Solutions to Problem Set 3 (Due October 20)

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

Maximum number of points for Problem set 3 is: 110

3.1

- (i) (2 pts.) *hsperc* is defined so that the smaller it is, the lower the students standing in high school. Everything else equal, the worse the students standing in high school, the lower is his/her expected college GPA.
- (ii) (2 pts.)
$$\widehat{colgpa} = 1.392 - .0135(20) + .00148(1050) = 2.676$$
.
- (iii) (2 pts.) The difference between A and B is 140 times the coefficient on *sat*, because *hsperc* is the same for both students. So A is predicted to have a score $.00148(140) \approx .207$ higher.
- (iv) (4 pts.) With *hsperc* fixed, $\Delta\widehat{colgpa} = .00148\Delta sat$. Now, we want to find Δsat such that $\Delta\widehat{colgpa} = .5$, so $.5 = .00148(\Delta sat)$ or $\Delta sat = .5/(.00148) \approx 338$. Perhaps not surprisingly, a large ceteris paribus difference in SAT score almost two and one-half standard deviations is needed to obtain a predicted difference in college GPA or a half a point.

3.3

- (i) (2 pts.) If adults trade off sleep for work, more work implies less sleep (other things equal), so $\beta_1 < 0$
- (ii) (2 pts.) The signs of β_2 and β_3 are not obvious, at least to me. One could argue that more educated people like to get more out of life, and so, other things equal, they sleep less ($\beta_2 < 0$). The relationship between sleeping and age is more complicated than this model suggests, and economists are not in the best position to judge such things.
- (iii) (4 pts.) Since *totwrk* is in minutes, we must convert five hours into minutes: $\Delta totwrk = 5(60) = 300$. Then *sleep* is predicted to fall by $.148(300) = 44.4 minutes$. For a week, 45 minutes less sleep is not an overwhelming change.

- (iv) (2 pts.) More education implies less predicted time sleeping, but the effect is quite small. If we assume the difference between college and high school is four years, the college graduate sleeps about 45 minutes less per week, other things equal.
- (v) (2 pts.) Not surprisingly, the three explanatory variables explain only about 11.3 percent of the variation in *sleep*. One important factor in the error term is general health. Another is marital status, and whether the person has children. Health (however we measure that), marital status, and number and ages of children would generally be correlated with *totwrk*. (For example, less healthy people would tend to work less.)

3.4

- (i) (2 pts.) A larger rank for a law school means that the school has less prestige; this lowers starting salaries. For example, a rank of 100 means there are 99 schools thought to be better.
- (ii) (2 pts.) $\beta_1 > 0$ $\beta_2 > 0$ Both $LSAT$ and GPA are measures of the quality of the entering class. No matter where better students attend law school, we expect them to earn more, on average. $\beta_3 > 0$ $\beta_4 > 0$ The number of volumes in the law library and the tuition cost are both measures of the school quality. (Cost is less obvious than library volumes, but should reflect quality of the faculty, physical plant, and so on.)
- (iii) (2 pts.) This is just the coefficient on GPA, multiplied by 100: 24.8 percent.
- (iv) (2 pts.) This is an elasticity: a one percent increase in library volumes implies a .095 percent increase in predicted median starting salary, other things equal.
- (v) (2 pts.) It is definitely better to attend a law school with a lower rank. If law school A has a ranking 20 less than law school B, the predicted difference in starting salary is $100(.0033)(20) = 6.6$percent higher for law school A.

4.1

- (i) (2 pts.) Heteroskedasticity generally causes the t statistics not to have a t distribution under H_0. Homoskedasticity is one of the CLM assumptions.
- (ii) (2 pts.) The CLM assumptions contain no mention of the sample correlations among independent variables, except to rule out the case where the correlation is one. If two independent variables are perfectly correlated, then the X matrix is not of full rank and we have a problem. Otherwise, partial correlations are acceptable (and likely).
- (iii) (2 pts.) An important omitted variable violates Assumption MLR.4 (zero conditional mean), so then the t statistics don't have a t distribution under H_0. For example, suppose we are trying to predict consumption of cigarettes. On the right hand side, we include income but we do not include education. Since income and education are almost surely positively correlated, then the errors would not have zero conditional mean. This would lead to biased estimates of β.

4.3

- (i) (4 pts.) Holding $profmarg$ fixed,

$$\widehat{\triangle rdintents} = .321\triangle log(sales) = (.321/100)[100\triangle log(sales)] \approx .00321(\%\triangle sales)$$

Therefore, if $\%\triangle sales = 10, \widehat{\triangle rdintens} \approx .032$, or only about 3/100 of a percentage point. For such a large percentage increase in sales, this seems like a very small effect.

- (ii) (4 pts.) $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 > 0$, where β_1 is the population slope on $log(sales)$. The t statistic is $.321/.216 \approx 1.486$. The 5% critical value for a one-tailed test, with $df = 32 - 3 = 29$, is obtained from Table G.2 as 1.699; so we cannot reject H_0 at the 5% level. But the 10% critical value is 1.311; since the t statistic is above this value, we reject H_0 in favor of H_1 at the 10% level.

- (iii) (2 pts.) With an increase of profit margin by 1 percentage point, expenditures on R&D rise by 0.05 percentage points. Economically that is quite significant, as given a 10 % increase in profit margin then they will increase expenditures on R& D by 0.5 percentage point.

- (iv) 2 pts.) Not really. Its t statistic is only 0.05/0.046=1.087, so we are not able to reject at even the 10% level.

4.5

- (i) (2 pts.) $.412 \pm 1.96(.094)$, or about [.228 , .596].

- (ii) (2 pts.) No, because the value .4 is well inside the 95% CI.

- (iii)(2 pts.) Yes, because 1 is well outside the 95% CI.

C3.8

- (i) (2 pts.)

```
. summarize prpblck income

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     prpblck |       409    .1134864    .1824165         0   .9816579
      income |       409    47053.78    13179.29     15919     136529
```

The average of *prpblck* is .113 with standard deviation .182; the average of *income* is 47,053.78 with standard deviation 13,179.29. It is evident that *prpblck* is a proportion and that *income* is measured in dollars.

- (ii) (2 pts.)

```
. regress psoda prpblck income

      Source |       SS       df       MS              Number of obs =     401
-------------+------------------------------           F(  2,   398) =   13.66
       Model |  .202552215     2  .101276107           Prob > F      =  0.0000
    Residual |  2.95146493   398  .007415741           R-squared     =  0.0642
-------------+------------------------------           Adj R-squared =  0.0595
       Total |  3.15401715   400  .007885043           Root MSE      =  .08611

------------------------------------------------------------------------------
       psoda |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     prpblck |   .1149882   .0260006     4.42   0.000     .0638724    .1661039
      income |   1.60e-06   3.62e-07     4.43   0.000     8.91e-07    2.31e-06
       _cons |   .9563196    .018992    50.35   0.000     .9189824    .9936568
------------------------------------------------------------------------------
```

The results from the OLS regression are

$$\widehat{psoda} = .956 + .115 prpblck + .0000016 income$$

$$n = 401,\ R^2 = .064$$

If say *prpblck* increases by .10 (ten percentage point), the price of soda is estimated to increase by .0115 dollars, or about 1.2 cents. While this does not seem large, there are communities with no black population and others that are almost all black, in which case the difference in psoda is estimated to be almost 11.5 cents.

- (iii) (2 pts.)

```
. regress psoda prpblck

      Source |       SS       df       MS              Number of obs =     401
-------------+------------------------------           F(  1,   399) =    7.34
       Model |  .057010466     1  .057010466           Prob > F      =  0.0070
    Residual |  3.09700668   399  .007761922           R-squared     =  0.0181
-------------+------------------------------           Adj R-squared =  0.0156
       Total |  3.15401715   400  .007885043           Root MSE      =   .0881
```

```
------------------------------------------------------------------------------
       psoda |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     prpblck |   .0649269    .023957     2.71   0.007     .0178292    .1120245
       _cons |   1.037399   .0051905   199.87   0.000     1.027195    1.047603
------------------------------------------------------------------------------
```

The simple regression estimate on *prpblck* is .065, so the simple regression estimate is actually lower. This is because *prpblck* and *income* are negatively correlated (-.43) and *income* has a positive coefficient in the multiple regression. You can see the negative correlation by using the command "corr prpblck income".

- (iv) (2 pts.)

```
. regress lpsoda prpblck lincome

      Source |       SS       df       MS              Number of obs =     401
-------------+------------------------------           F(  2,   398) =   14.54
       Model |  .196020672     2  .098010336           Prob > F      =  0.0000
    Residual |  2.68272938   398  .006740526           R-squared     =  0.0681
-------------+------------------------------           Adj R-squared =  0.0634
       Total |  2.87875005   400  .007196875           Root MSE      =   .0821

------------------------------------------------------------------------------
      lpsoda |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     prpblck |   .1215803   .0257457     4.72   0.000     .0709657    .1721948
     lincome |   .0765114   .0165969     4.61   0.000     .0438829    .1091399
       _cons |   -.793768   .1794337    -4.42   0.000    -1.146524   -.4410117
------------------------------------------------------------------------------
```

$$\widehat{log(psoda)} = -.794 + .122 prpblck + .077 lincome$$

$$n = 401,\ R^2 = .068$$

If *prpblck* increases by .20, log(psoda) is estimated to increase by .20(.122)=.0244, or about 2.44 percent.

- (v) (2 pts.)

```
. regress lpsoda prpblck lincome prppov

      Source |       SS       df       MS              Number of obs =     401
```

```
-------------+------------------------------           F(  3,   397) =   12.60
       Model |  .250340622     3  .083446874           Prob > F      =  0.0000
    Residual |  2.62840943   397  .006620679           R-squared     =  0.0870
-------------+------------------------------           Adj R-squared =  0.0801
       Total |  2.87875005   400  .007196875           Root MSE      =  .08137

------------------------------------------------------------------------------
      lpsoda |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     prpblck |   .0728072   .0306756     2.37   0.018     .0125003    .1331141
     lincome |   .1369553   .0267554     5.12   0.000     .0843552    .1895553
      prppov |     .38036   .1327903     2.86   0.004     .1192999    .6414201
       _cons |  -1.463333   .2937111    -4.98   0.000    -2.040756   -.8859092
------------------------------------------------------------------------------
```

$\hat{\beta}_{prpblck}$ falls to about .073 when *prppov* is added to the regression.

- (vi) (2 pts.)

```
    . corr lincome prppov
(obs=409)

             |  lincome   prppov
-------------+------------------
     lincome |   1.0000
      prppov |  -0.8385   1.0000
```

The correlation is about -.84, which makes sense because poverty rates are determined by income (but not directly in terms of median income).

- (vii) (2 pts.) There is no argument that they are highly correlated, but we are using them simply as controls to determine if there is price discrimination against blacks. In order to isolate the pure discrimination effect, we need to control for as many measures of income as we can; therefore, including both variables makes sense.

C4.1

- (i) (2 pts.) Holding other factors fixed,

$$\triangle voteA = \beta_1 \triangle log(expendA) = (\beta_1/100)[100 \triangle log(expendA)] \approx (\beta_1/100)(\%\triangle expendA) \quad (1)$$

So a .01 increase in expenditure will result in a $(\beta_1/100) * (100 * .01) = .01\beta_1$ change in the vote for A.

- (ii) (2 pts.) The null hypothesis is $H_0 : \beta_2 = -\beta_1$, which means a $z\%$ increase in expenditure by A and a $z\%$ increase in expenditure by B leaves voteA unchanged. We can equivalently write $H_0 : \beta_1 + \beta_2 = 0$.

- (iii) (4 pts.)

```
. reg  voteA lexpendA lexpendB prtystrA

      Source |       SS       df       MS              Number of obs =     173
-------------+------------------------------           F(  3,   169) =  215.23
       Model |  38405.1089     3   12801.703           Prob > F      =  0.0000
    Residual |  10052.1396   169  59.4801161           R-squared     =  0.7926
-------------+------------------------------           Adj R-squared =  0.7889
       Total |  48457.2486   172  281.728189           Root MSE      =  7.7123

------------------------------------------------------------------------------
       voteA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lexpendA |   6.083316     .38215    15.92   0.000     5.328914    6.837719
    lexpendB |  -6.615417   .3788203   -17.46   0.000    -7.363247   -5.867588
    prtystrA |   .1519574   .0620181     2.45   0.015     .0295274    .2743873
       _cons |   45.07893   3.926305    11.48   0.000     37.32801    52.82985
------------------------------------------------------------------------------
```

The estimated equation (with standard errors in parentheses below estimates) is

$$\widehat{voteA} = 45.08(3.93) + 6.08(0.38)log(expendA) - 6.62(0.39)log(expendB) + .15(0.06)prtystrA$$

$$n = 173, R^2 = .793$$

The coefficient on log(expendA) is very significant (t statistic ≈ 15.92), as is the coefficient on log(expendB) (t statistic ≈ -17.45). The estimates imply that a 10%, ceteris paribus, increase in spending by candidate A increases the predicted share of the vote going to A by about .61 percentage points. [Recall that, holding other factors fixed, $\triangle\widehat{voteA} \approx (6.083/100)\%\triangle log(expendA)$ Similarly, a 10% ceteris paribus increase in spending by B reduces A's vote by about .66 percentage points. These effects certainly cannot be ignored. While the coefficients on log(expendA) and log(expendB) are of similar magnitudes (and opposite in sign, as we expect), we do not have the standard error of $\hat{\beta}_1 + \hat{\beta}_2$, which is what we would need to test the hypothesis from part (ii).

- (iv) (2 pts.)

```
. test lexpendA=-lexpendB

 ( 1)  lexpendA + lexpendB = 0

       F(  1,   169) =    1.00
            Prob > F =    0.3196
```

So we fail to reject $\beta_1 + \beta_2 = 0$.

C4.3

- (i) (2 pts.) The estimated model is

```
. regress lprice sqrft bdrms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   60.73
       Model |  4.71671468     2  2.35835734           Prob > F      =  0.0000
    Residual |  3.30088884    85  .038833986           R-squared     =  0.5883
-------------+------------------------------           Adj R-squared =  0.5786
       Total |  8.01760352    87  .092156362           Root MSE      =  .19706

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sqrft |   .0003794   .0000432     8.78   0.000     .0002935    .0004654
       bdrms |   .0288844   .0296433     0.97   0.333    -.0300543    .0878232
       _cons |   4.766027   .0970445    49.11   0.000     4.573077    4.958978
------------------------------------------------------------------------------
```

$$\widehat{log(price)} = 4.766(0.10) + .000379(.000043)sqrft + .0289(.0296)bdrms$$

$$n = 88, R^2 = .588$$

Therefore, $\hat{\theta}_1 = 150(.000379) + .0289 = .858$, which means that an additional 150 square foot bedroom increases the predicted price by about 8.6 %.

- (ii) (2 pts.) $\beta_2 = \theta_1 - 150\beta_1$, and so $log(price) = \beta_0 + \beta_1 sqrft + (\theta_1 - 150\beta_1)bdrms + u$ $= \beta_0 + \beta_1(sqrft - 150bdrms) + \theta_1 bdrms + u$.

- (iii) (2 pts.) From part (ii) we run the regression

```
. gen sqrft150=sqrft-150*bdrms

. regress  lprice sqrft150 bdrms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   60.73
       Model |  4.71671468     2  2.35835734           Prob > F      =  0.0000
    Residual |  3.30088884    85  .038833986           R-squared     =  0.5883
-------------+------------------------------           Adj R-squared =  0.5786
       Total |  8.01760352    87  .092156362           Root MSE      =  .19706

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sqrft150 |   .0003794   .0000432     8.78   0.000     .0002935    .0004654
       bdrms |   .0858013   .0267675     3.21   0.002     .0325804    .1390223
```

```
       _cons |   4.766027   .0970445    49.11   0.000     4.573077    4.958978
------------------------------------------------------------------------------
```

Really, $\hat{\theta}_1 = .0858$; note we also get $se(\hat{\theta}_1) = .0268$.The 95% confidence interval is .0326 to .1390 (or about 3.3% to 13.9%).

Problem C4.5

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

$$\widehat{log(salary)} = \underset{(0.27)}{11.02} + \underset{(.0121)}{.0677}\, years + \underset{(.0016)}{.0158}\, gamesyr + \underset{(.0011)}{.0014}\, bavg + \underset{(.0072)}{.0359}\, hrunsyr$$

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

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

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

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

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

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

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

  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) (2 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) (2 points) The OLS regression results when $log(hseval)$ is added are

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

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

  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) (4 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) (2 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.