Cleaning Data

Advanced Econometrics and Applications

Antonio Jurlina

3/3/2021

Why talk about data cleaning?

Most econometrics classes and textbooks present students with nice, picture-perfect data sets for applied problem set (look at the data sets provided by our textbook). These data sets are "perfect":

- No missing data
- No values are the product of an obvious typographical error
- Data is already transformed, e.g. log wage
- All the data is contained in one neat file, and so on.

In most cases, your research data is not "clean":

- It will come in several files covering difference questionnaire modules across different years
- Monetary values will have been recorded in nominal values
- Some people will have refused to answer some questions
- Other will have "trolled the enumerators" with unrealistic answers, and
- Whoever entered the data will have made typos.

The list of possible issues is almost endless.

How do you clean your data?

Cleaning your data, usually, involves the following steps:

1. ALWAYS(!!!) save a "raw" copy of the un-adjusted data
2. Merge data files together (if applicable)
3. Inspect your data
   - Graph histograms and scatter plots
   - Look at summary statistics and correlations
   - Look for obvious irregularities
4. Drop/adjust some observations due to:
   - Missing values
   - Outliers
   - Typos, etc.
5. Transform variables
6. Generate new variables

Protocol. In this part of the lab we will focus on steps 3 and 4, inspecting you data set and making data adjustments.

★ **EXERCISE** ★

(a) Load the Stata data file, labeled *wageV1.dta*, into memory and generate a variable named 'lwage' which equals the $log(wage)$.

```
// Clear previously stored data and set global options
cls
clear all
cd "/Users/labteam/Google Drive/Spring 2021/ECO 531/data"

use "wageV1.dta" // load data

// Store results in a log file (diary)
cd "/Users/labteam/Google Drive/Spring 2021/ECO 531/logs"
log using "lab_03_log.txt", replace text

// generate new variable (if it isn't already there)
capture confirm variable lwage, exact
if _rc {
    generate lwage = log(wage)
}
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	526	5.896103	3.693086	.53	24.98
educ	451	12.62749	2.688006	0	18
exper	475	17.06737	13.60565	1	50
sex	469	1.520256	.500123	1	2
married	526	1.608365	.4885804	1	2
region	496	2.425403	1.039852	1	4

Figure 1: Descriptive statistics

(b) Use **browse** to examine the data – look for any patterns or flags that suggest something is amiss.

```
browse
```

(c) Generate descriptive statistics and correlations for the variables *wage*, *educ*, *exper*, *sex*, *married*, and *region*. Again, look for any patterns or flags that suggest something is amiss.

```
summarize wage educ exper sex married region

correlate wage educ exper sex married region
```

(d) Plot histograms for these variables to get a visual "feel" for the data.

```
histogram wage, name(wage)
histogram educ, name(educ)
histogram exper, name(exper)
histogram sex, name(sex)
```

	wage	educ	exper	sex	married	region
wage	1.0000					
educ	0.4188	1.0000				
exper	0.0818	-0.3421	1.0000			
sex	0.3506	0.0640	0.0060	1.0000		
married	0.2316	0.0431	0.3043	0.1973	1.0000	
region	0.0513	-0.0022	0.0132	-0.0016	0.0662	1.0000

Figure 2: Correlation matrix

```
histogram married, name(married)
histogram region, name(region)
```

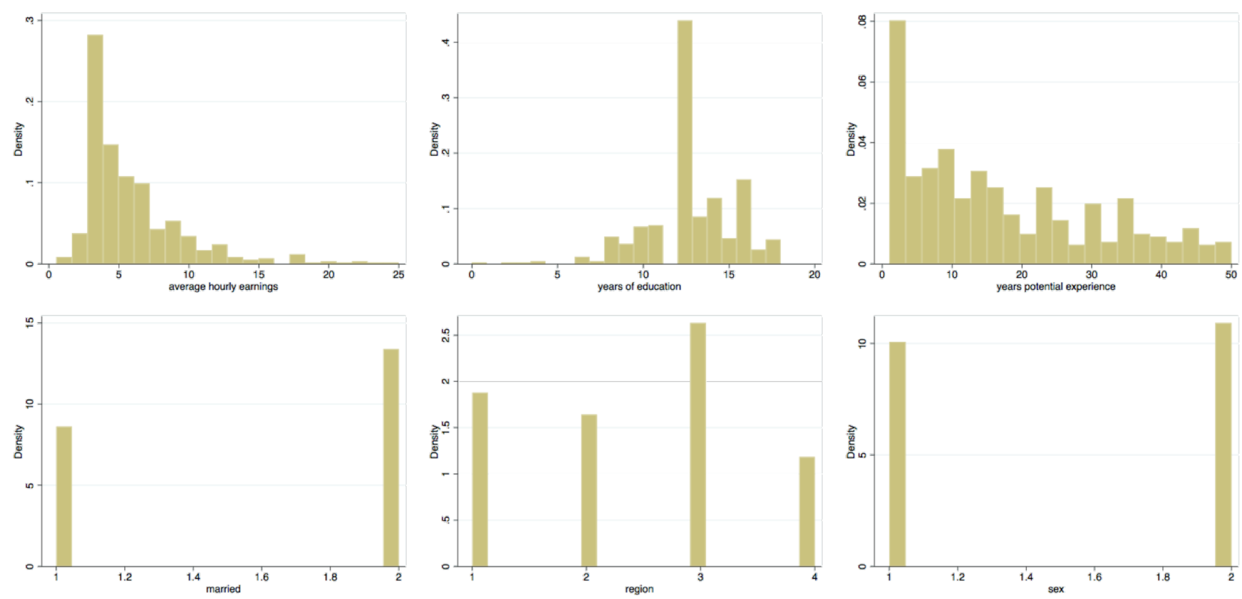

Figure 3: Histograms

(e) Use **inspect** to explore each variable to see whether there are obvious irregularities: missing values, outliers, censoring, truncation, etc.

```
inspect wage
inspect educ
inspect exper
inspect sex
inspect married
inspect region
```

(f) Graph scatter plots your dependent variables against each right-hand side variable to get a visual sense of what is going on as well as detect outliers and leverage points. e.g., **graph twoway (scatter wage educ) (lfit wage educ)**.

```
graph twoway (scatter wage educ) (lfit wage educ), name(wage_v_educ)
graph twoway (scatter wage exper) (lfit wage exper), name(wage_v_exper)
```

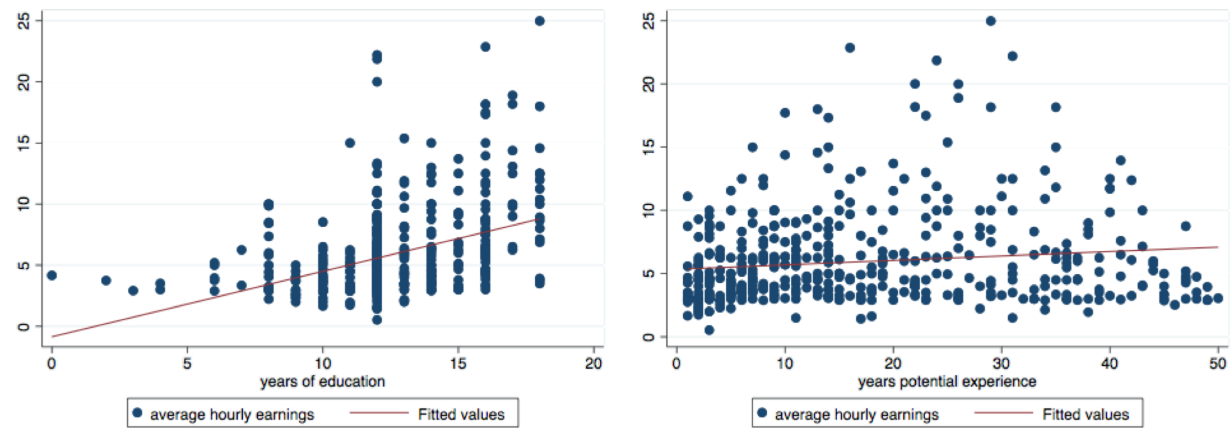

Figure 4: Scatter plots

Missing data in Stata

When working with missing data, you need to consider why that data is missing. In survey data, missing values may mean that the surveyor did not ask the question, that the respondent did not answer the question, or that the data are truly missing. (Some datasets have these three cases coded differently; others lump them together. Check your metadata/codebook to make sure you know what you are working with!) For numeric data, keep in mind that missing data are not the same as a value of zero. (This may seem obvious, but I have had many students nonchalantly say “oh, so we can just replace those with zeros. . . ” Nope.) Consider this in the context of gas mileage. $MPG = 0$ is very different from $MPG =$ “I’m not sure”.

Different statistical software code missing data differently. In Stata, if your variable is numeric and you are missing data, you will see ‘.’ in your dataset. If you are working with string variables, the data will appear as ’ ’ (a blank space).

Missing data values will affect how Stata handles your data.

- **summarize** - uses only non-missing values
- **tabulate** - missing values excluded by default; use missing option within tab to include missing values.
- **correlate** - calculated on pairs with non-missing data by default (pairwise deletion of missing data).
- **regress** - if an observation is missing data for a variable in the regression model, that observation is excluded from the regression (listwise deletion of missing data).

Explore patterns in the missing data

The **misstable** command allows a researcher to explore missing observations in the data – specifically, whether there are patterns across the missing data. This can be very useful as we need to make decisions about what to do about (if anything) the missing data.

★ EXERCISE ★

(a) Use **misstable patterns** to explore the patterns of missing data. Does the missing data appear to be random? Are some variables more commonly missing?

```
misstable patterns
```

65% of the data has no missing pattern to it (Figure 5). Variables *educ* and *exper* are often missing without any apparent connection to other variables, and together they also constitute the most missing data. These missing variables seem to have a random pattern to them. Variables

Missing-value patterns
(1 means complete)

Percent	Pattern 1	2	3	4	5	6	7	8	9	10
65%	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	0
8	1	1	1	1	1	1	1	0	0	1
6	1	1	1	1	1	1	0	1	1	1
5	1	1	0	0	0	0	1	1	1	1
2	1	1	1	1	1	1	0	1	1	0
2	1	1	1	1	1	1	1	0	0	0
1	0	0	1	1	1	1	1	1	1	1
<1	1	1	1	1	1	1	0	0	0	0
<1	1	1	0	0	0	0	1	1	1	0
<1	0	0	1	1	1	1	0	1	1	1
<1	1	1	0	0	0	0	1	0	0	0
<1	1	1	1	1	1	1	0	0	0	1
100%										

Variables are (1) **nonwhite** (2) **race**
(3) **northeast** (4) **region**
(5) **south** (6) **west** (7) **exper**
(8) **male** (9) **sex** (10) **educ**

Figure 5: Misstable patterns

race and *nonwhite*, as evident from Figures 5 and 6. are missing together. The same can be said for variables *west*, *region*, *south*, and *northeast*, as well as *sex* and *male*. This is only a pattern given that *nonwhite* is a dummy variable for *race*; *west*, *northeast*, and *south* are dummy variables for *region*; and *male* is one for *sex*. Since they are directly related, the missing pattern is not random, nor surprising.

(b) Type **help misstable** to explore more options. Try some and see what happens.

```
1.  race(7) <-> nonwhite(7)
2.  west(30) <-> south(30) <-> region(30) <-> northeast(30)
3.  exper(51)
4.  sex(57) <-> male(57)
5.  educ(75)
```

Figure 6: Misstable nested

```
//help misstable

misstable summarize
misstable tree
misstable nested
```

				Obs<.		
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
educ	75		451	17	0	18
exper	51		475	50	1	50
male	57		469	2	0	1
sex	57		469	2	1	2
nonwhite	7		519	2	0	1
race	7		519	2	1	2
northeast	30		496	2	0	1
south	30		496	2	0	1
west	30		496	2	0	1
region	30		496	4	1	4

Figure 7: Misstable summarize

There are four main approaches to dealing with missing data:

1. Do nothing – drop incomplete observations
   - Advantage: simple; works so long as data is missing at random
   - Disadvantage: may throw out lots of good data; will introduce bias if missing data not random
2. Replace missing observations with unconditional mean or mode * Advantage: simple; mean replacement

won't affect OLS slope estimates; works so long as data is missing at random * Disadvantage: will introduce bias if missing data not random; reduce variability in data; wrong standard errors

3. Replace missing observations with conditional mean
    - Advantage: uses all available data; can work, even if data not missing at random (not always!)
    - Disadvantage: overestimates model fit; wrong standard errors
4. Use statistical imputation methods (we won't do in this class)

Be aware:

- Choose carefully and thoughtfully
- No matter the choice, it has consequences on the results!

★ EXERCISE ★

(a) Let's start by creating three new variables, labeled 'educ_mean', 'educ_mode' and 'educ_ols' which replicate the data contained in *educ*. For example, generate $educ_mean = educ$ would do this for the first variable.

```
generate educ_mean = educ
generate educ_mode = educ
generate educ_ols = educ
```

(b) Now, let's replace the missing observations in *educ_mean* and *educ_mode* with the unconditional mean and mode, respectively. Hint: this information is contained in the **summarize** command and you will need to use the **replace** command.

```
summarize(educ), detail
replace educ_mean = r(mean) if educ == .
replace educ_mode = r(p50) if educ == .
```

(c) Now let's get more adventurous. Replace the missing observations in *educ_ols* with the conditional mean (OLS prediction). Hint: you will need to use the **regress** and **predict** commands.

```
regress educ exper wage male married race region numdep
predict educhat
summarize educhat
replace educ_ols = r(mean) if educ == .
drop educhat
```

(d) Suppose we are interested in estimating the following relationship,

$$ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u.$$

Run separate regressions using our four choices: do nothing (*educ*), mean replacement (*educ_mean*), mode replacement (*educ_mode*), and conditional mean replacement (*educ_ols*). Given this data set, and its unique characteristics, did your choice have much impact on the results? Which approach works "best" will depend on the characteristics of your data.

```
regress lwage educ c.exper##c.exper
regress lwage educ_mean c.exper##c.exper
regress lwage educ_mode c.exper##c.exper
regress lwage educ_ols c.exper##c.exper
```

Our choices did not impact the results in significant ways. Estimates on the education coefficient are fairly robust to our interventions, and R^2 and F statistics hold similar values across iterations.

```
. regress lwage educ c.exper##c.exper

      Source |       SS           df       MS      Number of obs   =       416
-------------+----------------------------------   F(3, 412)       =     57.42
       Model |  36.0370869         3  12.0123623   Prob > F        =    0.0000
    Residual |  86.1856949       412   .20918858   R-squared       =    0.2948
-------------+----------------------------------   Adj R-squared   =    0.2897
       Total |  122.222782       415  .294512727   Root MSE        =    .45737

---------------------------------------------------------------------------------
          lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
           educ |    .092606   .0089147    10.39   0.000      .075082    .1101299
          exper |   .0425937    .006023     7.07   0.000     .0307541    .0544333
                |
c.exper#c.exper |  -.0007405   .0001331    -5.57   0.000    -.0010021    -.000479
                |
          _cons |   .0949308   .1254162     0.76   0.450    -.1516047    .3414662
---------------------------------------------------------------------------------
```

```
. regress lwage educ_mean c.exper##c.exper

      Source |       SS           df       MS      Number of obs   =       475
-------------+----------------------------------   F(3, 471)       =     55.00
       Model |  35.6550945         3  11.8850315   Prob > F        =    0.0000
    Residual |   101.77465       471  .216082059   R-squared       =    0.2594
-------------+----------------------------------   Adj R-squared   =    0.2547
       Total |  137.429744       474  .289936169   Root MSE        =    .46485

---------------------------------------------------------------------------------
          lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
      educ_mean |   .0915229   .0090223    10.14   0.000      .073794    .1092519
          exper |   .0396841   .0057144     6.94   0.000     .0284552    .0509131
                |
c.exper#c.exper |  -.0006988   .0001272    -5.49   0.000    -.0009487   -.0004488
                |
          _cons |   .1272511   .1252495     1.02   0.310    -.1188657     .373368
---------------------------------------------------------------------------------
```

```
. regress lwage educ_mode c.exper##c.exper

      Source |       SS           df       MS      Number of obs   =       475
-------------+----------------------------------   F(3, 471)       =     55.77
       Model |  36.0225868         3  12.0075289   Prob > F        =    0.0000
    Residual |  101.407157       471   .21530182   R-squared       =    0.2621
-------------+----------------------------------   Adj R-squared   =    0.2574
       Total |  137.429744       474  .289936169   Root MSE        =    .46401

---------------------------------------------------------------------------------
          lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
      educ_mode |   .0917605   .0089557    10.25   0.000     .0741625    .1093585
          exper |   .0398703   .0057023     6.99   0.000     .0286652    .0510754
                |
c.exper#c.exper |  -.0007041   .0001269    -5.55   0.000    -.0009534   -.0004549
                |
          _cons |   .1307785   .1238357     1.06   0.291    -.1125603    .3741174
---------------------------------------------------------------------------------
```

```
. regress lwage educ_ols c.exper##c.exper

      Source |       SS           df       MS      Number of obs   =       475
-------------+----------------------------------   F(3, 471)       =     54.84
       Model |  35.5749642         3  11.8583214   Prob > F        =    0.0000
    Residual |   101.85478       471  .216252187   R-squared       =    0.2589
-------------+----------------------------------   Adj R-squared   =    0.2541
       Total |  137.429744       474  .289936169   Root MSE        =    .46503

---------------------------------------------------------------------------------
          lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
       educ_ols |   .0913745   .0090275    10.12   0.000     .0736353    .1091137
          exper |   .0396637   .0057169     6.94   0.000     .0284299    .0508975
                |
c.exper#c.exper |  -.0006983   .0001273    -5.49   0.000    -.0009483   -.0004482
                |
          _cons |   .1282186   .1254036     1.02   0.307    -.1182011    .3746383
---------------------------------------------------------------------------------
```

Figure 8: Regressions