Solutions to Problem Set 5 (Due November 22)

EC 228 02, Fall 2010 Prof. Baum, Ms Hristakeva

Maximum number of points for Problem set 5 is: 220

Problem 7.3

(i) (5 points) The t statistic on $hsize^2$ is over four in absolute value, so there is very strong evidence that it belongs in the equation. We obtain this by finding the turnaround point; this is the value of $hsize$ that maximizes $\widehat{sat}$ (other things fixed): $19.3/(22.19) \approx 4.41$. Because $hsize$ is measured in hundreds, the optimal size of graduating class is about 441.

(ii) (5 points) This is given by the coefficient on $female$ (since $black = 0$): nonblack females have SAT scores about 45 points lower than nonblack males. The t statistic is about 10.51, so the difference is very statistically significant. (The very large sample size certainly contributes to the statistical significance.)

(iii) (5 points) Because $female = 0$, the coefficient on $black$ implies that a black male has an estimated SAT score almost 170 points less than a comparable nonblack male. The t statistic is over 13 in absolute value, so we easily reject the hypothesis that there is no ceteris paribus difference.

(iv) (5 points) We plug in $black = 1$, $female = 1$ for black females and $black = 0$ and $female = 1$ for nonblack females. The difference is therefore $169.81 + 62.31 = 107.50$. Because the estimate depends on two coefficients, we cannot construct a t statistic from the information given. The easiest approach is to define dummy variables for three of the four race/gender categories and choose nonblack females as the base group. We can then obtain the t statistic we want as the coefficient on the black female dummy variable.

Problem 7.4

(i) (5 points) The approximate difference is just the coefficient on $utility$ times 100, or 28.3percent. The t statistic is $.283/.099 \approx 2.86$, which is very statistically significant.

(ii) (5 points) $100[e^{(.283)}1) \approx 24.7$, and so the estimate is somewhat smaller in magnitude.

(iii) (5 points) The proportionate difference is .181.158 = .023, or about 2.3 percent. One equation that can be estimated to obtain the standard error of this difference is

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \delta_1 consprod + \delta_2 utility + \delta_3 trans + u$$

where *trans* is a dummy variable for the transportation industry. Now, the base group is $finance$, and so the coefficient δ_1 directly measures the difference between the consumer products and finance industries, and we can use the t statistic on *consprod*.

Problem C7.10

(i) (p5ts) The estimated equation is

$$\begin{array}{lccccccccc} points = & 4.76 & + & 1.28 & exper - & .072 & exper^2 + & 2.31 & guard + & 1.54 \quad forward \\ & (1.18) & & (.33) & & (.024) & & (1.00) & & (1.00) \end{array}$$

$$n = 269, R^2 = .091, \bar{R}^2 = .077$$

(ii) (5 pts) Including all three position dummy variables would be redundant and would induce perfect collinearity, a violation of OLS assumptions. Each player falls into one of the three categories, and the overall intercept is the intercept for centers. The coefficients on the other positions are the points per game (PPG) relative to PPG of centers.

(iii) (5 pts) A guard is estimated to score about 2.3 points more per game, holding experience fixed. The t statistic is 2.31, so the difference is statistically different from zero at the 5% level, against a two-sided alternative.

(iv) (5 pts) When *marr* is added to the regression, its coefficient is about .584 (se=.740). Therefore, a married player is estimated to score just over half a point more per game (experience and position held fixed), but the estimate is not statistically different from zero (p-value=.43). So, based on points per game, we cannot conclude married players are more productive.

(v) (5 pts) Adding the terms $marr \cdot exper$ and $marr \cdot exper^2$ leads to complicated signs on the three terms involving *marr*. The F test for the joint significance, with 3 and 261 df, gives F= 1.44 and p-value=.23. Therefore, there is not very strong evidence that marital status has any partial effect on points scored, even at the 20 percent significance level.

(vi) (5 pts) If in the regression from part (iv) we use *assists* as the dependent variable, the coefficient on *marr* becomes .322 (se=.222). Therefore, holding experience and position fixed, a married man has almost one-third more assist per game. The p-value against a two-sided alternative is about .15, which is stronger, but not overwhelming, evidence that married men are more productive when it comes to assists.

Problem C7.12

(i) (5 points.) For women, the fraction rated as having above average looks is about .33; for men, it is .29. The proportion of women rated as having below average looks is only .135; for men, it is even lower at about .117.

(ii) (5 points.) The difference is about .04, that is, the percent rated as having above average looks is about four percentage points higher for women than men. A simple way to test whether the difference is statistically significant is to run a simple regression of *abvavg* on $female$ and do a t test (which is asymptotically valid). The t statistic is about 1.48 with two-sided $p - value = .14$. Therefore, there is not strong evidence against the null that the population fractions are the same, but there is some evidence.

(iii) (10 points.) The regression for men is:

$$\widehat{\log(wage)} = \underset{(0.024)}{1.884} + \underset{(.060)}{.199}\ belavg + \underset{(.042)}{-0.044}\ abvavg$$

$$n = 824, R^2 = .013$$

and the regression for women is

$$\widehat{\log(wage)} = \underset{(0.034)}{1.309} + \underset{(.076)}{.138}\ belavg + \underset{(.055)}{-0.034}\ abvavg$$

$$n = 436, R^2 = .011$$

Using the standard approximation, a man with below average looks earns almost 20percent less than a man of average looks, and a woman with below average looks earns about 13.8 percent less than a woman with average looks. (The more accurate estimates are about 18percent and 12.9 percent, respectively.) The null hypothesis H_0: $\beta_1 = 0$ against H_1: $\beta_1 < 0$ means that the null is that people with below average looks earn the same, on average, as people with average looks; the alternative is that people with below average looks earn less than people with average looks (in the population). The one-sided p-value for men is .0005 and for women it is .036. We reject H_0 more strongly for men because the estimate is larger in magnitude and the estimate has less sampling variation (as measured by the standard error).

(iv) (5 points.) Women with above average looks are estimated to earn about 3.4 percent more, on average, than women with average looks. But the one-sided p-value is .272, and this provides very little evidence against H_0: $\beta_2 = 0$.

(v) (5 points) Given the number of added controls, with many of them very statistically significant, the coefficients on the looks variables do not change by much. For men, the coefficient on *belavg* becomes .143 ($t = 2.80$) and the coefficient on *abvavg* becomes .001 ($t = .03$). For women, the changes in magnitude are similar: the coefficient on *belavg* becomes .115 ($t = 1.75$) and the coefficient on *abvavg* becomes .058 ($t = 1.18$). In both cases, the estimates on belavg move closer to zero but are still reasonably large.

Problem C8.2

(i) (10 pts) The estimated equation with both sets of standard errors (heteroskedasticity-robust standard errors in brackets) is

$$\begin{array}{lccccccccc} price = & -21.77 & + & .00207 & lotsize + & .123 & sqrft + & 13.85 & bdrms \\ & (29.48) & & (.00064) & & (.013) & & (9.01) & \\ & [37.13] & & [.00125] & & [.017] & & [8.48] & \end{array}$$

$$n = 88, R^2 = .672$$

The robust standard error on *lotsize* is almost twice as large as the usual standard error, making *lotsize* much less significant (the t statistic falls from about 3.23 to 1.70). The t statistic on *sqrft* also falls, but it is still very significant. The variable *bdrms* actually becomes somewhat more significant, but it is still barely significant. The most important change is in the significance of *lotsize*.

(ii) (10 pts) For the log-log model,

$$\begin{array}{lccccccccc} \widehat{\log(price)} = & -1.30 & + & .168 & \log(lotsize) + & .700 & \log(sqrft) + & .037 & bdrms \\ & (0.65) & & (.038) & & (.093) & & (.028) & \\ & [.78] & & [.041] & & [.103] & & [.030] & \end{array}$$

$$n = 54, R^2 = .643$$

Here, the heteroskedasticity-robust error is always slightly greater than the corresponding usual standard error, but the differences are relatively small. In particular, $\log(lotsize)$ and $\log(sqrft)$ still have very large t statistics, and the t statistic on bdrms is not significant at the 5% level against a one-sided alternative using either standard error.

(iii) (5 pts) As we discussed in Section 6.2, using the logarithmic transformation of the dependent variable often mitigates, if not entirely eliminates, heteroskedasticity. Taking log transformations allow you to interpret the coefficients as elasticities which don't depend on units. This is certainly the case here, as no important conclusions in the model for $\log(price)$ depend on the choice of the standard error. (We have also transformed two of the independent variables to make the model of the constant elasticity variety in *lotsize* and *sqrft*.

Problem C8.4

(i) (10 pts) The estimated equation is

$$\begin{array}{lccccccccc} voteA = & 37.66 & + & .252 & prtystrA + & 3.793 & democA + & 5.779 & \log(expendA) \\ & (4.74) & & (.071) & & (1.407) & & (.392) & \\ & & - & 6.238 & \log(expendB) & +\hat{u} & & & \\ & & & (.397) & & & & & \end{array}$$

$$n = 173, R^2 = .801, \bar{R}^2 = .796.$$

You can convince yourself that regressing the $\hat{u}_i$ on all of the explanatory variables yields an R-squared of zero, although it might not be exactly zero in your computer output due to rounding error. Remember, OLS works by choosing the estimates, $\hat{\beta}_j$, such that the residuals are uncorrelated in the sample with each independent variable (and the residuals have a zero sample average, too).

(ii) (5 pts) The Breusch-Pagan test entails regressing the $\hat{u}_i^2$ on the independent variables in part (i). The F-statistic for joint significant (with 4 and 168 *df*) is about 2.33 with p-value $\approx .058$. Therefore, there is some evidence of heteroskedasticity, but not quite at the 5 % level. However, we can reject the null of homoskedasticity at the 10 percent level.

(iii) (5 pts) Now we regress $\hat{u}_i^2$ on $\widehat{voteA}_i$ and $(\widehat{voteA}_i)^2$, where the $\widehat{voteA}_i$ are the OLS fitted values from part (i). The F-test, with 2 and 170 *df*, is about 2.79 with p-value $\approx .065$. This is slightly less evidence of heteroskedasticity than provided by the Breusch-Pagan test, but the conclusion is very similar: not significant at the 5 % level but significant at the 10 % level.

Problem C9.3

(i) (5 pts) If the grants were awarded to firms based on firm or worker characteristics, *grant* could easily be correlated with such factors that affect productivity. In the simple regression model, these are contained in u.

(ii) (5 pts) The simple regression estimates using the 1988 data are

$$\widehat{\log(scrap)} = \underset{}{.409} + \underset{(.406)}{.057}\ grant$$
$$\underset{(.241)}{}$$

$$n = 54, R^2 = .0004.$$

The coefficient on *grant* is actually positive, but not statistically different from zero.

(iii) (10 pts) When we add $\log(scrap_{87})$ to the equation, we obtain

$$\widehat{\log(scrap_{88})} = \underset{(.089)}{.021} - \underset{(.147)}{.254}\ grant_{88} + \underset{(.044)}{.831}\ \log(scrap_{87})$$

$$n = 54, R^2 = .873,$$

where the year subscripts are for clarity. The coefficient on grant is -.254 meaning that firms which received job training grants in 1988 had lower scrap rates in 1988. The t-statistic for $H_0 : \beta_{grant} = 0$ is $-.254/.147 \approx -1.73$. We use the 5 % critical value for 40 *df* in Table G.2: -1.68. Because $t = -1.73 < -1.68$, we reject H_0 in favor of $H_1 : \beta_{grant} < 0$ at the 5 % level.

(iv) (5 pts) The t-statistic is $(.831 - 1)/.044 \approx -3.84$, which is a strong rejection of H_0.

(v) (5 pts) With the heteroskedasticity-robust standard errors, the t-statistic for $grant_{88}$ is $-.254/.142 \approx -1.79$, so the coefficient is even more significantly less than zero when we use the heteroskedasticity-robust standard error. The t-statistic for $H_0 : \beta_{\log(scrap_{87})} = 1$ is $(.831 - 1)/.0735 \approx -2.29$, which is notably smaller than before, but it is still significant.

Problem C9.4

(i) (10 pts) Adding DC to the regression in equation (9.37) gives

$$\begin{array}{rccccccc} \widehat{infmort}) = & 23.95 & - & .567 & \log(pcinc) - & 2.74 & \log(physic) \\ & (12.42) & & (1.641) & & (1.19) & \\ & & + & .629 & \log(popul) + & 16.03 & DC \\ & & & (.191) & & (1.77) & \end{array}$$

$$n = 51, R^2 = .691, \bar{R}^2 = .664.$$

The coefficient on DC means that even if there was a state that had the same per capita income, per capita physicians, and population as Washington D.C., we predict that D.C. has an infant mortality rate that is about 16 deaths per 1000 live births higher. This is a very large "D.C. effect."

(ii) (10 pts) In the regression from part (i), the intercept and all slope coefficients, along with their standard errors, are identical to those in equation (9.38), which simply excludes D.C. [Of course, equation (9.38) does not have DC in it, so we have nothing to compare with its coefficient and standard error.] Therefore, for the purposes of obtaining the effects and statistical significance of the other explanatory variables, including a dummy variable for a single observation is identical to just dropping that observation when doing the estimation. The R-squareds and adjusted R-squareds from (9.38) and the regression in part (i) are not the same. They are much larger when DC is included as an explanatory variable because we are predicting the infant mortality rate perfectly for D.C. You might want to confirm that the residual for the observation corresponding to D.C. is identically zero.

Problem C9.8

(i) (5 pts) Use "summarize *stotal*" to see that its mean is .0474 and its standard deviation is .853.

(ii) (5 pts) Use "regress *stotal jc*" to see that the 95% confidence interval for *jc* includes zero, or use "corr *jc stotal*" to see that *jc* explains only 1.24% of the variation in *stotal*. Running the same commands for *univ*, we see that *univ* is positively statistically related to *stotal* with a p-val of 0.00 and that *univ* can explain 43.46 % of the variation in *stotal*. So only *univ* is statistically related to *stotal*.

(iii) (5 pts) Adding *stotal* to the regression in equation (4.17) gives

$$\widehat{\log(wage)} = \underset{(.021)}{1.495} + \underset{(.0068)}{.063}\,jc + \underset{(.0026)}{.069}\,univ + \underset{(.0002)}{.005}\,exper + \underset{(.0068)}{.049}\,stotal$$

$$n = 6763, R^2 = .228,$$

Then we can use "test *jc* = *univ*" where the null is that $\beta_1 = \beta_2$. We get an F-statistic with 1 and 6758 dfs and a p-val of .4205, so we fail to reject the null. In section 4.4 the p-val was about .07 so we could reject at the 10% level that the return to junior college was equal to the return of four-year college.

(iv) (5 pts) We generate the variable $stotal^2 = stotal * stotal$ and then run the regression from part (iii). The coefficient estimates are nearly identical for all of the variables as in part (iii) and the coefficient on $stotal^2$ has a p-val of .68, meaning it is pretty much insignificant. Thus, we don't seem to need it in the model. We also get a slightly lower adjusted R^2 which is also suggestive of the fact that adding the variable seems unnecessary.

(v) (5 pts) We generate the interaction terms $stotaljc = stotal * jc$ and $stotaluniv = stotal * univ$ then run the regression from part (iii) again. Then we test the joint significance of the interaction terms using the command "test *stotaljc stotaluniv*" and get an F-statistic with 2 and 6756 dfs with a p-val of .1410, meaning these interaction terms are jointly significant only at the 15 % level.

(vi) (5 pts) I would use the regression from part (iii) since we showed that the quadratic and interaction terms were not jointly significant. You could also use the variable totcoll instead of jc and univ since we showed that we cannot reject that they are different.