

## Soc 63993, Advanced Social Statistics II Homework No. 2 Multicollinearity/Missing Data

### I. *Multicollinearity*

[The following problem is adapted from Greene, *Econometric Analysis*, Fourth Edition.] The data in *longley.dta* (available at <http://www.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

A. **Diagnosis.** 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?

In addition, suppose the analysis ended a year sooner, i.e. only covered the period 1947-1961. How would the results have differed? Do these differences also imply a multicollinearity problem? If so, explain why. To answer this, you will need to run the regression

```
. regress employ price gnp armed year if year < 1962
```

B. Possible solutions to the problem. Run the following commands:

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

Explain what the variable `gnpadj` means (hint: `adj` is short for adjusted), and the rationale for doing this as a way of addressing the multicollinearity problem. (Try using the `list` command again if you are not sure what the above `gen` command has done.) After having done this, check the VIFs to see if there is still a multicollinearity problem. If so, try dropping one variable from the model and see if that makes a difference. Explain your choice of variables to drop and the pros and cons of your decision.

## II. Missing data

For this problem, you need to copy and run *missing.sps* and *missing.sav* from my web page. 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 SPSS program, but she needs your help to understand exactly what was done and how to interpret her results.

### Missing.sps

```
* Missing.sav is an extract from the 1991 General Social Survey.
get file = 'Missing.sav'.

* Part 1. Do frequencies on the original vars. Look at MD
* patterns, problems with coding.
FREQUENCIES VARS = RINCOME EDUC AGE SEX RACE PAEDUC/
              STATISTICS = DEFAULT.

* 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.
RECODE RINCOME (1=0.5) (2=2.0) (3=3.0) (4=4.5) (5=5.5)
              (6=6.5) (7=7.5) (8=9) (9=12.5) (10=17.5)
              (11=22.5) (12=25.0)
              (0 = 97) (98,99=99) (13=98) into INCOME.
MISSING VALUES INCOME (99 97 98).
VALUE LABELS INCOME
              97 "Not Applicable"
              98 "Refused"
              99 "NA or DK".
FREQUENCIES VARIABLES = INCOME/ STATISTICS = DEFAULT.

* 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) into WHITE/
              SEX (1 = 1) (ELSE = 0) INTO MALE.
```

```

FREQUENCIES VARIABLES = WHITE MALE/ STATISTICS = DEFAULT.

* Part 4. Create a modified PAEDUC2 that I can use later. Create
* an MD indicator.
DO IF (MISSING (PAEDUC)).
  COMPUTE MDPAEDUC=1.
  COMPUTE PAEDUC2=10.88.
ELSE.
  COMPUTE MDPAEDUC=0.
  COMPUTE PAEDUC2 = PAEDUC.
END IF.
FREQUENCIES VARIABLES = PAEDUC2 MDPAEDUC/ STATISTICS = DEFAULT.

* Part 5. Listwise deletion of MD.
REGRESSION VARS INCOME EDUC AGE MALE PAEDUC WHITE
  /MISSING LISTWISE
  /STATISTICS DEF CI
  /DESCRIPTIVES
  /DEP INCOME
  /ENTER EDUC AGE MALE PAEDUC WHITE .

* Part 6. Pairwise deletion of MD.
REGRESSION VARS INCOME EDUC AGE MALE PAEDUC WHITE
  /MISSING PAIRWISE
  /STATISTICS DEF CI
  /DESCRIPTIVES
  /DEP INCOME
  /ENTER EDUC AGE MALE PAEDUC WHITE .

* Part 7. Mean substitution of MD (both IVs and DVs). Seems questionable for the DV.
REGRESSION VARS INCOME EDUC AGE MALE PAEDUC WHITE
  /MISSING MEANSUBSTITUTION
  /STATISTICS DEF CI
  /DESCRIPTIVES
  /DEP INCOME
  /ENTER EDUC AGE MALE PAEDUC WHITE .

* 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.
REGRESSION VARS INCOME EDUC AGE MALE PAEDUC2 MDPAEDUC WHITE
  /MISSING LISTWISE
  /STATISTICS DEF CI
  /DESCRIPTIVES
  /DEP INCOME
  /ENTER EDUC AGE MALE PAEDUC2 WHITE
  /ENTER MDPAEDUC.

```

- 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. 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. In parts 5-8, why does her assistant run the regressions 4 different ways? 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. 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. 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. 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.

### III. [Optional] Using Stata for Missing Data

Replicate as much of Missing.sps as you can using Stata. You can download the file missing.dta for this purpose. Some things you need to know are:

\* 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.

\* 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, so you won't be able to replicate part 6 of Missing.sps, unless you want to go to a lot of extra trouble.

\* 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 a value of `.` for all 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.

<code>tab1</code>	<code>generate</code>	<code>if</code>	<code>summarize</code>
<code>replace</code>	<code>recode</code>	<code>impute</code>	