BOSTON COLLEGE
Department of Economics
EC 228 02 Econometric Methods
Fall 2009, Prof. Baum, Ms. Phillips (TA), Ms. Pumphrey (grader)
Problem Set 4
Due Tuesday 27 October 2009
Total Points Possible: 120

Problem 4.7

(i) (3 points) In this regression, we are mainly concerned with *hrsemp*.While the standard error on *hrsemp* has not changed very much, the magnitude of the coefficient has increased from -.029 to -.042. The t statistic on *hrsemp* has gone from about 1.47 to 2.21, so now the coefficient is statistically less than zero at the 5% level. (From Table G.2 the 5% critical value with 40 *df* is 1.684. The 1% critical value is 2.423, so the p-value is between .01 and .05.) The R-squared coefficient has also increased.

(ii) (3 points) We can add and subtract $\beta_2 log(employ)$ from the right-hand-side and collect terms to get

$$log(scrap) = \beta_0 + \beta_1 hrsemp + [\beta_2 log(sales) - \beta_2 log(empl)] + [\beta_2 log(empl) + \beta_3 log(empl)] + u =$$

$$\beta_0 + \beta_1 hrsemp + \beta_2 log(sales/employ) + (\beta_2 + \beta_3) log(employ) + u$$

where the second equality follows from the fact that $log(sales/employ) = log(sales) - log(employ)$. Defining $\theta_3 \equiv \beta_2 + \beta_3$ gives the result. Interpreting the null of $\theta_3 = 0$ is equivalent to testing if $\beta_2 + \beta_3 = 0$ or that the two coefficients on log(*sales*) and log(*employ*) are of the same magnitude and opposite sign. From the regression output, we see that $\beta_2 = -.951$ and $\beta_3 = .992$. With their given standard errors, we cannot conclude that these two coefficients are of differing magnitudes.

(iii) (3 points) No. We are interested in the coefficient on $log(employ)$, which has a t statistic of .2, which is very small. While θ_3 is of the correct sign, it is not significantly different from zero. Therefore, we conclude that the size of the firm, as measured by employees, does not matter, once we control for training and sales per employee (in a logarithmic functional form).

(iv) (3 points) The null hypothesis in the model from part (ii) is $H_0 : \beta_2 = -1$. The t statistic is $[-.951 - (-1)]/.37 \approx .132$; this is very small, and we fail to reject regardless of whether we specify a one- or two-sided alternative.

Problem C4.5

(i) (4 points) If we drop *rbisyr* the estimated equation becomes

$$\begin{array}{lccccccc} \widehat{log(salary)} = & 11.02 & + & .0677 & years+ & .0158 & gamesyr \\ & (0.27) & & (.0121) & & (.0016) & \\ & & + & .0014 & bavg+ & .0359 & hrunsyr \\ & & & (.0011) & & (.0072) & \end{array}$$

$$n = 353, R^2 = .625.$$

Now *hrunsyr* is very statistically significant (t-statistic $\approx$. 4.99), and its coefficient has increased by about two and one-half times.

(ii) (4 points) The equation with *runsyr*, *fldperc*, and *sbasesyr* added is

$$\begin{array}{lccccccccc} \widehat{log(salary)} = & 10.41 & + & .0700 & years+ & .0079 & gamesyr & & \\ & (0.20) & & (.0120) & & (.0027) & & & \\ & & + & .00053 & bavg+ & .0232 & hrunsyr & & \\ & & & (.00110) & & (.0086) & & & \\ & & + & .0174 & runsyr+ & .0010 & fldperc \text{ -} & .0064 & sbasesyr \\ & & & (.0051) & & (.0020) & & (.0052) & \end{array}$$

$$n = 353, R^2 = .639.$$

Of the three additional independent variables, only runsyr is statistically significant (t-statistic $= .0174/.0051 \approx 3.41$). The estimate implies that one more run per year, other factors fixed, increases predicted salary by about 1.74%, a substantial increase. The stolen bases variable even has the "wrong" sign with a t-statistic of about -1.23, while *fldperc* has a t-statistic of only .5. Most major league baseball players are pretty good fielders; in fact, the smallest *fldperc* is 800 (which means .800). With relatively little variation in *fldperc*, it is perhaps not surprising that its effect is hard to estimate.

(iii) (4 points) From their t-statistics, *bavg*, *fldperc*, and *sbasesyr* are individually insignificant. The F-statistic for their joint significance (with 3 and 345 *df*) is about .69 with p-value $\approx$ .56. Therefore, these variables are jointly very insignificant.

Problem C4.9

(i) (3 points) The results from the OLS regression, with standard errors in parentheses, are

$$\begin{array}{lccccccccc} \widehat{log(psoda)} = & \text{-1.46} & + & .073 & prpblck\ + & .137 & log(income) & +.380 & prppov \\ & (0.29) & & (.031) & & (.027) & & (.133) & \end{array}$$

$$n = 401R^2 = .087.$$

The p-value for testing $H_0 : \beta_1 = 0$ against the two-sided alternative is about .018, so that we reject H_0 at the 5% level but not at the 1% level.

(ii) (3 points) The correlation is about -.84, indicating a strong degree of multicollinearity. Yet each coefficient is very statistically significant: the t statistic for $\hat{\beta}log(income)$ is about 5.1 and that for $\hat{\beta}prppov$ is about 2.86 (two-sided p-value = .004).

(iii) (3 points) The OLS regression results when $log(hseval)$ is added are

$$\begin{array}{lcccccc} \widehat{log(psoda)} = & \text{-.84} & + & .098 & prpblck\text{ -} & .053 & log(income) \\ & (0.29) & & (.029) & & (.038) & \\ & & + & .052 & prppov\ + & .121 & log(hseval) \\ & & & (.134) & & (.018) & \end{array}$$

$$n = 401R^2 = .184.$$

The coefficient on $log(hseval)$ is an elasticity: a one percent increase in housing value, holding the other variables fixed, increases the predicted price by about .12 percent. The two-sided p-value is zero to three decimal places.

(iv) (3 points) Adding $log(hseval)$ makes $log(income)$ and $prppov$ individually insignificant (at even the 15% significance level against a two-sided alternative for $log(income)$, and $prppov$ is does not have a t statistic even close to one in absolute value). Nevertheless, they are jointly significant at the 5% level because the outcome of the $F_{2,396}$ statistic is about 3.52 with p-value = .030. All of the control variables - $log(income)$, $prppov$, and $log(hseval)$ - are highly correlated, so it is not surprising that some are individually insignificant.

(v) (3 points) Because the regression in (iii) contains the most controls, $log(hseval)$ is individually significant, and $log(income)$ and $prppov$ are jointly significant, (iii) seems the most reliable. It holds fixed three measure of income and affluence. Therefore, a reasonable estimate is that if the proportion of blacks increases by .10, $psoda$ is estimated to increase by 1%, other factors held fixed.

Problem C6.2

(i) (3 points) The estimated equation is

$$\widehat{\log(wage)} = \underset{(.106)}{.128} + \underset{(.0075)}{.0904}\ educ + \underset{(.0052)}{.0410}\ exper - \underset{(.000116)}{.000714}\ exper^2$$

$$n = 526, R^2 = .300, \bar{R}^2 = .296$$

(ii) (3 points) The t-statistic on $exper^2$ is about 6.16, which has a p-value of essentially zero. So $exper^2$ is definitely significant at the 1% level.

(iii) (3 points) To estimate the return to the fifth year of experience, we start at $exper = 4$ and increase $exper$ by one, so $\Delta exper = 1$:

$$\%\Delta\widehat{wage} \approx 100[.0410 - 2(.000714)4] \approx 3.53\%$$

Similarly, for the 20th year of experience,

$$\%\Delta\widehat{wage} \approx 100[.0410 - 2(.000714)19] \approx 1.39\%$$

(iv) (3 points) The turnaround point is about $.041/[2(.000714)] \approx 28.7$ years of experience. In the sample, there are 121 people with at least 29 years of experience. This is a fairly sizeable fraction of the sample.

Problem C6.3

(i) (3 points) Holding $exper$ (and the elements in u) fixed, we have

$$\Delta\log(wage) = \beta_1\Delta educ + \beta_3\Delta educ \cdot exper = (\beta_1 + \beta_3 exper)\Delta educ,$$

or

$$\frac{\Delta\log(wage)}{\Delta educ} = (\beta_1 + \beta_3 exper)$$

This is the approximate proportionate change in wage given one more year of education.

(ii) (3 points) $H_0 : \beta_3 = 0$. If we think that education and experience interact positively so that people with more experience are more productive when given another year of education then $\beta_3 > 0$ is the appropriate alternative.

(iii) (6 points) The estimated equation is

$$\begin{array}{lllllllll} \widehat{\log(wage)} = & 5.95 & + & .0440 & educ- & .0215 & exper+ & .00320 & educ \cdot exper \\ & (.24) & & (.0174) & & (.0200) & & (.00153) & \end{array}$$

$$n = 935, R^2 = .135, \bar{R}^2 = .132$$

The t-statistic on the interaction term is about 2.13,which gives a p-value below .02 against $H_1 : \beta_3 > 0$. Therefore, we reject $H_0 : \beta_3 = 0$ at the 2 % level.

(iv) (3 points) We rewrite the equation as

$$\log(wage) = \beta_0 + \theta_1 educ + \beta_2 exper + \beta_3 educ(exper - 10) + u,$$

and run the regression $\log(wage)$ on $educ$, $exper$, and $educ(exper - 10)$. We want the coefficient on $educ$. We obtain $\hat{\theta}_1 \approx .0761$ and $se\left(\hat{\theta}_1\right) \approx .0066$. The 95 % CI for θ_1 is about .063 to .089.

Problem C6.8

(i) (3 points) The estimated equation (where $price$ is in dollars) is

$$\begin{array}{lllllllll} \widehat{price} = & -21,770.3 & + & 2.068 & lotsize+ & 122.78 & sqrft+ & 13,852.5 & bdrms \\ & (29,475.0) & & (0.642) & & (13.24) & & (9,010.1) & \end{array}$$

$$n = 88, R^2 = .672, \bar{R}^2 = .661, \hat{\sigma} = 59,833$$

The predicted price at $lotsize = 10,000$, $sqrft = 2,300$, and $bdrms = 4$ is about \$336,714.

(ii) (3 points) The regression is $price_i$ on $(lotsize_i 10,000)$, $(sqrft_i 2,300)$, and $(bdrms_i 4)$. We want the intercept estimate and the associated 95% CI from this regression. The CI is approximately $336,706.7 \pm 14,665$, or about \$322,042 to \$351,372 when rounded to the nearest dollar.

(iii) (6 points) We must use equation (6.36) to obtain the standard error of $\hat{e}^0$ and then use equation (6.37) (assuming that *price* is normally distributed). But from the regression in part (ii), $se(\hat{y}^0) \approx 7,374.5$ and $\hat{\sigma} \approx 59,833$. Therefore, $se\,(\hat{e}^0) \approx [(7,374.5)^2 + (59,833)^2]^{1/2} \approx 60,285.8$. Using 1.99 as the approximate 97.5^{th} percentile in the t_84 distribution gives the 95% CI for $price^0$, at the given values of the explanatory variables, as $336,706.7 \pm 1.99(60,285.8)$ or, rounded to the nearest dollar, \$216,738 to \$456,675. This is a fairly wide prediction interval. But we have not used many factors to explain housing price. If we had more factors included in the regression, we could presumably reduce the error standard deviation, and therefore $\hat{\sigma}$, to obtain a narrower prediction interval.

Problem C7.2

(i) (6 points) The estimated equation is

$$\begin{array}{ccccccccc}
\widehat{\log(wage)} = & 5.40 & + & .0654 & educ+ & .0140 & exper+ & .0117 & tenure \\
 & (.11) & & (.0063) & & (.0032) & & (.0025) & \\
+ & .199 & married- & .188 & black- & .091 & south+ & .184 & urban \\
 & (.039) & & (.038) & & (.026) & & (.027) &
\end{array}$$

$$n = 935, R^2 = .253.$$

The coefficient on *black* implies that, at given levels of the other explanatory variables, black men earn about 18.8 % less than nonblack men. The t-statistic is about 4.95, and so it is very statistically significant.

(ii) (4 points) The F-statistic for joint significance of $exper^2$ and $tenure^2$, with 2 and 925 *df*, is about 1.49 with p-value $\approx .226$. Because the p-value is above .20, these quadratics are jointly insignificant at the 20 % level.

(iii) (6 points) We add the interaction $black \cdot educ$ to the equation in part (i). The coefficient on the interaction is about -.0226 (se $\approx$ .0202). Therefore, the point estimate is that the return to another year of education is about 2.3 percentage points lower for black men than nonblack men. (The estimated return for nonblack men is about 6.7 %.) This is nontrivial if it really reflects differences in the population. But the t statistic is only about 1.12 in absolute value, which is not enough to reject the null hypothesis that the return to education does not depend on race.

(iv) (6 points) We choose the base group to be single, nonblack. Then we add dummy variables *marrnonblck*, *singblck*, and *marrblck* for the other three groups. The result is

$$\begin{array}{rlllllll}
\widehat{\log(wage)} = & 5.40 & +.0655 & educ+ & .0141 & exper+ & .0117 & tenure \\
 & (.11) & (.0063) & & (.0032) & & (.0025) & \\
 & & -.092 & south+ & .184 & urban+ & .189 & marrnonblck \\
 & & (.026) & & (.027) & & (.043) & \\
 & & -.241 & singblck+ & .0094 & marrblck & & \\
 & & (.096) & & (.0560) & & &
\end{array}$$

$$n = 935, R^2 = .253.$$

We obtain the ceteris paribus differential between married blacks and married nonblacks by taking the difference of their coefficients: .0094 - .189 = -.1796, or about -.18. That is, a married black man earns about 18 % less than a comparable, married nonblack man.

Problem C7.6

(i) (6 points) The estimated equation for men is

$$\begin{array}{rllllll}
\widehat{sleep} = & 3,648.2 & - & .182 & totwrk- & 13.05 & educ \\
 & (310.0) & & (.024) & & (7.41) & \\
+ & 7.16 & age- & .0448 & age^2+ & 60.38 & yngkid \\
 & (14.32) & & (.1684) & & (59.02) &
\end{array}$$

$$n = 400, R^2 = .156$$

The estimated equation for women is

$$\begin{array}{rllllll}
\widehat{sleep} = & 4,238.7 & - & .140 & totwrk- & 10.21 & educ \\
 & (384.9) & & (.028) & & (9.59) & \\
- & 30.36 & age- & .368 & age^2- & 118.28 & yngkid \\
 & (18.53) & & (.223) & & (93.19) &
\end{array}$$

$$n = 306, R^2 = .098$$

There are certainly notable differences in the point estimates. For example, having a young child in the household leads to less sleep for women (about two hours a week) while men are estimated to sleep about an hour more. The quadratic in *age* is a hump-shape for men but a U-shape for women. The intercepts for men and women are also notably different.

(ii) (4 points) The F statistic (with 6 and 694 df) is about 2.12 with *p-value* $\approx$ .05, and so we reject the null that the sleep equations are the same at the 5 % level.

(iii) (4 points) If we leave the coefficient on *male* unspecified under H_0, and test only the five interaction terms, $male \cdot totwork$, $male \cdot educ$, $male \cdot age$, $male \cdot age^2$, and $male \cdot yngkid$, the F statistic (with 5 and 694 df) is about 1.26 and *p-value* $\approx$ .28.

(iv) (6 points) The outcome of the test in part (iii) shows that, once an intercept difference is allowed, there is not strong evidence of slope differences between men and women. This is one of those cases where the practically important differences in estimates for women and men in part (i) do not translate into statistically significant differences. We need a larger sample size to confidently determine whether there are differences in slopes. For the purposes of studying the sleep-work tradeoff, the original model with *male* added as an explanatory variable seems sufficient.