Multivariate probability inference

Adam Boult (www.bou.lt)

April 30, 2025

Contents

Preface

This is a live document, and is full of gaps, mistakes, typos etc.

Part I

Linear regression for inference

Chapter 1

Ordinary Least Squares for inference

1.1 Bias of OLS estimators

1.1.1 Expectation of OLS estimators

Expectation in terms of observables

We have: $\hat{\theta} = (X^T X)^{-1} X^T y$

Let's take the expectation.

$E[\hat{\theta}] = E[(X^T X)^{-1} X^T y]$

Expectation in terms of errors

Let's model y as a function of X. As we place no restrictions on the error terms, this is not as assumption.

$y = X\theta + \epsilon$.

$E[\hat{\theta}] = E[(X^T X)^{-1} X^T (X\theta + \epsilon)]$

$E[\hat{\theta}] = E[(X^T X)^{-1} X^T X\theta] + E[(X^T X)^{-1} X^T \epsilon)]$

$E[\hat{\theta}] = \theta + E[(X^T X)^{-1} X^T \epsilon)]$

$E[\hat{\theta}] = \theta + E[(X^T X)^{-1} X^T] E[\epsilon] + cov[(X^T X)^{-1} X^T, \epsilon]$

The Gauss-Markov: Expected error is 0

$E[\epsilon = 0]$

This means that:

$E[\hat{\theta}] = \theta + cov[(X^T X)^{-1} X^T, \epsilon]$

The Gauss-Markov: Errors and indepedent variables are uncorrelated

If the error terms and X are uncorrelated then $E[\epsilon|X] = 0$ and therefore:

$E[\hat{\theta}] = \theta$

So this is an unbiased estimator, so long as the condition holds.

1.2 Variance of OLS estimators

1.2.1 Variance of OLS estimators

Variance-covariance matrix

We know:

$\hat{\theta} = (X^T X)^{-1} X^T y$

$y = X\theta + \epsilon$

Therefore:

$\hat{\theta} = (X^T X)^{-1} X^T (X\theta + \epsilon)$

$\hat{\theta} = \theta + (X^T X)^{-1} X^T \epsilon$

$\hat{\theta} - \theta = (X^T X)^{-1} X^T \epsilon$

$Var[\hat{\theta}] = E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T]$

$Var[\hat{\theta}] = E[(X^T X)^{-1} X^T \epsilon (X^T X^{-1} X^T \epsilon)^T]$

$Var[\hat{\theta}] = E[(X^T X)^{-1} X^T \epsilon\epsilon^T X (X^T X)^{-1}]$

$Var[\hat{\theta}] = (X^T X)^{-1} X^T E[\epsilon\epsilon^T] X (X^T X)^{-1}$

We write:

$\Omega = E[\epsilon\epsilon^T]$

$Var[\hat{\theta}] = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}$

Depending on how we estimate Ω, we get different variance terms.

Variance under IID

If IID:

$\Omega = I\sigma_\epsilon^2$

$Var[\hat{\theta}] = (X^T X)^{-1} X^T I\sigma_\epsilon^2 X (X^T X)^{-1}$

$Var[\hat{\theta}] = \sigma_\epsilon^2 (X^T X)^{-1}$

1.2.2 Heteroskedasticity-Consistent (HC) standard errors

Variance of OLS estimators

$Var[\hat{\theta}] = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}$

Robust standard errors for heteroskedasticity

$\Omega_{ij} = \delta_{ij} \epsilon_i \epsilon_j$

These are also known as the Eicker-Huber-White standard errors, or the White correction.

These are also refered to as robust standard errors.

1.3 Properties of the OLS estimator

1.3.1 Maximum Likelihood Estimator (MLE) and OLS equivalence

The OLS estimator

$\hat{\theta}_{OLS} = (X^T X)^{-1} X^T y$

$E[\hat{\theta}_{OLS}] = w$

$Var[\hat{\theta}_{OLS}] = \sigma^2 (X^T X)^{-1}$

The MLE estimator

$y_i = \mathbf{x}_i \theta + \epsilon_i$

$P(y = y_i | x = x_i) = P(\epsilon_i = y_i - \mathbf{x}_i \theta)$

If we assume $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ we have:

$$P(y = y_i | x = x_i) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} e^{-\frac{(y_i - \mathbf{x}_i\theta)^2}{2\sigma_\epsilon^2}}$$

$$L(X, \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} e^{-\frac{(y_i - \mathbf{x}_i\theta)^2}{2\sigma_\epsilon^2}}$$

$$l(X, \theta) = \sum_{i=1}^n -\frac{1}{2} \ln(2\pi\sigma_\epsilon^2) - \frac{(y_i - \mathbf{x}_i\theta)^2}{2\sigma_\epsilon^2}$$

$$\frac{\delta l}{\delta \theta_j} = \sum_{i=1}^n 2x_{ij} \frac{y_i - \mathbf{x}_i\theta}{2\sigma_\epsilon^2}$$

$\sum_{i=1}^n x_{ij}(y_i - \hat{\theta}_{MLE} \mathbf{x}_i) = 0$

$X^T(y - X\hat{\theta}_{MLE}) = 0$

$X^T y = X^T X\hat{\theta}_{MLE}$

$\hat{\theta}_{MLE} = (X^T X)^{-1} X^T y$

Equivalence

If errors are normally IID then:

$\hat{\theta}_{OLS} = \hat{\theta}_{MLE}$

1.3.2 Gauss-Markov theorem

Mean of errors zero + If the model should only have errors on upside or downside for some reason, OLS will not provide this.

Homoscedastic (all have the same variance) + The results aren't biased, but variances etc are

Errors are uncorrelated + (this would mean you should add lagged variables etc)

show bias from each GM violation

OLS is BUE under normally distributed errors

OLS is BLUE for non-normally distribed errors

1.4 Selection

1.4.1 T-test selection

1.4.2 Post-LASSO

1.5 Heteroskedasticity

1.5.1 Checking for heteroskedasticity using the White test

1.5.2 Robust standard errors

1.5.3 Noise

1.5.4 Regression dilution

Noise in y doesn't cause bias.

Noise in x does cause bias.

Need to correct.

1.5.5 Causality

1.5.6 Introduction

Causality v correlation. If just getting correlation, could have bad out of sample performance

Section on causality. Difference between disease causes symptom and symptom causes disease

Linear models can be manipulated to have any variable on the left.

Chapter 2

Testing regression parameter estimates with Z-tests and T-tests

Chapter 3

Multiple hypothesis testing

3.1 Multiple hypothesis testing

3.1.1 P-hacking

Likely to see some significant results from random chance.

3.1.2 Family-Wise Error Rate (FWER)

What is the chance of making at least one false positive result?

Number of tests: m

Number of false positive results: V

$FWER = P(V > 0)$

3.1.3 False Discovery Rate (FDR)

The proportion of false discoveries is:

$Q = \frac{V}{V+S}$

Where: V is the number of false positives

S is the number of true positives

The FRD is $E[Q]$.

3.1.4 The Bonferroni correction

We change the significance level.

reject if $p \leq \frac{\alpha}{m}$

If $m = 1$ this is the standard test.

Chapter 4

Generalised Least Squares

4.1 Generalised Least Squares (GLS)

4.1.1 The Generalised Least Squares (GLS) estimator

Introduction

We make the same assumptions as OLS.

$\mathbf{y} = \mathbf{X}\theta + \boldsymbol{\epsilon}$

We assume:

- $E[\epsilon|\boldsymbol{X}] = 0$
- $Cov[\epsilon|\boldsymbol{X}] = \boldsymbol{\Omega}$

The GLS estimator

GLS estimator is:

$\hat{\theta}_{GLS} = argmin_b (y - Xb)^T \Omega^{-1} (y - Xb)$

$\hat{\theta}_{GLS} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$

This is the vector that minimises the Mahalanobis distance.

This is equivalent to doing OLS on a linearly transformed version of the data.

Identifying Ω

If Ω is known, we can proceed. Generally, however, Ω is not known, and so the GLS estimate in infeasible.

4.2 Feasible Generalised Least Squares (FGLS)

4.2.1 The Feasible Generalised Least Squares (FGLS) estimator

Introduction

We do OLS to get a consistent estimate of Ω, $\hat{\Omega}$.

We then plug this into the GLS estimator.

4.3 Heteroskedasticity

4.3.1 Weighted least squares

4.4 Bias and variance of the GLS estimator

4.4.1 Introduction

you have the same sandwich term as before, so same process, right?

4.5 Linear discriminant analysis

Chapter 5

General Linear Models

5.1 Cross-sectional regression

5.1.1 The cross-sectional model

Hierarchical data

Our standard linear model is:

$y_i = \alpha + X_i\theta + \epsilon_i$

If we had two sets of data we could view these as:

$y_{i,0} = \alpha_0 + X_{i,0}\theta_0 + \epsilon_{i,0}$

$y_{i,1} = \alpha_1 + X_{i,1}\theta_1 + \epsilon_{i,1}$

Here, the data data from 1 does not affect the parameters in 2.

Pooled data

If we think the data generating process is similar between models, then by restricting the freedom of parameters between models we can get more data for each estimate.

For example if we think that all parameters are the same between the models we can estimate:

$y_{i,0} = \alpha + X_{i,0}\theta + \epsilon_{i,0}$

$y_{i,1} = \alpha + X_{i,1}\theta + \epsilon_{i,1}$

Or:

$y_{ij} = \alpha + X_{ij}\theta + \epsilon_{ij}$

Fixed slopes

Intercepts may be different between the groups. In this case we can instead use the model:

$y_{ij} = \alpha + X_{ij}\theta + \xi_j + \epsilon_{ij}$

There are different ways of estimating this model:

- Pooled OLS
- Fixed effects
- Random effects

5.1.2 Unbalanced data

5.2 The pooled OLS estimator

5.2.1 Pooled OLS

Introduction

Our model is:

$y_{ij} = \alpha + X_{ij}\theta + \xi_j + \epsilon_{ij}$

The pooled OLS estimator

5.3 The fixed effects estimator

5.3.1 Within and between transformation

Introduction

We can group the data in two ways, one gets between differences and the other within differences.

In the above example, we could find the effects of schools, or of departments.

$y_{ij} = \alpha + X_{ij}\theta + \epsilon_{ij}$

$(y_{ij} - \bar{y}_j) = (\alpha - \bar{\alpha}) + (X_{ij} - \bar{X}_j)\theta + (\epsilon_{ij} - \bar{\epsilon}_j)$

$(y_{ij} - \bar{y}_j) = (X_{ij} - \bar{X}_j)\theta + (\epsilon_{ij} - \bar{\epsilon}_j)$

Or alternatively:

$(y_{ij} - \bar{y}_i) = (X_{ij} - \bar{X}_i)\theta + (\epsilon_{ij} - \bar{\epsilon}_i)$

Regardless of the form we choose, we can write this as:

$\ddot{y}_{ij} = \ddot{X}_{ij}\theta + \ddot{\epsilon}_{ij}$

5.3.2 The fixed effects estimator

Recap on the model

Our model is:

$y_{ij} = \alpha + X_{ij}\theta + \xi_j + \epsilon_{ij}$

The fixed effects estimator

With fixed effects we assume that U_{ij} is a constant for each group. That is:

$U_{ij} = \delta_{ij} U_j$

$y_{ij} = \alpha + X_{ij}\theta + \epsilon_{ij} + \delta_{ij} U_j$

We can use this in a regression if the standard assumptions of OLS are met. In particular, that group membership is uncorrelated with the error term.

We add these dummies to X_{ij} and regress:

$y_{ij} = \alpha + X_{ij}\theta + \epsilon_{ij}$

The parameter for the dummy is the fixed effect of group membership.

As we are including membership in the dependent variables, there is no problem if group membership correlates with other independent variables.

Using the within and between transformations

$(y_{ij} - \bar{y}_i) = (X_{ij} - \bar{X}_i)\theta + (U_{ij} - \bar{U}_i) + (\epsilon_{ij} - \bar{\epsilon}_i)$

Or:

$\ddot{y}_{ij} = \ddot{X}_{ij}\theta + \ddot{\epsilon}_{ij}$

This this get the same outcome, but is a different computational process.

5.4 The random effects estimator

5.4.1 The random effects estimator

Introduction

Our model is:

$y_{ij} = \alpha + X_{ij}\theta + \xi_j + \epsilon_{ij}$

FGLS recap

The random effects estimator

For fixed effects, we had the requirement that group membership be uncorrelated with the error term, but that it could be correlated with other independent

variables.

For random effects models, group membership cannot be correlated with other variables.

We have:

$y_{ij} = \alpha + X_{ij}\theta + \epsilon_{ij} + U_{ij}$

We now model $U_{ij} = \bar{U}_j + \rho_j$.

$y_{ij} = \alpha + X_{ij}\theta + \epsilon_{ij} + \bar{U}_j + \rho_j$

This randomness of the effect implies, for example, that if we ran the survey again we would expect a different effect

Clustering standard error

Estimation

We use GLS.

5.5 Choosing the model form

5.5.1 The Hausman specification test

Introduction

The Hausman specification test allows you to choose between a fixed effects model and a random effects model.

Efficiency

Random effects models are more efficient.

5.6 The mixed effects estimator

5.6.1 The mixed effects estimator

Introduction

5.7 Manipulating data

5.7.1 Disaggregation

Used in polls

5.7.2 Multilevel Regression with Poststratification (Mr P)

Part II

Advanced inference

Chapter 6

Analysis of variance (ANOVA)

6.1 Cross-sectional data

6.1.1 Cross-sectional data

6.1.2 Group means and the grand mean

Introduction

6.1.3 Within-group variance and between-group variance

Introduction

6.2 Analysis of variance (ANOVA)

6.2.1 Analysis of variance (ANOVA) table

Chapter 7

Instrumental Variables

7.1 Motivation

7.1.1 Bias of OLS estimator from ommitted variables

7.1.2 Bias of OLS estimator from measurement error

7.2 Parameter estimation for simultaneous equations

7.2.1 Structural and reduced forms

7.2.2 Parameter identification problem with simultaneous equations

Identification terminology

A system is under-identified if there are not enough estimators for all structural parameters.

A system is exactly identified if there are the same number of estimators as structural parameters.

A system is over-identified if there are more estimators than structural parameters.

In general we have in our structural form:

$\sum_i^n \beta_{ij} y_i = \sum_i^m \gamma_{ij} x_i + \epsilon_j$

This is a system with n endogeneous variables and m exogeneous variables.

We can write this in matrix form.

$B\mathbf{y} = \Gamma\mathbf{x} + \epsilon$

We can use this to get:

$\mathbf{y} = B^{-1}\Gamma\mathbf{x} + B^{-1}\epsilon$

We estimate by placing restrictions on Γ.

Strucutral models

If our data generating process is:

$Q = \alpha + \beta P + \epsilon$

We can estimate $\alpha$and β through measuring P and Q.

If, however the data generating process involves simulataneous equations, we can have:

$Q = \alpha_1 + \beta_1 P + \epsilon_1$

$Q = \alpha_2 + \beta_2 P + \epsilon_2$

Reduced form

We can reduce this:

$\alpha_1 + \beta_1 P + \epsilon_1 = \alpha_2 + \beta_2 P + \epsilon_2$

$(\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)P + (\epsilon_1 - \epsilon_2) = 0$

$$P = \frac{\alpha_2 - \alpha_1}{\beta_1 - \beta_2} + \frac{\epsilon_2 - \epsilon_1}{\beta_1 - \beta_2}$$

We can rewrite this as:

$P = \pi_1 + \tau_1$

Similarly we can reduce for Q:

$$Q = \frac{\alpha_2\beta_1 - \alpha_1\beta_2}{\beta_1 - \beta_2} + \frac{\beta_1\epsilon_2 - \beta_2\epsilon_1}{\beta_1 - \beta_2}$$

$Q = \pi_2 + \tau_2$

We can't directly estimate structural models

If P is correlated with $epsilon_1$ or ϵ_2 then our estimates for β_1 and β_2 will be biased.

This also affects Q.

From the reduced forms we can see that P will be correlated, due to simultaneity.

The identification problem

We can estimate π_1 and π_2, but this does not allow us to identify any of the structural parameters.

We have 2 estimators, but 4 parameters.

This is the identification problem.

7.3 2 Stage OLS

7.3.1 2 Stage OLS (2SOLS) estimator

Motivation

If x is correlated with the error term the OLS estimate will be biased.

2 Stage OLS - first stage

We have

$y_i = x_i\theta + \epsilon_i$

$x_i = z_i\rho + \mu_i$

We do OLS on the second to get $\hat{\rho}$.

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

We use this to get predicted values of X.

$\hat{X} = Z\rho = Z(Z^T Z)^{-1} Z^T X = P_Z X$

2 Stage OLS - second stage

We then regress y on the estimated X:

$y_i = \hat{x}_i\theta + \epsilon_i$

Our prediction is then:

$\hat{\theta_{2SOLS}} = (\hat{X^T}\hat{X})^{-1}\hat{X^T}y)$

$\hat{\theta_{2SOLS}} = ((P_Z X)^T P_Z X)^{-1}(P_Z X)^T y)$

$\hat{\theta_{2SOLS}} = (X^T P_Z X)^{-1} X^T P_Z y)$

If the dimension of Z is the same as X this collapses to:

$\hat{\theta_{2SOLS}} = (Z^T X)^{-1} Z^T y$

7.3.2 Bias of the 2SOLS estimator

7.3.3 Variance of the 2SOLS estimator

7.4 More

7.4.1 Identification through exogeneous variables

Previously our structural model was:

$Q = \alpha_1 + \beta_1 P + \epsilon_1$

$Q = \alpha_2 + \beta_2 P + \epsilon_2$

And our reduced form:

$P = \frac{\alpha_2 - \alpha_1}{\beta_1 - \beta_2} + \frac{\epsilon_2 - \epsilon_1}{\beta_1 - \beta_2}$

$Q = \frac{\alpha_2\beta_1 - \alpha_1\beta_2}{\beta_1 - \beta_2} + \frac{\beta_1\epsilon_2 - \beta_2\epsilon_1}{\beta_1 - \beta_2}$

Or:

$P = \pi_1 + \tau_1$

$Q = \pi_2 + \tau_2$

Adding another variable

This time we add another measured variable, I.

$Q = \alpha_1 + \beta_1 P + \theta_1 I + \epsilon_1$

$Q = \alpha_2 + \beta_2 P + \theta_2 I + \epsilon_2$

The reduced form is now:

$P = \frac{\alpha_2 - alpha_1}{\beta_1 - \beta_2} + \frac{\theta_2 - \theta_1}{\beta_1 - \beta_2} I + \frac{\epsilon_2 - \epsilon_1}{\beta_1 - \beta_2}$

$Q = \frac{\alpha_2\beta_1 - \alpha_1\beta_2}{\beta_1 - \beta_2} + \frac{\theta_2\beta_1 - \theta_1\beta_2}{\beta_1 - \beta_2} I + \frac{\beta_1\epsilon_2 - \beta_2\epsilon_1}{\beta_1 - \beta_2}$

Or:

$P = \pi_{11} + \pi_{12} I + \tau_1$

$Q = \pi_{21} + \pi_{22} I + \tau_2$

We can estimate π_1 and π_2 as $\hat{\pi}_1$ and $\hat{\pi}_2$ respectively.

We can now create estimators $\hat{\pi}_{11}$, $\hat{\pi}_{12}$, $\hat{\pi}_{21}$ and $\hat{\pi}_{22}$.

Identification with an exogeneous variable

We now have 4 estimators and 6 parameters, meaning that we still cannot identify the model.

Partial identification

Can we use $\hat{\pi}$ to identify any of the structural parameters?

We know that:

- $\pi_{11} = \dfrac{\alpha_2 - \alpha_1}{\beta_1 - \beta_2}$
- $\pi_{12} = \dfrac{\theta_2 - \theta_1}{\beta_1 - \beta_2}$
- $\pi_{21} = \dfrac{\alpha_2\beta_1 - \alpha_1\beta_2}{\beta_1 - \beta_2}$
- $\pi_{22} = \dfrac{\theta_2\beta_1 - \theta_1\beta_2}{\beta_1 - \beta_2}$

If the exogenous variable only affects one side of the equation, so $\theta_1 = 0$, we have:

- $\pi_{11} = \dfrac{\alpha_2 - \alpha_1}{\beta_1 - \beta_2}$
- $\pi_{12} = \dfrac{\theta_2}{\beta_1 - \beta_2}$
- $\pi_{21} = \dfrac{\alpha_2\beta_1 - \alpha_1\beta_2}{\beta_1 - \beta_2}$
- $\pi_{22} = \dfrac{\theta_2\beta_1}{\beta_1 - \beta_2}$

So we can see that:

$\hat{\beta}_1 = \dfrac{\hat{\pi}_{22}}{\hat{\pi}_{12}}$

This means we now have:

- $\pi_{11} = \dfrac{\pi_{12}(\alpha_2 - \alpha_1)}{\pi_{22} - \pi_{12}\beta_2}$
- $\pi_{12} = \dfrac{\pi_{12}\theta_2}{\pi_{22} - \pi_{12}\beta_2}$
- $\pi_{21} = \dfrac{\pi_{12}(\alpha_2\beta_1 - \alpha_1\beta_2)}{\pi_{22} - \pi_{12}\beta_2}$
- $\pi_{22} = \dfrac{\pi_{12}\theta_2\beta_1}{\pi_{22} - \pi_{12}\beta_2}$

We can use this to also identify α_1.

Complete identification

If we have independent variables for each of the two equations, we can fully identify the model.

We will have 6 estimators and 6 parameters.

We are estimating:

$Q = \alpha_1 + \beta_1 P + \theta_1 I + \epsilon_1$

$Q = \alpha_2 + \beta_2 P + \theta_2 J + \epsilon_2$

I and J are essentially instrumental variables for the model.

I is an instrumental variable for demand shocks, and J is an instrumental variable for supply shocks.

7.4.2 Power of instruments

7.5 The Instrumental Variable (IV) estimator

7.5.1 Instrumental Variable (IV) estimator

$\hat{\theta_{IV}} = (Z^T X)^{-1} Z^T y$

2SOLS collpases to IV in some circumstances.

7.5.2 Bias of the IV estimator

Equal to actual parameter so long as ϵ uncorrelated with Z.

7.5.3 Variance of the IV estimator

In OLS we had:

$\hat{\theta_{OLS}} = (X^T X)^{-1} X^T y$

$Var[\hat{\theta_{OLS}}] = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}$

With IV we have

$\hat{\theta_{IV}} = (Z^T X)^{-1} Z^T y$

$Var[\hat{\theta_{IV}}] = (Z^T X)^{-1} Z^T \Omega Z (Z^T X)^{-1}$

We can use weighted least squares for Ω.

7.6 Choosing instrumental variables

7.6.1 Double selection

7.7 Other

7.7.1 Natural experiments

7.7.2 Non-linear models in the first stage

7.7.3 Random Effects Instrumental Variables (REIV)

7.7.4 Fixed Effects Instrumental Variables (FEIV)

7.7.5 SORT

synthetic IV indep on nuisance as alternative to matching.

IV: h3 on non-linear models for first stage

discontinuity

controlled experiments

two sources: missing data and simultaneous

variations in government rollouts, lotteries

IV may only affect subset of individuals

For example IV of draft number for military service. This only is an instrument for conscripts, not volunteers

generally, need to rationalise this and time series. There's stuff there on natural experiments etc

define confounding in IV? or in dependent variables? is different issue to the one of correlation with error?

h3 on Limited Information Maximum Likelihood

h3 on K-class estimation

Contrast loss and Siamese h3? One shot classification

IV: frame around parameter estimation when don't observe some variables. This can mean the direct variable can't be measured, or that some controls can't be measured

which factors to include? All?

page on structural and reduced forms

h3 on simultaneous equations there? Eg $y = c_1 + \theta_1 X + \epsilon_1$ $y = c_2 + \theta_2 X + \rho Z \epsilon_2$

We can turn this into the reduced form: $y = c_3 + \theta_3 Z + \epsilon_3$ $y = c_4 + \theta_4 Z + \epsilon_4$

difference between confounding and correlation with error?

Chapter 8

Imputing missing data for inference

8.1 Introduction

8.1.1 Techniques for inference

Deleting whole row if missing data.

As with techniques for prediction, there is bias if not random.

Chapter 9

Measurement error and inference

9.1 Other

9.1.1 Omitted variable bias

9.2 Measurement error

Chapter 10

Semi-parametric regression

10.1 The Robinson estimator

10.1.1 Partially linear models

10.1.2 The Robinson estimator

Partialling out

$y_i = \mathbf{x}_i\theta + g(\mathbf{z}_i) + \epsilon_i$

Consider:

$E(y_i|\mathbf{z}_i) = E(\mathbf{x}_i\theta + g(\mathbf{z}_i) + \epsilon_i|\mathbf{z}_i)$

$E(y_i|\mathbf{z}_i) = E(\mathbf{x}_i\theta|\mathbf{z}_i) + E(g(\mathbf{z}_i)|\mathbf{z}_i) + E(\epsilon_i|\mathbf{z}_i)$

$E(y_i|\mathbf{z}_i) = E(\mathbf{x}_i|\mathbf{z}_i)\theta + g(\mathbf{z}_i)$

We can now remove the parametric part:

$y_i - E(y_i|\mathbf{z}_i) = \mathbf{x}_i\theta + g(\mathbf{z}_i) + \epsilon_i - E(\mathbf{x}_i|\mathbf{z}_i)\theta - g(\mathbf{z}_i)$

$y_i - E(y_i|\mathbf{z}_i) = (\mathbf{x}_i - E(\mathbf{x}_i|\mathbf{z}_i))\theta + \epsilon_i$

We define:

- $\bar{y}_i = y_i - E(y_i|\mathbf{z}_i)$
- $\bar{x}_i = \mathbf{x}_i - E(\mathbf{x}_i|\mathbf{z}_i)$

$\bar{y}_i = \bar{x}_i\theta + \epsilon_i$

Estimating $\bar{y}_i$ and $\bar{x}_i$

So we can use OLS if we can estimate.

- $E(y_i|\mathbf{z}_i)$
- $E(\mathbf{x}_i|\mathbf{z}_i)$

We can do this with non-parametric methods.

10.1.3 Bias and variance of the Robinson estimator

robinson: can't have confounded in dummy. but can in real. general result of propensity stuff?

Framing: Partialling out is an alternative to OLS where $n << p$ doesn't hold. alterntive to LASSO etc

$\hat{\theta} \approx N(\theta, V/n)$

$V = (E[\hat{D}^2)^{-1} E[\hat{D}^2 \epsilon^2](E[\hat{D}^2])^{-1}$

These are robust standard errors.

Moments of the Robinson estimator

If IID then

$$Var(\hat{\theta}) = \frac{\sigma_\epsilon^2}{\sum_i (x_i - \hat{X}_i)^2}$$

Otherwise, can use GLM

What are the properties of the estimator?

$$E[\hat{\theta}] = E[\frac{\sum_i (X_i - \hat{X}_i)(y_i - \hat{y}_i)}{\sum_i (x_i - \hat{X}_i)^2}]$$

10.1.4 Non-linear treatment effects in the Robinson estimator

Page on reformulating as non-linear. can do it. show can be estimated using arg min https://arxiv.org/pdf/1712.04912.pdf

10.1.5 DML

in DML. page on orthogonality scores, page on constructing them; page on using them to estimate parameters (GMM)

We have $P(X) = f(\theta, \rho)$ $\hat{\theta} = f(X, n)$ $\theta = g(\rho, X)$

So error is: $\hat{\theta} - \theta = f(X, n) - g(\rho, X)$

Bias is defined as: $Bias(\hat{\theta}, \theta) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta$ $Bias = E[\hat{\theta} - \theta] = E[f(X, n) - g(\rho, X)]$ $Bias = E[\hat{\theta} - \theta] = E[f(X, n)] - g(\rho, X)$

double ML: regression each parametric parameter on ML of other variables. eg: get $e(x|z)$ $e(d|x)$ $d = m(x) + v$ d is correlated with x so bias. v is corrleated with d but not x. use as "iv". Still need estimate for $g(x)$.

for iterative, process is: + estimate $g(x)$ + plug into other and estimate theta + this section should be in sample splitting. rename iterative estimation. separate pages for bias, variance + how does this work?? paper says random forest regression and OLS. intialise θ randomly? + page on bias, variance, efficiency? + page on sample splitting, why?

+ page on goal: x and z orthogonal for split sampling + page on $X = m_0(Z)+\mu$, first stage machine learning, synthetic instrumental variables? h3 on that for multiple variables on interest. regression for each

10.1.6 DML1

Divide into k.

For each do ML on nuicance (how???) use all instances outside of sample

Then do GMM using orthogonality condition to calculate θ. (how??) use instances in sample

Average θ from each class

10.1.7 Last stage Robinson

Separate page for last stage: note we can do OLS, GLS etc with choice of Ω.

10.2 Causal trees