6.2 This question asks you to use the results from column (1) in the table on page 213.

a. The first part of this question asks whether workers with college degrees earn more than workers with only a high school degree. Based on the regression results, workers with college degrees earn $5.46/hour more, on average, than workers with only high school degrees.

b. The second part of question asks a similar question, only in this case focusing on the wage differential for men versus women. The regression results indicate that men earn $2.64/hour more, on average, than women.

6.3 The next question asks you to use the results from column (2) in the table on page 213.

b. In the second part of the question, you are asked to predict the earnings for two individuals: Sally, who is a 29-year-old female college graduate, and Betsy, who is a 34-year-old female college graduate. Sally’s earnings prediction is $40 + 5.48 \times 1 - 2.62 \times 1 + 0.29 \times 29 = 15.67$ dollars per hour. Betsy’s earnings prediction is $40 + 5.48 \times 1 - 2.62 \times 1 + 0.29 \times 34 = 17.12$ dollars per hour. The difference is $1.45$ dollars per hour.

6.4 The next question asks you to use the results from column (3) in the table on page 213.

b. Here you are asked why the regressor West is excluded from the regression. The regressor West is omitted to avoid perfect multicollinearity. If West is included, then the intercept can be written as a perfect linear function of the four regional regressors. Because of perfect multicollinearity, the OLS estimator cannot be computed.

6.5 In question 6.5, you are to use the results from an analysis of housing prices.

b. Here you are asked to estimate the impact of an increase in house size by 100 square feet through the addition of a bathroom. In this case $\Delta BDR = 1$ and $\Delta H size = 100$. The resulting expected change in price is $23.4 + 0.156 \times 100 = 39.0$ thousand dollars or $39,000$.

6.7 This question continues question 6.4 above, providing standard errors for the estimated regression model, as reported in the table on page 247.

a. You are asked whether or not the regional differences appear to be important. The F-statistic testing the coefficients on the regional regressors are zero is 6.10. The 1% critical value (from the $F_{3,\infty}$ distribution) is 3.78. Because 6.10 > 3.78, the regional effects are significant at the 1% level.
The two empirical exercises in this homework use the same dataset: **CollegeDistance**. The data can be downloaded from the Web site listed in the assignment (which you can also reach from the class website). A program that carries all of the tasks for problems in E6.2 is appended to this answer sheet.

E6.2  

a. The first task you are asked to do is to regress the years of completed education (\( ED \)) on distance to the nearest college (\( Dist \)) and to report the estimated slope. The results are as follows: 

\[
\hat{ED} = 13.96 - 0.073 \times Dist, \ R^2 = 0.0074
\]

The slope, then, for this regression is -0.073.

b. Next, you are asked to run an additional regression including some of the other variables in the data set. The resulting parameter estimates are:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Est.</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>dist</td>
<td>0.032</td>
<td>0.012</td>
</tr>
<tr>
<td>bytest</td>
<td>0.093</td>
<td>0.0030</td>
</tr>
<tr>
<td>female</td>
<td>0.145</td>
<td>0.050</td>
</tr>
<tr>
<td>black</td>
<td>0.367</td>
<td>0.068</td>
</tr>
<tr>
<td>hispanic</td>
<td>0.398</td>
<td>0.074</td>
</tr>
<tr>
<td>incomehi</td>
<td>0.395</td>
<td>0.062</td>
</tr>
<tr>
<td>ownhome</td>
<td>0.152</td>
<td>0.065</td>
</tr>
<tr>
<td>dadcoll</td>
<td>0.696</td>
<td>0.071</td>
</tr>
<tr>
<td>cue80</td>
<td>0.023</td>
<td>0.009</td>
</tr>
<tr>
<td>stwmfg80</td>
<td>0.051</td>
<td>0.020</td>
</tr>
<tr>
<td>intercept</td>
<td>8.827</td>
<td>0.241</td>
</tr>
</tbody>
</table>

The estimated effect of \( Dist \) is now 0.032.

c. The coefficient has fallen by more than 50%. Thus, it seems that result in (a) did suffer from omitted variable bias.

d. The regression in (b) fits the data much better as indicated by the \( R^2 \), \( \bar{R}^2 \) and \( SER \). The \( R^2 \) and \( \bar{R}^2 \) are similar because the number of observations is large (\( n = 3796 \)).

e. Students with a \( dadcoll = 1 \) (so that the student’s father went to college) complete 0.696 more years of education, on average, than students with \( dadcoll = 0 \) (so that the student’s father did not go to college).

f. These terms capture the opportunity cost of attending college. As \( STWMFG80 \) increases, forgone wages increase, so that, on average, college attendance declines. The negative sign on the coefficient is consistent with this. As \( CUE80 \) increases, it is more difficult to find a job, which lowers the opportunity cost of attending college, so that college attendance increases. The positive sign on the coefficient is consistent with this.

g. Bob’s predicted years of education = \( 0.0315 \times 2 + 0.093 \times 58 + 0.145 \times 0 + 0.367 \times 1 + 0.398 \times 0 + 0.395 \times 1 + 0.152 \times 1 + 0.696 \times 0 + 0.023 \times 7.5 + 0.051 \times 9.75 + 8.827 = 14.75 \). The program computes this more precisely using the \textit{lincom} command.

h. Jim’s expected years of education is \( 2 \times 0.0315 = 0.0630 \) less than Bob’s. Thus, Jim’s expected years of education is \( 14.75 - 0.063 = 14.69 \).

E6.2  These are the answers to the additional questions.

a. The first additional question asks you to construct a 90% confidence interval around the predictions in parts g and h. This can be read directly from the Stata output using the \textit{lincom} command and the \textit{level(90)} option. Specifically, the 90% confidence interval for part g is given by: (14.63886, 14.94217). The 90% confidence interval for part h is (14.56512,14.88975).

b. The second question asks you to test the hypothesis that the additional variables in E6.2b are jointly significant. This is done using the \textit{test} command after the regression. In this case, the F-statistic is 215.43 and the p-value associate with the test being \(< 0.0001\), so we would reject this restriction. The more complicated model is a statistically significant improvement on the basic model at the 10%, 5%, and 1% levels.
c. Finally, you are asked to test the hypothesis that the coefficients on Black and Hispanic are the same. Again, we can use the test command after the regression to test this hypothesis. This gives us an F-statistic of 0.13, with a p-value of 0.7168. Clearly, we would not reject the null hypothesis. At least based on these data, the additional years of education completed by these two sub-populations, conditional on all the other factors, are the same.
Problem Set #4

```
# delimit ;
clear;
cap log close;
set more off;
cd "R:\users\jaherrig\My Documents\Classes\Economics 371\Stata";
log using Problemset4.log,replace;
define
********************************************************************************;
*       Specify the output file
*       *********************************************;
log using Problemset4.log,replace;
set more off;
********************************************************************************;
*       Read in and summarize the data
*       *********************************************;
use     CollegeDistance.dta;
describe;
summarize;
********************************************************************************;
*       Estimate the model for question E6.2a
*       *********************************************;
reg     ed dist,r;
reg     ed dist;
********************************************************************************;
*       Estimate the model for question E6.2b. Also, include a test of two
*       hypotheses:
*       First, that the additional variables jointly have zero coefficients
*       Second, that the black and hispanic coefficients are the same
*       *********************************************;
reg     ed dist bytest female black hispanic incomehi ownhome dadcoll cue80 stwmfg80,r;
test    bytest female black hispanic incomehi ownhome dadcoll cue80 stwmfg80;
test    black=hispanic;
reg     ed dist bytest female black hispanic incomehi ownhome dadcoll cue80 stwmfg80;
********************************************************************************;
*       Compute the fitted value of ED for E6.2g and E6.2h
*       *********************************************;
lincom  _cons + 2*dist + 58*bytest + 0*female + 1*black + 0*hispanic +
        1*incomehi + 1*ownhome + 0*dadcoll + 7.5*cue80 + 9.75*stwmfg80, level(90);
lincom  _cons + 4*dist + 58*bytest + 0*female + 1*black + 0*hispanic +
        1*incomehi + 1*ownhome + 0*dadcoll + 7.5*cue80 + 9.75*stwmfg80, level(90);
log close;
```
*clear;
*exit;
log:  R:\users\jaherrig\My Documents\Classes\Economics 371\Stata \Problemset4.log  
log type:  text  
opened on:  14 Oct 2009, 08:30:56

. set more off;

. ********************************************************************************;
. *       Read in and summarize the data
. * Read in and summarize the data
> ********************************************************************************;
. use  CollegeDistance.dta;

. describe;

Contains data from CollegeDistance.dta
obs:       3,796
vars:        14                  1 Aug 2006 17:31
size:   227,760 (78.3% of memory free)

-------------------------------------------------------------------------------------
storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------
female          float  %9.0g
black           float  %9.0g
hispanic        float  %9.0g
bytest          float  %9.0g
dadcoll         float  %9.0g
momcoll         float  %9.0g
ownhome         float  %9.0g
urban           float  %9.0g
cue80           float  %9.0g
stwmfg80        float  %9.0g
dist            float  %9.0g
tuition         float  %9.0g
incomehi        float  %9.0g
ed              float  %9.0g
-------------------------------------------------------------------------------------

Sorted by:

. summarize ;

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>3796</td>
<td>.5453109</td>
<td>.4980083</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>black</td>
<td>3796</td>
<td>.1925711</td>
<td>.394371</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>hispanic</td>
<td>3796</td>
<td>.1498946</td>
<td>.3570151</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>bytest</td>
<td>3796</td>
<td>51.00193</td>
<td>8.819251</td>
<td>28.95</td>
<td>71.36</td>
</tr>
<tr>
<td>dadcoll</td>
<td>3796</td>
<td>.2020548</td>
<td>.4015858</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>momcoll</td>
<td>3796</td>
<td>.1393572</td>
<td>.3463645</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ownhome</td>
<td>3796</td>
<td>.8192835</td>
<td>.3848338</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>urban</td>
<td>3796</td>
<td>.243941</td>
<td>.4295141</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>cue80</td>
<td>3796</td>
<td>7.654874</td>
<td>2.86577</td>
<td>1.4</td>
<td>24.9</td>
</tr>
</tbody>
</table>
Estimate the model for question E6.2a

```
. reg ed dist, r;
```

```
Linear regression
Number of obs = 3796
F( 1, 3794) = 29.83
Prob > F = 0.0000
R-squared = 0.0074
Root MSE = 1.8074

------------------------------------------------------------------------------
|               Robust
ed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dist |  -.0733727   .0134334    -5.46   0.000    -.0997101   -.0470353
_cons |   13.95586   .0378112   369.09   0.000     13.88172    14.02999
------------------------------------------------------------------------------
```

Estimate the model for question E6.2b. Also, include a test of two hypotheses:
First, that the additional variables jointly have zero coefficients
Second, that the black and hispanic coefficients are the same

```
. reg ed dist bytest female black hispanic incomehi ownhome dadcoll cue80
   stwmfg80, r;
```

```
Linear regression
Number of obs = 3796
F( 10, 3785) = 197.68
Prob > F = 0.0000
R-squared = 0.2788
Root MSE = 1.5425

------------------------------------------------------------------------------
|               Robust
ed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dist |  -.0733727   .0137498    -5.34   0.000    -.1003304    -.046415
_cons |   13.95586   .0377241   369.95   0.000     13.88189    14.02999
------------------------------------------------------------------------------
```
estat summarize

## estat summarize

| ed   | Coef. | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|------|-------|-----------|-------|------|----------------------|
| dist | -0.0315387 | 0.0116616 | -2.70 | 0.007 | -0.0544023           | -0.0086752 |
| bytest | 0.0938201 | 0.0029804 | 31.48 | 0.000 | 0.0879768           | 0.0996634 |
| female | 0.145408 | 0.0503939 | 2.89  | 0.004 | 0.0466061           | 0.2442098 |
| black | 0.367971  | 0.0675359 | 5.45  | 0.000 | 0.2355608           | 0.5003812 |
| hispanic | 0.3985196 | 0.0738763 | 5.39  | 0.000 | 0.2536785           | 0.5433608 |
| incomehi | 0.3951984 | 0.0738763 | 5.39  | 0.000 | 0.2536785           | 0.5433608 |
| ownhome | 0.1521313 | 0.0649193 | 2.34  | 0.019 | 0.0248511           | 0.2794115 |
| dadcoll | 0.6961324 | 0.0707602 | 9.84  | 0.000 | 0.5574006           | 0.8348641 |
| cue80 | 0.0232052 | 0.00931  | 2.49  | 0.013 | 0.0049521           | 0.0414583 |
| stwmfg80 | -0.0517777 | 0.0196751 | -2.63 | 0.009 | -0.0903526          | -0.0132029 |
| _cons | 8.827518 | 0.2413001 | 36.58 | 0.000 | 8.354427            | 9.300609  |

### . test bytest female black hispanic incomehi ownhome dadcoll cue80 stwmfg80;

( 1)  bytest = 0
( 2)  female = 0
( 3)  black = 0
( 4)  hispanic = 0
( 5)  incomehi = 0
( 6)  ownhome = 0
( 7)  dadcoll = 0
( 8)  cue80 = 0
( 9)  stwmfg80 = 0

F( 9, 3785) = 215.43
Prob > F =  0.0000

### . test black=hispanic;

( 1)  black - hispanic = 0

F( 1, 3785) = 0.13
Prob > F =  0.7168

### . reg ed dist bytest female black hispanic incomehi ownhome dadcoll cue80 > stwmfg80;

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 3796</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3481.95254</td>
<td>10</td>
<td>348.195254</td>
<td>F( 10, 3785) = 146.35</td>
</tr>
<tr>
<td>Residual</td>
<td>9005.42997</td>
<td>3785</td>
<td>2.37924173</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>12487.3825</td>
<td>3795</td>
<td>3.29048287</td>
<td>Adj R-squared = 0.2769</td>
</tr>
</tbody>
</table>

| ed   | Coef. | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|------|-------|-----------|-------|------|----------------------|
| dist | -0.0315387 | 0.0123703 | -2.55 | 0.011 | -0.0557918           | -0.0086752 |
| bytest | 0.0938201 | 0.0031622 | 29.67 | 0.000 | 0.0876204           | 0.1000199 |
| female | 0.145408 | 0.0503939 | 2.89  | 0.004 | 0.0466061           | 0.2442098 |
| black | 0.367971  | 0.071363  | 5.16  | 0.000 | 0.2280574           | 0.5078846 |
| hispanic | 0.3985196 | 0.0744617 | 5.35  | 0.000 | 0.2525308           | 0.5445085 |
| incomehi | 0.3951984 | 0.0744617 | 5.35  | 0.000 | 0.2525308           | 0.5445085 |
| ownhome | 0.1521313 | 0.0668075 | 2.28  | 0.023 | 0.0211492           | 0.2831135 |
. lincom _cons + 2*dist + 58*bytest + 0*female + 1*black + 0*hispanic +
   I*incomehi + 1*ownhome + 0*dadcoll + 7.5*cue80 + 9.75*stwmfg80,
   level(90);

( 1)  2 dist + 58 bytest + black + incomehi + ownhome + 7.5 cue80 + 9.75 stwmfg80 +
   _cons = 0

|     | Coef.  | Std. Err. |   t   |     P>|t|   |   [90% Conf. Interval] |
|-----|--------|-----------|-------|-------|------------------------|
| (1) | 14.79051 | .0921789 | 160.45 | 0.000 | 14.63886               | 14.94217 |

. lincom _cons + 4*dist + 58*bytest + 0*female + 1*black + 0*hispanic +
   I*incomehi + 1*ownhome + 0*dadcoll + 7.5*cue80 + 9.75*stwmfg80,
   level(90);

( 1)  4 dist + 58 bytest + black + incomehi + ownhome + 7.5 cue80 + 9.75 stwmfg80 +
   _cons = 0

|     | Coef.  | Std. Err. |   t   |     P>|t|   |   [90% Conf. Interval] |
|-----|--------|-----------|-------|-------|------------------------|
| (1) | 14.72744 | .0986563 | 149.28 | 0.000 | 14.56512               | 14.88975 |

. log close;
log:  R:\users\jaherrig\My Documents\Classes\Economics 371\Stata\Problemset4.log
log type:  text
closed on:  14 Oct 2009, 08:30:57