Soc 63993, Homework #2: 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
```

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.]

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.

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 (ologit) x2 = v1 v2 v3 v4 …
```

  where mlogit and ologit 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.

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`