Solutions to Problem Set 4 (Due April 1)

EC 2228 03, Spring 2015 Prof. Baum, Mr. Zhang

Maximum number of points for Problem set 4 is: 66

Problem C 6.1

(i) (3 pts.) If the presence of the incinerator depresses housing prices, then $\beta_1 > 0$: all else equal, the further away from the incinerator, the higher housing prices are. The estimated equation is:

$$\widehat{\log(price)} = \underset{(0.65)}{8.05} + \underset{(0.066)}{0.365}\ \log(dist)$$

$$n = 142, R^2 = 0.180, \bar{R}^2 = 0.174,$$

which means a 1% increase in distance from the incinerator is associated with a predicted price that is about 0.365% higher.

(ii) (1 pt.) When the variables are added to the regression, the coefficient on $\log(dist)$ becomes about 0.055 ($se \approx 0.058$). The effect is much smaller now, and statistically insignificant. This is because we have explicitly controlled for several other factors that determine the quality of a house and its location. This is consistent with the hypothesis that the incinerator was located near less desirable houses to begin with.

(iii) (2 pts.) When $[\log(intst)]^2$ is added to the regression in part (ii), the coefficient on the $\log(dist)$ is now very significant, with a t statistic of about three. The coefficients on $\log(intst)$ and $[\log(intst)]^2$ are both significant. Just adding $[\log(intst)]^2$ has had a very big effect on the coefficients important for policy purposes. This means that distance from the incinerator and distance from the interstate are correlated in some nonlinear way that also affects housing prices.

(iv) (1 pt.) When added to the regression in part (iii), the coefficient on $[\log(dist)]^2$ is about -0.0365, t statistic is only about -0.33. Therefore, it is not necessary to add this variable.

Problem C 6.6

(i) (2 pts.) The partial effect of $expendB$ on $voteA$ is $\beta_3 + \beta_4 expendA$; The partial effect of $expendA$ on $voteA$ is $\beta_2 + \beta_4 expendB$; But the sign of β_4 is ambiguous: Is the effect of more spending by B smaller or larger for higher levels of spending by A?

(ii) (2 pts.) The estimated equation is:

$$\widehat{voteA} = \underset{(4.59)}{32.12} \underset{(0.088)}{+0.342}\ prtystrA \underset{(0.0050)}{+0.0383}\ expendA \underset{(0.0046)}{-0.0317}\ expendB \underset{(0.0000072)}{-0.0000066}\ expendAex$$

$$n = 173, R^2 = 0.571, \bar{R}^2 = 0.561.$$

The interaction term is not statistically significant.

(iii) (1 pt.) The average value of $expendA$ is about 310.61, or \$310,610. If we set $expendA$ at 300, which is close to the average value, we have

$$\Delta\widehat{voteA} = [-0.0317 - 0.0000066(300)]\Delta expendB \approx -0.0337(\Delta expendB).$$

So when $\Delta expendB = 100$, $\Delta\widehat{voteA} \approx -3.37$, which is a fairly large effect.

(iv) (1 pt.) 3.76. This does make sense, and it is a nontrivial effect.

(v) (2 pts.) The new estimated equation is:

$$\begin{array}{lccccccccc} \widehat{voteA} = & 18.20 & +0.157 & prtystrA & -0.0067 & expendA & +0.0043 & expendB & +0.494 & shareA \\ & (2.57) & (0.050) & & (0.0028) & & (0.0026) & & (0.025) & \end{array}$$

$$n = 173, R^2 = 0.868, \bar{R}^2 = 0.865.$$

When holding both $expendA$ and $expendB$ fixed, there is no way to change $shareA$.

(vi) (3 pts.) Generally we have

$$\frac{\partial\widehat{voteA}}{\partial expendB} = \hat{\beta}_3 + \hat{\beta}_4\left(\frac{\partial shareA}{\partial expendB}\right),$$

where $shareA = 100[expendA/(expendA + expendB)]$. Now,

$$\frac{\partial shareA}{\partial expendB} = -100\left(\frac{expendA}{(expendA + expendB)^2}\right).$$

Evaluated at $expendA = 300$ and $expendB = 0$, the above partial derivative is -1/3, therefore

$$\frac{\partial\widehat{voteA}}{\partial expendB} = \hat{\beta}_3 + \hat{\beta}_4(-1/3) \approx -0.164.$$

So $\widehat{voteA}$ falls by 0.164 percentage points given the first thousand dollows of spending by condidate B, where A's spending is held fixed at 300. This is a fairly large effect, although it may not be the most typical scenario. The effect tapers off as $expendB$ grows.

Problem C 6.9

(i) (3 pts.) The estimated equation is

$$\begin{array}{lccccccccc} \widehat{points} = & 35.22 & +2.364 & exper & -0.077 & exper^2 & -1.074 & age & -1.286 & coll \\ & (6.99) & (0.405) & & (0.0235) & & (0.295) & & (0.451) & \end{array}$$

$$n = 269, R^2 = 0.141, \bar{R}^2 = 0.128.$$

(ii) (2 pts.) Take derivative with respect to $exper$. The turnaround point is $2.364/[2(0.077)] \approx 15.35$. So the increase from 15 to 16 years of experience would actually reduce points-per-game. It makes sense when plays are getting older their performance decreases. But notice that only two players in the sample of 269 have more than 15 years of experience.

(iii) (1 pt.) Many of the most promising players leave college early, or even no college, to play in the NBA. Thus the less years you stay in college means the better you play, hence *coll* has a significant negative effect on points-per-game.

(iv) (1 pt.) When age^2 is added, its coefficient is 0.0536 (se=0.0492). Its t statistic is barely above one, so we are justified to drop it.

(v) (2 pts.) The estimated equation is

$$\begin{array}{rccccccccccc} \widehat{\log(wage)} = & 6.78 & 0.078 & points & +0.218 & exper & -0.0071 & exper^2 & -0.048 & age & -0.040 & coll \\ & (0.85) & (0.007) & & (0.05) & & (0.0028) & & (0.035) & & (0.053) & \end{array}$$

$$n = 269, R^2 = 0.488, \bar{R}^2 = 0.478.$$

(vi) (2 pts.) *test age coll* The F statistic is about 1.19. With 2 and 263 *df*, this gives a p-value of roughly 0.31. Therefore, once productivity and seniority are accounted for, there is no evidence for wage differentials depending on age or years played in college.

Problem 7.3

(i) (1 pt.) The t statistic for $hsize^2$ is over four in absolute value, so there is very strong evidence that it belongs in the equation.

(ii) (2 pts.) The first difference 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 very statistically significant.

(iii) (2 pts.) Similar as above. $female = 0$. The difference is the coefficient on *black*, 170 points higher for nonblack male than black male. The t statistic is over 13, so easily reject the null hypothesis that there is no difference.

(iv) (2 pts.) $blackfemale - nonblackfemale = (-45.09 - 169.81 + 62.31) - (-45.09) = -107.50$. 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 C 7.2

(i) (3 pts.) The estimated model is:

$$\begin{array}{rccccccccc} \widehat{\log(wage)} = & 5.40 & +0.0654 & educ & +0.0140 & exper & +0.0117 & tenure & & \\ & (0.11) & (0.0063) & & (0.0032) & & (0.0025) & & & \\ +0.199 & married & -0.188 & black & -0.091 & south & +0.184 & urban & & \\ (0.039) & & (0.038) & & (0.026) & & (0.027) & & & \end{array}$$

$$n = 935, R^2 = 0.253.$$

The blacks earn about 18.8% less than nonblacks. The t statistic is about -4.95, so it is statistically significant.

(ii) (1 pts.) The F statistic for $exper^2$ and $tenure^2$ with 2 and 925 *df* is about 1.49 with p-value$\approx$0.226. So they are jointly insignificant at the 20% level.

(iii) (2 pts.) Add the interaction term $black \cdot educ$. The coefficient on the interaction is about -0.0226 ($se \approx 0.0202$). Therefore, the point estimate is that the return to another year of education is about 2.3 percentage points lower for blacks than nonblacks. 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) (2 pt.) We choose the base group to be single, nonblack. Then we add dummy variables *marrnonblck*, *singblck*, and *marrblck* for the other three groups. Run the regression with other control variables, we obtain the difference between married blacks and married nonblacks as -0.18. That is, a married black earns about 18% less than a comparable married nonblack.

Problem C 7.6

(i) (3 pts.) *reg sleep totwrk educ age age*^{2} *yngkid if male==1*:

$$\begin{array}{lllllllllll} \widehat{sleep} = & 3648.2 & -0.182 & totwrk & -13.05 & educ & +7.16 & age & -0.0448 & age^2 & +60.38 \quad yngkid \\ & (310.0) & (0.024) & & (7.41) & & (14.32) & & (0.1684) & & (59.02) \end{array}$$

$$n = 400, R^2 = 0.156$$

reg sleep totwrk educ age age^2 *yngkid if male==0*:

$$\begin{array}{lllllllllll} \widehat{sleep} = & 4238.7 & -0.140 & totwrk & -10.21 & educ & -30.36 & age & -0.368 & age^2 & -118.28 \quad yngkid \\ & (384.9) & (0.028) & & (9.59) & & (18.53) & & (0.223) & & (93.19) \end{array}$$

$$n = 306, R^2 = 0.098$$

There are notable differences. For example the effects of age and having young children.

(ii) (2 pts.) The F statistic (with 6 and 694 df) is about 2.12 with p-value$\approx$0.05, and so we reject the null that the sleep equations are the same at the 5% level.

(iii) (2 pts.) If we have the coefficient on *male* unspecified under H_0, and test only the five interaction terms with *male*, the F statistic (5 and 694 df) is about 1.26 and $p-$value $\approx$0.28.

(iv) (2 pts.) 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.

Problem 8.2

(3 pts.) Divide both sides of the original model by *inc*, the resulted new equation has a homoskedastic error term.

Problem 8.4

(i) (4 pts.) These variables have the anticipated signs. If a student takes courses where grades are, on average, higher as reflected by higher *crsgpa* then his/her grades will be higher. The better the student has been in the past as measured by *cumgpa*, the better the student does (on average) in the current semester. Finally, *tothrs* is a measure of experience, and its coefficient indicates an increasing return to experience.

The t statistic for *crsgpa* is very large, over five using the usual standard error (which is the largest of the two). Using the robust standard error for *cumgpa*, its t statistic is about 2.61, which is also significant at the 5% level. The t statistic for *tothrs* is only about 1.17 using either standard error, so it is not significant at the 5% level.

(ii) (3 pts.) This is easiest to see without other explanatory variables in the model. If *crsgpa* were the only explanatory variable, $H_0 : \beta_{crsgpa} = 1$ means that, without any information about the student, the best predictor of term GPA is the average GPA in the students courses; this holds essentially by definition. (The intercept would be zero in this case.) With additional explanatory variables it is not necessarily true that $\beta_{crsgpa} = 1$ because *crsgpa* could be correlated with characteristics of the student. (For example, perhaps the courses students take are influenced by ability as measured by test scores and past college performance.) But it is still interesting to test this hypothesis.

The t statistic using the usual standard error is $t = (.9001)/.175 \simeq -.57$; using the hetero- skedasticity-robust standard error gives $t \simeq -.60$. In either case we fail to reject $H_0 : \beta_{crsgpa} = 1$ at any reasonable significance level, certainly including 5%.

(iii) (3 pts.) The in-season effect is given by the coefficient on season, which implies that, other things equal, an athletes GPA is about .16 points lower when his/her sport is competing. The t statistic using the usual standard error is about 1.60, while that using the robust standard error is about 1.96. Against a two-sided alternative, the t statistic using the robust standard error is just significant at the 5% level (the standard normal critical value is 1.96), while using the usual standard error, the t statistic is not quite significant at the 10% level. So the standard error used makes a difference in this case. This example is somewhat unusual, as the robust standard error is more often the larger of the two.