BOSTON COLLEGE
Department of Economics
EC 228 02 Econometric Methods
Fall 2009, Prof. Baum, Ms. Phillips (tutor), Ms. Pumphrey (grader)
Problem Set 3
Due at classtime, Tuesday 13 Oct 2009

C3.1

(i) (2 pts.) Probably $\beta_2 > 0$, as more income typically means better nutrition for the mother and better prenatal care.

(ii) (4 pts.) Yes, they are likely correlated and an argument can be made for both positive or negative correlation. On the one hand, an increase in income generally increases the consumption of a good, and $cigs$ and $faminc$ could be positively correlated. On the other hand, family incomes are also higher for families with more education, and more education and cigarette smoking tend to be negatively correlated. The sample correlation between $cigs$ and $faminc$ is about -0.173, indicating a negative correlation.

(iii) (4 pts.)

```
. regress  bwght cigs

      Source |       SS       df       MS              Number of obs =     694
-------------+------------------------------           F(  1,   692) =   25.33
       Model |  10394.4794     1  10394.4794           Prob > F      =  0.0000
    Residual |  283941.338   692  410.319852           R-squared     =  0.0353
-------------+------------------------------           Adj R-squared =  0.0339
       Total |  294335.817   693  424.727009           Root MSE      =  20.256

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |   -.601789    .119565    -5.03   0.000    -.8365427   -.3670353
       _cons |   120.3839    .821228   146.59   0.000     118.7715    121.9963
------------------------------------------------------------------------------

. regress bwght cigs faminc
```

```
      Source |       SS       df       MS              Number of obs =     694
-------------+------------------------------           F(  2,   691) =   14.21
       Model |   11626.062     2  5813.03102           Prob > F      =  0.0000
    Residual |  282709.755   691  409.131339           R-squared     =  0.0395
-------------+------------------------------           Adj R-squared =  0.0367
       Total |  294335.817   693  424.727009           Root MSE      =  20.227

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5632265   .1214429    -4.64   0.000     -.801668   -.3247851
      faminc |    .073165   .0421699     1.74   0.083    -.0096316    .1559616
       _cons |   118.1664   1.518518    77.82   0.000      115.185    121.1479
------------------------------------------------------------------------------
```

For the regression without $faminc$:
$\widehat{bwght} = 120.3839 - .601789 cigs$, n=694, $R^2 = 0.0353$
For the regression with $faminc$:
$\widehat{bwght} = 118.1664 - .5632 cigs + .0732 faminc$, n=694, $R^2 = 0.0395$
The effect of cigarette smoking is slightly smaller when $faminc$ is added to the regression, but the difference is not great. This is due to the fact that $cigs$ and $faminc$ are not very correlated, and the coefficient on $faminc$ is practically small.(The variable $faminc$ is measured in thousands, so 10000 more dollars in 1988 inome increases predicted weight by only 0.93 ounces.)

C3.2

(i) (2 pts.)

```
. 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

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

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

(ii) (2 pts.) Holding square footage constant, $\triangle\widehat{price} = 15.20\triangle bdrms$, and so $\widehat{price}$ increases by 15.20, which means $15,200.

(iii) (2 pts.) Now $\triangle\widehat{price} = .128\triangle sqrft + 15.20\triangle bdrms = .128(140) + 15.20 = 33.12$,or $33,120. Because the size of the house is increasing, this is a much larger effect than in(ii).

(iv) (2 pts.) About 63.2% from R^2.

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

(vi) (2 pts.) 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.

C3.4

(i) (2 pts.)The minimum, maximum, and average values for these three variables are given in the table below. Use the command "summarize atndrte priGPA ACT".

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

(ii) (4 pts.)

```
. 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) (2 pts)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) (2 pts)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) (2 pts)The difference in predicted attendance rates for A and B is $17.26(3.1 - 2.1) - (21 - 26) = 25.86$.

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.

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

C4.1

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

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

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%).