

# Problem Set 4

This problem set will revisit some of the material covered in Handouts 3 and 4. You will be required to work with a ‘raw’ dataset, downloaded from an online repository. For this reason, you should take care to check how the data is coded.

You will be using a version of the US Current Population Survey (CPS) called the [Merged Outgoing Rotation Group](#) (MORG). This data is compiled by the National Bureau of Economic Research (NBER) and has been used in many famous studies of the US economy. The CPS has a rather unique rotating panel design: “The monthly CPS is a rotating panel design; households are interviewed for four consecutive months, are not in the sample for the next eight months, and then are interviewed for four more consecutive months.” ([source: IPUMS](#)). The NBER’s MORG keeps only the outgoing rotation group’s observations.

The MORG .dta files can be found at: <https://data.nber.org/morg/annual/>.

## Preamble

Create a do-file for this problem set and include a preamble that sets the directory and opens the data **directly from the NBER website**. Of course, this requires a good internet connection. For example,

```
clear
//or, to remove all stored values (including macros, matrices, scalars, etc.)
*clear all

* Replace $rootdir with the relevant path to on your local harddrive.
cd "$rootdir/problem-sets/ps-4"

cap log close
log using problem-set-4-log.txt, replace

use "https://data.nber.org/morg/annual/morg19.dta", clear
```

You can, of course, download the data and open it locally on your computer.

## Questions

1. Create a new variable **exper** equal to age minus (years of education + 6). This is referred to as potential years of experience. Check how each variable defines missing values before proceeding. You will need to create a years of education variable for this. Here is the suggested code:

```
tab grade92, m
gen eduyrs = .
  replace eduyrs = .3 if grade92==31
  replace eduyrs = 3.2 if grade92==32
  replace eduyrs = 7.2 if grade92==33
  replace eduyrs = 7.2 if grade92==34
  replace eduyrs = 9   if grade92==35
  replace eduyrs = 10  if grade92==36
  replace eduyrs = 11  if grade92==37
  replace eduyrs = 12  if grade92==38
  replace eduyrs = 12  if grade92==39
  replace eduyrs = 13  if grade92==40
  replace eduyrs = 14  if grade92==41
  replace eduyrs = 14  if grade92==42
  replace eduyrs = 16  if grade92==43
  replace eduyrs = 18  if grade92==44
  replace eduyrs = 18  if grade92==45
  replace eduyrs = 18  if grade92==46
  lab var eduyrs "completed education"
tab grade92, sum(eduyrs)
```

2. Keep only those between the ages of 18 and 54. Check the distribution of ‘exper’ and replace any negative values to 0.

3. Create a categorical variable that takes on 4 values: 1 “less than High School”; 2 “High School Diploma”; 3 “some Higher Education”; 4 “Bachelors”; 5 “Postgraduate”. This variable should be based on the **grade92** variable. You can find the value labels for this variable in this document: <https://data.nber.org/morg/docs/cpsx.pdf>. I suggest using the **recode** command, which allows you to create value labels while assigning values. Check the distribution of **exper** by education category.

4. Create the variable **lnwage** equal to the (natural) log of weekly earnings. Create a figure that shows the predicted *linear* fit of **lnwage** against **exper**, by **educat**. Try to place all 5 fitted lines in the same graph.

5. Estimate a linear regression model that allows the slope coefficient on `exper` and constant term to vary by education category (`educat`). Let the base (excluded) education category be 2 “High School diploma”.

$$\ln(Wage_i) = \alpha + \sum_{j \neq 2} \psi_j \mathbf{1}\{Educate_i = j\} + \beta Exper_i + \sum_{j \neq 2} \gamma_j Exper_i \times \mathbf{1}\{Educate_i = j\} + v_i$$

6. Show that after 13 years of experience, those with some Higher Education (but no Bachelors), out earn those with just a high school diploma. You can assume that there is a 2 year difference between the experience (education).

7. Use the post-estimation `test` command to test the null hypothesis:  $H_0 : 15\beta = 13(\beta + \gamma_3) + \psi_3$ .

8. Estimate a transformed version of the above model allowing you to test the above hypothesis using the coefficient from a single regressor. That is, the resulting test should be a simple t-test of  $H_0 : \phi = 0$ , where  $\phi$  is the coefficient on the interaction of `exper` and a dummy variable for `educat=3`. This will be easier to do if you estimate the model using only the relevant sample: those with High School diplomas and some Higher Education. I suggest avoiding the use of factor notation to create the dummy variables and interaction terms for this exercise. For example, the following should replicate the relevant coefficients from Q5.

```
gen hasHE = educat==3 if inlist(educat,2,3)
gen hasHEexp = hasHE*exper

reg lnwage exper hasHE hasHEexp
```

9. Verify that the F-statistic from Q7 is the square of the above T-statistic.

10. Use the restricted OLS approach to replicate the F-statistic and p-value from Q7.

11. Use the restricted OLS approach to test the following hypothesis corresponding to the model in Q5:

$$H_0 : \gamma_j = 0 \quad \text{for } j = 1, 3, 4, 5$$

Compute the F-statistic and p-value. Verify your result using the post-estimation `test` command.

12. Compute the relevant Chi-squared distributed test statistic and corresponding p-value for the above test, assuming  $n$  is large (enough).

13. Using the data from Problem Set 2, estimate the simple linear regression model using OLS,

$$\ln(Wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Female_i + \varepsilon_i$$

**14.** Estimate the model using Maximum Likelihood. Take a look at <https://www.stata.com/manuals13/rmlexp.pdf>, the documentation for the `mlexp` command. It has a discussion on estimating the CLRM using ML.<sup>1</sup>

**15.** Estimate the model using Method of Moments. You can use the `gmm` command in Stata. Hint: the regressors will be their own instruments and use the `onestep` option.<sup>2</sup>

## Postamble

```
log close
```

---

<sup>1</sup>You can also look at the following resource for a more flexible approach to ML estimation in Stata: <https://www.stata.com/features/overview/maximum-likelihood-estimation/>

<sup>2</sup>Here is a resource on GMM in Stata: <https://www.stata.com/features/overview/generalized-method-of-moments/>