BOSTON COLLEGE
Department of Economics
EC 228 Econometrics, Prof. Baum, Mr. Barbato, Spring 2003
Problem Set 4

Problem 6.6

Answer.

The extended model has $df = 690 - 9 = 671$,and we are testing two restrictions. Therefore, $F = [(.232 - .229)/(1 - .232)](671/2) \approx 1.31$,which is well below the 10% critical value in the F distribution with 2 and ∞ df : $cv = 2.30$. Thus, $atndrte^2$ and ACT**atndrte* are jointly insignificant. Because adding these terms complicates the model without statistical justification, we would not include them in the final model.

Problem 6.7

Answer.

The second equation is clearly preferred, as its adjusted R-squared is notably larger than that in the other two equations. The second equation contains the same number of estimated parameters as the first, and one fewer than the third. The second equation is also easier to interpret than the third.

Problem 6.9

Answer.

(i) The estimated equation is

$$\begin{aligned} \log(wage) &= \underset{(.106)}{.128} + \underset{(.0075)}{.0904}\, educ + \underset{(.0052)}{.0410} \exp er - \underset{(.000116)}{.000714} \exp er^2 \\ n &= 526,\ R^2 = .300,\ \bar{R}^2 = .296 \end{aligned}$$

(ii) The t statistic on $exper^2$ is about -6.16, which has a p-value of essentially zero. So *exper* is significant at the 1% level (and much smaller significance levels).

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

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

Similarly, for the 20th year of experience,

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

(iv) 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 6.12

Answer.

(i) The results of estimating the log-log model (but with *bdrms* in levels) are

$$\begin{aligned} \log(price) &= \underset{(0.65)}{5.61} + \underset{(.038)}{.168}\log(lotsize) + \underset{(.093)}{.700}\,(\log(sqrft) + \underset{(.028)}{.037}\,bdrms \\ n &= 88,\ R^2 = .634,\ \ \bar{R}^2 = .630 \end{aligned}$$

(ii) With $lotsize = 20,000$, $sqrft = 2,500$,and $bdrms = 4$,we have

$$lprice = 5.61 + .168\log(20,000) + .700\log(2,500) + .037(4) \approx 12.90$$

where we use *lprice* to denote log(*price*). To predict *price*, we use the equation $price = \alpha_0 \exp(lprice)$,where $\alpha_0$is the slope on $m_i = \exp(lprice)$ from the regression $price_i$ on $m_i, i = 1, 2, ..., 88$ (without an intercept). When we do this regression we get $\alpha_0 \approx 1.023$. Therefore, for the values of the independent variabes given above, $price \approx (1.023)\exp(12.90) \approx \$409,519$ (rounded to the nearest dollar). If we forget to multiply by α_0 the predicted price would be about $400,312.

(iii) When we run the regression with all variables in levels, the R-squared is about .672. When we compute the correlation between $price_i$ and m_i from part (ii), we obtain about .859. The square of this, or roughly .738, is the comparable goodness-of-fit measure for the model with log(*price*) as the dependent variable. Therefore, for predicting *price*, the log model is notably better.

Problem 6.16

Answer.

(i) The estimated equation is

$$\begin{aligned} points &= \underset{(6.99)}{35.22} + \underset{(.405)}{2.364}\exp er - \underset{(.0235)}{.0770}\exp er^2 - \underset{(.295)}{1.074}\,age - \underset{(.451)}{1.286}\,coll \\ n &= 269,\ R^2 = .141,\ \ \bar{R}^2 = .128 \end{aligned}$$

(ii) The turnaround point is 2.364/[2(.0770)] $\approx$ 15.35. So, the increase from 15 to 16 years of experience would actually reduce salary. This is a very high level of experience, and we can essentially ignore this prediction: only two players in the sample of 269 have more than 15 years of experience.

(iii) Many of the most promising players leave college early, or in some cases, forego college altogether, to play in the NBA. These top players command the highest salaries. It is not more college that hurts salary, but less indicative of super-star potential.

(iv) When $age^2$is added to the regression from part (i), its coefficient is .0536 (se = .0492). Its t statistic is barely above one, so we are justified in dropping it. The coefficient on *age* in the same regression is -3.984 (se = 2.689). Together, these estimates imply a negative, increasing, return to *age*. The turning point is roughly at 74 years old. In any case, the linear function of *age* seems sufficient.

(v) The OLS results are

$$\begin{aligned}\log(wage) &= \underset{(.85)}{6.78} + \underset{(.007)}{.078}\, points + \underset{(.050)}{.218} \exp er - \underset{(.0028)}{.0071} \exp er^2 \\ &\quad - \underset{(.035)}{.048}\, age - \underset{(.053)}{.040}\, coll \\ n &= 269,\ R^2 = .488,\ \bar{R}^2 = .478\end{aligned}$$

(vi) The joint F test produced by Stata is about 1.19. With 2 and 263 *df*, this gives a *p*-value of roughly .31. Therefore, once scoring and years played are controlled for, there is no evidence for wage differentials depending on age or years played in college.

Problem 7.3

Answer.

(i) 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 *sat* (other things fixed): $19.3/(2(2.19)) \approx 4.41$. Because *hsize* is measured in hundreds, the optimal size of graduating class is about 441.

(ii) This 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) 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 difference, ceteris paribus.

(iv) 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 females dummy variable.

Problem 7.8

Answer.

(i) We want to have a constant semi-elasticity model, so a standard wage equation with marijuana usage included would be

$$\log(wage)=\beta_0+\beta_1 usage+\beta_2 educ+\beta_3\exp er+\beta_4\exp er^2+\beta_5 female+u$$

The 100*β_1 is the approximate percentage change in *wage* when marijuana usage increases by one time per month.

(ii) We would add an interaction term in female and usage:

$$\log(wage)=\beta_0+\beta_1 usage+\beta_2 educ+\beta_3\exp er+\beta_4\exp er^2+\beta_5 female+\beta_6 female*usage+u$$

The null hypothesis that the effect on marijuana usage does not differ by gender is $H_0:\beta_6=0$.

(iii) We take the base group to be nonuser. Then we need dummy variables for the other three groups:*lghtuser*, *moduser*, and *hvyuser.* Assuming no interactive effect with gender, the model would be

$$\begin{aligned}\log(wage) &= \beta_0+\delta_1\lg htuser+\delta_2\bmod user+\delta_3 hvyuser+\beta_2 educ\\ +\beta_3\exp er+\beta_4\exp er^2+ & \quad \beta_5 female+u\end{aligned}$$

(iv) The null hypothesis is $H_0:\delta_1=0,\delta_2=0,\delta_3=0$,for a total of q = 3 restrictions. If n is the sample size, the *df* in the unrestricted model - the denominator *df* in the F distribution - is n - 8. So we would obtain the critical value from the $F_{q,n-8}$ distribution.

(v) The error term could contain factors, such as family background (including parental history of drug abuse) that could directly affect wages and also be correlated with marijuana usage. We are interested in the effects of a person's drug usage on his or her wage, so we would like to hold other confounding factors fixed. We could try to collect data on relevant background information.

Problem 7.11

Answer.

(i) $H_0 : \beta_{13} = 0$. Using the data in MLB1.raw gives $\beta_{13} \approx .254$, $se(\beta_{13}) \approx .131$. The t statisticis about 1.94, which gives a p-value against a two-sided alternative of just over .05. Therefore, we would reject the H_0 at just about the 5% significance level. Controlling for the performance and experience variables, the estimated salary differential between catchers and outfielders is huge, on the order of $100[\exp(.254) - 1] \approx 28.9\%$ [using equation (7.10)].

(ii) This is a joint null, $H_0 : \beta_9 = 0, \beta_{10} = 0, ..., \beta_{13} = 0$. The F statistic, with 5 and 339 *df*, is about 1.78, and its p-vaueis about .117. Thus, we cannot reject H_0at the 10% level.

(iii) Parts (i) and (ii) are roughly consistent. The evidence against the joint null in part (ii) is weaker because we are testing, along with the marginally significant *catcher*, several other insignificant variables (especially *thrdbase* and *shrtstop*, which has absolute t statistics well below one).