Micro Econometrics

Sol Yates

February 11, 2024

Contents

Lecture 1: First Lecture

Wed 31 Jan 11:21

0.1 Classical Linear Model

Studying the relationship between one (dependent) variable y and k other independent variables x_j, $(j = 1, \dots, k)$

1. Does the coved vaccine work
2. What are the returns to schooling
3. What is the effect of having internet at home on student's grades
4. Does a job training program decrease the time of getting out of unemployment

Sometimes we are interested in a single variable, and other regress ors are included as controls
Regression is the workhorse for many sophisticated identification procedures
Linear regression

- Relies on 5 main Gauss-Markov assumptions
- In small samples is unbiased and BLUE
- in large samples is consistent and asymptotically normal. There is no need for the normality assumption to establish asymptotic distribution

0.2 Multiple Regression

classical linear model

$$\begin{aligned} y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_k x_{i,k} + u_i, i = 1, \ldots, n \\ = x_i \beta + u_i \quad \text{vector notation} \\ Y = X\beta + U \quad \text{matrix notation.} \end{aligned}$$

1. where β_0 is the intercept, β_j is the parameter (slope) associated with x_j
2. a is the unobserved error term : containing factors other than x_j 's explaining y
3. n is the number of observations

Least Squares Estimator

Objective

: to estimate the effect of x_j on y, we need to estimate the population parameters $\beta_0, \ldots, \beta_k$

Ordinary Least squares estimates

β by minimising the sum of squared residuals :

$$\hat{\beta} = argmin \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_k x_{i,k})^2 = \|Y - X\beta\|^2$$

taking first order conditions

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

It can be shown that

1. Residuals : $\hat{u}_i = y_i - x_i \beta$ with $\sum_{i=1}^{n} \hat{u}_i = 0$
2. fitted values : $\hat{y}_i = x_i \beta$

in the single regress or model
Model $y = \beta_0 + \beta_1 x_1 + u$

$$\begin{aligned} \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_{i,1} - \overline{x}_1 (y_i - \overline{y})}{\sum_{i=1}^{n} (x_{i,1} - \overline{x}_1)^2} \\ = cov \frac{x_1, y}{\hat{v}(x_1)} \hat{\beta}_0 = \overline{y} - \hat{\beta_1} \overline{x}_1 \end{aligned}$$

0.3 Gauss Markov Assumptions

Assumption 1 : Linear in parameters (*MLR.1*)
The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + B_k x_k + u$$

Where $\beta_0, \beta_1, \ldots \beta_k$ are the unknown parameters (constants) of interest and u is an unobservable random error or disturbance term

Assumption 2 : Random Sampling (*MLR.2*)
we have a random sample of n observations, following the population model assumption in MLR.1.

- often referred to as IID assumption
- ensures that our sample's representative for the population
- would fail if we observed only part of the population in our sample

Assumption 3 : No perfect colinearity *MLR.3*
In the sample (thus pop too), none of the independent variables are constant, and there are no exact linear relationships between the independent variables

- often referred to as full rank assumption
- dummy variable trap - not to include a binary for both male and female
- not to confuse with highly but not perfectly correlated variables (multicollinearity)

Assumption 4 : zero conditional mean *MLR.4*
The error u has an expected value of zero given any value of the explanatory variable,

$$E[u|x_1, x_2, \ldots, x_k] = 0$$

- key to deriving unbiasedness
- if it holds for variable x_j, the variable is exogenous
- requires at a minimum : all factors in the observed error term must be uncorrelated with the explanatory variables
- any problem that causes u to be correlated with any of the x_j 's causes this assumption to fail and OLS to be biased !
- examples for endogeneity : misspecified functional form, omitting important variables, measurement error and any x_j being jointly determined with y

Assumption 5 : Homoskedacity *MLR.5*
The error u has the same variance given any value of explanatory variables, in other words

$$V[u|x_1, \ldots x_k] = \sigma^2$$

- the variance of the unobserved error u conditional on the explanatory variables is the same for all combinations of the outcomes of the explanatory variables
- if this assumption fails, we speak of heteroskedatic errors
- this assumption is not needed for unbiased/ consistency but for efficiency of OLS
- this also means that $V[y|x] = \sigma^2$

0.4 Small sample properties

Unbiasedness of OLS

under assumption 1-4, the OLS estimator is unbiased

$$E[\hat{\beta}_j] = \beta_j \text{for } j = 0, \dots, k$$

for any values of the population parameter β_j.

The OLS estimators are unbiased estimators of the population parameters

- might not exactly be the population value
- deviations from the population value are not systematic
- if we were to repeat the estimation on several random samples the deviations should avergae out to zero

Variance

sampling variance of the LOS slope estimators

Under assumptions 1-5, conditional on the sample values of the independent variables the variance is

$$V[\hat{\beta}_j] = \frac{\sigma^2}{SST_j(1 - R_j^2} \text{for} j = 0, \dots, k$$

Where $SST_j = \sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2$ is the sum of total sample variation in x_j and R_j^2 is the R-squared from regressing x_j on all other independent variables (including an intercept)

- the standard error formulas make it apparent that we need variation in the regress ors to increase precision
- the R_j^2 representation makes it also apparent that a high multicollinearity increases the variance of the estimator

Matrix Representation

General formula in matrix form (including the intercept)

$$V[\hat{B}_j] = \sigma^2(X'X)^{-1}$$

the variance of the j-the parameter estimate

$$\sigma^2(X'X)^{-1}_{[j+1,j+1}$$

Gauss Markov Theorem

Theorem 1 : Gauss Markov Theorem
Under assumptions 1-5 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are the *best linear unbiased estimators* (BLUE)s of

$$\beta_0, \beta_1, \dots, \beta_k$$

- if the assumptions hold, we do not need to look for another unbiased estimator since this see the best
- best meaning most efficient with smallest variance

Small Sample Inference

we are interested in performing inference, we need : variance (standard error) and distribution of parameter estimator

Firstly Estimation of the error variance : $\sigma^2 = \frac{1}{n-k-1}\hat{u}_i^2 = \frac{SSR}{n-k-1}$
we can show this estimator is unbiased under

Theorem 2 : unbiased estimation of σ^2
under the GM assumptions (1-5),

$$E[\hat{\sigma^2}] = \sigma^2$$

Standard Errors

is $\sqrt{}$ of variance

$$sd(\hat{\beta}_j) = \sqrt{\frac{\sigma^2}{SST_j(1-r_j^2}}$$

$$se(\hat{\beta}_j) = \sqrt{\frac{\sigma^2}{SST_j(1-R_j^2}} = \frac{\hat{\sigma}}{\sqrt{n}sd(x_j)\sqrt{1-R_j^2}} \quad \text{where} sd(x_j) = \sqrt{n^{-1}\sum_i (x_{ij} - \overline{x}_j)^2}$$

standard errors shrink to zero at the rate $\frac{1}{\sqrt{n}}$ (since in denominator)

Assumption 6 : Normality MLR.6
The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2 : $u \sim \mathcal{N}(0, \sigma^2)$

This is a stronger assumption than 1-5 and means we are necessarily assuming zero conditional mean (4) and homoskedacity (5).

Theorem 3 : Normal Sampling Distributions
Under assumptions 1-6, conditional on the sample values of the independent variables

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, V(\hat{\beta}_j)$$

(variance expression)

Therefore,

$$(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j) \sim \mathcal{N}(0,1)$$

Or, $\hat{\beta}|X \sim MVN(\beta, \sigma^2(X'X)^{-1}$ (matrix notation)

on testing

this doesn't give us a test start since it depends on unobservable error variance

0.5 asymptotic properties

Assumption 7 : Zero Mean and Zero correlation *(MLR.4'*

$$E[u] = 0 \quad and \quad Cov[x_j, u] = 0, \quad \text{for} \quad j = 1, 2, \ldots, k.$$

if we are only interested in consistency : this replace zero conditional mean (MLR.4)

- However, zero conditional mean important for finite sample and to ensure that we have properly modelled the population regression function $E[y|x] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- this gives us the average or partial effects of x_j on the expected value of y.

Theorem 4 : Consistency of OLS

Under assumptions MLR.1 - MLR.4, (or replacing 4 with 7), the OLS estimator $\hat{\beta}_j$ is consistent for β_j for all $j = 1, 2, \ldots, k$

Consistency means that when n goes to ∞, the estimator will recover the population value in probability :

$$\underset{n\to\infty}{plim} \hat{\beta}_j = \beta_j$$

Essentially, the asymptotic bias shrinks to 0.
On the consistency of OLS
For the simple model with one regressor, show consistency : $y_i = \beta_0 + \beta_1 x_{i,1} + u_i$

1. Write down the formula for $\hat{\beta}_1$ and plug in y_i :

$$\hat{\beta}_1 = (\sum_{i=1}^{n}(x_{i,1} - \overline{x})(y_i - \overline{y})/(\sum_{i=1}^{n}(x_{i,1} - \overline{x}_1)^2)$$
$$= \beta_1 + (\frac{1}{n}\sum_{i=1}^{n}(x_{i,1} - \overline{x}_1)(u_i - \overline{u})/(\frac{1}{n}\sum_{i=1}^{n}(x_{i,1} - \overline{x}_1)^2)$$

2. apply the LLN :

$$\underset{n\to\infty}{plim} \hat{\beta}_j = \beta_1 + \frac{cov[u, x_1]}{V(x_1)} = \beta_1$$

since $Cov[u, x_1] = 0$ (previous assumption)

However, we need to assume finite moments for the LLN to hold. Since it assumes iid observations
LLN : $\bar{X}_n \xrightarrow{P} X$ as $n \to \infty$
Remarks

- in a single regressor model : $\beta_1 = \frac{Cov(y,x_1)}{V(x_1)}$
- including more regressor changes this handy expression for the population estimate β_j bit since the effect of the other covariates is partialled out, one still recovers β_j
- multicollinearity only affects the variance of the estimator but not consistency

where $se(\hat{\beta}_j)$ is the usual OLS estimator

Lecture 2: Standard Errors

Sun 04 Feb 17:54

1 Introduction

- after a point estimate, we want to know the statistical significance
- requiring the standard error and the distribution
- if the standard errors are wrong we cannot use the usual t-dist statistics for drawing inference
- under too large SES,
  - zero might be included in the CI when it should not be
  - there is a risk of not detecting an effect even there was
- too small SEs
  - zero might not be in the CI when it should be
  - we may claim the existence of an effect when in reality there is none
- **wrong** SE can lead to **wrong** conclusions!
- Robust SE
  - traditional inference assumes homoskedacity
  - but the variance of error terms might be different for different observations depending on their characteristics
  - heteroskedacity robust SE to the rescue
- SE
  - traditional estimation relied on random sampling
  - in the case of data with a group structure, the error terms might be correalted
  - to account we use clustered SE
- bootstrap
  - bootstrap is a re sampling method that offers an alternative to inference based on asymptotic formulas convenient in cases where the sampling distribution is unknown

Note.

1. if we can estimate a model parameter consistently, why do we care about inference?
2. do heteroskedatic errors or clustering affect the OLS point estimate for model parameters
3. an example where heteroskedacity / clustering occurs
4. when would bootstrap be useful?

2 Heteroskedacity - Robust Standard Errors

2.1 Heteroskedacity Problems

- traditional inference assumes homoskedatic errors

$$V(u|x) = \sigma^2$$

- this implies that the variance of the unobserved error a, is constant for all possible values of all the regressor x's
- since the proofs for unbiasedness and consistency do not depend on this assumption we still obtain unbiased and consistent OLS estimates
- however, if this is not true (σ_i^2) then the errors are called **heteroskedatic** and traditional variance estimators are biased
- heteroskedacity robust SE specifically in the CS case
- if the degeree of heteroskedacity is low ,the traditional variance estimator might be less biased

> **Example** (Returns to education). if we regress $wage \sim educ$
>
> it is reasonable to believe that the variance is unobserved factors hidden in the error term differers by educational attainment
>
> individuals with higher education : potentially more diverse interests and more job opportunities affecting their wage
>
> individuals with very low education : fewer opportunities and often must work at the minimum wage, the error variance is lower

variance estimation with heteroskedacity

simple regression : $y = \beta_0 + \beta_1 x + u$
we know $\hat{b}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
therefore :

$$V(\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x)^2\sigma_i^2}}{SST_x^2}$$

where $SST_x = \sum_{i=1}^{n}(x_i - \overline{x})^2$

- if $\sigma_i^2 = \sigma^2$ the formula reduces to the traditional formula : $V(\hat{\beta}_1 = \frac{\sigma^2}{SST_x}$
- this leads to the following heteroskedacity robust estimator :

$$\hat{V}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2\overline{u}_i^2}{SST_x^2}$$

where u_i^2 are the OLS residuals

Generalisation

the formula generalises to

$$\hat{V}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

where $\hat{r_{ij}}$ is the i-th residual from regressing x_j on all other independent variables and SSR_j the sum of squared residuals from this regression

- robust to heteroskedacity **of any form** (inc homoskedacity)
- often also called white, huber, eicker SE
- sometimes degrees of freedom adjustment by multiplying $\frac{n}{n-k-1}$
- but with **drawback** that it only has asymptotic justification

2.2 Breusch-Pagan Test for Heteroskedacity

- testing hypothesis

$$H_0 : V(u|x_1, \ldots, x_k) = E(u^2|x_1, \ldots, x_k) = \sigma^2$$

- assume a linear relationship :

$$u^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + v, E[v|x_1, \ldots, x_k] = 0$$

- since we cannot observe the errors, we replace them with the residuals and estimate the regression

1. estimate
$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \ldots + error$$
Recover the $R^2_{\hat{u}}$
2. Hypotheses : $\delta_1 = \ldots = \delta_k = 0$
3. test stat : $F = \frac{R^2_{\hat{u}}/k}{(1-R^2_{\hat{u}})/(n-k-1)} \sim^{H_0} \mathcal{F}_{k,n-k-1}$ or $LM = nR^2_{\hat{u}^2} \sim^{H_0} \chi^2_k$
4. Decision : if the p-value is small enough (typ <0.05), we **reject** the null of homoskedacity

Example. model : price $= \beta_0 + \beta_1$lotsize $+ \beta_2$sqrft $+ \beta_3$bdrms $+ u$

where price is the housing price, lotsize the size of the lot, size size of house in sq ft

we want to estimate teh above regression and test for heteroskedacity and see whether using logs in the dependent variable changes our conclusion

- use robust SE when heteroskedatic errors
- but there is a danger of small sample bias from robust SE (arising from asymptotic justification)
- under homoskedacity or little heteroskedacity, it might be preferable to use the traditional OLS variance estimator
- it is recommended to report both the robust and conventional standard error and suggest to take the maximum of both for inference

- white test for heteroskedacity includes the squares and cross-products of the independent variables
- LPM : built in heteroskedacity $\rightarrow$ need to compute robust SEs
- using logs in the dependent variables has been seen to improve in terms of heteroskedacity in many applications

3 Clustered Standard Errors

Illustration of Moulton Problem

- Pillar assumption is random sampling
- there is potential dependence of data within a group structure
  - exam grades of children from same class or school : grades are correlated because of the same school, teacher and background / class environment
  - health outcomes in the same village, Errors are correalted because of the same medical and food supply and similar cultural background
  - earning in the same region might be correlated because of the same industrial structure
  - analysing workers in firms (earnings, tenure, promotion) will suffer from common firm effects

The problem

- illustration using a simple model with a group structure
- intuitively, effect of a macro variable on an individual level outcomes
  - effect of school-type on exam-grades
  - effect of regional unemployment on individuals' wages
- model

$$y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}$$

  - with $g = 1, \ldots, G$ and $i = 1, \ldots, n$
  - y_{ig} is the outcome for individual i in group g
  - here x_g varies only at the group level

> **Note.** Lecture : if we estimate a model parameter consistently, why do we care about inference?
>
> - we would like to investigate a problem
> - policymaker would like to know whether to implement school building program
> - but what is decision rule? Typically, think about Statistical significance and sufficient magnitude then the policymaker wants to adopt the program, if not then not adoptable.
> - we need CI or at least a statistical test. For this we need SE and distribution
> - need to estimate SE correctly to get correct CI, if we have too large CI (SE wrong), the implication/ error is that we risk not detecting an effect, when there is

- but the other way around too small SE (forgot to cluster), might think building schools help and invest a lot of money, but the effect is 0
- this is a danger and the problem is *incorrect standard errors lead to incorrect confidence intervals*

Note. Lecture : Once we have accounted for clustering using the Moulton approach compared to the standard errors, is it more likely that the clustered standard errors are larger or smaller than the OLS

> **Note.** larger

Note. Lecture : what solutions exist to account for clustering

- group averages (only valid for regressors that don't vary within each individual within a group)
- parametric - estimate the moulton factor
- clustering SE
- block bootstrap

Hello world