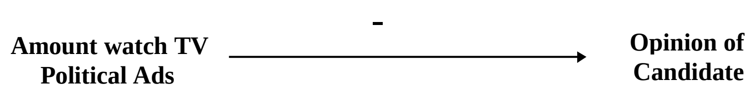
Soc 63993, Homework #5 Answer Key: Model Mis-Specification/Equality Constraints/Group Comparisons

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised February 15, 2015

1. Model mis-specification. A campaign manager has found that the amount of time spent watching TV political ads is negatively correlated with favorable opinion of her candidate. Two models have been proposed to explain this relationship:

(i)

(ii)



A. Suppose that model (i) is correct. What harm will result from estimating model (ii) and relying on the results? If appropriate, discuss such things as biased coefficients, inflated standard errors, and misguided policy decisions (particularly with regards to the use of TV advertising). Similarly, discuss the harm that will result if Model (ii) is correct and model (i) is mistakenly estimated and relied upon.

If model (ii) is estimated, the data will seem to support it even if Model (i) is correct. Model two predicts a negative effect of TV on opinion. Since the correlation between TV and opinion is negative, the bivariate regression coefficient will indeed be negative.

However, as the “true” model (i) shows, the effect of TV on opinion is actually positive. The negative correlation between TV and opinion arises from the fact that they share a common cause, Education. Better educated people are less likely to watch TV ads and more likely to like the candidate. Hence, those who watch TV ads tend to be disproportionately composed of the lesser-educated who are less likely to like the candidate. However, they would like the candidate even less if they didn’t watch the TV ads. Put another way, suppressor effects are present.

From a policy standpoint, this could lead to a grave mistake. The campaign could mistakenly conclude that it should turn away from TV advertising, when that advertising is actually helpful.

Probably less harm is done if model (ii) is correct but model (i) is estimated instead. The expected effect of Education on Opinion is zero, and the expected effect of TV on Opinion is negative. That is, adding extraneous variables to the model does not bias coefficients. Hence, when model (i) is estimated, the campaign will hopefully discover that the data don’t support it (whereas in the previous case the data did seem to support the model, even though it was wrong). However, adding extraneous variable does tend to increase standard errors and make estimates less precise. Hence, there is a greater risk that the campaign will conclude that TV does not have a significant effect (i.e. it is “neutral”) when in fact the TV ads are harmful. For that matter, because the estimates are less precise, there may even be a small chance that the estimated effect of TV winds up being positive, as Model (i) suggests.

B. Model (i) is estimated, yielding the following results. Based on this information, determine what the regression coefficient would be for model (ii). Compute the regression coefficient using both the formula for omitted variable bias and the formula for the slope coefficient in a bivariate regression.

```
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     opinion |       200          79         9.4   57.15274   99.79181
        educ |       200          14         2.7   6.597328   20.61872
          tv |       200          15         5.6  -.8872261   34.85936

. corr
(obs=200)

             |  opinion     educ       tv
-------------+---------------------------
     opinion |   1.0000
        educ |   0.3500   1.0000
          tv |  -0.2200  -0.9000   1.0000


. corr, cov
(obs=200)

             |  opinion     educ       tv
-------------+---------------------------
     opinion |    88.36
        educ |    8.883     7.29
          tv | -11.5808  -13.608    31.36


. reg  opinion educ tv, beta

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  2,   197) =   20.17
       Model |  2989.21851     2  1494.60926           Prob > F      =  0.0000
    Residual |   14594.421   197  74.0833553           R-squared     =  0.1700
-------------+------------------------------           Adj R-squared =  0.1616
       Total |  17583.6395   199  88.3599975           Root MSE      =  8.6072

------------------------------------------------------------------------------
     opinion |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |   2.785185   .5184337     5.37   0.000                       .8
          tv |   .8392856   .2499591     3.36   0.001                       .5
       _cons |   27.41812   10.77459     2.54   0.012                        .
------------------------------------------------------------------------------
```

We are asked to compute the coefficients for the incorrectly-specified bivariate regression. I’ll do this for both TV and Education as the IVs.

Opinion regressed on TV only

$$b = \frac{s_{TV,Opinion}}{s_{TV}^2} = \frac{-11.581}{31.360} = -.369$$

$$b_{TV}^* = b_{TV} + b_{Educ}\frac{Cov(TV, EDUC)}{V(TV)}$$

$$= .839286 + 2.785185\frac{-13.608}{31.360} = -.369$$

(Bonus) Opinion regressed on education only

$$b = \frac{s_{Educ,Opinion}}{s_{Educ}^2} = \frac{8.883}{7.29} = 1.219$$

$$b_{Educ}^* = b_{Educ} + b_{TV}\frac{Cov(TV, EDUC)}{V(Educ)}$$

$$= 2.785185 + .839286\frac{-13.608}{7.290} = 1.219$$

To confirm – note that we are given the means, correlations and standard deviations, so we can use the `corr2data` command to create a pseudo-replication of the data.

```
. matrix input means = (79\14\15)
. matrix input sds = (9.4\2.7\5.6)
. matrix input corr = (1,.35,-.22\.35,1,-.90\-.22,-.90,1)
. corr2data opinion educ tv, n(200) means(means) sds(sds) corr(corr)
. reg  opinion educ tv, beta

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  2,   197) =   20.17
       Model |  2989.2186     2  1494.6093           Prob > F      =  0.0000
    Residual |  14594.4214   197  74.0833575           R-squared     =  0.1700
-------------+------------------------------           Adj R-squared =  0.1616
       Total |    17583.64   199  88.3600001           Root MSE      =  8.6072

------------------------------------------------------------------------------
     opinion |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |   2.785185   .5184337     5.37   0.000                       .8
          tv |   .8392857   .2499591     3.36   0.001                       .5
       _cons |   27.41812   10.77459     2.54   0.012                        .
------------------------------------------------------------------------------

. reg opinion tv, beta

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =   10.07
       Model |  851.048099     1  851.048099           Prob > F      =  0.0017
    Residual |  16732.5919   198    84.50804           R-squared     =  0.0484
-------------+------------------------------           Adj R-squared =  0.0436
       Total |    17583.64   199  88.3600001           Root MSE      =  9.1928

------------------------------------------------------------------------------
     opinion |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          tv |  -.3692857   .1163682    -3.17   0.002                     -.22
       _cons |   84.53929   1.862631    45.39   0.000                        .
------------------------------------------------------------------------------
```

```
. reg  opinion educ, beta

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =   27.64
       Model |  2153.99576     1  2153.99576           Prob > F      =  0.0000
    Residual |  15429.6443   198  77.9274963           R-squared     =  0.1225
-------------+------------------------------           Adj R-squared =  0.1181
       Total |    17583.64   199  88.3600001           Root MSE      =  8.8277

------------------------------------------------------------------------------
     opinion |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |   1.218518   .2317688     5.26   0.000                      .35
       _cons |   61.94074   3.304259    18.75   0.000                        .
------------------------------------------------------------------------------
```

C. Based on these results, which model do you think is most plausible? Why?

Model (i) gets a clear edge. All coefficients are in the predicted direction, and all effects are statistically significant.

D. The campaign manager is concerned by the large correlation between educ and tv. Suppose the manager decided to "solve" the problem of multicollinearity by excluding education from the model. What would be the consequence of that decision? Do you think this would be a good idea in this case?

It would be a terrible mistake if you decided to "solve" the problem of multicollinearity by excluding education from the model. As noted above, this serious mis-specification would lead to very erroneous conclusions concerning TV ads. Further, even with this high correlation, effects are statistically significant. Stick with Model (i).

Incidentally, keep in mind that omitted variable bias can cause the magnitude of the coefficients for the remaining variables to be inflated either upwards or downwards. In this case, omitting education would cause the effect of TV to go down so much that the estimated effect actually switches from being positive to negative. This is because there are suppressor effects present in this example: TV and Education both positively affect opinion, but they are negatively correlated with each other.

2. Equality constraints. From the course web page, download gender.dta. This is yet another modified version of our income/education/job experience example. The sample now consists of 225 men and 275 women. Regress income on education and job experience. Test the following hypotheses:

H_0: $\beta_{Educ} = \beta_{Jobexp}$
H_A: $\beta_{Educ} \neq \beta_{Jobexp}$

Perform a Wald test, an incremental F test, and a likelihood ratio chi-square test. The results should all be identical or nearly identical.

(i) Wald test:

```
. use https://www3.nd.edu/~rwilliam/xsoc63993/statafiles/gender.dta, clear
. * Unconstrained model
. reg  income educ jobexp

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  239.86
       Model |  22352.7545     2  11176.3773           Prob > F      =  0.0000
    Residual |  23157.8824   497  46.5953368           R-squared     =  0.4912
-------------+------------------------------           Adj R-squared =  0.4891
       Total |  45510.6369   499  91.2036811           Root MSE      =  6.8261

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.309229   .0838474    15.61   0.000      1.14449    1.473968
      jobexp |   .8533107   .0670888    12.72   0.000     .7214982    .9851233
       _cons |  -1.076636   1.205717    -0.89   0.372    -3.445568    1.292295
------------------------------------------------------------------------------


. test educ = jobexp

 ( 1)  educ - jobexp = 0

       F(  1,   497) =   15.63
            Prob > F =    0.0001
```

To confirm that Stata got it right:

```
. vce

             |     educ   jobexp    _cons
-------------+---------------------------
        educ |   .00703
      jobexp | -.000883  .004501
       _cons | -.065025 -.049566  1.45375
```

$$F_{1,N-K-1} = \left(\frac{(b_{Educ} - b_{Job\exp})}{\sqrt{s^2{}_{b_{Educ}} + s^2{}_{b_{Job\exp}} - 2s_{b_{Educ},b_{Job\exp}}}} \right)^2 = \left(\frac{(1.309229 - .8533107)}{\sqrt{.00703 + .004501 - 2 * -.000883}} \right)^2$$

$$= \left(\frac{.4559183}{.115312619} \right)^2 = 3.95375897^2 = 15.63$$

(ii) Incremental F test:

```
. * Unconstrained model
. reg  income educ jobexp

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  239.86
       Model |  22352.7545     2  11176.3773           Prob > F      =  0.0000
    Residual |  23157.8824   497  46.5953368           R-squared     =  0.4912
-------------+------------------------------           Adj R-squared =  0.4891
       Total |  45510.6369   499  91.2036811           Root MSE      =  6.8261

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.309229   .0838474    15.61   0.000      1.14449    1.473968
      jobexp |   .8533107   .0670888    12.72   0.000     .7214982    .9851233
       _cons |  -1.076636   1.205717    -0.89   0.372    -3.445568    1.292295
------------------------------------------------------------------------------

. est store unconstrained
. * Constrained model
. gen jobed = educ + jobexp
. reg  income jobed

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  450.84
       Model |    21624.34     1    21624.34           Prob > F      =  0.0000
    Residual |  23886.2969   498  47.9644516           R-squared     =  0.4751
-------------+------------------------------           Adj R-squared =  0.4741
       Total |  45510.6369   499  91.2036811           Root MSE      =  6.9256

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       jobed |   1.037904   .0488816    21.23   0.000     .9418644    1.133944
       _cons |  -.5465906   1.215718    -0.45   0.653    -2.935159    1.841978
------------------------------------------------------------------------------

. est store constrained
. * Use Buis's -ftest- command
. ftest constrained unconstrained
Assumption: constrained nested in unconstrained

F(  1,     497) =     15.63
       prob > F =    0.0001
```

If you prefer to do things the hard way – From the unconstrained model, we get

$$SSE_u = 23157.8824,\ R_u^2 = .4912,\ N = 500,\ K = 2.$$

From the constrained model, we get

$$SSE_c = 23886.2969,\ R^2_c = .4751,\ J = 1.$$

Using the incremental F test, we get

$$F_{1,N-K-1} = \frac{(SSE_c - SSE_u)*(N-K-1)}{SSE_u*1} = \frac{(R_u^2 - R_c^2)*(N-K-1)}{(1-R_u^2)*1}$$

$$= \frac{(23886.30 - 23157.88)*497}{23157.88} = \frac{(.4912 - .4751)*497}{1-.49115} = 15.63$$

(iii) Likelihood ratio chi square test:

```
. lrtest constrained unconstrained, stats

Likelihood-ratio test                                 LR chi2(1)  =     15.48
(Assumption: constrained nested in unconstrained)     Prob > chi2 =    0.0001

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
 constrained |    500   -1837.243   -1676.082      2     3356.165    3364.594
unconstrai~d |    500   -1837.243    -1668.34      3      3342.68    3355.324
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

The test statistics are all highly significant. It is very unlikely that the effects of education and job experience are equal.

3. Group comparisons. Using the same data as in problem 2, do the following:

(a) Do T-tests of whether the means of men and women significantly differ on education, job experience, and income. If using Stata, use commands such as

```
. ttest  educ, by(female)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    male |     225    11.22222     .298438     4.47657    10.63412    11.81033
  female |     275    10.63636    .1733252    2.874273    10.29515    10.97758
---------+--------------------------------------------------------------------
combined |     500        10.9    .1650287    3.690154    10.57576    11.22424
---------+--------------------------------------------------------------------
    diff |            .5858586    .3310136               -.0644967    1.236214
------------------------------------------------------------------------------
Degrees of freedom: 498

                  Ho: mean(male) - mean(female) = diff = 0

     Ha: diff < 0             Ha: diff != 0              Ha: diff > 0
       t =   1.7699             t =   1.7699              t =   1.7699
   P < t =   0.9613         P > |t| =   0.0774        P > t =   0.0387
```

Men have slightly more education than women do. The difference is significant if you use a 1-tailed test.

```
. ttest  jobexp, by(female)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    male |     225    14.11111    .3569664    5.354497    13.40767    14.81455
  female |     275    12.36364    .2249718    3.730735    11.92074    12.80653
---------+--------------------------------------------------------------------
combined |     500       13.15    .2062525    4.611945    12.74477    13.55523
---------+--------------------------------------------------------------------
    diff |            1.747475    .4075443                .9467565    2.548193
------------------------------------------------------------------------------
Degrees of freedom: 498

                  Ho: mean(male) - mean(female) = diff = 0

     Ha: diff < 0               Ha: diff != 0              Ha: diff > 0
       t =   4.2878                t =   4.2878              t =   4.2878
   P < t =   1.0000          P > |t| =   0.0000          P > t =   0.0000
```

On average, men have almost 2 more years of job experience than do women. The difference is highly significant.

```
. ttest  income, by(female)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    male |     225    27.81111      .76553    11.48295    26.30255    29.31967
  female |     275    21.63636    .3865032     6.40943    20.87547    22.39726
---------+--------------------------------------------------------------------
combined |     500      24.415    .4270917    9.550062    23.57588    25.25412
---------+--------------------------------------------------------------------
    diff |            6.174747    .8135835                4.576268    7.773227
------------------------------------------------------------------------------
Degrees of freedom: 498

                  Ho: mean(male) - mean(female) = diff = 0

     Ha: diff < 0               Ha: diff != 0              Ha: diff > 0
       t =   7.5896                t =   7.5896              t =   7.5896
   P < t =   1.0000          P > |t| =   0.0000          P > t =   0.0000
```

Men make more than $6,000 a year more than women, and the difference is highly significant.

(b) Test the following. Use a likelihood ratio chi square test. Performing an incremental F test and/or a Wald test using `suest` is optional.

H_0: Model parameters are the same for both men and women
H_A: Model parameters are not the same for both men and women.

```
. * Constrained model: No Gender differences
. reg  income educ jobexp

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  239.86
       Model |  22352.7545     2  11176.3773           Prob > F      =  0.0000
    Residual |  23157.8824   497  46.5953368           R-squared     =  0.4912
-------------+------------------------------           Adj R-squared =  0.4891
       Total |  45510.6369   499  91.2036811           Root MSE      =  6.8261

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.309229   .0838474    15.61   0.000      1.14449    1.473968
      jobexp |   .8533107   .0670888    12.72   0.000     .7214982    .9851233
       _cons |  -1.076636   1.205717    -0.89   0.372    -3.445568    1.292295
------------------------------------------------------------------------------
```

Note that the constrained model was the unconstrained model in problem 2. In problem 2, we viewed it as unconstrained because the effects of education and job experience were free to differ. In this problem, we view it as constrained because the coefficients are constrained to be the same for both men and women.

```
. * Unconstrained - Effects differ by gender
. reg  income educ jobexp if female == 0

      Source |       SS       df       MS              Number of obs =     225
-------------+------------------------------           F(  2,   222) =  210.87
       Model |  19350.4582     2  9675.22912           Prob > F      =  0.0000
    Residual |  10185.7638   222  45.8818188           R-squared     =  0.6551
-------------+------------------------------           Adj R-squared =  0.6520
       Total |   29536.222   224  131.858134           Root MSE      =  6.7736

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .8195378   .1070818     7.65   0.000     .6085108    1.030565
      jobexp |   1.384972   .0895246    15.47   0.000     1.208545    1.561398
       _cons |  -.9294128    1.49777    -0.62   0.536     -3.88108    2.022254
------------------------------------------------------------------------------
. est store male

. reg  income educ jobexp if female == 1

      Source |       SS       df       MS              Number of obs =     275
-------------+------------------------------           F(  2,   272) =  120.03
       Model |  5276.94296     2  2638.47148           Prob > F      =  0.0000
    Residual |  5979.19312   272  21.9823276           R-squared     =  0.4688
-------------+------------------------------           Adj R-squared =  0.4649
       Total |  11256.1361   274  41.0807886           Root MSE      =  4.6885

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.525582   .1004096    15.19   0.000     1.327903    1.723261
      jobexp |  -.0049199   .0773587    -0.06   0.949    -.1572178    .1473779
       _cons |   5.470545   1.589722     3.44   0.001     2.340821    8.600269
------------------------------------------------------------------------------
. est store female
```

Likelihood Ratio Test. Doing a Likelihood Ratio Chi Square Test,

```
. lrtest (male female) both

Likelihood-ratio test                                 LR chi2(3)  =    213.10
                                                      Prob > chi2 =    0.0000

Assumption: (both) nested in (male, female)
```

[Note that the LR Chi Square / Degrees of Fredom = 213.10/3 = 71.03; compare with the incremental F value calculated below]

This is highly significant, ergo we reject the null and conclude that the coefficients for men and women likely are different. This conclusion is not surprising, since, by looking at the coefficients in the separate male and female models, you can see that the effects appear to be very different. Nonetheless, keep in mind that all we know for sure is that at least one parameter (including possibly the intercept) differs between men and women.

Incremental F Test. If we want to be masochistic and do an incremental F test, from the constrained model we get

$$SSE_c = 23158, N = 500, K = 2.$$

From the regressions for males only and females only we get

$$SSE_{Males} = 10186, N_{Males} = 225$$

$$SSE_{Females} = 5979, N_{Females} = 275.$$

Hence, by adding up the figures for men and women, for the unconstrained model we get

$$SSE_u = 16165, N_u = 500.$$

Also, note that J = K + 1 = 3, i.e. the constrained model estimates 2 betas and 1 intercept, while the unconstrained model estimates 4 betas and 2 intercepts.

Hence, for the incremental F, we get

$$F_{K+1, N_1+N_2-2K-2} = \frac{(SSE_c - SSE_u) * (N_1 + N_2 - 2K - 2)}{SSE_u * (K+1)} = \frac{(23158 - 16165) * 494}{16165 * 3} = 71.24$$

[Note too that this is almost identical to LR chi square/3 shown above]

Suest. If we wanted to do this with a Wald chi-square test and the `suest` command,

```
. quietly reg income educ jobexp if female == 0
. est store male
. quietly reg income educ jobexp if female == 1
. est store female
. suest male female

Simultaneous results for male, female

                                                  Number of obs   =        500

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
male_mean    |
        educ |   .8195378   .1000803     8.19   0.000      .623384    1.015692
      jobexp |   1.384972   .1096212    12.63   0.000     1.170118    1.599825
       _cons |  -.9294128    .494266    -1.88   0.060    -1.898156    .0393307
-------------+----------------------------------------------------------------
male_lnvar   |
       _cons |   3.826069   .0705412    54.24   0.000     3.687811    3.964327
-------------+----------------------------------------------------------------
female_mean  |
        educ |   1.525582   .0930839    16.39   0.000     1.343141    1.708023
      jobexp |  -.0049199   .0400294    -0.12   0.902    -.0833761    .0735362
       _cons |   5.470545   1.626955     3.36   0.001     2.281772    8.659318
-------------+----------------------------------------------------------------
female_lnvar |
       _cons |   3.090239   .1010872    30.57   0.000     2.892112    3.288366
------------------------------------------------------------------------------

. test [male_mean = female_mean], constant coef

 ( 1)  [male_mean]educ - [female_mean]educ = 0
 ( 2)  [male_mean]jobexp - [female_mean]jobexp = 0
 ( 3)  [male_mean]_cons - [female_mean]_cons = 0

           chi2(  3) =  180.32
         Prob > chi2 =    0.0000


Constrained coefficients

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
male_mean    |
        educ |   1.579664   .0309782    50.99   0.000     1.518948     1.64038
      jobexp |   .0646106   .0225819     2.86   0.004     .0203509    .1088703
       _cons |   3.960761   .3267024    12.12   0.000     3.320436    4.601085
-------------+----------------------------------------------------------------
male_lnvar   |
       _cons |   3.813277   .0572104    66.65   0.000     3.701146    3.925407
-------------+----------------------------------------------------------------
female_mean  |
        educ |   1.579664   .0309782    50.99   0.000     1.518948     1.64038
      jobexp |   .0646106   .0225819     2.86   0.004     .0203509    .1088703
       _cons |   3.960761   .3267024    12.12   0.000     3.320436    4.601085
-------------+----------------------------------------------------------------
female_lnvar |
       _cons |   2.949792    .092498    31.89   0.000     2.768499    3.131085
------------------------------------------------------------------------------
```

(c) Based on your results, explain whether men make more than women and if so why. [Note: these are hypothetical data, and the results are a little peculiar in some respects!]

We know from the T-Tests that, on average, men make significantly more money than do women. We also know from the T-Tests that men benefit from having higher levels of education and job experience than do women. The regressions add additional insights as to why differences exist. For men, both education and job experience have significant effects, with job experience actually having a larger effect than education does. For women, on the other hand, job experience has virtually no effect whatsoever; only education is important. Education actually appears to have a larger effect on women than it does men! But this is more than offset by the advantages men have from higher levels of job experience and education and the much greater effect job experience has on men than women. Perhaps women are more likely to be in dead-end jobs where additional experience does not help you to get promoted into higher paying positions.

It is true that, under certain conditions, a woman would be expected to make more than a comparable man, e.g. when jobexp = 0. However, no such person exists in the sample (the lowest value of jobexp is 3), and overall, men have the advantage.

These would be <u>extremely</u> interesting and important findings, if it weren't for the fact that I made these data up.

(d) Suppose there were no gender-related compositional differences, i.e. women had the same levels of education and job experience as men did. If education and job experience continued to have the same effects on women that they do now, how much would the gap in income between men and women be affected?

We are asking a "what if" question. The following analysis addresses this.

```
. tabstat income educ jobexp, by(female) columns(variables)

Summary statistics: mean
  by categories of: female

female |    income      educ    jobexp
-------+------------------------------
  male |  27.81111  11.22222  14.11111
female |  21.63636  10.63636  12.36364
-------+------------------------------
 Total |    24.415      10.9     13.15
--------------------------------------
```

As we saw before, men make $6174.75 more than women on average.

```
. reg income educ jobexp if female == 1

      Source |       SS       df       MS              Number of obs =     275
-------------+------------------------------           F(  2,   272) =  120.03
       Model |  5276.94296     2  2638.47148           Prob > F      =  0.0000
    Residual |  5979.19312   272  21.9823276           R-squared     =  0.4688
-------------+------------------------------           Adj R-squared =  0.4649
       Total |  11256.1361   274  41.0807886           Root MSE      =  4.6885

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.525582   .1004096    15.19   0.000     1.327903    1.723261
      jobexp |  -.0049199   .0773587    -0.06   0.949    -.1572178    .1473779
       _cons |   5.470545   1.589722     3.44   0.001     2.340821    8.600269
------------------------------------------------------------------------------
```

```
. margins, at (educ = 11.2222 jobexp = 14.11111)

Adjusted predictions                              Number of obs   =        275
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : educ            =     11.2222
               jobexp          =    14.11111

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   22.52151   .3236064    69.60   0.000     21.88442     23.1586
------------------------------------------------------------------------------
```

As the final numbers show, if women had the same levels of education and job experience as men, while the effects of education and job experience on women stayed the same, women would make $22,521.54 on average, an increase of $22,521.54 - $21,636.36 = $885.18. Hence, of the original difference of $6174.75, $885.18/$6174.75 = 14.34% is due to compositional factors. The rest is due to differences in effects, in particular the fact that women get virtually no benefit from their years of job experience.

Some alternative approaches that will also work:

Using the `predict` command,

```
. reg income educ jobexp if female == 1
. predict mcompfcoef if !female
(option xb assumed; fitted values)
(275 missing values generated)
. sum  mcompfcoef

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  mcompfcoef |       225    22.52154     6.82074   8.506949   31.35624
```

Using the `adjust` command,

```
. quietly reg income educ jobexp if female == 1
. adjust educ jobexp if female == 0

--------------------------------------------------------------------------------------
     Dependent variable: income     Command: regress
 Covariates set to mean: educ = 11.222222, jobexp = 14.111111
--------------------------------------------------------------------------------------

----------------------
      All |         xb
----------+-----------
          |    22.5215
----------------------
     Key:  xb  =  Linear Prediction
```

Using `margins` with `atmeans`,

```
. quietly reg income educ jobexp if female == 1
. margins if female == 0, atmeans noesample

Adjusted predictions                              Number of obs   =        225
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : educ            =    11.22222 (mean)
               jobexp          =    14.11111 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   22.52154    .323607    69.60   0.000     21.88445    23.15863
------------------------------------------------------------------------------
```

Using `margins` with precise values,

```
. sum educ if female == 0, meanonly
. scalar malemeaneduc = r(mean)
. sum jobexp if female == 0, meanonly
. scalar malemeanjobexp = r(mean)
. scalar list
malemeanjobexp =  14.111111
malemeaneduc =  11.222222

. quietly reg income educ jobexp if female == 1
. margins, at (educ = `=malemeaneduc' jobexp = `=malemeanjobexp')

Adjusted predictions                              Number of obs   =        275
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : educ            =    11.22222
               jobexp          =    14.11111

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   22.52154    .323607    69.60   0.000     21.88445    23.15863
------------------------------------------------------------------------------
```