Predictive Analysis for Decision Making

Week 3
Extending Linear Regression

Dr Issam Malki
School of Finance and Accounting
Westminster Business School

February 2024

1

1

Classical Linear Regression
RECAP

2

2

Linear Regression

- The multiple linear regression model can be expressed as:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + u_i$$

i: subscript refers to the units and takes values from 1 to n.
x_{ij}: explanatory variables where j=1, 2, …, k with x_{i1}=1.
β_j: unknown parameters to be estimated.
u_i: random error component. Unobserved with variance σ^2.

We aim to estimate β_j and σ^2.

3

3

Linear Regression

- The multiple linear in a matrix notation:

$$Y = X\beta + \mathbf{u}$$

where Y is $(n \times 1)$, X is $(n \times k)$, $\mathbf{u}$ is $(n \times 1)$ and β is $(k \times 1)$.

- The OLS estimator is define as follows:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Why? See notes.

4

4

Linear Regression

- Assumptions
  - X is fixed (non stochastic) with rank k (full column rank, meaning?)
  - $\mathbf{u}$ is random vector with $\boldsymbol{E(u) = 0}$ and $\boldsymbol{\mathbf{var}(u) = E(u'u) = \sigma^2 I}$. Or:

$$var[u_i] = var[u] = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \ddots & \cdots & \vdots \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

5

5

Linear Regression

- Assumptions – Implications
  - Assumption 1:
    - Explanatory variables are strictly exogenous (i.e. E(X**u**)=0).
    - The X's are linearly independent (i.e. no multicollinearity).
  - Assumption 2:
    - The fitted line is indeed in the middle.
    - The errors are homoscedastic (no heteroskedasticity).
    - The errors are serially uncorrelated (no autocorrelation)
  - The validity of these assumptions is a must for the OLS to be a reliable estimator.

6

6

Linear Regression

- Key results I: unbiasedness

$$E(\hat{\beta}) = \beta$$

See full proof in the notes

- Key results II:

$$\boldsymbol{var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

where σ^2 is estimated using the sum of squared residuals, $e'e = \sum_{i=1}^{n} e_i^2$, as follows:

$$\widehat{\sigma^2} = s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k}$$

- Key results III: Gauss-Markov Theorem:

  OLS estimator is the Best Linear Unbiased Estimator (BLUE)

7

7

Linear Regression in Python

- Data: nls80.xls. See problem set for details on data and variables definitions.
- You need the following libraries for the basic linear regression
  - Pandas
  - Numpy
  - Statsmodels.api
- For today, we need one more: matplotlib.pyplot
- Use the file: computer_seminar 1
- The model we are about to estimate is:

$$\ln(wage_i) = \beta_1 + \beta_2 educ_{i2} + \beta_3 hours_{i3} + u_i$$

8

8

Linear Regression in Python

OLS Regression Results

Dep. Variable:	logwage	R-squared:	0.103
Model:	OLS	Adj. R-squared:	0.101
Method:	Least Squares	F-statistic:	53.62
Date:	Wed, 03 Feb 2021	Prob (F-statistic):	9.12e-23
Time:	10:54:46	Log-Likelihood:	-466.72
No. Observations:	935	AIC:	939.4
Df Residuals:	932	BIC:	954.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.1504	0.109	56.534	0.000	5.937	6.364
educ	0.0612	0.006	10.243	0.000	0.049	0.073
hours	-0.0044	0.002	-2.448	0.015	-0.008	-0.001

Omnibus:	28.032	Durbin-Watson:	1.765
Prob(Omnibus):	0.000	Jarque-Bera (JB):	34.519
Skew:	-0.340	Prob(JB):	3.19e-08
Kurtosis:	3.651	Cond. No.	388.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

9

Classical Linear Regression

Extension II: Relaxing Assumptions About the Error Term and Matrix of Covariates

10

10

Endogeneity

- Fixed X

$$\mathrm{E}(\hat{\beta}) = E\{(X'X)^{-1}X'Y\} = \beta \text{ (We need to prove this!)}$$

- Random X (See lectures Week 1 and 2 for examples)
  - Keep all other assumptions (including homogeneity – or stability of the model).
  - We do not know what would be E(X)
  - We also what would be E($X\boldsymbol{u}$)

11

Endogeneity

- Under Random X:

$$\mathrm{E}(\hat{\beta}) = E\{(X'X)^{-1}X'Y\} = \beta + E\{(X'X)^{-1}X'\boldsymbol{u}\}$$

- The OLS estimator is no longer unbiased (i.e. $\mathrm{E}(\hat{\beta}) \neq \beta$)
- Even when the sample size is large (i.e. $n \to \infty$), the OLS estimator is still biased. This is to say:

$$plim\hat{\beta} \neq \beta$$

This last expression refers states that the probability limit of $\hat{\beta}$ is not consistent and does not provide a good estimate of β.

12

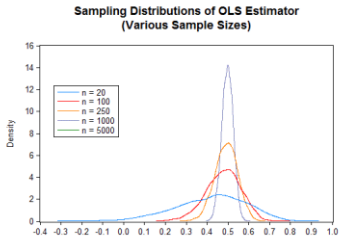
Endogeneity – Properties and Uses of Plims

- Suppose the true parameter $\beta = 0.5$. A model is estimated using the same assumptions but different sample sizes. The aim is to get an unbiased OLS estimator. Here are the results of this simulation:

Sample	$\hat{\beta}$	β	Bias = $(\hat{\beta} - \beta)$
20	0.417166	0.5	-0.08283
100	0.477581	0.5	-0.02242
250	0.491151	0.5	-0.00885
1000	0.497703	0.5	-0.0023
5000	0.499518	0.5	-0.00048

13

Endogeneity – Properties and Uses of Plims



14

Endogeneity – Properties and Uses of Plims

- The above experiment is testing the consistency of OLS estimator. This is to say, the limit of estimated value when the sample size is getting larger or approaching infinity:
  - Step 1:

$$\hat{\beta} = \{(X'X)^{-1}X'Y\} = \beta + (X'X)^{-1}X'\boldsymbol{u}$$

  - Step 2:

$$\hat{\beta} = \beta + (\frac{1}{N}X'X)^{-1}(\frac{1}{N}X'\boldsymbol{u})$$

  - Step 3:

$$plim\hat{\beta} = \beta + (plim\,\frac{1}{N}X'X)^{-1}(plim\,\frac{1}{N}X'\boldsymbol{u}) = \beta$$

$$(plim\,\frac{1}{N}X'\boldsymbol{u}) = \mathbf{0}$$

15

Endogeneity – Properties and Uses of Plims

- If X is random, then:

$$(plim\frac{1}{N}X'\boldsymbol{u}) \neq \mathbf{0}$$

and therefore:

$$plim\hat{\beta} = \beta + (plim\frac{1}{N}X'X)^{-1}(plim\frac{1}{N}X'\boldsymbol{u}) \neq \beta$$

This means that even when increasing the sample to infinity, the OLS estimator will never approach the true value. Thus under endogeneity, the OLS estimator is biased and inconsistent.

16

Instrumental Variable Estimation

- We need to find additional variables that might serve as 'instrumental variables' or 'instruments'.
- These variables are collected in a single matrix Z $(n \times m)$, where the rank of Z is equal to m the number of instruments (full column rank – explain!).
- For Z to be valid, two conditions must be satisfied:
  - **Relevance:** Z and X are correlated

$$(plim\frac{1}{N}Z'X) \neq \mathbf{0}$$

  - **Exogeneity:** Z and **u** are not correlated

$$(plim\frac{1}{N}Z'\boldsymbol{u}) = \mathbf{0}$$

17

Instrumental Variable Estimation

- The variables in X that are uncorrelated with **u** are called **exogenous.** They can be used instruments for themselves.
- The variables in X that are correlated with **u** are called **endogenous.** For each endogenous variable, we must find at least one instrument in order to proceed.
- This means that $m \geq k$:
  - $m = k$: exactly identified.
  - $m > k$: over identified.

18

Instrumental Variable Estimation- The principle

- Recall the linear model:

$$Y = X\beta + \mathbf{u}$$

Pre multiply by Z':

$$Z'Y = Z'X\beta + Z'\mathbf{u}$$

The error term has: $\text{var}(Z'\mathbf{u}) = \sigma^2 Z'Z$

The IV estimator is then derived to be:

$$\hat{\beta}_{IV} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y$$

Note that this is the same principle as OLS.

19

19

Instrumental Variable Estimation- The principle

- Consistency can be proved using the conditions (relevance and exogeneity) and the plim:
  - Step 1:

$$\hat{\beta}_{IV} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y$$
$$= \beta + (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'\boldsymbol{u}$$

  - Step 2:

$$\hat{\beta}_{IV} = \beta + (N^{-1}X'Z(N^{-1}Z'Z)^{-1}N^{-1}Z'X)^{-1}N^{-1}X'Z(N^{-1}Z'Z)^{-1}N^{-1}Z'\boldsymbol{u}$$

  - Step 3:

$$plim\hat{\beta}_{IV} = \beta + (plimN^{-1}X'Z(plimN^{-1}Z'Z)^{-1}plimN^{-1}Z'X)^{-1}plim\, N^{-1}X'Z(plimN^{-1}Z'Z)^{-1}plimN^{-1}Z'\boldsymbol{u}$$

Since: $(plim\frac{1}{N}Z'\boldsymbol{u}) = \mathbf{0}$ then:

$$plim\hat{\beta}_{IV} = \beta$$

20

Two Stage Least Square

- A more general approach is to perform OLS in two stages:
  - Stage 1: Regress each variable in X on Z and collect the predicted values into one matrix:

$$\hat{X} = Z(Z'Z)^{-1}Z'X$$

  - Stage 2: Regress $\boldsymbol{y}$ on $\hat{X}$ and obtain the OLS slope estimates:

$$\hat{\beta}_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y$$

21

When do we use IV estimator

- Endogeneity can be due to one or more of the following:
  - Measurement error in explanatory variables
  - Autoregression with autocorrelated errors
  - Simultaneity
  - Omitted explanatory variable
  - Sample selection

22

OLS Vs. 2SLS

OLS Regression Results

Dep. Variable:	logwage	R-squared:	0.103
Model:	OLS	Adj. R-squared:	0.101
Method:	Least Squares	F-statistic:	53.62
Date:	Wed, 03 Feb 2021	Prob (F-statistic):	9.12e-23
Time:	10:54:46	Log-Likelihood:	-466.72
No. Observations:	935	AIC:	939.4
Df Residuals:	932	BIC:	954.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.1504	0.109	56.534	0.000	5.937	6.364
educ	0.0612	0.006	10.243	0.000	0.049	0.073
hours	-0.0044	0.002	-2.448	0.015	-0.008	-0.001

Omnibus:	28.032	Durbin-Watson:	1.765
Prob(Omnibus):	0.000	Jarque-Bera (JB):	34.519
Skew:	-0.340	Prob(JB):	3.19e-08
Kurtosis:	3.651	Cond. No.	388.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

23

OLS Vs. 2SLS

IV-2SLS Estimation Summary

Dep. Variable:	wage	R-squared:	-0.0127
Estimator:	IV-2SLS	Adj. R-squared:	-0.0148
No. Observations:	935	F-statistic:	24.339
Date:	Wed, Feb 10 2021	P-value (F-stat)	0.0000
Time:	10:58:59	Distribution:	chi2(2)
Cov. Estimator:	unadjusted		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	5.3475	0.3398	15.737	0.0000	4.6815	6.0135
hours	-0.0063	0.0021	-3.0422	0.0023	-0.0103	-0.0022
educ	0.1267	0.0268	4.7202	0.0000	0.0741	0.1793

Endogenous: educ
Instruments: sibs
Unadjusted Covariance (Homoskedastic)
Debiased: False

24

Other Estimators

- Generalized Method of Moments:

$$\hat{\beta}_{GMM} = (X'ZW^{-1}Z'X)^{-1}X'ZW^{-1}Z'Y$$

where W is symmetric positive definite matrix, possibly stochastic.

- Note that if $W=Z'Z$, then the GMM and 2SLS estimators are equivalent.
  - The GMM estimator is more efficient.
  - The GMM estimator is preferred if we have more than one instrument.
- There are other variations of GMM, such as Iterative GMM.
- One popular estimator is Maximum Likelihood Estimator (MLE).

25

Testing and Inference

- Testing for relevance assumption:
  - Start with the correlation matrix to see if there is any meaningful correlation
  - Test the statistical significance of the instrument on the endogenous variable.
  - It can be a simple t test or joint significance F test.
- Exogeneity tests:
  - Wu and Hausman, Wooldridge tests
  - The null implies that the set of endogenous variables are uncorrelated with the error term.
- Overidentifying restrictions:
  - Sargan test: the null states that the model is not overidentified (valid instruments).
  - J test: for GMM.

26

THANK YOU

27

27