Soc 63993, Homework #2 Answer Key: Multicollinearity/Missing Data

Richard Williams, University of Notre Dame, https://academicweb.nd.edu/~rwilliam/
Last revised January 22, 2015

I. Multicollinearity

[The following problem is adapted from Greene, Econometric Analysis, Fourth Edition.] The data in *longley.dta* (available at https://academicweb.nd.edu/~rwilliam/xsoc63993/index.html) were collected by James W. Longley ("An Appraisal of Least Squares Programs for the Electronic Computer from the point of view of the User," Journal of the American Statistical Association, Vol. 62, No. 319 (Sep. 1967), pp. 819-841) for the purpose of assessing the accuracy of least squares computations by computer programs. (If you want to see how they did things before the advent of modern computers, the article is available on JSTOR in the statistics journals.) Economic data were collected for the US for each of the years 1947-1962. The variables are:

Variable	Description
employ	Number of people employed (in thousands). This is the dependent variable in the analysis
price	Gross National Product Implicit Price Deflator. This is an adjustment for inflation. It equals 100 in the base year, 1954. Because of inflation, it is higher in years after 1954, and lower in years before that. A value of 110 would mean that, in that particular year, it cost $110 to buy the same goods that cost $100 in 1954.
gnp	Gross National Product (in millions of dollars)
armed	Size of armed forces (in thousands)
year	Year the data are from

Analyze these data with Stata. First, give the commands

```
. list
. summarize
```

just so you can get a feel for the characteristics of the data. Then give the command

```
. regress  employ price gnp armed year
```

Here are the initial results:

```
. list

     +----------------------------------------+
     | employ   price      gnp   armed   year |
     |----------------------------------------|
  1. |  60323      83   234289    1590   1947 |
  2. |  61122    88.5   259426    1456   1948 |
  3. |  60171    88.2   258054    1616   1949 |
  4. |  61187    89.5   284599    1650   1950 |
  5. |  63221    96.2   328975    3099   1951 |
     |----------------------------------------|
  6. |  63639    98.1   346999    3594   1952 |
  7. |  64989      99   365385    3547   1953 |
  8. |  63761     100   363112    3350   1954 |
  9. |  66019   101.2   397469    3048   1955 |
 10. |  67857   104.6   419180    2857   1956 |
     |----------------------------------------|
 11. |  68169   108.4   442769    2798   1957 |
 12. |  66513   110.8   444546    2637   1958 |
 13. |  68655   112.6   482704    2552   1959 |
 14. |  69564   114.2   502601    2514   1960 |
 15. |  69331   115.7   518173    2572   1961 |
     |----------------------------------------|
 16. |  70551   116.9   554894    2827   1962 |
     +----------------------------------------+
```

```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      employ |        16       65317    3511.968      60171      70551
       price |        16    101.6812    10.79155         83      116.9
         gnp |        16    387698.4    99394.94     234289     554894
       armed |        16    2606.688    695.9196       1456       3594
        year |        16      1954.5    4.760952       1947       1962

. regress  employ price gnp armed year

      Source |       SS       df       MS              Number of obs =      16
-------------+------------------------------           F(  4,    11) =  101.11
       Model |   180110100     4    45027525           Prob > F      =  0.0000
    Residual |  4898726.13    11  445338.739           R-squared     =  0.9735
-------------+------------------------------           Adj R-squared =  0.9639
       Total |   185008826    15  12333921.7           Root MSE      =  667.34

------------------------------------------------------------------------------
      employ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -19.76811   138.8927    -0.14   0.889     -325.469    285.9328
         gnp |    .064394   .0199519     3.23   0.008     .0204802    .1083078
       armed |  -.0101452   .3085695    -0.03   0.974     -.689302    .6690116
        year |  -576.4642   433.4875    -1.33   0.210    -1530.564    377.6353
       _cons |    1169087   835902.5     1.40   0.189    -670721.5     3008896
------------------------------------------------------------------------------
```

The data suggest steady growth across time in employment, GNP, and inflation. This is not surprising, given that these were postwar boom years. The size of the armed forces fluctuated somewhat. There was a big boost during the Korean War and then troop sizes declined a bit.

In the regression, only gnp has a significant effect on employment. However, given the way these variables all changed together across time, it would not be surprising to find that they are highly correlated and that multicollinearity might be an issue.

Then, do further examination to determine what evidence, if any, suggests that multicollinearity may or may not be present in these data. Estimate and examine the bivariate correlations, tolerances/VIFs, condition numbers, the sample size, and anything else that you think would help to diagnose a problem of multicollinearity if it existed. For everything you do, be sure to explain what it means and how it applies to multicollinearity; don't just give numbers without explanation. If you find that multicollinearity is present, offer a substantive explanation for it, i.e. why are these variables so highly correlated with each other? [Optional - Offer any suggestions you may have for dealing with the problem.]

```
. corr  employ price gnp armed year

(obs=16)

             |   employ    price      gnp    armed     year
-------------+---------------------------------------------
      employ |   1.0000
       price |   0.9709   1.0000
         gnp |   0.9836   0.9916   1.0000
       armed |   0.4573   0.4647   0.4464   1.0000
        year |   0.9713   0.9911   0.9953   0.4172   1.0000
```

Except for armed, these variables have very high intercorrelations with each other, .97 and above.

```
. collin  price gnp armed year

  Collinearity Diagnostics

                        SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
     price     75.67    8.70    0.0132      0.9868
       gnp    132.46   11.51    0.0075      0.9925
     armed      1.55    1.25    0.6438      0.3562
      year    143.46   11.98    0.0070      0.9930
----------------------------------------------------
  Mean VIF     88.29

                           Cond
        Eigenval          Index
---------------------------------
    1     4.9199          1.0000
    2     0.0450         10.4553
    3     0.0349         11.8684
    4     0.0001        198.1631
    5     0.0000      15824.1489
---------------------------------
 Condition Number     15824.1489
 Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
 Det(correlation matrix)    0.0001

. collin  price gnp armed year, corr

[Repetive material deleted]

                           Cond
        Eigenval          Index
---------------------------------
    1     3.2471          1.0000
    2     0.7397          2.0952
    3     0.0090         18.9611
    4     0.0042         27.9611
---------------------------------
 Condition Number        27.9611
 Eigenvalues & Cond Index computed from deviation sscp (no intercept)
 Det(correlation matrix)    0.0001
```

Except for armed, the vifs are all extremely high, well over the rule of thumb figure of 10. For price, gnp and year, their standard errors will be 8.7 to 11.98 times larger than they would be if the variables were uncorrelated. The raw score Condition index may be the most appropriate of the two indices because the variables are all ratio level, and its value is almost 16,000! Even the centered condition index is very large. The N is extremely small, so that won't help us much either.

Also, lets take a look at the standardized betas:

```
. reg, beta

      Source |       SS       df       MS              Number of obs =      16
-------------+------------------------------           F(  4,    11) =  101.11
       Model |   180110100     4    45027525           Prob > F      =  0.0000
    Residual |  4898726.13    11  445338.739           R-squared     =  0.9735
-------------+------------------------------           Adj R-squared =  0.9639
       Total |   185008826    15  12333921.7           Root MSE      =  667.34

------------------------------------------------------------------------------
      employ |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
       price |  -19.76811   138.8927    -0.14   0.889                -.0607433
         gnp |    .064394   .0199519     3.23   0.008                 1.822464
       armed |  -.0101452   .3085695    -0.03   0.974                -.0020103
        year |  -576.4642   433.4875    -1.33   0.210                -.7814759
       _cons |    1169087   835902.5     1.40   0.189                        .
------------------------------------------------------------------------------
```

Even though price, gnp and year have almost identical correlations with employ, there is a vast difference in their standardized effects. Also, a standardized effect larger than 1 is extremely unusual, and is further evidence of multicollinearity.

As far as possible solutions go, you might try something like

```
. gen gnpadj = gnp/(price/100)
. reg  employ gnpadj armed year
```

gnpadj is gnp adjusted for inflation, i.e. it is the value of the gnp in 1954 dollars. The use of inflation-adjusted dollars gives us a clearer picture of how gnp was really changing across time. Conceptually, it probably makes more sense to be using adjusted gnp anyway, and this will eliminate one of the highly collinear variables from the model. Rerunning some of our earlier analyses with this new measure,

```
. regress  employ gnpadj armed year, beta

      Source |       SS       df       MS              Number of obs =      16
-------------+------------------------------           F(  3,    12) =  173.04
       Model |   180828691     3  60276230.3           Prob > F      =  0.0000
    Residual |  4180135.09    12  348344.591           R-squared     =  0.9774
-------------+------------------------------           Adj R-squared =  0.9718
       Total |   185008826    15  12333921.7           Root MSE      =  590.21

------------------------------------------------------------------------------
      employ |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      gnpadj |   .0863357   .0213993     4.03   0.002                 1.450322
       armed |  -.4148106   .3017286    -1.37   0.194                -.0821974
        year |   -315.743   253.5094    -1.25   0.237                -.4280328
       _cons |   651097.1   487959.6     1.33   0.207                        .
------------------------------------------------------------------------------
```

```
. collin  gnpadj armed year

  Collinearity Diagnostics

                        SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
    gnpadj     68.63    8.28    0.0146      0.9854
     armed      1.90    1.38    0.5267      0.4733
      year     62.73    7.92    0.0159      0.9841
----------------------------------------------------
  Mean VIF     44.42

                           Cond
        Eigenval          Index
---------------------------------
    1     3.9451          1.0000
    2     0.0423          9.6595
    3     0.0126         17.6742
    4     0.0000       9361.8280
---------------------------------
 Condition Number      9361.8280
 Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
 Det(correlation matrix)    0.0120

. collin  gnpadj armed year, corr

[Repetitive material deleted]

                           Cond
        Eigenval          Index
---------------------------------
    1     2.3072          1.0000
    2     0.6852          1.8351
    3     0.0076         17.4095
---------------------------------
 Condition Number        17.4095
 Eigenvalues & Cond Index computed from deviation sscp (no intercept)
 Det(correlation matrix)    0.0120
```

The collinearity measures are not as extreme as they were before, but they are still quite large. Looking at the correlations of the remaining xs, we see

```
. corr  gnpadj armed year

(obs=16)

             |   gnpadj    armed     year
-------------+---------------------------
      gnpadj |   1.0000
       armed |   0.4951   1.0000
        year |   0.9885   0.4172   1.0000
```

gnpadj and year are very highly correlated; furthermore, the effect of year is not statistically significant. Conceptually, we might wonder if year is really important, or is the important thing those variables that tend to change by year. All of this suggests that year may not be an essential variable in the model. Hence, lets see what happens when we drop it:

```
. regress  employ gnpadj armed, beta

      Source |       SS       df       MS              Number of obs =      16
-------------+------------------------------           F(  2,    13) =  248.25
       Model |   180288324     2  90144162.2           Prob > F      =  0.0000
    Residual |  4720501.68    13  363115.514           R-squared     =  0.9745
-------------+------------------------------           Adj R-squared =  0.9706
       Total |   185008826    15  12333921.7           Root MSE      =  602.59

------------------------------------------------------------------------------
      employ |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      gnpadj |   .0599416   .0030354    19.75   0.000                 1.006937
       armed |  -.2082112    .257329    -0.81   0.433                -.0412584
       _cons |   43350.33   1007.374    43.03   0.000                        .
------------------------------------------------------------------------------

. collin gnpadj armed

  Collinearity Diagnostics

                        SQRT                                   Cond     Cond         R-
  Variable       VIF    VIF    Tolerance       Eigenval       Index1   Index2     Squared
-------------------------------------------------------------------------------------------
    gnpadj      1.32    1.15    0.7548      1.4951      1      1.0000      1.0000      0.2452
     armed      1.32    1.15    0.7548      0.5049      2      1.7209      9.2708      0.2452
                                                        3                 16.7569
-------------------------------------------------------------------------------------------
  Mean VIF      1.32                     Condition Number    1.7209    16.7569
                     Determinant of correlation matrix    0.7548
                     Cond Index1 from deviation SSCP (no intercept)
                     Cond Index2 from scaled raw SSCP (w/ intercept)
```

There no longer appear to be any multicollinearity issues. (We might want to consider dropping armed too, because its effect is not significant.)

In short, by using a more appropriate measure of inflation-adjusted gnp, and by dropping the questionable year variable, we were able to resolve the issues of multicollinearity with these data. (A remaining issue may be the appropriateness of using OLS regression in the first place; while the gnp probably affects employment, employment also probably affects gnp, i.e. the causal relationships do not just run one way. We'll talk about such issues later in the semester.)

II. Multiple Imputation

A. Run the following commands:

```
use "https://academicweb.nd.edu/~rwilliam/statafiles/md.dta", clear
sum income educ jobexp black other
reg income educ jobexp black other
```

Now use multiple imputation to impute the missing values for educ and rerun the regression. You will need to use the `mi set`, `mi register`, `mi impute`, and `mi estimate` commands. When running the imputations you should specify 50 imputations with an rseed of 2232 (otherwise everybody will get different results!). Briefly explain your reasoning behind each step, e.g. why did you choose the imputation method that you did, how did you choose the variables for the imputation model, what is the purpose of the command you are using? You should find that, in this case, the results from using multiple imputation are not that different from the results using listwise deletion.

```
. use "https://academicweb.nd.edu/~rwilliam/statafiles/md.dta", clear
. sum income educ jobexp black other

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      income |       500       27.79    8.973491          5       48.3
        educ |       405    13.01728    3.974821          2         21
      jobexp |       500       13.52    5.061703          1         21
       black |       500          .2    .4004006          0          1
       other |       500          .1    .3003005          0          1
```

Educ has missing data on 95 cases but the other variables have complete data. Those 95 cases get dropped from the regression, even though the other variables are not missing data.

```
. reg income educ jobexp black other

      Source |       SS       df       MS              Number of obs =     405
-------------+------------------------------           F(  4,   400) =  608.74
       Model |  27795.9439     4  6948.98598           Prob > F      =  0.0000
    Residual |  4566.17485   400  11.4154371           R-squared     =  0.8589
-------------+------------------------------           Adj R-squared =  0.8575
       Total |  32362.1188   404  80.1042544           Root MSE      =  3.3787

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.762008   .0482888    36.49   0.000     1.667076    1.856939
      jobexp |   .6132015   .0360704    17.00   0.000     .5422903    .6841127
       black |   -3.71989    .485472    -7.66   0.000    -4.674285   -2.765494
       other |  -5.162724    .566557    -9.11   0.000    -6.276525   -4.048923
       _cons |  -2.370497   .9712102    -2.44   0.015    -4.279811   -.4611829
------------------------------------------------------------------------------

. mi set mlong
```

The mi set command tells Stata that this is going to be an mi data set. The style mlong is good because it is memory efficient, i.e. it requires less storage space.

```
. mi register imputed educ
(95 m=0 obs. now marked as incomplete)

. mi register regular income jobexp black other white
```

The missing values of educ will be imputed. The values of the other variables, missing or non-missing, will be left as is.

```
. mi impute regress educ income jobexp black other, add(50) rseed(2232)

Univariate imputation                       Imputations =       50
Linear regression                                 added =       50
Imputed: m=1 through m=50                       updated =        0

------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
              educ |        405           95        95 |       500
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

Educ is imputed using all the variables in the analytic model, both dependent and independent. If some were excluded relationships involving that variable would be biased toward 0. The add option causes fifty imputations to be done. The rseed option will let us reproduce the exact same results later if we wish.

```
. mi estimate, dots: regress income educ jobexp black other

Imputations (50):
  .........10.........20.........30.........40.........50 done

Multiple-imputation estimates                     Imputations     =         50
Linear regression                                 Number of obs   =        500
                                                  Average RVI     =     0.1753
                                                  Largest FMI     =     0.2164
                                                  Complete DF     =        495
DF adjustment:   Small sample                     DF:     min     =     284.36
                                                          avg     =     363.51
                                                          max     =     390.93
Model F test:       Equal FMI                     F(   4,  463.3) =     630.52
Within VCE type:          OLS                     Prob > F        =     0.0000

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.785123   .0459292    38.87   0.000      1.69481    1.875435
      jobexp |   .6217021    .035947    17.29   0.000     .5509462     .692458
       black |  -3.466247   .4677852    -7.41   0.000    -4.385946   -2.546547
       other |  -5.047868   .5528513    -9.13   0.000    -6.134801   -3.960934
       _cons |  -2.839132   .9140529    -3.11   0.002    -4.636369   -1.041895
------------------------------------------------------------------------------
```

We redo the estimation with the imputed data. All 500 cases are now used. In this particular case, the changes from listwise appear fairly minor, but that will not always be true.

B. This problem is adapted from Paul Allison's 2009 book *Fixed Effects Regression Models*. Data are from the National Longitudinal Study of Youth (NLSY). This subset of the data set has 1151 teenage girls who were interviewed annually for 5 years beginning in 1979. Only the fifth and final wave is used here. I have modified the data set so that some values are missing.

- id is the subject id number and is the same across each wave of the survey
- pov is coded 1 if the subject was in poverty during that time period, 0 otherwise.
- age is the age at last interview.
- mother is coded 1 if the respondent currently has at least 1 child, 0 otherwise.
- spouse is coded 1 if the respondent is currently living with a spouse, 0 otherwise.
- hours is the hours worked during the week of the survey.

Start with the command

```
use "https://academicweb.nd.edu/~rwilliam/statafiles/mdpov2.dta", clear
```

You eventually want to run the commands

```
mi xeq 0: logit pov age mother spouse hours
mi estimate, dots: logit pov age mother spouse hours
```

Before you can do that though, you must do the following. Briefly explain your reasoning behind each step, e.g. why did you choose the imputation method that you did, how did you choose the variables for the imputation model, what is the purpose of the command you are using?

- mi set the data.

- Identify the two variables that have missing data, and decide what imputation method is appropriate, e.g. regress, logit, mlogit. [NOTE: Different methods will be required.] The `mi misstable summarize` command is one way of doing this, but there are other ways that will work just as well.
- Register the variables to be imputed.
- Use `mi impute chained` to impute the two variables. Since two variables are imputed and different methods are being used, the syntax will be something like

```
mi impute chained (mlogit) x1 (poisson) x2 = v1 v2 v3 v4 …
```

where mlogit and poisson and the variable names are replaced by appropriate values.

- Do 20 imputations using an rseed of 2232. If everybody doesn't use the same rseed, you will get different results.

After doing the above, note any differences between the imputed and unimputed results, e.g. differences in sample size, coefficients, and standard errors. Most of the differences are modest in this case.

Here is one way to do all of this.

```
. use "https://academicweb.nd.edu/~rwilliam/statafiles/mdpov2.dta", clear
. mi set mlong
. mi misstable summarize
                                                               Obs<.
                                                +------------------------------
               |                                | Unique
      Variable |     Obs=.     Obs>.     Obs<.  | values        Min         Max
  -------------+--------------------------------+------------------------------
           age |       228                 923  |      4         18          21
        mother |       338                 813  |      2          0           1
  -----------------------------------------------------------------------------

. mi misstable patterns

   Missing-value patterns
     (1 means complete)

              |   Pattern
    Percent   |  1  2
  ------------+-------------
       57%    |  1  1
              |
       23     |  1  0
       13     |  0  1
        6     |  0  0
  ------------+-------------
      100%    |

  Variables are  (1) age  (2) mother
```

We see that the problem variables are age and mother. About 43% of the cases are missing data on either or both. Just to make sure of their coding, we can use the `fre` command (which needs to be installed; if it isn't `tab1` will work).

```
. fre age mother

age -- age of r at interview date curr yr
-----------------------------------------------------------
              |      Freq.    Percent      Valid       Cum.
--------------+--------------------------------------------
Valid   18    |        153      13.29      16.58      16.58
        19    |        257      22.33      27.84      44.42
        20    |        269      23.37      29.14      73.56
        21    |        244      21.20      26.44     100.00
        Total |        923      80.19     100.00
Missing .     |        228      19.81
Total         |       1151     100.00
-----------------------------------------------------------

mother
-----------------------------------------------------------
              |      Freq.    Percent      Valid       Cum.
--------------+--------------------------------------------
Valid   0     |        539      46.83      66.30      66.30
        1     |        274      23.81      33.70     100.00
        Total |        813      70.63     100.00
Missing .     |        338      29.37
Total         |       1151     100.00
-----------------------------------------------------------
```

Regress and logit would appear to be reasonable choices for imputation models. We could also try using pmm (Predictive Mean Matching) for age.

```
. mi register imputed age mother
(492 m=0 obs. now marked as incomplete)
. mi register regular id pov spouse hours
. mi impute chained (regress) age (logit) mother = pov spouse hours, add(20) rseed(2232)

Conditional models:
               age: regress age i.mother pov spouse hours
            mother: logit mother age pov spouse hours

Performing chained iterations ...

Multivariate imputation                     Imputations =       20
Chained equations                                 added =       20
Imputed: m=1 through m=20                       updated =        0

Initialization: monotone                     Iterations =      200
                                                burn-in =       10

               age: linear regression
            mother: logistic regression

------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
               age |        923          228       228 |      1151
            mother |        813          338       338 |      1151
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

Note that the imputation models include all of the variables in the analytic model, including the dependent variable pov. That is, the analytic model and the imputation model are congenial. If we did not do this, relationships with the variables that have been omitted would be biased toward 0, e.g. if we left out pov we would likely underestimate how strongly related it is to age and mother.

```
. mi xeq 0: logit pov age mother spouse hours

m=0 data:
-> logit pov age mother spouse hours

Iteration 0:   log likelihood = -442.43908
Iteration 1:   log likelihood = -397.43515
Iteration 2:   log likelihood = -396.74436
Iteration 3:   log likelihood = -396.74254
Iteration 4:   log likelihood = -396.74254

Logistic regression                               Number of obs   =        659
                                                  LR chi2(4)      =      91.39
                                                  Prob > chi2     =     0.0000
Log likelihood = -396.74254                       Pseudo R2       =     0.1033

------------------------------------------------------------------------------
         pov |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1744982   .0837387    -2.08   0.037     -.338623   -.0103734
      mother |   1.002909   .2046794     4.90   0.000     .6017444    1.404073
      spouse |  -1.278553   .2583428    -4.95   0.000    -1.784895   -.7722099
       hours |  -.0338663   .0058632    -5.78   0.000     -.045358   -.0223746
       _cons |   3.227516   1.624431     1.99   0.047     .0436898    6.411343
------------------------------------------------------------------------------

. mi estimate, dots: logit pov age mother spouse hours

Imputations (20):
  .........10.........20 done

Multiple-imputation estimates                     Imputations     =         20
Logistic regression                               Number of obs   =       1151
                                                  Average RVI     =     0.0775
                                                  Largest FMI     =     0.1662
DF adjustment:   Large sample                     DF:     min     =     707.56
                                                          avg     =   45657.50
                                                          max     =  190649.98
Model F test:       Equal FMI                     F(   4,10366.5) =      28.78
Within VCE type:          OIM                     Prob > F        =     0.0000

------------------------------------------------------------------------------
         pov |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1563587   .0705805    -2.22   0.027    -.2949247   -.0177926
      mother |    1.09854   .1620649     6.78   0.000     .7805984    1.416482
      spouse |  -1.175587   .1975277    -5.95   0.000    -1.562747   -.7884262
       hours |  -.0324061   .0044964    -7.21   0.000     -.041219   -.0235933
       _cons |   2.689299   1.370235     1.96   0.050    -.0009151    5.379512
------------------------------------------------------------------------------
```

The imputed data uses 492 more cases in the analysis. Mother becomes more significant, probably because we picked up cases with data on mother that were missing on age. Spouse and hours also become more significant. The changes in coefficients are pretty modest. I set the problem up so that missing data were MCAR, so it isn't too surprising that the changes mostly involve smaller standard errors and greater statistical significance.

If for some reason you had this mad urge to do predictive mean matching instead:

```
. * Use pmm instead
. use "https://academicweb.nd.edu/~rwilliam/statafiles/mdpov2.dta", clear
. mi set mlong
. mi register imputed age mother
(492 m=0 obs. now marked as incomplete)
. mi register imputed id pov spouse hours
. mi impute chained (pmm, knn(5)) age (logit) mother = pov spouse hours, add(20)
rseed(2232)

Conditional models:
               age: pmm age i.mother pov spouse hours , knn(5)
            mother: logit mother age pov spouse hours

Performing chained iterations ...

Multivariate imputation                     Imputations =       20
Chained equations                                 added =       20
Imputed: m=1 through m=20                       updated =        0

Initialization: monotone                     Iterations =      200
                                                burn-in =       10

               age: predictive mean matching
            mother: logistic regression

------------------------------------------------------------------
                   |               Observations per m             
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
               age |        923          228       228 |      1151
            mother |        813          338       338 |      1151
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

```
. mi estimate, dots: logit pov age mother spouse hours

Imputations (20):
  .........10.........20 done

Multiple-imputation estimates                   Imputations     =         20
Logistic regression                             Number of obs   =       1151
                                                Average RVI     =     0.2033
                                                Largest FMI     =     0.4382
DF adjustment:   Large sample                   DF:     min     =     103.95
                                                        avg     =    6109.13
                                                        max     =   27747.83
Model F test:       Equal FMI                   F(   4, 1713.0) =      25.09
Within VCE type:          OIM                   Prob > F        =     0.0000

------------------------------------------------------------------------------
         pov |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1576892   .0724014    -2.18   0.030    -.2999786   -.0153998
      mother |   1.100815   .2012856     5.47   0.000     .7016558    1.499974
      spouse |  -1.188204    .206692    -5.75   0.000    -1.593591   -.7828171
       hours |  -.0324396   .0045338    -7.16   0.000     -.041326   -.0235532
       _cons |   2.717243    1.39775     1.94   0.052    -.0291912    5.463677
------------------------------------------------------------------------------
```

There are no obvious advantages to using PMM instead of regress in this case.

III. *Missing data (Traditional Methods)*

For this problem, you need to copy and run *missing.do* and *missing.dta* from my web page. You may need to tweak the code to get the right location for the data file. This question tests your understanding of missing data concepts, but it also illustrates some basic data manipulation techniques.

A rookie researcher is investigating how several major demographic factors affect one's income. She uses the General Social Survey of 1991. Her assistant has included many comments in the following programs, but she needs your help to understand exactly what was done and how to interpret her results.

a. Based on the frequencies from part 1 of the program, how prevalent is missing data? Does it exist primarily in the DV (Income), one or more of the IVs, or both?

b. In part 2, why do you think her assistant decided to recode the income variable? Why didn't the assistant think MD was being handled correctly in the original coding?

c. [Optional] What exactly is her assistant doing in part 3, and why? Why did she create a variable called WHITE, but not create a variable called BLACK? (Careful – be sure you look at the frequencies for RACE before answering this.)

d. Likewise, in part 4, why does the assistant create the PAEDUC2 and MDPAEDUC variables? Why are they coded that way?

e. [Optional] In parts 5-8, why does her assistant run the regressions 3 different ways (a fourth is possible in SPSS)? Why does the sample size differ in the various approaches? Do the different results seem to lead to different conclusions, and if so, why?

f. [Optional] In part 7, why does the assistant make the comment that mean substitution on the DV seems questionable?

g. In part 8, the assistant comments that "The final regression will give us an idea of whether or not the MD in PAEDUC is missing on a random basis." How does the regression do that??? What does the coefficient for MDPAEDUC supposedly tell you? Would Allison approve or disapprove of what the assistant is doing here? Why?

h. [Optional] Given the nature of the missing data, which approach do you think is most appropriate in this case? Why? Why are the other approaches less desirable? Briefly describe what the main substantive conclusions are from your preferred model (e.g. which variables are important, what effect do the main variables have on income, etc.)

i. [Optional] Do you have any other suggestions for deciding how to handle the MD? Present any additional analyses you think might be helpful. For example, you might examine whether men or women are more likely to have missing data on income.

Here is the Stata program:

missing.do

```
version 9.2
set more off

* Change the -use- command if you want to use a local copy of the data.
use "https://academicweb.nd.edu/~rwilliam/statafiles/missing.dta", clear

* Part 1.  Do frequencies/descriptives on the original vars. Look at MD
* patterns, problems with coding. The -fre- command, available from
* ssc, needs to be installed.
sum rincome educ age sex race paeduc
fre rincome educ age sex race paeduc, tab(10)

* Part 2.  I don't like the way RINCOME is coded.  I also don't think the
* MD categories are quite right.  Create a new variable, INCOME,
* that is coded better. I won't distinguish between MD codes.
recode rincome  (1=.5) (2=2) (3=3) (4=4.5) (5=5.5) (6=6.5) (7=7.5) (8=9) ///
   (9=12.5) (10=17.5) (11=22.5) (12=25) (else=.), gen(income)
fre income

* Part 3.  Let's fix the RACE and SEX variables too.  Even though race
* has 3 categories, I think it is better to only make one dummy.
recode race (1=1)(else=0), gen(white)
recode sex (1=1)(else=0), gen(male)
fre white male

* Part 4.  Create a modified PAEDUC2 that I can use later.  Create
* an MD indicator. Using the impute command makes it
* easy and also more precise.
gen one = 1
gen mdpaeduc = missing(paeduc)
impute paeduc one, gen(paeduc2)
fre   paeduc2 mdpaeduc

* Part 5.  Listwise deletion of MD.
reg  income educ age male paeduc white

* Part 6.  Sorry, unlike SPSS, no easy way to do pairwise in Stata.  If I was a fanatic
* about it, I could probably use the pwcorr and corr2data commands.

* Part 7.  Mean substitution of MD (both IVs and DVs).  Seems questionable for
* the DV. I'll use the impute command to create new vars
* with the mean substituted for MD.
impute income one, gen(incomex)
impute educ one, gen(educx)
impute age one, gen(agex)
impute male one, gen(malex)
impute paeduc one, gen(paeducx)
impute white one, gen(whitex)
reg  incomex educx agex malex paeducx whitex

* Part 8.  Mean substitution, Father's education only, without and then with an MD indicator.
* The final regression will give us an idea of whether or not the MD in PAEDUC is missing
* on a random basis.
reg  income educ age male paeduc2 white
reg  income educ age male paeduc2 white mdpaeduc
```

```
* Part 9. Add any additional analyses you think are useful.
```

A few other comments about how you might extend the analysis using Stata, and the differences between Stata and SPSS:

* The `tab1` and `summarize` commands in Stata are some of the many ways you can get descriptive statistics, such as SPSS gives you with the Frequencies command. You may have to run tab1 twice, both with and without the `nolabel` option. The `fre` command, available from SSC, is often much better than the `tab1` command.

* As explained in the class notes, there are various ways to plug in values for missing data, some of which are easier or at least different than their SPSS counterparts

* Stata does not have a pairwise deletion option, which is why Part 6 could be easily done in SPSS but not Stata.

* SPSS lets you use whatever values you want as missing, e.g. 97, 98, 99. Stata does things differently. Missing data has values of ., .a, .b, etc., through .z. As a consequence, missing.dta uses the values .a, .b and .c for the missing data, rather than the values used in the original SPSS file. Stata does not have a separate missing values command like SPSS does; if you want data to be missing, you have to code or recode it to the values ., .a, .b, etc.

* Here are some of the commands you may find useful. Use `help` if you need help for any of them. You can also use the Stata menus, of course.

```
tab1        generate    if          summarize
replace     recode      impute      fre
```

Here is how you can solve the problem using Stata. I sometimes rearrange or edit the output.

a. Based on the frequencies from part 1 of the program, how prevalent is missing data? Does it exist primarily in the DV (Income), one or more of the IVs, or both?

```
. * Part 1.  Do frequencies/descriptives on the original vars. Look at MD
. * patterns, problems with coding. The -fre- command, available from
. * ssc, needs to be installed.

. sum rincome educ age sex race paeduc

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     rincome |       952    9.338235    3.357915          1         13
        educ |      1510    12.88411    2.984022          0         20
         age |      1514    45.62616    17.80842         18         89
         sex |      1517    1.580751    .4935988          1          2
        race |      1517    1.199077    .4734917          1          3
-------------+--------------------------------------------------------
      paeduc |      1069     10.8812    4.128542          0         20

. fre rincome educ age sex race paeduc, tab(10)
[Output is interspersed below]
```

Most of the missing data is in rincome and paeduc.

b. In part 2, why do you think her assistant decided to recode the income variable? Why didn't the assistant think MD was being handled correctly in the original coding?

```
rincome -- RESPONDENTS INCOME
-----------------------------------------------------------------------
                             |      Freq.    Percent      Valid       Cum.
-----------------------------+-----------------------------------------
Valid   1  LT $1000          |         36       2.37       3.78       3.78
        2  $1000 TO 2999     |         34       2.24       3.57       7.35
        3  $3000 TO 3999     |         35       2.31       3.68      11.03
```

```
        4  $4000 TO 4999     |         29       1.91       3.05      14.08
        5  $5000 TO 5999     |         35       2.31       3.68      17.75
        6  $6000 TO 6999     |         16       1.05       1.68      19.43
        7  $7000 TO 7999     |         14       0.92       1.47      20.90
        8  $8000 TO 9999     |         41       2.70       4.31      25.21
        9  $10000 - 14999    |        119       7.84      12.50      37.71
        10 $15000 - 19999    |        127       8.37      13.34      51.05
        11 $20000 - 24999    |        105       6.92      11.03      62.08
        12 $25000 OR MORE    |        321      21.16      33.72      95.80
        13 refused           |         40       2.64       4.20     100.00
        Total                |        952      62.76     100.00
Missing .a MD-Not Applicable |        463      30.52
        .b MD-Don't Know     |          7       0.46
        .c MD-No Answer      |         95       6.26
        Total                |        565      37.24
Total                        |       1517     100.00
--------------------------------------------------------------------------
```

```
. * Part 2.  I don't like the way RINCOME is coded.  I also don't think the
. * MD categories are quite right.  Create a new variable, INCOME,
. * that is coded better. I won't distinguish between MD codes.
. recode rincome  (1=.5) (2=2) (3=3) (4=4.5) (5=5.5) (6=6.5) (7=7.5) (8=9) ///
>    (9=12.5) (10=17.5) (11=22.5) (12=25) (else=.), gen(income)
(1448 differences between rincome and income)

. fre income

income -- RECODE of rincome (RESPONDENTS INCOME)
-----------------------------------------------------------
              |      Freq.    Percent      Valid       Cum.
--------------+--------------------------------------------
Valid   .5    |         36       2.37       3.95       3.95
        2     |         34       2.24       3.73       7.68
        3     |         35       2.31       3.84      11.51
        4.5   |         29       1.91       3.18      14.69
        5.5   |         35       2.31       3.84      18.53
        6.5   |         16       1.05       1.75      20.29
        7.5   |         14       0.92       1.54      21.82
        9     |         41       2.70       4.50      26.32
        12.5  |        119       7.84      13.05      39.36
        17.5  |        127       8.37      13.93      53.29
        22.5  |        105       6.92      11.51      64.80
        25    |        321      21.16      35.20     100.00
        Total |        912      60.12     100.00
Missing .     |        605      39.88
Total         |       1517     100.00
-----------------------------------------------------------
```

The original coding was ordinal at best – distance between categories was not the same. In the new coding, the midpoint of the original intervals is used. Category 13 (Refused) was not being treated as MD in the original, which is a mistake.

c. [Optional] What exactly is her assistant doing in part 3, and why? Why did she create a variable called WHITE, but not create a variable called BLACK? (Careful – be sure you look at the frequencies for RACE before answering this.)

```
race -- RACE OF RESPONDENT
-------------------------------------------------------------
                |      Freq.    Percent      Valid       Cum.
----------------+--------------------------------------------
Valid   1 white |       1264      83.32      83.32      83.32
        2 black |        204      13.45      13.45      96.77
        3 other |         49       3.23       3.23     100.00
        Total   |       1517     100.00     100.00
-------------------------------------------------------------

. * Part 3.  Let's fix the RACE and SEX variables too.  Even though race
. * has 3 categories, I think it is better to only make one dummy.
. recode race (1=1)(else=0), gen(white)
(253 differences between race and white)
. recode sex (1=1)(else=0), gen(male)
(881 differences between sex and male)

. fre white male

white -- RECODE of race (RACE OF RESPONDENT)
-----------------------------------------------------------
              |      Freq.    Percent      Valid       Cum.
--------------+--------------------------------------------
Valid   0     |        253      16.68      16.68      16.68
        1     |       1264      83.32      83.32     100.00
        Total |       1517     100.00     100.00
-----------------------------------------------------------

male -- RECODE of sex (RESPONDENTS SEX)
-----------------------------------------------------------
              |      Freq.    Percent      Valid       Cum.
--------------+--------------------------------------------
Valid   0     |        881      58.08      58.08      58.08
        1     |        636      41.92      41.92     100.00
        Total |       1517     100.00     100.00
-----------------------------------------------------------
```

She is computing dummy vars out of race and gender. Although race has 3 categories, only a very small number of cases fall into the "other" category, which could create multicollinearity problems if 3 dummies were used.

d. Likewise, in part 4, why does the assistant create the PAEDUC2 and MDPAEDUC variables? Why are they coded that way?

```
paeduc -- HIGHEST YEAR SCHOOL COMPLETED, FATHER
------------------------------------------------------------
               |      Freq.    Percent      Valid       Cum.
---------------+--------------------------------------------
Valid   0      |         17       1.12       1.59       1.59
        2      |          7       0.46       0.65       2.25
        3      |         31       2.04       2.90       5.14
        4      |         22       1.45       2.06       7.20
        5      |         22       1.45       2.06       9.26
        6      |         61       4.02       5.71      14.97
        7      |         27       1.78       2.53      17.49
        8      |        165      10.88      15.43      32.93
        9      |         39       2.57       3.65      36.58
        10     |         49       3.23       4.58      41.16
        11     |         38       2.50       3.55      44.71
        12     |        300      19.78      28.06      72.78
        13     |         28       1.85       2.62      75.40
        14     |         77       5.08       7.20      82.60
```

```
        15     |         12       0.79       1.12      83.72
        16     |        103       6.79       9.64      93.36
        17     |         12       0.79       1.12      94.48
        18     |         24       1.58       2.25      96.73
        19     |         13       0.86       1.22      97.94
        20     |         22       1.45       2.06     100.00
        Total  |       1069      70.47     100.00
Missing .a nap |        205      13.51
        .b dk  |        211      13.91
        .c na  |         32       2.11
        Total  |        448      29.53
Total          |       1517     100.00
-----------------------------------------------------------

. * Part 4.  Create a modified PAEDUC2 that I can use later.  Create
. * an MD indicator. Using the impute command makes it
. * easy and also more precise.
. gen one = 1
. gen mdpaeduc = missing(paeduc)
. impute paeduc one, gen(paeduc2)
 29.53% (448) observations imputed

. fre   paeduc2 mdpaeduc

paeduc2 -- imputed paeduc
------------------------------------------------------------
                |      Freq.    Percent      Valid       Cum.
----------------+-------------------------------------------
Valid   0       |         17       1.12       1.12       1.12
        2       |          7       0.46       0.46       1.58
        3       |         31       2.04       2.04       3.63
        4       |         22       1.45       1.45       5.08
        5       |         22       1.45       1.45       6.53
        6       |         61       4.02       4.02      10.55
        7       |         27       1.78       1.78      12.33
        8       |        165      10.88      10.88      23.20
        9       |         39       2.57       2.57      25.77
        10      |         49       3.23       3.23      29.00
        10.8812 |        448      29.53      29.53      58.54
        11      |         38       2.50       2.50      61.04
        12      |        300      19.78      19.78      80.82
        13      |         28       1.85       1.85      82.66
        14      |         77       5.08       5.08      87.74
        15      |         12       0.79       0.79      88.53
        16      |        103       6.79       6.79      95.32
        17      |         12       0.79       0.79      96.11
        18      |         24       1.58       1.58      97.69
        19      |         13       0.86       0.86      98.55
        20      |         22       1.45       1.45     100.00
        Total   |       1517     100.00     100.00
------------------------------------------------------------

mdpaeduc
----------------------------------------------------------
              |      Freq.    Percent      Valid       Cum.
--------------+-------------------------------------------
Valid   0     |       1069      70.47      70.47      70.47
        1     |        448      29.53      29.53     100.00
        Total |       1517     100.00     100.00
----------------------------------------------------------
```

She wants to use the mean substitution technique with a missing data dummy variable indicator. The 448 missing data cases in PAEDUC are set equal to the mean (10.88).

e. [Optional] In parts 5-8, why does her assistant run the regressions 3 different ways (a fourth is possible SPSS)? Why does the sample size differ in the various approaches? Do the different results seem to lead to different conclusions, and if so, why?

```
. * Part 5.  Listwise deletion of MD.
. reg  income educ age male paeduc white

      Source |       SS       df       MS              Number of obs =     694
-------------+------------------------------           F(  5,   688) =   38.74
       Model |  10869.4508     5  2173.89017           Prob > F      =  0.0000
    Residual |  38604.7369   688  56.1115361           R-squared     =  0.2197
-------------+------------------------------           Adj R-squared =  0.2140
       Total |  49474.1877   693  71.3913242           Root MSE      =  7.4908

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .9206479   .1203655     7.65   0.000     .6843201    1.156976
         age |   .1703887   .0255263     6.68   0.000     .1202699    .2205074
        male |   4.777683   .5729088     8.34   0.000     3.652824    5.902542
      paeduc |   .0180433    .085851     0.21   0.834    -.1505182    .1866047
       white |   .1643811   .9440889     0.17   0.862     -1.68926    2.018022
       _cons |  -4.994316   2.076236    -2.41   0.016    -9.070836   -.9177955
------------------------------------------------------------------------------

. * Part 6.  Sorry, unlike SPSS, no easy way to do pairwise in Stata.  If I was a
fanatic
. * about it, I could probably use the pwcorr and corr2data commands.
.
. * Part 7.  Mean substitution of MD (both IVs and DVs).  Seems questionable for
. * the DV. I'll use the impute command to create new vars
. * with the mean substituted for MD.
. impute income one, gen(incomex)
 39.88% (605) observations imputed
. impute educ one, gen(educx)
  0.46% (7) observations imputed
. impute age one, gen(agex)
  0.20% (3) observations imputed
. impute male one, gen(malex)
  0.00% (0) observations imputed
. impute paeduc one, gen(paeducx)
 29.53% (448) observations imputed
. impute white one, gen(whitex)
  0.00% (0) observations imputed
```

```
. reg  incomex educx agex malex paeducx whitex

      Source |       SS       df       MS              Number of obs =    1517
-------------+------------------------------           F(  5,  1511) =   43.02
       Model |  8121.88153     5  1624.37631           Prob > F      =  0.0000
    Residual |  57053.9143  1511  37.7590432           R-squared     =  0.1246
-------------+------------------------------           Adj R-squared =  0.1217
       Total |  65175.7958  1516  42.9919497           Root MSE      =  6.1448

------------------------------------------------------------------------------
     incomex |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       educx |   .5104091   .0583069     8.75   0.000     .3960381      .62478
        agex |   .0623961   .0095878     6.51   0.000     .0435892    .0812029
       malex |   3.133283   .3221433     9.73   0.000     2.501387    3.765178
     paeducx |    .008691    .050987     0.17   0.865    -.0913218    .1087038
      whitex |   .1019889   .4308004     0.24   0.813    -.7430412     .947019
       _cons |   5.738923     1.0351     5.54   0.000     3.708538    7.769308
------------------------------------------------------------------------------

. * Part 8.  Mean substitution, Father's education only, without and then with an MD
indicator.
. * The final regression will give us an idea of whether or not the MD in PAEDUC is
missing on a random basis.
. reg  income educ age male paeduc2 white

      Source |       SS       df       MS              Number of obs =     911
-------------+------------------------------           F(  5,   905) =   52.95
       Model |  14735.0402     5  2947.00803           Prob > F      =  0.0000
    Residual |  50371.0427   905  55.6586107           R-squared     =  0.2263
-------------+------------------------------           Adj R-squared =  0.2220
       Total |  65106.0829   910   71.545146           Root MSE      =  7.4605

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .9728974   .0973281    10.00   0.000     .7818824    1.163912
         age |   .1331499    .020681     6.44   0.000     .0925616    .1737382
        male |   5.195091   .4969931    10.45   0.000     4.219698    6.170484
     paeduc2 |  -.0333754   .0814587    -0.41   0.682    -.1932452    .1264944
       white |   .4738556   .6972264     0.68   0.497    -.8945131    1.842224
       _cons |  -4.340494   1.717974    -2.53   0.012    -7.712171   -.9688166
------------------------------------------------------------------------------
```

```
. reg  income educ age male paeduc2 white mdpaeduc

      Source |       SS       df       MS              Number of obs =     911
-------------+------------------------------           F(  6,   904) =   44.42
       Model |  14823.9809     6  2470.66348           Prob > F      =  0.0000
    Residual |   50282.102   904  55.6217943           R-squared     =  0.2277
-------------+------------------------------           Adj R-squared =  0.2226
       Total |  65106.0829   910   71.545146           Root MSE      =   7.458

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .9380886    .101115     9.28   0.000     .7396412    1.136536
         age |   .1352963   .0207437     6.52   0.000     .0945848    .1760077
        male |   5.218512   .4971738    10.50   0.000     4.242763    6.194261
     paeduc2 |  -.0267418   .0816005    -0.33   0.743    -.1868903    .1334067
       white |   .2642979   .7164261     0.37   0.712    -1.141754     1.67035
    mdpaeduc |  -.7894665   .6243181    -1.26   0.206    -2.014748     .435815
       _cons |  -3.673691   1.796537    -2.04   0.041    -7.199559   -.1478227
------------------------------------------------------------------------------
```

She is using different approaches for handling MD. The sample sizes differ, because with some techniques whole cases are deleted, while with others as many cases as possible are retained. The results are not all that different from model to model, except that the mean substitution approach differs a lot (perhaps because it is the most questionable choice).

f. [Optional] In part 7, why does the assistant make the comment that mean substitution on the DV seems questionable?

Many cases were MD because the question was “not applicable.” Perhaps these subjects had no income, or there were other reasons the question was not asked. You should understand the coding better before using mean substitution; it sounds like these cases should be dropped or perhaps even coded as 0.

g. In part 8, the assistant comments that “The final regression will give us an idea of whether or not the MD in PAEDUC is missing on a random basis.” How does the regression do that??? What does the coefficient for MDPAEDUC supposedly tell you? Would Allison approve or disapprove of what the assistant is doing here? Why?

According to Cohen and Cohen, the coefficient for the MDPAEDUC variable indicates whether or not the MD cases for father’s education are randomly missing. Since the coefficient is not significant, there doesn’t seem to be much problem (although that may just reflect the fact that PAEDUC’s effects are so trivial). Allison, however, cautions against this technique, on the grounds that it produces biased coefficient estimates. I might still be tempted to use it if the data were missing, say, because the respondent had no father, but it is not clear that that is the case here, i.e. the not applicables might be because there is no father, but some of the missing data is also due to Don’t Know responses.

h. [Optional] Given the nature of the missing data, which approach do you think is most appropriate in this case? Why? Why are the other approaches less desirable? Briefly describe what the main substantive conclusions are from your preferred model (e.g. which variables are important, what effect do the main variables have on income, etc.)

In the past (before I read Allison) I said I probably liked the last model the best (Mean substitution for Father's education only, without and then with an MD indicator). It doesn’t use the “not applicable” income cases, nor does it cause you to lose data because of PAEDUC.

Among other things, the model shows that race and Father's education do not significantly affect Income. Those who are better educated, older, and male make more than those who are not. I might still be tempted to use it if the data were missing, say, because the respondent had no father, but it is not clear that that is the case here, i.e. the not applicables might be because there is no father, but some of the missing data is also due to Don't Know responses.

Post-Allison, I lean more towards the model from part 5, listwise deletion:

```
. reg  income educ age male paeduc white

      Source |       SS       df       MS              Number of obs =     694
-------------+------------------------------           F(  5,   688) =   38.74
       Model |  10869.4508     5  2173.89017           Prob > F      =  0.0000
    Residual |  38604.7369   688  56.1115361           R-squared     =  0.2197
-------------+------------------------------           Adj R-squared =  0.2140
       Total |  49474.1877   693  71.3913242           Root MSE      =  7.4908

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .9206479   .1203655     7.65   0.000     .6843201    1.156976
         age |   .1703887   .0255263     6.68   0.000     .1202699    .2205074
        male |   4.777683   .5729088     8.34   0.000     3.652824    5.902542
      paeduc |   .0180433    .085851     0.21   0.834    -.1505182    .1866047
       white |   .1643811   .9440889     0.17   0.862     -1.68926    2.018022
       _cons |  -4.994316   2.076236    -2.41   0.016    -9.070836   -.9177955
------------------------------------------------------------------------------
```

Luckily, you get similar results either way. The same coefficients are significant, and the coefficients are pretty similar to each other. If you were writing up these results for a paper, you might note that a variety of approaches were tried and they all yielded similar results. If you've made a mistake with your preferred approach, it doesn't seem to be a very costly one.

i. [Optional] Do you have any other suggestions for deciding how to handle the MD? Present any additional analyses you think might be helpful. For example, you might examine whether men or women are more likely to have missing data on income.

It may be wise to simply drop PAEDUC, since it has no direct effect and is a major source of MD. If you do that using listwise deletion, you get 911 cases (up from 694 when paeduc is included) and you get the following results:

```
. * Other suggestions.  Drop paeduc completely!
. reg  income educ age male white

      Source |       SS       df       MS              Number of obs =     911
-------------+------------------------------           F(  4,   906) =   66.20
       Model |  14725.6966     4  3681.42416           Prob > F      =  0.0000
    Residual |  50380.3862   906  55.6074903           R-squared     =  0.2262
-------------+------------------------------           Adj R-squared =  0.2228
       Total |  65106.0829   910   71.545146           Root MSE      =   7.457

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .9601524    .092181    10.42   0.000     .7792393    1.141065
         age |   .1353336   .0199733     6.78   0.000     .0961343    .1745328
        male |   5.180144   .4954247    10.46   0.000      4.20783    6.152458
       white |   .4488951   .6942408     0.65   0.518     -.913612    1.811402
       _cons |  -4.602027   1.594254    -2.89   0.004    -7.730888   -1.473166
------------------------------------------------------------------------------
```

Note that these coefficients are not too much different from when PAEDUC was included, and the T values are all higher.

You may also want to examine more whether the MD in Income is random. Create a new variable coded 1 if Income is missing, 0 otherwise. Crosstab it with gender and race. If there is no association, that suggests data are missing randomly. If there is an association, it might indicate that, say, women are more likely to have missing data than men are. (If you do this, you find women are significantly more likely to have MD on income. Nonwhites are a little more likely to have MD, but, as the chi-square tests show, the difference is not significant. This might reflect their reduced likelihood that women and nonwhites will be employed.)

```
. * Try to id where the MD is.
. gen mdinc = missing(income)
. tabulate male mdinc, chi2 exact lrchi2 row

+-----------------+
| Key             |
|-----------------|
|   frequency     |
| row percentage  |
+-----------------+

 RECODE of |
       sex |
(RESPONDEN |         mdinc
   TS SEX) |         0          1 |     Total
-----------+----------------------+----------
         0 |       462        419 |       881
           |     52.44      47.56 |    100.00
-----------+----------------------+----------
         1 |       450        186 |       636
           |     70.75      29.25 |    100.00
-----------+----------------------+----------
     Total |       912        605 |     1,517
           |     60.12      39.88 |    100.00

          Pearson chi2(1) =  51.6713   Pr = 0.000
 likelihood-ratio chi2(1) =  52.5111   Pr = 0.000
           Fisher's exact =                 0.000
   1-sided Fisher's exact =                 0.000
```

```
. tabulate white mdinc, chi2 exact lrchi2 row

+-----------------+
| Key             |
|-----------------|
|   frequency     |
| row percentage  |
+-----------------+

 RECODE of |
race (RACE |
        OF |
RESPONDENT |         mdinc
         ) |         0          1 |     Total
-----------+----------------------+----------
         0 |       140        113 |       253 
           |     55.34      44.66 |    100.00 
-----------+----------------------+----------
         1 |       772        492 |     1,264 
           |     61.08      38.92 |    100.00 
-----------+----------------------+----------
     Total |       912        605 |     1,517 
           |     60.12      39.88 |    100.00 

          Pearson chi2(1) =   2.8968   Pr = 0.089
 likelihood-ratio chi2(1) =   2.8700   Pr = 0.090
           Fisher's exact =                 0.092
   1-sided Fisher's exact =                 0.052
```