BOSTON COLLEGE
Department of Economics
EC 228 01 Econometric Methods
Fall 2008, Prof. Baum, Ms. Phillips (tutor), Mr. Dmitriev (grader)
Problem Set 3
Due at classtime, Thursday 14 Oct 2008

Problem 4.1

(i) (5 marks) generally cause the t statistics not to have a t distribution under H0. Homoskedasticity is one of the CLM assumptions.

(ii) (5 marks) The CLM assumptions contain no mention of the sample correlations among independent variables, except to rule out the case where the correlation is one.

(iii) (5 marks) An important omitted variable violates Assumption MLR.4 (zero conditional mean), t statistics doesn't have distribution under H_0.

Problem 4.3

(i) (10 marks) Holding $profmarg$ fixed, $\triangle \widehat{rdintents} = .321 \triangle log(sales) = (.321/100)[100 \triangle log(sales)] \approx .00321(\% sales)$. Therefore, if $\% \triangle sales = 10, \triangle \widehat{rdintens} \approx .032$, or only about 3/100 of a percentage point. For such a large percentage increase in sales, this seems like a practically small effect.

(ii) (10 marks) $H_0 : \beta_1 = 0$ 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% criticavalue is 1.311; since the t statistic is above this value, we reject H0 in favor of H_1 at the 10% level.

(iii) (5 marks) With an increase of profit margin by 1 percentage point expenditures on R&D rise by 0.05 percentage points. Economically it is quite large, as for 10 % difference in profit margin difference will increase expendetures on R& D by 0.5 percentage point, which is really big, given that for a company with 100 million dollars sales they will be around 2 %, so they will rise by somewhat around quarter.

(iv) (5 marks) Not really. Its t statistic is only 0.05/0.046=1.087, which is well below even the 10% critical value for a one-tailed test.

Problem 4.5

(i) (5 marks) .412 1.96(.094), or about .228 to .596.
(ii) (5 marks) No, because the value .4 is well inside the 95% CI.
(iii)(5 marks) Yes, because 1 is well outside the 95% CI.

Problem 4.6

(i) (10 marks) With $df = n-2 = 86$, we obtain the 5% critical value from Table G.2 with $df = 90$. Because each test is two-tailed, the critical value is 1.987. The t statistic for $H_0 : \beta_0 = 0$ is about -.89, which is much less than 1.987 in absolute value. Therefore, we fail to reject $\beta_0 = 0$. The t statistic for H_0: $\beta_1 = 1$ is $(.976 - 1)/.049 \approx -.49$, which is even less significant. (Remember, we reject H0 in favor of H1 in this case only if $|t| > 1.987$.)

(ii)(5 marks) We use the SSR form of the F statistic. We are testing q = 2 restrictions and the df in the unrestricted model is 86. We are given SSR_r = 209,448.99 and SSR_ur = 165,644.51. Therefore,

$$F = \frac{(209,448.99165,644.51)}{165,644.51}(\frac{86}{2}) \approx 11.37$$

which is a strong rejection of H_0: from Table G.3c, the 1% critical value with 2 and 90 df is 4.85.

(iii) (10 marks) We use the R-squared form of the F statistic. We are testing q = 3 restrictions and there are 88 . 5 = 83 df in the unrestricted model. The F statistic is $[(.829 - .820)/(1 - .829)](83/3) \approx 1.46$. The 10% critical value (again using 90 denominator df in Table G.3a) is 2.15, so we fail to reject H0 at even the 10% level. In fact, the p-value is about .23.

(iv)(5 marks)If heteroskedasticity were present, Assumption MLR.5 would be violated (homoskedasticity), and the F statistic would not have an F distribution under the null hypothesis. Therefore, comparing the F statistic against the usual critical values, or obtaining the p-value from the F distribution, would not be especially meaningful.

Problem 4.7

(i) (5 marks) While the standard error on $hrsemp$ has not changed, the magnitude of the coefficient has increased by half. 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.)

(ii) (5 marks) if we add and subtract $\beta_2 log(employ)$ from the right-hand-side and collect terms, we have

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

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

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.

(iii) (5 marks) No. We are interested in the coefficient on $log(employ)$, which has a t statistic of .2, which is very small. 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) (5 marks) The null hypothesis in the model from part (ii) is $H_0 : \beta_2 = -1$. The t sratistic is $[-.951 - (-1)]/.37 \approx .132$; this is very small, and we fail to reject whether we specify a one- or two-sided alternative.

Problem C3.2

(i) (5 marks)

```
. regress price sqrft bdrms
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   72.96
       Model |  580009.152     2  290004.576           Prob > F      =  0.0000
    Residual |  337845.354    85  3974.65122           R-squared     =  0.6319
-------------+------------------------------           Adj R-squared =  0.6233
       Total |  917854.506    87  10550.0518           Root MSE      =  63.045

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sqrft |   .1284362   .0138245     9.29   0.000     .1009495    .1559229
       bdrms |   15.19819   9.483517     1.60   0.113    -3.657582    34.05396
       _cons |    -19.315   31.04662    -0.62   0.536    -81.04399      42.414
------------------------------------------------------------------------------
```

The estimated equation is

$$\widehat{price} = -19.32 + .128 sqrft + 15.20 bdrms$$

$$n = 88,\ R^2 = .632$$

(ii) (5 marks) Holding square footage constant, $\widehat{\triangle price} = 15.20 \triangle bdrms$, and so $\widehat{price}$ increases by 15.20, which means \$15,200.

(iii) (5 marks) Now $\widehat{\triangle price} = .128 \triangle sqrft + 15.20 \triangle bdrms = .128(140) + 15.20 = 33.12$,or \$33,120. Because the size of the house is increasing, this is a much larger effect than in(ii).

(iv) (5 marks) About 63.2%

(v) (5 marks) The predicted price is $-19.32 + .128(2,438) + 15.20(4) = 353.544$, or \$353,544.

(vi) (5 marks) From part (v), the estimated value of the home based only on square footage and number of bedrooms is \$353,544. The actual selling price was \$300,000, which suggests the buyer underpaid by some margin. But, of course, there are many other features of a house (some that we cannot even measure) that affect price, and we have not controlled for these.

Problem C3.4

(i) (5 marks)The minimum, maximum, and average values for these three variables are given in the table below:

Variable	Average	Minimum	Maximum
atndrte	81.71	6.25	100
priGPA	2.59	0.86	3.93
ACT	22.51	13	32

(ii) (5 marks)

```
. regress atndrte priGPA ACT

      Source |       SS       df       MS              Number of obs =     680
-------------+------------------------------           F(  2,   677) =  138.65
       Model |  57336.7612     2  28668.3806           Prob > F      =  0.0000
    Residual |  139980.564   677  206.765974           R-squared     =  0.2906
-------------+------------------------------           Adj R-squared =  0.2885
       Total |  197317.325   679   290.59989           Root MSE      =  14.379

------------------------------------------------------------------------------
     atndrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |   17.26059   1.083103    15.94   0.000     15.13395    19.38724
         ACT |  -1.716553    .169012   -10.16   0.000    -2.048404   -1.384702
       _cons |    75.7004   3.884108    19.49   0.000     68.07406    83.32675
------------------------------------------------------------------------------
```

The estimated equation is

$$\widehat{atndrte} = 75.70 + 17.26 priGPA - 1.72 ACT$$

$$n = 680, R^2 = 0.291$$

The intercept means that, for a student whose prior GPA is zero and ACT score is zero, the predicted attendance rate is 75.7%. But this is clearly not an interesting segment of the population. (In fact, there are no students in the college population with $priGPA = 0$ and $ACT = 0$, or with values even close to zero.)

(iii) (5 marks)The coefficient on $priGPA$ means that, if a students prior GPA is one point higher (say, from 2.0 to 3.0), the attendance rate is about 17.3 percentage points higher. This holds ACT fixed. The negative coefficient on ACT is, perhaps initially a bit surprising. Five more points on the ACT is predicted to lower attendance by 8.6 percentage points at a given level of $priGPA$. As $priGPA$ measures performance in college (and, at least partially, could reflect, past attendance rates), while ACT is a measure of potential in college, it appears that students that had more promise (which could mean more innate ability) think they can get by with missing lectures.

(iv) (5 marks)We have $\widehat{atndrte} = 75.70 + 17.267(3.65) - 1.72(20) \approx 104.3$. Of course, a student cannot have higher than a 100% attendance rate. Getting predictions like this is always possible when using regression methods for dependent variables with natural upper or lower bounds. In practice, we would predict a 100% attendance rate for this student. (In fact, this student had an actual attendance rate of 87.5%.)

(v) (5 marks)The difference in predicted attendance rates for A and B is $17.26(3.1 - 2.1) - (21 - 26) = 25.86$.

Problem C3.8

(i) (5 marks)

```
. 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) (5 marks)

```
. 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

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

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

. 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) (5 marks)

```
. 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.

(iv) (5 marks)

```
. regress lpsoda prpblck income
```

```
      Source |       SS       df       MS              Number of obs =     401
-------------+------------------------------           F(  2,   398) =   14.08
       Model |  .190231453     2  .095115727           Prob > F      =  0.0000
    Residual |   2.6885186   398  .006755072           R-squared     =  0.0661
-------------+------------------------------           Adj R-squared =  0.0614
       Total |  2.87875005   400  .007196875           Root MSE      =  .08219

------------------------------------------------------------------------------
      lpsoda |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     prpblck |   .1111178   .0248154     4.48   0.000     .0623321    .1599035
      income |   1.56e-06   3.45e-07     4.51   0.000     8.79e-07    2.24e-06
       _cons |  -.0456777   .0181263    -2.52   0.012    -.0813129   -.0100425
------------------------------------------------------------------------------
```

$$\widehat{log(psoda)} = -.045 + .111 prpblck + 1.56e - 06(income)$$
$$n = 401,\ R^2 = .067$$

If *prpblck* increases by .20, log(psoda) is estimated to increase by .20(.111)=.0222, or about 2.22 percent.

(v) (5 marks)

```
. regress lpsoda prpblck income prppov

      Source |       SS       df       MS              Number of obs =     401
-------------+------------------------------           F(  3,   397) =   10.05
       Model |  .203184207     3  .067728069           Prob > F      =  0.0000
    Residual |  2.67556584   397  .006739461           R-squared     =  0.0706
-------------+------------------------------           Adj R-squared =  0.0636
       Total |  2.87875005   400  .007196875           Root MSE      =  .08209

------------------------------------------------------------------------------
      lpsoda |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     prpblck |   .0861628   .0306334     2.81   0.005     .0259388    .1463868
      income |   1.97e-06   4.55e-07     4.33   0.000     1.07e-06    2.86e-06
      prppov |   .1505201   .1085741     1.39   0.166    -.0629319    .3639722
       _cons |   -.072912   .0267156    -2.73   0.007    -.1254337   -.0203904
------------------------------------------------------------------------------
```

$\hat{\beta}_{prpblck}$ falls to about .086 when *prppov*is added to the regression.

(vi) (5 marks)

```
    . 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) (5 marks)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; including both variables makes sense.

Problem C4.1

(i) (5 marks) Holding other factors fixed,

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

(ii) (5 marks) 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) (10 marks)

```
. 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

$$\widehat{voteA} = 45.08(3.93) + 6.083(0.382) log(expendA) - 6.615(0.379) log(expendB) + .152(0.062) prtystrA$$

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

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 by about .66 percentage points. These effects certainly cannot be ignored. ..voteA 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) (5 marks)

```
. test lexpendA=-lexpendB

 ( 1)  lexpendA + lexpendB = 0

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

or, equivalently,

```
. gen diffBA= lexpendB- lexpendA

. reg  voteA lexpendA diffBA 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 |  10052.1397   169  59.4801165           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 |   -.532101   .5330858    -1.00   0.320    -1.584466     .520264
      diffBA |  -6.615417   .3788203   -17.46   0.000    -7.363246   -5.867588
    prtystrA |   .1519574   .0620181     2.45   0.015     .0295274    .2743873
       _cons |   45.07893   3.926305    11.48   0.000     37.32801    52.82985
------------------------------------------------------------------------------
```

Write $\theta_1 = \beta_1 + \beta_2$, or $\beta_1 = \theta_1 - \beta 2$. Plugging this into the original equation, and rearranging, gives

$$\widehat{voteA} = \beta_0 + \theta_1 log(expendA) + \beta_2[log(expendB) - log(expendA)] + \beta_3 prtystrA + u$$

When we estimate this equation we obtain $\hat{\theta}_1 \approx -.532$ and $se(\hat{\theta}_1) \approx .533$. The t statistic for the hypothesis in part (ii) is $-.532/.533 \approx -1$. Therefore, we fail to reject $H_0 : \beta_2 = -\beta_1$.

C4.3(i) (5 marks) 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) (5 marks)$\beta_2 = \theta_1 - 150\beta_1$, 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) (5 marks) 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, $\hat{\theta_1} = .0858$; no we also get $se(\hat{\theta_1}) = .0268$.The 95% confidence interval reported by my software package is .0326 to .1390 (or about 3.3% to 13.9%).

Problem C4.5

(i) (5 marks) If we drop $rbisyr$ the estimated equation becomes

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

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

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

(ii) (5 marks) The equation with *runsyr*, *fldperc*, and *sbasesyr* added is

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

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

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) (5 marks) 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) (5 marks) The results from the OLS regression, with standard errors in parentheses, are

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

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

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

(iii) (5 marks) The OLS regression results when $log(hseval)$ is added are

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

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

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) (5 marks) 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) (marks) 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.