Multivariate time series

Adam Boult (www.bou.lt)

February 10, 2023

Contents

Preface

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

Part I

Stochastic processes

Chapter 1

Stochastic processes and their moments

1.1 Introduction to processes

1.1.1 Stochastic processes

In a stochastic process we have a mapping from a variable (time) to a random variable.

Discrete and continuous time

Time could be discrete, or continuous.

Temperature over time is a stochastic process, as is the number of cars sold each day.

Discrete and continous state space

The state space for temperature is continous, the number of people on the moon is discrete.

1.1.2 Stochastic evolution

We can describe processes by their evolution.

$p(x_t|x_{t-1}...)$

1.1.3 Gaussian processes

1.1.4 Moments of stochastic processes

1.1.5 Autocovariance and autocorrelation

Autocovariance

$AC(a, b) = cov(X_a, X_b)$

Autocorrelation

The autocorrelation between two time periods is their covariance, normlised by their variances

$$AC(a, b) = \frac{E[(X_a - \mu_a)(X_b - \mu_b)]}{\sigma_a \sigma_b}$$

This is also called serial correlation.

Chapter 2

White noise, and weak- and wide-sense stationarity

2.1 Stationarity

2.1.1 Weak- and wide-sense stationarity

Unconditional probabilities don't change over time.

So GDP would not be stationary, but random noise would. A random walk is not stationary, because the variance increases over time.

2.1.2 Weak-sense stationary

Mean and autocovarinance don't change over time.

2.1.3 Wide-sense stationary

All moments are the same.

2.1.4 Unit roots

2.2 Introduction

2.2.1 White noise

Variables at each time are indepdendent.

Chapter 3

Random walks

3.1 Random walks

3.1.1 Random walks

Chapter 4

Martingale processes

4.1 Introduction

4.1.1 Martingale property

For a process with the Martingale property, the expected value of all future variables is the current state.

This only restricts expectations.

$E(X_{n+1}|X_0, ..., X_n) = X_n$

Chapter 5

Markov processes

5.1 Introduction

5.1.1 Markov property

For a process with the Markov property, only the current state matters for all probability distributions.

$P(x_{t+n}|x_t) = P(x_{t+n}|x_t, x_{t-1}...)$

5.2 Markov chains

5.2.1 Finite state Markov chains

Transition matrices

This shows the probability for moving between discrete states.

We can show the probability of being in a state by multiplying the vector state by the transition matrix.

Mv

Time-homogenous Markov chains

For time-homogenous Markov chains the transition matrix is independent of time.

For these we can calculate the probability of being in any given state in the future:

$M^n v$

This becomes independent of v as we tend to infinity. The initial starting state does not matter for long term probabilities.

How to find steady state probability?

$Mv = v$

The eigenvectors! With associated eigenvector 1. There is only one eigenvector. We can find it by iteratively multiplying any vector by M.

5.2.2 Infinite state Markov chains

Markov model description We can represent the transition matrix as a series of rules to reduce the number of dimensions $P(x_t|y_{t-1}) = f(x, y)$

can represent states as number, rather than atomic. could be continuous, or even real.

in more complex, can use vectors.

5.3 Hidden Markov Models

5.3.1 Introduction

As well as the Markov process X, we have another process Y which depends on X.

5.4 Dynamic Bayesian networks

5.4.1 Introduction

Chapter 6

Multivariate time series

6.1 Multiple time series

6.1.1 Cointegration

If we have multiple variables, we can explore the order of integration of linear combinations.

If two series have time trends, a linear combination of them could remove this.

6.1.2 Exogeneity

Contemporaneous exogeneity

$Cov(x_{it}, u_{it}) = 0$

Strict exogeneity

$Cov(x_{is}, u_{it}) = 0)$

This is stronger than contemporeous, all periods.

Shocks don't affect future outcomes.

Sequential exogeneity

Sequential exogeneity: a bit looser than strict exogeneity. only holds when $s \leq t$.

So shocks can affect, but only in future.

6.1.3 Introduction

Weak stationary processes can be decomposed to a deterministic and a stochastic component.

Chapter 7

Bayesian networks

7.1 Bayesian networks

7.1.1 Bayesian networks

Chapter 8

Survival functions

8.1 Introduction

8.1.1 Survival functions

Part II

Continuous-time stochastic processes

Chapter 9

Wiener processes and Brownian motion

9.1 Wiener proceses

9.1.1 Independent increments

The changes in any non-overlapping time increments are independent.

Formally:

$t_0 < t_1 < t_2 < ... < t_m$

With X_t

$X_{t_1} - X_{t_0}$ is indepentent from $X_{t_2} - Xt_1$ etc.

9.1.2 Wiener processes

A Wiener process is a process W_t with independent increments, which: + Is continuous + Has normally distributed increments.

Can be constructed as limit of random walk. Can also be constructed as integral of Gaussian noise?

9.2 Brownian motion

9.2.1 Brownian motion

brownian motion in stats. given we start at a, what is chance be end up at b? normal. do 1d then multi d

Chapter 10

Stochastic differential equations

Part III

Discrete-time stochastic processes

Chapter 11

Orders of integration

11.1 Introduction

11.1.1 Orders of integration

How many diffs do you need to do to get a stationary process?

If something is first order integrated it is $I(1)$.

11.1.2 Trend stationary

If we can remove the trend as a function, eg linear or non-linear growth, and the rest is stationary, then the process is trend stationary

11.1.3 Seasonal and non-seasonal trends

We can model the process as:

$y_t = \mu_t + f(t) + \epsilon_t$

11.1.4 Cyclical fluctuations

We can have shocks having effects over time.

This is separate to trends.

Chapter 12

Auto-Regressive processes, Moving-Average processes and Wold's theorem

12.1 Autoregressive model

12.1.1 Autoregressive models (AR)

AR(1)

Our basic model was:

$x_t = \alpha + \epsilon_t$

We add an autoregressive component by adding a lagged observation.

$x_t = \alpha + \beta x_{t-1} + \epsilon_t$

AR(p)

AR(p) has p previous dependent variables.

$x_t = \alpha + \sum_{i=1}^{p} \beta_i x_{t-i}$

Propagation of shocks

A shock bumps up the output variable, which bumps up output variables forever, at a decreasing rate.

12.1.2 Testing for stationarity with Dickey-Fuller (DF) and Augmented Dicky-Fuller (ADF)

Stationarity

Unit roots

Integration order

Dickey-Fuller

The Dickey-Fuller test tests if there is a unit root.

The AR(1) model is:

$y_t = \alpha + \beta y_{t-1} + \epsilon_t$

We can rewrite this as:

$\Delta y_t = \alpha + (\beta - 1) y_{t-1} + \epsilon_t$

We test if $\beta - 1) = 0$.

If the coefficient on the last term is 1 we have a random walk, and the process is non-stationary.

If the last term is < 1 then we have a stationary process.

Variation: Removing the drift

If our model has no intercept it is:

$y_t = \beta y_{t-1} + \epsilon_t$

$\Delta y_t = (\beta - 1) y_{t-1} + \epsilon_t$

Variation: Adding a deterministic trend

If our model has a time trend it is:

$y_t = \alpha\beta y_{t-1} + \gamma t + \epsilon_t$

$\Delta y_t = \alpha + (\beta - 1) y_{t-1} + \gamma t + \epsilon_t$

Augmented Dickey-Fuller

We include more lagged variables.

$y_t = \alpha + \beta t + \sum_i^p \theta_i y_{t-i} + \epsilon_t$

If no unit root, can do normal OLS?

12.1.3 Autoregressive Conditional Heteroskedasticity (ARCH)

Variance of the AR(1) model

The standard AR(1) model is:

$y_t = \alpha + \beta y_{t-1} + \epsilon_t$

The variance is:

$Var(y_t) = Var(\alpha + \beta y_{t-1} + \epsilon_t)$

$Var(y_t)(1 - \beta^2) = Var(\epsilon_t)$

Assuming the errors are IID we have:

$$Var(y_t)) = \frac{\sigma^2}{1 - \beta^2}$$

This is independent of historic observations, which may not be desirable.

Conditional variance

Consider the alternative formulation:

$y_t = \epsilon_t f(y_{t-1})$

This allows for conditional heteroskedasticity.

12.2 Moving average models

12.2.1 Moving Average models (MA)

We add previous error terms as input variables

MA(q) has q previous error terms in the model

Unlike AR models, the effects of any shocks wear off after q terms.

This is harder to fit the OLS, the error terms themselves are not observed.

12.3 Autoregressive Moving Average models

12.3.1 Autoregressive Moving Average models (ARMA)

We include both AR and MA

Estimted using Box-Jenkins

12.3.2 Autoregressive Integrated Moving Average models (ARIMA)

Uses differences to remove non statiority

Also estiamted with box-jenkins

12.3.3 Seasonal ARIMA

12.4 Wold's theorem

12.4.1 Introduction

Chapter 13

Vector Autoregression (VAR)

13.1 Vector Autoregression (VAR)

13.1.1 Vector Autoregression (VAR)

We consider a vector of observables, not just one

Autoregressive (AR) model for a vector.

VAR(p) looks p back.

The AR(p) model is:

$y_t = \alpha + \sum_{i=1}^{p} \beta y_{t-i} + \epsilon_t$

VAR(p) generalises this to where y_t is a vector. We define VAR(p) as:

y_t

$y_t = c + \sum_{i=1}^{p} A_i y_{t-i} + \epsilon_t$

13.1.2 VAR impulse response

13.1.3 Bayesian VAR

13.2 Structural models

13.2.1 Autoregressive Distributed Lag (ARDL) model

Include lagged y and lagged x (and current x)

Chapter 14

ARMAX

14.1 ARMAX

14.1.1 ARMAX

14.1.2 ARIMAX

14.1.3 SARIMA

Chapter 15

Partial Adjustment Model (PAM)

15.1 Partial Adjustment Model

15.1.1 Partial Adjustment Model

Estimating a static model

We start by estimating a static model.

$y_t = \alpha + \theta x_t + \gamma_t$

Equilibrium

We then use this form an equilibrium for y_t, y_t^*.

$y_t^* = \hat{\alpha} + \hat{\theta} x_t$

The process depends on the difference from this equilibrium.

$y_t - y_{t-1} = \beta(y_t^* - y_{t-1}) + \epsilon_t$

$y_t - y_{t-1} = \beta(\hat{\alpha} + \hat{\theta} x_t - y_{t-1}) + \epsilon_t$

$y_t = \beta\hat{\alpha} + \beta\hat{\theta} x_t + (1-\beta) y_{t-1} + \epsilon_t$

$y_t = \alpha y_{t-1} + (1-\beta)(y_t^* - y_{t-1}) + \epsilon$

The higher β, the slower the adjustment.

If stationary, can we can use OLS.

Chapter 16

Error Correction Model

16.1 Error Correction Model

16.1.1 Error Correction Model

Static model

Like PAM we start with static estimator.

The ECM

The ECM does a regression with first differences, and includes lagged error terms.

We start with a basic first-difference model.

$\Delta y_t = \Delta x_t$

We could also expand this to include laggs for both x and y. Here we don't.

We know that long term $y_t = \theta x_t$. We use the error from this in a first difference model.

$\Delta y_t = \alpha \Delta x_t + \beta(y_{t-1} - \theta x_{t-1})$

Page on identifying error terms

Also, page on Vector Error Correction Model (VECM)

Part IV

Signal processing

Chapter 17

Quantisation and sample rates

17.1 Introduction

17.1.1 Quantisation

17.1.2 Sample rate

Chapter 18

Discrete Fourier Transform

18.1 Introduction

18.1.1 Discrete Fourier Transform

Chapter 19

Down sampling

19.1 Introduction

19.1.1 Down sampling

Chapter 20

Fast Fourier Transform

20.1 Introduction

20.1.1 Fast Fourier Transform

Chapter 21

Noisy networks

21.1 Introduction

21.1.1 Noisy networks

Part V

Estimating time series models

Chapter 22

Estimating Markov chains

22.1 Estimating Markov chains

22.1.1 Estimating the Markov chain stochastic matrix

Introduction

Given a sequence: $x_1, ...x_n$.

The likelihood is:

$L = \prod_{i=2}^{n} p_{x_{i-1}, x_i}$

If there are k states we can rewrite this as:

$L = prod_{i=1}^{k} \prod_{j=1}^{k} n_{ij} p_{ij}$

Where p_{ij} is the chance of moving from state i to state j, and n_{ij} is the number of transtions between i and j.

The log likelhood is:

$\ln L = \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} \ln p_{ij}$

Constrained optimisation

Not all parameters are free. All probabilities must sum to 1.

$\ln L = \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} \ln p_{ij} - sum_{i=1} \lambda_i (\sum_{j=1} p_{ij} - 1)$

This gives us:

$\hat{p}_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}$

22.1.2 Estimating infinite state Markov chains

We can represent the transition matrix as a series of rules to reduce the number of dimensions

$P(x_t|y_{t-1}) = f(x, y)$

can represent states as number, rather than atomic. could be continuous, or even real.

in more complex, can use vectors.

22.2 Ergodic processes

22.2.1 Ergodic processes

Sample moments must converge to generating momements. Not guaranteed.

Eg process with path dependence. 50

Generating average is £50, but sample will only convergen to £100 or £0

Chapter 23

Estimating Hidden Markov Models (HMMs)

23.1 Estimating Hidden Markov Models (HMMs)

23.1.1 Recap of Hidden Markov Models (HMMs)

We don't see state

Each state produces a visible output. this output is drawn from a distribution for each state.

We observe a sequence of outputs, not states.

23.1.2 Estimating HMMs with the Viterbi algorithm

Assume we know transition matrix. and starting probls

Given we observe sequence of outputs, what were most likely actual paths?

Virbiti returns this

23.1.3 Estimating HMMs with the forward algorithm

Given we have observed outputs, what is the chance of being in a certain state at a certain time?

23.1.4 Estimating HMMs with the forward-backward algorithm

We calculate state x at time t given all obs.

23.1.5 Baum-Welch algorithm

23.1.6 Kalman filters

Chapter 24

Univariate forecasting

24.1 Introduction

24.1.1 Seasonal and non-seasonal trends

We can model the process as:

$y_t = \mu_t + f(t) + \epsilon_t$

24.1.2 Identifying the order of integration using Augmented Dickey-Fuller

The Dickey-Fuller test with deterministic time trend was:

$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \epsilon_t$

The Augmented Dickey-Fuller model adds lags for the differences.

$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_i^p \delta_i \Delta y_{t-i} + \epsilon_t$

24.1.3 Cyclical fluctuations

We can have shocks having effects over time.

This is separate to trends.

24.1.4 Identifying serial correlation using the Durbin-Watson statistic

24.1.5 Introduction to forecasting

We observe a series of observations:

$x_1, x_2, ..., x_t)$

What can we say about x_{t+1}?

If the data was drawn iid then the past data then we would just want to identify moments.

However if the data is not iid, for example because it is increasing in time, then this is not the best way.

Regression formation

We can model

$x_t = \alpha + \epsilon_t$

24.2 Autoregressive model

24.2.1 Autoregressive models (AR)

AR(1)

Our basic model was:

$x_t = \alpha + \epsilon_t$

We add an autoregressive component by adding a lagged observation.

$x_t = \alpha + \beta x_{t-1} + \epsilon_t$

AR(p)

AR(p) has p previous dependent variables.

$x_t = \alpha + \sum_{i=1}^{p} \beta_i x_{t-i}$

Propagation of shocks

A shock bumps up the output variable, which bumps up output variables forever, at a decreasing rate.

24.2.2 Testing for stationarity with Dickey-Fuller (DF) and Augmented Dicky-Fuller (ADF)

Stationarity

Unit roots

Integration order

Dickey-Fuller

The Dickey-Fuller test tests if there is a unit root.

The AR(1) model is:

$y_t = \alpha + \beta y_{t-1} + \epsilon_t$

We can rewrite this as:

$\Delta y_t = \alpha + (\beta - 1) y_{t-1} + \epsilon_t$

We test if $\beta - 1) = 0$.

If the coefficient on the last term is 1 we have a random walk, and the process is non-stationary.

If the last term is < 1 then we have a stationary process.

Variation: Removing the drift

If our model has no intercept it is:

$y_t = \beta y_{t-1} + \epsilon_t$

$\Delta y_t = (\beta - 1) y_{t-1} + \epsilon_t$

Variation: Adding a deterministic trend

If our model has a time trend it is:

$y_t = \alpha\beta y_{t-1} + \gamma t + \epsilon_t$

$\Delta y_t = \alpha + (\beta - 1) y_{t-1} + \gamma t + \epsilon_t$

Augmented Dickey-Fuller

We include more lagged variables.

$y_t = \alpha + \beta t + \sum_i^p \theta_i y_{t-i} + \epsilon_t$

If no unit root, can do normal OLS?

24.2.3 Autoregressive Conditional Heteroskedasticity (ARCH)

Variance of the AR(1) model

The standard AR(1) model is:

$y_t = \alpha + \beta y_{t-1} + \epsilon_t$

The variance is:

$Var(y_t) = Var(\alpha + \beta y_{t-1} + \epsilon_t)$

$Var(y_t)(1 - \beta^2) = Var(\epsilon_t)$

Assuming the errors are IID we have:

$$Var(y_t)) = \frac{\sigma^2}{1 - \beta^2}$$

This is independent of historic observations, which may not be desirable.

Conditional variance

Consider the alternative formulation:

$y_t = \epsilon_t f(y_{t-1})$

This allows for conditional heteroskedasticity.

24.3 Moving average models

24.3.1 Moving Average models (MA)

We add previous error terms as input variables

MA(q) has q previous error terms in the model

Unlike AR models, the effects of any shocks wear off after q terms.

This is harder to fit the OLS, the error terms themselves are not observed.

24.4 Autoregressive Moving Average models

24.4.1 Autoregressive Moving Average models (ARMA)

We include both AR and MA

Estimted using Box-Jenkins

24.4.2 Autoregressive Integrated Moving Average models (ARIMA)

Uses differences to remove non statiority

Also estiamted with box-jenkins

24.4.3 Seasonal ARIMA

24.5 Forecasting

24.5.1 Monte carlo simulations

24.5.2 N-step ahead

24.5.3 Consensus forecasting

24.6 Other

24.6.1 Identifying the order of integration using Augmented Dickey-Fuller

The Dickey-Fuller test with deterministic time trend was:

$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \epsilon_t$

The Augmented Dickey-Fuller model adds lags for the differences.

$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_i^p \delta_i \Delta y_{t-i} + \epsilon_t$

24.6.2 Identifying serial correlation using the Durbin-Watson statistic

Chapter 25

Multivariate forecasting

25.1 Introduction to multiple time series

25.1.1 Testing for cointegration with Johansen

25.2 Vector Autoregression (VAR)

25.2.1 Vector Autoregression (VAR)

We consider a vector of observables, not just one

Autoregressive (AR) model for a vector.

VAR(p) looks p back.

The AR(p) model is:

$y_t = \alpha + \sum_{i=1}^{p} \beta y_{t-i} + \epsilon_t$

VAR(p) generalises this to where y_t is a vector. We define VAR(p) as:

y_t

$y_t = c + \sum_{i=1}^{p} A_i y_{t-i} + \epsilon_t$

25.2.2 VAR impulse response

25.2.3 Bayesian VAR

25.3 Structural models

25.3.1 Autoregressive Distributed Lag (ARDL) model

Include lagged y and lagged x (and current x)

If the processes are stationary, then we can use OLS. THIS IS A BROADER POINT! INTRO??

25.4 ARMAX

25.4.1 ARMAX

25.4.2 Error Correction Model

Static model

Like PAM we start with static estimator.

The ECM

The ECM does a regression with first differences, and includes lagged error terms.

We start with a basic first-difference model.

$\Delta y_t = \Delta x_t$

We could also expand this to include laggs for both x and y. Here we don't.

We know that long term $y_t = \theta x_t$. We use the error from this in a first difference model.

$\Delta y_t = \alpha \Delta x_t + \beta(y_{t-1} - \theta x_{t-1})$

Page on identifying error terms

Also, page on Vector Error Correction Model (VECM)

25.4.3 Partial Adjustment Model

Estimating a static model

We start by estimating a static model.

$y_t = \alpha + \theta x_t + \gamma_t$

Equilibrium

We then use this form an equilibrium for y_t, y_t^*.

$y_t^* = \hat{\alpha} + \hat{\theta} x_t$

The process depends on the difference from this equilibrium.

$y_t - y_{t-1} = \beta(y_t^* - y_{t-1}) + \epsilon_t$

$y_t - y_{t-1} = \beta(\hat{\alpha} + \hat{\theta} x_t - y_{t-1}) + \epsilon_t$

$y_t = \beta\hat{\alpha} + \beta\hat{\theta}x_t + (1-\beta)y_{t-1} + \epsilon_t$

$y_t = \alpha y_{t-1} + (1-\beta)(y_t^* - y_{t-1}) + \epsilon$

The higher β, the slower the adjustment.

If stationary, can we can use OLS.

Chapter 26

Inference with time series

26.1 OLS on time series data

26.1.1 Bias of static models and spurious correlations

Static models

Static models are of the form:

$y_t = \alpha + \beta x_t + \epsilon_t$

These have no lagged variables or difference operators.

Bias of static models

26.1.2 Heteroskedasticity and Autocorrelation (HAC) adjusted standard errors

26.2 Time series

26.2.1 Taking differences

What we use should depend on I(1), I(0) etc from ADF

if we're missing time invariant data, we can do first differences and this isn't a problem if we do diff in diff this removes trends?

page on first difference estiamtion? OLS on first differences. No other lags page on first difference ESTIMATOR

26.2.2 Discontinuity

Create a dummy for before/after a date.

26.3 Panel data

26.3.1 Difference-in-difference

Consider the grouped linear model:

$y_{ij} = \mu + \tau_i + X_j\theta + \epsilon_{ij}$

By taking differences with another observation in the same group we remove the average terms.

$y_{ij} - y_{ik} = (\mu + \tau_i + X_j\theta + \epsilon_{ij}) - (\mu + \tau_i + X_k\theta + \epsilon_{ik})$

$y_{ij} - y_{ik} = (X_j\theta - X_k) + (\epsilon_{ij} - \epsilon_{ik})$

diff in diff: control group and treated group. page on leakiness? are control affected too? Assumption: in absense of treatment, price would have evolved like control

26.3.2 Controlled experiments

26.3.3 Natural experiments

26.3.4 Structural breaks

Testing for structural breaks with the Chow test.

26.3.5 Dynamic or lagged independent variables

Static panel data: No lags of independent variables. Dynamic panel data: Lags of independent variables.

OLS is consistent for static panel data, not for dynamic This results in Nickell's bias for dynamic panel data

Dynamic panel data: y_{t-1} is a regressor Panel data estimation: LSDV. Least squares dummy variable estimator arnello bond

Chapter 27

Survival analysis

27.1 Introduction

27.1.1 Cox-hazard

Part VI

Advanced inference (time)

Chapter 28

Imputing missing data for time series

28.1 Time series

28.1.1 ARIMA interpolation

28.1.2 Last Observation Carried Forward (LOCF)

28.1.3 Next Observation Carried Backward (NOCB)

28.1.4 Other

Multi period averages for imputation on time series.

Chapter 29

Homogeneous treatment effects

29.1 Introduction

29.1.1 Treatment data

Recap

With multilevel data with fixed coefficients we have:

$y_{ij} = \mathbf{x}_{ij}\theta + m_j + \epsilon_{ij}$

We can estimate m_j using fixed effects or similar methods.

Treatment data

If the data is grouped by whether an entity was treated then will have:

- y_{i0} - the outcome if the entity was not treated
- y_{i1} - the outcome if the entity was treated

However we only observe y_i and D_i.

$y_i = y_{i0} + D_i(y_{i1} - y_{10})$

29.1.2 Average Treatment Effects (ATE, ATET, ATEUT)

Average Treatment Effect (ATE)

$ATE = E[y_{i1} - y_{i0}]$

Average Treatment Effect on the Treated (ATET)

$ATE = E[y_{i1} - y_{i0}|D_i = 1]$

$ATE = E[y_{i1}|D_i = 1] - E[y_{i0}|D_i = 1]$

Average Treatment Effect on the Untreated (ATEUT)

29.1.3 Conditional Average Treatment Effect (CATE)

$E[y_{i1} - y_{i0}|\mathbf{x}_i]$

29.2 Exogenous treatment

29.2.1 Randomly Controlled Trials (RCTs)

If the model is:

$y_i = D_i\theta + g(X) + \epsilon_i$

And D is randomly assigned, then we can estimate

$y_i = D_i\theta + \epsilon_i$

To get an estimate for θ without collecting data on X.

29.2.2 Calculating CATEs in RCTs with interaction terms

29.2.3 Calculating CATEs in RCTs with subgroup analysis

29.3 Calculating treatment effects without estimating missing data

29.3.1 Regression

We can simply regress outcomes on variables, including treatment.

This assumes treatment effects are constant.

This also assumes that outcomes y_{1i} and y_{0i} are independent of D_i, conditional on X.

If we are missing variables in X then we will have biased estimates.

This also assumes the effects of X are linear.

We assume: $E[y_{0i}|\mathbf{x}_i, D_i] = \mathbf{x}_i\theta$.

29.3.2 Instrumental Variables and natural experiments

29.3.3 Regression discontinuity

29.3.4 Synthetic controls

29.4 Calculating treatment effects by estimating missing data

29.4.1 Matching

Matching is similar to regression. We assume that effects are constant, and the effect of treatment on y_{0i} and y_{1i} are independent of treatment, once controlling for X.

Again, this is biased if this is not the case.

We however do not have to assume a linear form for X.

We assume: $E[y_{ji}|\mathbf{x}_i, D_i] = E[y_{ji}|\mathbf{x}_i]$

For each entity, find a near entity which had the opposite treatment.

29.4.2 Propensity score matching

Match on the chance of getting treatment, given covariates.

29.4.3 Matrix completion

$E[y_{i1} - y_{i0}|\mathbf{x}_i]$

29.5 Using semi-parametric

29.6 Other

29.6.1 Estimating ATE using MCMC

29.6.2 Local Average Treatment Effect (LATE)

We have IVs for treatment.

29.6.3 Treatment effects

+ propsentiy score weighting + regression adjustemnt + matching + IV + Regression discontinuity

29.6.4 Meta analysis

big page in advanced analytics? Random effects meta analysis?

meta analysis: fixed effect v random effects model

types of study: + RCT + cohort studies + case-control studies + cross sectional studies

29.6.5 Dose response curve

29.6.6 Sensitivity analysis

29.6.7 Page on Rubin causal model

Chapter 30

Heterogeneous treatment effects

30.1 Heterogenous treatment effects

30.1.1 Introduction

30.1.2 subgroup analysis

30.1.3 interaction terms

30.1.4 efficient policy learning

30.1.5 Het DML

$y = a(z) + db(z)$ Het effects is $b(z)$ We build groups instead of arbitrary function. So we estimate $E[b(z)|G]$

Use part of the data set to estimate

$\hat{y} = \hat{a}(z) + D\hat{b}(z)$

Use $s = \hat{b}(z)$ to stratify. Key point is defining subgroups algorithmically. Less opportunity for hacking

30.1.6 Continuous treatment effects

30.1.7 Intent-to-treat

30.2 (LATE, causal tree (from CART))

30.2.1 Introduction

bart causal is different to causal tree

In stuff now two problems: + non random but constant effect + Random but heterogenous effect

causal trees can find heterogenous treatment effects

Approaches: We have treated and untreated. X and y Estimate y—x for treated, and untreated separately. Then take differnece for a given x to be the estimated treatment effect

2nd approach: have treatment as input diffence is again y—x - y—x treatment minus no treatment

3rd approach: (type of single tree) split not by predictive power, but by treatment effect differnece

4th approach: cross validation at each leaft we note the sample average treatment effect goal is to choose hyper parameters which minimise sum of diffence between these and cross valid data

Once we have the trees from the last one, calculate the effect using test data. nb: separate creating of tree to estimation of treatment effect

30.2.2 Instrumental forests

Estimate LATE

like causal forest, but do IV regression on leaf.

Chapter 31

Causal trees

31.1 Causal trees

31.1.1 Measuring treatment effects in leaves

31.1.2 Sample splitting for treatment effects

31.1.3 Honest trees

We use part of the sample to estimate Θ, and another part of the sample to estimate the treatment effect.

31.1.4 Estimating ATE using MCMC

31.2 Ensemble methods for causal trees

31.2.1 Causal forests

31.2.2 Bayesian causal forests

Part VII

Sampling

Chapter 32

Markov chain Monte Carlo sampling

32.1 Markov Chain Monte Carlo (MCMC) methods

32.2 Metropolis-Hastings algorithm

32.2.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm creates a set of samples x such that the distribution of the samples approaches the goal distribution.

Initialisation

The algorithm takes an arbitrary starting sample x_0. It then must decide which sample to consider next.

Generation

It does this using a Markov chain. That is, there is a map $g(x_j, x_i)$.

This distribution is generally a normal distribution around x_i, making the process a random walk.

Acceptance

Now we have a considered sample, we can either accept or reject it. It is this step that makes the end distribution approximage the function.

We accept if $\frac{f(x_j)}{f(x_i)} > u$, where u is a random variable between 0 and 1, generated each time.

We can calculate this because we know this function.

Properties

32.3 Gibb's sampling

32.3.1 Gibb's sampling

Introduction

As with Metropolis-Hastings, we want to generate samples for $P(X)$ and use this to approximate its form.

We do this by using the conditional distribution. If X is a vector then we also have:

$P(x_j|x_0, ..., x_{j-1}, x_{j+1}, ..., x_n)$

We use our knowledge of this distribution.

Start with vector x_0.

This has components $x_{0,j}$

To form the next vector x_1 we loop through each component.

$P(x_{1,0}|x_{0,0}, x_{0,1}, ..., x_{0,n})$

We use this to form $x_{1,0}$

However after th the first component we update this so it uses the updated variables.

$P(x_{1,k}|x_{1,0}, ..., x_{1,k-1}, x_{0,k}, ..., x_{0,n}$

This means we only need to know the conditional distributions.

Chapter 33

Sampling from processes

33.1 Introduction

Chapter 34

Forecasting stochastic processes

34.1 Forecasting

34.1.1 Introduction to forecasting

We observe a series of observations:

$x_1, x_2, ..., x_t)$

What can we say about x_{t+1}?

If the data was drawn iid then the past data then we would just want to identify moments.

However if the data is not iid, for example because it is increasing in time, then this is not the best way.

Regression formation

We can model

$x_t = \alpha + \epsilon_t$

34.1.2 Monte carlo simulations

34.1.3 N-step ahead

34.1.4 Consensus forecasting