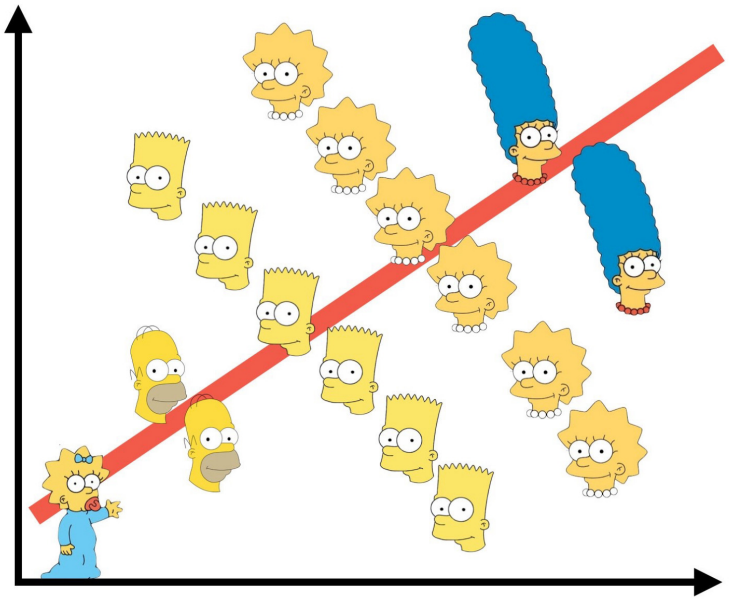
Causal Inference

By Hedibert Lopes

October 2024



The Simpsons Paradox

These notes are taken from various sources, mainly Pearl, Glymour and Jewell's Causal Inference in Statistics: Chapters 1 and 2, and Daniels, Linero and Roy's Bayesian Nonparametrics for Causal Inference and Missing Dara: Chapter 1.

Causal Inference

Contents

1 Readings in Statistics and Econometrics 2015: Causality

• In the Fall semester of 2015 (almost a decade ago!) I organized a readings in statistics seminars on causality.

• The list of presenters and papers can be found here:

`https://hedibert.org/previous-teaching/`

• In the following link you will find links to textbooks and edited books, special issues, articles with discussion and web material: slides of lectures, discussion of causality, video lectures and more (in chronological order):

`http://hedibert.org/wp-content/uploads/2015/10/annotatedbibliography.pdf`

• In the following link you will find only articles and book chapters (in alphabetical order).

`http://hedibert.org/wp-content/uploads/2015/10/annotatedbibliography-articles.pdf`

• I reproduce below the **Outline of the lectures** (all talks have slides, but the 7th)

1. September 29th, 2015 – Hedibert Lopes, Insper
   Haavelmo (1943) The statistical implications of a system of simultaneous equations. *Econometrica*, 11, 1-12.

2. October 6th, 2015 – Hedibert Lopes, Insper
   Rubin (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology,* 56, 688-701.

3. October 13th, 2015 – André Yoshizumi, IME/USP
   Holland (1986) Statistics and causal inference (with discussion). *JASA*, 81, 945-970.

4. October 20th, 2015 – Paloma Uribe, IME/USP
   Pearl (1995) Causal diagrams for empirical research (with discussion). *Biometrika*, 82, 669-710.

5. November 3rd, 2015 – Sergio Firpo, EESP/FGV
   Angrist, Imbens and Rubin (1996) Identification of causal effects using IVs (with discussion). *JASA*, 91, 444-472.

6. November 10th – Julio Trecenti, IME/USP
   Dawid (2000) Causal inference without counterfactuals (with discussion). *JASA*, 95, 407-424.

7. November 24th, 2015 – Manasses Nóbrega, UFABC
   Vansteelandt and Goetghebeur (2003) Causal inference with generalized structural mean models. *JRSS-B*, 65, 817-835.

8. December 1st, 2015 – Hedibert Lopes, Insper
   Heckman and Pinto (2015) Causal analysis after Haavelmo. *Econometric Theory*, 31,115-151.

2 Simpson's Paradox

"Named after Edward Simpson (born 1922), the statistician who first popularized it, the paradox refers to the existence e of data in which a statistical association that holds for an entire population is reserved in every subpopulation."

2.1 Example 1

"We record the recovery rates of 700 patients (343 women and 357 men) who were given access to the drug. A Total of 350 patients chose to take the drug and 350 patients did not."

	Drug	No drug
Patients	273 out of 350 - 78%	289 out of 350 - 83%

Total and percentage of recovered.

Question: Based on this data, should a doctor recommend the drug or not?

2.2 Example 2

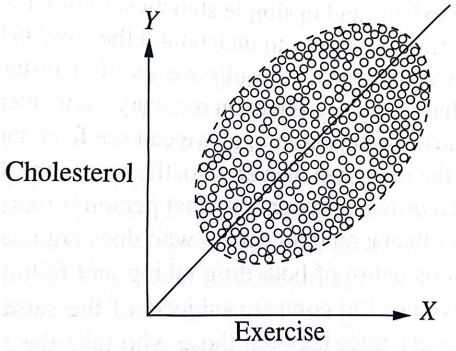
	Drug	No drug
Low BP	81 out of 87 - 93%	234 out of 270 - 87%
High BP	192 out of 263 - 73%	55 out of 80 - 69%

Total and percentage of recovered.

Question: Based on this data, should a doctor recommend the drug or not?

2.3 Example 3

The more a person exercises, the higher their cholesterol is!



3 Causal inference

3.1 Why study causation

- "We study causation because we need to make sense of data, to guide actions and policies, and to learn from our sucess and failures."
- "We need to estimate the effect of

  i) Smoking on lung cancer;

  ii) Education on salaries;

  iii) Carbon emissions on the climate."
- "We need to understand HOW and WHY causes influence effects"

3.2 Back to example 1

"We record the recovery rates of 700 patients (343 women and 357 men) who were given access to the drug. A Total of 350 patients chose to take the drug and 350 patients did not."

	Drug	No drug
Men	81 out of 87 - 93%	234 out of 270 - 87%
Women	192 out of 263 - 73%	55 out of 80 - 69%
Combined	273 out of 350 - 78%	289 out of 350 - 83%

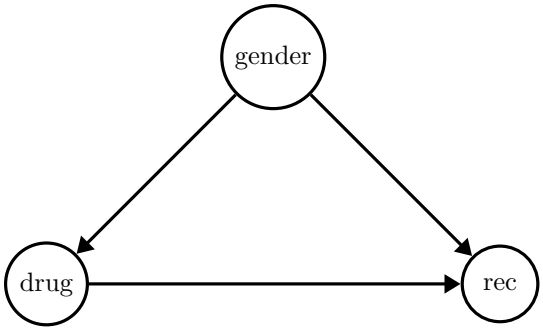
Total and percentage of recovered.

"The data seem to say that if we know the patient's gender – male or female – we can prescribe the drug, but if the gender is unknown we should not! Obviously, that conclusion is ridiculous."

"Should a doctor prescribe the drug for a woman? A man? A patient of unknown gender? Or consider a policy maker who is evaluating the drug's overall effectiveness on the population. Should he/she use the recovery rate for the general population? Or should he/she use the recovery rates for the gendered subpopulations?"

"The answer is nowhere to the found in simple statistics. In order to decide whether the drug will harm or help a patient, we first have to understand the story behind the data – the causal mechanism that let to, or generated, the results we see."

"Suppose we knew an additional fact: Estrogen has a negative effect on recovery, so women are less likely to recover than men, regardless of the drug. In addition, as we can see from the data, women are significantly more likely to take the drug than men are."



3.3 Back to example 2

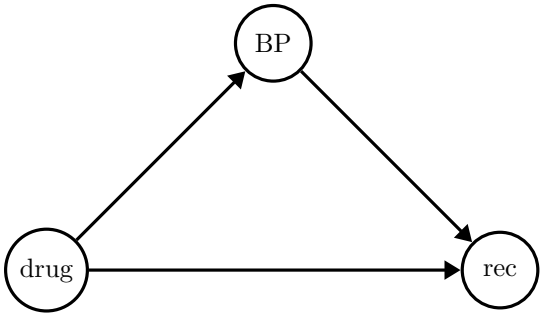
"Suppose we looked at the same numbers from our first example of drug taking to recovery, instead of recording participants' gender, patient's blood pressure (BP) were recorded at the end of the experiment. In this case, we know that the drug affects recovery by lowering the BP of those who take it – but unfortunately, it also has a toxic effect."

	Drug	No drug
Low BP	81 out of 87 - 93%	234 out of 270 - 87%
High BP	192 out of 263 - 73%	55 out of 80 - 69%
Combined	273 out of 350 - 78%	289 out of 350 - 83%

Total and percentage of recovered.

"In the general population, the drug might improve recovery rates because of its effect on the BP. But in the subpopulations – the group of people whose posttreatment BP is high and the group whose posttreatment BP is low – we, of course, would not see that effect; we would only see the drug's toxic effect."

"Remarkably, though the numbers are the same in the gender and blood pressure examples, the correct result lies in the segregated data for the former and the aggregated data for the latter."



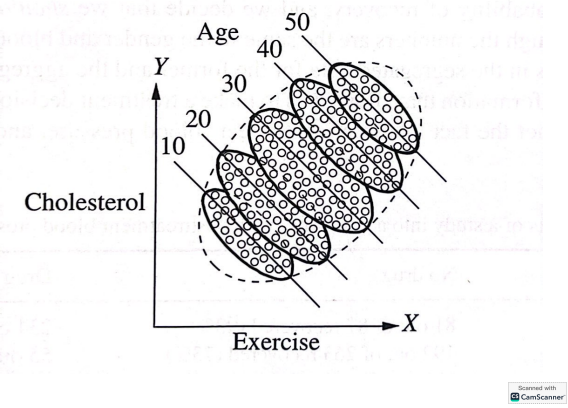
3.4 Back to example 3

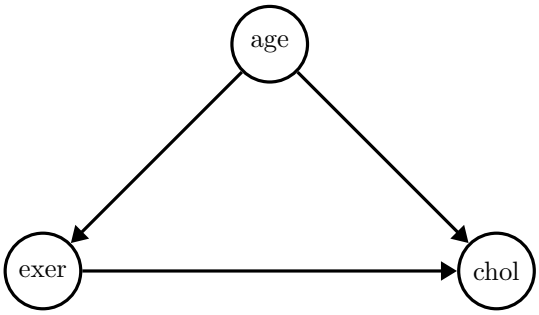
"Consider a study that measures weekly exercise and cholesterol in various age groups. When we plot exercise on the X-axis and cholesterol on the Y-axis and seggragate by age, we see that there is a general trend downward in each group; the more young people exercise, the lower their cholesterol is, and the sample applies for middle-aged people and elderly."

"If, however, we use the sample scatter plot, but we don't segregate by gender, we see a general trend upward; the more a person exercises, the higher their cholesterol is."

"If we know that that older people, who are more likely to exercise, are also more likely to have high cholesterol regardless of exercise, then the reversal is easily explained, and easily resolved."

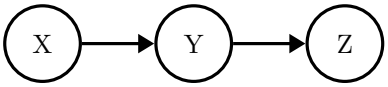
"Age is a common cause of both treatment (exercise) and outcome (cholesterol)."





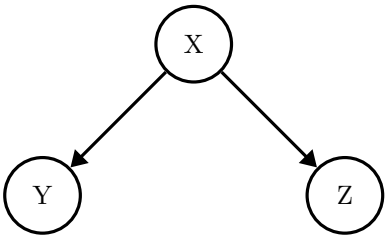
4 Chains, Forks and Colliders

4.1 Chains



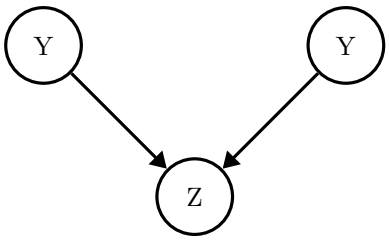
Rule 1: Conditional independence chains: X and Z are conditionally independent given Y, if there is only one unidirectional path between X and Z and Y is any set of variables that intercepts that path.

4.2 Forks



Rule 2: Conditional independence forks: If X is a common cause of Y and Z, and there is only one path between Y and Z, then Y and Z are independent conditional on X.

4.3 Colliders

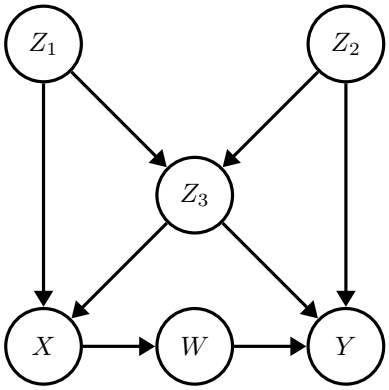


Rule 3: Conditional independence of colliders: If Z is a collision between X and Y and there is only one path between X and Y, then X and Y are unconditionally independent but are conditionally dependent given Z and any descendants of Z.

5 More on chains, forks and colliders

Graham Harrison, Towards Data Science, Jan 31, 2024. Explaining junctions using correlation, independence and regression to understand their critical importance in causal inference. `https://search.app/xNWjnZeghtJ1Jayx8`

"I have constructed the fictitious DAG below because it is sufficiently simple to effectively explore the concepts and sufficiently complex to contain all 3 types of junctions: chains, forks and colliders."



5.1 Paths and junctions

"Paths always start at the treatment (X), always end at the outcome (Y) and are acyclic (i.e. they do not loop back). A junction has exactly 3 nodes and 2 connections. The 5 paths visualised above can be expressed in this form as follows:"

1) $X \rightarrow W \rightarrow Y$ One junction.
2) $X \leftarrow Z1 \rightarrow Z3 \rightarrow Y$ Two junction.
3) $X \leftarrow Z1 \rightarrow Z3 < -Z2 \rightarrow Y$ Three junction.
4) $X \leftarrow Z3 \rightarrow Y$
5) $X \leftarrow Z3 \leftarrow Z2 \rightarrow Y$

All junctions from the above DAG:

1) $X \rightarrow W \rightarrow Y$

2) $Z_1 \rightarrow X \rightarrow W$

3) $Z_3 \rightarrow X \rightarrow W$

4) $Z_1 \rightarrow Z_3 \rightarrow X$

5) $Z_1 \rightarrow Z_3 \rightarrow Y$

6) $Z_2 \rightarrow Z_3 \rightarrow X$

7) $Z_2 \rightarrow Z_3 \rightarrow Y$

8) $X \leftarrow Z_1 \rightarrow Z_3$

9) $X \leftarrow Z_3 \rightarrow Y$

10) $Y \leftarrow Z_2 \rightarrow Z_3$

11) $W \rightarrow Y \leftarrow Z_2$

12) $W \rightarrow Y \leftarrow Z_3$

13) $Z_2 \rightarrow Y \leftarrow Z_3$

14) $Z_1 \rightarrow X \leftarrow Z_3$

15) $Z_1 \rightarrow Z_3 \leftarrow Z_2$

"It should be apparent from a quick review of all junctions within our DAG that there are just 3 possible patterns or types.

Chain: Junction 1 ($X \rightarrow W \rightarrow Y$) is an example of a **chain** where the first node points to the intermediary ($X \rightarrow W$) and the intermediary "points" to the final node ($W \rightarrow Y$). Junctions 2 to 7 are also chains.

Fork: Junction 8 ($X \leftarrow Z_1 \rightarrow Z_3$) is an example of a **fork** where the intermediary node points to both the first node ($X \leftarrow Z_1$) and the final node ($Z_1 \rightarrow Z_3$). Junctions 9 and 10 are also forks.

Collider: Junction 11 ($W \rightarrow Y \leftarrow Z_2$) is an example of a **collider** where the first node points to the intermediary ($W \rightarrow Y$) and the final node also points to the intermediary ($Y \leftarrow Z_2$). Junctions 12 to 15 are also colliders.

5.2 Causal Discovery and Causal Validation

"**Causal Discovery** is the concept of automatically generating a DAG from the data and **Causal Validation** is the process of testing a proposed DAG against a dataset. It is typically possible for more than one DAG to satisfy the causal validation tests against a given dataset, hence these approaches are complex and uncertain. "

5.3 Simulating a dataset from a DAG

$$
\begin{aligned}
Z_1 &\sim N(4.75, 1.72^2) \quad \text{and} \quad Z_2 \sim N(3.29, 1.89^2) \\
Z_3 &= 3Z_1 - 1.5Z_2 + \epsilon_{z_3} \\
X &= 2Z_1 + 2.5Z_3 + \epsilon_x \\
W &= 3X + \epsilon_w \\
Y &= 2W + 2Z_2 - 3Z_3 + \epsilon_y
\end{aligned}
$$

R code

```
set.seed(4321)
n = 100
z1 = rnorm(n,4.75,1.72)
z2 = rnorm(n,3.29,1.89)
z3 = 3*z1 -1.5*z2
x = 2*z1 + 2.5*z3
w = 3*x
y  = 2*w + 2*z2 -3*z3
sig.z3 = sqrt(var(z3))
sig.x  = sqrt(var(x))
sig.w = sqrt(var(w))
sig.y = sqrt(var(y))
z3 = z3 + rnorm(n,0,sig.z3)
x = x + rnorm(n,0,sig.x)
w = w + rnorm(n,0,sig.w)
y = y + rnorm(n,0,sig.y)
cbind(z1,z2,w,x,y)
```

6 d-separation (directional)

- *d*-connected: there exists a connecting path between them.
- *d*-separated: there exists no such path.
- If we are not conditioning on any variable, then only colliders can block a path.

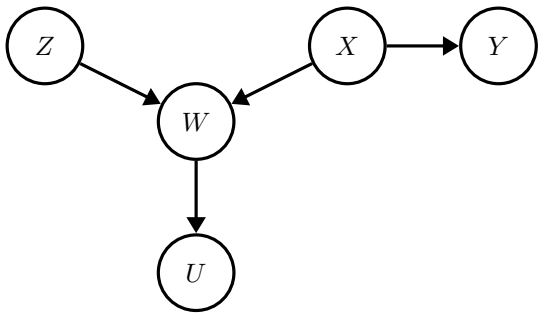
(d-separation) A path p is blocked by a set of nodes Z if and only if

1. p contains a chain of nodes $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that the middle node B is in Z (i.e., B is conditioned on), or
2. p contains a collider $A \to B \leftarrow C$ such that the collision node B is not in Z, and no descendants of B is in Z.

If Z blocks every path between X and Y, then X and Y are d-separated, conditional on Z, and thus are independent conditional on Z.

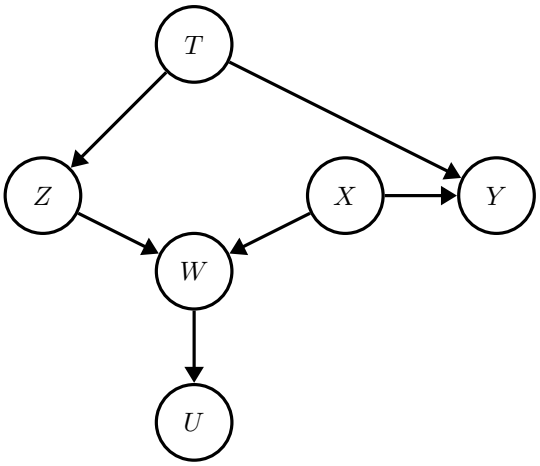
6.1 Example

The graphical model below contains a collider with child and a fork.



- Z and Y are d-separated, so $Z \perp\!\!\!\perp Y$.
- Conditioning on W (collider) "unblocks" Z and Y, so Z and Y are conditionally dependent.
- Conditioning on U (descendant of a collider) "unblocks" Z and Y, so Z and Y are conditionally dependent.
- Conditioning on $\{W, X\}$ "blocks" Z and Y, so Z and Y are conditionally independent.

The graphical model below contains an additional forked path between Z and Y.



- $Z \not\perp\!\!\!\perp Y$, since path between Z and Y with no colliders.
- $Z \perp\!\!\!\perp Y|T$
- $Z \not\perp\!\!\!\perp Y|T, W$
- $Z \perp\!\!\!\perp Y|T, W, X$
- Z and Y are d-connected conditional on $W, U, \{W, U\}, \{W, T\}, \{U, T\}, \{W, U, T\}, \{W, X\}, \{U, X\}, \{W, U, X\}$.
- Z and Y are d-separated conditional on $T, \{X, T\}, \{W, X, T\}, \{U, X, T\}, \{W, U, X, T\}$.

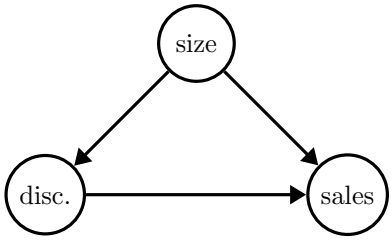
7 Examples

7.1 Estimating the effect of a marketing campaign

This example is taken from Heinrich Kögel's Medium article *Causal Machine Learning in Marketing*, from July 31st, 2023. `https://medium.com/@heinrichkoegel/causal-machine-learning-in-marketing-12dcd91ec24e`

1. The company sells computer equipment to other businesses.
2. Marketing campaign offering discounts to certain customer firms.
3. Marketing managers: should we continue providing these discounts?
4. Comparing average sales from firms that received or not the discount.
5. Larger firms (more employees) were more likely to receive the discount.
6. The company has higher sales with larger firms.

The following graph summarizes the dependences between the variables.



Effect of receiving the discount on sales: disc → sales.
Firm size influences the likelihood of receiving the discount: size → disc.
Larger firms tend to result in higher sales: size → sales.

Discount is the *treatment*, sales is the *outcome*, and firm size is the *confounding variable* is associated with both discount and sales. If we do not account for the influence of this confounding variable when estimating the effect of the discount, we will obtain an incorrect estimate. Confounding variables lead to false conclusions in naive estimations. To isolate the causal effect, we need to **control for** or **hold constant** the confounding variables in our estimation.

7.2 Comparing two classes of medication for hypertension

The figure below is the beginning of Section 1.1, page 4, of the 2024 book *Bayesian Nonparametrics for Causal Inference and Missing Data*, by Michael Daniels, Antonio Linero and Jason Roy. Also, the reference [8] in the text is the 2015 book *Causal Inference for Statistics, Social, and Biomedical Sciences*, by Guido Imbens and Donald Rubin. The reference [14] in the text is the 1974 paper *Estimating causal effects of treatments in randomized and nonrandomized studies*, by Donald Rubin, that appeared in the volume 66, number 5, pages 688, of "Journal of Educational Psychology".

1.1 Introduction

Often a goal of researchers is to learn about the causal effect of exposures, treatments, or policies (referred to generically as "treatments") on outcomes of interest: for example, does smoking cause cancer? Or does exercising reduce depression? A key distinction between causal inference and standard statistical inference is that causal inference requires assumptions that cannot be checked from the data.

Example 1.1.1. Consider an observational study comparing two classes of medication for treating hypertension: angiotensin-converting enzyme inhibitor (ACEIs) and angiotensin II receptor blockers (ARBs). Suppose we are interested in systolic blood pressure (SBP) 3 months after the start of treatment (Y). From the data, using standard statistical methods, we could learn about the average value of Y for people who took ACEIs or for people who took ARBs. We could also learn about conditional averages, such as the average SBP among men, age 55, who took ACEIs. However, additional assumptions are needed to compare the average SBP in the hypothetical world where *everyone* in the population had been prescribed an ACEI to the average SBP in the world where everyone had been prescribed an ARB; this comparison requires assumptions that can only be assessed via subject matter considerations and that cannot be checked from data.

The fundamental problem in the above example is that we are attempting to draw conclusions about quantities that we did not observe: we cannot know what would have happened to someone who was prescribed an ACEI had they taken an ARB instead because physicians may systematically prescribe ACEIs to men with more severe symptoms or vice versa. One way to identify the effect of interest is to assume that we have measured and controlled for all *confounders*, i.e., variables that influence both the assigned treatment (medication here) and SBP, but it is important to recognize that this assumption cannot be tested strictly from the data itself.

In this chapter we review the *potential outcomes* framework [14] for causal inference, which will be used throughout. We will introduce the framework, use it to define commonly used causal estimands, and describe common estimation procedures. A more in-depth treatment of the potential outcomes framework itself can be found in [8].

8 Potential outcome, ATE, QTE, CATE, STUVA

"Causal inference requires assumptions that cannot be checked from the data."

Suppose that A is a binary treatment, then

$$A = \begin{cases} 1 & \text{if unit is treated} \\ 0 & \text{if unit is untreated} \end{cases}.$$

In the previous example, $A = 1$ if prescribed ACEIs (treated) and $A = 0$ if prescribed ARBs (untreated). The outcome Y is the systolic blood pressure (SBP). Recall that ACEI stands for *angiotensin-converting enzyme inhibitor* and ARB stands for *angiotensin II receptor blockers*.

- $Y(a)$ is a **potential outcome**, i.e. the outcome that would be observed if the individual received treatment $a \in \{0, 1\}$.

- Potential outcomes can be used to derive causal estimands. For instance, the **average treatment effect (ATE)** is defined as
$$E\{Y(1)\} - E\{Y(0)\},$$
which, for the example, is "a population-level parameter and can be interpreted as the answer to the question, *how much higher would the average SBP be in the population if everyone had been treated with ACEIs compared to if everyone had been treated with ARBs?*

- The **average treatment effect on the treated (ATT)** is defined as
$$E\{Y(1)|A = 1\} - E\{Y(0)|A = 1\},$$
which is "a contrast between the average outcome under treatment and under no treatment within the subpopulation of treated individuals. For treated individuals, we can think of $Y(0)$ as the counterfactual outcome – the outcome that would have been observed had the subject, contrary to fact, not been treated.

- The **quantile treatment effect (QTE)** can be defined as
$$F_1^{-1}(p) - F_0^{-1}(p),$$
where $F_a(y) = Pr(Y(a) \leq y)$. "QTEs are particularly usefl when the outcome of interest is skewed, or when the treatment only has a large effect for a small subset of the population. Check Brantly Callaway's notes on the QTE at `https://cran.r-project.org/web/packages/qte/vignettes/R-QTEs.html`.

- For binary outcomes, one might define the **causal relative risk**
$$\frac{E\{Y(1)\}}{E\{Y(0)\}},$$
which can be read as "how many times more likely would high BPS be if everyone was treated with ACEIs rather than ARBs?"

- Computing ATE for subpopulations, the **conditional ATE (CATE)**:
$$E\{Y(1) - Y(0)|V = v\},$$
where V is a set of covariates of interest and $V = v$ defines a subpopulations.

8.1 Identifiability and causal assumptions

"For most studies, we observe the treatment received (A_i) and the outcome (Y_i) for N total subjects ($i = 1, \ldots, N$). To link observed data (no potential outcomes!) to potential outcomes, we need *causal assumptions*."

SUTVA: The *Stable Unit Treatment Value Assumption* (SUTVA) is said to hold when the potential outcome of any subject i does not depend on the treatment received by the other subjects. That is if $a = (a_1, \ldots, a_N)$ and $a' = (a'_1, ..., a'_N)$ are any two possible assignments of subjects to treatments such that $a_i = a'_i$ and $Y_i(a)$ is the potential outcome of subject i under a, we have

$$Y_i(a) = Y_i(a').$$

Consistency: The consistency assumption holds if $Y_i = Y_i(a)$, if $A_i = a$. That is, if subject i is observed to have received treatment a then their observed outcome is just their potential outcome for treatment a.

Randomized trials are the gold standard for establishing causation because randomizing the treatment assignment guarantees that $\{Y_i(0), Y_i(1)\}$ is independent of A, so

$$E\{Y_i(a)\} = E\{Y_i(a)|A_i = a\} = E(Y_i|A_i = a),$$

so that $E\{Y_i(a)\}$ can be identified in terms of the observable quantities (Y_i, A_i).

Ignorability: The ignorability assumption holds if

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid L,$$

where L is a set of observed covariates that influence both Y and A (remember when we defined a **fork**?). "The selection of confounders L is, at least in part, based on subject matter knowledge. These pre-treatment variables should be chosen to completely capture the association between the treatment assignment and the outcome. This assumption is sometimes refereed to as a *no unmeasured confounders* assumption, an *unconfoundedness* assumption, or as a *exchangeability* assumption."

"The validity of ignorability cannot be assessed from the observed data, regardless of how large the sample size is. This is because we can never rule out the possibility that an apparent association between the treatment and outcome is due to some variable that we happen to not have measured."

Positivity: The positivity assumption holds if

$$Pr(A = a|L = l) > 0 \quad \text{for all } a \text{ and } l.$$

Identifiability of the ATE: Combining consistency, positivity, and ignorability can be used to identify the ATE with binary treatments.

8.2 Propensity scores

The **propensity score** is defined as

$$e(l) = Pr(A = 1|L = l)$$

Proposition: Suppose that the SUTVA, consistency, ignorability, and positivity assumptions hold and that the treatment A_i is binary. Then the ATE can be written as

$$ATE = E\left\{\frac{AY}{e(L)} - \frac{(1-A)Y}{1-e(L)}\right\} = E\left\{Y\frac{A-e(L)}{e(L)\{1-e(L)\}}\right\}$$

"If we can estimate the propensity score sufficiently well, then the plug-in estimate

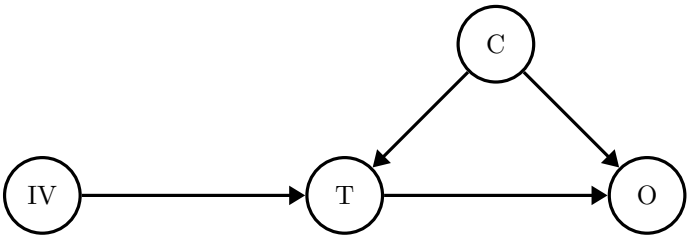
$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{A_i - \widehat{e}(L_i)}{\widehat{e}(L_i)\{1 - \widehat{e}(L_i)\}}$$

is consistent for the ATE."

"The propensity score plays a prominent role in causal inference. It has the property of being a *balancing scores*, which makes it a useful one-dimensional summary of L."

9 Instrumental variables

The following DAG illustrate the well-known instrumental variable solution to the endogeneity in the treatment/outcome scenario.



IV: Instrumental variable/Instrument
T: Treatment/Program/Policy/Risk Factors
C: Unmeasured Confounders/Measured Confounders/Confounding Factors
O: Outcome

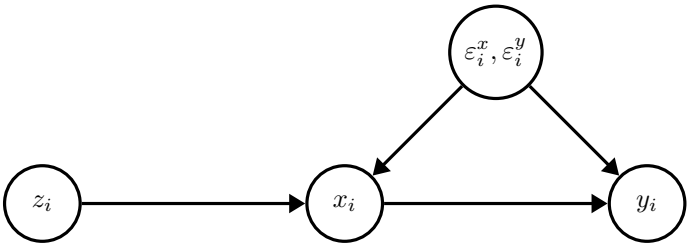
	Confounders (C)	Treatment (T)	Outcome (O)	IV
i	Maternal characteristics	Smoking during pregnancy	Low birth weight	Cigarettes taxes
ii	Smoking, caffeine, alcohol	body mass index	Parkinson disease	FTO gene variant
iii	Prognostic factors	Catheter use	Mortality	Patients with catheter at facility
iv	Ability	Education	Earnings	Mother's education
v	Proximity	Tutoring program	GPA	Library hours

9.1 Bivariate Gaussian linear regression

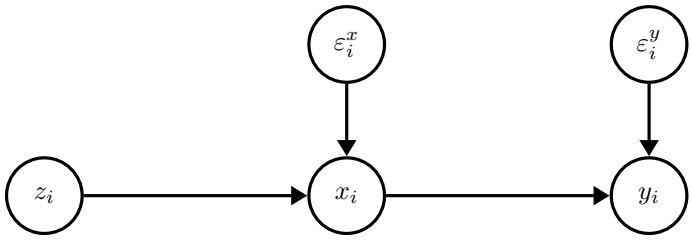
In the most popular IV problem, the goal is to measure the effect of treatment x on the outcome y:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i^y,$$

for $i = 1, \ldots, n$. However, there is dependence between x_i and ε_i^y as the following DAG illustrates. In words, the treatment variable x_i is *endogenous*, and both x_i and the outcome variable y_i are affected by unmeasured or measured confounders $(\varepsilon_i^x, \varepsilon_i^y)$



If the error terms (the unmeasured confounders) were uncorrelated, i.e. $cov(\varepsilon_i^x, \varepsilon_i^y) = 0$, then endogeneity disappears and β_1 is a *causal effect*. From the below DAG becomes obvious why. Notice that $z_i \to x_i \to y_i$ is a fork, so y_i and z_i are independent, conditionally on x_i.



Now, let us go back to the first DAG where $cov(\varepsilon_i^x, \varepsilon_i^y) \neq 0$. We have have a system with two equations:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i^y \\ x_i &= \delta_0 + \delta_1 z_i + \varepsilon_i^x, \end{aligned}$$

with $\varepsilon_i = (\varepsilon_i^x, \varepsilon_i^y)$ following a bivariate normal distribution with zero mean vector, variances σ_y^2 and σ_x^2 and covariance $\sigma_{xy} = \rho\sigma_y\sigma_x$. We can now characterize lack of endogeneity and the strength of the instrumental variables:

$$\begin{aligned} \rho = 0 &\Rightarrow \text{no endogeneity} \\ \delta = 0 &\Rightarrow \text{no instrument} \end{aligned}$$

To understand how one is able to learn the effect of x on y, let us replace x_i in the outcome equation by the treatment equation, i.e.

$$\begin{aligned} y_i &= \beta_0 + \beta_1[\delta_0 + \delta_1 z_i + \varepsilon_i^x] + \varepsilon_i^y \\ &= (\beta_0 + \beta_1\delta_0) + (\beta_1\delta_1) z_i + (\beta_1\varepsilon_i^x + \varepsilon_i^y) \\ &= \theta_0 + \theta_1 z_i + u_i, \end{aligned}$$

where u_i are iid $N(0, \tau^2)$, where $\tau^2 = \beta^2\sigma_x^2 + \sigma_y^2 + 2\rho\sigma_x\sigma_y$. We are now ready to obtain the (two-stage) estimator of β_1:

1. Obtain $\widehat{\delta}_1$ from the treatment equation (first stage)
2. Obtain $\widehat{\theta}_1$ from the outcome equation (2nd stage)
3. Since $\theta_1 = \beta_1\delta_1$, it follows that

$$\text{effect of } x \text{ on } y = \widehat{\beta}_1^{IV} = \frac{\widehat{\theta}_1}{\widehat{\delta}_1} = \frac{cov(z,y)}{cov(z,x)} = \frac{\text{effect of } z \text{ on } y}{\text{effect of } z \text{ on } x}$$

9.2 Regression of y on (x, z)

Since $(\varepsilon_i^x, \varepsilon_i^y)$ is bivariate normal with zero mean, variances σ_y^2 and σ_x^2 and covariance $\sigma_{xy} = \rho\sigma_y\sigma_x$, it can be easily shown that

$$\varepsilon_i^y | \varepsilon_i^x \sim N\left(\frac{\sigma_{xy}}{\sigma_x^2}\varepsilon_i^x, \sigma_y^2(1-\rho^2)\right),$$

or

$$\varepsilon_i^y = \frac{\sigma_{xy}}{\sigma_x^2}\varepsilon_i^x + \omega_i,$$

for $\omega_i \sim N(0, \sigma_y^2(1-\rho^2))$. Now, replacing ε_i^x by $x_i - \delta_0 - \delta_1 z_i$, if follows that

$$\begin{aligned}\varepsilon_i^y &= \frac{\sigma_{xy}}{\sigma_x^2}[x_i - \delta_0 - \delta_1 z_i] + \omega_i, \\ &= \frac{\sigma_{xy}}{\sigma_x^2}x_i - \frac{\sigma_{xy}}{\sigma_x^2}\delta_0 - \frac{\sigma_{xy}}{\sigma_x^2}\delta_1 z_i + \omega_i,\end{aligned}$$

so the outcome equation becomes

$$y_i = \beta_0 + \beta_1 x_i + \frac{\sigma_{xy}}{\sigma_x^2}x_i - \frac{\sigma_{xy}}{\sigma_x^2}\delta_0 - \frac{\sigma_{xy}}{\sigma_x^2}\delta_1 z_i + \omega_i,$$

or

$$y_i = \left(\beta_0 - \frac{\sigma_{xy}}{\sigma_x^2}\delta_0\right) + \left(\beta_1 + \frac{\sigma_{xy}}{\sigma_x^2}\right)x_i + \left(-\frac{\sigma_{xy}}{\sigma_x^2}\delta_1\right)z_i + \omega_i.$$

In words, by simply including the instrumental variable z_i in the outcome regression will not suffice to estimate β_1.

9.3 A few examples

A few examples can be found here, `http://hedibert.org/wp-content/uploads/2016/05/iv-workedexamples.pdf`, back in 2015 when I taught *Introduction to Econometrics* to Economics undergraduate students.

A) Simulated exercise: the omitted variable problem

B) Estimating the return to education for married women: 753 observations and 22 variables
Outcome: log(wage)
Treatment: Education in years
IV: Father's education in years
OLS: 11% return for another year of education
2SLS: 5.9% return for another year of education
OLS suffers from omitted ability bias
More details: `http://hedibert.org/wp-content/uploads/2016/05/return-to-education-women.pdf`
`R` code: `http://hedibert.org/wp-content/uploads/2016/05/return-to-education-women-R.txt`

C) Estimating the effect of smoking on birth weight: 1388 observations and 14 variables
Outcome: Child birth weight
Treatment: Cigarette smoking (number of packs smoked by the mother per day)
IV: Average price of cigarettes in the state of residence
The IV fails the one requirement of an IV that we can always test.

D) College Proximity as IV: 3010 observations and 31 variables
Card (1995)[1] used wage and education data for a sample of men in 1976 to estimate the return to education.
Outcome: log(wage)
Treatment: Education
IV: Dummy variable for whether someone grew up near a four-year college.
Controls: i) experience, ii) a black dummy variable, iii) dummy variables for living in an Standard Metropolitan Statistical Area (SMSA), and iv) living in the South, and a few others.
$\widehat{\beta}_{ols} = 0.075$ and $\widehat{\beta}_{iv} = 0.132$ (twice as big!) - 95% CI: $(0.069, 0.081)$.
$se(\widehat{\beta}_{ols}) = 0.003$ and $se(\widehat{\beta}_{iv}) = 0.055$ (twenty times as big!) = 95% CI: $(0.022, 0.242)$.
Larger CIs: Price paid for consistent estimator of the return to (endogenous) education.

[1] Card (1995) Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, ed. Christophides, Grant and Swidinsky, 201-222. Toronto: University of Toronto Press.

9.4 A few of my own papers on Bayesian IV modeling

- Hahn, He and Lopes (2018)
  Bayesian factor model shrinkage for linear IV regression with many instruments,
  Journal of Business and Economic Statistics, 36(2), 278-287.

  Abstract: A Bayesian approach for the many instruments problem in linear instrumental variable models is presented. The new approach has two components. First, a slice sampler is developed, which leverages a decomposition of the likelihood function that is a Bayesian analogue to two-stage least squares. The new sampler permits non-conjugate shrinkage priors to be implemented easily and efficiently. The new computational approach permits a Bayesian analysis of problems that were previously infeasible due to computational demands that scaled poorly in the number of regressors. Second, a new predictor-dependent shrinkage prior is developed specifically for the many instruments setting. The prior is constructed based on a factor model decomposition of the matrix of observed instruments, allowing many instruments to be incorporated into the analysis in a robust way. Features of the new method are illustrated via a simulation study and three empirical examples.

- Lopes and Polson (2014)
  Bayesian instrumental variables: likelihoods and priors,
  Econometric Reviews, 33, 100-121.

  Abstract: Instrumental variable (IV) regression provides a number of statistical challenges due to the shape of the likelihood. We review the main Bayesian literature on instrumental variables and highlight these pathologies. We discuss Jeffreys priors, the connection to the errors-in-the-variables problems and more general error distributions. We propose, as an alternative to the inverted Wishart prior, a new Cholesky-based prior for the covariance matrix of the errors in IV regressions. We argue that this prior is more f lexible and more robust thanthe inverted Wishart prior since it is not based on only one tightness parameter and therefore can be more informative about certain components of the covariance matrix and less informative about others. We show how prior-posterior inference can be formulated in a Gibbs sampler and compare its performance in the weak instruments case for synthetic as well as two illustrations based on well-known real data.

- Heckman, Lopes and Piatek (2014)
  Treatment effects: a Bayesian perspective,
  Econometric Reviews, 33, 36-67.

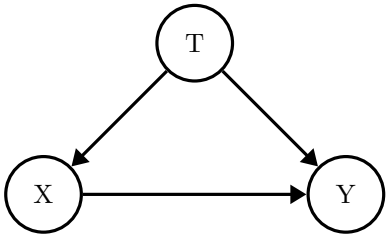
  Abstract: This paper contributes to the emerging Bayesian literature on treatment effects. It derives treatment parameters in the framework of a potential outcomes model with a treatment choice equation, where the correlation between the unobservable components of the model is driven by a low-dimensional vector of latent factors. The analyst is assumed to have access to a set of measurements generated by the latent factors. This approach has attractive features from both theoretical and practical points of view. Not only does it address the fundamental identification problem arising from the inability to observe the same person in both the treated and untreated states, but it also turns out to be straightforward to implement. Formulae are provided to compute mean treatment effects as well as their distributional versions. A Monte Carlo simulation study is carried out to illustrate how the methodology can easily be applied.

10 Difference-in-differences (DiD)

This section is partially based on Nick Huntington-Klein's notes at `https://www.nickchk.com/causalgraphs.html`.

> "There's a group of people, let's call them Treated, who at a certain point had a new policy applied to them. We can observe them both Before the treatment went into effect, and After. We think that the policy treatment might have had an effect on Y. Ideally, we could just look at whether Y went up After Treatment, compared to Before, and call it a day. However, there are plenty of reasons this might not work! Y might have risen for all groups at the same time that treatment was imposed, not just for the Treated group."

Below, X is Treatment, Y is the outcome and T is the measured confounder.

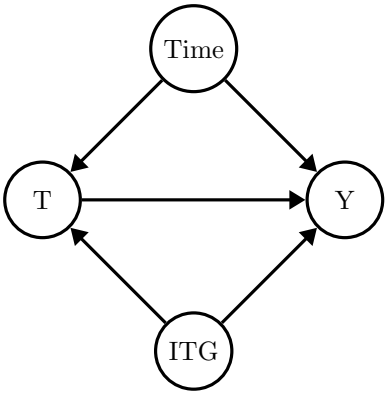


Example: Effect of switching from cubicles to open office on productivity

> "For example, say Treatment is a particular office switching from cubicles to an open office, and Y is productivity. They make the switch on January 1, 2017, so Before Treatment might be 2016 and After Treatment might be 2017. But the economy also improved from 2016 to 2017, so maybe the increase in productivity has nothing to do with the open office."
>
> "When this happens, the difference between Y Before treatment and Y After treatment for the Treated group will reflect two things:
>
> - *We want:* The effect of Treatment on Y, and
> - *We don't want:* The way that Y may have changed over Time for reasons unrelated to Treatment.
>
> Time gives us a **back-door path** from Treatment to Y. We can get from Treatment to Y either through the Treatment → Y path (which we want), or the Treatment ← Time → Y path (which we don't)."

Question: What can we do? Answer: Adding a control group

> We can add a Control group that never gets treated (in our example, an office that keeps its cubicles throughout 2016 and 2017). This is going to let us control for Time, but introduces the problem that now we have another back door, since the Control and Treatment groups may be different.
>
> In the below diagram, a person receives "Treatment" only if they are in the Treated group AND in the Time period AFTER treatment is applied.
>
> In addition to our Time back door, we also have a back door from Treatment ← In Treated Group → Y that we need to close:



"We can close both back-door paths through Time and In Treated Group using Difference-in-Differences. The idea is that we look at how much Y changed from Before to After in the Treated group, and also how much Y changed from Before to After in the Control group (those are the Differences)."

10.1 DiD linear regression and parameter interpretation

The following four graphs were borrowed from here:
`https://bookdown.org/cuborican/RE_STAT/difference-in-differences.html#regression-did`
In fact, check the Section of 11.2.4 (Difference in Differences: Animated).

There are two equivalent strategies to think about the two "differences":

	Strategy #1	Strategy #2
Difference 1	**Average change of treated over time**	**Average change between treated and control in post-treatment period**
	$E(Y_{it}\|T_i = 1, P_t = 1) - E(Y_{it}\|T_i = 1, P_t = 0)$	$E(Y_{it}\|T_i = 1, P_t = 1) - E(Y_{it}\|T_i = 0, P_t = 1)$
Difference 2	**Average change of control over time**	**Average change between treated and control in pre-treatment period**
	$E(Y_{it}\|T_i = 0, P_t = 1) - E(Y_{it}\|T_i = 0, P_t = 0)$	$E(Y_{it}\|T_i = 1, P_t = 0) - E(Y_{it}\|T_i = 0, P_t = 0)$
Difference in Differences	**Difference 1 – Difference 2**	

Y_{it}: Outcome of unit i at time t
T_i: Dummy variable indicating 1 if treatment is assigned and 0 otherwise
P_t: Dummy variable indicating 1 if post − treatment period and 0 otherwise

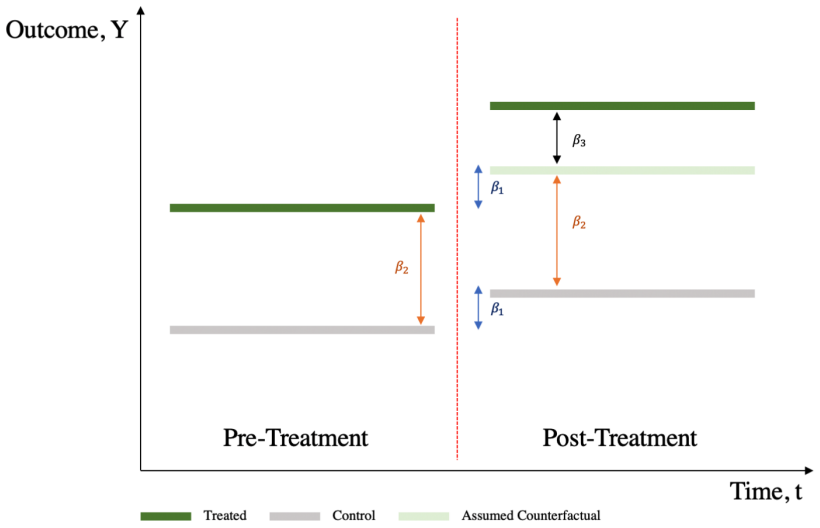
The following linear regression summarizes the above discussion:

$$y_{ti} = \beta_0 + \beta_1 P_t + \beta_2 T_i + \beta_3 P_t \times T_i + x'_{ti}\gamma + \varepsilon_{ti},$$

where $P_t = 1$ for the post-treatment period and $P_t = 0$ for the pre-treatment period, while $T_i = 1$ if individual i is in the treatment group and $T_i = 0$ if individual i is in the control group. The components of x_{ti} are additional control variables and will be cancelled out.

"While it is possible to obtain the DiD estimator by calculating the means by hand, using a regression framework may be more advantageous as it: i) outputs standard errors for hypothesis testing, ii) can be easily extended to include multiple periods and groups, and iii) allows the addition of covariates."

	Treatment Group ($T_i = 1$) (1)	Control Group ($T_i = 0$) (2)	Difference (1) – (2)
Post-Treatment Period ($P_t = 1$) (a)	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1$	$\beta_2 + \beta_3$
Pre-Treatment Period ($P_t = 0$) (b)	$\beta_0 + \beta_2$	β_0	β_2
Difference (a) – (b)	$\beta_1 + \beta_3$	β_1	$\boldsymbol{\beta_3}$



10.2 Example: Increase in the state minimum wage on the employment

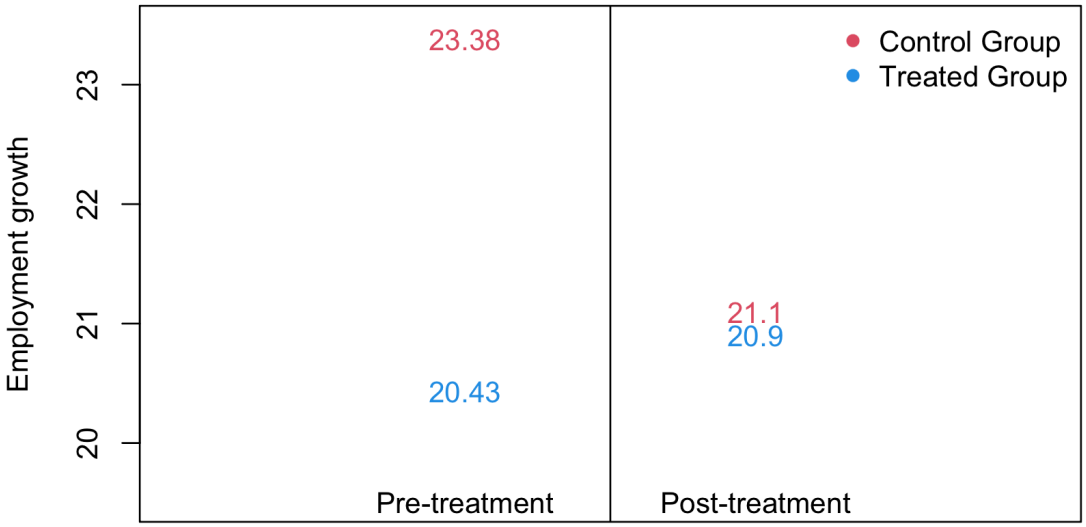
"The data is adapted from the dataset in Card and Krueger (1994), which estimates the causal effect of an increase in the state minimum wage on the employment. On April 1, 1992, New Jersey raised the state minimum wage from $4.25 to $5.05 while the minimum wage in Pennsylvania stays the same at $4.25. Data about the employment in the fast food restaurants in NJ (0) and PA (1) were collected in February 1992 and in November 1992. Total 384 restaurants after removing null values.

Source: Card and Krueger (1994) Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania, *The American Economic Review*, 84(4), 772-793. `https://davidcard.berkeley.edu/papers/njmin-aer.pdf`

- time= 0 : February 1992
- time= 1 : November 1992
- treatment= 1 : New Jersey raised the state minimum wage from $4.25 to $5.05
- treatment= 0 : Pennsylvania minimum wage stays the same at $4.25

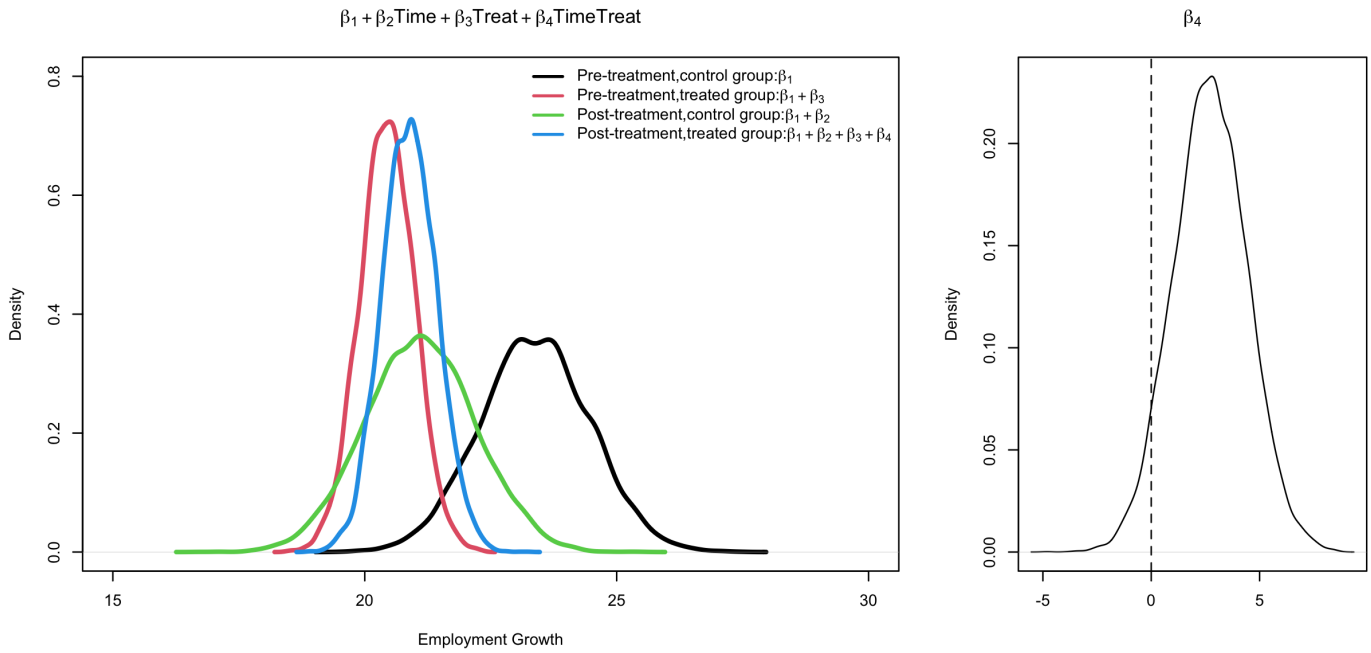
10.2.1 R script - Classical approach

```
data = read.table("https://hedibert.org/wp-content/uploads/2024/10/card-krueger.txt",header=TRUE)
attach(data)
pretreatment.untreated  = mean(outcome[time==0 & treatment==0])
pretreatment.treated    = mean(outcome[time==0 & treatment==1])
posttreatment.untreated = mean(outcome[time==1 & treatment==0])
posttreatment.treated   = mean(outcome[time==1 & treatment==1])

A = posttreatment.treated   - pretreatment.treated
B = posttreatment.untreated - pretreatment.untreated
effect = A-B
c(A,B,effect)
#[1]  0.4666667 -2.2833333  2.7500000

interaction = time*treatment

summary(lm(outcome~time+treatment+interaction))

#lm(formula = outcome ~ time + treatment + interaction)
#
#Residuals:
#    Min      1Q  Median      3Q     Max
#-21.097  -6.472  -0.931   4.603  64.569
#
#Coefficients:
#            Estimate Std. Error t value Pr(>|t|)
#(Intercept)   23.380      1.098  21.288   <2e-16 ***
#time          -2.283      1.553  -1.470   0.1419
#treatment     -2.949      1.224  -2.409   0.0162 *
#interaction    2.750      1.731   1.588   0.1126
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#Residual standard error: 9.511 on 764 degrees of freedom
#Multiple R-squared:  0.007587,Adjusted R-squared:  0.00369
#F-statistic: 1.947 on 3 and 764 DF,  p-value: 0.1206


plot(c(0,0),xlim=c(0,3),ylim=c(19.5,23.5),col=0,axes=FALSE,xlab="",ylab="Employment growth")
box();axis(2)
abline(v=1.5)
text(1,pretreatment.untreated,round(pretreatment.untreated,2),col=2)
text(2,posttreatment.untreated,round(posttreatment.untreated,2),col=2)
text(1,pretreatment.treated,round(pretreatment.treated,2),col=4)
text(2,posttreatment.treated,round(posttreatment.treated,2),col=4)
text(1,19.5,"Pre-treatment")
text(2,19.5,"Post-treatment")
legend("topright",legend=c("Control Group","Treated Group"),col=c(2,4),pch=16,bty="n")
```



10.2.2 R script - Bayesian approach

```
# Bayesian fit
n     = length(outcome)
y     = outcome
X     = cbind(1,time,treatment,interaction)
sig2  = sum((y-X%*%beta1)^2)/(n-4)
iXtX  = solve(t(X)%*%X)
Vbeta = sig2*iXtX
Ebeta = iXtX%*%t(X)%*%y
Lbeta = chol(Vbeta)
M     = 10000
betas = matrix(Ebeta,M,4,byrow=TRUE)+matrix(rnorm(4*M),M,4)%*%Lbeta

layout(matrix(c(1,1,2,1,1,2), 2, 3, byrow = TRUE))
layout(matrix(c(1,1,2), 1, 3, byrow = TRUE))

plot(density(betas[,1]),xlim=c(15,30),ylim=c(0,0.8),xlab="Employment Growth",main="",lwd=3)
lines(density(betas[,1]+betas[,3]),col=2,lwd=3)
lines(density(betas[,1]+betas[,2]),col=3,lwd=3)
lines(density(betas[,1]+betas[,2]+betas[,3]+betas[,4]),col=4,lwd=3)
legend("topright",legend=c(
expression(paste("Pre-treatment,control group:",beta[1],sep="")),
expression(paste("Pre-treatment,treated group:",beta[1]+beta[3],sep="")),
expression(paste("Post-treatment,control group:",beta[1]+beta[2],sep="")),
expression(paste("Post-treatment,treated group:",beta[1]+beta[2]+beta[3]+beta[4],sep=""))),
col=1:4,lwd=2,bty="n")
title(expression(beta[1]+beta[2]*Time+beta[3]*Treat+beta[4]*Time*Treat))

plot(density(betas[,4]),xlab="",main=expression(beta[4]))
abline(v=0,lty=2)
```



11 Regression Discontinuity Design (RDD)

Most of the text and examples in this section were taken primarily from Chapter 4 (Regression Discontinuity Designs), pages 147-164, of Angrist and Pischke's (2015) book enttitle *Mastering 'Metrics: The Path from Cause to Effect*. They start the chapter with the following paragraph:

> Human behavior is constrained by rules.
>
> - The State of California limits elementary school class size to 32 students; 33 is one too many.
> - The Social Security Adm. won't pay you a penny in retirement benefits until you've reached age 62.
> - Potential armed forces recruits with test scores in the lower deciles are ineligible for military service.
>
> They continue by saying:
>
> - For rules that constrain the role of chance in human affairs often generate interesting experiments.
> - Masters of 'metrics exploit these experiments with **regression discontinuity (RD) design**.
> - RD doesn't work for all causal questions, but it works for many.
> - And when it does, **the results have almost the same causal force as those from a randomized trial**.

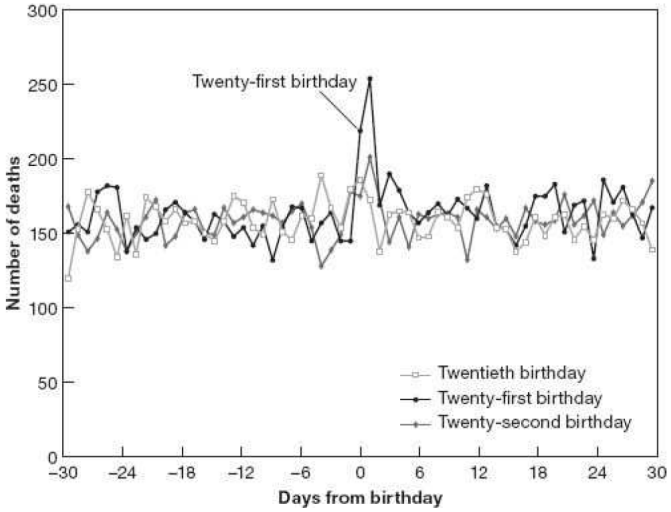
11.1 Birthdays and Funerals

In the US the minimum legal drinking age (MLDA) is 21 years of age. Therefore, "the history of the MLDA generates a natural experiment that can be used for a sober assessment of alcohol policy."

The MLDA experiment emerges from the fact that a small change in age (measured in months or even days) generates a big change in legal access. The difference a day makes can be seen in the figure below, which plots the relationship between birthdays and funerals, i.e. number of deaths among Americans aged 20-22 between 1997 and 2003.

FIGURE 4.1

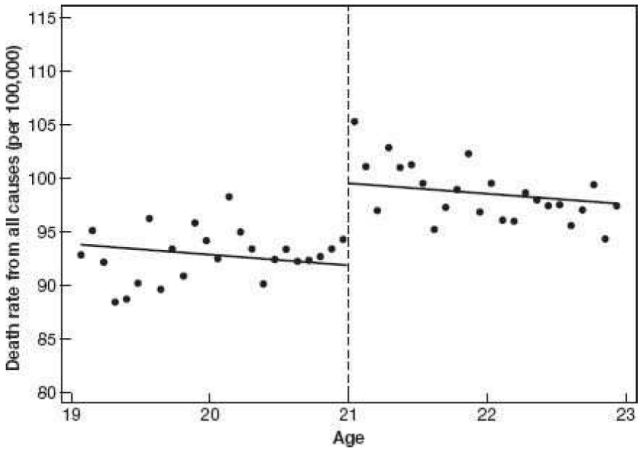
Birthdays and funerals



Mortality risk shoots up on and immediately following a twenty-first birthday. This spike adds about 100 deaths to a baseline level of about 150 per day. There's something special about the twenty-first birthday.

FIGURE 4.2

A sharp RD estimate of MLDA mortality effects



Notes: This figure plots death rates from all causes against age in months. The lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months (the vertical dashed line indicates the minimum legal drinking age (MLDA) cutoff).

Death rates fluctuate from month to month, but few rates to the left of the age-21 cutoff are above 95. At ages over 21, however, death rates shift up, and few of those to the right of the age-21 cutoff are below 95.

The causal question addressed by Figure 4.2 is the effect of legal access to alcohol on death rates. The treatment variable in this case can be written D_a, where $D_a = 1$ indicates legal drinking and is $D_a = 0$ otherwise. D_a is a function of age, a: the MLDA transforms 21-year-olds from underage minors to legal alcohol consumers. We capture this transformation in mathematical notation by writing

$$D_a = \begin{cases} 1 & \text{if } a \geq 21 \\ 0 & \text{if } a < 21 \end{cases}$$

- Treatment status is a deterministic function of a, so that once we know a, we know D_a.
- Treatment status is a discontinuous function of a, because no matter how close a gets to the cutoff, D_a remains unchanged until the cutoff is reached (**sharp RD**).
- The variable that determines treatment, age in this case, is called the **running variable**.
- **Sharp RD designs:** treatment switches cleanly off or on as the running variable passes a cutoff.
- The MLDA is a sharp function of age, so an investigation of MLDA effects on mortality is a sharp RD study.

11.2 Controling for smooth variation in death rates

Mortality clearly changes with the running variable, a, for reasons unrelated to the MLDA.

Death rates from disease-related causes like cancer (known to epidemiologists as internal causes) are low but increasing for those in their late teens and early 20s, while deaths from external causes, primarily car accidents, homicides, and suicides, fall.

To separate this trend variation from any possible MLDA effects, an RD analysis controls for smooth variation in death rates generated by a.

RD gets its name from the practice of using regression models to implement this control.

A simple RD analysis of the MLDA estimates causal effects using a regression like

$$y_a = \alpha + \rho D_a + \gamma a + \epsilon_a$$

where y_a is the death rate in month a (month is defined as a 30-day interval counting from the 21st birthday).

The regression equation includes the treatment dummy, D_a, as well as a linear control for age in months.

Fitted values from the above regression produce the lines drawn in Figure 4.2.

The negative slope, captured by γ, reflects smoothly declining death rates among young people as they mature.

The parameter ρ captures the jump in deaths at age 21.

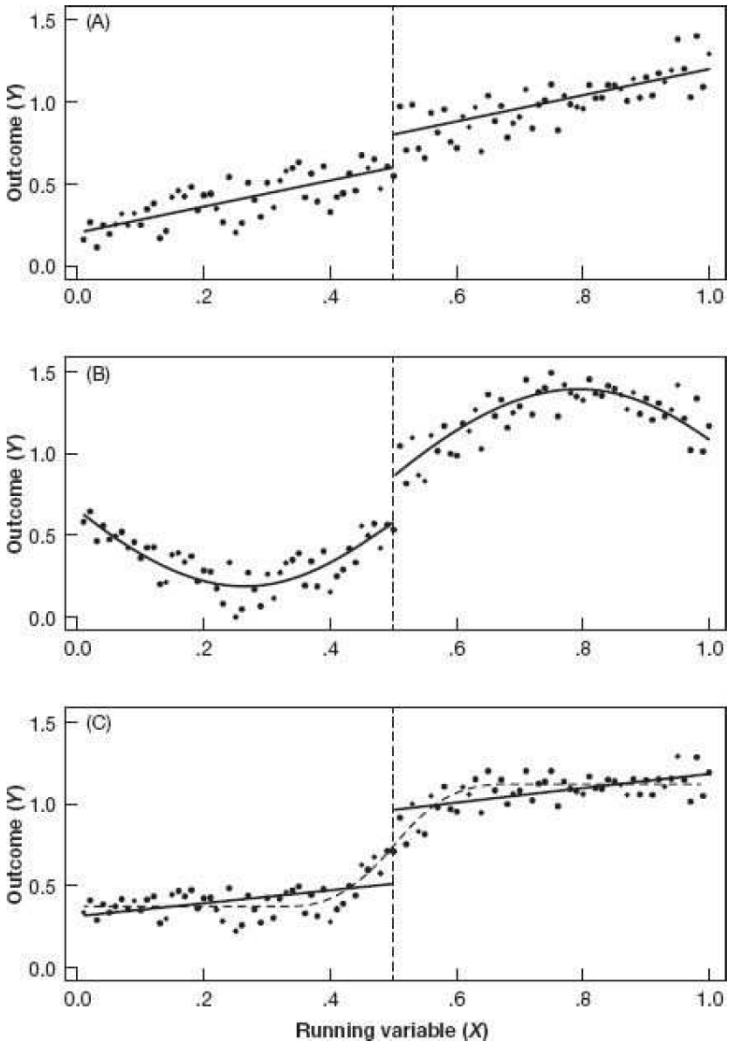
The regression generates an estimate of ρ equal to 7.7

This estimate indicates a substantial increase in risk at the MLDA cutoff.

11.3 Nonlinearity mistaken for a discontinuity

Panel A shows RD with a linear model for $E(y_i|X_i)$; panel B adds some curvature. Panel C shows nonlinearity mistaken for a discontinuity. The vertical dashed line indicates a hypothetical RD cutoff.

FIGURE 4.3
RD in action, three ways



11.4 Comparing two curves

Linear regression

$$y_a = \alpha + \rho D_a + \gamma a + \epsilon_a$$

Quadratic regressions with interactions

$$y_a = \alpha + \rho D_a + \gamma_1(a - a_0) + \gamma_2(a - a_0)^2 + \delta_1[(a - a_0)D_a] + \delta_2[(a - a_0)^2 D_a] + \epsilon_a$$

TABLE 4.1

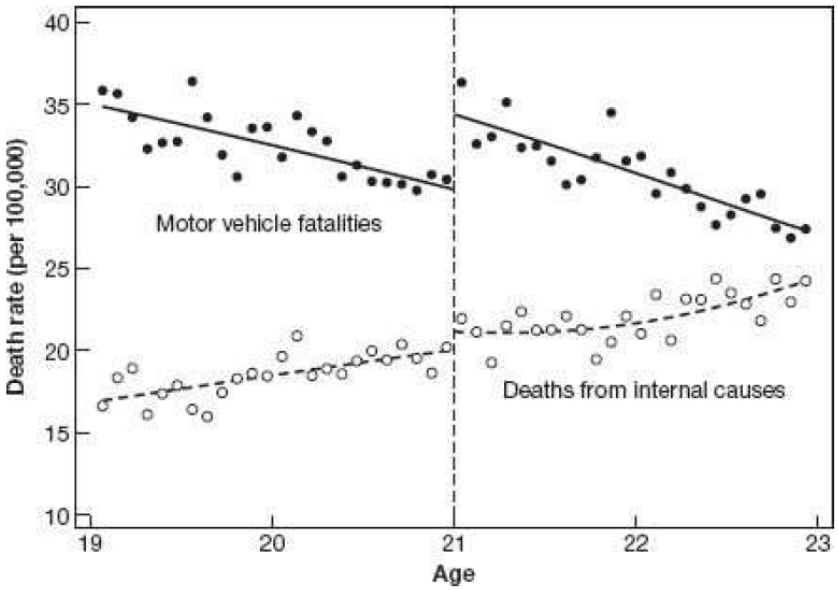
Sharp RD estimates of MLDA effects on mortality

Dependent variable	Ages 19–22		Ages 20–21	
	(1)	(2)	(3)	(4)
All deaths	7.66	9.55	9.75	9.61
	(1.51)	(1.83)	(2.06)	(2.29)
Motor vehicle accidents	4.53	4.66	4.76	5.89
	(.72)	(1.09)	(1.08)	(1.33)
Suicide	1.79	1.81	1.72	1.30
	(.50)	(.78)	(.73)	(1.14)
Homicide	.10	.20	.16	−.45
	(.45)	(.50)	(.59)	(.93)
Other external causes	.84	1.80	1.41	1.63
	(.42)	(.56)	(.59)	(.75)
All internal causes	.39	1.07	1.69	1.25
	(.54)	(.80)	(.74)	(1.01)
Alcohol-related causes	.44	.80	.74	1.03
	(.21)	(.32)	(.33)	(.41)
Controls	age	age, age^2, interacted with over-21	age	age, age^2, interacted with over-21
Sample size	48	48	24	24

Notes: This table reports coefficients on an over-21 dummy from regressions of month-of-age-specific death rates by cause on an over-21 dummy and linear or interacted quadratic age controls. Standard errors are reported in parentheses.

FIGURE 4.5

RD estimates of MLDA effects on mortality by cause of death



Notes: This figure plots death rates from motor vehicle accidents and internal causes against age in months. Lines in the figure plot fitted values from regressions of mortality by cause on an over-21 dummy and a quadratic function of age in months, interacted with the dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

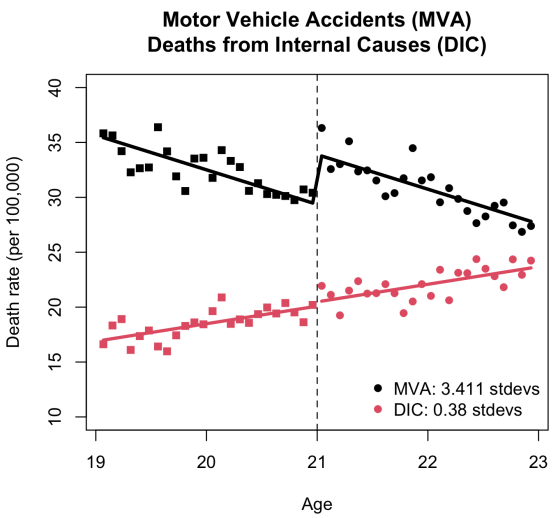
11.5 Replicating Figure 4.5

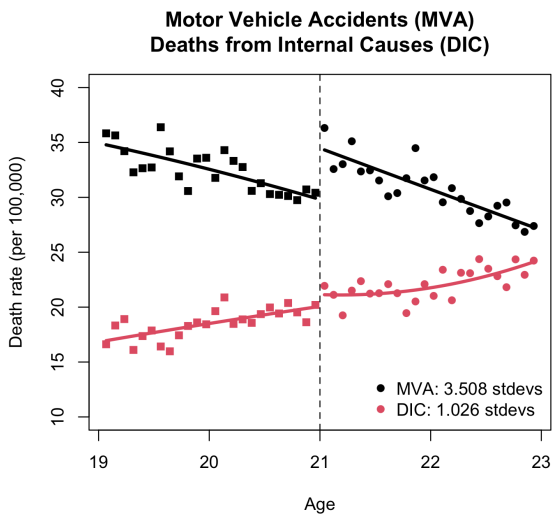
R script replicating Figure 4.5, Chapter 4: Regression discontinuity design, from Angrist and Pischke (2015).

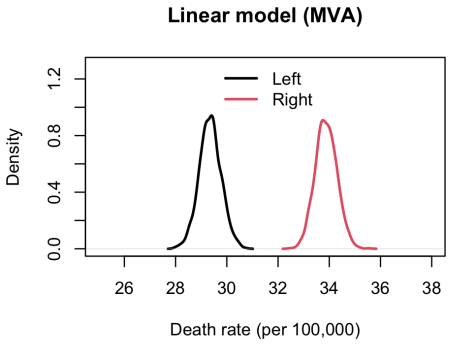
```
data = read.table("age-mva-internal.txt",header=TRUE)
attach(data)
n = nrow(data)
over21 = rep(0,n)
over21[age>21]=1

par(mfrow=c(1,1))
plot(age,mva,pch=over21+15,ylim=c(10,40),xlab="Age",ylab="Death rate (per 100,000)")
points(age,internal,col=2,pch=over21+15)
abline(v=21,lty=2)
legend("bottomright",legend=c("Motor Vehicle Accidents","Deaths from Internal Causes"),col=1:2,pch=16,bty="n")
abline(v=21,lty=2)

# Linear models
fit  = lm(mva~age+over21)
right  = c(1,21,1)%*%fit$coef
left   = c(1,21,0)%*%fit$coef
delta  =  right-left
sigma = summary(fit)$sigma
nsig = round(delta/sigma,3)

fit1  = lm(internal~age+over21)
right1 = c(1,21,1)%*%fit1$coef
left1  = c(1,21,0)%*%fit1$coef
delta1 =  right1-left1
sigma1 = summary(fit1)$sigma
nsig1 = round(delta/sigma1,3)

par(mfrow=c(1,2))
plot(age,mva,pch=over21+15,ylim=c(10,40),xlab="Age",ylab="Death rate (per 100,000)")
points(age,internal,col=2,pch=over21+15)
abline(v=21,lty=2)
lines(age,fit$fit,lwd=3)
lines(age,fit$fit,lwd=3)
lines(age[over21==0],fit$fit[over21==0],lwd=3)
lines(age[over21==1],fit$fit[over21==1],lwd=3)
lines(age[over21==0],fit1$fit[over21==0],lwd=3,col=2)
lines(age[over21==1],fit1$fit[over21==1],lwd=3,col=2)
legend("bottomright",legend=c(paste("MVA: ",nsig," stdevs",sep=""),paste("DIC: ",nsig1," stdevs",sep="")),col=1:2,pch=16,bty="n")
title("Motor Vehicle Accidents (MVA)\nDeaths from Internal Causes (DIC)")


# Quadratic models
age2       = age^2
over21age  = over21*age
over21age2 = over21*age2

fit  = lm(mva~age+over21+age2+over21age+over21age2)
right  = c(1,21,1,21^2,21,21^2)%*%fit$coef
left   = c(1,21,0,21^2,0,0)%*%fit$coef
delta  =  right-left
sigma = summary(fit)$sigma

fit1  = lm(internal~age+over21+age2+over21age+over21age2)
right1 = c(1,21,1,21^2,21,21^2)%*%fit1$coef
left1  = c(1,21,0,21^2,0,0)%*%fit1$coef
delta1 =  right1-left1
sigma1 = summary(fit1)$sigma

plot(age,mva,pch=over21+15,ylim=c(10,40),xlab="Age",ylab="Death rate (per 100,000)")
points(age,internal,col=2,pch=over21+15)
abline(v=21,lty=2)
lines(age[over21==0],fit$fit[over21==0],lwd=3)
lines(age[over21==1],fit$fit[over21==1],lwd=3)
lines(age[over21==0],fit1$fit[over21==0],lwd=3,col=2)
lines(age[over21==1],fit1$fit[over21==1],lwd=3,col=2)
legend("bottomright",legend=c(
       paste("MVA: ",round(delta/sigma,3)," stdevs",sep=""),
       paste("DIC: ",round(delta1/sigma1,3)," stdevs",sep="")),
       col=1:2,pch=16,bty="n")
title("Motor Vehicle Accidents (MVA)\nDeaths from Internal Causes (DIC)")
```

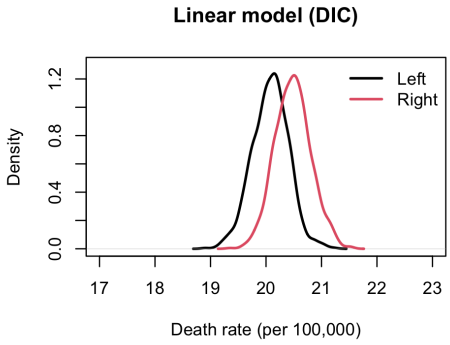
Bayesian approach

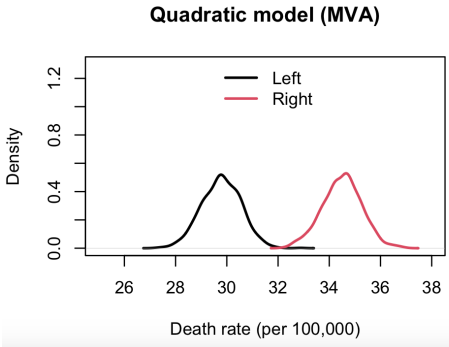
```
M = 1000
par(mfrow=c(2,2))

# Linear models
y     = mva
X     = cbind(1,age,over21)
iXtX  = solve(t(X)%*%X)
b     = iXtX%*%t(X)%*%y
sigma = mean((y-X%*%b)^2)
Vb    = sigma*iXtX
L     = chol(Vb)
betas = matrix(b,M,3,byrow=TRUE)+matrix(rnorm(M*3),M,3)%*%L
right = betas%*%c(1,21,1)
left  = betas%*%c(1,21,0)

plot(density(left),xlim=c(25,38),xlab="Death rate (per 100,000)",main="",lwd=2,ylim=c(0,1.3))
lines(density(right),col=2,lwd=2)
title("Linear model (MVA)")
legend("top",legend=c("Left","Right"),col=1:2,lwd=2,bty="n")

y     = internal
X     = cbind(1,age,over21)
iXtX  = solve(t(X)%*%X)
b     = iXtX%*%t(X)%*%y
sigma = mean((y-X%*%b)^2)
Vb    = sigma*iXtX
L     = chol(Vb)
betas = matrix(b,M,3,byrow=TRUE)+matrix(rnorm(M*3),M,3)%*%L
right = betas%*%c(1,21,1)
left  = betas%*%c(1,21,0)
plot(density(left),xlim=c(17,23),xlab="Death rate (per 100,000)",main="",lwd=2,ylim=c(0,1.3))
lines(density(right),col=2,lwd=2)
title("Linear model (DIC)")
legend("topright",legend=c("Left","Right"),col=1:2,lwd=2,bty="n")


# Quadratic models
y     = mva
X     = cbind(1,age,over21,age2,over21age,over21age2)
iXtX  = solve(t(X)%*%X)
b     = iXtX%*%t(X)%*%y
sigma = mean((y-X%*%b)^2)
Vb    = sigma*iXtX
L     = chol(Vb)
betas = matrix(b,M,6,byrow=TRUE)+matrix(rnorm(M*6),M,6)%*%L
right = betas%*%c(1,21,1,21^2,21,21^2)
left  = betas%*%c(1,21,0,21^2,0,0)

plot(density(left),xlim=c(25,38),xlab="Death rate (per 100,000)",main="",lwd=2,ylim=c(0,1.3))
lines(density(right),col=2,lwd=2)
title("Quadratic model (MVA)")
legend("top",legend=c("Left","Right"),col=1:2,lwd=2,bty="n")

y     = internal
X     = cbind(1,age,over21,age2,over21age,over21age2)
iXtX  = solve(t(X)%*%X)
b     = iXtX%*%t(X)%*%y
sigma = mean((y-X%*%b)^2)
Vb    = sigma*iXtX
L     = chol(Vb)
betas = matrix(b,M,6,byrow=TRUE)+matrix(rnorm(M*6),M,6)%*%L
right = betas%*%c(1,21,1,21^2,21,21^2)
left  = betas%*%c(1,21,0,21^2,0,0)

plot(density(left),xlim=c(17,23),xlab="Death rate (per 100,000)",main="",lwd=2,ylim=c(0,1.3))
lines(density(right),col=2,lwd=2)
title("Quadratic model (DIC)")
legend("top",legend=c("Left","Right"),col=1:2,lwd=2,bty="n")
```

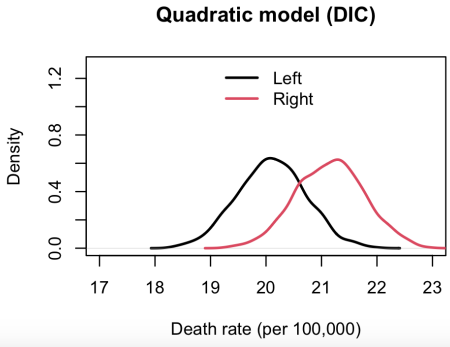

Motor Vehicle Accidents (MVA)
Deaths from Internal Causes (DIC)
Death rate (per 100,000)
Age
MVA: 3.411 stdevs
DIC: 0.38 stdevs


Motor Vehicle Accidents (MVA)
Deaths from Internal Causes (DIC)
Death rate (per 100,000)
Age
MVA: 3.508 stdevs
DIC: 1.026 stdevs


Linear model (MVA)
Density
Death rate (per 100,000)
Left
Right


Linear model (DIC)
Density
Death rate (per 100,000)
Left
Right


Quadratic model (MVA)
Density
Death rate (per 100,000)
Left
Right


Quadratic model (DIC)
Density
Death rate (per 100,000)
Left
Right

11.6 Non-parametric RDD

Local behavior: For observations in $[a_0 - b, a_0 + b]$, for small b, nonlinear trends need not concern us at all.

Strategy: Comparing averages in a narrow window just to the left and just to the right of the cutoff.

High variance: Very narrow window $\rightarrow$ few observations $\rightarrow$ too imprecise estimates.

Trade-off: bias reduction near the boundary against the increased variance.

Local linear regression: In this case, one would consider the linear regression

$$y_a = \alpha + \rho D_a + \gamma a + \epsilon_a,$$

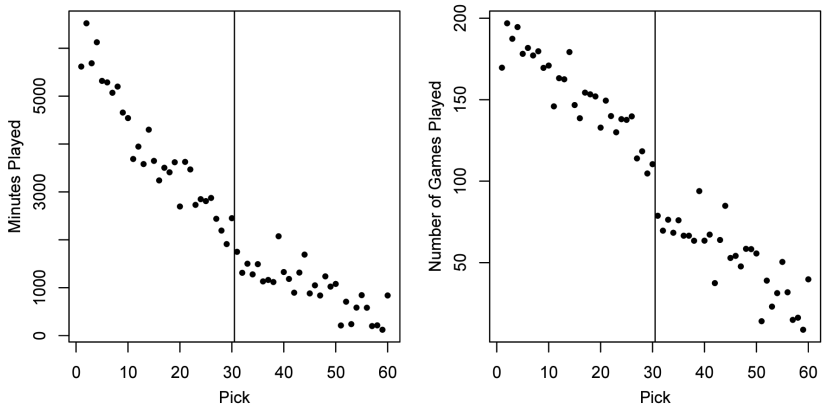
for the subset of observations where $a \in [a_0 - b, a_0 + b]$.

11.7 Example: The NBA draft

Branson, Z., M. Rischard, L. Bornn, & L.W. Miratrix (2019) A nonparametric Bayesian methodology for RDDs. *Journal of Statistical Planning and Inference*, 202, 14-30. `https://arxiv.org/abs/1704.04858`.

1,238 NBA basketball players drafted between 1995 and 2016.

> The National Basketball Association (NBA) draft, held annually, is divided into 2 rounds, where each NBA teams gets one selection per round to draft a player of their choice. Because players are picked sequentially, there is no reason to believe there is a marked skill difference between the last pick of the first round and first pick of the second round. However, because of the difference in the perceived value of first-round versus second-round picks, as well as differing contract structures between the two rounds, we suspect that first-round picks are treated more favorably and given more playing time than their second-round colleagues, above and beyond what can be explained by differences in skill. As such, we seek to explore if there is a difference between first- and second-round picks in both skill and playing time.



Bayesian Non-Parametric (BNP) Regression Discontinuity Design (RDD) via Gaussian Process Regression (GPR) - "In summary, using our GPR methodology, we find that the treatment effect of being a second-round pick significantly reduces the number of games played and marginally reduces the number of minutes played."

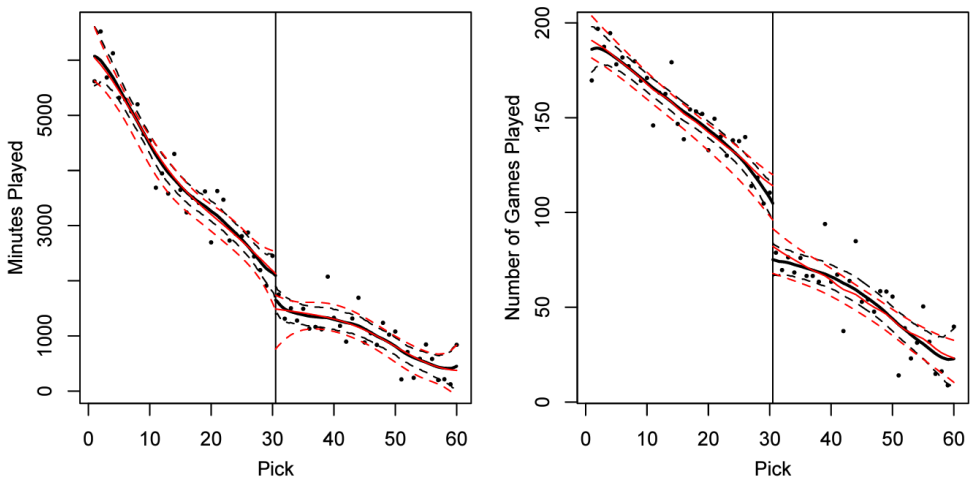


Figure 5: The estimated mean functions (solid lines) and corresponding confidence intervals (dashed lines) for LLR (black lines) and GPR (red lines). The lines for LLR were produced by the `rdd` R package (Dimmery, 2013), but using the bandwidth estimated by the `rdrobust` R package. The two treatment groups are the first round of picks (picks 30 and below) and second round of picks (picks 31 and above). We set the boundary to be $b = 30.5$ to minimize the amount of extrapolation that needs to be conducted on both sides of the boundary to estimate the treatment effect.

Table 1: Treatment effect estimation for LLR, robust LLR, and GPR on NBA data

	LLR		**Robust LLR**		**GPR**	
Outcome	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Box Plus-Minus	**-3.06**	[-5.20, -0.92]	**-3.37**	[-5.86, -0.88]	-2.42	[-5.02, 0.01]
Win Shares	-1.58	[-4.09, 0.93]	-1.73	[-4.76, 1.29]	-1.08	[-3.04, 0.88]
Minutes Played	-446.29	[-1042.81, 150.23]	-406.63	[-1123.99, 310.72]	-640.75	[-1259.65, 5.51]
Games Played	**-29.71**	[-40.20, -19.22]	**-28.16**	[-40.81, -15.51]	**-32.00**	[-45.69, -18.14]

Point estimates and 95% confidence intervals for the treatment effect on each of the four outcomes: box plus-minus, win shares played, number of minutes played, and number of games played. Statistically significant point estimates are in bold.

11.8 References

- Alcantara, Wang, Hahn and Lopes (2024)
  Modified BART for Learning Heterogeneous Effects in RDDs
  https://arxiv.org/abs/2407.14365

- Chib, Greenberg and Simoni (2023)
  Nonparametric Bayes Analysis of the Sharp and Fuzzy RDDs
  Econometric Theory, 39(3), 481-533.
  doi:10.1017/S0266466622000019

- Cattaneo, M.D., N. Idrobo, & R. Titiunik (2020)
  A Practical Introduction to RDDs: Foundations.
  https://arxiv.org/abs/1911.09511

- Gelman and Imbens (2019)
  Why High-Order Polynomials Should Not Be Used in RDDs,
  Journal of Business & Economic Statistics, 37(3), 447-456.
  https://doi.org/10.1080/07350015.2017.1366909

- Gelman (2019)
  Another Regression Discontinuity Disaster and what can we learn from it
  https://statmodeling.stat.columbia.edu/2019/06/25/another-regression-discontinuity-disaster-and/

- Cattaneo, M.D., R. Titiunik, & G. Vazquez-Bare (2017)
  Comparing inference approaches for RDDs: A reexamination of the effect of head start on child mortality.
  Journal of Policy Analysis and Management, 36, 643-681.

- Gelman and Zelizer (2015)
  Evidence on the deleterious impact of sustained use of polynomial regression on causal inference,
  Research and Politics, 2(1).
  https://doi.org/10.1177/2053168015569830

- Cattaneo, M.D., B.R. Frandsen, & R. Titiunik (2015)
  Randomization inference in the RDD: An application to party advantages in the US senate.
  Journal of Causal Inference, 3, 1-24.

- Calonico, S., M.D. Cattaneo, & R. Titiunik (2014)
  Robust nonparametric confidence intervals for RDDs.
  Econometrica, 82, 2295-2326.

- Imbens, G. & K. Kalyanaraman (2012)
  Optimal bandwidth choice for the regression discontinuity estimator.
  Review of Economic Studies, 79, 933-959.

- Imbens, G.W. & T. Lemieux (2008)
  Regression discontinuity designs: A guide to practice.
  Journal of Econometrics, 142, 615-635.

- Hahn, J.Y., P. Todd, & W. Van der Klaauw (2001)
  Identification and estimation of treatment effects with a RDD.
  Econometrica, 69, 201-209.

- Thistlethwaite, D.L. & D.T. Campbell (1960)
  Regression-discontinuity analysis: An alternative to the ex-post facto experiment.
  Journal of Educational Psychology, 51, 309-317.