens to the standard errors of each coefficient estimate (compared to Models 1 and 2 above):

```
MODEL 4: OLS estimates using the 36 observations 1959-1994
Dependent variable: gdp

      VARIABLE      COEFFICIENT      STDERROR       T STAT    2Prob(t > |T|)

   0)    const          -312618        131351       -2.380       0.023245 **
   1)     year         162.0968       69.8562        2.320       0.026644 **
   3)      pop         -15.9244       30.6164       -0.520       0.606449

Mean of dep. var.           4269.528  S.D. of dep. variable          1330.370
Error Sum of Sq (ESS)         482036  Std Err of Resid. (sgmahat)    120.8600
Unadjusted R-squared           0.992  Adjusted R-squared                0.992
F-statistic (2, 33)           2103.9  p-value for F()                0.000000
Durbin-Watson stat.            0.598  First-order autocorr. coeff       0.697
```

This is the result of multicollinearity between YEAR and POP.

2. a. Ho is that the coefficients for HSGPA, VSAT, and MSAT are all zero. Ha is that at least one of the coefficients is nonzero. The test statistic is given in Equation (4.4) as: Fc = (0.22/3)/(o.78/423) = 39.8. Under the null hypothesis this has an F-distribution with 3 and 423 df The critical F*(0.01) = 3.8 which is well below Fc. Therefore, we reject the null hypothesis and conclude that at least on of the regression coefficients is nonzero.

b. A single regression coefficient is tested with a t-test. The critical t (with 423 d.f.) is t*(0.01) = 2.33 (note that the alternative is one-sided). The t-statistics for the coefficients of constant, HSGPA, VSAT, and MSAT are obtained by dividing the corresponding regression coefficients by their standard errors. These values are 1.92, 6.52, 2.63, and 3.46. Except for the constant term, all the rest are above 2.33. Therefore, we conclude that all the coefficients are significant at the one per cent level with the exception of the constant term.

c. The expected average increase in COLGPA is 0.175.

d. Let the general unrestricted model (U) be:

$$COLGPA = \beta_1 + \beta_2 HSGPA + \beta_3 VSAT + \beta_4 MSAT + \mu$$

The marginal effect of VSAT is β_3 and the marginal effect of MSAT is β_4. The test is therefore $\beta_3 = \beta_4$. The alternative is that these two are unequal.

Method 1 (the Wald test): Assume this condition and obtain the restricted model (R) as:

$$COLGPA = \beta_1 + \beta_2 HSGPA + \beta_3 (VSAT + MSAT) + \mu$$

Generate the new variable Z = MSAT + VSAT. Now regress COLGPA on a constant, HSGPA and Z. Save the error sum of squares. The Wald F-statistic is given by Equation (4.3). We reject the null hypothesis, $\beta_3 = \beta_4$ if Fc > F*, where F* is obtained from F(1,423) such that the area to the right is equal to the level of significance.

Method 2 (an indirect t-test): Let $\beta = \beta_3 - \beta_4$ or $\beta_4 = \beta_3 - \beta$. The modified model is:

$$COLGPA = \beta_1 + \beta_2 HSGPA + \beta_3 VSAT + (\beta_3 - \beta) MSAT + \mu$$

and combining together the β_3 terms we get:

$$COLGPA = \beta_1 + \beta_2 HSGPA + \beta_3 Z - \beta MSAT + \mu$$

where Z is as defined above. The test is conducted by estimating the equation directly above and using the regular t-test on the coefficient β.

Method 3 (direct t-test): The variance of the estimated difference $\hat{\beta}_3 + \hat{\beta}_4$ is given by:

$$Var(\hat{\beta}_3) + Var(\hat{\beta}_4) - 2Cov(\hat{\beta}_3, \hat{\beta}_4)$$

Therefore, the computed t-statistic is:

$$t_c = \frac{\hat{\beta}_3 - \hat{\beta}_4}{[Var(\hat{\beta}_3) + Var(\hat{\beta}_4) - 2Cov(\hat{\beta}_3, \hat{\beta}_4)]^{1/2}}$$

For a two-tailed test, H_o is rejected if the numerical value of t_c exceed $t^*_{n-k}(\alpha/2)$.

e. The major of a student is an important determinant of the GPA because some disciplines are easier to get good grades. Also, perhaps a variable for institution would be useful, as would one indicating the number of hours per week worked during the semester. Recall, omitted variable bias causes biased and inconsistent estimates and thus tests of hypotheses are invalid.

3. *Note: This question can be answered two ways (although only one is correct). The first way assumes that we should have positive coefficients on the z-variables (since they are all productivity enhancing). The second (and correct way) assumes negative coefficients on the z-variables since, it ranges between 0 and 4 with 4 being the worst. Thus, the higher the z-variable, the less an individual should earn, hence negative coefficient values. Some of you caught this in the question, some did not. In fact, the instructor's manual that accompanies the text book also got it wrong. In either case, however, the best model is the same. You were not penalized as long as your answers remained consistent throughout. The following uses the correct interpretation of the z-variables.*

We would expect all of these variables to have positive coefficients since they are all human capital characteristics that should increase earnings.

```
MODEL 1: OLS estimates using the 25 observations 1-25
Dependent variable: slrygain

      VARIABLE      COEFFICIENT      STDERROR       T STAT    2Prob(t > |T|)

   0)    const          60.8988       22.5130        2.705       0.014495 **
   2)  tuition           0.3141        0.7497        0.419       0.680184
   3)       z1          -3.9483        2.7560       -1.433       0.169105
   4)       z2          -2.0158        2.1654       -0.931       0.364214
   5)       z3          -2.4017        2.9485       -0.815       0.425978
   6)       z4          -0.6125        3.0624       -0.200       0.843719
   7)       z5          -5.3250        3.7731       -1.411       0.175207

Mean of dep. var.              41.399  S.D. of dep. variable            9.981
Error Sum of Sq (ESS)       1288.2140  Std Err of Resid. (sgmahat)     8.4598
Unadjusted R-squared            0.461  Adjusted R-squared               0.282
F-statistic (6, 18)           2.56741  p-value for F()               0.056379
Durbin-Watson stat.             1.993  First-order autocorr. coeff     -0.022
```

Notice that some all of the variables have the correct signs. Also, the F-statistic has a p-value of 0.56379, indicating that we cannot reject the null that all coefficients are zero at the 5 per cent level of significant (although we can at the 10 per cent level).

Using the data-based model reduction procedure (in which the insignificant variables with the highest p-values are eliminated one at a time) we end up with the following estimates (if we stop where all coefficients are significant at at least the 10 per cent level):

```
MODEL 2: OLS estimates using the 25 observations 1-25
Dependent variable: slrygain
```

VARIABLE	COEFFICIENT	STDERROR	T STAT	2Prob(t > |T|)
0) const	62.4375	6.2315	10.020	0.000000 ***
3) z1	-6.3977	2.1036	-3.041	0.005991 ***
7) z5	-5.5160	2.6833	-2.056	0.051873 *

Mean of dep. var.	41.399	S.D. of dep. variable	9.981
Error Sum of Sq (ESS)	1494.5115	Std Err of Resid. (sgmahat)	8.2421
Unadjusted R-squared	0.375	Adjusted R-squared	0.318
F-statistic (2, 22)	6.59599	p-value for F()	0.005699
Durbin-Watson stat.	1.851	First-order autocorr. coeff	0.060

Or if we stop at the 5 per cent level we end up with:

MODEL 3: OLS estimates using the 25 observations 1-25
Dependent variable: slrygain

VARIABLE	COEFFICIENT	STDERROR	T STAT	2Prob(t > |T|)
0) const	52.9856	4.4913	11.797	0.000000 ***
3) z1	-6.2972	2.2457	-2.804	0.010073 **

Mean of dep. var.	41.399	S.D. of dep. variable	9.981
Error Sum of Sq (ESS)	1781.5793	Std Err of Resid. (sgmahat)	8.8011
Unadjusted R-squared	0.255	Adjusted R-squared	0.222
F-statistic (1, 23)	7.86334	p-value for F()	0.010073
Durbin-Watson stat.	1.795	First-order autocorr. coeff	0.071

In either case, the results are not really credible. Only the MBA skills in being analysts and (in the first case) the curriculum evaluation rating had significant effects. The model is probably misspecified, perhaps we could have included non-linear effects, or perhaps there are omitted variables (with the low R-squared values). Also, multicollinearity may be a problem (resulting in large standard errors and difficulty in coefficient interpretation).