Statistical Reasoning

Week 12

Sciences Po - Louis de Charsonville

Spring 2018

Outline

Research Paper

Timeline

Final draft	1^{st} **May**

Review of regression

Regression models produce **fitted** (predicted) values and residuals that hold the unexplained variance for each data point. Issues that arise in that context are :

- unreliable coe icients due to **multicollinearity**, i.e. interactions between independent variables
- unreliable significance tests due to **heteroskedasticity**, i.e. heterogeneous variance in the residuals
- unreliable predictions due to **outliers and influential points** in the data that either do not fit or 'overfit' the model

Note : The model still assumes a linear, additive relationship between Y and $X_1, X_2, \ldots X_k$. That assumption can also be violated among other matters.

The model also **fits** a linear function to the data, of the form :

$$Y = \alpha + \beta_1 X_1 + beta_2 X_2 + \cdots + beta_k X_k + \epsilon \quad (1)$$

where :

- Y is the **dependent variable** (response)
- X is a vector of **independent variable** (predictors)
- α is the constant
- $\beta_1 X_1 + beta_2 X_2 + \cdots + beta_k X_k$ is a vector of regression coefficients
- ϵ is the **error term**

```
reg births schooling log_gdpc
```

The `reg` command can take any number of continuous variables as arguments, and shows **unstandardised** coefficients by default, using their original metric and possible transformation :

```
. reg births schooling log_gdpc

      Source |       SS           df       MS      Number of obs   =        86
-------------+----------------------------------   F(2, 83)        =     88.51
       Model |  150.301883         2  75.1509417   Prob > F        =    0.0000
    Residual |   70.475313        83  .849100157   R-squared       =    0.6808
-------------+----------------------------------   Adj R-squared   =    0.6731
       Total |  220.777196        85  2.59737878   Root MSE        =    .92147

------------------------------------------------------------------------------
      births |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   schooling |  -.1976117   .0724595    -2.73   0.008    -.3417306   -.0534927
    log_gdpc |  -.4703416   .1324501    -3.55   0.001    -.7337796   -.2069036
       _cons |   7.950304   .6861182    11.59   0.000     6.585642    9.314965
------------------------------------------------------------------------------
```

```
reg births schooling log_gdpc, beta
```

The `beta` option provides **standardised coefficients**, which use the standard deviation of regressors (or predictor, i.e. the independent variables) in order to provide coefficients with comparable units :

births	Coef.	Std. Err.	t	P>|t|	Beta
schooling	-.1976117	.0724595	-2.73	0.008	-.3686479
log_gdpc	-.4703416	.1324501	-3.55	0.001	-.4800156
_cons	7.950304	.6861182	11.59	0.000	.

```
reg births schooling i.region
```

Categorical variables can be used as **dummies**, i.e. binary recodes of each category that are tested against a reference category to provide regression coefficients for net effect of that category alone :

births	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
schooling	-.0415563	.0639718	-0.65	0.518	-.1688888	.0857763
log_gdpc	-.742187	.1380037	-5.38	0.000	-1.016876	-.4674975
region						
2	-.6523485	.5803126	-1.12	0.264	-1.807432	.5027349
3	.3682404	.254364	1.45	0.152	-.1380585	.8745393
4	1.411177	.2486027	5.68	0.000	.9163457	1.906008
5	1.167491	.337383	3.46	0.001	.4959471	1.839035
_cons	8.315004	.8006456	10.39	0.000	6.721359	9.908649

Instrumental Variables

Motivation

- Some variables might be *unobserved*.
- OLS is inconsistent under omitted variables (Week 10).
- Omitted variables bias can be mitigated using **proxy variable** for the unobserved variable.
- Suitable proxy variable are not always available.
- When treatment is not randomly assigned, the causal effect of the treatment cannot be recovered from simple regression methods

Example

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 ability + \epsilon \tag{2}$$

- $ability$ is unobserved
- no proxy $\rightarrow$ $log(wage) = \beta_0 + \beta_1 educ + u$
- u contains $ability$ and β_1 is biaised if $educ$ and $ability$ are correlated.

Simple OLS model

$$log(wage) = \beta_0 + \beta_1 educ + \epsilon \quad (3)$$

```
. reg lwage educ

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(1, 426)       =     56.93
       Model |  26.3264237         1  26.3264237   Prob > F        =    0.0000
    Residual |  197.001028       426  .462443727   R-squared       =    0.1179
-------------+----------------------------------   Adj R-squared   =    0.1158
       Total |  223.327451       427  .523015108   Root MSE        =    .68003

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1086487   .0143998     7.55   0.000     .0803451    .1369523
       _cons |  -.1851969   .1852259    -1.00   0.318    -.5492674    .1788735
------------------------------------------------------------------------------
```

- One additional year of education is associated with earnings 11% higher.
- **Bias** : Self-selection into education → individuals who have the most to gain from education are the most likely to stay.
- Ability is unobserved and is correlated with both education and wages.
- OLS estimates are not consistent.

Solutions

- One additional year of education is associated with earnings 11% higher.
- **Bias** : Self-selection into education → individuals who have the most to gain from education are the most likely to stay.
- Ability is unobserved and is correlated with both education and wages.
- OLS estimates are not consistent.

Solutions

- Randomized control trial (RCT) : allocate education randomly to individuals and observe the difference in their wages.
  - However : RCT is infeasible on ethical grounds.
- Quasi-natural experiments can alter individuals choices and can be used as instruments.

A **valid instrument** (or instrumental variable, IV) is :

1. Significantly correlated with the treatmenf of interest (**instrument relevance**)
2. Only affect the outcome via its effect on the treatment (exclusion restriction or **instrument exogeneity**)

Formally :

$$y = \alpha + \beta x + \epsilon \tag{4}$$

z is a valid instrument if :

1. Instrument relevance $\Leftrightarrow Cov(z,x) \neq 0$
2. Instrument exogeneity $\Leftrightarrow Cov(z,\epsilon) = 0$

While we can test whether the first condition is satisfied the second condition cannot be tested.

IV - Example 5/

Examples of instruments ?

- IQ (Intelligence Quotient) ?
- Mother's education ?
- Number of siblings ?
- Legislative change increasing number of minimum schooling

Example 1 - Father's education

- Assume father's education is uncorrelated with ϵ
- We can check father's education is indeed correlated with education

```
reg educ fatheduc if !mi(lwage)

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(1, 426)       =     88.84
       Model |  384.841983         1  384.841983   Prob > F        =    0.0000
    Residual |  1845.35428       426  4.33181756   R-squared       =    0.1726
-------------+----------------------------------   Adj R-squared   =    0.1706
       Total |  2230.19626       427  5.22294206   Root MSE        =    2.0813

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fatheduc |   .2694416   .0285863     9.43   0.000     .2132538    .3256295
       _cons |   10.23705   .2759363    37.10   0.000     9.694685    10.77942
------------------------------------------------------------------------------
```

Example 1 - Father's education

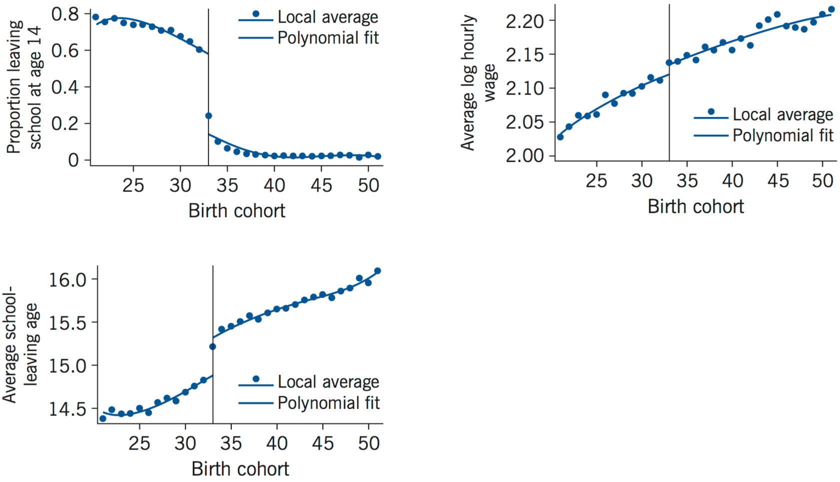
- We use father's education as a IV for educ :

```
. ivreg lwage (educ = fatheduc)

Instrumental variables (2SLS) regression

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(1, 426)       =      2.84
       Model |  20.8673618         1  20.8673618   Prob > F        =    0.0929
    Residual |  202.460089       426  .475258426   R-squared       =    0.0934
-------------+----------------------------------   Adj R-squared   =    0.0913
       Total |  223.327451       427  .523015108   Root MSE        =    .68939

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0591735   .0351418     1.68   0.093    -.0098994    .1282463
       _cons |   .4411035   .4461018     0.99   0.323    -.4357311    1.317938
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   fatheduc
------------------------------------------------------------------------------
```

Example 2 - Leglislative change in number of mandatory schooling

- In 1947, a legislative change in the UK increased the minimum school leaving age from 14 to 15
- Children who wanted to leave school at 14 are prevented from doing so and have to do one additional year of schooling.
- Let assume :
  - children under the two legislations are similar
  - Children face similar labor market conditions
- Quasi-natural experiment : independent of their ability, some individuals will need to stay one more year in schooling.
- Instrument variable : binary variable for being affected by the reform.

Example 2 - Leglislative change in number of mandatory schooling



Example 2 - Leglislative change in number of mandatory schooling

- Impact of the IV (the reform) on the treatment (education) (1^{st} stage) :
  - Reform increased the average years of schooling for men by 0.397 years
- Impact of the IV on the dependent variable (wages) (Reduced-form estimate)
  - Reform increased wages by 1.2%
- IV estimates is $\frac{0.012}{0.937} = 0.03$ or 3% (Wald estimates).

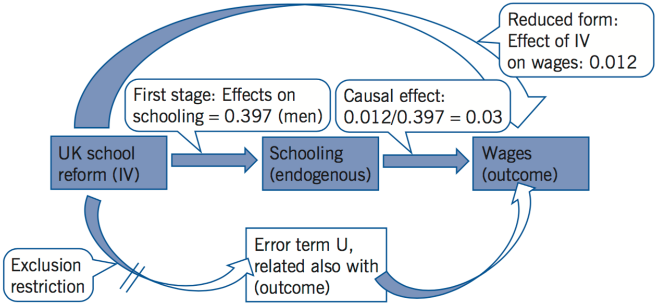
Example 2 - Leglislative change in number of mandatory schooling

1. If the reform has an effect on education
2. If the reform affects wages *exclusively* through its effect on education

⇒ The IV estimates can be interpreted as the **causal effect of the treatment on the outcome.**

Schematic depiction of IV estimation

Example 2 - Leglislative change in number of mandatory schooling



Note : A UK reform that increased minimum school leaving age is used as the Instrumental variable (IV) ; it should affect the outcome only via its effect on the endogenous variable but not in other ways

Credits : Sascha O. Becker, University of Warwick

- Causal relationship of interest :

$$Y = \alpha + \beta X + \epsilon$$

- First-Stage regression :

$$X = \eta + \gamma Z + u$$

- Second-Stage regression :

$$Y = \mu + \rho \hat{X} + \nu$$

- Reduced form :

$$Y = \delta + \phi Z + \upsilon$$

- **Wald estimate** is the ratio of the reduced form estimate and the first stage estimate
- Can be easily computed when the instrument takes only two values
- In general case, a "two stage least squares" (**2SLS**) estimate will be computed
- Only the variation in the treatment coming from the instrument is used to explain the variance in the outcome.

Difficulties

- Finding a valid instrument
- Interpreting the results

1. Relevance

- Correlation between the instrument and the change in treatment allocation is strong.
- **Weak instruments** = instruments that are only *weakly* correlated with the treatment.
- Weak instruments induce a bias that can be larger than the bias of the OLS estimates.

2. Exclusion restriction

- Cannot be statistically tested
- Need to be supported by a convincing narrative

Intepreting IV results can be difficult...

Why is the IV estimate much lower than OLS estimate ?

Intepreting IV results can be difficult...

Why is the IV estimate much lower than OLS estimate ?

- OLS estimate describes the average difference in earnings for those whose education differs by one year
- IV estimate is the effect of increasing education *only* for the population whose choise of the treatment was *affected* by the instrument.
- Such effect is known as **Local Average Treatment Effect (LATE)**
- In this case, treatment effects are heterogeous.
- For IV to estimate LATE, another assumption need to be satisfied :
  - While the instrument may have no effect on some people, all those who are affected are affected in the same way.
    Monotonicity assumption

Some LATE's specific jargon :

- **Always-taker** : They always take the treatment independently of the IV.
- **Compliers** : Their treatment status is affected by the instrument in the right direction.
- **Never-takers** : They never take the treatment independently of IV.
- **Defiers** : Their treatment status is affected by the instrument in the "wrong" direction.

$\Rightarrow$ **Monotonicity ensures that there are no defiers.**

- With defiers, effects on compliers could be party cancelled out by opposite effects on defiers
- Reduced form effect could be close to O although treatment effets are positive for everyone (but the compliers are pushed in one direction by the instrument and the defiers in the other direction)

Example

		Old regime $Educ = 14$	$Educ \geq 14$
New regime	$Educ = 14$	Never-taker	Defier
	$Educ \geq 14$	Complier	Always-taker

IV - Wrapping up

Pros

- IV are useful to address :
  - Omitted variable bias
  - Measurement error
  - Simultaneity or reverse causality

Cons

- Finding valid instrumental variables that affect treatment but do not have a direct effect on the outcome is difficult.
- Estimated treatment effects do not generally apply to the whole population
- Estimated treatment effects may vary across different instruments.
- In case of "weak" instruments, instrumental variable estimates are biased.

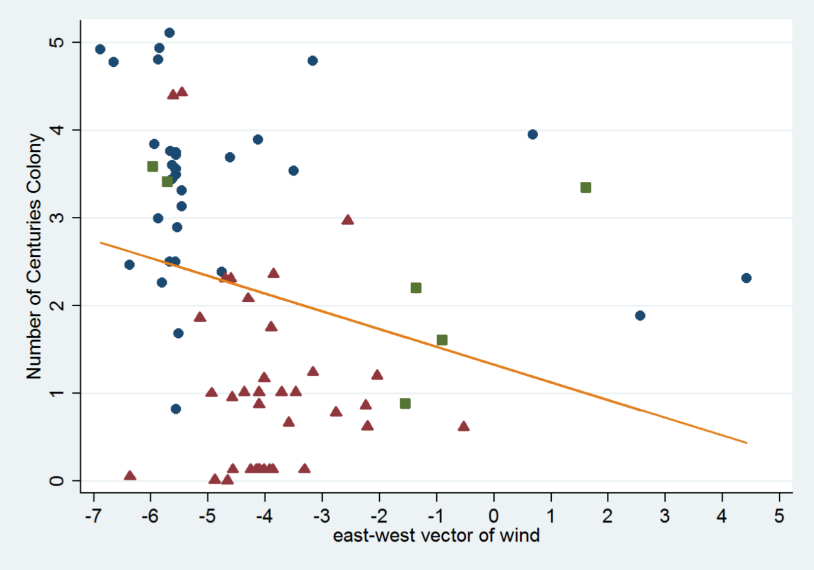
Example of IV in the litterature 1/2

Institutions and prosperity

- In rich economies institutions (rules that govern society) function rather well on the whole while in poor ones they don't.
- Is good institutions a cause of economics progress or a consequence ?
- Find an IV which is link to institutions but not to economic success.

Institutions and prosperity

- In rich economies institutions (rules that govern society) function rather well on the whole while in poor ones they don't.
- Is good institutions a cause of economics progress or a consequence ?
- Find an IV which is link to institutions but not to economic success.
- Feyrer and Sacerdote (2006) uses **winds and currents** as an IV.
- Early colonists went where their sails took them. Some islands were colonized earier because there lay on natural sailing routes


Number of Centuries Colony
0
1
2
3
4
5
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
east-west vector of wind

Example of IV in the litterature 1/2

Institutions and prosperity - Findings

- A robust positive relationship between the years of European colonialism and current levels of income : a century as a colony is worth a 40% increase in today's GDP.
- Years under US and Dutch colonial rule are significantly better than years under the Spanish and Portuguese.
- Later years of colonialism are associated with a much larger increase in modern GDP than years before 1700.

Review of Instructions

Univariate statistics

- Introduction
- Datasets
- Distribution
- Estimation

Bivariate statistics

- Significance
- Crosstabs
- Correlation
- Regression

Statistical modelling

- Basics
- Extensions
- Diagnostics
- Conclusion

Instructions

Tablets of Stones

1. Interpret your results
2. Reference your sources
3. Proofread your work


THIS
IS
STATA
Warner Bros

Thank you

`exit, clear`

Credits

- Francois Briatte & Ivaylo Petev, Stata Guide
- Urdan, Statistics in Plain English
- Jeffrey M. Wooldridge, Introductory Econometrics : A Modern Approach, 5e Ed.
- Marcelo Coca Perraillon, Health Services Research Methods I, University of Colorado
- Michael Visser, Econometrics I, ENSAE ParisTech
- Sasha Becker, Using instrumental variables to establish causality
- Fabian Waldinger, Applied Econometrics, University of Warwick