

A Guide to Getting Started on Empirical Work

It is difficult to get far in applied work without having a solid grasp of the issues and appropriate estimation techniques needed when working with real data. The following is intended as a guide to the main issues that you are likely to encounter when you begin applied work. Most of the examples below deal with cross-section data, though many of the issues addressed are also applicable to users of time series data, which in addition has its own concerns. There are several summary references now around in the literature and I have tried to list the main ones in the reference section. As always, it usually helps to see examples and I have tried to illustrate the text with issues that I have encountered in my own work.

1. Know your data

Assuming you have managed to come up with a sensible idea to test and managed to get hold of a data set that will enable you to test your ideas, (in itself a non-trivial task), it really is important to get a feel for the variables to be used in your analysis **before** you begin any rigorous empirical work. This entails examining each variable in detail, checking units of measurement (which will help you interpret the meaning of any estimated regression coefficients); looking at the means, standard deviations, minima and maxima of the data - since the latter can often pick up the presence of outliers - which may or may not prove to be valid observations - and/or measurement error in the data.

It is a good idea to get into the habit of describing the basic trends and features of your data set, (this can be used to motivate the issue you are trying to address). This requires:

- 1) Graphing the key variables of interest – the dependent variable and the principal explanatory variables – can often be useful in detecting patterns in your data and putting across a central argument in your work.

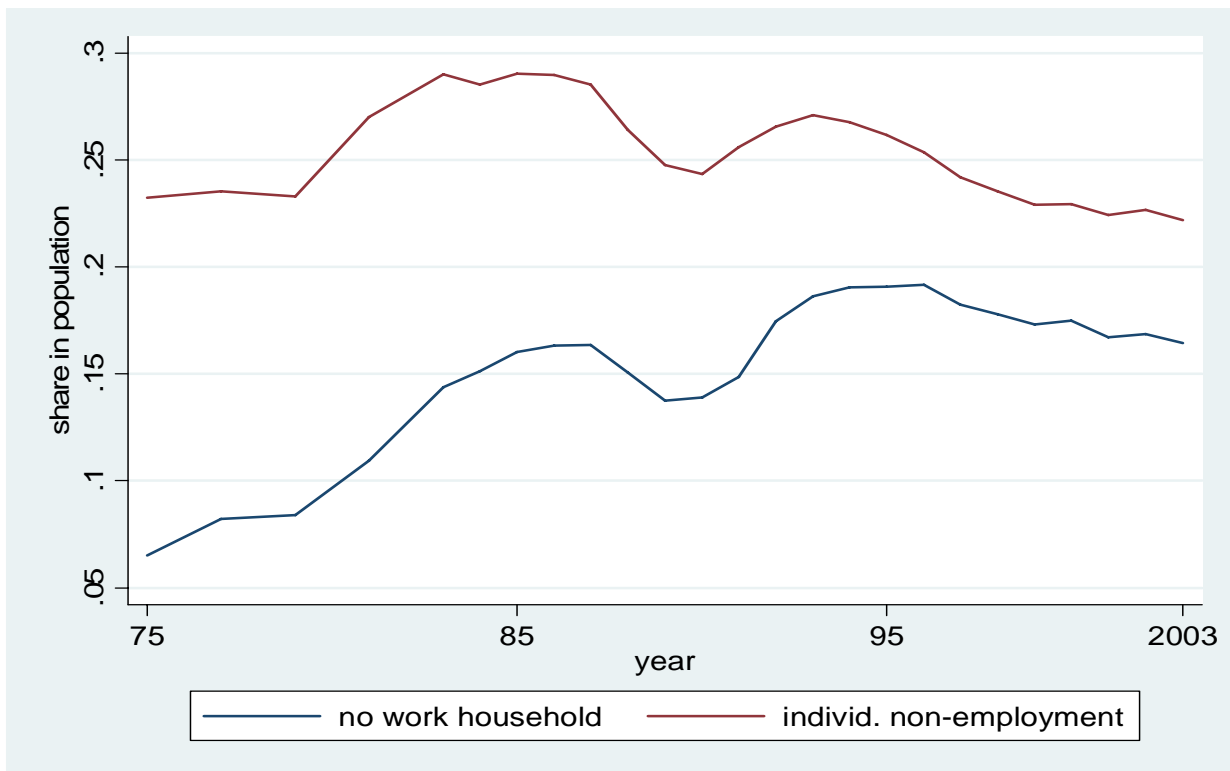
Stata's graphic package is quite sophisticated and you should learn the basics of the "graph" and "kdensity" commands

Example 1. Use of graph command

This example is taken from a paper where we measure non-employment from a time series of cross-section household survey data, using a) the individual b) the household as the base unit from which to aggregate up.

```
. lab var nonemp "individual non-employment"  
. lab var wless "workless household"
```

```
twoway (line noworkh year) (line nonemp year), ytitle(share in population) xscale(range(75 103))  
xlabel(75 85 95 103, valuelabel)
```



It is immediately obvious from inspection of the graph that the 2 units of aggregation give conflicting signals about British labour market performance over the last 25 years or so. On one measure (the individual) non-employment moves around over the cycle but shows no trend. On the household aggregation, non-employment has tripled over the same period.

- 2) Tabulating the mean values and the standard errors of the variables used in your analysis. You can do this using the *summary variable-name, detail* command in Stata. Be careful with standard errors of proportions – Stata command to get the means and standard errors of binary variables is
`ci variable name(s)`

Stata's command to get means and variances of the distribution for each variable is
`sum variable name(s), detail`

It is good practice to include a table giving the mean and standard deviations of all the variables you use in your study along with their units of measurement along with a note on how many observations you have.

Sometimes this can help you detect values that may have either been mis-typed or, if the data has come from a secondary source, which contain unexpected values

Example 1. Suppose you wish to measure the returns to education by regressing the log of hourly wages on years of education from the data set *lfs00.dta*

The raw variable in this particular data set is called "edage" which measures the age at which the individual completed full-time education

Using the summary command we get the following output

```
. su edage, detail
      age when compltd cont. ft education
-----
Percentiles      Smallest
1%                14                6
5%                15                6
10%               15                7
25%               16                7
50%               16
75%               19                97
90%               23                97
95%               96                97
99%               96                97
Obs              82763
Sum of Wgt.      82763
Mean             22.59312
Std. Dev.        20.16311
Variance         406.551
Skewness         3.313285
Kurtosis         12.1779
```

your suspicions should be aroused by the fact that at least 5% of the sample appear to have completed their education at age 96. For the time being ignore this and create the variable to be used in the study. Years of education is calculated as

```
. g yearsed=edage-6
(2813 missing values generated)
/* missing values because some are not asked the question in addition to not
replying and being coded 96 or 97. */
```

```
. reg lhw yearsed
Source |      SS      df      MS                Number of obs = 16087
-----+-----
Model | 146.197845      1 146.197845            F( 1, 16085) = 430.22
Residual | 5466.06331 16085  .339823644            Prob > F      = 0.0000
-----+-----
Total | 5612.26115 16086  .348891033            R-squared     = 0.0260
                                           Adj R-squared = 0.0260
                                           Root MSE    = .58294

lhw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
yearsed | -.0063071   .0003041   -20.742  0.000   -.0069031   -.005711
_cons   |  2.114567   .0063196   334.602  0.000    2.10218    2.126954
```

The regression suggest that 1 extra year of education **reduces** pay by around 0.6%

The reason is that values 96 & 97 in the edage data are **missing value codes**
Removing observations from the sample gives

```
. reg lhw yearsed if edage<96
Source |      SS      df      MS                Number of obs = 15487
-----+-----
Model | 596.675541      1 596.675541            F( 1, 15485) = 2002.10
Residual | 4614.90882 15485  .298024463            Prob > F      = 0.0000
-----+-----
Total | 5211.58436 15486  .336535216            R-squared     = 0.1145
                                           Adj R-squared = 0.1144
                                           Root MSE    = .54592

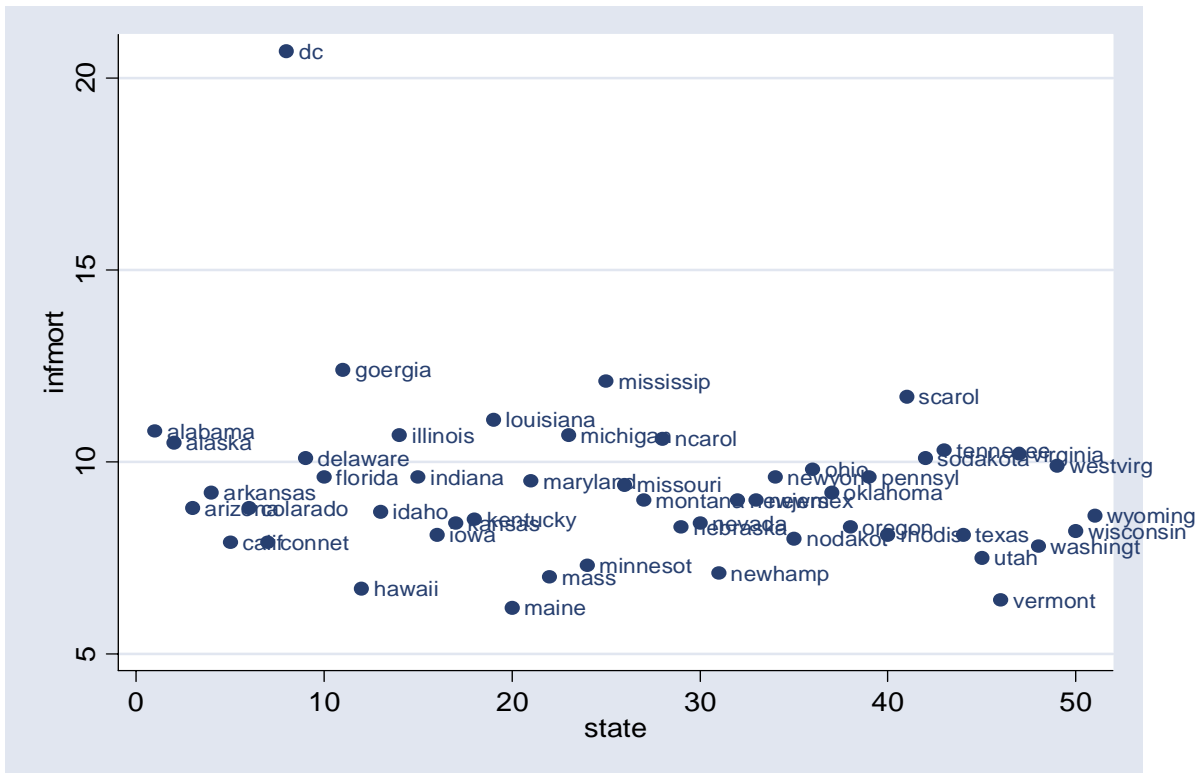
lhw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
yearsed |  .07752    .0017325    44.745  0.000    .0741241    .0809159
_cons   |  1.171652   .020113    58.253  0.000    1.132228    1.211075
```

which looks more sensible

Estimates can be sensitive to the presence of **outliers**. Outliers can have big effects on regression estimates, particularly in small samples and you should always admit to the presence (or absence) of outliers in the data. This can be caused either by a genuine observation being a long way from the main body of observations.

Eg. The data set *cex2.dta* contains data on infant mortality across U.S. states. Graphing this variable shows that Washington DC appears to be something of an outlier

```
. twoway (scatter infmort state, mlabel(state)), ytitle(infmort)
ylabel(, labels) xtitle(state) xlabel(, labels)
```



This is an example of a genuine observation but it does have a large influence in a regression. If we include DC we get

```
. reg infmort lpop ldocs lpcap
Source |          SS          df          MS          Number of obs =          51
-----+-----+-----+-----+-----+-----+-----+-----
      Model |    32.162998         3    10.7209993      F(  3,   47) =          2.53
Residual |   199.084471        47     4.23583981    Prob > F       =          0.0684
-----+-----+-----+-----+-----+-----+-----
      Total |   231.247469        50     4.62494938    R-squared      =          0.1391
                                           Adj R-squared  =          0.0841
                                           Root MSE     =          2.0581

-----+-----+-----+-----+-----+-----+-----+-----
      infmort |          Coef.      Std. Err.      t      P>|t|      [95% Conf. Interval]
-----+-----+-----+-----+-----+-----+-----+-----
      lpop    |   -0.0878224      .28725      -0.306   0.761      -0.6656948      .4900499
      ldocs   |    4.153261      1.512659     2.746   0.009      1.110185      7.196338
      lpcap   |   -4.684662      2.604124    -1.799   0.078     -9.923484      .5541606
      _cons   |   33.85931      20.42785     1.658   0.104     -7.236219     74.95484
```

The RHS variables are in logs, the dependent variable in levels so coefficients are semi-elasticities ($dy/d\log x_i = b_i$ so $dy = b_i * dx_i / x_i$ so $b_i / 100$ is the unit change in the dependent variable when x_i changes by 1%.)

Regression suggests more doctors leads to a **rise** in infant mortality – strange.
 (a 1% rise in doctors appears to increase infant mortality by 4 in every 100,000 **not** 4 in every 1000 – be careful to divide the estimated coefficient by 100).

In cross-section data it is not uncommon to delete suspect observations (though you must have a good reason for doing so.) If we exclude DC we get

```
. reg infmort lpop ldocs lpcap if state~=8
```

Source	SS	df	MS			
Model	26.8600265	3	8.95334216	Number of obs =	50	
Residual	71.4631754	46	1.55354729	F(3, 46) =	5.76	
Total	98.3232019	49	2.00659596	Prob > F =	0.0020	
				R-squared =	0.2732	
				Adj R-squared =	0.2258	
				Root MSE =	1.2464	

infmort	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lpop	.6292349	.1911062	3.293	0.002	.2445581	1.013912
ldocs	-2.741837	1.190773	-2.303	0.026	-5.138739	-.3449347
lpcap	-.5669275	1.641216	-0.345	0.731	-3.870524	2.736669
_cons	23.95479	12.41946	1.929	0.060	-1.044287	48.95388

Now doctor variable is negative and significant (makes more intuitive sense)
 A 1% rise in the number of doctors (per 100,000 inhabitants) leads to a fall in infant mortality of 2.7 per 100,000. The t value also suggests that the effect is statistically significantly different from zero (at the 5% level).

The other case is when the data are measured with error. With income or pay data it is quite common that individuals will mis-report income (due to rounding error or often reporting by 3rd party respondents or a badly worded question in the survey). Remember measurement error in the dependent variable just means higher standard errors on the rhs variables, but measurement error in the rhs variables leads to attenuation bias in the coefficients (coefficients biased toward zero) and you may have to worry about finding an instrument, (see section below).

Whatever the cause it is always a good idea to test the sensitivity of your estimates to outliers, which will help you decide whether to drop them or not

(Stata also has a series of commands to help you identify outliers – see the LVR2PLOT and DFITS commands – the latter used with the predict command eg
 predict dfits, dfits)

Interpreting Regression Coefficients

Many students do not put enough effort into describing the estimated effects from their regression models. You should always do 2 things when evaluating the effects of your main variables of interest.

- 1) Is it significantly different from zero? (use the t value for this)

2) If it is what is the estimated impact of the variable (even though a variable is statistically significant it may not have a large economic impact on the dependent variable)

Often knowing the units in which the rhs variable is measured will help interpretation. Again summarising the data set will give you a firm idea of the units of measurement

Example. The US infant mortality data contains information on 1990 infant mortality rates (per 1000), in each of the 51 states of the U.S. the number of doctors per 100,000 of the population in each state and the average per capita income (in dollars) of each state

If we summarise the data, the means give us a sense of the units of measurement

summ

Variable	Obs	Mean	Std. Dev.	Min	Max
year	51	1990	0	1990	1990
infmort	51	9.284314	2.15057	6.2	20.7
afdcnum	51	234.4118	335.1369	16	2023
pop	51	4876.647	5439.203	454	29760
pcapinc	51	17836.25	2967.524	12700	25528
docspop	51	205.1569	76.06034	125	615
afdcper	51	4.255785	1.462805	1.688183	8.896211
dc	51	.0196078	.140028	0	1
state	51	26	14.86607	1	51
ldocs	51	5.278746	.2800004	4.828314	6.421622
lpcap	51	9.775915	.1621253	9.449357	10.14753
lpop	51	7.995111	1.032828	6.118097	10.30092

A regression of the infant mortality rate (mean value 9.3 in every 1000) on the population level in each state (measured in 1000, so the mean value is 4876 ie 4,876,000), the number of doctors per 100,000 (mean: 205 in every 100,000) and state per capita average income (mean: \$17836) gives

```
. reg infmort pop docspop pcapinc
```

Source	SS	df	MS	Number of obs = 50		
Model	14.5822657	3	4.86075524	F(3, 46) =	2.67	
Residual	83.7409362	46	1.82045513	Prob > F =	0.0585	
-----				R-squared =	0.1483	
Total	98.3232019	49	2.00659596	Adj R-squared =	0.0928	
-----				Root MSE =	1.3492	
infmort	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pop	.0000651	.0000377	1.73	0.091	-.0000108	.0001411
docspop	-.0088853	.005938	-1.50	0.141	-.020838	.0030673
pcapinc	-.0000533	.0000995	-0.54	0.595	-.0002536	.0001471
_cons	11.42652	1.22515	9.33	0.000	8.96042	13.89262

None of the variables is statistically significant so we need not waste time describing their effects (though for the record the estimated coefficient on doctors suggests that if the number of doctors increases by 1 in every 100,000 then infant mortality would fall by 0.009 in every 1000 - $dy/dx_i = b_i$ so $dy = b_i * dx_i = -.009 * 1 = -0.009$)

If we regress the *level* of infant mortality on the *logs* of doctors, population and income

```
. reg linfm pop docspop pcapinc
```

Source	SS	df	MS	Number of obs = 50		
Model	.189703233	3	.063234411	F(3, 46)	=	2.81
Residual	1.03366681	46	.022471018	Prob > F	=	0.0496
				R-squared	=	0.1551
				Adj R-squared	=	0.1000
Total	1.22337004	49	.024966736	Root MSE	=	.1499

linfm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pop	7.69e-06	4.19e-06	1.83	0.073	-7.48e-07	.0000161
docspop	-.001053	.0006597	-1.60	0.117	-.002381	.0002749
pcapinc	-5.06e-06	.0000111	-0.46	0.650	-.0000273	.0000172
_cons	2.450128	.1361165	18.00	0.000	2.176139	2.724116

Again the t values show nothing significant so no need to comment further (but you should know how to interpret the estimates from a log-lin regression like this as $d\log(y)/dx_i = b_i$ so $dy/y = b_i \cdot dx_i$ ie so (% change in y)/ 100 = $b_i \cdot dx_i$ or % change in y = $100 \cdot b_i \cdot dx_i$ and if the number of doctors increases by 1 in every 100,000 then the infant mortality rate would fall by $100 \cdot -.001 \cdot 1 = -0.1$ ie about 0.1%

While a lin-log regression of the *level* of infant mortality on the *logs* of doctors, income and population gives

```
. reg infmort lpop ldocs lpcap
```

Source	SS	df	MS	Number of obs = 50		
Model	26.8600265	3	8.95334216	F(3, 46)	=	5.76
Residual	71.4631754	46	1.55354729	Prob > F	=	0.0020
				R-squared	=	0.2732
				Adj R-squared	=	0.2258
Total	98.3232019	49	2.00659596	Root MSE	=	1.2464

infmort	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lpop	.6292349	.1911062	3.29	0.002	.2445581	1.013912
ldocs	-2.741837	1.190773	-2.30	0.026	-5.138739	-.3449347
lpcap	-.5669275	1.641216	-0.35	0.731	-3.870524	2.736669
_cons	23.95479	12.41946	1.93	0.060	-1.044287	48.95388

the t values on 2 of the variables now suggest estimated effects of these variables (population and doctors) are significantly different from zero.

The RHS variables are now in logs, the dependent variable in levels so coefficients are semi-elasticities ($dy/d\log x_i = b_i$ so $dy = b_i \cdot dx_i / x_i$ so $b_i / 100$ is the unit change in the dependent variable when x_i changes by 1%.)

A 1% rise in the number of doctors (per 100,000 inhabitants) leads to a fall in infant mortality of 2.7 per 100,000 (**not** 2.7 in every 1000 – be careful to divide the estimated coefficient by 100).

The t value also suggests that the effect is statistically significantly different from zero (at the 5% level). – The column to the right of the t values are p values – which give the exact significance level at which the null hypothesis (that the effect is zero) can be rejected. For the doctor variable the p value suggests we could go down to a significance level of 2.6% and still reject the null of zero effect.

Time Series Data

As the name suggests, anything which consists of a set of observations on a set of variables over time. The time dimension varies with the nature of the variables under study. Aggregate economy data tend to be quarterly or even annual. Stock price data can be daily or even hourly.

The main advantage of time series data is that it can be used to examine whether past events influence current outcomes.

Time series data often come with their own unique econometric problems. Many time series display seasonality or trends. These issues must be dealt with before convincing econometric results can be reported.

Organising Time Series Data

The first thing to do is to ensure that all the data are input in chronological order, (earliest time period first, latest observation last). Failure to do this simple task will give spurious results.

For example, constructing change variables makes no sense if the data are not ordered chronologically. The best way to do this is to create a variable that denotes the time period relating to each observation. For annual data the following structure should look like:

```
list year cons income
      year  cons  income
1.    55 167320 162396
2.    56 168163 166385
3.    57 171626 169167
4.    58 176417 172361
5.    59 184046 181361
6.    60 191078 193426
7.    61 195233 201546
8.    62 199627 203972
9.    63 208902 212683
10.   64 215334 221808
```

For quarterly data, it may be better to have one time related variable that increases continuously and others to denote year and quarter. It will also be useful to create a set of seasonal dummy variables that can be used to “de-seasonalise” the data. This can be done manually when you input the data or by using the command

```
tab quarter, gen(q)
```

which will create a set of dummy variables, (one for each category in quarter – in this case 4)

For example in the data set *bopq.dta*

```
. list
```

	var1	bop	quarter	time	xchange1	year	q1	q2	q3	q4
1.	119.2	-451	1	1	.	1970	1	0	0	0
2.	122.1	-187	2	2	119.2	1970	0	1	0	0
3.	125.1	854	3	3	122.1	1970	0	0	1	0
4.	131.4	1113	4	4	125.1	1970	0	0	0	1
5.	135.9	1683	1	5	131.4	1971	1	0	0	0
6.	131.9	1276	2	6	135.9	1971	0	1	0	0
7.	123.1	-141	3	7	131.9	1971	0	0	1	0
8.	120.9	420	4	8	123.1	1971	0	0	0	1
9.	123.6	265	1	9	120.9	1972	1	0	0	0
10.	122.7	127	2	10	123.6	1972	0	1	0	0

The data are sorted by year and then quarter. These time related variables will be used by Stata in the calculation of several time-series related statistics, (eg Durbin-Watson test) Failure to do this simple task will give spurious results if you generate things like change variables.

To tell stata that you have time series data, once you have input the data, type the command

```
tsset name of time variable , yearly (or quarterly or monthly)
```

For example to get stata to recognise the quarterly data above type

```
tsset time, quarterly
```

Nominal or real?

One of the first things you should check is whether the variables are measured in real or nominal terms. This can make a big difference to your interpretation of your estimates. It is usual to use real variables when comparing changes over time. If you have nominal data then you will have to index them using an appropriate deflator. (Remember re-scaling all the variables won't change the value of the estimates, just the interpretation). Make sure you know the base year in which the units are measured in, (eg. 1990 pounds or 2002 pounds?).

Example: the data set *gdpuk.dta* contains both nominal and real levels of gdp. A regression of the log of gdp on a time trend will give the (approximate) annual growth rate of gdp

Using nominal gdp

```
. reg lgdp year
```

Source	SS	df	MS	
Model	123.799553	1	123.799553	Number of obs = 56
Residual	2.12232554	54	.039302325	F(1, 54) = 3149.93
				Prob > F = 0.0000
				R-squared = 0.9831
				Adj R-squared = 0.9828
				Root MSE = .19825
Total	125.921878	55	2.2894887	

lgdp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
year	.0919893	.001639	56.12	0.000	.0887033	.0952754
_cons	-170.0515	3.238013	-52.52	0.000	-176.5433	-163.5597

the coefficient suggests that nominal gdp grew, on average, by around 9% a year.

Using the real gdp variable regressed on the *same* right hand side variable

```
. reg lrgdp year
```

Source	SS	df	MS	Number of obs =	56
Model	9.31466539	1	9.31466539	F(1, 54) =	8145.64
Residual	.061749869	54	.001143516	Prob > F =	0.0000
Total	9.37641526	55	.170480277	R-squared =	0.9934
				Adj R-squared =	0.9933
				Root MSE =	.03382

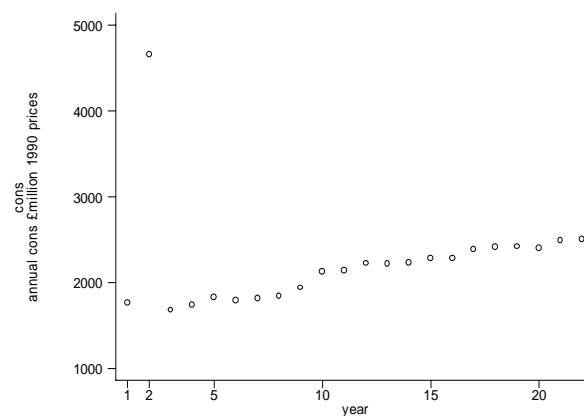
lrgdp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
year	.0252326	.0002796	90.25	0.000	.0246721	.0257931
_cons	-44.36674	.5523197	-80.33	0.000	-45.47408	-43.25941

the estimated coefficient now says that real gdp grew by 2.5% a year

What to do with outliers in a time series data set is a difficult issue. Discarding an observation in the middle of a time series may not be sensible. Some people effectively remove the influence of outliers by including a dummy variable for specific data points. (It can be shown that this is also equivalent to using the sample mean of the x variable to replace the suspect observation, which applies also to the case of missing data below).

With time series data it is not usually sensible to drop data. You can effectively ignore outliers in time series data, however, by including a dummy variable that equals 1 for that particular time period and 0 otherwise. The effect is as if you had estimated the regression over the entire sample period excepting the period causing concern, (though the R^2 will be typically lower in the latter).

Example. When trying to estimate a consumption function, you check the data by a graph of consumption over time



You should be struck by the observation on consumption in year 2. It seems much higher than in other years (and is in fact the result of a typing the 1st digit as “4” instead of “1”

when entering the data interactively). Yet with the data point included the regression estimate of the consumption function is

```
. reg cons income
```

Source	SS	df	MS	Number of obs = 22		
Model	5734.25183	1	5734.25183	F(1, 20)	=	0.01
Residual	7688939.20	20	384446.96	Prob > F	=	0.9040
-----				R-squared	=	0.0007
Total	7694673.45	21	366413.022	Adj R-squared	=	-0.0492
-----				Root MSE	=	620.04
cons	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	-.0604679	.4951133	-0.12	0.904	-1.093256	.9723204
_cons	2373.672	1154.948	2.06	0.053	-35.50723	4782.85

which doesn't look very sensible. So you decide to create a dummy variable covering the suspicious data period and include it as an extra variable in the regression.

```
g d2=year==2
```

```
. reg cons income d2
```

Source	SS	df	MS	Number of obs = 22		
Model	7041552.67	2	3520776.34	F(2, 19)	=	102.42
Residual	653120.781	19	34374.7779	Prob > F	=	0.0000
-----				R-squared	=	0.9151
Total	7694673.45	21	366413.022	Adj R-squared	=	0.9062
-----				Root MSE	=	185.40
cons	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.8171838	.1602557	5.10	0.000	.4817647	1.152603
d2	2938.774	205.4135	14.31	0.000	2508.839	3368.71
_cons	206.2528	377.1215	0.55	0.591	-583.0716	995.5772

Now the estimated marginal propensity to consume looks much better (and is the same as if you had excluded the data point from your estimates)

```
. reg cons income if year~=2
```

Source	SS	df	MS	Number of obs = 21		
Model	893824.457	1	893824.457	F(1, 19)	=	26.00
Residual	653120.781	19	34374.7779	Prob > F	=	0.0001
-----				R-squared	=	0.5778
Total	1546945.24	20	77347.2619	Adj R-squared	=	0.5556
-----				Root MSE	=	185.40
cons	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.8171838	.1602557	5.10	0.000	.4817647	1.152603
_cons	206.2528	377.1215	0.55	0.591	-583.0716	995.5772

Missing observations are also a common feature of time series data. Often this is because some data are not collected at the same rate of others, (eg quarterly observations on one variable v. monthly observations on another). In general trying to fill in gaps is not a good idea – because it introduces measurement error which will bias the results of an estimation.

Multicollinearity is also a common issue in time series data. The shorter the time series, in general, the worse the potential problems. Remember to look out for the signs :

- 1) small additions to the data lead to large differences in the coefficient estimates
- 2) large standard errors insignificant t values and high R² value
- 3) coefficients with implausibly large values

Unfortunately the only sensible options to the problem are to get more data or drop variables.

(Remember in a 3 variable model $y_t = b_0 + b_1X_{1t} + b_2X_{2t} + e_t$ the estimated variance on the coefficient b_1 is $Var(\hat{b}_1) = \frac{\sigma^2}{NVar(X_1)*(1-r_{1x2}^2)}$)

Example. The data set *emp.dta* is used to estimate the effect of GDP and inflation on the level of employment. The estimates of GDP and price are highly colinear, (both trend upward)

list year employ price gdp

	year	employment	price	gdp
1.	1947	60323	83	234289
2.	1948	61122	88.5	259426
3.	1949	60171	88.2	258054
4.	1950	61187	89.5	284599
5.	1951	63221	96.2	328975
6.	1952	63639	98.1	346999
7.	1953	64989	99	365385
8.	1954	63761	100	363112
9.	1955	66019	101.2	397469
10.	1956	67857	104.6	419180
11.	1957	68169	108.4	442769
12.	1958	66513	110.8	444546
13.	1959	68655	112.6	482704
14.	1960	69564	114.2	502601
15.	1961	69331	115.7	518173
16.	1962	70551	116.9	554894

and a simple correlation coefficient shows this to be true

. corr employ price gdp

	employ~t	price	gdp
employment	1.0000		
price	0.9709	1.0000	
gdp	0.9836	0.9916	1.0000

A regression over the whole sample gives

```
. reg emp price gdp
```

Source	SS	df	MS	Number of obs = 16		
Model	179184631	2	89592315.5	F(2, 13)	=	199.98
Residual	5824194.93	13	448014.995	Prob > F	=	0.0000
-----				R-squared	=	0.9685
Total	185008826	15	12333921.7	Adj R-squared	=	0.9637
-----				Root MSE	=	669.34
employment	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
price	-85.10657	123.7364	-0.69	0.504	-352.4228	182.2097
gdp	.0439148	.0134344	3.27	0.006	.0148916	.072938
_cons	56945.04	7449.448	7.64	0.000	40851.49	73038.59

and excluding the last observation gives

```
. reg emp price gdp if year<1962
```

Source	SS	df	MS	Number of obs = 15		
Model	151328327	2	75664163.7	F(2, 12)	=	203.61
Residual	4459425.58	12	371618.798	Prob > F	=	0.0000
-----				R-squared	=	0.9714
Total	155787753	14	11127696.6	Adj R-squared	=	0.9666
-----				Root MSE	=	609.61
employment	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
price	-244.6421	140.1079	-1.75	0.106	-549.911	60.62672
gdp	.0629839	.0157709	3.99	0.002	.0286221	.0973457
_cons	65878.65	8231.824	8.00	0.000	47943.04	83814.25

Clearly the estimates vary considerably. The coefficient on price is always imprecisely measured, (because it is so colinear with the GDP variable).

N.B. the variable price is an Index variable so be careful how you interpret the coefficient. An increase of 1 unit in an index value does not have immediate economic meaning (ie it is **not** equivalent to a 1 percentage (point) increase in price)

Stationarity

Ultimately whether you can sensibly include lags of either the dependent or explanatory variables in a regression also depends on whether the time series data that you are analysing are **stationary**

A variable is said to be (weakly) stationary if

- 1) its mean
- 2) its variance
- 3) its autocovariance $Cov(Y_t, Y_{t-s})$ where $s \neq t$

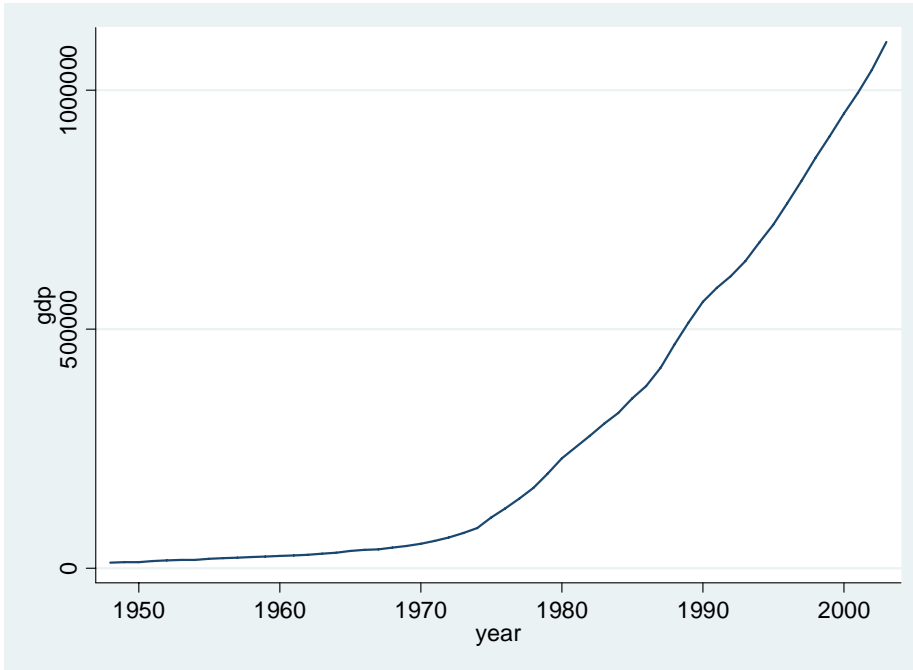
do not change over time

Stationarity is need if the Gauss-Markov conditions need for unbiased, efficient OLS estimation are to be met by time series data

(Essentially any variable that is **trended** is unlikely to be stationary)

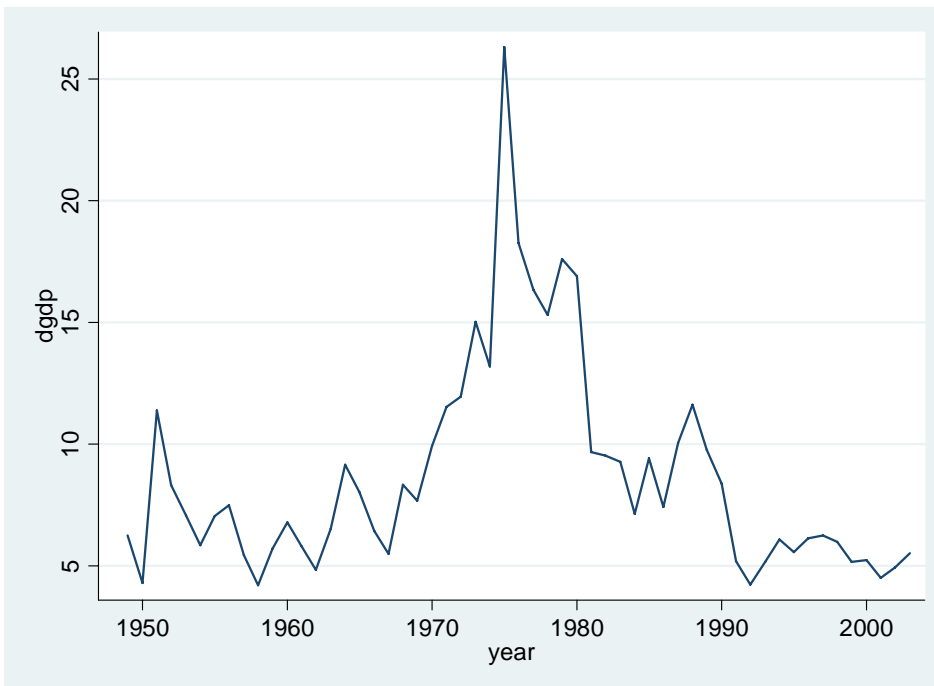
So for example graphing the data in *gdpuk.dta* shows that the level of gdp is clearly not stationary

`twoway (line gdp year), yscale(range(. .)) ylabel(, value1label) xscale(range(. .)) xlabel(, value1label)`



But the **change** in $\text{gdp} = \text{GDP}_t - \text{GDP}_{t-1}$ is

`twoway (line dgdp year), xscale(range(. .)) xlabel(, value1label)`



Remember: Regressing non-stationary variables on each other will generate **spurious regression** estimates (pick up common trends rather than any causal relationship). So always use **stationary** variables in **time series regressions**.

Endogeneity

One of the basic requirements of an OLS regression is that the relationship you estimate must be **causal**. In other words variation in the value of the right hand side X variables should explain variation in the explanatory (y) variable **and** there should be no reverse feedback ie y should not explain changes in X. If this happens then the X variable is said to be endogenous and OLS will give biased estimates.

Whether caused by measurement error, 2-way causality or omitted variable bias, it is quite common in empirical work to find that the error term is correlated with one or more of the right hand side variables, $Cov(X,u) \neq 0$.

Solution: try and find an *instrument* that is correlated with the right hand side variable of interest but uncorrelated with the error term (and the dependent variable by extension). There is some debate about what constitutes a good instrument. Some people prefer to rely on strict, but untested, assumptions about the precise structural form. Others rely on stories that can be tested empirically. In practice, a researcher should always try to demonstrate detailed knowledge of the relevant institutional details used to justify any instrument along with careful empirical investigation and quantification.

The usual way to estimate in the presence of endogeneity is 2SLS. In the 1st stage the endogenous rhs variable is regressed on all the instruments. The predicted value from this regression is then included instead of the endogenous variable in the 2nd stage.

Things that practitioners should be aware of when using IV are that:

2SLS estimates are consistent but can be biased in small samples. In general the bias will increase with the degree of over-identification (number of instruments relative to number of endogenous right hand side variables) in **small** samples, so using less instruments will reduce bias.

2SLS estimates generally have larger standard errors than OLS estimates. This is because (using the 2 variable model to illustrate), the variances of the slope coefficient using the 2 estimation strategies are

$$Var(\hat{\beta}_{OLS}) = \frac{\sigma^2}{N * Var(X)} \qquad Var(\hat{\beta}_{2SLS}) = \frac{\sigma^2}{N * Var(X) * \rho_{X,Z}^2}$$

where $\rho_{X,Z}^2$ is the square of the correlation coefficient between the endogenous variable and its instrument. The smaller the correlation of the instrument, Z, with the endogenous rhs variable, X, the larger the standard error in IV regression.

A poor correlation can also mean that the IV estimate of the coefficient is also biased even asymptotically. The probability limit of the IV estimator (in the 2 variable model) is given by

$$p \lim \hat{\beta}_{IV} = \beta + \frac{Cov(Z, u)}{Cov(Z, X)}$$

so that, even if the correlation between the error term and the instrument is small, the inconsistency in the IV estimate can be large if $Cov(Z, X)$ is small. In this case OLS may be preferable, even if it is also inconsistent

Moral: - always check the correlation of an instrument and its regressor (and with that of the other rhs variables – it may be highly correlated with other X variables and therefore not doing anything extra) before proceeding. Can do this using correlation coefficients or simple regression.

Bound et. al (1995) suggest 2 tests for the quality of any instrumental variable

1) an F test on the joint explanatory power of all the instruments (not the original exogenous variables in the structural form equation), when regressed on the endogenous variable. An F value >1 suggests, to the authors, that the instruments are sufficiently strong.

2) Adding the instruments to the reduced form equation should increase the adjusted R^2 in this regression. (An equivalent test is to look at the partial $R^2 = \frac{R^2_{yxz} - R^2_{yx}}{1 - R^2_{yx}}$, where R^2_{yx}

is obtained from a regression of the endogenous variable on the original x variables and R^2_{yxz} is obtained from a regression of the endogenous variable on the original x variables and the instruments.

Example

Consider the example of the effect of years of education on (log) hourly wages from a sample of British men and women taken from the 1998 GHS, (the dataset *ivex.dta* is available). Policy makers are often interested in the costs and benefits of education. Some people argue that education is endogenous (because it picks up the effects of omitted variables like ability or motivation) and so is correlated with the error term.

The OLS estimates suggest that for men

```
. reg lhw yearsed if lhw>1 & sex==1
```

Source	SS	df	MS	Number of obs = 3174		
Model	72.518738	1	72.518738	F(1, 3172)	=	251.84
Residual	913.400547	3172	.287957297	Prob > F	=	0.0000
-----				R-squared	=	0.0736
Total	985.919285	3173	.310721489	Adj R-squared	=	0.0733
-----				Root MSE	=	.53662
lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearsed	.0457013	.0028798	15.87	0.000	.0400547	.0513478
_cons	1.64272	.0350984	46.80	0.000	1.573903	1.711538

1 extra year of education is associated with 4.6% increase in earnings.

If endogeneity is a problem, then these estimates are biased, (biased upward if ability and education are positively correlated – see notes on omitted variable bias).

It is usually worth also establishing the endogeneity of the suspicious right-hand side variable. Can test for possible endogeneity using the Hausman-Wu test, (though this sort of relies on having a good instrument in the first place).

First regress the potentially endogenous variable on all the exogenous variables in the system. Save the residuals and add them to the original structural equation. Wooldridge (2001) shows that a test of the orthogonality of the right hand side variable becomes a test

of whether the residuals from this augmented regression are statistically significant, (with more than one potentially endogenous variable just add a residual for each auxiliary equation).

So try to instrument instead. You choose whether the individual owns a black and white television.

To be a good instrument the variable should be uncorrelated with the error term in the original structural equation, (if the variable has explanatory power then it should have been in the original structural equation), but correlated with the endogenous right hand side variable (education).

Correlation between a potential instrument and an (unobserved) true residual can never be tested. An indirect test often used is to check whether the instrument is partially correlated with the dependent variable in the structural equation. ie after other exogenous variables are included the addition of the potential instrument should not be statistically significant, nor should it change the coefficients on the original exogenous variables. To test this you first examine the correlation between wages and tv in a regression.

```
. reg lhw bw if lhw>1 & sex==1 & e(sample)
```

Source	SS	df	MS			
Model	.106187883	1	.106187883	Number of obs =	3174	
Residual	985.813097	3172	.31078597	F(1, 3172) =	0.34	
Total	985.919285	3173	.310721489	Prob > F =	0.5589	
				R-squared =	0.0001	
				Adj R-squared =	-0.0002	
				Root MSE =	.55748	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lhw						
bw	.0166282	.0284471	0.58	0.559	-.0391484	.0724047
_cons	2.176468	.0106755	203.88	0.000	2.155537	2.1974

The regression shows that it seems to satisfy the first requirement since it is poorly correlated with wages (black and white tv's are relatively cheap now so that ownership is more a matter of taste than income).

However, when you instrument education with tv ownership the IV estimates are now very large and insignificant. This does not seem very sensible. (note sample size is large so might expect asymptotic property of IV consistency to hold.)

```
. ivreg lhw (years=ed=bw) if lhw>1 & sex==1 & e(sample)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS			
Model	-75.3525064	1	-75.3525064	Number of obs =	3174	
Residual	1061.27179	3172	.334574966	F(1, 3172) =	0.32	
Total	985.919285	3173	.310721489	Prob > F =	0.5732	
				R-squared =	.	
				Adj R-squared =	.	
				Root MSE =	.57842	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lhw						
years=ed	.1109609	.1969606	0.56	0.573	-.275222	.4971439

```

      _cons |      .8772044    2.310431    0.38    0.704    -3.652886    5.407295
-----+-----
Instrumented:  yearsed
Instruments:   bw
-----+-----

```

Note also no R^2 reported in IV estimation – because can't decompose TSS into $\beta^2\text{Var}(X) + \text{Var}(u)$ when $\text{Cov}(X,u)$ non-zero)

The reason is that video ownership is hardly correlated with education as the 1st stage of the 2SLS IV regression below shows.

```
. reg yearsed bw if e(sample)
```

Source	SS	df	MS			
Model	8.62451882	1	8.62451882	Number of obs =	3174	
Residual	34712.5198	3172	10.9434173	F(1, 3172) =	0.79	
Total	34721.1443	3173	10.9426865	Prob > F =	0.3747	
				R-squared =	0.0002	
				Adj R-squared =	-0.0001	
				Root MSE =	3.3081	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearsed						
bw	.1498561	.1688043	0.89	0.375	-.1811206	.4808329
_cons	11.7092	.0633481	184.84	0.000	11.585	11.83341

Note that in this case the F test is less than 1 (and p value on the variable is .38).

It is good practice therefore to always check these two criteria before you present the results of IV estimation.

In large samples it is generally more efficient to use more than one instrument. In small samples it is often better to use a minimum number of instruments, since the bias in 2SLS is proportional to the extent of over-identification, $(L-k)$.

Testing

No estimation should be complete without a thorough testing of the model specification. This means using the battery of t, F tests that can be used to test hypotheses on single coefficients or subsets of coefficients. Always make sure you can give the economic meaning of any coefficients of interest as well as its sign and statistical significance. Can the coefficients be interpreted as elasticities? If so say what the elasticity is. This also means checking the nature of the residuals to see whether your standard errors are biased, (heteroskedasticity if working with cross-section data, autocorrelation if working with time series data) and fixing the standard errors up if problems are revealed. Examples of which tests to use and how can be found in your QM2 notes and exercises.

Testing also means applying the model to different subsets (time periods, subsets of individuals) to see whether the results are robust across the different groups/periods.

Your final tables should generally consist of more than one specification demonstrating the robustness of your results to changes in the number and/or type of right hand side variables

At the bottom of the table it is good practice to include a set of diagnostic statistics (sample size, R^2 , F test of goodness of fit of the model are the usual ones)

TABLE 3(b)—continued

Independent variable	Sample mean	Model		
		1	2	3
Wage rate	4.66	0.11 (0.24)	0.18 (0.21)	0.02 (0.22)
Unemployment rate	2.13	-0.12 (0.10)	-0.21* (0.09)	0.01 (0.11)
Selectivity correction	3.49	0.12 (0.09)	0.18** (0.09)	0.15* (0.09)
Claimant × duration 3-12 months	0.22	—	—	-0.14 (0.13)
Claimant × duration 12-24 months	0.10	—	—	0.14 (0.15)
Claimant × duration 24+ months	0.12	—	—	0.24* (0.14)
Claimant × vacancy rate	0.38	—	—	0.34* (0.20)
Claimant × wage rate	2.43	—	—	0.37** (0.13)
Claimant × unemployment rate	1.13	—	—	-0.45** 0.15
Diagnostics				
\bar{R}^2		0.083	0.051	0.060
Standard error		1.418	1.279	1.274
F-value ($k, N - k$)		10.570	6.329	6.004
Mean of dependent variable		2.718	2.116	2.116
Sample size		2963	2963	2963

Note: Standard errors in parentheses ** indicates significance at 95% level; * indicates significance at 90% level 2-tailed *t*-test.

Natural Experiments as Instruments

One of the principal aims which underlies much empirical work is the need to establish the “treatment effect” of a policy intervention. Usually this means comparing outcomes for those affected by an event, (eg teenage motherhood), - called the “treatment” group, and those not, (the “control” group). One issue that has attracted a lot of attention recently is whether these comparisons can ever measure the “effect of treatment on the treated” ie do the control group, (non-teenage mothers), represent a suitable counterfactual comparator group which would adequately represent the behaviour of the treatment group (teenage mothers) if they had not experienced the event. If there are systematic differences between treatment and control groups then a simple comparison of the behaviour of the two will give a biased estimate of the “effect of treatment on the treated” – the coefficient b in .

$$\text{LnW} = a + b \cdot \text{Treatment Dummy} + \gamma X$$

The idea then is to try and purge the regression estimate of all these potential behavioural and environmental differences. If you believe that with a suitably large number of right hand side control variables, X , you can reasonably capture all the possible heterogeneity between the treatment and control groups, then the treatment effect is said to be “ignorable”. If you believe that there is likely to be some unobserved heterogeneity not captured by these controls, then you need to find an instrument that predicts whether someone will be observed in the control or the treatment group and is at the same time uncorrelated with the dependent variable of interest.

There has been a large drive in recent years to try and find instruments to deal with omitted variable bias based on the outcome of “natural experiments” – where government policy changes or changes to the general environment affect certain groups but not others. Examples of each, taken from the US literature, are a) the vietnam war draft – which assigned a random number to all men of fighting age from which individuals were chosen on the basis of their randomly chosen draft number. Since the number were random, it seems reasonable to compare the subsequent outcomes of draftees with others b) quarter of birth as an instrument for years of schooling. Since the US school year begins in January, all those born toward the beginning of the year receive more schooling than those who are not. Quarter of birth is therefore correlated with schooling, but should be uncorrelated with earnings (net of schooling).

A related literature concerns “difference in difference” estimation. This technique can also be applied to the evaluation of policy interventions or (medical) experiments. Here an event (the “treatment”) occurs between time t and t+1 that affects the potentially endogenous rhs variable, x, but is exogenous to the dependent Y variable. Typically this event will affect a subset of the x population only, (eg teenagers if x=age). The idea is then to compare the change in Y for the treatment group who experienced the shock (subset t) with the change in Y of the control group who did not, (subset c).

$$[Y_t^2 - Y_t^1] - [Y_c^2 - Y_c^1]$$

This is the raw difference in difference estimator. Note that this approach does not need regressions conditioning on a set of X variables. The implicit assumption here is that all other X variables changed to the same extent and have the same effect on the Y variable, so that differencing removes this effect., $[X_t^2 - X_t^1] - [X_c^2 - X_c^1] = 0$ if $X_k^1 = X_k^2$

This is quite a strong assumption to make and requires the careful consideration of a comparable control group.

In practice this estimator can be obtained from pooled cross-section data or panel data from 2 or more periods – one observed before a program was implemented and the other in the period after, the following regression.

$$\ln W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1$$

$$\ln W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2$$

The coefficients b_1 and b_2 give the differential impact of the treatment group on wages in each period and so the difference between the coefficients gives the “difference in difference” estimator – the change in the treatment effect following an intervention.

Note however that there is no standard error associated with this method. However this can be obtained by pooling the estimates and running the following regression.

$$\ln W = a + a_2 \text{Year}_2 + b_1 \text{Treatment Dummy} + b_2 \text{Year}_2 * \text{Treatment Dummy}$$

Where now a is the average wage of the control group in the base year, a_2 , is the average wage of the control group in the second year, b_1 gives the difference on wages between the treatment and control group in the base year and b_2 is the “difference in difference” estimator – the additional change in wages for the treatment group relative to the control group in the second period.

If Year₂=0 and Treatment Dummy = 0, LnW = a
 If Year₂=0 and Treatment Dummy = 1, LnW = a + b₁
 If Year₂=1 and Treatment Dummy = 0, LnW = a + a₂
 If Year₂=1 and Treatment Dummy = 1, LnW = a + a₂ + b₁ + b₂

So the change in wages for the treatment group is (a + a₂ + b₁ + b₂) – (a + b₁) = a₂ + b₂
 and the change in wages for the control group is (a + a₂) – (a) = a₂

so the “difference in difference” estimator = Change in wages for treatment – change in wages for control = (a₂ + b₂) - (a₂) = b₂

Example: In April 2000 the UK government introduced the Working Families Tax Credit aimed at increasing the income in work relative to out of work for groups of traditionally low paid

Individuals. If successful the scheme could have been expected to increase the hours worked of those who benefited most from the scheme- namely single parents. By comparing hours of worked for this group before and after the change with a suitable control group, it should be possible to obtain a difference in difference estimate of the policy effect. The example uses other single women as a control group.

```
. tab year, g(y)                                /* set up year dummies */
```

year	Freq.	Percent	Cum.
98	29399	50.46	50.46
2000	28868	49.54	100.00
Total	58267	100.00	

```
. g lonepy2=lonep*y2                            /* create interaction variable */
```

```
. reg hours lonep if year==98
```

Source	SS	df	MS	Number of obs =	29026
Model	1159891.90	1	1159891.90	F(1, 29024) =	3041.43
Residual	11068703.6	29024	381.363824	Prob > F	= 0.0000
Total	12228595.5	29025	421.312507	R-squared	= 0.0949
				Adj R-squared	= 0.0948
				Root MSE	= 19.529

hours	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lonep	-13.14152	.2382905	-55.15	0.000	-13.60858 -12.67446
_cons	27.88671	.1436816	194.09	0.000	27.60509 28.16834

```
. reg hours lonep if year==2000
```

Source	SS	df	MS	Number of obs =	28369
Model	969891.29	1	969891.29	F(1, 28367) =	2905.13
Residual	9470465.62	28367	333.855029	Prob > F	= 0.0000
Total	10440356.9	28368	368.032886	R-squared	= 0.0929
				Adj R-squared	= 0.0929
				Root MSE	= 18.272

hours	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lonep	-12.10205	.2245309	-53.90	0.000	-12.54214 -11.66195

```
_cons | 26.56678 .1368139 194.18 0.000 26.29861 26.83494
```

The coefficient on lone parents gives the difference in average hours worked between lone parents and the control group. So comparing the lone parent coefficient across both periods, lone parents worked 13 hours less than other single women in 1998 before the policy, (27.9-13.1 = 14.8 hours for single parents on average) and 12 hours less than other single women immediately after the introduction of WFTC, (26.6-12.1 = 14.5 hours for lone parents in 2000, on average). So the change (difference in difference)

$$\begin{aligned}
 &= -13.1 - (-12.1) = 1.0 \\
 &= (\text{Hours}^{\text{LonePar}}_{2000} - \text{Hours}^{\text{LonePar}}_{1998}) - (\text{Hours}^{\text{Single}}_{2000} - \text{Hours}^{\text{Single}}_{1998}) \\
 &= (14.5 - 14.8) - (26.6 - 27.9) = -0.3 - (-0.7) = 1.0
 \end{aligned}$$

Which suggests lone parents worked relatively about 1 hour more as a result of the policy. (Note that hours worked actually fall for both groups, they just fall less for lone parents).

To obtain standard errors, pool the data and estimate the following

```
. reg hours y2 lonep lonepy2
```

Source	SS	df	MS			
Model	2145163.25	3	715054.418	Number of obs = 57395		
Residual	20539169.2	57391	357.881362	F(3, 57391) = 1998.02		
				Prob > F = 0.0000		
				R-squared = 0.0946		
				Adj R-squared = 0.0945		
				Root MSE = 18.918		
Total	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hours						
y2	-1.319938	.1985909	-6.65	0.000	-1.709177	-.9306989
lonep	-13.14152	.2308375	-56.93	0.000	-13.59396	-12.68908
lonepy2	1.039477	.3276099	3.17	0.002	.3973598	1.681594
_cons	27.88671	.1391877	200.35	0.000	27.6139	28.15952

So now the coefficient on the interaction term, lonep*Year2, is the difference in difference estimator = 1.03 and the associated standard error suggests that this effect is statistically significant.

The constant gives hours worked by the control group in 1998. The coefficient, y2, says that hours worked for the control group were 1.3 hours lower in 2000. The coefficient, lonep, confirms that hours worked were around 13.1 lower than the control group in 1998.

Control variables

It is usually a good idea to try and account for observable differences between the treatment and control groups. In the example above, inspection of the sample means shows that lone parents are generally younger than other single women.

```
su age if lonep==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	21289	22.91883	7.532616	16	63

```
. su age if lonep==0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	36714	29.52571	10.09665	16	64

Since older workers generally work longer hours if we did not account for differences in age across the groups we may wrongly attribute some of this effect to the difference in difference estimator.

```
. reg hours y2 lonep lonepy2 age
```

Source	SS	df	MS			
Model	2286701.31	4	571675.328	Number of obs =	57395	
Residual	20397631.2	57390	355.421348	F(4, 57390) =	1608.44	
Total	22684332.5	57394	395.238744	Prob > F =	0.0000	
				R-squared =	0.1008	
				Adj R-squared =	0.1007	
				Root MSE =	18.853	

hours	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
y2	-1.368384	.197922	-6.91	0.000	-1.756312	-.9804557
lonep	-11.98638	.237214	-50.53	0.000	-12.45132	-11.52143
lonepy2	.9716015	.3264997	2.98	0.003	.3316604	1.611543
age	.1700081	.0085193	19.96	0.000	.1533102	.186706
_cons	22.8912	.2861919	79.99	0.000	22.33026	23.45214

Including age as a control variable therefore reduces the difference in difference estimate, a little (because age and lone parents are negatively correlated omitting age biases up the lone parent effect)

Panel Data

Since unobserved or omitted variables are often responsible for endogeneity and inconsistent OLS estimates, the availability of data sets that follow the same set of agents (individuals, firms) over time can be used to remove the influence of *unobservables* from regressions and so produce consistent estimates. In addition panel data can be used to answer questions that cross-section data alone can not. For example is a 10% unemployment rate caused by 10% of the population being permanently unemployed or by everyone in the population having a 10% chance of being unemployed at some point? Only by following individuals over time can this issue be solved.

Given

$$Y_{it} = a_t + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

The idea is to somehow purge the equation of the individual-specific component ϕ_i . One simple approach is to assume that ϕ_i is constant over time (a *fixed effect*), so that by differencing the above the unobservable effect disappears and an OLS regression on the 1st difference gives consistent estimates of b_1 .

$$[Y_{i2} - Y_{i1}] = (a_2 + b_1 X_{i2} + \phi_i + e_{i2}) - (a_1 + b_1 X_{i1} + \phi_i + e_{i1})$$

so
$$\Delta Y = \delta + b_1 \Delta X + \Delta e \quad (\text{if } a_1 = a_2)$$

(note that this techniques also removes *any* variable that stays constant over time. It is also important to remember to remove the first observation for each cross-section unit in the data after differencing). With more than two time periods, then subtract data on each

unit at time 1 from data on the same unit at time 2 *and* subtract data on each unit at time 2 from data on the same unit at time 3 and then pool these differenced observations. The easiest way to do this is to sort your data by individual unit and then time, (using a command like `sort idcode time` in Stata), so that the 1st observation in the data is the 1st unit at time 1, the 2nd observation is the 1st unit at time 2 etc. It is also normal to include a constant in these differenced regressions, (even though differencing removes all constants). The way to interpret the constant is that it represents the *change* in the value of the intercept over time, ie $a_1 \neq a_2$. Remember also that the absence of a constant in a regression no longer restricts the R^2 coefficient to lie between 0 and 1.

Panel data can be useful addition to the problem of **policy evaluation** outlined above. If the same agents appear in the data before and after an event then the difference in difference estimator will also net out any fixed effects that might otherwise influence the results. (Note that this also applies if the data are pooled and year dummy/policy interactions used instead, since the fixed effects drop out in this formulation).

Example.

Using the *train.dta* dataset, (the data is taken from the wooldridge website), you wish to investigate the effect workforce training grants on the rate at which firm's scrap defective items – presumably more training should lead to fewer scrap rates). Since differences in scrap rates could be due to unobserved fixed effects you decide to rid your estimates of fixed effects, by either

a) 1st differencing

```
. reg dscrap dgrant if year==1988
```

Source	SS	df	MS			
Model	6.73345587	1	6.73345587	Number of obs =	54	
Residual	298.400031	52	5.73846213	F(1, 52) =	1.17	
Total	305.133487	53	5.7572356	Prob > F =	0.2837	
				R-squared =	0.0221	
				Adj R-squared =	0.0033	
				Root MSE =	2.3955	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dscrap						
dgrant	-.7394436	.6826276	-1.08	0.284	-2.109236	.6303488
_cons	-.5637143	.4049149	-1.39	0.170	-1.376235	.2488068

or b) pooling the data over two years and adding a second period year dummy and an interaction of the year dummy with a dummy for whether the firm received a grant.

```
. reg scrap treatfm y88treat y88
```

Source	SS	df	MS			
Model	36.6138737	3	12.2046246	Number of obs =	108	
Residual	4060.68502	104	39.0450483	F(3, 104) =	0.31	
Total	4097.2989	107	38.292513	Prob > F =	0.8162	
				R-squared =	0.0089	
				Adj R-squared =	-0.0197	
				Root MSE =	6.2486	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
scrap						
mcont	-.4085865	1.780611	-0.23	0.819	-3.939605	3.122432
y88cont	-.7394436	2.518164	-0.29	0.770	-5.733057	4.25417

y88		- .5637143	1.493702	-0.38	0.707	-3.525781	2.398353
_cons		4.755429	1.056207	4.50	0.000	2.660931	6.849926

The 2 estimation techniques give identical results.

One problem with the 1st differencing approach is that it can generate autocorrelation in the differenced error term. It may also make it harder to assume that the (differenced) X variables are uncorrelated with (differenced) errors – if policy responds to changes in observables for example. One solution is to include more time-varying variables that could account for the autocorrelation, (which often stems from missing variables in an equation).

There are at least two other ways of obtaining fixed effects estimates of b_1 .

The first is to pool the data across years and estimate (1) directly by including a dummy variable for each individual in the data to capture the fixed effect, (*least squares dummy variables*). This may be rather wasteful of degrees of freedom and will usually produce inconsistent estimates of the dummy variables (ie the fixed effects) if the time dimension of the panel is small, (which is usually the case).

For these reasons, the most-commonly used alternative method – which can be used to obtain consistent estimates of the fixed effects if desired - is to calculate the mean value for each observation for each individual and subtract the observation at time t from this mean.

$$\bar{Y}_i = a_1 + b_1 \bar{X}_i + \phi_i + \bar{e}_i$$

$$Y_{it} - \bar{Y}_i = b_1 (X_{it} - \bar{X}_i) + (e_{it} - \bar{e}_i)$$

This *within-group estimator* approach also removes the fixed effect, (because the mean of the fixed effect is the same as the individual fixed effect value), and avoids the problem of introducing autocorrelation into the residuals, (since the mean value of the residual should be zero).

Problems with this approach arise if there is a variation in the X variables across individuals, but less variation over time. Differencing will not produce an estimate for any X variable that is constant. Likewise a within-groups or least squares dummy variable estimators. Even for variables that do vary a little over time, inclusion of fixed effects will produce estimates on the X variables that are close to zero. The fixed effect picks up the possibly true impact of variables that move only a little over time.

Within-Groups or First Difference?

While within-groups does reduce problems of autocorrelation, it is easier to get heteroskedastic adjusted standard errors using first differences, (just use the `, robust` option at the end of a regression command line in Stata). Within-groups estimates can also be quite sensitive, (though remain consistent), with large T and small N dimensions to the panel. In practice it is probably better to do both to test the sensitivity of the results.

Note that the two methods will produce identical estimates when there are 2 time periods in the data.

Example

Consider the 1st difference regression in a 2 year panel, (*panel2.dta*) of the change in log sales on the change in log employment, (the `noconst` option in Stata removes the constant from the regression).

```
. reg clsales clempl if year==1988, noconst
```

Source	SS	df	MS			
Model	6.61023185	1	6.61023185	Number of obs =	115	
Residual	23.5097718	114	.206226068	F(1, 114) =	32.05	
-----				Prob > F =	0.0000	
Total	30.1200036	115	.261913075	R-squared =	0.2195	
-----				Adj R-squared =	0.2126	
				Root MSE =	.45412	
clsales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
clempl	.8395683	.1482926	5.66	0.000	.5458018	1.133335

Now the within-group estimate gives identical slope estimate

```
. xtreg lsales lempl if year<1989, fe i(fcode)
```

```
Fixed-effects (within) regression      Number of obs   =      230
Group variable (i) : fcode             Number of groups =      115

R-sq:  within = 0.2195                 Obs per group:  min =       2
        between = 0.6957                avg   =       2.0
        overall = 0.6704                max   =       2

corr(u_i, Xb) = 0.0794                 F(1,114)        =      32.05
                                         Prob > F         =      0.0000
```

lsales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lemploy	.8395692	.1482924	5.66	0.000	.5458032	1.133335
_cons	12.06415	.5150927	23.42	0.000	11.04375	13.08454

sigma_u	.60128965					
sigma_e	.32111175					
rho	.77809084 (fraction of variance due to u_i)					

```
F test that all u_i=0:      F(114, 114) =      6.97      Prob > F = 0.0000
```

However when the data are extended to 3 time periods, the results no longer coincide.

```
. reg clsales clempl if year==1988 | year==1989, noc
```

Source	SS	df	MS			
Model	11.1708402	1	11.1708402	Number of obs =	235	
Residual	38.7381982	234	.165547856	F(1, 234) =	67.48	
-----				Prob > F =	0.0000	
Total	49.9090384	235	.212378887	R-squared =	0.2238	
-----				Adj R-squared =	0.2205	
				Root MSE =	.40688	
clsales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
clempl	.7845282	.0955053	8.21	0.000	.5963681	.9726882

```
. xtreg lsales lempl, fe i(fcode)
```

```

Fixed-effects (within) regression                Number of obs      =       345
Group variable (i) : fcode                      Number of groups   =       115
R-sq:  within = 0.3029                          Obs per group: min =         3
        between = 0.7162                          avg               =       3.0
        overall = 0.6901                          max               =         3
                                                F(1,229)          =      99.52
corr(u_i, Xb) = 0.1611                          Prob > F           =      0.0000
-----+-----
      lsales |          Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
      lemploy |   .8092035   .0811145     9.98  0.000    .6493773   .9690297
      _cons   |  12.19437   .2850389    42.78  0.000   11.63273   12.756
-----+-----
      sigma_u |   .58411543
      sigma_e |   .28530622
      rho     |   .80737938   (fraction of variance due to u_i)
F test that all u_i=0:      F(114, 229) =    12.25      Prob > F = 0.0000

```

References and Guides

As always it is always a good idea to keep in touch with ideas and programs that other people have written. The following web sites are useful

<http://www.nber.org>

<http://ideas.uqam.ca/ideas/data/bocbocode.html>

<http://www.stata.com>

Self-help type articles/books

Angrist, J. and Kreuger, A., 'Instrumental Variables and the Search for Identification', *Journal of Economic Perspectives*, 2002, also <http://www.nber.org/papers/w8456>

Angrist, J. and A. Kreuger 'Empirical Strategies in Labor Economics' in the Handbook of Labor Economics, (eds.) O. Ashenfelter and D. Card, Vol. 3A, Elsevier Press

<http://www.irs.princeton.edu/wpframe.html>

Blundell, R. and M. Costa Dias, (2000), 'Evaluation Methods for Non-Experimental Data', Fiscal Studies, Vol. 21, No. 4, pp.427-468

<http://www.ifs.org.uk/publications/fiscalstudies/fsabs21dias.shtml>

Deaton, A., (1997), 'The Analysis of Household Surveys', Johns Hopkins University Press.

DiNardo, J. and T. Lemieux, 'Diverging Male Wage Inequality in the U.S. and Canada 1981{1988: Do Institutions Explain the Difference?', Industrial and Labor Relations Review, July 1997.

Hamermesh, D., (1999), 'The Art of Labormetrics', NBER Working Paper No. 6927

<http://www.nber.org/papers/w6297>

Kennedy, P., (2002), Sinning in the Basement: The ten Commandments of Applied Econometrics, Journal of Economic Surveys Vol. 16, No. 4, p. 569-621

General Econometric Textbooks

C. Dougherty, "Introduction to Econometrics 2nd Edition", Oxford University Press, (Library Code: 330.01 DOU)

D. Gujarati, "Basic Econometrics", McGraw-Hill Press, (Library Code: 330.01 GUJ)
 Wooldridge, J., *Introductory Econometrics*, South Western College Press, 2000