BOSTON COLLEGE
Department of Economics
EC 228 Econometrics, Prof. Baum, Ms. Yu, Fall 2003

**Problem Set 3 Solutions**

Problem sets should be your own work. You may work together with classmates, but if you're not figuring this out on your own, you will eventually regret it.

**1.** (**3.2**)

(i) Yes. Because of budget constraints, it makes sense that, the more siblings there are in a family, the less education any one child in the family has. To find the increase in the number of siblings that reduces predicted education by one year, we solve $1 = .094(\Delta sibs)$, so $\Delta sibs = 1/.094 \approx 10.6$.

(ii) Holding *sibs* and $feduc$ fixed, one more year of mother's education implies .131 years more of predicted education. So if a mother has four more years of education, her son is predicted to have about a half a year (.524) more years of education.

(iii) Since the number of siblings is the same, but *meduc* and $feduc$ are both different, the coefficients on *meduc* and $feduc$ both need to be accounted for. The predicted difference in education between $B$ and $A$ is $.131(4) + .210(4) = 1.364$.

**2.** (**3.5**)

(i) No. By definition, $study + sleep + work + leisure = 168$. So if we change *study*, we must change at least one of the other categories so that the sum is still 168.

(ii) From part (i), we can write, say, *study* as a perfect linear function of the other independent variables: $study = 168 - sleep - work - leisure$. This holds for every observation, so MLR.4 is violated.

(iii) Simply drop one of the independent variables, say *leisure*:

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + u.$$

Now, for example, $\beta_1$ is interpreted as the change in $GPA$ when *study* increases by one hour, where *sleep*, *work*, and $u$ are all held fixed. If we are holding *sleep* and *work* fixed but increase *study* by one hour, then we must be reducing *leisure* by one hour. The other slope parameters have a similar interpretation.

**3.** (**3.7**)

Only (ii), omitting an important variable, can cause bias, and this is true only when the omitted variable is correlated with the included explanatory variables. The homoskedasticity assumption, MLR.5, played no role in showing that the OLS estimators are unbiased. (Homoskedasticity was used to obtain the standard variance formulas for the $\hat{\beta}_j$.) Further, the degree of collinearity between the explanatory variables in the sample, even if it is reflected in a correlation as high as .95, does not affect the Gauss-Markov assumptions. Only if there is a *perfect* linear relationship among two or more explanatory variables is MLR.4 violated.

**4.** (**3.13**)

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

(ii)
```
.  use http://fmwww.bc.edu/ec-p/data/wooldridge/BWGHT50

.  correlate cigs faminc
(obs=694)

             |     cigs   faminc
-------------+------------------
        cigs |   1.0000
      faminc |  -0.1830   1.0000
```

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, 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 $-.183$, indicating a negative correlation.

(iii)
```
.  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    -.8016679   -.3247851
      faminc |    .073165   .0421699     1.74   0.083    -.0096316    .1559616
       _cons |   118.1664   1.518518    77.82   0.000      115.185    121.1479
------------------------------------------------------------------------------
```

The regressions without and with $faminc$ are:

$$\begin{aligned} \widehat{bwght} &= 120.38 - .602 cigs \\ n &= 694, R^2 = .035 \end{aligned}$$

and

$$\begin{aligned} \widehat{bwght} &= 118.17 - .563 cigs + .073 faminc \\ n &= 496, R^2 = .0395. \end{aligned}$$

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 $10,000 more in 1988 income increases predicted birth weight by only .93 ounces.)

**5.** (**3.16**)

(i)
```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/ATTEND

. summarize atndrte priGPA ACT

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     atndrte |       680    81.70956    17.04699       6.25        100
      priGPA |       680    2.586775    .5447141       .857       3.93
         ACT |       680    22.51029    3.490768         13         32
```

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 | .86 | 3.93 |
| *ACT* | 22.51 | 13 | 32 |

(ii)
```
. 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
```

```
-----------------------------------------------------------------------------

.  list if priGPA>3.64 & ACT==20

     +-----------------------------------------------------------------+
569. | attend | termgpa | priGPA | ACT | final | atndrte | hwrte | frosh |
     |     28 |     3.5 |   3.65 |  20 |    29 |    87.5 |    50 |     1 |
     |-----------------------------------------------------------------|
     |        soph         |          skipped         |       stndfnl     |
     |          0          |                4         |      .6827731     |
     +-----------------------------------------------------------------+
```

**Note:** You can also use the command:

```
.  list if priGPA==float(3.65)
```

to find the same student record.

The estimated equation is

$$\widehat{atndrte} = 75.70 + 17.26 priGPA - 1.72 ACT$$
$$n = 680, R^2 = .291.$$

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 $ACT = 0$.)

(iii) The coefficient on $priGPA$ means that, if a student's 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 coefficent on $ACT$ is, perhaps initially a bit surprising. Five more points on the $ACT$ is predicted to lower attenance 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) 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 with natural upper or lower bounds on the dependent variable. In practice, we would predict a 100% attendance rate for this student. (In fact, this student had an attendance rate of only 87.5%.)

(v) The difference in predicted attendance rates for A and B is $17.26(3.1 - 2.1) - 1.72(21 - 26) = 25.86$

**6.** (**4.1**) (i) and (iii) generally cause the $t$ statistics not to have a $t$ distribution under $H_0$. Homoskedasticity is one of the CLM assumptions. An important omitted variable violates Assumption MLR.3. The CLM assumptions contain no mention of the sample correlations among independent variables, except to rule out the case where the correlation is one.

**7.** (**4.8**)

(i) We use Property VAR.3 from Appendix B: $\text{Var}(\hat{\beta}_1 - 3\hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + 9\text{Var}(\hat{\beta}_2) - 6\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$.

(ii) $t = (\hat{\beta}_1 - 3\hat{\beta}_2 - 1)/\text{se}(\hat{\beta}_1 - 3\hat{\beta}_2)$, so we need the standard error of $\hat{\beta}_1 - 3\hat{\beta}_2$.

(iii) Because $\theta_1 = \beta_1 - 3\beta_2$, we can write $\beta_1 = \theta_1 + 3\beta_2$. Plugging this into the population model gives

$$\begin{aligned} y &= \beta_0 + (\theta_1 + 3\beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ &= \beta_0 + \theta_1 x_1 + \beta_2(3x_1 + x_2) + \beta_3 x_3 + u. \end{aligned}$$

This last equation is what we would estimate by regressing $y$ on $x_1$, $3x_1 + x_2$, and $x_3$. The coefficient and standard error on $x_1$ are what we want.

**8.** (**4.16**)

(i)

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/MLB1

. regress lsalary years gamesyr bavg hrunsyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  4,   348) =  145.24
       Model |  307.800712     4  76.950178            Prob > F      =  0.0000
    Residual |  184.374856   348  .529812806           R-squared     =  0.6254
```

```
-------------+------------------------------           Adj R-squared =  0.6211
       Total |  492.175568   352  1.39822605           Root MSE      =  .72788

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |   .0677325   .0121128     5.59   0.000     .0439089     .091556
     gamesyr |   .0157595   .0015636    10.08   0.000     .0126841    .0188348
        bavg |   .0014185   .0010658     1.33   0.184    -.0006776    .0035146
     hrunsyr |   .0359435   .0072408     4.96   0.000     .0217022    .0501847
       _cons |   11.02091   .2657191    41.48   0.000      10.4983    11.54353
------------------------------------------------------------------------------
```

If we drop *rbisyr*, the estimated equation becomes

$$\begin{aligned}
\log(\widehat{salary}) &= \underset{(0.27)}{11.02} + \underset{(.0121)}{.0677}\, years + \underset{(.0016)}{.0158}\, gamesyr \\
&\quad + \underset{(.0011)}{.0014}\, bavg + \underset{(.0072)}{.0359}\, hrunsyr \\
n &= 353, R^2 = .625.
\end{aligned}$$

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

(ii)

```
. regress lsalary years gamesyr bavg hrunsyr runsyr fldperc sbasesyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  7,   345) =   87.25
       Model |  314.510484     7  44.9300691           Prob > F      =  0.0000
    Residual |  177.665085   345   .51497126           R-squared     =  0.6390
-------------+------------------------------           Adj R-squared =  0.6317
       Total |  492.175568   352  1.39822605           Root MSE      =  .71761

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |   .0699848   .0119756     5.84   0.000     .0464305     .093539
     gamesyr |   .0078995   .0026775     2.95   0.003     .0026333    .0131657
        bavg |   .0005296   .0011038     0.48   0.632    -.0016414    .0027007
     hrunsyr |   .0232107   .0086392     2.69   0.008     .0062186    .0402028
      runsyr |   .0173921   .0050641     3.43   0.001     .0074318    .0273525
```

```
    fldperc |   .0010351   .0020046     0.52   0.606    -.0029077    .0049778
   sbasesyr |  -.0064191   .0051842    -1.24   0.216    -.0166156    .0037775
      _cons |   10.40827   2.003255     5.20   0.000     6.468142     14.3484
------------------------------------------------------------------------------
```

The equation with *runsyr*, *fldperc*, and *sbasesyr* added is

$$\begin{aligned}\log(\widehat{salary}) &= \underset{(2.00)}{10.41} + \underset{(.0120)}{.0700}\, years + \underset{(.0027)}{.0079}\, gamesyr \\ &\quad + \underset{(.00110)}{.0053}\, bavg + \underset{(.0086)}{.0232}\, hrunsyr \\ &\quad + \underset{(.0051)}{.0174}\, runsyr + \underset{(.0020)}{.0010}\, fldperc - \underset{(.0052)}{.0064}\, sbasesyr \\ n &= 353, R^2 = .639.\end{aligned}$$

Of the three additional independent variables, only *runsyr* is statistically significant ($t$ statistic $= .0175/.0051 \approx 3.41$). 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" sing 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)
```
. test bavg fldperc sbasesyr

 ( 1)  bavg = 0
 ( 2)  fldperc = 0
 ( 3)  sbasesyr = 0

       F(  3,   345) =    0.68
            Prob > F =    0.5617
```

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 .68 with $p$-value $\approx .56$. Therefore, these variables are jointly very insignificant.