Labor Economics with STATA

Liyousew G. Borga

December 2, 2015

Estimating the Human Capital Model Using Artificial Data

Outline

1. The Human Capital Model

2. Generating Synthetic Data

3. Estimation and Diagnostics

4. Wage Gap Decomposition

The Mincer Wage Equation

A complete wage equation model would include the following human capital variables

$$log(wages_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + \cdots + \mu_i$$

where

- the term μ_i contains factors such as ability, quality of education, family background and other factors influencing a person's wage
- $exper = Age - Education - 6$

The Mincer Wage Equation

A complete wage equation model would include the following human capital variables

$$log(wages_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + \cdots + \mu_i$$

where

- the term μ_i contains factors such as ability, quality of education, family background and other factors influencing a person's wage
- $exper = Age - Education - 6$
- For some specific purposes, you may also include gender and union status
- You may think of the relationship between wages and their determinants, including institutions and industrial characteristics, as the *wage structure*

Creating Artificial Data

- Generate right hand side variables
- variables should have plausible (in line with empirical evidence) range of values, distributions and should have reasonable V-C matrix (reflecting likely correlations of RHS variables),
- set reasonable coefficients $\beta_0, \beta_1, \beta_2, \beta_3$,
- generate stochastic term e of plausible *iid*, and generate left-hand side variable
- Generate log of earnings
- Estimate underlying model by OLS using underlying functional form

Random Data Generation

Stata's random-number generation functions, such as "runiform()" and "rnormal()", are deterministic algorithms that produce numbers that can pass for random

- Stata's random-number functions are formally called pseudorandom-number functions.
- The sequences these functions produce are determined by the seed, which is just a number and which is set to *123456789* every time Stata is launched.
- This means that runiform() produces the same sequence each time you start Stata.
- To obtain different pseudo-random sequences from the pseudo-random number functions, you must specify different seeds using the "set seed" command.
- It does not really matter how you set the seed, as long as there is no obvious pattern in the seeds that you set and as long as you do not set the seed too often during a session.
- The drawnorm command provides an alternative way to generate multiple normal variables, and optionally to specify the correlations between them

Stata Codes for Random Data Generation

- Generating matrix of means and covariance of RHS variables: *edu, exp, error*

```
clear
mat in m=(12,20,0)
mat in c=(5,-.6, 0 \ -.6,119,0 \ 0,0,.1)

matrix list m /* displays matrix of means */
matrix list c /* displays covariance matrix */

set seed 12345 /*Specify initial value of random-number seed */
```

- Drawing a sample of 2300 observations from a multivariate normal distribution with desired means and covariance matrix

```
drawnorm edu exp e, n(2300) means(m) cov(c)
```

- Generate log of earnings

```
gen logY = 7.6 + edu*.07 + exp*.012 - exp2*.0005 + e
drop if exp<0 | exp>40 /*observations with extreme values of exp */
correlate edu exp e, cov m
/* Compare means and covariance matrix for generated data and parameters
     you required */
```

Estimation

```
eststo: reg logY edu exp
eststo: reg logY edu exp exp2
```

Table : Regression table

	(1) logY	(2) logY
edu	0.0736***	0.0722***
	(24.68)	(24.91)
exp	-0.00751***	0.0111***
	(-11.93)	(6.54)
exp2		-0.000473***
		(-11.77)
Constant	7.686***	7.575***
	(197.46)	(194.37)
Observations	2300	2300
R^2	.25	.29

t statistics in parentheses

- If relevant variables are omitted from the model, the common variance they share with included variables may be wrongly attributed to those variables, and the error term is inflated

Diagnostics

Model Specification

- Check for error in model specification

```
linktest
ovtest
```

- The link test reveals no problems with our specification
- No error in model specification from the "ovtest"

Multicollinearity

- We can ask STATA to compute the Variance Inflation Factor, $VIF = (1 - R_k^2)^{-1}$, which measures the degree to which the variance has been inflated because regressor k is not orthogonal to the other regressors

```
vif
collin edu exp exp2, corr
```

Diagnostics

Linearity

```
gen Y=exp(logY)

graph matrix Y edu exp exp2
acprplot edu, lowess
acprplot exp, lowess // exp^2 is collinear with exp
acprplot exp2, lowess

graph matrix logY edu exp exp2
acprplot edu, lowess
acprplot exp, lowess // exp^2 is collinear with exp
acprplot exp2, lowess
```

Diagnostics

Distribution

```
graph box Y, saving(box_Y, replace)
graph box logY, saving(box_logY, replace)

graph combine box_Y.gph box_logY.gph, rows(1)

// OR

kdensity Y, normal saving(Y, replace)
kdensity logY, normal saving(logY, replace)

graph combine Y.gph logY.gph, rows(1)
```

Diagnostics

Unusual and influential data

```
gen id=_n
scatter logY edu, mlabel(id)
scatter logY exp, mlabel(id)
scatter logY exp2, mlabel(id)

reg logY edu exp exp2

lvr2plot, mlabel(id)

dfbeta
disp 2/sqrt(2151)
scatter _dfbeta_1 _dfbeta_2 _dfbeta_3 id, ylabel(-1(.5)3) yline(.04 -.04)
    mlabel(id id id)
```

Heteroskedasticity

- One of the main assumptions for OLS regression is the homogeneity of variance of the residuals
- If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values

```
reg logY edu exp exp2
predict res, res
gen res2=res^2
predict logY_h
// Plot the residuals versus fitted (predicted) values.
rvfplot, yline(0)
estat imtest
estat hettest
```

- Both tests could not reject the null hypothesis $H_0 : Constant\ \ Variance$

Diagnostics

Heteroskedasticity

Cases when heteroskedasticity is an issue:

- Heteroskedastic error term: variance is a function of edu

```
gen e_a=sqrt(edu)*e
graph twoway scatter e_a edu, yline(0) title("Heter=f(edu)") saving(
    graph_e_a_edu, replace)

gen logY_a=7.6+ edu*.07 + exp*.012- exp2*.0005 + e_a
reg logY_a edu exp exp2
predict logY_ah
predict res_a, res
gen res_a2=res_a^2

graph twoway scatter res_a logY_ah, yline(0)  title("Heter=f(edu)")
    saving(graph_edu_heter, replace)

estat hettest
estat hettest edu
reg res_a2 edu, noconstant
```

Diagnostics

Heteroskedasticity

Cases when heteroskedasticity is an issue:

- Heteroskedastic error term: variance is a function of external variable

```
gen x=runiform()
gen e_b=e*(x+.01) /*Heteroskedastic error: var =f(external variable x)*/
graph twoway scatter e_b x, yline(0) title("Heter=f(x)")

gen logY_b=7.6+ edu*.07 + exp*.012- exp2*.0005 + e_b
reg logY_b edu exp exp2
predict logY_bh
predict res_b, res
gen res_b2=res_b^2
graph twoway scatter res_b logY_bh, yline(0)  title("Heter=f(x)")

estat hettest
estat hettest, rhs
estat hettest x
```

- The Stata rreg command performs a robust regression using iteratively re-weighted least squares (assigns a weight to each observation with higher weights given to better behaved observations)

Errors in variables

Measurement Error

```
gen error=rnormal() /* Measurement error*/
gen logYX=logY+.2*error /* logY with error */
dotplot logY logYX, ny(25) saving(logY_logYX, replace)
reg logY edu exp exp2
reg logYX edu exp exp2
```

Stochastic Error

```
gen eduX=edu+2*error  /* Education years with error */
dotplot edu eduX, ny(25) saving(edu_eduX, replace)
reg logY edu exp exp2
reg logY eduX exp exp2
```

Systematic Error

```
gen eduQ=.8*edu  /* Education years with error */
dotplot edu eduQ, ny(25) saving(edu_eduQ, replace)
reg logY edu exp exp2
reg logY eduQ exp exp2
```

Errors in variables

Table : Regression with Errors in Measurement

	(1) logY	(2) logYX	(3) logY	(4) logY
edu	0.0722*** (0.00290)	0.0723*** (0.00339)		
exp	0.0111*** (0.00170)	0.0124*** (0.00198)	0.0111*** (0.00180)	0.0111*** (0.00170)
exp2	-0.000473*** (0.0000402)	-0.000485*** (0.0000470)	-0.000487*** (0.0000427)	-0.000473*** (0.0000402)
eduX			0.0392*** (0.00232)	
eduQ				0.0903*** (0.00363)
Constant	7.575*** (0.0390)	7.553*** (0.0455)	7.980*** (0.0331)	7.575*** (0.0390)
Observations	2300	2300	2300	2300
Adjusted R^2	0.293	0.228	0.202	0.293

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Decomposition

Oaxaca (1973)

- Question: how much of the wage gap can be "explained" by observable differences in human capital (education and labour market experience, occupational choices, etc)?
- Estimate OLS regressions of (log) wages on covariates/charactistics
- Use the estimates to construct a *counterfactual* wage such as "what would be the average wage of women (blacks) if they had the same characteristics as men (whites)?"
- This forms the basis of the decomposition

Decomposition

We want to decompose the difference in the mean of an outcome variable Y between two groups A and B;

- Postulate linear model for Y, with conditionally independent errors $(E(\nu|X) = 0)$

  $$Y_{gi} = \beta_{g0} + \sum_{k=1}^{K} X_{ik}\beta_{gk} + \nu_{gi}$$

  where g=A,B

- To get the Oaxaca decomposition start with (in simplified notation):

  $$\Delta\bar{Y} = \bar{Y}_A - \bar{Y}_B$$
  $$\Delta\bar{Y} = \bar{X}_A'\hat{\beta}_A - \bar{X}_B'\hat{\beta}_B$$

- This expression can, in turn, be written as the sum of the following three terms:

$$\Delta\bar{Y} = \underbrace{(\bar{X}_A - \bar{X}_B)'\hat{\beta}_B}_{\text{endowments}} + \underbrace{\bar{X}_A'(\hat{\beta}_A - \hat{\beta}_B)}_{\text{coefficients}} + \underbrace{(\bar{X}_A - \bar{X}_B)'(\hat{\beta}_A - \hat{\beta}_B)}_{\text{interaction}}$$

Decomposition

$$\Delta\bar{Y} = \underbrace{(\bar{X}_A - \bar{X}_B)'\hat{\beta}_B}_{\text{endowments}} + \underbrace{\bar{X}_A'(\hat{\beta}_A - \hat{\beta}_B)}_{\text{coefficients}} + \underbrace{(\bar{X}_A - \bar{X}_B)'(\hat{\beta}_A - \hat{\beta}_B)}_{\text{interaction}}$$

The raw differential $y_A - y_B$ is decomposed into a part:

- due to differences in endowments (E)
- due to differences in coefficients (including the intercept) (C)
- due to interaction between coefficients and endowments (CE)

Depending on the model which is assumed to be non-discriminating, these terms may be used to determine the “unexplained” (i.e., discrimination) and the “explained” part of the differential

Decomposition: Stata procedure

- install the decompose command from the web

```
ssc install decompose
```

- generate artificial data, draw random samples of “white” and “black” employees

```
mat m_W=(12,18,0) /* matrix of means of RHS vars for Whites*/
mat c_W=(5,-.6, 0 \ -.6,119,0 \ 0,0,.1) /*cov. matrix of RHS vars*/
mat m_B=(8,23,0) /* matrix of means of RHS vars for Blacks*/
mat c_B=(5,-.6, 0 \ -.6,119,0 \ 0,0,.1) /*cov. matrix */
```

- Draw a sample of 2000 obs. for Whites and 1000 obs. for Blacks

```
set seed 10000
set obs 2000
gen black=0
drawnorm edu exp e, means(m_W) cov(c_W)
save Whites1.dta, replace
set seed 20000
set obs 1000
gen black=1
drawnorm edu exp e ,means(m_B) cov(c_B)
append using Whites1.dta
```

Decomposition: Stata procedure

```
drop if (exp<0 | exp>40)    /*Drop obs. with extreme values of exp */
gen exp2=exp^2
gen logY=7.6+ edu*.07 + exp*.012  + e if black==0 /*log of earnings for
    Whites */
replace logY=4.6+ edu*.04 + exp*.012 + e if black==1 /*log of earnings for
    Blacks */
table black, contents(mean logY mean edu mean exp mean e)

decompose logY edu exp, by(black) detail estimates
```

Decomposition: Stata procedure

- The first block of output reports the mean values of y for the two groups, and the difference between them. It then shows the contribution attributable to the gaps in endowments (E), the coefficients (C), and the interaction (CE)
- The second block of output shows how the explained and unexplained portions of the outcome gap vary depending on the decomposition used
- The third block of output allows the user to see how far gaps in individual x's contribute to the overall explained gap
- The fourth and final block of output gives the coefficient estimates, means, and predictions for each x for each group