Sociology 593
Exam 1 Answer Key
February 13, 2004

I. *True-False.* (20 points) These questions all pertain to the following analysis: A researcher is interested in the relationship between race (white, black, and other; dummy variables have been computed for race) and an attitudinal scale she has constructed (psyscale). Her results are as follows:

```
. reg  psyscale white black

      Source |       SS       df       MS              Number of obs =     438
-------------+------------------------------           F(  2,   435) =  155.90
       Model |  334.812471     2  167.406236           Prob > F      =  0.0000
    Residual |  467.105132   435   1.0738049           R-squared     =  0.4175
-------------+------------------------------           Adj R-squared =  0.4148
       Total |  801.917603   437  1.83505172           Root MSE      =  1.0362

------------------------------------------------------------------------------
    psyscale |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   1.897789   .1212834    15.65   0.000     1.659415    2.136163
       black |   1.808343   .1212834    14.91   0.000     1.569969    2.046718
       _cons |   7.081524   .0857603    82.57   0.000     6.912968     7.25008
------------------------------------------------------------------------------

. test white = black

 ( 1)  white - black = 0

       F(  1,   435) =    0.54
            Prob > F =    0.4612
```

Indicate whether the following statements are true or false. If false, briefly explain why.

1. Based on these results, she should conclude that race does not significantly affect attitudes.

False. As the F and T values show, race has highly significant effects. However, while whites and blacks significantly differ from others, they do not significantly differ from each other.

2. The researcher is confident that race is well measured, but she also believes that psyscale suffers from random measurement error. Increasing the sample size would help to make this less problematic.

True. Random measurement error in the DV increases standard errors. A larger sample size will help to reduce the standard errors.

3. The researcher has conducted a GQ test and the test statistic is not significant. This shows that heteroscedasticity is not a problem in her data.

False. GQ only tests for a specific type of heteroscedasticity; other types could be present.

4. The researcher has decided to re-estimate her regression model, this time using robust standard errors. This will probably cause her coefficient estimates, standard errors, and t values to change.

False. The use of robust standard errors will not change the coefficient estimates. It will probably change the standard errors and t values.

5. The researcher has again decided to re-estimate her regression model, this time using backwards stepwise selection. Hence, in the next step, black will be dropped from her model.

False. All variables now in the model are highly significant, so none will be dropped.

II. Short answer. Discuss <u>three</u> of the following five problems. (15 points each, 45 points total, up to 5 points extra credit for each additional problem.) In each case, the researcher has used SPSS or Stata to test for a possible problem, concluded that there is a problem, and then adopted a strategy to address that problem. Explain (a) what problem the researcher was testing for, why the test or tests used were appropriate, and why she concluded that there was a problem, (b) the rationale behind the solution she chose, i.e. how does it try to address the problem, and (c) at least one or two alternative solutions she could have tried, and why.

II-1.

```
. reg  y x1 x2 x3

      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  3,   396) =   70.33
       Model |  126.086202     3  42.0287339           Prob > F      =  0.0000
    Residual |  236.656252   396  .597616798           R-squared     =  0.3476
-------------+------------------------------           Adj R-squared =  0.3426
       Total |  362.742454   399  .909128957           Root MSE      =  .77306

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.0749123   .7937044    -0.09   0.925    -1.635313    1.485489
          x2 |  -.0834071   .7894182    -0.11   0.916    -1.635382    1.468567
          x3 |    .390442    .790224     0.49   0.622    -1.163117    1.944001
       _cons |  -.0138723   .0387045    -0.36   0.720    -.0899643    .0622196
------------------------------------------------------------------------------

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
          x3 |   1358.71    0.000736
          x1 |    454.95    0.002198
          x2 |    445.10    0.002247
-------------+----------------------
    Mean VIF |    752.92


. reg y x1 x2

      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  2,   397) =  105.57
       Model |  125.940308     2  62.9701542           Prob > F      =  0.0000
    Residual |  236.802145   397  .596478956           R-squared     =  0.3472
-------------+------------------------------           Adj R-squared =  0.3439
       Total |  362.742454   399  .909128957           Root MSE      =  .77232

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .3166639   .0433117     7.31   0.000     .2315151    .4018128
          x2 |   .3060416   .0435516     7.03   0.000     .2204209    .3916622
       _cons |  -.0130207   .0386293    -0.34   0.736    -.0889642    .0629228
------------------------------------------------------------------------------
```

There are several indications that multicollinearity is a concern. The overall F value is significant but the individual t values are not. The variance inflation factors are huge.

To solve the problem, the researcher decided to drop one variable, and when she did this, the multicollinearity problem seemed to go away. This may be a good idea if x3 was not that substantively important to begin with or if it was somehow computed from x1 and x2 (which, incidentally, it was) or if it was simply a different way of measuring the same concept. If, however, we felt x3 was an important variable to have in the model, dropping it could lead to specification error and omitted variable bias. Some other solutions we might consider, then, are joint hypothesis tests involving two or more variables; or somehow creating a scale from the variables (it looks, for example, like x1 and x2 could probably be added together). We would need to know more about the data and the problem before deciding what strategy was best.

II-2.

```
. reg inc  educ jobexp

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =   77.17
       Model |  32482.8173     2  16241.4087           Prob > F      =  0.0000
    Residual |  104599.191   497  210.461149           R-squared     =  0.2370
-------------+------------------------------           Adj R-squared =  0.2339
       Total |  137082.008   499  274.713444           Root MSE      =  14.507

------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    1.99003    .164487    12.10   0.000     1.666854    2.313206
      jobexp |   .5617026   .1293588     4.34   0.000      .307545    .8158602
       _cons |  -5.911479   3.021675    -1.96   0.051    -11.84831    .0253521
------------------------------------------------------------------------------

. hettest educ

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: educ

         chi2(1)      =    78.73
         Prob > chi2  =   0.0000

. reg inc  educ jobexp [aw = 1/educ^2]
(sum of wgt is   6.3404e+00)

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  338.25
       Model |  59996.7523     2  29998.3762           Prob > F      =  0.0000
    Residual |  44077.9356   497  88.6879992           R-squared     =  0.5765
-------------+------------------------------           Adj R-squared =  0.5748
       Total |  104074.688   499  208.566509           Root MSE      =  9.4174

------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.884388   .0830109    22.70   0.000     1.721293    2.047484
      jobexp |   .5667619   .0905118     6.26   0.000      .388929    .7445948
       _cons |  -4.628755   1.219253    -3.80   0.000     -7.02428   -2.233229
------------------------------------------------------------------------------
```

The researcher is testing for heteroscedasticity. Specifically, with the Breusch-Pagan test, she is testing to see whether there is a linear relationship between the error variances and education, e.g. do the error variances go up as education goes up? The highly significant chi-square value indicates that heteroscedasticity is a problem. This will cause her coefficient estimates to be inefficient and the estimated standard errors to be biased.

To solve the problem, she uses weighted least squares. This causes the cases with the largest error variances (i.e. the cases with larger education values) to be weighted less heavily. Assuming she has done the weighting right, her parameter estimates will now be efficient and the estimated standard errors will be unbiased. Note that, in this particular case, when she does this, the coefficient estimates change little, but the t values, standard errors and confidence intervals change quite a bit. Particularly in a borderline situation, these changes might make the difference between accepting and rejecting various null hypotheses.

She might have also considered using SPSS's WLS routine, which would allow her to determine what the optimal weighting values were. She could have also taken the more simple route of simply using robust standard errors. This would not give her the most efficient parameter estimates, but it would give her unbiased standard errors.

Before doing any of this, though, she probably should have done some additional checking as to why the data seemed heteroscedastic. It may be that important variables are omitted from the model; or, it may be that the variables should be transformed in some way, e.g. use log of income instead of income. It is generally a bad idea to just leap to a solution before you understand what the cause of the problem is.

II-3.

```
. reg y x

      Source |       SS       df       MS              Number of obs =     300
-------------+------------------------------           F(  1,   298) =    7.34
       Model |  12944.8562     1  12944.8562           Prob > F      =  0.0071
    Residual |  525247.207   298  1762.57452           R-squared     =  0.0241
-------------+------------------------------           Adj R-squared =  0.0208
       Total |  538192.063   299  1799.97345           Root MSE      =  41.983

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    2.04354    .754064     2.71   0.007     .5595746    3.527505
       _cons |   -4.00445   3.877035    -1.03   0.303    -11.63429    3.625385
------------------------------------------------------------------------------

. predict rstandard, rstandard

. dfbeta
                             DFx:  DFbeta(x)
```

```
. extremes  rstandard DFx  y x

  +--------------------------------------------------------+
  | obs:   rstandard        DFx           y           x |
  |--------------------------------------------------------|
  | 131.   -.4009594   -.0302183   -4.013718   8.195821 |
  | 204.   -.3840846   -.0318037   -2.460587   8.606113 |
  |  42.   -.3784217    -.032272   -1.941961    8.74255 |
  | 134.   -.3583736   -.0484251    4.542933   11.46649 |
  | 151.   -.3270369   -.0070271   -7.064314   5.208639 |
  +--------------------------------------------------------+

  +--------------------------------------------------------+
  |  96.    .2108727   -.0019709    11.97055    3.492507 |
  | 285.    .2430508   -.0384514   -3.518938   -4.685962 |
  | 276.    .2586429   -.0271533    3.099044   -1.799608 |
  | 286.    .3552958   -.0373423     7.11564     -1.8055 |
  |   3.    17.18468    14.61801    731.2714    8.495018 |
  +--------------------------------------------------------+

. rreg y x, nolog

Robust regression estimates                            Number of obs =     299
                                                       F(  1,   297) =  194.24
                                                       Prob > F      =  0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.021697   .0733091    13.94   0.000     .8774255    1.165968
       _cons |  -2.244073   .3758284    -5.97   0.000    -2.983697   -1.504449
------------------------------------------------------------------------------
```

The researcher is testing whether there are any extreme outliers in the data. Standardized residuals greater than 3 or dfbetas greater than 1 may indicate a problem. In this case, it is clear that case 3 is an extreme outlier; its standardized residual is more than 17 and the dfbeta is almost 15. No other cases stand out as being extreme outliers.

To solve the problem, the researcher has turned to the rreg (robust regression) routine. rreg uses an iterative weighting procedure that causes outliers to have less influence on the regression estimates; in this instance it causes case 3 to be dropped altogether (notice how the N in rreg is only 299, compared to the earlier 300).

Rreg often works well as a means for dealing with outliers, but before turning to it other things should probably be tried first. First, check for coding errors. Case 3 has a y value of 731.2714 while all the other y values listed range between about -7 and 12. This suggests that the decimal place may be off by 2 spots for case 3. If you don't trust the coding of case 3 and aren't sure how to fix it, you could just drop the case yourself and rerun the OLS regression. If you believe the code 731.2714 is legitimate, you may want to examine that case further; perhaps there is some additional variable that could be added to the model that would make case 3 less of an outlier, or perhaps case 3 is not really a member of the population of interest and should be excluded. You could also consider using median regression (qreg). Median regression is less affected by outliers than OLS regression is, and sometimes the median is of greater theoretical interest than the mean anyway.

II-4.

```
SUMMARIZE
  /TABLES=y x1 x2
  /FORMAT=VALIDLIST NOCASENUM TOTAL LIMIT=100
  /TITLE='Case Summaries'
  /MISSING=VARIABLE
  /CELLS=COUNT .
```

Summarize

Case Processing Summary[a]

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Y	40	100.0%	0	.0%	40	100.0%
X1	28	70.0%	12	30.0%	40	100.0%
X2	28	70.0%	12	30.0%	40	100.0%

a. Limited to first 100 cases.

Case Summaries[a]

	Y	X1	X2
1	3.02	.57	.
2	3.28	.	1.71
3	3.34	.	1.07
4	3.38	3.05	2.32
5	3.54	.	.
6	3.67	3.24	1.93
7	4.02	1.89	2.90
8	4.23	3.70	.
9	4.29	.	1.94
10	4.31	3.06	.41
11	4.43	3.26	.51
12	4.53	3.07	1.31
13	4.53	.	1.54
14	4.55	2.93	.
15	4.57	.	1.83
16	4.60	3.12	2.49
17	4.65	4.44	3.08
18	4.73	2.83	1.05
19	4.74	4.78	1.95
20	4.80	2.35	.44
21	5.06	3.65	.
22	5.19	.	.70
23	5.28	4.09	3.17
24	5.32	.	2.30
25	5.36	.	2.31
26	5.37	2.99	2.18
27	5.52	3.96	.
28	5.66	3.90	.
29	5.71	3.83	.
30	5.74	1.40	.72
31	5.91	.	1.81
32	5.93	3.32	2.91
33	6.15	4.41	.
34	6.17	3.33	.
35	6.19	3.73	1.54
36	6.22	2.10	.
37	6.28	3.53	3.50
38	6.36	.	.
39	6.57	3.75	2.83
40	6.80	.	2.45
Total N	40	28	28

a. Limited to first 100 cases.

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING PAIRWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT y
  /METHOD=ENTER x1 x2  .
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Y	5.0000	1.00000	40
X1	3.2237	.92467	28
X2	1.8892	.87957	28

Correlations

		Y	X1	X2
Pearson Correlation	Y	1.000	.356	.294
	X1	.356	1.000	.421
	X2	.294	.421	1.000
Sig. (1-tailed)	Y	.	.031	.065
	X1	.031	.	.041
	X2	.065	.041	.
N	Y	40	28	28
	X1	28	28	18
	X2	28	18	28

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.390[a]	.152	.039	.98034

a. Predictors: (Constant), X2, X1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.584	2	1.292	1.344	.290[a]
	Residual	14.416	15	.961		
	Total	17.000	17			

a. Predictors: (Constant), X2, X1

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.639	.879		4.141	.001
	X1	.306	.283	.283	1.079	.298
	X2	.198	.298	.175	.666	.516

a. Dependent Variable: Y

The researcher is basically relying on a visual inspection of the data to see whether missing data is a problem. While there are 40 cases in the data, you can tell from the case summaries that only 18 cases have complete data on all three variables. But, the missing data appear to be randomly scattered across the X1 and X2 variables (only one case is missing data on both x1 and x2).

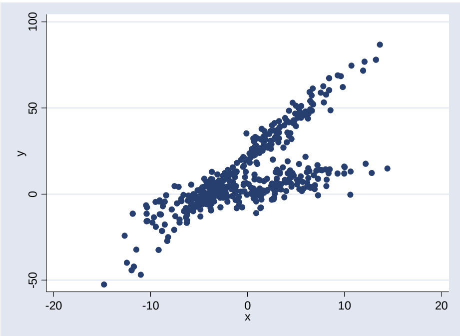
To solve the problem, the researcher relies on pairwise deletion of missing data. The most obvious indication of this is the missing pairwise option on the regression card; however the differing Ns on the descriptive statistics and correlations also indicate that pairwise deletion has been used. This lets the researcher use all available information (even though SPSS takes a very conservation approach by using N=18, i.e. the minimum number of cases that were used in computing any of the correlations, in this case the correlation between x1 and x2).

If the researcher is convinced that data are missing completely at random, this may not be a bad strategy, as it lets her use all the available information from all 40 cases. An alternative, of course, is listwise deletion, in which only the 18 cases with complete information would be used in the calculations. She could also try to impute estimated values for x1 and x2 by regressing them on each other and any other relevant variables that may be in the data. However, this strategy can be problematic in that the significance tests do not adequately take into account the fact that not all the data are "real." Before deciding on a strategy though, the researcher really needs to know more about why the data are missing. Perhaps they are missing because the questions did not apply to the respondent; or, perhaps skip patterns caused the same or similar questions to be asked at different points in the interview, and it would be possible to construct composite measures that had much less missing data. Once again, before you adopt a solution to a problem, you need to have a better idea of what is causing it in the first place.

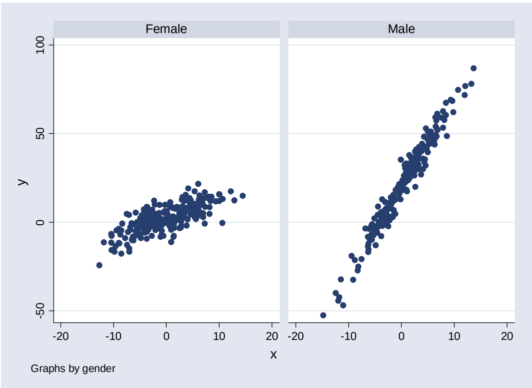
II-5.

```
. reg y x

      Source |       SS       df       MS              Number of obs =     460
-------------+------------------------------           F(  1,   458) =  487.50
       Model |  99566.5528     1  99566.5528           Prob > F      =  0.0000
    Residual |  93541.5301   458  204.239149           R-squared     =  0.5156
-------------+------------------------------           Adj R-squared =  0.5145
       Total |  193108.083   459  420.714778           Root MSE      =  14.291

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   2.847209   .1289532    22.08   0.000     2.593796    3.100622
       _cons |    10.9082   .6663333    16.37   0.000     9.598751    12.21765
------------------------------------------------------------------------------

. hettest
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of y

         chi2(1)      =   166.77
         Prob > chi2  =   0.0000

. twoway (scatter y x)
```




```
. twoway (scatter y x), by(gender )
```




```
. reg y x if gender == 1

      Source |       SS       df       MS              Number of obs =     253
-------------+------------------------------           F(  1,   251) =  292.60
       Model |  7547.24475     1  7547.24475           Prob > F      =  0.0000
    Residual |  6474.24443   251  25.7938025           R-squared     =  0.5383
-------------+------------------------------           Adj R-squared =  0.5364
       Total |  14021.4892   252  55.6408301           Root MSE      =  5.0788

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.047045   .0612109    17.11   0.000     .9264922    1.167597
       _cons |   2.252601   .3198793     7.04   0.000     1.622611     2.88259
------------------------------------------------------------------------------

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of y

         chi2(1)      =     0.58
         Prob > chi2  =   0.4480
```

```
. reg y x if gender == 2

      Source |       SS       df       MS              Number of obs =     207
-------------+------------------------------           F(  1,   205) = 4983.35
       Model |  128707.047     1  128707.047           Prob > F      =  0.0000
    Residual |     5294.62   205  25.8274147           R-squared     =  0.9605
-------------+------------------------------           Adj R-squared =  0.9603
       Total |  134001.667   206  650.493528           Root MSE      =  5.0821

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   4.906713   .0695072    70.59   0.000     4.769672    5.043754
       _cons |   20.05133   .3541185    56.62   0.000     19.35314    20.74951
------------------------------------------------------------------------------

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of y

         chi2(1)      =     0.07
         Prob > chi2  =   0.7977
```

Once again, the researcher is testing for the presence of heteroscedasticity, and the initial Breusch-Pagan test suggests that it is. But, rather than immediately adopting WLS as a solution, she does some additional checks to try to identify the problem. In the initial scatterplot, she notices that there seem to be two different clusterings of the data, with one having a much sharper slope than the other. Suspecting that gender might be a factor, and knowing that a failure to consider subpgroup differences can create the appearance of heteroscedasticity, she then does separate scatterplots by gender. She finds that indeed, the female data appear to have a much smaller slope than the male data. She then runs separate models for men and women. When she does this, the male slope (i.e. the effect of x on y) is almost 5 times as large, and the Breusch-Pagan tests are not significant for either men or women.

The researcher could, of course, have tried WLS, robust standard errors, transformed the variables in some way, or otherwise modified the model. But, the visual patterns she observed and her subsequent analyses strongly suggest that she took the correct route of examining whether there were subgroup differences and then estimating separate models accordingly. Later in the semester, we will show how to formally test whether the subgroup differences that appear to exist in the scatterplots are, indeed, statistically significant.

III. *Computation and interpretation.* (35 points total)

A research is interested in the relationship between health, socio-economic status, race and gender. She has collected data from 600 individuals on the following variables:

Variable	Description
health	Physical health, measured on a scale ranging from 0 to 1500. Higher scores indicate better health.
black	Coded 1 if the respondent is black, 0 otherwise
male	Coded 1 if the respondent is male, 0 otherwise
ses	Socio-economic status, measured on a scale that ranges from a low of 0 to a high of 200.

She obtains the following results.

```
. corr  health ses male black, means
(obs=600)

    Variable |         Mean    Std. Dev.          Min          Max
-------------+----------------------------------------------------
      health |     763.4698     166.7891     197.6368     1243.452
         ses |     107.6862     18.52201     53.78384     151.8031
        male |           .5     .5004172            0            1
       black |          .15     .3573694            0            1


             |   health      ses     male    black
-------------+------------------------------------
      health |   1.0000
         ses |   0.2901   1.0000
        male |  -0.0772   0.5102   1.0000
       black |  -0.3280  -0.2542   0.0093   1.0000


. * Model 1:

. reg  health black

      Source |       SS       df       MS              Number of obs =     600
-------------+------------------------------           F(  1,   598) =   [ 1 ]
       Model |   1792453.8     1   1792453.8           Prob > F      =  0.0000
    Residual |  14870886.7   598  24867.7035           R-squared     =  0.1076
-------------+------------------------------           Adj R-squared =  0.1061
       Total |  16663340.5   599  27818.5985           Root MSE      =  157.69

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -153.0711   18.02964    -8.49   0.000    -188.4802    -117.662
       _cons |   786.4305   6.982851   112.62   0.000     772.7166    800.1444
------------------------------------------------------------------------------
```

```
. * Model 2:

. reg  health ses male black, beta

      Source |       SS       df       MS              Number of obs =     600
-------------+------------------------------           F(  3,   596) =   50.32
       Model |  3367869.52     3  1122623.17           Prob > F      =  0.0000
    Residual |    13295471   596  22307.8372           R-squared     =  [  2 ]
-------------+------------------------------           Adj R-squared =  0.1981
       Total |  16663340.5   599  27818.5985           Root MSE      =  149.36

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
         ses |   3.276419    .401744     8.16   0.000                   [  3  ]
        male |  -86.89344   14.38207    [ 4 ]   0.000                -.2607064
       black |  -108.7724   17.90827    -6.07   0.000                -.2330604
       _cons |   470.4073    41.2488    11.40   0.000                        .
------------------------------------------------------------------------------

. test ses male

 ( 1)  ses = 0
 ( 2)  male = 0

       F(  2,   596) =   35.31
            Prob > F =    0.0000

. collin  ses male black

  Collinearity Diagnostics

                        SQRT                                   Cond     Cond          R-
  Variable      VIF     VIF    Tolerance        Eigenval       Index1   Index2     Squared
-------------------------------------------------------------------------------------------
       ses      1.49    1.22    [  5  ]    1.5664      1       1.0000     1.0000     0.3274
      male      1.39    1.18    0.7190     1.0075      2       1.2469     1.8602     0.2810
     black      1.10    1.05    0.9093     0.4262      3       1.9171     2.8634     0.0907
                                                       4                 16.5776
-------------------------------------------------------------------------------------------
  Mean VIF      1.33                        Condition Number   1.9171    16.5776
                       Determinant of correlation matrix    0.6725
                       Cond Index1 from deviation SSCP (no intercept)
                       Cond Index2 from scaled raw SSCP (w/ intercept)

. pcorr2  health ses male black
(obs=600)

Partial and Semipartial correlations of health with

    Variable |  Partial    SemiP     Sig.
-------------+---------------------------
         ses |   0.3168   0.2984    0.000
        male |  -0.2402  -0.2211    0.000
       black |  -0.2414  -0.2222    0.000
```

a) (15 pts) Fill in the missing quantities [1] – [5].

First off, here are the uncensored parts of the printout:

```
. reg health black

      Source |       SS       df       MS              Number of obs =     600
-------------+------------------------------           F(  1,   598) =   72.08
       Model |   1792453.8     1   1792453.8           Prob > F      =  0.0000
    Residual |  14870886.7   598  24867.7035           R-squared     =  0.1076
-------------+------------------------------           Adj R-squared =  0.1061
       Total |  16663340.5   599  27818.5985           Root MSE      =  157.69

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -153.0711   18.02964    -8.49   0.000    -188.4802    -117.662
       _cons |   786.4305   6.982851   112.62   0.000     772.7166    800.1444
------------------------------------------------------------------------------

. reg  health ses male black, beta

      Source |       SS       df       MS              Number of obs =     600
-------------+------------------------------           F(  3,   596) =   50.32
       Model |  3367869.52     3  1122623.17           Prob > F      =  0.0000
    Residual |    13295471   596  22307.8372           R-squared     =  0.2021
-------------+------------------------------           Adj R-squared =  0.1981
       Total |  16663340.5   599  27818.5985           Root MSE      =  149.36

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
         ses |   3.276419    .401744     8.16   0.000                  .363848
        male |  -86.89344   14.38207    -6.04   0.000                -.2607064
       black |  -108.7724   17.90827    -6.07   0.000                -.2330604
       _cons |   470.4073    41.2488    11.40   0.000                        .
------------------------------------------------------------------------------

. collin ses male black

  Collinearity Diagnostics

                        SQRT                                   Cond     Cond          R-
  Variable       VIF    VIF    Tolerance        Eigenval      Index1   Index2      Squared
-------------------------------------------------------------------------------------------
       ses      1.49    1.22    0.6726     1.5664      1     1.0000      1.0000     0.3274
      male      1.39    1.18    0.7190     1.0075      2     1.2469      1.8602     0.2810
     black      1.10    1.05    0.9093     0.4262      3     1.9171      2.8634     0.0907
                                                         4              16.5776
-------------------------------------------------------------------------------------------
  Mean VIF      1.33                      Condition Number    1.9171    16.5776
                       Determinant of correlation matrix    0.6725
                       Cond Index1 from deviation SSCP (no intercept)
                       Cond Index2 from scaled raw SSCP (w/ intercept)
```

To confirm that Stata got it right:

[1] F = MSR/MSE = 1,792,453.8/ 24,867.7035 = 72.08

[2] R^2 = SSR/SST = 3,367,869.52/16,663,340.5 = .2021

[3] $b'_{ses} = b_{ses} * s_{ses}/s_{health}$ = 3.276419 * 18.522/166.789 = .363848

[4] $t_{male} = b_{male}/se_{male}$ = -86.89344/14.38207 = -6.04

[5] $tol_{ses} = 1/vif_{ses}$ = 1/1.49 = .67; or, $tol_{ses} = 1 – R^2_{xkgk}$ = 1 - .3274 = .6726

b) (15 points) Interpret the results. Be sure to answer the following questions, explaining how the printout supports your conclusions.

1. What percentage of the sample is black? What percentage is male?

From the means, you can tell that 15% of the sample is black and 50% are males.

2. Who has higher socio-economic status – men or women?

Men do. You can tell that from the positive correlation (.5102) between male and ses.

3. Which variable has the strongest impact on health? Cite at least two or three pieces of evidence from the printout to support your conclusion on this point.

Socio-economic status. It has the largest t value, the largest standardized beta, and the largest partial and semi-partial values.

4. The effect of black declines once ses and male are added to the model (compare Model 1 with Model 2). Why? Offer a substantive explanation that is supported by the printout.

Note that black is negatively correlated with ses (-.2542) and that ses is positively correlated with health (.2901). This suggests that part of the reason blacks have poorer health than whites is because blacks tend to be of lower socio-economic status. As a result, they may be less able to afford quality health care, may be in more dangerous occupations, and be more likely to be exposed to problems related to poverty and health. Nonetheless, even after controlling for SES, significant racial differences in health remain. Perhaps there are racial barriers to health care or cultural differences between blacks and whites that affect their health.

5. According to the model, which types of individuals will tend to have the worst health?

Low ses black males will tend to have the worst health. High ses white women will have the best health.

c) (5 points) In the first regression, health is regressed on black only. In the second regression, health is regressed on black, ses, and male. Test whether the joint effects of male and ses significantly differ from zero, i.e. test

H_0: $\beta_{ses} = \beta_{male} = 0$
H_A: β_{ses} and/or $\beta_{male} \neq 0$

The kindly researcher has already done the work for you with the test command, which yields an F value of 35.31 with d.f. 2, 596. This value is highly significant, meaning we should reject the null. This is hardly surprising, given that the individual T values were so large.

For those who just don't trust computers to get these things right, you can do the calculations on your own:

$$F_{J,N-K-1} = \frac{(SSE_c - SSE_u) * (N - K - 1)}{SSE_u * J} = \frac{(14870886.7 - 13295471) * (600 - 3 - 1)}{13295471 * 2} = 35.31$$