

EC831 Lecture:
Working on an Applied Project(in Stata)

Paul Fisher

14th November 2011

Some Key Points for an Applied Project

- **Find data** that can address your research question (given that you have your research question)
- Find an **econometric method** that with your data can address your question
- **Get to know your data** in Stata
- Undertake analysis in **Stata**
- **Interpret** your Stata analysis

Plan for the Lectures

- The first part of the lecture will hopefully get you to think about the **types of data** and **econometric method** you may want to use in your project.
- In the second half(next week) we will see some **example** outputs from Stata which will hopefully get you thinking about the process of getting to **know your own data** and to introduce you to some of the **outputs that can be produced** in stata
- (We will see how to produce some of the outputs in the labs starting next week)

Econometric Method to Use?

- Should be jointly decided with:
 - ▶ the **type of data** we choose to use and
 - ▶ the type of **dependent variable** our research question relates to

Econometric Method to Use?

- Does your research question relate to a **continuous dependent variable**?
 - ▶ eg. Income, Wages, GDP/GNP, Sales, Consumption
- Or a **discrete dependent variable**?
 - ▶ eg. Employment status, Whether using Childcare?, Happiness: 1,2,3,4

Econometric Method to Use?

- For **continuous dependent variables** we will want to use multiple linear regression models:
 - ▶ **OLS** is used in the majority of applied work in economics. (even papers published in top academic journals)
- For **discrete dependent variables** we will want to use limited dependent variable models:
 - ▶ eg. **probit** with the *probit* command for binary dependent variables (or we could estimate a linear probability model using *regress*)
 - ▶ **ordered probit** with the *oprobit* command for dependent variables that are ordered eg. Happiness on a scale of 1-5
 - ▶ (If we are not comfortable working with the more exotic ordered probit then the dependent variable could be recoded as a zero one variable with 1 indicating happiness greater than 2 say)

Econometric Method to Use?

- You may have seen another econometric method you wish to use. There are many!
- Search the **stata help file** first if you have a particular model in mind.
- Using a complicated model is **not a guarantee of a successful project**. (The majority of applied economics uses OLS - including applied work in top academic journals).
- The goal is to **answer your research question** - use a model which is appropriate to that.
- Make sure you know the **key assumptions** of your method and can **interpret the coefficients** correctly!

Econometric Method to Use?

- We must also consider the **type of data** we will use when choosing our econometric method
- The type of data we have will determine the tools we use to manipulate and analyse the data
- There are four data types:
 - ▶ **Cross-sectional data**
 - ▶ **Time-Series data**
 - ▶ **Cross-section/time-series data**
 - ▶ **Panel data**

Cross-sectional data

- Contains measurements on distinct units at a **given point in time**
- eg. **individuals, countries, firms**
- Sample size is given by the number of cross-sectional units N
- Measurements are typically for a given period eg. Income in one month, share price in one day
- Data are indexed i and not t hence the **order of observations is arbitrary**
- Typically this data has an identifier variable eg. Country name, Person ID, Firm code

Pooled Cross-Sectional data

- Every observation is indexed to i and t . eg. James family in year 2000
- i is **randomly sampled in each period**. Therefore we **do not** follow the same units over time. eg. James family in year 2000, Hernandez family in 2001
- Eg. A **different** 1000 households report their consumption each year. An example dataset is the Expenditure and Food Survey
- Can examine individual and time dimensions. However we look at average changes and not changes of each individual as with panel data

Time-Series data

- Sequence of observations on a variable **eg. A stock price every day, monthly interest rates, real GDP each year, monthly rpi**
- Data is collected at specific points in time with observations indexed to t eg. Y_t Stock price at time t .
- **Observations are ordered**
- We may need to use a method accounting for **autocorrelation**
- Periods are identified by a **date variable** eg. day, month, year
- Time series operators in Stata allow us to specify **lags, leads, differences and seasonal differences**
- Sample size is given by T the total number of time periods we observe eg. If we have 10 years of data $T=10$

Panel data

- We follow a **group of units over time** eg. a households is followed over many years
- Observations are indexed i and t eg. If Y is GDP Y_{it} is GDP of country i at time period t
- Eg. British Household Panel Survey (BHPS) here at Essex
 - ▶ In this case we have large i and small t
- Sample size is given by the number of units N multiplied by the number of time periods T
- An **advantage** is that we can look at **transitions** eg. Movements from unemployment to employment
- Also we can account for certain types of **omitted variables bias**

Panel and Time Series data

- Note. For Time Series and Panel data we need to let Stata know which variables indicate the **time and panel variables**.
- We use the *tsset* command for time series data and *xtset* command for panel data - but more on this in the labs.

Econometric Model to Use?

- Cross sectional data/Pooled Cross-sectional data
 - ▶ dependent variable is **continuous**: OLS (*regress*)
 - ▶ dependent variable is **discrete**: Could use OLS, *probit*, *oprobit*

OLS

- We model the **dependent variable** as a linear function of the **independent variables**
 - ▶ eg. $Unemployment = \alpha + \beta_1 Inflation + \beta_2 Population + \beta_3 GDP + \mu$
- Use the *regress* command in Stata
- Can you **justify the assumptions** of your method? How about:
 - ▶ **Zero Conditional Mean Assumption** $E[x|\mu]=0$
 - ▶ Are you **controlling** for enough factors in μ ? Or are there **unobserved variables** related to the independent variables?
- Is **heteroskedasticity** a problem?
 - ▶ Stata can implement a Breush-Pagan test with the *hettest* command
 - ▶ Stata can also calculate heteroskedasticity **robust standard errors**
- Can you justify the functional form you choose?
 - ▶ Are the variables in **Levels? Logs?**
 - ▶ Do you include **Squares? Interactions?**

Time Series data

- Again our favourite method: **OLS**
- Maybe you want to test an economic theory with a **static model**?
 - ▶ eg. $ExchangeRate_t = \alpha + \rho RelativeInflation_t + \mu_t$
- Or maybe your research question involves a **lagged effect**?
 - ▶ eg. $Birthrate_{year2011} = \alpha + \rho ChildBenefitValue_{year2011} + \gamma ChildBenefitValue_{year2010} + \mu_t$
- How do you **interpret the coefficients** from this model?
- Again we should think carefully about our **assumptions** when using time series data
 - ▶ Particularly $corr(u_t, u_s | X) = 0$ (No **serial correlation**)
 - ▶ Could try serial correlation robust standard errors
- Functional form issues: Do you want to **difference your variables**?
How many **lags**? **Time trends**? **Seasonality**? **Logs**?
- Advanced time-series methods include **unit root** testing and **cointegration**.

Econometric Model to Use?

- Panel Data

- ▶ Working with Panel Data can be **very challenging**
- ▶ Given your limited time it may be better to work with a **single year of the panel** and treat it as a cross-section
- ▶ Alternatively you could **pool different years of the panel** together and use OLS
 - ★ This will increase your sample size but your errors are likely to be **serially correlated**
 - ★ Including **lags of the dependent variable** in your regression would allow you to control for time invariant types of omitted variable bias

Once you have implemented your method...

- A good project will undertake some **sensitivity analysis**
- The idea is to **modify** your model/method and see if your **results still hold**

Sensitivity Analysis

- This could involve trying:
 - ▶ a **different dependent variable** eg. Change from Hours Worked to an employment dummy
 - ▶ **Restricting the sample** in some way eg. Restrict your data to a particular year and see if the results are the same
 - ▶ Try **dropping outliers**
 - ▶ If your results do not hold **think about why** this may be.....Research is never easy!
 - ▶ **Explain your thinking** clearly when writing up

- Part II: **An Applied Example**

- ▶ Research Question: What is the causal effect of education on wages?

Data

Research Question: What is the causal effect of education on wages?

- We will answer this question with data from the British Household Panel Survey (BHPS) (Available through the **UK data archive** website)
- Annual interviewing of the same individuals
- We will look at one year of the panel - 2007 (**Cross-section**)
- We will further **restrict our sample** to working age individuals age 25-55
- BHPS has a rich set of control variables. As we have limited time we will just consider variables for **wages, education, age, sex and marital status** in this lecture.

Methodology

Research Question: What is the causal effect of education on wages?

- Economic theory suggests that the log-level model is the correct specification
- We will estimate the relationship using an **OLS earnings equation**:
 - ▶ $\text{Log}(\text{Wage}) = \alpha + \beta \text{YearsofEducation} + \rho \text{Age} + \delta \text{Married} + \gamma \text{Sex} + \mu$
- Key assumption: $E[\mu|x]=0$ (we will return to this later)

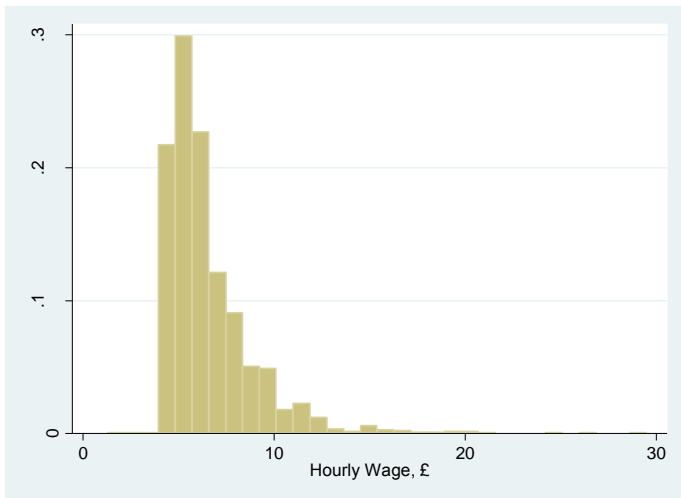
Data Cleaning

Research Question: What is the causal effect of education on wages?

- But first we need to get **familiar with our data**.
- BHPS is available in stata format(.dta extension). We can easily open the main data set in stata (if your data is in another format we have to use the **insheet** command - more on this in the labs)
- Think about: **The structure of the data/how it is saved? Conventions? Missing value codes (could be . or -99 for example)? Units of the variables? Binary variables? Ordinal variables? Real/Nominal values? Base period? Rates/Percentages/Proportions? Logs? Are the time series chronologically ordered? Outliers?**
- Begin by **summarising means, standard deviations, number of observations, minimum and maximum values**

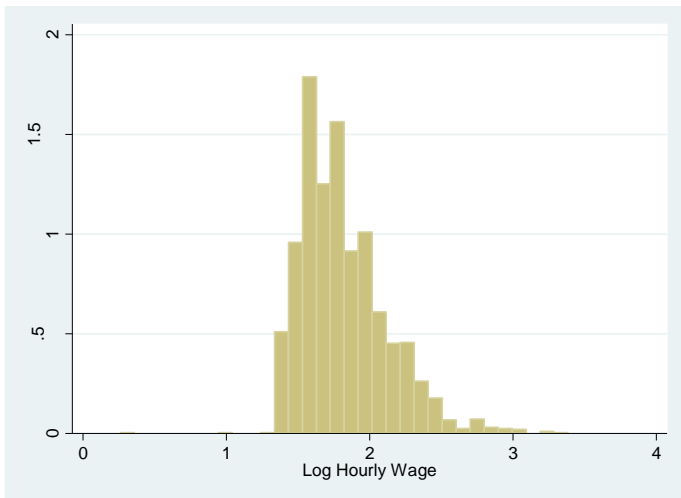
Variable	Obs	Mean	Std. Dev.	Min	Max
wage	8821	-4.548512	6.101695	-9	29.5
education	8821	3.654613	3.411847	-9	7
age	8821	39.81013	8.555553	25	55
sex	8821	.4661075	.4988782	0	1
married	8821	.6091589	.4879665	0	1

- **Summarising** the variables reveals some odd patterns.
- Focussing on the wage we see a negative mean wage and negative minimum value
- This reveals that we have not dealt with missing wage values
- A quick look in the BHPS supporting documentation tells us that wages are measured in pounds per hour and missing values are coded as negative
- We should set the negative values to missing which is indicated by a "." in stata
- (Also we notice a minimum value for education which is negative. We will explore this later.)



- It is also useful to look at **plots of our variables** - Stata has various graph options
- The **histogram** confirms that we have removed the missing wage values but reveals some outliers.

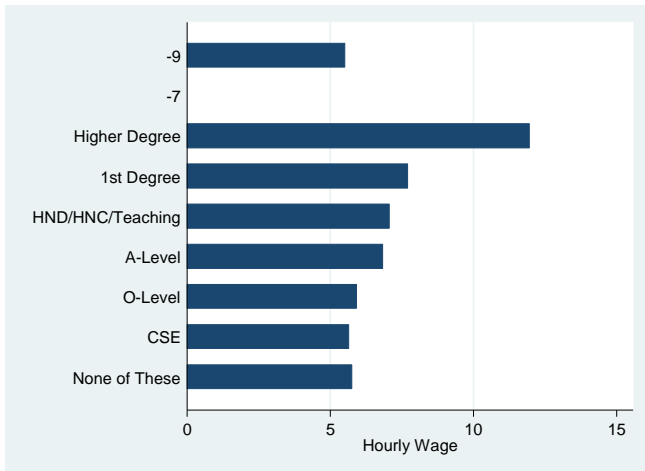
- But...our dependent variable should be the **log** of the hourly wage(from Economic theory)
- We should **generate a new variable** in stata equal to the log of the hourly wage
- And plotting the histogram.....



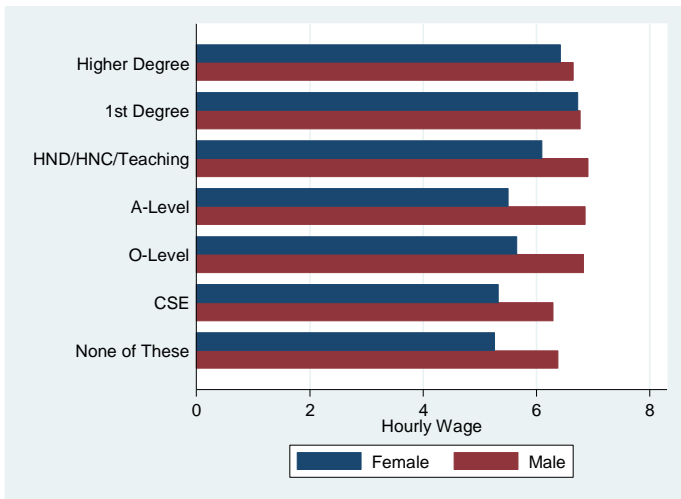
- We can see that this has **compressed the distribution**
- The effect of outlying observations on regression estimates will be lessened
- Our coefficients will have a **semi-elasticity interpretation**

Highest Qualification	Freq.	Percent	Cum.
-9. missing or wild	96	1.09	1.09
-7.proxy respondent	505	5.72	6.81
1. higher degree	332	3.76	10.58
2. 1st degree	1187	13.45	24.03
3.hnd,hnc,teaching	683	7.74	31.78
4. a level	1769	20.05	51.83
5. o level	2217	25.14	76.98
6. cse	501	5.69	82.67
7. none of these	1528	17.33	100.00
Total	8821	100.00	

- Returning to the education variable: a **tabulation** reveals that missing and proxy values are set as -9 and -7.
- The education variable is **categorical**
- We should either break it down into **dummy variables** or create a **years of education variable**



- We can also plot a **bar graph** of the mean of the hourly wage by highest qualification
- We can see that the **hourly wage increases with education level**
- We still have the missing categories labelled -9 and -7



- (It may also be interesting to look at mean wages by gender)

- Lets deal with the **categorical** education variable by creating a years of education variable
- Assuming that people begin school at age 4: Lets assign 18 years of education to people reporting a higher degree, 17 years to people with a first degree, 14 for an a-level, and 12 years of education to people reporting GCSE/O-level/none of these
- These are **assumptions** which we should **discuss** in our writing up
 - ▶ Do all people begin school at age 4? Are GCSEs comparable to "none of these"? Length of degrees? Higher degrees? O-levels to GCSEs?
- A good project would **experiment** with different education variables available in BHPS

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	1946	6.518541	2.441973	1.3	29.5
yearseducation	8220	13.81024	2.173776	12	18
age	8821	39.81013	8.555553	25	55
sex	8821	.4661075	.4988782	0	1
married	8821	.6091589	.4879665	0	1

- Once these changes are implemented things look more normal
- We now have a positive mean hourly wage
- We have created a **years of education variable** ranging from 12 to 18
- The number of observations on the wage has fallen as we have dropped the missing wage information

```
. regress lnwage yearseducation sex married age
```

Source	SS	df	MS			
Model	31.0278622	4	7.75696555	Number of obs =	1922	
Residual	150.331341	1917	.078420105	F(4, 1917) =	98.92	
				Prob > F =	0.0000	
				R-squared =	0.1711	
				Adj R-squared =	0.1694	
Total	181.359203	1921	.094408747	Root MSE =	.28004	

lnwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearseduca~n	.0455416	.0040687	11.19	0.000	.0375621	.0535212
sex	.2030527	.0129421	15.69	0.000	.1776707	.2284348
married	.057049	.0135691	4.20	0.000	.0304372	.0836608
age	.0017205	.00079	2.18	0.030	.0001712	.0032698
_cons	.8641667	.0795334	10.87	0.000	.7081857	1.020148

- Here is the stata output from an OLS regression of log wages on our independent variables
- The Stata output presents **N**, the **coefficient estimates**, **standard errors** and the **adjusted R squared**
- Results of **t-tests**, **F-tests** and **confidence intervals** are also given
- The coefficient on education is **highly significant** and implies that one more year of education raises the hourly wage by 4.55 percent

. regress lnwage yearseducation sex married age age2

Source	SS	df	MS	
Model	31.9238531	5	6.38477061	Number of obs = 1922
Residual	149.43535	1916	.077993398	F(5, 1916) = 81.86
Total	181.359203	1921	.094408747	Prob > F = 0.0000
				R-squared = 0.1760
				Adj R-squared = 0.1739
				Root MSE = .27927

lnwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearseduca~n	.0466291	.0040703	11.46	0.000	.0386464	.0546118
sex	.2043484	.0129125	15.83	0.000	.1790244	.2296724
married	.0528193	.0135896	3.89	0.000	.0261673	.0794712
age	.0266999	.0074118	3.60	0.000	.0121638	.041236
age2	-.0003122	.0000921	-3.39	0.001	-.0004928	-.0001316
_cons	.5569599	.1582317	3.52	0.000	.2466353	.8672844

- Maybe we should allow for **non-linear terms** in age
- We can **generate** a variable equal to the square of age
- The age squared variable is **significant** in our regression
- The coefficient on education remains **highly significant** and the magnitude increases slightly (from 0.0455 to 0.0466)

```
. hettest
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
```

```
Ho: Constant variance
```

```
Variables: fitted values of lnwage
```

```
chi2(1)      = 113.58
```

```
Prob > chi2  = 0.0000
```

```
.
```

- We can conduct a Breush-Pagan Test for **Heteroskedasticity**
- **What does this result imply?**
- (If we had time series data maybe we would want to test for **serial correlation**)

- How about our key Assumption: $E[\mu|x]=0$
- Maybe we could search for some more **controls**? Ability? Family Background?

. regress lnwage yearseducation daddegree sex married age

Source	SS	df	MS	Number of obs = 1922	
Model	31.0541551	5	6.21083102	F(5, 1916) =	79.17
Residual	150.305048	1916	.078447311	Prob > F =	0.0000
				R-squared =	0.1712
				Adj R-squared =	0.1691
Total	181.359203	1921	.094408747	Root MSE =	.28008

lnwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearseduca~n	.0453064	.0040897	11.08	0.000	.0372857	.053327
daddegree	.0248909	.0429942	0.58	0.563	-.0594295	.1092112
sex	.2033203	.0129526	15.70	0.000	.1779177	.2287229
married	.0567121	.013584	4.17	0.000	.0300712	.0833531
age	.0017222	.0007901	2.18	0.029	.0001727	.0032718
_cons	.8676011	.0797681	10.88	0.000	.7111597	1.024042

- Including a measure of fathers education in our regression (a dummy indicating father has a degree) has little impact on the estimated years of education coefficient (0.0453 vs. 0.0455)
- Looking at the p-value tells us that it remains **statistically significant**
- Our result thus seems to be **robust**. With more time we could **add further controls** to our baseline regression.

```
. regress lnwage yearseducation sex married age if wage<20
```

Source	SS	df	MS			
Model	28.9476184	4	7.2369046	Number of obs =	1915	
Residual	140.330373	1910	.073471399	F(4, 1910) =	98.50	
Total	169.277991	1914	.088442002	Prob > F =	0.0000	
				R-squared =	0.1710	
				Adj R-squared =	0.1693	
				Root MSE =	.27106	

lnwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearseduca~n	.0380483	.0039958	9.52	0.000	.0302117	.0458849
sex	.2090268	.0125461	16.66	0.000	.1844214	.2336322
married	.0563658	.0131494	4.29	0.000	.0305771	.0821545
age	.0013596	.0007662	1.77	0.076	-.000143	.0028622
_cons	1.150853	.0635573	18.11	0.000	1.026204	1.275502

- How about **sensitivity checks**? We could try dropping the outlying observations on wages
- Our results seem to hold

- Then writing up.....(not the topic of this lecture)
- But we should document **all** the things we have tried
- (Note. **Do not** copy and paste stata tables directly into your project. You should make your own tables showing only the relevant information as in an academic paper.)

In Summary

- You have a great opportunity to present some evidence on a real world problem!
- Take advantage of the opportunity!
- Think jointly about the:
 - ▶ type of dependent variable your question relates to
 - ▶ type of data available to you
 - ▶ methodology you will use
- Spend some time getting to know your data
- Try to justify the assumptions of your method
- Try some sensitivity analysis

Finally.....

- Don't get scared - it can be quite daunting when starting a big project for the first time!
- Enjoy the opportunity to undertake your own piece of research and say something about a real-world problem!