Bayesuvius,
a small visual dictionary of Bayesian Networks

Robert R. Tucci
www.ar-tiste.xyz

February 7, 2021

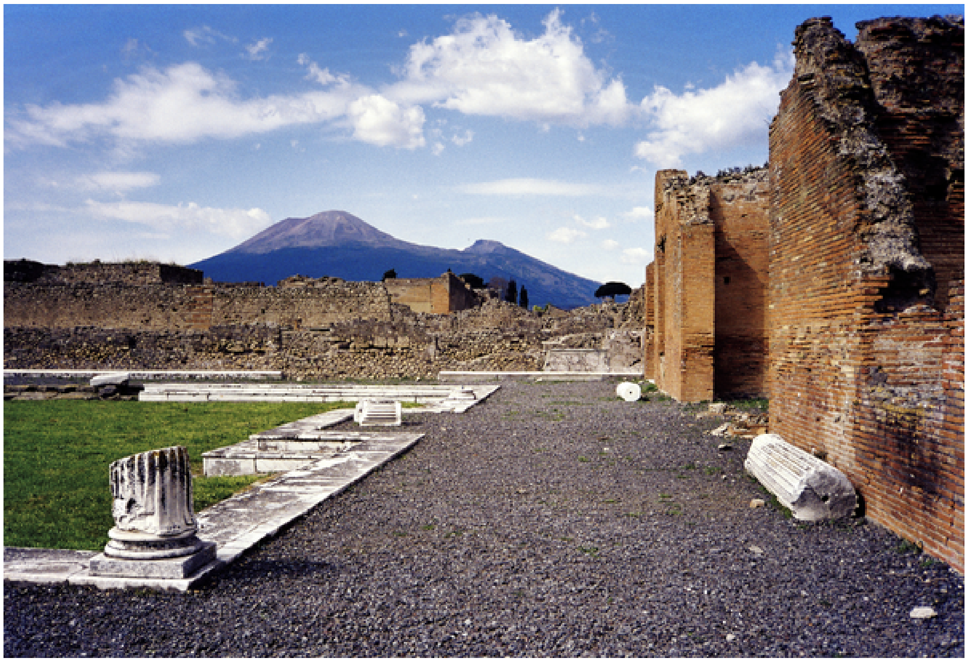


Figure 1: View of Mount Vesuvius from Pompeii

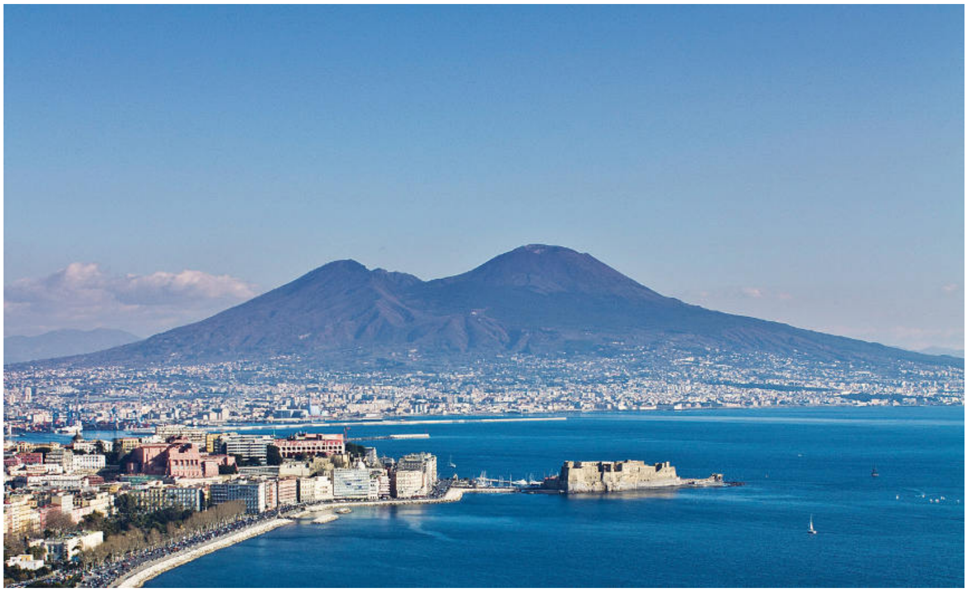


Figure 2: Mount Vesuvius and Bay of Naples

Contents

Foreword **8**

Navigating the ocean of Judea Pearl's Books **9**

Notational Conventions and Preliminaries **10**
0.1 Some abbreviations frequently used throughout this book. 10
0.2 $\mathcal{N}(!a)$ 10
0.3 One hot 10
0.4 Special sets 10
0.5 Kronecker delta function 11
0.6 Dirac delta function 11
0.7 Indicator function (aka Truth function) 11
0.8 Underlined letters indicate random variables 11
0.9 Probability distributions 11
0.10 Discretization of continuous probability distributions 12
0.11 Samples, i.i.d. variables 12
0.12 Normal Distribution 13
0.13 Uniform Distribution 13
0.14 Sigmoid and logit functions 13
0.15 Expected Value and Variance 13
0.16 Conditional Expected Value 14
0.17 Law of Total Variance 14
0.18 Notation for covariances 15
0.19 Linear regression, Ordinary Least Squares (OLS) 16
0.19.1 LR, assuming x_σ are non-random 17
0.19.2 LR, assuming x_σ are random and i.i.d. 19
0.20 Short Summary of Boolean Algebra. 20
0.21 Entropy, Kullback-Liebler divergence 21
0.22 Definition of various entropies used in Shannon Information Theory 22

Definition of a Bayesian Network **23**

1 ARACNE-structure learning **26**

2 Backdoor Adjustment **28**
2.1 Examples 29

3 Back Propagation (Automatic Differentiation) **32**
3.1 General Theory 32
3.1.1 Jacobians 32
3.1.2 bnets for function composition, forward propagation and back propagation . 32
3.2 Application to Neural Networks 34
3.2.1 Absorbing b_i^λ into $w_{i|j}$. 34
3.2.2 bnets for function composition, forward propagation and back propagation for NN 35
3.3 General bnets instead of Markov chains induced by layered structure of NNs 38

4 Basic Curve Fitting Using Gradient Descent **39**

5 Bell and Clauser-Horne Inequalities in Quantum Mechanics **41**

6 Binary Decision Diagrams **42**

7 Chow-Liu Trees and Tree Augmented Naive Bayes (TAN) **46**
7.1 Chow-Liu Trees 46
7.2 Tree Augmented Naive Bayes (TAN) 50

8 Counterfactual Reasoning **52**
8.1 The 3 Rungs of Causal AI 52
8.2 Two kinds of intervention operators 52
8.3 Do operator for DEN diagrams 53
8.4 Mediation Analysis 56

9 Decision Trees **59**

10 Difference-in-Differences **62**
10.1 John Snow, DID and a cholera transmission pathway 62
10.2 PO analysis 63
10.3 Linear Regression 66

11 Digital Circuits **68**
11.1 Mapping any dcircuit to a bnet 69
11.1.1 Option A of Fig.11.2 69
11.1.2 Option B of Fig.11.2 69

12 Do-Calculus **70**
12.1 Parent Adjustment 73
12.2 3 Rules of do-calculus 74

12.3 Backdoor Adjustment . . . 74
12.4 Front Door Adjustment . . . 76

13 D-Separation 80

14 Dynamic Bayesian Networks 83

15 Expectation Maximization 84
15.1 The EM algorithm: . . . 85
15.1.1 Motivation . . . 86
15.2 Minorize-Maximize (MM) algorithms . . . 86
15.3 Examples . . . 88
15.3.1 Gaussian mixture . . . 88
15.3.2 Blood Genotypes and Phenotypes . . . 89
15.3.3 Missing Data/Imputation . . . 90

16 Front-door Adjustment 91
16.1 Examples . . . 92

17 Generative Adversarial Networks (GANs) 93

18 Gaussian Nodes with Linear Dependence on Parents 98

19 Hidden Markov Model 101

20 Influence Diagrams & Utility Nodes 105

21 Instrumental Inequality and beyond 107
21.1 I-inequality . . . 107
21.1.1 I-inequality for binary z,d,y . . . 109
21.2 Bounds on Effect of IV on treatment outcome y . . . 109

22 Instrumental Variables 112
22.1 δ with unnmeasured confounder . . . 112
22.2 δ (with unmeasured confounder) can be inferred via IV . . . 113
22.3 More general bnets with IVs . . . 114
22.4 Instrumental Inequality . . . 114

23 Junction Tree Algorithm 116

24 Kalman Filter 117
24.1 Problem . . . 118
24.2 Solution . . . 118

25 Linear and Logistic Regression 120
25.1 Generalization to x with multiple components(features) 122
25.2 Alternative $V(b, m)$ for logistic regression 122

26 Linear Deterministic Bnets with External Noise 124
26.1 Example of LDEN diagram 124
26.2 Fully Connected LDEN diagrams 125
26.2.1 Fully connected LDEN diagram with $nx = 2$ 126
26.2.2 Fully connected LDEN diagram with $nx = 3$ 126
26.2.3 Fully connected LDEN diagram with arbitrary nx 127
26.3 Non-linear DEN diagrams 128

27 Markov Blankets 130

28 Markov Chain Monte Carlo (MCMC) 132
28.1 Inverse Cumulative Sampling 132
28.2 Rejection Sampling 134
28.3 Metropolis-Hastings Sampling 135
28.4 Gibbs Sampling 138
28.5 Importance Sampling 139

29 Markov Chains 141

30 Message Passing (Belief Propagation) 142
30.1 Distributed Soldier Counting 142
30.2 Spring Systems 144
30.3 BP for Markov Chains 144
30.4 BP Algorithm for Polytrees 151
30.4.1 How BP algo for polytrees reduces to the BP algo for Markov chains 155
30.5 Derivation of BP Algorithm for Polytrees 155
30.6 Example of BP algo for a Tree 159
30.7 Bipartite bnets 163
30.8 BP for bipartite bnets (BP-BB) 165
30.8.1 BP-BB and general BP agree on Markov chains 167
30.8.2 BP-BB and general BP agree on tree bnets. 169
30.9 BP-BB and sum-product decomposition 170

31 Missing Data, Imputation 172
31.1 Imputation via EM 173
31.2 Imputation via MCMC 176
31.3 Multiple Imputations 177

32 Monty Hall Problem 178

33 Naive Bayes 180

34 Neural Networks 181
34.1 Activation Functions $\mathcal{A}_i^\lambda : \mathbb{R} \to \mathbb{R}$ . . . 182
34.2 Weight optimization via supervised training and gradient descent . . . 184
34.3 Non-dense layers . . . 186
34.4 Autoencoder NN . . . 187

35 Noisy-OR gate 188
35.1 3 ways to interpret the parameters π_i . . . 189

36 Non-negative Matrix Factorization 192
36.1 Bnet interpretation . . . 192
36.2 Simplest recursive algorithm . . . 193

37 Observational Equivalence of DAGs 194
37.1 Examples . . . 194

38 Potential Outcomes 197
38.1 G and G_{den}, bnets, the starting point bnets . . . 198
38.2 G_{do+} bnet . . . 200
38.3 G_{im+} bnet . . . 201
38.4 Translation Dictionary . . . 203
38.5 $\mathcal{Y}_{d|\tilde{d}}$ differences (aka treatment effects) . . . 203
38.6 Ignorability . . . 205
38.7 Hypothesis testing for sharp null . . . 206
38.8 Matching Strata . . . 208
38.8.1 1-1 strata-match . . . 208
38.8.2 Exact and approximate strata-match . . . 209
38.8.3 Positivity . . . 210
38.9 Propensity Score . . . 210

39 Program evaluation and review technique (PERT) 213
39.1 Example . . . 215

40 Recurrent Neural Networks 218
40.1 Language Sequence Modeling . . . 221
40.2 Other types of RNN . . . 222
40.2.1 Long Short Term Memory (LSTM) unit (1997) . . . 223
40.2.2 Gated Recurrence Unit (GRU) (2014) . . . 225

41 Regression Discontinuity Design 227
41.1 PO analysis . . . 227
41.2 Linear Regression . . . 228

42 Reinforcement Learning (RL) 230
42.1 Exact RL bnet 233
42.2 Actor-Critic RL bnet 235
42.3 Q function learning RL bnet 237

43 Reliability Box Diagrams and Fault Tree Diagrams 239
43.1 Minimal Cut Sets 246

44 Restricted Boltzmann Machines 247

45 Scoring the Nodes of a Learned Bnet 249
45.1 Probability Distributions and Special Functions 250
45.2 Single node with no parents 251
45.3 Multiple nodes with any number of parents 253
45.4 Bayesian Scores 255
45.5 Information Theoretic scores 256

46 Simpson's Paradox 257
46.1 Pearl Causality 259
46.2 Numerical Example 260

47 Structure and Parameter Learning for Bnets 261
47.1 Overview 261
47.2 Score based SL algorithms 263
47.3 Constraint based SL algorithms 263
47.4 Pseudo-code for some bnet learning algorithms 265

48 Turbo Codes 267
48.1 Decoding Algorithm 270
48.2 Message Passing Interpretation of Decoding Algorithm 272

49 Variational Bayesian Approximation 273
49.1 Free Energy $\mathcal{F}(\vec{x})$ 275

50 Zero Information Transmission (Graphoid Axioms) 278
50.1 Consequences of Eq.(50.2) 278

Bibliography 281

Notational Conventions and Preliminaries

0.1 Some abbreviations frequently used throughout this book.

- bnet= B net= Bayesian Network
- CPT = Conditional Probabilities Table, same as TPM
- DAG = Directed Acyclic Graph
- i.i.d.= independent identically distributed.
- RCT= Randomized Controlled Trial, aka A/B testing.
- TPM= Transition Probability Matrix, same as CPT

0.2 $\mathcal{N}(!a)$

$\mathcal{N}(!a)$ will denote a normalization constant that does not depend on a. For example, $P(x) = \mathcal{N}(!x)e^{-x}$ where $\int_0^\infty dx\ P(x) = 1$.

0.3 One hot

A **one hot** vector of zeros and ones is a vector with all entries zero with the exception of a single entry which is one. A **one cold** vector has all entries equal to one with the exception of a single entry which is zero. For example, if $x^n = (x_0, x_1, \ldots, x_{n-1})$ and $x_i = \delta(i, 0)$ then x^n is one hot.

0.4 Special sets

Define $\mathbb{Z}, \mathbb{R}, \mathbb{C}$ to be the integers, real numbers and complex numbers, respectively.

For $a < b$, define $I_\mathbb{Z}$ to be the integers in the interval I, where $I = [a, b], [a, b), (a, b], (a, b)$ (i.e, I can be closed or open on either side).

$A_{>0} = \{k \in A : k > 0\}$ for $A = \mathbb{Z}, \mathbb{R}$.

0.5 Kronecker delta function

For x, y in discrete set S,

$$\delta(x, y) = \begin{cases} 1 \text{ if } x = y \\ 0 \text{ if } x \neq y \end{cases} \tag{1}$$

0.6 Dirac delta function

For $x, y \in \mathbb{R}$,

$$\int_{-\infty}^{+\infty} dx\ \delta(x - y) f(x) = f(y) \tag{2}$$

0.7 Indicator function (aka Truth function)

$$\mathbb{1}(\mathcal{S}) = \begin{cases} 1 \text{ if } \mathcal{S} \text{ is true} \\ 0 \text{ if } \mathcal{S} \text{ is false} \end{cases} \tag{3}$$

For example, $\delta(x, y) = \mathbb{1}(x = y)$.

0.8 Underlined letters indicate random variables

Random variables will be indicated by underlined letters and their values by non-underlined letters. Each node of a bnet will be labelled by a random variable. Thus, $\underline{x} = x$ means that node $\underline{x}$ is in state x.

It is more conventional to use an upper case letter to indicate a random variable and a lower case letter for its state. Thus, $X = x$ means that random variable X is in state x. However, we have opted in this book to avoid that notation, because we often want to define certain lower case letters to be random variables or, conversely, define certain upper case letters to be non-random variables.

0.9 Probability distributions

$P_{\underline{x}}(x) = P(\underline{x} = x) = P(x)$ is the probability that random variable $\underline{x}$ equals $x \in S_{\underline{x}}$. $S_{\underline{x}}$ is the set of states (i.e., values) that $\underline{x}$ can assume and $n_{\underline{x}} = |S_{\underline{x}}|$ is the size (aka cardinality) of that set. Hence,

$$\sum_{x \in S_{\underline{x}}} P_{\underline{x}}(x) = 1 \tag{4}$$

$$P_{\underline{x},\underline{y}}(x, y) = P(\underline{x} = x, \underline{y} = y) = P(x, y) \tag{5}$$

$$P_{\underline{x}|\underline{y}}(x|y) = P(\underline{x} = x|\underline{y} = y) = P(x|y) = \frac{P(x,y)}{P(y)} \tag{6}$$

0.10 Discretization of continuous probability distributions

The TPM of a node of a bnet can be either a discrete or a continuous probability distribution. To go from continuous to discrete, one replaces integrals over states of a node by sums over new states, and Dirac delta functions by Kronecker delta functions. More precisely, consider a function $f : [a,b] \to \mathbb{R}$. Express $[a,b]$ as a union of small, disjoint (except for one point) closed sub-intervals (bins) of length Δx. Name one point in each bin to be the representative of that bin, and let $S_{\underline{x}}$ be the set of all the bin representatives. This is called discretization or binning. Then

$$\frac{1}{(b-a)} \int_{[a,b]} dx\ f(x) \to \frac{\Delta x}{(b-a)} \sum_{x \in S_{\underline{x}}} f(x) = \frac{1}{n_{\underline{x}}} \sum_{x \in S_{\underline{x}}} f(x)\ . \tag{7}$$

Both sides of last equation are 1 when $f(x) = 1$. Furthermore, if $y \in S_{\underline{x}}$, then

$$\int_{[a,b]} dx\ \delta(x-y) f(x) = f(y) \to \sum_{x \in S_{\underline{x}}} \delta(x,y) f(x) = f(y)\ . \tag{8}$$

0.11 Samples, i.i.d. variables

$$\vec{x} = (x[0], x[1], x[2] \ldots, x[nsam(\vec{x}) - 1]) = x[:] \tag{9}$$

$nsam(\vec{x})$ is the number of samples of $\vec{x}$. $\underline{x}[\sigma] \in S_{\underline{x}}$ are i.i.d. (independent identically distributed) samples with

$$x[\sigma] \sim P_{\underline{x}} \text{ (i.e. } P_{\underline{x[\sigma]}} = P_{\underline{x}}) \tag{10}$$

$$P(\underline{x} = x) = \frac{1}{nsam(\vec{x})} \sum_{\sigma} \mathbb{1}(x[\sigma] = x) \tag{11}$$

Hence, for any $f : S_{\underline{x}} \to \mathbb{R}$,

$$\sum_{x} P(\underline{x} = x) f(x) = \frac{1}{nsam(\vec{x})} \sum_{\sigma} f(x[\sigma]) \tag{12}$$

If we use two sampled variables, say $\vec{x}$ and $\vec{y}$, in a given bnet, their number of samples $nsam(\vec{x})$ and $nsam(\vec{y})$ need not be equal.

$$P(\vec{x}) = \prod_{\sigma} P(x[\sigma]) \tag{13}$$

$$\sum_{\vec{x}} = \prod_{\sigma} \sum_{x[\sigma]} \tag{14}$$

$$\partial_{\vec{x}} = [\partial_{x[0]}, \partial_{x[1]}, \partial_{x[2]}, \ldots, \partial_{x[nsam(\vec{x})-1]}] \tag{15}$$

$$\begin{aligned} P(\vec{x}) &\approx [\prod_x P(x)^{P(x)}]^{nsam(\vec{x})} & (16) \\ &= e^{nsam(\vec{x})\sum_x P(x)\ln P(x)} & (17) \\ &= e^{-nsam(\vec{x})H(P_{\underline{x}})} & (18) \end{aligned}$$

0.12 Normal Distribution

For $x, \mu, \sigma \in \mathbb{R}$, $\sigma > 0$

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{19}$$

0.13 Uniform Distribution

For $a < b$, $x \in [a, b]$

$$\mathcal{U}(x; a, b) = \frac{1}{b-a} \tag{20}$$

0.14 Sigmoid and logit functions

The sigmoid function sig:$\mathbb{R} \to [0, 1]$ is defined by

$$\text{sig}(x) = \frac{1}{1 + e^{-x}} \tag{21}$$

sig() is monotonically increasing with sig$(-\infty) = 0$ and sig$(+\infty) = 1$.

The logit or log-odds function logit:$[0, 1] \to \mathbb{R}$ is defined by

$$\text{logit}(p) = \ln \frac{p}{1-p} \tag{22}$$

logit() is the inverse of sig():

$$\text{logit}[\text{sig}(x)] = x \tag{23}$$

0.15 Expected Value and Variance

Given a random variable $\underline{x}$ with states $S_{\underline{x}}$ and a function $f : S_{\underline{x}} \to \mathbb{R}$, define

$$E_{\underline{x}}[f(\underline{x})] = E_{x\sim P(x)}[f(x)] = \sum_x P(x)f(x) \tag{24}$$

$$\begin{aligned} Var_{\underline{x}}[f(\underline{x})] &= E_{\underline{x}}\left[(f(\underline{x}) - E_{\underline{x}}[f(\underline{x})])^2\right] && (25) \\ &= E_{\underline{x}}[f(\underline{x})^2] - (E_{\underline{x}}[f(\underline{x})])^2 && (26) \end{aligned}$$

$$E[\underline{x}] = E_{\underline{x}}[\underline{x}] \tag{27}$$

$$Var[\underline{x}] = Var_{\underline{x}}[\underline{x}] \tag{28}$$

0.16 Conditional Expected Value

Given a random variable $\underline{x}$ with states $S_{\underline{x}}$, a random variable $\underline{y}$ with states $S_{\underline{y}}$, and a function $f : S_{\underline{x}} \times S_{\underline{y}} \to \mathbb{R}$, define

$$E_{\underline{x}|\underline{y}}[f(\underline{x}, \underline{y})] = \sum_x P(x|\underline{y}) f(x, \underline{y}) \ , \tag{29}$$

$$E_{\underline{x}|\underline{y}=y}[f(\underline{x}, y)] = E_{\underline{x}|y}[f(\underline{x}, y)] = \sum_x P(x|y) f(x, y) \ . \tag{30}$$

Note that

$$\begin{aligned} E_{\underline{y}}[E_{\underline{x}|\underline{y}}[f(\underline{x}, \underline{y})]] &= \sum_{x,y} P(x|y)P(y)f(x,y) && (31) \\ &= \sum_{x,y} P(x,y) f(x,y) && (32) \\ &= E_{\underline{x},\underline{y}}[f(\underline{x}, \underline{y})] \ . && (33) \end{aligned}$$

0.17 Law of Total Variance

Claim 1 *Suppose $P : S_{\underline{x}} \times S_{\underline{y}} \to [0, 1]$ is a probability distribution. Suppose $f : S_{\underline{x}} \times S_{\underline{y}} \to \mathbb{R}$ and $f = f(x, y)$. Then*

$$Var_{\underline{x},\underline{y}}(f) = E_{\underline{y}}[Var_{\underline{x}|\underline{y}}(f)] + Var_{\underline{y}}(E_{\underline{x}|\underline{y}}[f]) \ . \tag{34}$$

In particular,

$$Var_{\underline{x}}(x) = E_{\underline{y}}[Var_{\underline{x}|\underline{y}}(x)] + Var_{\underline{y}}(E_{\underline{x}|\underline{y}}[x]) \ . \tag{35}$$

proof:

Let

$$A = \sum_y P(y) \left(\sum_x P(x|y) f \right)^2 \ . \tag{36}$$

Then

$$
\begin{aligned}
Var_{\underline{x},\underline{y}}(f) &= \sum_{x,y} P(x,y)f^2 - \left(\sum_{x,y} P(x,y)f\right)^2 && (37) \\
&= \left\{ \begin{array}{l} \sum_{x,y} P(x,y)f^2 - A \\ + \left(A - \left(\sum_{x,y} P(x,y)f\right)^2\right) \end{array} \right. && (38)
\end{aligned}
$$

$$
\begin{aligned}
E_{\underline{y}}[Var_{\underline{x}|\underline{y}}(f)] &= \sum_{y} P(y) \left(\sum_{x} P(x|y)f^2 - \left(\sum_{x} P(x|y)f\right)^2\right) && (39) \\
&= \sum_{x,y} P(x,y)f^2 - A && (40)
\end{aligned}
$$

$$
\begin{aligned}
Var_{\underline{y}}(E_{\underline{x}|\underline{y}}[f]) &= \sum_{y} P(y) \left(\sum_{x} P(x|y)f\right)^2 - \left(\sum_{y} P(y) \sum_{x} P(x|y)f\right)^2 && (41) \\
&= A - \left(\sum_{x,y} P(x,y)f\right)^2 && (42)
\end{aligned}
$$

QED

0.18 Notation for covariances

Consider two random variables $\underline{x}, \underline{y}$.

- Mean value of $\underline{x}$

$$\langle \underline{x} \rangle = E_{\underline{x}}[\underline{x}] \tag{43}$$

- Signed distance of $\underline{x}$ to its mean value

$$\Delta \underline{x} = \underline{x} - \langle \underline{x} \rangle \tag{44}$$

- Covariance of $(\underline{x}, \underline{y})$

$$Cov(\underline{x}, \underline{y}) = \langle \underline{x}, \underline{y} \rangle = \langle \Delta \underline{x} \Delta \underline{y} \rangle = \langle \underline{x}\underline{y} \rangle - \langle \underline{x} \rangle \langle \underline{y} \rangle \tag{45}$$

$\langle \underline{x}, \underline{y} \rangle$ is bilinear.

- Variance of $\underline{x}$
$$Var(\underline{x}) = \langle \underline{x}, \underline{x} \rangle \tag{46}$$

- Standard deviation or $\underline{x}$
$$\sigma_{\underline{x}} = \sqrt{\langle \underline{x}, \underline{x} \rangle} \tag{47}$$

- Correlation of $(\underline{x}, \underline{y})$
$$\rho_{\underline{x},\underline{y}} = \frac{\langle \underline{x}, \underline{y} \rangle}{\sqrt{\langle \underline{x}, \underline{x} \rangle \langle \underline{y}, \underline{y} \rangle}} \tag{48}$$

0.19 Linear regression, Ordinary Least Squares (OLS)

Wikipedia articles

1. Linear Regression (LR)
    - linear regression, Ref.[56]
    - simple linear regression, Ref.[70]
    - errors in variable, Ref.[43]
2. Least squares (LS)
    - least squares, Ref.[55]
    - ordinary least squares (OLS), Ref.[67]

In LR, the **dependent variables** y equal a linear combination of some **independent variables** x plus some external noise variables ϵ called the **residuals**.

Below, we consider two types of LR:

1. LR in which the independent variables are non-random.
2. LR in which the independent variables are random and i.i.d.

Once one assumes that certain variables are random, a "model" (i.e., a bnet) for the random variables must be specified.

For LR of type 2, there is randomness in y coming from the randomness in x and in the residuals. For LR of type 1, there is randomness in y too, but it comes from the residuals only.

OLS provides a cost function which when minimized, yields LR. The term OLS is often used to refer to LR of type 1.

0.19.1 LR, assuming x_σ are non-random

Let

$\sigma \in \{0, 1, 2, \ldots, nsam - 1\}$: sample index
$y_\sigma \in \mathbb{R}$: dependent variables
$x_{\sigma j} \in \mathbb{R}$: independent variables
$\epsilon_\sigma \in \mathbb{R}$: residuals
β_0, β_j: real coefficients

$$y_\sigma = \beta_0 + \sum_{j=1}^{n} x_{\sigma j}\beta_j + \epsilon_\sigma \tag{49}$$

If we define

$$x_{\sigma 0} = 1 \tag{50}$$

for all σ, then

$$y_\sigma = \sum_{j=0}^{n} x_{\sigma j}\beta_j + \epsilon_\sigma \ . \tag{51}$$

If y and ϵ are $nsam \times 1$ column vectors and β is an $(n+1) \times 1$ column vector, then can write previous equation in matrix form as:

$$y = X\beta + \epsilon \ . \tag{52}$$

Define the **projection matrices**

$$\wedge = X(X^T X)^{-1} X^T \ , \ \ \vee = 1 - \wedge \tag{53}$$

A square matrix M is symmetric if $M^T = M$ and is idempotent if $M^2 = M$. $\wedge$ is symmetric and idempotent and so is $\vee$. Note that $\wedge$ and $\vee$ also satisfy:

$$\vee\wedge = \wedge\vee = 0 \tag{54}$$

and

$$\wedge X = X \ , \ \ \vee X = 0 \ . \tag{55}$$

One has

$$\beta = (X^T X)^{-1} X^T (y - \epsilon) \ . \tag{56}$$

Define

$$\hat{\beta} = (X^T X)^{-1} X^T y = By \ , \tag{57a}$$

$$\hat{y} = X\hat{\beta} = \wedge y \ , \tag{57b}$$

and

$$\hat{\epsilon} = y - X\hat{\beta} = y - \hat{y} = (1 - \wedge)y = \vee y \ . \tag{57c}$$

$\wedge$ is sometimes called the **hat matrix**, because it gives y a hat.

Given any function $f = f(y, X, \epsilon)$ and a scalar factor $\xi \in \mathbb{R}$, suppose $f(\xi y, \xi X, \xi\epsilon) = \xi^{\mathcal{O}} f(y, X, \epsilon)$. Then we will say that $f(\cdot)$ is of **order $\mathcal{O}$ under scaling**. Note that $\{X, y, \hat{y}, \epsilon, \hat{\epsilon}\}$ are all of order 1 under scaling, $\{\beta, \hat{\beta}, \wedge, \vee\}$ are all of order 0 under scaling, and B is of order -1 under scaling. Thus, the estimator variables (i.e, those with a hat) scale the same way as the variables without a hat that they are estimating. Furthermore, β, its estimator $\hat{\beta}$, and the projection matrices $\wedge, \vee$ are invariant ($\mathcal{O} = 0$) under scaling.

Fig.3 illustrates that y can be expressed as a sum of 2 estimators:

$$y = \underbrace{\hat{y}}_{\wedge y} + \underbrace{\hat{\epsilon}}_{\vee y} \ . \tag{58}$$

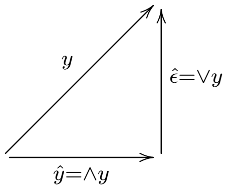


Figure 3: Decomposition of y into sum of two estimators, $\hat{y}$ and $\hat{\epsilon}$.

---

model dependent results:

Assume the components of ϵ are random over σ and

$$E_\sigma[\underline{\epsilon}] = \langle \underline{\epsilon} \rangle = 0 \tag{59}$$

Assume X and β are not random. This makes $\underline{y} = X\beta + \underline{\epsilon}$ and $\underline{\hat{\beta}} = (X^T X)^{-1} X^T \underline{y}$ random. One finds that

$$\langle \underline{y} \rangle = X\beta \tag{60}$$

$$\langle \underline{\hat{y}} \rangle = \wedge \langle \underline{y} \rangle = \langle \underline{y} \rangle \tag{61a}$$

$$\langle \underline{\hat{\epsilon}} \rangle = \vee \langle \underline{y} \rangle = 0 \tag{61b}$$

$$\langle \underline{\hat{\beta}} \rangle = \beta \tag{61c}$$

So far, we have assumed a zero mean value for ϵ. Next, assume **"homoscedasticity" (HS)**, which means that

$$\langle \underline{\epsilon}, \underline{\epsilon}^T \rangle = \xi^2 I_{nsam} \tag{61d}$$

where $\xi \geq 0$, $nsam = \sum_\sigma$ and I_{nsam} is the $nsam \times nsam$ identity matrix. It follows that

$$\langle \underline{y}, \underline{y}^T \rangle = \langle \underline{\epsilon}, \underline{\epsilon}^T \rangle = \xi^2 I_{nsam} \ , \tag{62}$$

$$\langle \hat{\underline{\epsilon}}, \hat{\underline{\epsilon}}^T \rangle = \vee \langle \underline{y}, \underline{y}^T \rangle \vee^T = \xi^2 \vee \ , \tag{63}$$

$$\langle \hat{\underline{y}}, \hat{\underline{y}}^T \rangle = \wedge \langle \underline{y}, \underline{y}^T \rangle \wedge^T = \xi^2 \wedge \tag{64}$$

and

$$\langle \hat{\underline{\beta}}, \hat{\underline{\beta}}^T \rangle = B \langle \underline{y}, \underline{y}^T \rangle B^T = \xi^2 (X^T X)^{-1} \ . \tag{65}$$

The goodness of fit for this model is often measured using the **coefficient of determination** R^2. R^2 is defined by

$$R^2 = \frac{\parallel \hat{\underline{y}} - \langle \hat{\underline{y}} \rangle \parallel^2}{\parallel \underline{y} - \langle \underline{y} \rangle \parallel^2} = \frac{\text{tr} \langle \hat{\underline{y}}, \hat{\underline{y}}^T \rangle}{\text{tr} \langle \underline{y}, \underline{y}^T \rangle} \tag{66}$$

If HS holds, then R^2 reduces to

$$R^2 = \frac{\text{tr} \wedge}{nsam} \ . \tag{67}$$

0.19.2 LR, assuming x_σ are random and i.i.d.

Let

$\underline{y}$ = true value
$\hat{\underline{y}}$ = estimator
$\underline{\epsilon}$ = residual

$$\hat{\underline{y}} = \beta_0 + \sum_{j=1}^{n} \beta_j \underline{x}_j \tag{68}$$

$$\underline{y} = \hat{\underline{y}} + \underline{\epsilon} \tag{69}$$

Assume

$$\langle \underline{\epsilon} \rangle = 0 \tag{70}$$

and

$$\langle \underline{x}_j, \underline{\epsilon} \rangle = 0 \tag{71}$$

for all j.

For $k = 1, \ldots, n$,

$$\langle \underline{x}_k, \underline{y} \rangle = \sum_{j=1}^{n} \beta_j \langle \underline{x}_k, \underline{x}_j \rangle \; . \tag{72}$$

Let $\underline{x}^n$ and β^n be column vectors. Then

$$\langle \underline{x}^n, \underline{y} \rangle = \langle \underline{x}^n, (\underline{x}^n)^T \rangle \beta^n \; , \tag{73}$$

$$\beta^n = \langle \underline{x}^n, (\underline{x}^n)^T \rangle^{-1} \langle \underline{x}^n, \underline{y} \rangle \; . \tag{74}$$

$$\beta_0 = \langle \underline{y} \rangle - (\beta^n)^T \langle \underline{x}^n \rangle \tag{75}$$

0.20 Short Summary of Boolean Algebra.

See Ref.[38] for more info about this topic.

Suppose $x, y, z \in \{0, 1\}$. Define

$$x \text{ or } y = x \vee y = x + y - xy \; , \tag{76}$$

$$x \text{ and } y = x \wedge y = xy \; , \tag{77}$$

and

$$\text{not } x = \overline{x} = 1 - x \; , \tag{78}$$

where we are using normal addition and multiplication on the right hand sides.[1]

Actually, since $x \wedge y = xy$, we can omit writing the symbol $\wedge$. The symbol $\wedge$ is useful to exhibit the symmetry of the identities, and to remark about the analogous identities for sets, where $\wedge$ becomes intersection $\cap$ and $\vee$ becomes union $\cup$. However, for practical calculations, $\wedge$ is an unnecessary nuisance.

Since $x \in \{0, 1\}$,

$$P(\overline{x}) = 1 - P(x) \; . \tag{79}$$

Clearly, from analyzing the simple event space $(x, y) \in \{0, 1\}^2$,

$$P(x \vee y) = P(x) + P(y) - P(x \wedge y) \; . \tag{80}$$

[1] Note the difference between $\vee$ and modulus 2 addition $\oplus$. For $\oplus$ (aka XOR): $x \oplus y = x + y - 2xy$.

Associativity	$x \vee (y \vee z) = (x \vee y) \vee z$ $x \wedge (y \wedge z) = (x \wedge y) \wedge z$
Commutativity	$x \vee y = y \vee x$ $x \wedge y = y \wedge x$
Distributivity	$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ $x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$
Identity	$x \vee 0 = x$ $x \wedge 1 = x$
Annihilator	$x \wedge 0 = 0$ $x \vee 1 = 1$
Idempotence	$x \vee x = x$ $x \wedge x = x$
Absorption	$x \wedge (x \vee y) = x$ $x \vee (x \wedge y) = x$
Complementation	$x \wedge \overline{x} = 0$ $x \vee \overline{x} = 1$
Double negation	$\overline{(\overline{x})} = x$
De Morgan Laws	$\overline{x} \wedge \overline{y} = \overline{(x \vee y)}$ $\overline{x} \vee \overline{y} = \overline{(x \wedge y)}$

Table 1: Boolean Algebra Identities

0.21 Entropy, Kullback-Liebler divergence

For probabilty distributions $p(x), q(x)$ of $x \in S_{\underline{x}}$

- Entropy:

$$H(p) = -\sum_x p(x) \ln p(x) \geq 0 \tag{81}$$

- Kullback-Liebler divergence:

$$D_{KL}(p \parallel q) = \sum_x p(x) \ln \frac{p(x)}{q(x)} \geq 0 \tag{82}$$

- Cross entropy:

$$\begin{aligned} CE(p \to q) &= -\sum_x p(x) \ln q(x) && (83) \\ &= H(p) + D_{KL}(p \parallel q) && (84) \end{aligned}$$

0.22 Definition of various entropies used in Shannon Information Theory

- **(plain) Entropy of $\underline{x}$**

$$H(\underline{x}) = -\sum_x P(x) \ln P(x) \tag{85}$$

This quantity measures the spread of $P_{\underline{x}}$.

- **Conditional Entropy of $\underline{y}$ given $\underline{x}$**

$$\begin{aligned} H(\underline{y}|\underline{x}) &= -\sum_{x,y} P(x,y) \ln P(y|x) & (86) \\ &= H(\underline{y},\underline{x}) - H(\underline{x}) & (87) \end{aligned}$$

This quantity measures the conditional spread of $\underline{y}$ given $\underline{x}$.

- **Mutual Information (MI) of $\underline{x}$ and $\underline{y}$**

$$\begin{aligned} H(\underline{y}:\underline{x}) &= \sum_{x,y} P(x,y) \ln \frac{P(x,y)}{P(x)P(y)} & (88) \\ &= H(\underline{x}) + H(\underline{y}) - H(\underline{y},\underline{x}) & (89) \end{aligned}$$

This quantity measures the correlation between $\underline{x}$ and $\underline{y}$.

- **Conditional Mutual Information (CMI)[2] of $\underline{x}$ and $\underline{y}$ given $\underline{\lambda}$**

$$\begin{aligned} H(\underline{y}:\underline{x}|\underline{\lambda}) &= \sum_{x,y,\lambda} P(x,y,\lambda) \ln \frac{P(x,y|\lambda)}{P(x|\lambda)P(y|\lambda)} & (90) \\ &= H(\underline{x}|\underline{\lambda}) + H(\underline{y}|\underline{\lambda}) - H(\underline{y},\underline{x}|\underline{\lambda}) & (91) \end{aligned}$$

This quantity measures the conditional correlation of $\underline{x}$ and $\underline{y}$ given $\underline{\lambda}$.

- **Kullback-Liebler Divergence from $P_{\underline{x}}$ to $P_{\underline{y}}$.**

Assume random variables $\underline{x}$ and $\underline{y}$ have the same set of states $S_{\underline{x}} = S_{\underline{y}}$. Then

$$D_{KL}(P_{\underline{x}} \parallel P_{\underline{y}}) = \sum_x P_{\underline{x}}(x) \ln \frac{P_{\underline{x}}(x)}{P_{\underline{y}}(x)} \tag{92}$$

This measures a non-symmetric distance between the probability distributions $P_{\underline{x}}$ and $P_{\underline{y}}$. $D_{KL}(P_{\underline{x}} \parallel P_{\underline{y}})$ is non-negative and equals zero iff $P_{\underline{x}} = P_{\underline{y}}$.

[2] CMI can be read as "see me".

Chapter 10

Difference-in-Differences

This chapter is based on Ref.[3].

This chapter assumes that the reader has read Chapter 38 on Potential Outcomes. DID theory applies the basic single-time PO theory described in Chapter 38 to 2 well separated times in which different conditions prevail.

The Difference-in-Differences (DID) method was first used by John Snow in an 1854 report that argued that cholera in London was being transmitted by sewage polluted water rather than, as others at the time believed, by air (in fetid vapors called miasmas). In general, one can apply DID to discover causal effects in historical data. By **historical data** (aka a **natural experiment**. See Ref.[65]) we mean data that is collected long after the treatment (rather than during it) and is thus not subject to active intervention by the experimenter.

10.1 John Snow, DID and a cholera transmission pathway

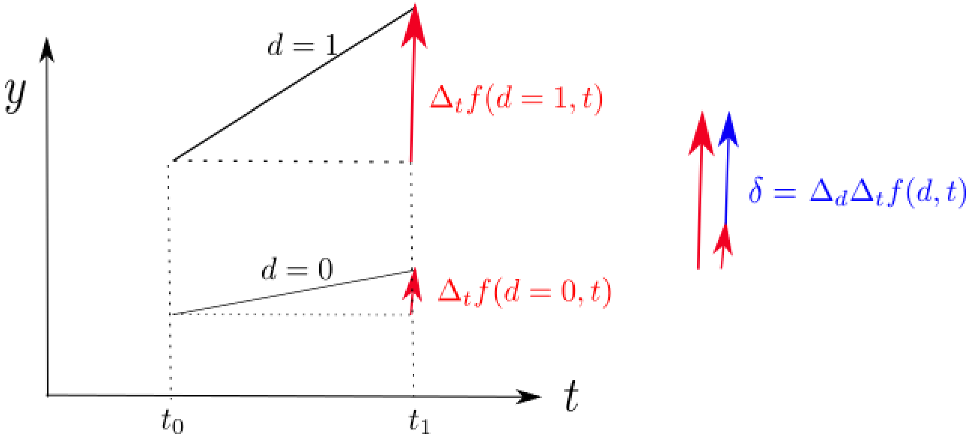


Figure 10.1: Pictorial representation of Difference-in-differences (DID) as a difference of two differences (i.e., a difference of two slopes).

Let

$d \in \{0, 1\}$
$t \in \{t_0, t_1\}$, $t_0 < t_1$
$y = f(d, t) \in \mathbb{R}$

$$\Delta_t f(d, t) = f(d, t_1) - f(d, t_0) \tag{10.1}$$

$$\Delta_d f(d, t) = f(1, t) - f(0, t) \tag{10.2}$$

$$DID = \delta = \Delta_d \Delta_t f(d, t) \tag{10.3}$$

DID is illustrated in Fig.10.1.

A **time series** is any function of time (the domain of the function is usually a discrete set of times).

In DID, one calculates the slope, over the same time interval, of two time series. One of the time series ($d = 0$) is for the control (i.e., untreated) population and the other ($d = 1$) is for the treated population. Then the difference δ of those 2 slopes is taken. The idea is that if there is no causal difference between the 2 time series, then both time series will have the same slope, and δ will be zero.

	$t = t_0$ (1849)	$t = t_1$ (1854)
$\tilde{d} = 1$ (town 1)	85 deaths, polluted DW	19 death, unpolluted DW
$\tilde{d} = 0$ (town 0)	135 deaths, polluted DW	147 deaths, polluted DW

Table 10.1: A condensation of the data collected by John Snow in 1854, to test the hypothesis that cholera in London was being spread by polluted drinking water (DW).

A condensation of the data collected by John Snow in 1854 is given in Table 10.1. From that data, we find that

$$\delta = \Delta_d \Delta_t f(d, t) = (19 - 85) - (147 - 135) = -66 - 12 = -78 \tag{10.4}$$

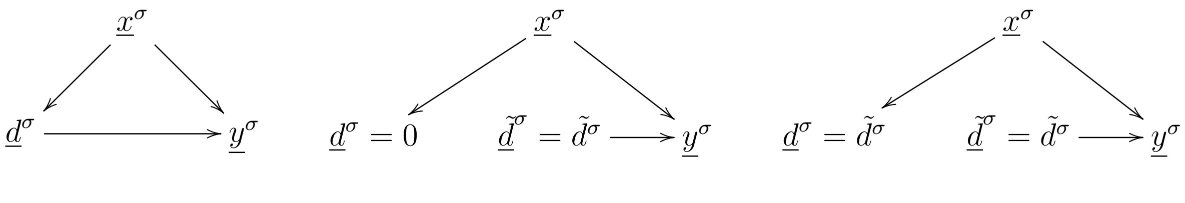
10.2 PO analysis

In this section, we show how to analyze DID using the formalism of PO theory.

We will speak of a treatment outcome $\underline{y}^\sigma(\underline{\tilde{d}}, x^\sigma; t, g^\sigma)$ for individual σ that depends, not just on the treatment dose $\tilde{d}^\sigma \in \{0, 1\}$ and the confounder state x^σ, but also on a group parameter (i.e., which population or town) $g^\sigma \in \{0, 1\}$ and on a time parameter $t \in \{t_0, t_1\}$ (note t is independent of σ). Actually, we will assume $g^\sigma = \tilde{d}^\sigma$, so we will just speak of $\underline{y}^\sigma(\underline{\tilde{d}}, x^\sigma; t)$ with no explicit g^σ dependence. As usual for PO theory, we will consider expected values of y^σ:

$$E_{\sigma|d,\tilde{d},x}[y^\sigma(\tilde{d}^\sigma, x^\sigma; t)] = E_{y|d,\tilde{d},x}[\underline{y}(\tilde{d}, x; t)] = \mathcal{Y}_{d|\tilde{d},x}(t) \tag{10.5}$$

To calculate these expected values, we need a "model" with probability distributions. In this case, the needed model and probability distributions are provided by the bnets depicted in Fig.10.2. The TPMs, printed in blue, for the two bnets $G(t_0)$ and $G(t_1)$ shown in Fig.10.2, are as follows. Note that the TPMs for the bnets $G(t_0)$ and $G(t_1)$ are defined in terms of the TPMs for the bnet G.


$\underline{x}^\sigma$
$\underline{d}^\sigma$
$\underline{y}^\sigma$
$\underline{x}^\sigma$
$\underline{d}^\sigma = 0$
$\underline{\tilde{d}}^\sigma = \tilde{d}^\sigma$
$\underline{y}^\sigma$
$\underline{x}^\sigma$
$\underline{d}^\sigma = \tilde{d}^\sigma$
$\underline{\tilde{d}}^\sigma = \tilde{d}^\sigma$
$\underline{y}^\sigma$

G $\quad$ $G(t_0) = G_{im}|_{d^\sigma=0}$ $\quad$ $G(t_1) = G_{im}|_{d^\sigma=\tilde{d}^\sigma}$

Figure 10.2: Bnet $G_{im} = \kappa_{\underline{d}^\sigma \to \underline{y}^\sigma}(\tilde{d}^\sigma)G$ is obtained by applying the imagine operator to arrow $\underline{d}^\sigma \to \underline{y}^\sigma$ of bnet G. Bnet $G_{im}(t_0)$ is obtained by setting $\underline{d}^\sigma = 0$ in bnet G_{im}. Bnet $G_{im}(t_1)$ is obtained by setting $\underline{\tilde{d}}^\sigma = \underline{d}^\sigma$ in bnet G_{im}.

$$P(x^\sigma) = P_{\underline{x}}(x^\sigma) \tag{10.6}$$

$$P(d^\sigma|x^\sigma) = \left\{ \begin{array}{ll} P_{\underline{d}|\underline{x}}(d^\sigma|x^\sigma)\delta(d^\sigma, 0) & \text{for graph } G(t_0) \\ P_{\underline{d}|\underline{x}}(d^\sigma|x^\sigma)\delta(d^\sigma, \tilde{d}^\sigma) & \text{for graph } G(t_1) \end{array} \right. \tag{10.7}$$

$$P(y^\sigma|\tilde{d}^\sigma, x^\sigma) = P_{\underline{y}|\underline{\tilde{d}},\underline{x}}(y^\sigma|\tilde{d}^\sigma, x^\sigma) \tag{10.8}$$

$$P((\tilde{d}^\sigma)') = \delta((\tilde{d}^\sigma)', \tilde{d}^\sigma) \tag{10.9}$$

Henceforth, for simplicity, we will omit the confounder state x from the indices of $\mathcal{Y}$; i.e., we will write $\mathcal{Y}_{d|\tilde{d}}(t)$ instead of $\mathcal{Y}_{d|\tilde{d},x}(t)$. The fact that we will not explicitly mention x does not mean that it doesn't exist or that it doesn't affect our analysis. John Snow does not seem to have considered any confounders in his cholera study, or else he tried to collect data restricted to a single stratum x. If there are confounders, they cannot be neglected. As discussed in Chapter 38 under the subject of strata-matching in PO, one must condition $\mathcal{Y}$ on a single x stratum and, later on, one must average over all the possible x strata.

Let $\mathcal{MY}_{d|\tilde{d}}(t)$ denote the **measured** $\mathcal{Y}_{d|\tilde{d}}(t)$. We define this quantity as

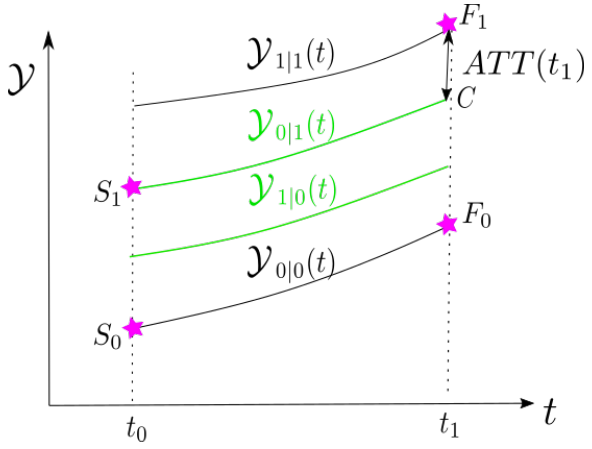


Figure 10.3: Four different time-dependent expected values $\mathcal{Y}_{d|\tilde{d}}(t)$ of y^σ. Bnets $G(t_0)$ and $G(t_1)$ are defined in Fig.10.1. The $\mathcal{Y}$ coordinates of the 2 magenta stars at time $t = t_0$ (resp., $t = t_1$) can be calculated using bnet $G(t_0)$ (resp., $G(t_1)$).

$$
\begin{aligned}
\mathcal{M}\mathcal{Y}_{d|\tilde{d}}(t) &= \left\{ \begin{array}{ll} \mathcal{Y}_{0|\tilde{d}}(t)\mathbb{1}(d=0) & \text{if } t = t_0 \\ \mathcal{Y}_{\tilde{d}|\tilde{d}}(t)\mathbb{1}(d=\tilde{d}) & \text{if } t = t_1 \end{array} \right. && (10.10) \\
&= \mathcal{Y}_{0|\tilde{d}}(t)\mathbb{1}(d=0, t=t_0) + \mathcal{Y}_{\tilde{d}|\tilde{d}}(t)\mathbb{1}(d=\tilde{d}, t=t_1) && (10.11)
\end{aligned}
$$

Now we claim that the DID δ calculated in the previous section for John Snow's data, can be expressed in PO formalism as follows:

$$
\delta = \Delta_{\tilde{d}}\Delta_t \sum_d \mathcal{M}\mathcal{Y}_{d|\tilde{d}}(t) \; . \tag{10.12}
$$

Fig.10.3 depicts the four functions $\mathcal{Y}_{d|\tilde{d}}(t)$ for t in the interval $[t_0, t_1]$ and for $d, \tilde{d} \in \{0, 1\}$. The $\mathcal{Y}$ coordinates of the four magenta stars in Fig.10.3 can be calculated using bnets $G(t_0)$ and $G(t_1)$.

Define the **parallel trends** (PT) by

$$
PT = \Delta_{\tilde{d}}\Delta_t \mathcal{Y}_{0|\tilde{d}}(t) \; . \tag{10.13}
$$

We will say the **parallel trends assumption (PTA)** holds if $PT = 0$.

Next we prove that the DID δ equals the sum of an ATT[1] and PT.

[1] ATT stands for the average treatment effect of the treated. ATT is defined in Chapter 38

$$
\begin{align}
\delta &= \Delta_{\tilde{d}}\Delta_t \sum_d \mathcal{M}\mathcal{Y}_{d|\tilde{d}}(t) \tag{10.14}\\
&= \sum_d \left[\Delta_t \mathcal{M}\mathcal{Y}_{d|1}(t) - \Delta_t \mathcal{M}\mathcal{Y}_{d|0}(t)\right] \tag{10.15}\\
&= \sum_d \left[\mathcal{M}\mathcal{Y}_{d|1}(t_1) - \mathcal{M}\mathcal{Y}_{d|1}(t_0)\right] - \sum_d \left[\Delta_t \mathcal{M}\mathcal{Y}_{d|0}(t_1) - \Delta_t \mathcal{M}\mathcal{Y}_{d|0}(t_0)\right] \tag{10.16}\\
&= \mathcal{Y}_{1|1}(t_1) - \mathcal{Y}_{0|1}(t_0) - \{\mathcal{Y}_{0|0}(t_1) - \mathcal{Y}_{0|0}(t_0)\} \tag{10.17}\\
&= \mathcal{Y}_{1|1}(t_1) - \mathcal{Y}_{0|1}(t_0) - \{\mathcal{Y}_{0|0}(t_1) - \mathcal{Y}_{0|0}(t_0)\} + \underbrace{\{\mathcal{Y}_{0|1}(t_1) - \mathcal{Y}_{0|1}(t_1)\}}_{\text{zero}} \tag{10.18}\\
&= \underbrace{\mathcal{Y}_{1|1}(t_1) - \mathcal{Y}_{0|1}(t_1)}_{ATT(t_1)} - \mathcal{Y}_{0|1}(t_0) - \{\mathcal{Y}_{0|0}(t_1) - \mathcal{Y}_{0|0}(t_0)\} + \mathcal{Y}_{0|1}(t_1) \tag{10.19}\\
&= ATT(t_1) - \Delta_t \mathcal{Y}_{0|0}(t) + \Delta_t \mathcal{Y}_{0|1}(t) \tag{10.20}\\
&= ATT(t_1) + \underbrace{\Delta_{\tilde{d}}\Delta_t \mathcal{Y}_{0|\tilde{d}}(t)}_{\text{zero if PTA holds}} \tag{10.21}
\end{align}
$$

10.3 Linear Regression

In this section, we show how to apply linear regression (LR) to the PO analysis of DID.

As before, let y^σ be the treatment outcome for individual σ, who receives a treatment dose $\tilde{d}^\sigma$ at times $t \in \{t_0, t_1\}$. $y^\sigma(\tilde{d}^\sigma; t)$ can be fitted as follows. Here ϵ^σ is the residual for individual σ and $b_0, m_0, b_1, m_1 \in \mathbb{R}$ are the fit parameters.

$$
y^\sigma = [b_0 + m_0 t](1 - \tilde{d}^\sigma) + [b_1 + m_1 t]\tilde{d}^\sigma + \epsilon^\sigma \; . \tag{10.22}
$$

Note that Eq.(10.22) yields a straight line in the $y^\sigma - t$ plane for $d^\sigma = 0$, and another straight line for $d^\sigma = 1$. We are using the standard symbols b to denote the y-intercept, and m to denote the slope of a straight line.

Taking the expected value of Eq.(10.22), we get

$$
\mathcal{Y}_{d|\tilde{d}}(t) = [b_0 + m_0 t](1 - \tilde{d}) + [b_1 + m_1 t]\tilde{d} \; . \tag{10.23}
$$

with $d = 0$ for $t = t_0$ and $d = \tilde{d}$ for $t = t_1$.

Let $T = t_1 - t_0$. Since $\Delta_t t = T$ and $\Delta_d d = 1$, one gets

$$
\delta = \Delta_d \Delta_t \mathcal{M}\mathcal{Y}_{d|\tilde{d}}(t) = \Delta_d \Delta_t y^\sigma = (m_1 - m_0)T \; . \tag{10.24}
$$

Figs.10.3 and 10.4 define points S_0, S_1, F_0, F_1, I, C. The $\mathcal{Y}$ coordinates of points S_0, S_1, F_0, F_1 are given by Table 10.2. The $\mathcal{Y}$ coordinates of points C, I are given by Eqs.10.25.

$$
\mathcal{Y}(C) = \mathcal{Y}_{0|1}(t_1) \tag{10.25a}
$$

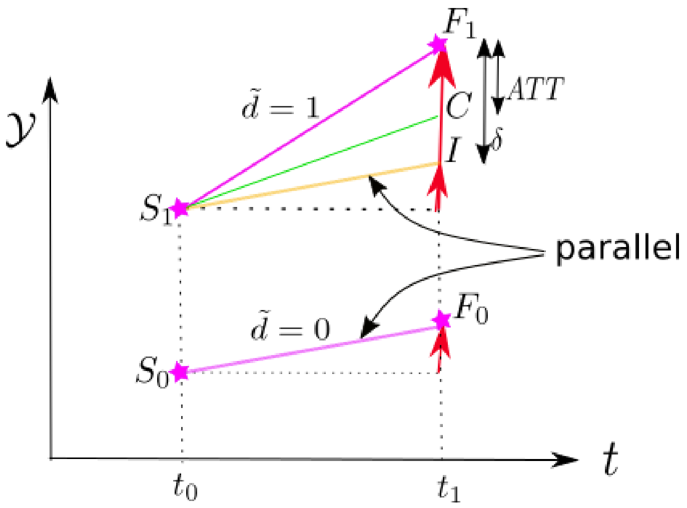


Figure 10.4: We use Linear Regression to fit a straight line between points S_0 and F_0, and between points S_1 and F_1. (S=starting, F=finishing). The $\mathcal{Y}$ coordinates of S_0, S_1, F_0, F_1 are given by Table 10.2. The $\mathcal{Y}$ coordinates of points I (image of point F_0) and C (counterfactual point) are given by Eqs.(10.25).

	$t = t_0$	$t = t_1$
$\tilde{d} = 1$	$\mathcal{Y}(S_1) = \mathcal{Y}_{0\|1}(t_0)$	$\mathcal{Y}(F_1) = \mathcal{Y}_{1\|1}(t_1)$
$\tilde{d} = 0$	$\mathcal{Y}(S_0) = \mathcal{Y}_{0\|0}(t_0)$	$\mathcal{Y}(F_0) = \mathcal{Y}_{0\|0}(t_1)$

Table 10.2: $\mathcal{Y}$ coordinates of points S_0, S_1, F_0, F_1 in Figs.10.3 and 10.4.

$$\mathcal{Y}(I) = \mathcal{Y}(F_0) + [\mathcal{Y}(S_1) - \mathcal{Y}(S_0)] \tag{10.25b}$$

We can express ATT and the δ for DID in terms of the $\mathcal{Y}$ of the points S_0, S_1, F_0, F_1, I, C. Indeed,

$$\begin{aligned} \delta &= \mathcal{Y}(F_1) - \mathcal{Y}(I) && (10.26) \\ &= \mathcal{Y}(F_1) - \mathcal{Y}(F_0) - [\mathcal{Y}(S_1) - \mathcal{Y}(S_0)] && (10.27) \end{aligned}$$

$$ATT = \mathcal{Y}(F_1) - \mathcal{Y}(C) \tag{10.28}$$

Hence,

$$\delta = ATT \iff \mathcal{Y}(I) = \mathcal{Y}(C) \iff \text{PTA holds} \tag{10.29}$$

Chapter 21

Instrumental Inequality and beyond

This chapter is based on Refs. [4] and [19].

Instrumental Variables (IVs) are discussed in Chapter 22. This chapter will discuss the original Instrumental inequality (I-inequality) discovered by Pearl, and other related inequalities. The I-inequality arises in bnets that use an IV. The I-inequality bounds the effect that an IV $\underline{z}$ can have on the outcome $\underline{y}$ of a treatment $\underline{d} \rightarrow \underline{y}$. Since there is a path $\underline{z} \rightarrow \underline{d} \rightarrow \underline{y}$, the treatment dose $\underline{d}$ acts as a mediator between the IV $\underline{z}$ and the treatment outcome $\underline{y}$. The I-inequality is reminiscent of the data processing inequality $H(\underline{z} : \underline{y}) \leq H(\underline{d} : \underline{y})$ which is valid for a simple Markov chain bnet $\underline{z} \rightarrow \underline{d} \rightarrow \underline{y}$. The data processing inequality is saying that the endpoint $\underline{y}$ receives more information from $\underline{d}$ than from $\underline{z}$. This is reasonable, since $\underline{y}$ is "closer" to $\underline{d}$ than to $\underline{z}$.

21.1 I-inequality

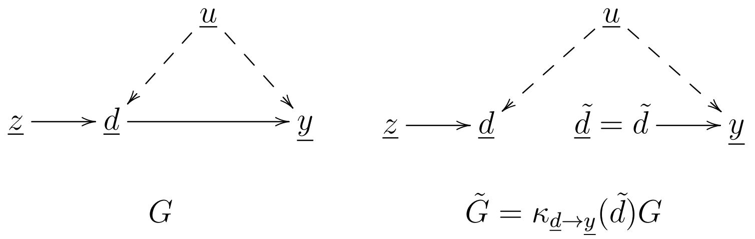


Figure 21.1: In bnet G, an IV $\underline{z}$ acts on a treatment $\underline{d} \rightarrow \underline{y}$. Bnet $\tilde{G}$ is obtained by applying an imagine operator to arrow $\underline{d} \rightarrow \underline{y}$ of bnet G.

Claim 18 *The TPMs for the bnet G in Fig.21.1 satisfy*

$$\max_d \sum_y \max_z P(d, y|z) \leq 1 \tag{21.1}$$

proof:

Below, any probability that alludes to a value $\tilde{d}$ refers to bnet $\tilde{G}$. Otherwise, if it doesn't allude to $\tilde{d}$, then it refers to G (or to $\tilde{G}$, since the TPMs of $\tilde{G}$ are defined from those of G in a consistent manner.)

G satisfies

$$P(d,y|z) = \sum_u P(u)P(y|u,d)P(d|u,z) \ , \tag{21.2}$$

and $\tilde{G}$ satisfies

$$P(d,y|z,\tilde{d}) = \sum_u P(u)P(y|u,\tilde{d})P(d|u,z) \ . \tag{21.3}$$

Note that Eqs.(21.2) and (21.3) imply that

$$P(d,y|z,d) = P(d,y|z) \tag{21.4}$$

and that

$$\boxed{P(\tilde{d},y|z,\tilde{d}) \leq \sum_d P(d,y|z,\tilde{d}) = P(y|\tilde{d})} \ . \tag{21.5}$$

Thus,

$$\begin{aligned}
\max_{\tilde{d}} \sum_y \max_z P(\tilde{d},y|z,\tilde{d}) &\leq \max_{\tilde{d}} \sum_y \max_z P(y|\tilde{d}) && (21.6) \\
&\leq \max_{\tilde{d}} \sum_y P(y|\tilde{d}) && (21.7) \\
&\leq \max_{\tilde{d}} 1 && (21.8) \\
&\leq 1 && (21.9)
\end{aligned}$$

QED

As pointed out in Ref.[4] from which I learned the above proof, the above proof is highly generalizable.

Fig.21.2 gives a graphical representation of the boxed Eq.(21.5) which is crucial to the proof.

And here is a meta-description of the steps in the proof:

1. Use imagine operator to create a non-negative matrix $M_{d,\tilde{d}}$.
2. Use fact that row or column sum of $M_{d,\tilde{d}}$ is larger than diagonal element in sum: $\sum_d M_{d,\tilde{d}} \geq M_{\tilde{d},\tilde{d}}$.

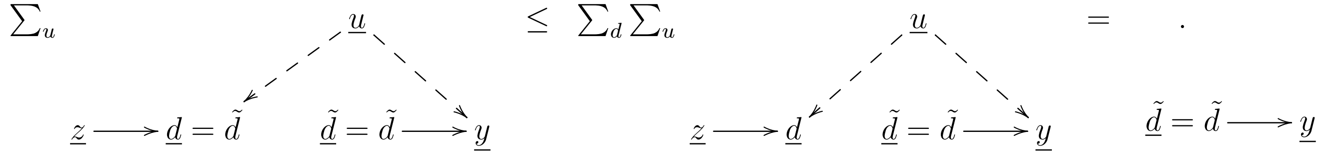

$\sum_u$
u
$\underline{z} \longrightarrow \underline{d} = \tilde{d}$
$\underline{\tilde{d}} = \tilde{d} \longrightarrow \underline{y}$
$\leq \sum_d \sum_u$
u
$\underline{z} \longrightarrow \underline{d}$
$\underline{\tilde{d}} = \tilde{d} \longrightarrow \underline{y}$
$=$
$\underline{\tilde{d}} = \tilde{d} \longrightarrow \underline{y}$
$.$

Figure 21.2: Graphical representation of the boxed equation Eq.(21.5).

$$P(d=1, y|z) = \begin{array}{c|c|c|} & z=0 & z=1 \\ \hline y=0 & a & b \\ \hline y=1 & c & d \\ \hline \end{array}$$

$$a + d \leq 1$$

$$b + c \leq 1$$

Figure 21.3: I-inequality for binary $\underline{z}, \underline{d}, \underline{y}$. The same picture except with $d = 0$ is also true.

21.1.1 I-inequality for binary z,d,y

It is enlightening to write down the I-inequality for the special case that $\underline{z}, \underline{d}, \underline{y}$ are binary.

In the binary case, the I-inequality implies 4 different inequalities. These are as follows. One gets two inequalities by setting $d = 1$ in the next 2 equations.

$$\sum_{y=0}^{1} \sum_{z=0}^{1} \mathbb{1}(y = z) P(d, y|z) \ , \tag{21.10a}$$

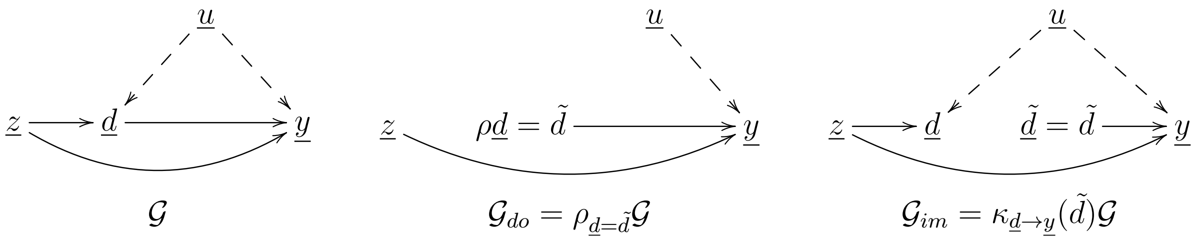
$$\sum_{y=0}^{1} \sum_{z=0}^{1} \mathbb{1}(y \neq z) P(d, y|z) \ . \tag{21.10b}$$

One gets an additional 2 inequalities by setting $d = 0$ in Eqs.(21.10). These 4 inequalities are illustrated in Fig.21.3.

What do they mean? That at fixed $\underline{d}$, the correlation between $\underline{z}$ and $\underline{y}$ is limited.

21.2 Bounds on Effect of IV on treatment outcome y

In this section, we will assume that random variables $\underline{z}, \underline{d}, \underline{y}$ are binary. Just as with the binary case of the I-inequality, we will find an inequality for each value of $\underline{d} \in \{0, 1\}$.



$\mathcal{G}$ $\mathcal{G}_{do} = \rho_{\underline{d}=\tilde{d}}\mathcal{G}$ $\mathcal{G}_{im} = \kappa_{\underline{d}\rightarrow\underline{y}}(\tilde{d})\mathcal{G}$

Figure 21.4: Bnet $\mathcal{G}$ is obtained from the bnet G in Fig.21.1 by adding to G an arrow from the IV $\underline{z}$ to the treatment outcome $\underline{y}$. Bnet $\mathcal{G}_{do}$ is obtained by applying a do operator to node $\underline{d}$ of $\mathcal{G}$. Bnet $\mathcal{G}_{im}$ is obtained by applying an imagine operator to arrow $\underline{d} \rightarrow \underline{y}$ of $\mathcal{G}$.

Below, we will use the following 3 shorthand notations:

$$P_{y|z}(d) = P(d, y|z) \;, \tag{21.11}$$

$$P_{|z}(d) = \sum_y P(d, y|z) \;, \tag{21.12}$$

and

$$\pi_{|z}(d) = 1 - P_{|z}(d) \;. \tag{21.13}$$

For the bnet $\mathcal{G}_{do}$ in Fig.21.4, define the IV effect at fixed $\rho\underline{d} = \tilde{d}$ by

$$IVE(\tilde{d}) = P(y = 1|z = 1, \rho\underline{d} = \tilde{d}) - P(y = 1|z = 0, \rho\underline{d} = \tilde{d}) \;. \tag{21.14}$$

Claim 19 *The TPMs for the bnet $\mathcal{G}_{do}$ in Fig.21.4 satisfy*

$$\pi_{|0}(d) \leq \left[IVE(d) - \{P_{1|1}(d) - P_{1|0}(d)\}\right] \leq \pi_{|1}(d) \tag{21.15}$$

proof:

$$\begin{aligned}
P(y|z, \rho\underline{d} = \tilde{d}) &= \sum_u P(u)P(y|u, z, \tilde{d}) && (21.16)\\
&= \sum_u P(u)\sum_d P(d, y|u, z, \tilde{d}) && (21.17)\\
&\geq \sum_u P(u)P(\tilde{d}, y|u, z, \tilde{d}) && (21.18)\\
&= \sum_u P(u)P(\tilde{d}, y|u, z) && (21.19)\\
&= P_{y|z}(\tilde{d}) && (21.20)
\end{aligned}$$

Next note that $P(d,y|z,\tilde{d}) \geq 0$, and $\sum_{d,y} P(d,y|z,\tilde{d}) = 1$. If we write a table for $P(d,y|z,\tilde{d})$ at fixed $z, \tilde{d}$ with row and column indices (d,y), then a partial sum of the entries of that table must be ≤ 1:

$$\sum_{d \neq \tilde{d}} P(d,y|z,\tilde{d}) + \underbrace{\sum_{y'} P(\tilde{d},y'|z,\tilde{d})}_{P_{|z}(\tilde{d})} \leq 1 \ . \tag{21.21}$$

Using the definitions of $P_{|z}$ and $\pi_{|z}$, we can rewrite the last equation as

$$\sum_{d \neq \tilde{d}} P(d,y|z,\tilde{d}) \leq \pi_{|z}(\tilde{d}) \ . \tag{21.22}$$

Next note that

$$\begin{aligned}
P(y|z, \rho\underline{d} = \tilde{d}) &= \sum_u P(u) P(y|u,z,\tilde{d}) & (21.23)\\
&= \sum_u P(u) \sum_d P(d,y|u,z,\tilde{d}) & (21.24)\\
&= P(\tilde{d},y|z,\tilde{d}) + \sum_{d \neq \tilde{d}} P(d,y|z,\tilde{d}) & (21.25)\\
&= P_{y|z}(\tilde{d}) + \sum_{d \neq \tilde{d}} P(d,y|z,\tilde{d}) & (21.26)\\
&\leq P_{y|z}(\tilde{d}) + \pi_{|z}(\tilde{d}) \ . & (21.27)
\end{aligned}$$

Hence,

$$P_{y|z}(\tilde{d}) \leq P(y|z, \rho\underline{d} = \tilde{d}) \leq P_{y|z}(\tilde{d}) + \pi_{|z} \tag{21.28}$$

$$P_{1|1}(\tilde{d}) \leq P(y=1|z=1, \rho\underline{d} = \tilde{d}) \leq P_{1|1}(\tilde{d}) + \pi_{|1} \tag{21.29}$$

$$-P_{1|0}(\tilde{d}) - \pi_{|0} \leq -P(y=1|z=0, \rho\underline{d} = \tilde{d}) \leq -P_{1|0}(\tilde{d}) \tag{21.30}$$

QED

Chapter 22

Instrumental Variables

This chapter is based on Refs.[3] and [50].

The theory of potential outcomes (PO) discussed in Chapter 38 assumes that confounders can be ignored by conditioning on them. However, there are cases when that is not possible, as when there are some unmeasured (i.e., unobserved, hidden) confounder nodes in the bnet, because one can only condition on observed random variables, by definition. So what if confounders can't be ignored? Are we then precluded from using PO theory? Not necessarily. It might still be possible to use PO theory if one can find a suitable instrumental variable (IV) for the problem.

IVs were actually invented by Sewall Wright and his father Philip Wright long before PO theory was invented by Rubin. The reason why IVs save PO theory is greatly clarified by using Pearl causal DAGs and his d-separation theorem (see Chapter 13).

Most of the discussion in this chapter is limited to LDEN (linear deterministic bnets with external noise). These are discussed in Chapter 26. However, as will become obvious to the reader, IVs are also applicable and useful in general bnet modeling.

22.1 δ with unnmeasured confounder

In this section, we explain using LDENs why unmeasured confounders prejudice PO calculations.

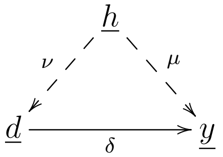


Figure 22.1: An LDEN bnet. The direct path $\underline{d} \rightarrow \underline{y}$ is confounded by a hidden variable $\underline{h}$.

Consider the LDEN bnet of Fig.22.1, For some $\delta, \mu \in \mathbb{R}$, we have

$$\underline{y} = \delta \underline{d} + \underbrace{\mu \underline{h} + \underline{u}_{\underline{y}}}_{\underline{n}_y} \ . \tag{22.1}$$

If $\langle \underline{n}_{\underline{y}}, \underline{d}\rangle = 0$, then

$$\langle \underline{y}, \underline{d}\rangle = \delta_0 \langle \underline{d}, \underline{d}\rangle \ , \tag{22.2}$$

whereas if $\langle \underline{n}_{\underline{y}}, \underline{d}\rangle \neq 0$, then

$$\langle \underline{y}, \underline{d}\rangle = \delta_1 \langle \underline{d}, \underline{d}\rangle + \langle \underline{n}_{\underline{y}}, \underline{d}\rangle \ . \tag{22.3}$$

Therefore,

$$\delta_0 = \frac{\langle \underline{y}, \underline{d}\rangle}{\langle \underline{d}, \underline{d}\rangle} \ , \tag{22.4}$$

$$\delta_1 = \underbrace{\frac{\langle \underline{y}, \underline{d}\rangle}{\langle \underline{d}, \underline{d}\rangle}}_{\delta_0} - \frac{\langle \underline{n}_{\underline{y}}, \underline{d}\rangle}{\langle \underline{d}, \underline{d}\rangle} \ . \tag{22.5}$$

If we assume no confounders and there is one, this gives the difference between the estimate δ_1 of δ for the truth, versus the naive estimate δ_0.

If the confounder $\underline{h}$ had been measured, then we would calculate the covariances at fixed $\underline{n}_{\underline{y}}$, and the conditional covariance $\langle \underline{n}_{\underline{y}}, \underline{d}\rangle_{|\underline{n}_{\underline{y}}} = 0$

22.2 δ (with unmeasured confounder) can be inferred via IV

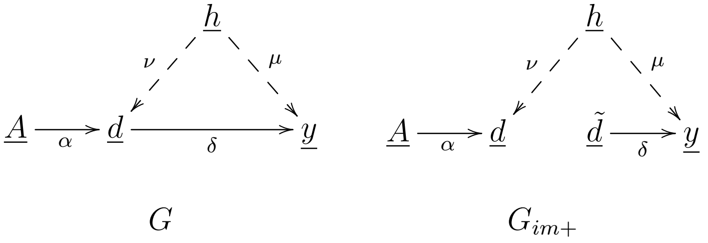


Figure 22.2: Two LDEN bnets. The direct path $\underline{d} \rightarrow \underline{y}$ is confounded by a hidden variable $\underline{h}$, but by using the IV $\underline{A}$, we are still able to identify (i.e. calculate) δ.

Now consider the two LDEN bnets shown in Fig.22.2. Note that there are no arrows $\underline{A} \rightarrow \underline{y}$ or $\underline{A} \rightarrow \underline{h}$. Note that node $\underline{d}$ is a collider in the path $\underline{A} - \underline{d} - \underline{h} - \underline{y}$, Therefore, the only unblocked path from $\underline{A}$ to $\underline{y}$ in G is $\underline{A} \rightarrow \underline{d} \rightarrow \underline{y}$ and that path has been removed in G_{im+}. These observations are encapsulated in the following statements.

$$\underline{d} \perp_G \underline{y} = \text{false}, \ \ \underline{A} \perp_G \underline{y} = \text{false} \ . \tag{22.6}$$

$$\underline{d} \perp_{G_{im+}} \underline{y} = \text{false}, \;\; \underline{A} \perp_{G_{im+}} \underline{y} = \text{true} \; . \tag{22.7}$$

The following is true for G:

$$\underline{y} = \delta \underline{d} + \underbrace{\mu \underline{h} + \underline{u}_{\underline{y}}}_{\underline{n}_{\underline{y}}} \tag{22.8}$$

$$\underline{d} = \alpha \underline{A} + \underbrace{\nu \underline{h} + \underline{u}_{\underline{d}}}_{\underline{n}_{\underline{d}}} \; . \tag{22.9}$$

Since $\langle \underline{n}_{\underline{y}}, \underline{A} \rangle = \langle \underline{n}_{\underline{d}}, \underline{A} \rangle = 0$ is true in G, we have

$$\langle \underline{y}, \underline{A} \rangle = \delta \langle \underline{d}, \underline{A} \rangle \tag{22.10}$$

and

$$\langle \underline{d}, \underline{A} \rangle = \alpha \langle \underline{A}, \underline{A} \rangle \; . \tag{22.11}$$

Note that $\langle \underline{y}, \underline{A} \rangle = \delta = 0$ for G_{im+} but not for G, so we are speaking about G from here on. It follows that

$$\alpha = \frac{\langle \underline{d}, \underline{A} \rangle}{\langle \underline{A}, \underline{A} \rangle} \tag{22.12}$$

and

$$\begin{aligned}
\delta &= \frac{\langle \underline{y}, \underline{A} \rangle}{\langle \underline{d}, \underline{A} \rangle} && (22.13) \\
&= \frac{\langle \underline{y}, \underline{A} \rangle}{\langle \underline{A}, \underline{A} \rangle} \frac{\langle \underline{A}, \underline{A} \rangle}{\langle \underline{d}, \underline{A} \rangle} && (22.14) \\
&= \frac{\langle \underline{y}, \underline{A} \rangle}{\langle \underline{A}, \underline{A} \rangle} \frac{1}{\alpha} && (22.15) \\
&= \frac{\langle \underline{y}, \alpha \underline{A} \rangle}{\langle \alpha \underline{A}, \alpha \underline{A} \rangle} \; . && (22.16)
\end{aligned}$$

22.3 More general bnets with IVs

Figs.22.3 and 22.4 are examples of other bnets for which the effect δ is identifiable thanks to the IV $\underline{A}$.

22.4 Instrumental Inequality

Pearl's instrumental inequality and related inequalities are discussed in Chapter 21.

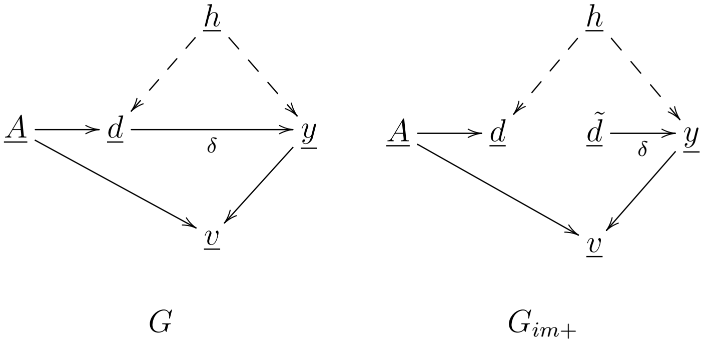


Figure 22.3: The 2 paths in G_{im+} from IV variable $\underline{A}$ to $\underline{y}$ are blocked by colliders $\underline{v}$ and $\underline{d}$. Thus, $\underline{d} \perp_{G_{im+}} \underline{y} =$ false, $\underline{A} \perp_{G_{im+}} \underline{y} =$ true

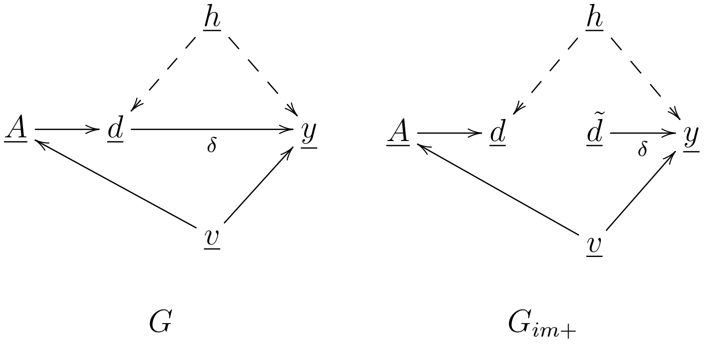


Figure 22.4: There are 2 paths in G_{im+} from IV variable $\underline{A}$ to $\underline{y}$. One is blocked by the collider $\underline{d}$ and the other can be blocked by conditioning on $\underline{v}$. Thus, $\underline{d} \perp_{G_{im+}} \underline{y}|\underline{v} =$ false, $\underline{A} \perp_{G_{im+}} \underline{y}|\underline{v} =$ true

Chapter 38

Potential Outcomes

This chapter is based on Ref.[3], a book by Stephen Cunningham entitled "Causal inference: the mixtape".

The theory of potential outcomes (PO) was for the most part invented in a seminal 1974 paper by Donald B. Rubin. Rubin has also made important extensions to PO theory since 1974. However, he refuses to use Pearl's causal DAGs to discuss PO theory. Pearl has shown that PO theory can be substantially clarified and extended by using the language of causal DAGs. The d-separation theorem that we discuss in Chapter 13 is especially useful in this regard.

In this chapter, we stress the connection of PO theory to bnets, and, in particular, to the do and imagine operators defined in Chapter 8. Hence, before reading this chapter, the reader is expected to have at least skimmed Chapter 8, so that he/she understands the definition of do and imagine operators.

σ	$\underline{d}^\sigma$	$\underline{y}^\sigma$	$(1-\underline{d}^\sigma)\underline{y}^\sigma$	$\underline{d}^\sigma\underline{y}^\sigma$
Andy	1	10	.	10
Ben	1	5	.	5
Chad	1	16	.	16
Daniel	1	3	.	3
Edith	0	5	5	.
Frank	0	7	7	.
George	0	8	8	.
Hank	0	10	10	.

Table 38.1: Dataset describing whether individual σ took a treatment dose ($d^\sigma = 1$) or didn't ($d^\sigma = 0$). The treatment outcome is measured by the real number y^σ.

Suppose a **population of individuals** $\sigma = 0, 1, 2, \ldots, nsam - 1$ is given ($d^\sigma = 1$) or not given ($d^\sigma = 0$) a **treatment drug dose** d^σ, and that the **treatment outcome (i.e., response)** is measured by a real number y^σ. Table 38.1 gives a possible dataset for this scenario. As you can see from that table, each individual either takes a drug dose or doesn't, but not both. PO theory can be viewed as a **missing data (MD) problem**. MD problems are discussed in Chapter 31. However, the PO MD problem is much more specialized than the generic MD problems discussed in Chapter 31. In the PO MD problem, we can fill in the blank cells by matching each individual that took the drug with another *similar* individual that didn't. We will have much more to say about this matching strategy later in this chapter.

One can define similar individuals as individuals that have the same value for nx features $x^\sigma = (x^\sigma_i)_{i=0,1,\ldots,nx-1}$. One can add to Table 38.1 nx extra columns giving the value of the feature vector x^σ for each individual. Members of a population with the same x^σ are referred to as a **subpopulation or stratum (ie., layer)**.

In a **randomized clinical trial (RCT)**, the effect of the variable x^σ on the value of d^σ is eliminated by randomizing the population and therefore making the effect of x^σ average out to zero. However, there are many situations in which carrying out an RCT is not possible. PO theory is a way of predicting the result of an RCT in situations where doing a real RCT is not physically possible.

In this chapter, x^σ will be called the confounders. Implicit throughout this chapter is the assumption that there are **no unmeasured confounders**. Because if there are some unmeasured confounders, those can send secret messages that influence the value that d^σ takes. This would ruin the predictions of someone trying to predict the results of an RCT without being privy to those secret messages. When there are **some unmeasured confounders**, it might still be possible to predict the effect of an RCT. This might be possible using instrumental variables. See Chapter 22 for a discussion of **instrumental variables**.

38.1 G and G_{den}, bnets, the starting point bnets

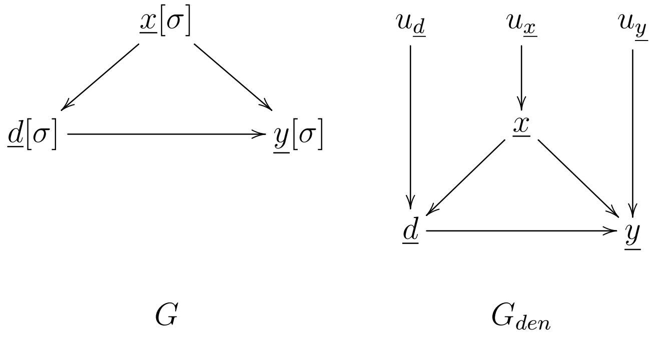


Figure 38.1: Bnets G and G_{den} are our starting point in discussing PO theory. G is for a single individual σ of the population. Bnet G_{den} is the DEN counterpart to G. DEN (Deterministic with External Noise) bnets are discussed in Chapter 26.

In this chapter, we will abbreviate $\underline{X}[\sigma] = \underline{X}^\sigma$ for $X \in \{d, x, y\}$ and for $\sigma = \{0, 1, 2, \ldots, nsam - 1\}$.

For each individual (aka unit, sample) $\sigma = 0, 1, 2, \ldots nsam - 1$, let:

$\underline{d}^\sigma \in \{0, 1\}$: treatment discrete drug dose, 1 if treated and 0 if untreated

$\underline{y}^\sigma \in \mathbb{R}$: treatment potential outcome

$\underline{x}^\sigma$: column vector of treatment confounders (aka covariates, because they are often used as covariates (i.e., independent variables) in linear regression.)

Consider bnets G and G_{den} in Fig.38.1. G reflects the language used in Ref.[3] to discuss PO theory. And G_{den} reflects the language that Judea Pearl prefers to use to discuss PO theory. Both languages are equivalent. To go from one language to the other, one need only perform the following swaps, where $\underline{u}$ is the external noise of the DEN bnet.

$\underline{X}^\sigma \leftrightarrow \underline{X}(\underline{u})$ for $X \in \{d, x, y\}$.
$P(\sigma) = \frac{1}{nsam} \leftrightarrow P(u)$
$\sum_\sigma P(\sigma)(\cdot) \leftrightarrow \sum_u P(u)(\cdot)$
The TPMs, printed in blue, for the bnet G in Fig.38.1, are as follows:

$$P(x^\sigma) = P_{\underline{x}}(x^\sigma) \tag{38.1}$$

$$P(d^\sigma|x^\sigma) = P_{\underline{d}|\underline{x}}(d^\sigma|x^\sigma) \tag{38.2}$$

$$P(y^\sigma|x^\sigma, d^\sigma) = P_{\underline{y}|\underline{x},\underline{d}}(y^\sigma|x^\sigma, d^\sigma) \tag{38.3}$$

Now let:
$\underline{d} \in \{0, 1\}$: treatment discrete drug dose, 1 if treated and 0 if untreated
$\underline{y} \in \mathbb{R}$: treatment potential outcome
$\underline{x}$: column vector of treatment confounders (aka covariates)
$\underline{u} = (\underline{u}_{\underline{d}}, \underline{u}_{\underline{x}}, \underline{u}_{\underline{y}})$: external noise
The TPMs, printed in blue, for the bnet G_{den} in Fig.38.1, are as follows:

$$P(x|u_{\underline{x}}) = \mathbb{1}(\ \ x = u_{\underline{x}}\ \ ) \tag{38.4}$$

$$P(d|x, u_{\underline{d}}) = \mathbb{1}(\ \ d = f_{\underline{d}}(x, u_{\underline{d}})\ \ ) \tag{38.5}$$

$$P(y|d, x, u_{\underline{y}}) = \mathbb{1}(\ \ y = f_{\underline{y}}(d, x, u_{\underline{y}})\ \ ) \tag{38.6}$$

If we linearize $f_{\underline{y}}$ in Eq.(38.6), we get

$$\underline{y} = \delta\underline{d} + \beta\underline{x} + \underline{u}_{\underline{y}}\ , \tag{38.7}$$

where $\delta, \beta \in \mathbb{R}$. Assuming that $\underline{x}, \underline{y} \in \mathbb{R}$ and $\underline{d} \in \{0, 1\}$, Eq.(38.7) can be plotted. The resulting plot is given in Fig.38.2. This plot is a very special case of the PO problem, but it gives a crude idea of the "effects" $\delta = y(1) - y(0)$ that PO theory gives estimates for. Any individual in the experiment experiences either $y(1)$ or $y(0)$, but not both.

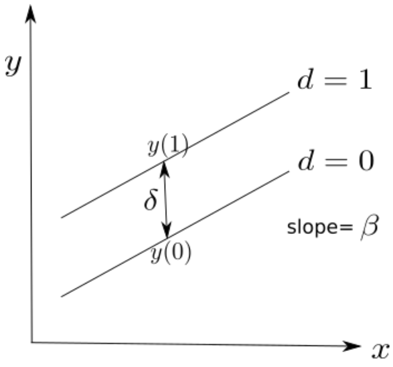
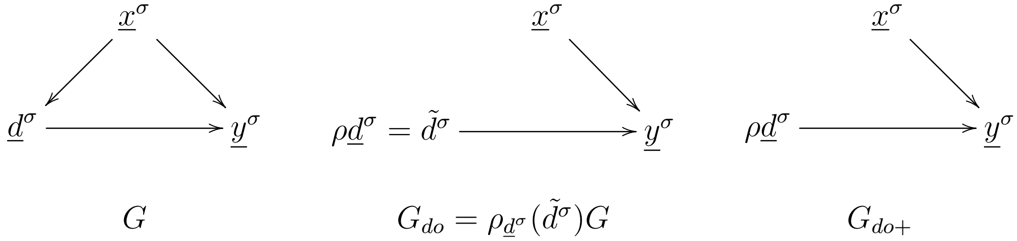


Figure 38.2: Plot of Eq.(38.7)

38.2 G_{do+} bnet



G

$G_{do} = \rho_{\underline{d}^\sigma}(\tilde{d}^\sigma)G$

G_{do+}

Figure 38.3: Bnet $G_{do} = \rho_{\underline{d}^\sigma}(\tilde{d}^\sigma)G$ is obtained by applying the do operator to node $\underline{d}^\sigma$ of bnet G. Bnet G_{do+} is obtained by adding a prior probability distribution $P(\tilde{d}^\sigma)$ to node $\rho\underline{d}^\sigma$ of bnet G_{do}.

Fig.38.3 shows how bnet G_{do} is obtained by applying the do operator to bnet G, and how bnet G_{do+} is obtained by adding a prior probability distribution to one of the nodes of G_{do}. In bnet G_{do}, node $\underline{d}^\sigma$ has been stripped of all outside influences and fixed to a specific state $\tilde{d}^\sigma$. This is what an RCT does.

The TPMs, printed in blue, for the bnets G_{do} and G_{do+}, are as follows. Note that the TPMs for bnets G_{do} and G_{do+} are defined in terms of the TPMs of bnet G.

$$P(x^\sigma) = P_{\underline{x}}(x^\sigma) \tag{38.8}$$

$$P_{\rho\underline{d}}(d) = \sum_x P_{\underline{d}|\underline{x}}(d|x)P_{\underline{x}}(x) \tag{38.9}$$

$$P(\tilde{d}^\sigma) = \left\{ \begin{array}{ll} \delta(\tilde{d}^\sigma, (\tilde{d}^\sigma)') & \text{for } G_{do} \\ P_{\rho\underline{d}}(\tilde{d}^\sigma) & \text{for } G_{do+} \end{array} \right. \tag{38.10}$$

$$P(y^\sigma | x^\sigma, \tilde{d}^\sigma) = P_{\underline{y}|\underline{x},\underline{d}}(y^\sigma | x^\sigma, \tilde{d}^\sigma) \tag{38.11}$$

It is convenient to define the following expected values of $\underline{y}^\sigma$ in terms of the TPMs of bnet G_{do+}:

$$\mathcal{Y}_{|\tilde{d},x} = E_{\sigma|\tilde{d},x}[\underline{y}^\sigma(\tilde{d})] \to E_{y|\tilde{d},x}[\underline{y}(\tilde{d})] = \sum_y yP(y|\tilde{d},x) \tag{38.12}$$

$$\mathcal{Y}_{|\tilde{d}} = E_{\sigma|\tilde{d}}[\underline{y}^\sigma(\tilde{d})] \to E_{y|\tilde{d}}[\underline{y}(\tilde{d})] = \sum_x \mathcal{Y}_{|\tilde{d},x} P(x) \tag{38.13}$$

$$\mathcal{Y}_{|x} = E_{\sigma|x}[\underline{y}^\sigma(\tilde{d})] \to E_{y|x}[\underline{y}(\tilde{d})] = \sum_{\tilde{d}} \mathcal{Y}_{|\tilde{d},x} P(\tilde{d}) \tag{38.14}$$

$$\mathcal{Y} = E_\sigma[\underline{y}^\sigma] \to E[\underline{y}] = \sum_{\tilde{d},x} \mathcal{Y}_{|\tilde{d},x} P(\tilde{d})P(x) \tag{38.15}$$

38.3 G_{im+} bnet

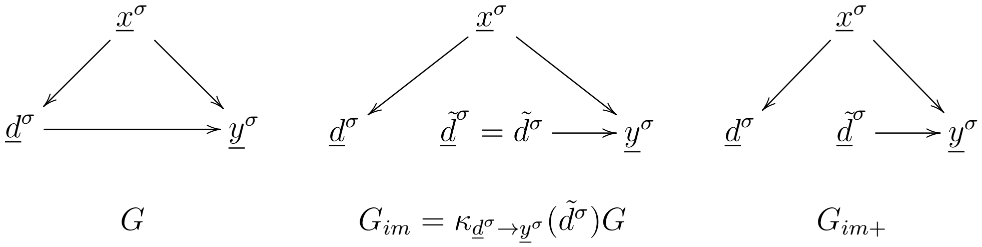


Figure 38.4: Bnet $G_{im} = \kappa_{\underline{d}^\sigma \to \underline{y}^\sigma}(\tilde{d}^\sigma)G$ is obtained by applying the imagine operator to arrow $\underline{d}^\sigma \to \underline{y}^\sigma$ of bnet G. Bnet G_{im+} is obtained by adding a prior probability distribution $P(\tilde{d}^\sigma)$ to node $\underline{\tilde{d}}^\sigma$ of bnet G_{im}.

Fig.38.4 shows how bnet G_{im} is obtained by applying an imagine operator to bnet G, and how bnet G_{im+} is obtained from bnet G_{im} by adding a prior probability distribution to one of the nodes of G_{im}. $\underline{d} \in \{0, 1\}$ represents the dose that a patient is told to take by a doctor, and $\underline{\tilde{d}} \in \{0, 1\}$ represents the dose he actually takes. If $\underline{d} = \underline{\tilde{d}}$, the patient is compliant, and if $\underline{d} \neq \underline{\tilde{d}}$, he is non-compliant.

It is convenient to define $\underline{y}^\sigma(\tilde{d}^\sigma)$ so that

$$\underline{y}^\sigma(\tilde{d}^\sigma) = \underline{y}^\sigma_{G_{im}} \tag{38.16}$$

and

$$\underline{y}^\sigma = \underline{y}^\sigma(\tilde{d}^\sigma) = \underline{y}^\sigma(1)\tilde{d}^\sigma + \underline{y}^\sigma(0)(1-\tilde{d}^\sigma)\ . \tag{38.17}$$

The TPMs, printed in blue, for the nodes of bnets G_{im} and G_{im+}, are as follows. Note that the TPMs for bnets G_{im} and G_{im+} are defined in terms of the TPMs of bnet G. Note that the prior $P(\tilde{d})$ is not arbitrary; it's calculated from the TPMs of bnet G.

$$P(x^\sigma) = P_{\underline{x}}(x^\sigma) \tag{38.18}$$

$$P(d^\sigma|x^\sigma) = P_{\underline{d}|\underline{x}}(d^\sigma|x^\sigma) \tag{38.19}$$

$$\pi_{\tilde{d}} = P(\tilde{d}) = \sum_x P_{\underline{d}|\underline{x}}(\tilde{d}|x)P_{\underline{x}}(x) \tag{38.20}$$

$$P(\tilde{d}^\sigma) = \left\{ \begin{array}{ll} \delta(\tilde{d}^\sigma, (\tilde{d}^\sigma)') & \text{for } G_{im} \\ \pi_{\tilde{d}^\sigma} & \text{for } G_{im+} \end{array} \right. \tag{38.21}$$

$$P(y^\sigma(\tilde{d}^\sigma)|x^\sigma, \tilde{d}^\sigma) = P_{\underline{y}|\underline{x},\underline{d}}(y^\sigma(\tilde{d}^\sigma)|x^\sigma, \tilde{d}^\sigma) \tag{38.22}$$

It is convenient to define the following expected values of $\underline{y}^\sigma$ in terms of the TPMs of bnet G_{im+}:

$$\mathcal{Y}_{d|\tilde{d},x} = E_{\sigma|d,\tilde{d},x}[\underline{y}^\sigma(\tilde{d})] \rightarrow E_{y|d,\tilde{d},x}[\underline{y}(\tilde{d})] = \sum_y yP(y|\tilde{d},x)P(d|x) \tag{38.23}$$

$$\mathcal{Y}_{d|\tilde{d}} = E_{\sigma|d,\tilde{d}}[\underline{y}^\sigma(\tilde{d})] \rightarrow E_{y|d,\tilde{d}}[\underline{y}(\tilde{d})] = \sum_x \mathcal{Y}_{d|\tilde{d},x}P(x) \tag{38.24}$$

$$\mathcal{Y}_{d|x} = E_{\sigma|d,x}[\underline{y}^\sigma(\tilde{d})] \rightarrow E_{y|d,x}[\underline{y}(\tilde{d})] = \sum_{\tilde{d}} \mathcal{Y}_{d|\tilde{d},x}P(\tilde{d}) \tag{38.25}$$

$$\mathcal{Y}_{d} = E_{\sigma|d}[\underline{y}^\sigma(\tilde{d})] \rightarrow E_{y|d}[\underline{y}(\tilde{d})] = \sum_{\tilde{d},x} \mathcal{Y}_{d|\tilde{d},x}P(\tilde{d})P(x) \tag{38.26}$$

$\mathcal{Y}_{0|0}, \mathcal{Y}_{1|1}$ are said to be **factual** (indicating compliant patients) whereas $\mathcal{Y}_{0|1}, \mathcal{Y}_{1|0}$ are said to be **counterfactual** (indicating non-compliant patients).

38.4 Translation Dictionary

In standard PO notation	In our notation (for G_{im+}, unless otherwise specified)
i, individual (i.e., unit, sample) index	σ
$D_i = d_i$, treatment dose	$\underline{d}^\sigma = d^\sigma$
$Y_i = y_i$, treatment outcome	$\underline{y}^\sigma = y^\sigma$
$X_i = x_i$, treatment confounders	$\underline{x}^\sigma = x^\sigma$
$Y_i(\tilde{d})$	$\underline{y}^\sigma = \underline{y}^\sigma(\tilde{d})$ for G_{im}
$E[Y_i(d)]$	$E_\sigma[\underline{y}^\sigma(d)] = \mathcal{Y}_d$
$E[Y_i \vert D_i = \tilde{d}]$	$E_{\sigma\vert\tilde{d}}[\underline{y}^\sigma] = \mathcal{Y}_{\vert\tilde{d}}$
$E[Y_i(d) \vert D_i = \tilde{d}]$	$E_{\sigma\vert d,\tilde{d}}[\underline{y}^\sigma] = \mathcal{Y}_{d\vert\tilde{d}}$
$E[Y_i(d) \vert D_i = \tilde{d}, X = x]$	$E_{\sigma\vert d,\tilde{d},x}[\underline{y}^\sigma] = \mathcal{Y}_{d\vert\tilde{d},x}$

Table 38.2: Dictionary for translating from standard PO notation of Ref.[3] to our notation.

Table 38.2 gives a dictionary for translating from the standard PO notation of Ref.[3] to our notation. $d, \tilde{d} \in \{0, 1\}$. I find the standard PO notation confusing because it often uses D_i to represent two different nodes, $\underline{d}^\sigma$ and $\underline{\tilde{d}}^\sigma$ in G_{im+}. This confusion becomes particularly distressing when we are told in PO notation that

$$Y_i = D_i Y_i(0) + (1 - D_i) Y_i(1) \ . \tag{38.27}$$

This could mean that

$$\underline{y}^\sigma = \underline{\tilde{d}}^\sigma y^\sigma(1) + (1 - \underline{\tilde{d}}^\sigma)\underline{y}^\sigma(0) \tag{38.28}$$

which is always true, or it could mean

$$\underline{y}^\sigma = \underline{d}^\sigma \underline{y}^\sigma(1) + (1 - \underline{d}^\sigma)\underline{y}^\sigma(0) \tag{38.29}$$

which is not necessarily true.

38.5 $\mathcal{Y}_{d|\tilde{d}}$ differences (aka treatment effects)

Note the $\mathcal{Y}_d$ and $\mathcal{Y}_{|\tilde{d}}$ are not the same thing.

$$\mathcal{Y}_{|\tilde{d}} = \sum_d \mathcal{Y}_{d|\tilde{d}} = \mathcal{Y}_{0|\tilde{d}} + \mathcal{Y}_{1|\tilde{d}} \tag{38.30}$$

whereas

$$\mathcal{Y}_d = \sum_{\tilde{d}} \mathcal{Y}_{d|\tilde{d}} P(\tilde{d}) = \mathcal{Y}_{d|0}\pi_0 + \mathcal{Y}_{d|1}\pi_1 \ . \tag{38.31}$$

$\mathcal{Y}_{|\tilde{d}}$ is connected to the do operator as follows.

$$P(\underline{y} = y|\rho\underline{d} = \tilde{d}) = \sum_x P(y|\tilde{d}, x)P(x) \tag{38.32}$$

so

$$\mathcal{Y}_{|\tilde{d}} = \sum_y yP(\underline{y} = y|\rho\underline{d} = \tilde{d}) \ . \tag{38.33}$$

In particular, when $\underline{y}$ is binary (i.e., $\underline{y} \in \{0, 1\}$), Eq.(38.33) becomes

$$\mathcal{Y}_{|\tilde{d}} = P(\underline{y} = 1|\rho\underline{d} = \tilde{d}) \ . \tag{38.34}$$

It is convenient to define the following effects. Note that we use the word **"effect"** to refer to a difference of two $\mathcal{Y}_{d|\tilde{d}}$.

- average controlled effect (ACE), used when doing an RCT.

$$ACE = \mathcal{Y}_{|1} - \mathcal{Y}_{|0} \tag{38.35}$$

- average treatment effect[1] (ATE).

$$ATE = \mathcal{Y}_1 - \mathcal{Y}_0 = \delta \tag{38.36}$$

- average treatment effect of the treated (ATT)

$$ATT = \mathcal{Y}_{1|1} - \mathcal{Y}_{0|1} \tag{38.37}$$

- average treatment effect of the untreated (ATU)

$$ATU = \mathcal{Y}_{1|0} - \mathcal{Y}_{0|0} \tag{38.38}$$

- simple difference in outcomes (SDO)

$$SDO = \mathcal{Y}_{1|1} - \mathcal{Y}_{0|0} \tag{38.39}$$

- selection bias (SB)

$$SB = \mathcal{Y}_{0|1} - \mathcal{Y}_{0|0} \tag{38.40}$$

[1] Note that effects in which $\tilde{d}$ varies are called "controlled", whereas those in which d varies instead, are called simply "treatments". y is averaged over in both cases.

Note that some of these effects are linearly related

$$\underbrace{\mathcal{Y}_{1|1} - \mathcal{Y}_{0|0}}_{SDO} = \underbrace{(\mathcal{Y}_{1|1} - \mathcal{Y}_{0|1})}_{ATT} + \underbrace{\mathcal{Y}_{0|1} - \mathcal{Y}_{0|0}}_{SB} \tag{38.41}$$

$$\underbrace{\mathcal{Y}_{1} - \mathcal{Y}_{0}}_{ATE} = \underbrace{(\mathcal{Y}_{1|1} - \mathcal{Y}_{0|1})}_{ATT}\pi_1 + \underbrace{(\mathcal{Y}_{1|0} - \mathcal{Y}_{0|0})}_{ATU}\pi_0 \tag{38.42}$$

$$\underbrace{\mathcal{Y}_{1|1} - \mathcal{Y}_{0|0}}_{SDO} \quad = \quad \underbrace{(\mathcal{Y}_{1|1} - \mathcal{Y}_{0|1})\pi_1 + (\mathcal{Y}_{1|0} - \mathcal{Y}_{0|0})\pi_0}_{ATE} \tag{38.43}$$

$$+ \underbrace{\mathcal{Y}_{0|1} - \mathcal{Y}_{0|0}}_{SB} \tag{38.44}$$

$$+ \underbrace{(\mathcal{Y}_{1|1} - \mathcal{Y}_{0|1})}_{ATT}\pi_0 \tag{38.45}$$

$$- \underbrace{(\mathcal{Y}_{1|0} - \mathcal{Y}_{0|0})}_{ATU}\pi_0 \tag{38.46}$$

38.6 Ignorability

Confounders are said to be **ignorable** (i.e., there is independence from confounders) if

$$\underline{y}^{\sigma}(\underline{\tilde{d}}^{\sigma}) \perp_P \underline{d}^{\sigma} \ . \tag{38.47}$$

This is satisfied by G_{do}. To prove so, check that

$$\underline{y}^{\sigma}(\underline{\tilde{d}}^{\sigma}) \perp_{G_{do}} \underline{d}^{\sigma} \tag{38.48}$$

and then invoke the d-separation theorem (see Chapter 13). When confounders are ignorable, we have

$$\mathcal{Y}_{d|\tilde{d}} = \mathcal{Y}_{d|y(\tilde{d})} = \mathcal{Y}_{d} \ , \tag{38.49}$$

or, equivalently,

$$\mathcal{Y}_{d|\tilde{d}=0} = \mathcal{Y}_{d|\tilde{d}=1} \ . \tag{38.50}$$

Therefore,

$$ATE = ATT = ATU = SDO \tag{38.51a}$$

and

$$SB = 0 \ . \tag{38.51b}$$

Let $\mathcal{E} \in \{ACE, ATE, ATT, ATU, SDO, SB\}$. $\mathcal{E}$ can be defined for a fixed stratum x by replacing $\mathcal{Y}_{d|\tilde{d}}$ with $\mathcal{Y}_{d|\tilde{d},x}$. We will denote such an extension by $\mathcal{E}_x$, or, sometimes, simply by $\mathcal{E}$.

Confounders are said to be **conditionally ignorable** (i.e., there is conditional independence from confounders) if

$$\underline{y}^\sigma(\underline{\tilde{d}}^\sigma) \perp_P \underline{d}^\sigma|\underline{x}^\sigma \; . \tag{38.52}$$

This is satisfied by G_{im}. To prove this, check that

$$\underline{y}^\sigma(\underline{\tilde{d}}^\sigma) \perp_{G_{im}} \underline{d}^\sigma|\underline{x}^\sigma \tag{38.53}$$

and then invoke the d-separation theorem (see Chapter 13). When confounders are conditionally ignorable, we have

$$\mathcal{Y}_{d|\tilde{d},x} = \mathcal{Y}_{d|y(\tilde{d}),x} = \mathcal{Y}_{d|x} \; , \tag{38.54}$$

or, equivalently,

$$\mathcal{Y}_{d|\tilde{d}=0,x} = \mathcal{Y}_{d|\tilde{d}=1,x} \; . \tag{38.55}$$

If conditional ignorability holds, then Eqs.(38.51) are valid at fixed x.

38.7 Hypothesis testing for sharp null

In this section, we assume no x dependence, or else we assume that the whole discussion refers to a single stratum x. Hence, we will omit the x subscript in this section.

Assume $\underline{d}^\sigma = \underline{\tilde{d}}^\sigma$. In this section, we will discuss hypothesis testing between the following two opposite hypotheses:

$$\begin{array}{l} H_0 : \underline{y}^\sigma(1) = \underline{y}^\sigma(0) \;\; \forall \sigma \\ H_1 = \; !H_0 \end{array} \; . \tag{38.56}$$

H_0 is called the **sharp null hypothesis**.

Table 38.1 becomes Table 38.3 if we fill the blank cells according to the sharp null hypothesis H_0. And Table 38.3 becomes Table 38.4 by permuting the entries of column $\underline{d}^\sigma$.

Define

$$N_1 = \sum_\sigma d^\sigma, \quad \pi_1 = \frac{N_1}{nsam} \; , \tag{38.57}$$

$$N_0 = \sum_\sigma (1 - d^\sigma) = nsam - N_1, \quad \pi_0 = \frac{N_0}{nsam} \; , \tag{38.58}$$

$$\vec{d} = (d^\sigma)_{\sigma=01,2,\ldots,nsam-1} \; , \tag{38.59}$$

σ	$\underline{d}^\sigma$	$\underline{y}^\sigma$	$(1-\underline{d}^\sigma)\underline{y}^\sigma$	$\underline{d}^\sigma\underline{y}^\sigma$
Andy	1	10	10	10
Ben	1	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	0	7	7	7
George	0	8	8	8
Hank	0	10	10	10

Table 38.3: Table 38.1 with blank cells filled according to the sharp null hypothesis H_0. Note that $\vec{d}_0 = (1,1,1,1,0,0,0,0)$. If $\xi(\vec{d})$ is defined by Eq.(38.60), then $\xi = |34-30|/4 = 1$

σ	$\underline{d}^\sigma$	$\underline{y}^\sigma$	$(1-\underline{d}^\sigma)\underline{y}^\sigma$	$\underline{d}^\sigma\underline{y}^\sigma$
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	1	7	7	7
George	0	8	8	8
Hank	0	10	10	10

Table 38.4: Table 38.3 after permuting the entries of column $\underline{d}^\sigma$. Note that $\vec{d}_1 = (1,0,1,1,0,1,0,0)$. If $\xi(\vec{d})$ is defined by Eq.(38.60), then $\xi = |36-28|/4 = 2$

and

$$\xi(\vec{d}) = \left| \frac{\sum_\sigma d^\sigma y^\sigma}{N_1} - \frac{\sum_\sigma (1-d^\sigma) y^\sigma}{N_0} \right| \rightarrow |\mathcal{Y}_1 - \mathcal{Y}_0| \; . \tag{38.60}$$

$\xi(\vec{d})$ is the statistic that we will use to test the sharp null hypothesis H_0. There are other possible functions $\xi(\vec{d})$ that are commonly used to test H_0; for example, the Kolgomorov-Smirmov statistic (see Ref.[3])

According to Tables 38.3 and 38.4, $\xi(\vec{d}) = 1$ for the true (i.e., H_0-satisfying) vector $\vec{d} = \vec{d}_0$, and $\xi(\vec{d}) = 2$ for one possible H_1-satisfying vector $\vec{d} = \vec{d}_1$. Let $\mathcal{D}$ be the set of all permutations of $\vec{d}_0$, and define $F : \mathbb{R} \rightarrow [0,1]$ by

$$\begin{aligned} F(\xi) &= \frac{1}{|\mathcal{D}|} \sum_{\vec{d}\in\mathcal{D}} \mathbb{1}(\xi(\vec{d}) \leq \xi) && (38.61) \\ &= E_{\vec{d}}[\mathbb{1}(\xi(\vec{d}) \leq \xi)] \quad \text{where } P(\vec{d}) = \frac{1}{|\mathcal{D}|} && (38.62) \end{aligned}$$

The function $F(\xi)$ is monotonically increasing from $F(-\infty) = 0$ to $F(+\infty) = 1$ so it can be interpreted to be a cumulative distribution for ξ. Then one can define a p value for the sharp null hypothesis by

$$p = 1 - F(\xi(\vec{d_0})) \ . \tag{38.63}$$

$p \in [0, 1]$ measures the likelihood that H_0 is true. The smaller it is, the less likely H_0 is.

Often, $|\mathcal{D}|$ = the number of permutations of $\vec{d_0}$, is too large to average over all the elements of $\mathcal{D}$. In that case, one can use random sampling methods. For example, one can choose a $\vec{d}$ at random from $\mathcal{D}$, and calculate a step function $F_i(\xi)$ from that. Do this ni times. Then average all the ni step functions to obtain an estimate of $F(\xi)$.

38.8 Matching Strata

For a situation described by the bnet G_{im+}, we can match *similar* individuals to fill the blank cells of Table 38.1. By "similar", we mean that they have the same or almost the same value of $\underline{x}^\sigma$.

38.8.1 1-1 strata-match

In 1-1 strata-match, we match each individual with $d^\sigma = 1$ with exactly one individual with $d^\sigma = 0$.

Recall that

$$SDO = \mathcal{Y}_{1|1,x} - \mathcal{Y}_{0|0,x} \tag{38.64a}$$

$$ATT = \mathcal{Y}_{1|1,x} - \mathcal{Y}_{0|1,x} \tag{38.64b}$$

$$ATU = \mathcal{Y}_{1|0,x} - \mathcal{Y}_{0|0,x} \tag{38.64c}$$

$$ATE = ATT\pi_1 + ATU\pi_0 \tag{38.64d}$$

Note that π_0, π_1 do not depend on x. This can be justified by looking at the bnet G_{im+}, for which $\pi_{\tilde{d}} = P(\tilde{d})$ is the prior of node $\underline{\tilde{d}}$.

Eqs.(38.64) can be estimated from the data via the following estimators. In these estimators, $s(\sigma)$ is the single match for individual σ.

$$\widehat{SDO} = \frac{1}{N_1}\sum_\sigma d^\sigma y^\sigma - \frac{1}{N_0}\sum_\sigma (1 - d^\sigma) y^\sigma \tag{38.65}$$

$$\begin{aligned} \widehat{ATT} &= \frac{1}{N_1}\sum_\sigma d^\sigma y^\sigma - \frac{1}{N_1}\sum_\sigma d_\sigma y^{s(\sigma)} & (38.66) \\ &= \frac{1}{N_1}\sum_\sigma d^\sigma (y^\sigma - y^{s(\sigma)}) & (38.67) \end{aligned}$$

$$\widehat{ATU} = \frac{1}{N_0}\sum_\sigma (1-d^\sigma)(y^{s(\sigma)}-y^\sigma) \tag{38.68}$$

$$\begin{aligned}
\widehat{ATE} &= \widehat{ATT}\pi_1 + \widehat{ATU}\pi_0 & (38.69)\\
&= \frac{1}{nsam}[\widehat{ATT}N_1 + \widehat{ATU}N_0] & (38.70)\\
&= \frac{1}{nsam}\left[\sum_\sigma d^\sigma(y^\sigma - y^{s(\sigma)}) + \sum_\sigma (1-d^\sigma)(y^{s(\sigma)}-y^\sigma)\right] & (38.71)\\
&= \frac{1}{nsam}\sum_\sigma (2d^\sigma - 1)(y^\sigma - y^{s(\sigma)}) & (38.72)
\end{aligned}$$

38.8.2 Exact and approximate strata-match

It is very often the case that one can't find for a given individual σ another individual that has exactly the same value of x^σ. In such cases, one can discard all matchless individuals. But that would entail a loss of precious information. Instead of discarding orphans, a better way is to relax our demands and match individual σ with another individual s such that x^s and x^σ are very close in some metric.

More precisely, for some arbitrary parameter $\epsilon > 0$, and an individual σ with $d^\sigma = 1$, define the **strata-matching set** $\mathcal{M}_\epsilon(\sigma)$ by[2]

$$\mathcal{M}_\epsilon(\sigma) = \{s : d^\sigma = 1, d^s = 0, dist(x^\sigma, x^s) \leq \epsilon\} \ , \tag{38.73}$$

where

$$dist(x^\sigma, x^s) = [x^\sigma]^T[\Sigma]^{-1}x^s \ , \tag{38.74}$$

where $\Sigma = \langle \underline{x}^\sigma, [\underline{x}^s]^T\rangle$. This metric $dist(x^\sigma, x^s)$ is called the **Mahalanobis distance**. We will call the case $\epsilon = 0$ an **exact strata-match**, and the case $\epsilon \neq 0$ an **approximate strata-match.**. To do an approximate strata-match, replace $y^{s(\sigma)}$ by $\langle y\rangle^\sigma$ in the estimators given above for a 1-1 strata-match. $\langle y\rangle^\sigma$ is defined by

$$\langle y\rangle^\sigma = \frac{1}{|\mathcal{M}_\epsilon(\sigma)|}\sum_{s\in\mathcal{M}_\epsilon(\sigma)} y^s \ . \tag{38.75}$$

Ref.[3] calculates the mean and variance of estimator $\widehat{ATT}$. The mean is biased, but one can define a new bias-corrected estimator.

[2] One can use an ϵ that depends on σ. Let $\epsilon(\sigma, 5)$ be the radius necessary so that $\mathcal{M}_{\epsilon(\sigma,5)}(\sigma)$ contains exactly 5 elements s.. Thus, $\mathcal{M}_{\epsilon(\sigma,5)}(\sigma)$ contains the s of the 5 points x^s that are the nearest neighbors of x^σ in the $dist()$ metric.

38.8.3 Positivity

Positivity is defined as the requirement that for all layers x,

$$0 < P(d^\sigma = 1 | x^\sigma = x) < 1 \tag{38.76}$$

or, equivalently,

$$P(d^\sigma = 1 | x^\sigma = x) > 0 \quad \text{and} \quad P(d^\sigma = 0 | x^\sigma = x) > 0 \; . \tag{38.77}$$

In other words, for each layer x, there is a non-zero probability of being both treated and untreated.

If positivity is violated, then for some layer x, $\mathcal{Y}_{0|\tilde{d},x}$ or $\mathcal{Y}_{1|\tilde{d},x}$ is zero, so we can't calculate effect estimators such as $\widehat{ATE}$. If $\widehat{ATE}$ can be calculated, one says it is identifiable (i.e., calculable). Positivity is a requirement for identifiability of $\widehat{ATE}$.

When $P(d^\sigma | x^\sigma = x)$ becomes 0 or 1 for some x, the arrow $\underline{x} \to \underline{d}$ becomes deterministic for some x. This situation is the very antithesis of RCTs, wherein the influence exerted by $\underline{x}^\sigma$ on $\underline{d}^\sigma$ is uniformly random and therefore ignorable. Hence, it is perhaps not too surprisingly that a violation of positivity makes $\widehat{ATE}$ non-identifiable.

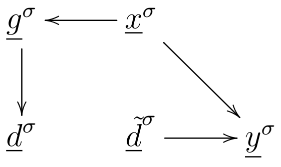
38.9 Propensity Score

It is often the case that the discrete vector $\underline{x}^\sigma$ has too many possible values to make matching possible. In such cases, it is convenient to map the space of vectors $\underline{x}^\sigma$ to the real line. One very convenient choice for that map is the **propensity score**, which is defined as

$$g(x^\sigma) = P(d^\sigma = 1 | x^\sigma) \; . \tag{38.78}$$

The propensity score is usually approximated by a sigmoid function using logistic regression[3]

$$g(x^\sigma) = \text{sig}(\alpha + \beta x^\sigma) \tag{38.79}$$



G_{ps}

Figure 38.5: Bnet G_{ps} used when doing propensity scoring.

[3] The sigmoid function is defined in Chapter Notational Conventions and Preliminaries to be $\text{sig}(x) = 1/(1 - e^{-x})$.

To use the propensity score, one replaces the bnet G_{im+} by the bnet G_{ps} shown in Fig.38.5. The TPMs, printed in blue, for the 2 nodes of G_{ps} that differ from the nodes of G_{im+}, are as follows:

$$P(g^\sigma|x^\sigma) = \delta(g^\sigma, g(x^\sigma)) \tag{38.80}$$

$$P(d^\sigma|g^\sigma) = g^\sigma d^\sigma + (1-g^\sigma)(1-d^\sigma) \tag{38.81}$$

Note that these TPMs are self-consistent because

$$\begin{aligned}
P(d|x) &= \sum_g P(d|g)P(g|x) && (38.82)\\
&= g(x)d + [1-g(x)](1-d) && (38.83)\\
&= P(d=1|x)d + [1-P(d=1|x)](1-d) && (38.84)\\
&= P(d|x) && (38.85)
\end{aligned}$$

We would like to do **propensity score strata-matching** by matching g-strata instead of x-strata. PO calculations for x-strata matching use the TPMs for $P(d|x)$, $P(x)$ and $P(y|d,x)$. To do g-strata matching using the same equations, but with x replaced by g, we would need to solve for $P(d|g)$, $P(g)$ and $P(y|d,g)$ in terms of $P(d|x)$, $P(x)$ and $P(y|d,x)$. We solve for those next.

From the TPMs for G_{ps}, one has

$$\boxed{P(d|g) = gd + (1-g)(1-d)} \tag{38.86}$$

and

$$\boxed{P(g) = \sum_x \overbrace{\delta(g, g(x))}^{P(g|x)} P(x)}\ . \tag{38.87}$$

Next, note that

$$P(y|d,g) = \sum_x P(y|d,x)P(x|g) \tag{38.88}$$

so we need to find $P(x|g)$. Since

$$\begin{aligned}
P(x|g) &= \frac{P(g|x)P(x)}{P(g)} && (38.89)\\
&= \frac{\delta(g,g(x))P(x)}{P(g)} && (38.90)
\end{aligned}$$

we finally get

$$\boxed{P(y|d,g) = \sum_x P(y|d,x)\frac{\delta(g,g(x))P(x)}{P(g)}}\ . \tag{38.91}$$

Chapter 41

Regression Discontinuity Design

This chapter is based on Ref.[3].

This chapter assumes that the reader has read Chapter 38 on Potential Outcomes.

In Regression Discontinuity Design (RDD), one switches the treatment dose $\underline{d}$ from 0 when $\underline{x} < \xi$ to 1 where $\underline{x} > \xi$, where $\underline{x}$ is an observed confounder (call it the **switch confounder**) and ξ is a threshold value for $\underline{x}$. One measures the jump δ in the treatment outcome $\underline{y}$ as $\underline{x}$ passes through $\underline{x} = \xi$. Then one makes the very reasonable assumption that δ equals[1] $\mathcal{Y}_{1|x=\xi} - \mathcal{Y}_{0|x=\xi} = ATE_{|x=\xi}$ for an imaginary experiment in which the confounder $\underline{x}$ acts as a normal confounder that doesn't switch the treatment dose $\underline{d}$.

For example, d^σ might be whether an individual is admitted to Harvard Univ., y^σ might be how much money the individual earns for the first 20 years after graduating from Harvard, and x^σ might be his SAT scores. We assume Harvard only admits students with an SAT score higher than ξ.

41.1 PO analysis

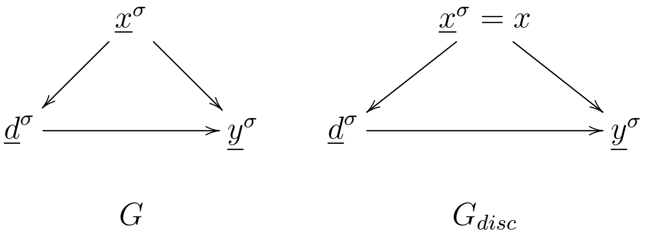


Figure 41.1: 2 bnets used in the PO analysis of RDD. The TPMs for G_{disc} are defined in terms of the TPMs for G. The TPM $P(d^\sigma|x^\sigma)$ for G_{disc} is discontinuous in x^σ.

The TPMs, printed in blue, for the bnet G_{disc} shown in Fig.41.1, are as follows. Note that

[1] ATE, which stands for "average treatment effect", is defined in Chapter 38.

the TPMs for the bnet G_{disc} are defined in terms of the TPMs for the bnet G.

$$P(x^\sigma) = \delta(x^\sigma, x) \tag{41.1}$$

$$P(y^\sigma|d^\sigma, x^\sigma = x) = P_{\underline{y}|\underline{d},\underline{x}}(y^\sigma|d^\sigma, x) \tag{41.2}$$

$$P(d^\sigma|x^\sigma = x) = \left\{ \begin{array}{ll} P_{\underline{d}|\underline{x}}(d^\sigma|x^\sigma = x) & \text{for } x > \xi \\ \delta(d^\sigma, 0) & \text{for } x < \xi \end{array} \right. \tag{41.3}$$

Define

$$E_{\sigma|d,x}[y^\sigma] = E_{y|d,x}[\underline{y}(d,x)] = \mathcal{Y}_{d|x} \tag{41.4}$$

and

$$\xi\pm = \xi \pm \epsilon \tag{41.5}$$

for some infinitesimal $\epsilon > 0$.

See Fig.41.2. In RDD, we assume that if we define the following 2 δ's, one for bnet G and the other for bnet G_{disc}, then the two δ's are equal, and they equal a conditional ATE.

$$\delta_{G_{disc}} = \mathcal{Y}_{1|x=\xi+} - \mathcal{Y}_{0|x=\xi-} \tag{41.6}$$

$$\delta_G = \mathcal{Y}_{1|x=\xi} - \mathcal{Y}_{0|x=\xi} \tag{41.7}$$

$$\delta_G = \delta_{G_{disc}} = \delta \tag{41.8}$$

$$\delta = ATE_{|x=\xi} \tag{41.9}$$

41.2 Linear Regression

In this section, we show how to apply linear regression (LR) to the PO analysis of RDD.

y^σ can be fitted as a function of $x \in \mathbb{R}$, for $d^\sigma \in \{0, 1\}$, as follows. Here ϵ^σ is the residual for individual σ and $b_0, m_0, b_1, m_1 \in \mathbb{R}$ are the fit parameters.

$$y^\sigma = [b_0 + m_0(x - \xi)](1 - d^\sigma) + [b_1 + m_1(x - \xi)]d^\sigma + \epsilon^\sigma\ . \tag{41.10}$$

Note that Eq.(41.10) yields a straight line in the $y^\sigma - x$ plane for $d^\sigma = 0$, and another straight line for $d^\sigma = 1$. These 2 lines are colored magenta in Fig.41.2. We are using the standard symbols b to denote the y-intercept, and m to denote the slope of a straight line.

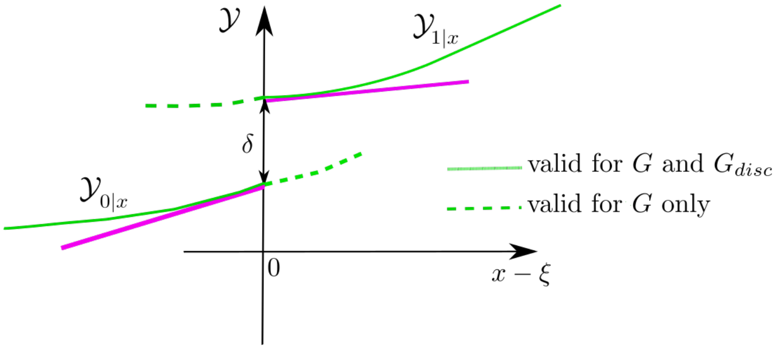


Figure 41.2: The jump δ between $\mathcal{Y}_{1|x}$ and $\mathcal{Y}_{0|x}$ is the same for G and G_{disc}.

Taking the expected value of Eq.(41.10), we get

$$\mathcal{Y}_{d|x} = [b_0 + m_0(x-\xi)](1-d) + [b_1 + m_1(x-\xi)]d \;. \tag{41.11}$$

Hence,

$$\mathcal{Y}_{1|x=\xi+} = b_1 \;,\;\; \mathcal{Y}_{0|x=\xi-} = b_0 \tag{41.12}$$

$$\delta = b_1 - b_0 \tag{41.13}$$

Bibliography

[1] Dan Bendel. Metropolis-Hastings: A comprehensive overview and proof. https://similarweb.engineering/mcmc/.

[2] Alexandra M Carvalho. Scoring functions for learning Bayesian networks. http://www.lx.it.pt/~asmc/pub/talks/09-TA/ta_pres.pdf.

[3] Scott Cunningham. *Causal inference: The mixtape*. Yale University Press, 2021. https://mixtape.scunning.com/index.html.

[4] Robin J. Evans. Graphical methods for inequality constraints in marginalized DAGs. https://arxiv.org/abs/1209.2978.

[5] Charles Fox, Neil Girdhar, and Kevin Gurney. A causal Bayesian network view of reinforcement learning. https://www.aaai.org/Papers/FLAIRS/2008/FLAIRS08-030.pdf".

[6] Bruno Gonçalves. Model testing and causal search. blog post https://medium.com/data-for-science/causal-inference-part-vii-model-testing-and-causal-search-536b796f

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, David Warde-Farley Bing Xu, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. https://arxiv.org/abs/1406.2661.

[8] Christina Heinze-Deml. Causality, spring semester 2019 at ETH Zurich. https://stat.ethz.ch/lectures/ss19/causality.php#course_materials.

[9] Cecil Huang and Adnan Darwiche. Inference in belief networks: A procedural guide. *International journal of approximate reasoning*, 15(3):225–263, 1996. http://www.ar-tiste.com/Huang-Darwiche1996.pdf.

[10] Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988. http://www.eecis.udel.edu/~shatkay/Course/papers/Lauritzen1988.pdf.

[11] Sergey Levine. Course CS 285 at UC Berkeley, Deep reinforcement learning. http://rail.eecs.berkeley.edu/deeprlcourse/.

[12] Dimitris Margaritis. Learning Bayesian network model structure from data (thesis, 2003, Carnegie Mellon Univ). https://apps.dtic.mil/sti/citations/ADA461103.

[13] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, page S7. Springer, 2006. https://link.springer.com/article/10.1186/1471-2105-7-S1-S7.

[14] Robert J. McEliece, David J. C. MacKay, and Jung-Fu Cheng. Turbo decoding as an instance of Pearls belief propagation algorithm. http://authors.library.caltech.edu/6938/1/MCEieeejstc98.pdf.

[15] Richard E Neapolitan. *Learning Bayesian networks*. Pearson Prentice Hall, 2004.

[16] Andrew Ng. Lecture at deeplearning.ai on recurrent neural networks. http://www.ar-tiste.com/ng-lec-rnn.pdf.

[17] Gregory Nuel. Tutorial on exact belief propagation in Bayesian networks: from messages to algorithms. https://arxiv.org/abs/1201.4724.

[18] Judea Pearl. Mediating instrumental variables. https://ftp.cs.ucla.edu/pub/stat_ser/r210.pdf.

[19] Judea Pearl. On the testability of causal models with latent and instrumental variables. https://arxiv.org/abs/1302.4976.

[20] Judea Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. https://www.aaai.org/Papers/AAAI/1982/AAAI82-032.pdf, 1982.

[21] Judea Pearl. *Probabilistic Inference in Intelligent Systems*. Morgan Kaufmann, 1988.

[22] Judea Pearl. *Causality: Models, Reasoning, and Inference, Second Edition*. Cambridge University Press, 2013.

[23] Judea Pearl. Causal and counterfactual inference. *The Handbook of Rationality*, pages 1–41, 2019. https://ftp.cs.ucla.edu/pub/stat_ser/r485.pdf.

[24] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[25] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[26] ReliaSoft. System analysis reference. http://reliawiki.org/index.php/System_Analysis_Reference.

[27] Marco Scutari. bnlearn. https://www.bnlearn.com/.

[28] Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, 115:235–253, 2019. https://arxiv.org/abs/1805.11908.

[29] Nitish Srivastava, G E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. http://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf.

[30] Masayoshi Takahashi. Statistical inference in missing data by MCMC and non-MCMC multiple imputation algorithms: Assessing the effects of between-imputation iterations. *Data Science Journal*, 16, 2017. https://datascience.codata.org/articles/10.5334/dsj-2017-037/.

[31] theinvestorsbook.com. Pert analysis. https://theinvestorsbook.com/pert-analysis.html.

[32] Robert R. Tucci. Bell's inequalities for Bayesian statisticians. blog post in blog Quantum Bayesian Networks, https://qbnets.wordpress.com/2008/09/19/bells-inequaties-for-bayesian-statistician/.

[33] Robert R. Tucci. Quantum Fog. https://github.com/artiste-qb-net/quantum-fog.

[34] W. E. Vesely, F. F. Goldberg, N. H. Roberts, and D. F. Haasl. Fault tree handbook nureg-0492. https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0492/.

[35] Wikipedia. Belief propagation. https://en.wikipedia.org/wiki/Belief_propagation.

[36] Wikipedia. Beta function. https://en.wikipedia.org/wiki/Beta_function.

[37] Wikipedia. Binary decision diagram. https://en.wikipedia.org/wiki/Binary_decision_diagram.

[38] Wikipedia. Boolean algebra. https://en.wikipedia.org/wiki/Boolean_algebra.

[39] Wikipedia. Categorical distribution. https://en.wikipedia.org/wiki/Categorical_distribution.

[40] Wikipedia. Chow-Liu tree. https://en.wikipedia.org/wiki/Chow%E2%80%93Liu_tree.

[41] Wikipedia. Data processing inequality. https://en.wikipedia.org/wiki/Data_processing_inequality.

[42] Wikipedia. Dirichlet distribution. https://en.wikipedia.org/wiki/Dirichlet_distribution.

[43] Wikipedia. Errors in variables models. https://en.wikipedia.org/wiki/Errors-in-variables_models.

[44] Wikipedia. Expectation maximization. https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm.

[45] Wikipedia. Gamma function. https://en.wikipedia.org/wiki/Gamma_function.

[46] Wikipedia. Gated recurrent unit. https://en.wikipedia.org/wiki/Gated_recurrent_unit.

[47] Wikipedia. Gibbs sampling. https://en.wikipedia.org/wiki/Gibbs_sampling.

[48] Wikipedia. Hidden Markov model. https://en.wikipedia.org/wiki/Hidden_Markov_model.

[49] Wikipedia. Importance sampling. https://en.wikipedia.org/wiki/Importance_sampling.

[50] Wikipedia. Instrumental variables estimation. https://en.wikipedia.org/wiki/Instrumental_variables_estimation.

[51] Wikipedia. Inverse transform sampling. https://en.wikipedia.org/wiki/Inverse_transform_sampling.

[52] Wikipedia. Junction tree algorithm. https://en.wikipedia.org/wiki/Junction_tree_algorithm.

[53] Wikipedia. k-means clustering. https://en.wikipedia.org/wiki/K-means_clustering.

[54] Wikipedia. Kalman filter. https://en.wikipedia.org/wiki/Kalman_filter.

[55] Wikipedia. Least squares. https://en.wikipedia.org/wiki/Least_squares.

[56] Wikipedia. Linear regression. https://en.wikipedia.org/wiki/Linear_regression.

[57] Wikipedia. Long short term memory. https://en.wikipedia.org/wiki/Long_short-term_memory.

[58] Wikipedia. Markov blanket. https://en.wikipedia.org/wiki/Markov_blanket.

[59] Wikipedia. Metropolis-Hastings method. https://en.wikipedia.org/wiki/Metropolis%E2%80%93Hastings_algorithm.

[60] Wikipedia. Minimum spanning tree. https://en.wikipedia.org/wiki/Minimum_spanning_tree.

[61] Wikipedia. Monte Carlo methods. https://en.wikipedia.org/wiki/Category:Monte_Carlo_methods.

[62] Wikipedia. Multinomial distribution. https://en.wikipedia.org/wiki/Multinomial_distribution.

[63] Wikipedia. Multinomial theorem. https://en.wikipedia.org/wiki/Multinomial_theorem.

[64] Wikipedia. Multivariate normal distribution. https://en.wikipedia.org/wiki/Multivariate_normal_distribution.

[65] Wikipedia. Natural experiment. https://en.wikipedia.org/wiki/Natural_experiment.

[66] Wikipedia. Non-negative matrix factorization. https://en.wikipedia.org/wiki/Non-negative_matrix_factorization.

[67] Wikipedia. Ordinary least squares. https://en.wikipedia.org/wiki/Ordinary_least_squares.

[68] Wikipedia. Program evaluation and review technique. https://en.wikipedia.org/wiki/Program_evaluation_and_review_technique.

[69] Wikipedia. Rejection sampling. https://en.wikipedia.org/wiki/Rejection_sampling.

[70] Wikipedia. Simple linear regression. https://en.wikipedia.org/wiki/Simple_linear_regression.

[71] Wikipedia. Simpson's paradox. https://en.wikipedia.org/wiki/Simpson's_paradox.

[72] Wikipedia. Spring system. https://en.wikipedia.org/wiki/Spring_system.

[73] Wikipedia. Variational Bayesian methods. https://en.wikipedia.org/wiki/Variational_Bayesian_methods.

[74] Hao Wu and Zhaohui Steve Qin. course notes, BIOS731: Advanced statistical computing, 2016 Emory Univ. http://web1.sph.emory.edu/users/hwu30/teaching/statcomp/statcomp.html.