Homework #4

A. Shawn Bandy

March 7th, 2013

1 Lab Problems

L1 STATA code:

```
/*  A. Shawn Bandy
Lab #4  - L1
*/
/* close previous run do-files */
cap log close
set more 1
clear
#delimit ;

cd "C:\Users\cla-spa206.CAMPUS-DOMAIN\Desktop\econ485-lab4\lab4";
log using arclab1.log , replace;
use "arclab";

/*a. Create a dummy variable, arc that is equal to one if the arc_county variable
 indicates it is an ARC county. Otherwise, non-ARC counties should have a
 value of zero. This variable will be used to see if there is a difference
 in employment growth between the ARC and non-ARC counties.*/

 gen arc=1 if arc_county=="ARC";
 replace arc=0 if arc==.;

 /*b. Create a set of dummy variables for each of the 13 states in the ARC region.
Use tabulate with the prefix state to create these variables. Hint: the state
variable has the unique names of the 13 states for each county in the dataset.*/

tabulate state, gen(state);

/*c. Create a new variable empgrowth_9006 that is the percent change in employment
 from 1990 to 2006. Use the label command to label this Percent change in
 employment from 1990 to 2006. Make sure you save the data as arcdata1.dta so you
 can use it for the next problem.*/

 gen empgrowth_9006 = (emp06 - emp90) / emp90;
```

```
save arcdata_n.dta, replace;

/*d. Use the summarize command to look at the description of the variables in
 this dataset. */
summarize;

/*e. Use the regression command to estimate the following linear
 regression model, under the assumption that college educated people
 contribute to employment growth: */
regress empgrowth_9006 percoll90;

/*f. What is the estimated slope coefficient for percoll90?
 What is the interpretation of this slope coefficient?

 INTERPRETATION

 */

 /*g. Use the regression command to estimate the following
linear regression model, where we now test whether college
educated people and having a higher percentage of self-employed
individuals are important to employment growth in this region:*/
regress empgrowth_9006 percoll90 perse90;

/*h. What is the estimated slope coefficient for percoll90 in
 the model from part g? What is the interpretation of this slope coefficient?

 INTERPRETATION

 */

/*i. Using your results from parts e and g, does the regression model in part
e suffer from omitted variable bias? Explain.

INTERPRETATION

*/


/*j. Use the regression command to estimate the following linear
 regression model:*/

 regress percoll90 perse90;

/*k. Use the predict command to capture the residuals from the regression in
 part j in a new variable named res.*/

 predict residual, res;
```

```
/*1. Use the covariance option of the correlate command to examine the
 covariance between res and empgrowth_9006.*/
 correlate res empgrowth_9006, covariance;

save arcdata1.dta, replace;
log close;
```

Log output:

```
-------------------------------------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:\Users\cla-spa206.CAMPUS-DOMAIN\Downloads\arclab.log
  log type:  text
 opened on:   5 Mar 2013, 11:44:36

. use "arclab";

. /*a. Create a dummy variable, arc that is equal to one if the arc_county variable
>  indicates it is an ARC county. Otherwise, non-ARC counties should have a
>  value of zero. This variable will be used to see if there is a difference
>  in employment growth between the ARC and non-ARC counties.*/
>
>  gen arc=1 if arc_county=="ARC";
(138 missing values generated)

.  replace arc=0 if arc==.;
(138 real changes made)

.  /*b. Create a set of dummy variables for each of the 13 states in the ARC region.
> Use tabulate with the prefix state to create these variables. Hint: the state
> variable has the unique names of the 13 states for each county in the dataset.*/
>
> tabulate state, gen(state);

      state |      Freq.     Percent        Cum.
------------+-----------------------------------
         AL |         47        8.44        8.44
         GA |         50        8.98       17.41
         KY |         74       13.29       30.70
         MD |          4        0.72       31.42
         MS |         36        6.46       37.88
         NC |         36        6.46       44.34
         NY |         33        5.92       50.27
         OH |         45        8.08       58.35
         PA |         59       10.59       68.94
         SC |         11        1.97       70.92
         TN |         69       12.39       83.30
         VA |         38        6.82       90.13
         WV |         55        9.87      100.00
------------+-----------------------------------
```

```
      Total |        557      100.00

. /*c. Create a new variable empgrowth_9006 that is the percent change in employment
>  from 1990 to 2006. Use the label command to label this Percent change in
>  employment from 1990 to 2006. Make sure you save the data as arcdata1.dta so you
>  can use it for the next problem.*/
>
>  gen empgrowth_9006 = (emp06 - emp90) / emp90;
(4 missing values generated)

. save arcdata_n.dta, replace;
file arcdata_n.dta saved

. /*d. Use the summarize command to look at the description of the variables in
>  this dataset. */
> summarize;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fips |       558    33813.85    15810.15       1001      54109
       state |         0
      county |         0
      manu90 |       554    6446.449    11222.06         13     118484
      farm90 |       554    925.0144    653.6925          0       4762
-------------+--------------------------------------------------------
   percoll90 |       554    11.39375    5.642082   3.689338   41.72341
       emp90 |       557     35498.5    79429.55        795     819868
       emp06 |       554    44022.81     94828.9        897     963372
  arc_county |         0
    totpop90 |       554    67051.16    125637.7       2124    1336449
-------------+--------------------------------------------------------
       pop60 |       554    54965.85    114567.6       2443    1628587
  popsqmi_60 |       554    112.2741     239.375          8       3735
       rural |       554    .3122744    .4638399          0          1
   permanu90 |       554    21.17014    10.98425   .7445443   53.52263
   perfarm90 |       554    7.988959    7.926842          0   55.84906
-------------+--------------------------------------------------------
       pci90 |       554    14090.18    2814.397       7825      25984
     perse90 |       557    16.13464    4.663423   4.079602   38.20309
pci90_thou~s |       554    14.09018    2.814397      7.825     25.984
         arc |       558    .7526882    .4318366          0          1
      state1 |       557    .0843806    .2782076          0          1
-------------+--------------------------------------------------------
      state2 |       557    .0897666     .286104          0          1
      state3 |       557    .1328546    .3397226          0          1
      state4 |       557    .0071813    .0845138          0          1
      state5 |       557     .064632    .2460963          0          1
      state6 |       557     .064632    .2460963          0          1
-------------+--------------------------------------------------------
```

```
      state7 |       557     .059246    .2362967          0          1
      state8 |       557    .0807899    .2727572          0          1
      state9 |       557    .1059246    .3080177          0          1
     state10 |       557    .0197487    .1392604          0          1
     state11 |       557    .1238779    .3297384          0          1
-------------+--------------------------------------------------------
     state12 |       557    .0682226    .2523542          0          1
     state13 |       557    .0987433    .2985852          0          1
empgrow~9006 |       554    .3210966    .8151637  -.7296684   15.54692

. /*e. Use the regression command to estimate the following linear
>  regression model, under the assumption that college educated people
>  contribute to employment growth: */
> regress empgrowth_9006 percoll90;

      Source |       SS       df       MS              Number of obs =     554
-------------+------------------------------           F(  1,   552) =    3.42
       Model |  2.26518362     1  2.26518362           Prob > F      =  0.0648
    Residual |  365.198846   552  .661592112           R-squared     =  0.0062
-------------+------------------------------           Adj R-squared =  0.0044
       Total |   367.46403   553  .664491916           Root MSE      =  .81338

------------------------------------------------------------------------------
empgrow~9006 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   percoll90 |   .0113436   .0061305     1.85   0.065    -.0006983    .0233855
       _cons |   .1918507     .07793     2.46   0.014     .0387751    .3449264
------------------------------------------------------------------------------

. /*f. What is the estimated slope coefficient for percoll90?
>  What is the interpretation of this slope coefficient?
>
>  INTERPRETATION
>
>  */
>
>  /*g. Use the regression command to estimate the following
> linear regression model, where we now test whether college
> educated people and having a higher percentage of self-employed
> individuals are important to employment growth in this region:*/
> regress empgrowth_9006 percoll90 perse90;

      Source |       SS       df       MS              Number of obs =     554
-------------+------------------------------           F(  2,   551) =   24.61
       Model |  30.1386418     2  15.0693209           Prob > F      =  0.0000
    Residual |  337.325388   551  .612205785           R-squared     =  0.0820
-------------+------------------------------           Adj R-squared =  0.0787
       Total |   367.46403   553  .664491916           Root MSE      =  .78244
```

```
------------------------------------------------------------------------------
empgrow~9006 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   percoll90 |   .0180694   .0059809     3.02   0.003     .0063213    .0298175
     perse90 |   .0490121   .0072637     6.75   0.000     .0347442      .06328
       _cons |  -.6767699   .1489679    -4.54   0.000    -.9693844   -.3841554
------------------------------------------------------------------------------

. /*h. What is the estimated slope coefficient for percoll90 in
>  the model from part g? What is the interpretation of this slope coefficient?
>
>  INTERPRETATION
>
>  */
>
> /*i. Using your results from parts e and g, does the regression model in part
> e suffer from omitted variable bias? Explain.
>
> INTERPRETATION
>
> */
>
>
> /*j. Use the regression command to estimate the following linear
>  regression model:*/
>
>  regress percoll90 perse90;

      Source |       SS       df       MS              Number of obs =     554
-------------+------------------------------           F(  1,   552) =   15.77
       Model |  488.962959     1  488.962959           Prob > F      =  0.0001
    Residual |  17114.7368   552  31.0049579           R-squared     =  0.0278
-------------+------------------------------           Adj R-squared =  0.0260
       Total |  17603.6997   553  31.8330917           Root MSE      =  5.5682

------------------------------------------------------------------------------
   percoll90 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     perse90 |  -.2024086   .0509691    -3.97   0.000    -.3025257   -.1022916
       _cons |   14.66448   .8569131    17.11   0.000     12.98127    16.34769
------------------------------------------------------------------------------

.  /*k. Use the predict command to capture the residuals from the regression in
>  part j in a new variable named res.*/
>
>  predict residual, res;
(4 missing values generated)

.  /*l. Use the covariance option of the correlate command to examine the
```

```
>  covariance between res and empgrowth_9006.*/
>  correlate res empgrowth_9006, covariance;
(obs=554)

             | residual emp~9006
-------------+------------------
    residual |  30.9489
empgrow~9006 |  .559228  .664492


.  save arcdata_n.dta, replace;
file arcdata_n.dta saved

. log close;
      name:  <unnamed>
       log:  C:\Users\cla-spa206.CAMPUS-DOMAIN\Downloads\arclab.log
  log type:  text
 closed on:   5 Mar 2013, 11:44:36
-------------------------------------------------------------------------------------------------------------------
```

Answers:

f. What is the estimated slope coefficient for percoll90? What is the interpretation of this slope coefficient?

The estimated slope is .0113436. For each unit change in percoll90, there is a 0.113436 unit change in employment growth between 1990 and 2006. It should be noted that in this regression the value for percoll90 is not within the 95% confidence interval and so there may be no realistic interpretation for the coefficient.

h. What is the estimated slope coefficient for percoll90 in the model from part g? What is the interpretation of this slope coefficient?

The estimate slope is .0180694. For each unit change in percoll90, there is a .0180694 unit change in employment growth between 1990 and 2006, holding perse90 constant. In this case the coefficient is in the 95% confidence interval although it has very little impact on the dependent variable.

i. Using your results from parts e and g, does the regression model in part e suffer from omitted variable bias? Explain.

The model in part e suffers from omitted bias only if the additional variable, perse90, has a zero coefficient in the model in part g *and* the omitted variable is correlated with percoll90. The coefficient for perse90 in model g is *not* zero within the 95% confidence interval and there is a negative correlation between the variables. We can conclude that there is omitted variable bias in model e.

L2 STATA code:

```
/*  A. Shawn Bandy
Lab #4
*/
/* close previous run do-files */
cap log close
set more 1
clear
#delimit ;

cd "C:\Users\cla-spa206.CAMPUS-DOMAIN\Desktop\econ485-lab4\lab4";
log using arclab2.log , replace;
use "arcdata1";

/* a. Now, we want to test if additional variables would add to the explanatory
value of employment growth in the region. Add the following additional variables
 to the model from L1 part g: pci90_thousands and pci90. */

 regress percoll90 perse90 pci90_thousands pci90;

/* b. Does the STATA output from part a) include both the new variables?
Explain what happened.*/

/*INTERPRETATION */

/* c. Now, use the regression command to estimate a linear regression where
 y=empgrowth_9006 and x includes percoll90, perse90, and the dummy variable arc.*/

 regress empgrowth_9006 percoll90 perse90 arc;

/* d. What is the coefficient on arc? What is the interpretation of this
coefficient in our model? (Hint: Look at the t-statistic!) */

/* INTERPRETATION */

/* e. Now, use the regression command to estimate a linear regression where
 y=empgrowth_9006 and x includes percoll90, perse90, and dummy variables for
 each of the states in the model (state1, state13). Exclude the dummy
 variable for state2 (Georgia) from the model.*/

 regress empgrowth_9006 percoll90 perse90 state1 state3-state13;

/* f. Why did we exclude Georgia from the model in part e? What is
the interpretation of the coefficient on state1 (Alabama)?*/

/* INTERPRETATION */

/* g. Now, add the following additional variables to the model from part e:
 popsqmi_60 and rural. */
```

```
  regress empgrowth_9006 percoll90 perse90 state1 state3-state13 popsqmi_60 rural;

/* h. Test for multicollinearity in your data by running the vif command.*/

vif;

/* i. Are there any problems with multicollinearity in your model? Explain.*/

/* INTERPRETATION */

/* j. Compare the final model in part g to the model in L1, part g in terms of
 how much they explain the variance in employment growth. Explain.*/

 /*INTERPRETATION */
```

Log output:

```
-------------------------------------------------------------------------------
      name:  <unnamed>
       log:  /Users/shawn/src/econ485/lab4/arclab2.log
  log type:  text
 opened on:  11 Mar 2013, 20:28:01

. use "arcdata1";

. /* a. Now, we want to test if additional variables would add to the explanatory
> value of employment growth in the region. Add the following additional variables
>  to the model from L1 part g: pci90_thousands and pci90. */
>
>  regress empgrowth_9006 percoll90 perse90 pci90_thousands pci90;
note: pci90_thousands omitted because of collinearity

      Source |       SS       df       MS              Number of obs =     551
-------------+------------------------------           F(  3,   547) =   16.76
       Model |   30.917101     3  10.3057003           Prob > F      =  0.0000
    Residual |   336.41239   547   .61501351           R-squared     =  0.0842
-------------+------------------------------           Adj R-squared =  0.0791
       Total |  367.329491   550  .667871801           Root MSE      =  .78423

------------------------------------------------------------------------------
 empgrowth_9006 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
      percoll90 |   .0123651    .007772     1.59   0.112    -.0029015    .0276317
        perse90 |   .0500949   .0073594     6.81   0.000     .0356387    .0645511
pci90_thousands |          0  (omitted)
          pci90 |   .0000186   .0000156     1.19   0.233     -.000012    .0000492
          _cons |  -.8912981    .233067    -3.82   0.000    -1.349114   -.4334821
------------------------------------------------------------------------------

.  /* b. Does the STATA output from part a) include both the new variables?
```

```
> Explain what happened.*/
>
> /*INTERPRETATION */
>
> /* c. Now, use the regression command to estimate a linear regression where
>  y=empgrowth_9006 and x includes percoll90, perse90, and the dummy variable arc.*/
>
>  regress empgrowth_9006 percoll90 perse90 arc;

      Source |       SS       df       MS              Number of obs =     554
-------------+------------------------------           F(  3,   550) =   16.49
       Model |  30.3162271     3   10.105409           Prob > F      =  0.0000
    Residual |  337.147802   550  .612996004           R-squared     =  0.0825
-------------+------------------------------           Adj R-squared =  0.0775
       Total |   367.46403   553  .664491916           Root MSE      =  .78294

------------------------------------------------------------------------------
empgrow~9006 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   percoll90 |   .0171481   .0062247     2.75   0.006      .004921    .0293752
     perse90 |   .0495403   .0073343     6.75   0.000     .0351336    .0639471
         arc |  -.0441713   .0820664    -0.54   0.591    -.2053732    .1170306
       _cons |  -.6413213   .1629652    -3.94   0.000    -.9614317   -.3212109
------------------------------------------------------------------------------

.  /* d. What is the coefficient on arc? What is the interpretation of this
> coefficient in our model? (Hint: Look at the t-statistic!) */
>
> /* INTERPRETATION */
>
> /* e. Now, use the regression command to estimate a linear regression where
>  y=empgrowth_9006 and x includes percoll90, perse90, and dummy variables for
>  each of the states in the model (state1, state13). Exclude the dummy
>  variable for state2 (Georgia) from the model.*/
>
>  regress empgrowth_9006 percoll90 perse90 state1 state3-state13;

      Source |       SS       df       MS              Number of obs =     554
-------------+------------------------------           F( 14,   539) =    7.87
       Model |  62.3617058    14  4.45440755           Prob > F      =  0.0000
    Residual |  305.102324   539  .566052549           R-squared     =  0.1697
-------------+------------------------------           Adj R-squared =  0.1481
       Total |   367.46403   553  .664491916           Root MSE      =  .75236

------------------------------------------------------------------------------
empgrow~9006 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   percoll90 |   .0194194   .0062607     3.10   0.002     .0071211    .0317178
     perse90 |   .0504565   .0073927     6.83   0.000     .0359344    .0649785
```

```
      state1 |   -.337751    .156141    -2.16   0.031    -.6444705   -.0310316
      state3 |  -.4122454   .1414116    -2.92   0.004    -.6900307   -.1344601
      state4 |  -.3204748    .391161    -0.82   0.413    -1.088862    .4479122
      state5 |  -.2342322   .1698213    -1.38   0.168    -.5678249    .0993604
      state6 |  -.4584991   .1645499    -2.79   0.006    -.7817368   -.1352615
      state7 |   -.696958   .1727595    -4.03   0.000    -1.036322   -.3575936
      state8 |  -.4095359   .1557558    -2.63   0.009    -.7154986   -.1035731
      state9 |  -.5738907   .1446458    -3.97   0.000    -.8580292   -.2897521
     state10 |   -.303029   .2549405    -1.19   0.235    -.8038276    .1977697
     state11 |  -.3601028   .1420734    -2.53   0.012    -.6391883   -.0810173
     state12 |   .2990636   .1622676     1.84   0.066    -.0196908    .6178179
     state13 |  -.5381868   .1482697    -3.63   0.000    -.8294441   -.2469296
       _cons |  -.3670907    .197466    -1.86   0.064     -.754988    .0208067
------------------------------------------------------------------------------

.  /* f. Why did we exclude Georgia from the model in part e? What is
> the interpretation of the coefficient on state1 (Alabama)?*/
>
> /* INTERPRETATION */
>
> /* g. Now, add the following additional variables to the model from part e:
>  popsqmi_60 and rural. */
>
>   regress empgrowth_9006 percoll90 perse90 state1 state3-state13 popsqmi_60 rural;

      Source |       SS       df       MS              Number of obs =     554
-------------+------------------------------           F( 16,   537) =    6.89
       Model |  62.5846817    16   3.9115426           Prob > F      =  0.0000
    Residual |  304.879348   537  .567745527           R-squared     =  0.1703
-------------+------------------------------           Adj R-squared =  0.1456
       Total |   367.46403   553  .664491916           Root MSE      =  .75349

------------------------------------------------------------------------------
empgrow~9006 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   percoll90 |   .0195789   .0065059     3.01   0.003     .0067988     .032359
     perse90 |   .0496949   .0077024     6.45   0.000     .0345643    .0648255
      state1 |  -.3446213   .1569492    -2.20   0.029     -.652931   -.0363117
      state3 |  -.3966821   .1444483    -2.75   0.006     -.680435   -.1129291
      state4 |   -.331701   .3922187    -0.85   0.398    -1.102172    .4387699
      state5 |  -.2158291   .1750308    -1.23   0.218    -.5596581    .1279998
      state6 |  -.4553793   .1649681    -2.76   0.006    -.7794412   -.1313174
      state7 |  -.7026335   .1733895    -4.05   0.000    -1.043238   -.3620286
      state8 |  -.4128032   .1567536    -2.63   0.009    -.7207286   -.1048778
      state9 |   -.574764   .1455941    -3.95   0.000    -.8607678   -.2887602
     state10 |  -.3148877   .2561012    -1.23   0.219    -.8179707    .1881953
     state11 |  -.3582187   .1423211    -2.52   0.012     -.637793   -.0786444
     state12 |   .2898573   .1638023     1.77   0.077    -.0319146    .6116291
     state13 |  -.5252413   .1500485    -3.50   0.001    -.8199954   -.2304872
```

```
  popsqmi_60 |   -.000048     .00015    -0.32   0.749    -.0003427    .0002467
       rural |   -.045763   .0802845    -0.57   0.569    -.2034732    .1119472
       _cons |  -.3396657   .2030335    -1.67   0.095     -.738503    .0591716
------------------------------------------------------------------------------

.  /* h. Test for multicollinearity in your data by running the vif command.*/
>
> vif;

    Variable |       VIF       1/VIF
-------------+----------------------
      state3 |      2.36    0.424391
     state11 |      2.13    0.469872
      state9 |      1.97    0.508064
     state13 |      1.96    0.509022
      state1 |      1.87    0.535845
      state8 |      1.79    0.558854
      state5 |      1.77    0.565196
     state12 |      1.67    0.597847
      state6 |      1.61    0.619771
      state7 |      1.60    0.626321
       rural |      1.35    0.740336
   percoll90 |      1.31    0.761969
  popsqmi_60 |      1.26    0.795968
     perse90 |      1.25    0.801824
     state10 |      1.25    0.802875
      state4 |      1.08    0.929361
-------------+----------------------
    Mean VIF |      1.64

. /* i. Are there any problems with multicollinearity in your model? Explain.*/
>
> /* INTERPRETATION */
>
> /* j. Compare the final model in part g to the model in L1, part g in terms of
>  how much they explain the variance in employment growth. Explain.*/
>
>  /*INTERPRETATION */

end of do-file

. correlate empgrowth_9006 percoll90 perse90 state1 state3-state13 popsqmi_60 rural
(obs=554)

             | emp~9006 perco~90  perse90   state1   state3   state4   state5
-------------+---------------------------------------------------------------
empgrow~9006 |   1.0000
   percoll90 |   0.0785   1.0000
     perse90 |   0.2585  -0.1667   1.0000
```

```
      state1 |  -0.0424   0.0055  -0.1638   1.0000
      state3 |  -0.0634  -0.1902  -0.0244  -0.1195   1.0000
      state4 |   0.0052   0.0344  -0.0081  -0.0260  -0.0335   1.0000
      state5 |  -0.0126  -0.0423  -0.1507  -0.0791  -0.1020  -0.0221   1.0000
      state6 |  -0.0160   0.0637   0.0384  -0.0803  -0.1035  -0.0225  -0.0685
      state7 |  -0.0668   0.2771   0.0067  -0.0754  -0.0972  -0.0211  -0.0643
      state8 |  -0.0284  -0.0262  -0.0091  -0.0905  -0.1167  -0.0254  -0.0772
      state9 |  -0.0581   0.0751   0.0953  -0.1051  -0.1356  -0.0294  -0.0897
     state10 |  -0.0310   0.0361  -0.1523  -0.0433  -0.0559  -0.0121  -0.0370
     state11 |  -0.0048  -0.1238   0.0597  -0.1139  -0.1469  -0.0319  -0.0971
     state12 |   0.2256   0.0314   0.0196  -0.0826  -0.1066  -0.0231  -0.0705
     state13 |  -0.0656  -0.0809   0.0786  -0.1011  -0.1304  -0.0283  -0.0862
  popsqmi_60 |  -0.0507   0.3227  -0.2656  -0.0575  -0.0742   0.0076  -0.0780
       rural |  -0.0794  -0.2025  -0.0268  -0.0794   0.2735  -0.0575   0.2573

             |   state6   state7   state8   state9  state10  state11  state12
-------------+---------------------------------------------------------------
      state6 |   1.0000
      state7 |  -0.0653   1.0000
      state8 |  -0.0784  -0.0736   1.0000
      state9 |  -0.0910  -0.0855  -0.1027   1.0000
     state10 |  -0.0375  -0.0352  -0.0423  -0.0491   1.0000
     state11 |  -0.0986  -0.0926  -0.1112  -0.1291  -0.0532   1.0000
     state12 |  -0.0715  -0.0672  -0.0807  -0.0937  -0.0386  -0.1015   1.0000
     state13 |  -0.0875  -0.0822  -0.0987  -0.1146  -0.0473  -0.1242  -0.0901
  popsqmi_60 |  -0.0203   0.0846   0.0835   0.1109  -0.0057  -0.0509   0.0518
       rural |   0.0120  -0.1334  -0.1291  -0.1316  -0.0680  -0.0028  -0.1829

             |  state13 popsq~60    rural
-------------+---------------------------
     state13 |   1.0000
  popsqmi_60 |  -0.0200   1.0000
       rural |   0.1540  -0.1821   1.0000


. do "/var/folders/8r/6p_8585x5435jsc5wc_xdv5w0000gn/T//SD66697.000000"

. /*      A. Shawn Bandy
>         Lab #4  - L1
> */
. /* close previous run do-files */
. cap log close
```

Answers:

b. Does the STATA output from part a) include both the new variables? Explain what happened.

STATA omits an independent variable when there is a dependency between it and one or more other variables. Running *regress pci90_thousands percoll90 perse90 pci90* shows that there is a dependency between pci90_thousands and pci90.

d. What is the coefficient on arc? What is the interpretation of this coefficient in our model?

The coefficient for arc is -.0441713, so for each unit change in the variable arc there is about a 4% drop in employment growth between 1990 and 2006. Because the t-stat is -0.54 and compounded by having a coefficient fairly close to zero, we should interpret this as almost certainly having no meaning in our model.

f. Why did we exclude Georgia from the model in part e? What is the interpretation of the coefficient on state1 (Alabama)?

We excluded Georgia because not doing so for at least one categorical dummy variable leads to perfect multicollinearity. In other words, if we included all the dummy variables then the sum of all dummy variables for each observation. In another sense, the Georgia variable becomes the basis by which all other dummy variables are measured.

i. Are there any problems with multicollinearity in your model? Explain.

The VIF($\hat{\beta}_i$) for all variables in the regression is less than 5 (or less than 10) so I would say that our model is reasonably free of multicollinearity. As a rule-of-thumb, multicollinearity is not considered high when VIF($\hat{\beta}_i$) is less than 5 (or less than 10, depending on the particular thumb).

j. Compare the final model in part g to the model in L1, part g in terms of how much they explain the variance in employment growth. Explain.

R^2 is the measure of how much variation in the dependent variable is explained by the regression model. In the L2.g model, adjusted R^2 is 0.1456. In the L1.g model, adjusted R^2 is 0.0787 which is about half of the L2.g model. The F-stat for both would lead us to reject the null hypothesis for the model at the 95% confidence level. In the L2.g model, the t-stat is low enough for popsqmi_60 and rural that we cannot reject the null hypothesis, but these variables are correlated with others in the model and so should be left in.

2 Questions

Q1 Suppose you are interested in whether there is a gender bias in setting wages.

a. You get data from the Current Population Survey. You then use STATA to estimate a regression function as follows:

$$wage_i = \beta_0 + \beta_1 Female_i + \beta_2 Nonwhite_i + \beta_3 UnionMember_i + \beta_4 Education_i + \beta_5 Experience_i + u_i$$

Some of the Stata output is as follows:

Source	SS	df	MS	Number of obs	1289
				F(5, 1283)	122.61
Model	25967.3	5	5193.46	Prob > F	0
Residual	54342.5	1283	42.3558	R-squared	
				Adj R-squared	
Total	80309.8	1288	62.3523	Root MSE	6.5081

Wage	Coef.	SE	t	P>t
Female	-3.0749	0.36462		
Nw	-1.5653	0.50919		
Un	1.09598	0.50608		
education	1.3703	0.0659		
experience	0.16661	0.01605		
_cons	-7.1833	1.01579		

What is the coefficient on female? What is the interpretation of this coefficient? Calculate the t-statistic and test whether it is statistically significant at the 5% level. Based on the regression, do you think that women earn less than men?

The coefficient for female is -3.0749. Holding all other variables in the model constant, being female reduces one's wages by -3.0749 dollars.[1] The t-statistic is calculated as $|\frac{\hat{\beta}_1}{SE}| = |\frac{-3.0749}{0.36462}| = |-8.43| > 1.96$ so this variable is statistically significant at the 5% level. Yes, I would say based on this regression, women earn less than men.

b. What are the R2 and Adjusted R2 of this regression model?

R^2 equals 0.3233 and $\bar{R^2}$ equals 0.3207.
$R^2 \equiv 1 - \frac{SS_{residuals}}{SS_{total}} = 1 - \frac{54342.5}{80309.8} = 0.3233$
$\bar{R^2} = 1 - \frac{SS_{residuals}}{SS_{total}} * \frac{df_t}{df_e} = 1 - \frac{54342.5}{80309.8} * \frac{1288}{1283} = 0.3207$

c. As an alternative to the regression in part a, you collect data about gender and salaries from people who stop by a table at the local mall. You then use a "difference in means" test to see if the average salary for women is less than the average salary for men. You find that there is a statistical difference in the means with women's average salary statistically lower than men's.

I see.

[1] I am assuming the unit here is dollars.

d. Even though both approaches give you the same answer, explain which method is a better way to test if women earn less than men.

Using the Current Population Survey is a much better method. Sampling from a population is, at best, a means of estimating population parameters. In this case, the sample size may be insufficient and the sample may in some way be self-selecting but more importantly a table at a local mall almost certainly does not adequately represent the population.

e. What if you also want to know whether women and men get the same additional wages for each additional year of school? To test this, you generate a new variable (female*education) which interacts the female dummy variable and education. The results of the regression are below:

```
      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  6,  1282) =  103.19
       Model |  26154.5202     6  4359.0867            Prob > F      =  0.0000
    Residual |  54155.3045  1282  42.2428273           R-squared     =  0.3257
-------------+------------------------------           Adj R-squared =  0.3225
       Total |  80309.8247  1288  62.3523484           Root MSE      =  6.4994

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .5205381   1.746147     0.30   0.766    -2.905081    3.946157
          nw |  -1.597437   .5087366    -3.14   0.002    -2.595484   -.5993888
          un |   1.171601   .5066774     2.31   0.021     .1775931    2.165609
   education |   1.492911   .0878829    16.99   0.000     1.320501    1.665321
  experience |   .1643591   .0160616    10.23   0.000     .1328491    .1958691
  women_educ |  -.2733575     .12984    -2.11   0.035    -.5280798   -.0186352
       _cons |  -8.773084   1.264614    -6.94   0.000    -11.25402   -6.292143
```

What is the coefficient on women_educ? Interpret the meaning of this coefficient in this regression model.

The coefficient on women_educ is -0.2734 and it is statistically significant at the 5% confidence level. All other variables held constant, women receive -0.2734 fewer dollars per unit of education. [2]

Q2 Use the results from L1 to do the following. Follow the directions carefully in terms of what to calculate. Even if other parts of L1 include the answer, show how these elements are calculated, assuming you only had certain results. Please show the formula you use and the steps you take to get to the final answer. You can always use the output to see if you did it right!

a. Use the results from L1 parts j and l to calculate the coefficient of percoll90 in the following regression model:

$$empgrowth\_9006_i = \beta_0 + \beta_1 percoll90_i + \beta_2 perse90_i + u_i$$

I used the following formula to estimate β_1 : $\hat{\beta}_1 = \frac{COV(empgrowth\_9006, residual)}{VAR(residual)} = \frac{.559228}{5.5682^2} = 0.01804$, where VAR(residual) $= RootMSE^2$

[2] I would have thought doing this would lead to a dependency issue if women_educ is calculated directly from two other independent variables, but this does not appear to be an issue.

b. Calculate the t-statistic to test whether the coefficient on percoll90 is different than zero. Note: In this part it is OK to use the standard error calculated by the regression output rather than having to calculate it.

The t-statistic for $\hat{\beta}_1$ is 3.0157 which is greater than 1.96, making this statistically significant at the 95% confidence interval and we can reject H_0.

I used the following formula to estimate the t-stat value: $t - stat = \frac{\hat{\beta}_1}{SE} = \frac{0.01804}{.0059809} = 3.0157$.