Statistical Data Science in Action

Experiencing Realworld Data Analytics

Jun Yan
Department of Statistics
University of Connecticut

January 22, 2019

Preliminaries

- R: SIAM workshop on R by Wenjie Wang
  - session one; source repo
  - session two; source repo
- Python
  - Learning Python in one video
- Bookdown by Yihui Xie
  - R package source
  - online book
- Git/GitHub
  - Learn Git in 20 minues

R Packages

```
##' Load needed packages, and install them if not installed.
##'
##' @usage need.packages(pkg)
##' @param pkg A character vector specifying the packages needed to
##'     reproduce this document.
##' @param ... Other arguments passed to function
##'     \code{\link[base]require}.
##' @return \code{NULL} invisibly.
##' @examples
##' need.pacakges(c("ggplot2", "geepack"))
need.packages <- function(pkg, ...)
{
    new.pkg <- pkg[! (pkg %in% installed.packages()[, "Package"])]
    if (length(new.pkg))
        install.packages(new.pkg, repos = "https://cloud.r-project.org")
    foo <- function(a, ...) suppressMessages(require(a, ...))
    sapply(pkg, foo, character.only = TRUE)
    invisible(NULL)
}
pkgs <- c("bookdown", "revealjs")
need.packages(pkgs)
```

```
## Warning: package 'bookdown' was built under R version 3.5.2
```

```
## Warning: package 'revealjs' was built under R version 3.5.2
```

Motivation

- Hands on realworld data analytics
- Tools for analytics as well as tools for project management
- Workflow ensuring reproducibility
- Version control
- Learning on the fly

- Communications (oral, written, critical review, etc.)
- Data science competitions / data challenges

Operation

- Instructor is a coach
- Everyone is both an instructor, a learner, and a reviewer
- Projects (training/real; group/individual; common/open; repository)
- Peer review like a journal review system
- Open source book on data science in action
- Guest lecturers
- Outreach (partnership with CT Open Data, CT Data Collaborative, etc.)
- Google

Topics

- Clusterd data analysis (GEE, copulas, NLME)
- Causal inference
- Deep learning
- Propensity score

Preliminaries: Statistical and Causal Models

Simpson's Paradox

- Edward Simpson (1951, JRSSb)
- A statistical association that holds for an entire population is reversed in every subpopulation.
- Table 1.1 Results of a study into a new drug, with gender being taken into account
- The reason the drug appears to be harmful overall is that, if we select a drug user at random, that person is more likely to be a woman and hence less likely to recover than a random person who does not take the drug.
- Put differently, being a woman is a common cause of both drug taking and failure to recover.

Simpson's Paradox: Continuous varaible

- Figure 1.1 Results of the exerciseâ "cholesterol study, segregated by age
- However, segregated data does not always give the correct answer.
- Suppose we looked at the same numbers from our first example of drug taking and recovery, instead of recording participantsâ gender, patientsâ blood pressure were recorded at the end of the experiment.
  - the drug affects recovery by lowering the blood pressure of those who take itâ but unfortunately, it also has a toxic effect.
  - would you recommend the drug to a patient?

Segregated data does not always give the correct answer

- Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account
- Since lowering blood pressure is one of the mechanisms by which treatment affects recovery, it makes no sense to separate the results based on blood pressure.
- So we consult the results for the general population, we find that treatment increases the probability of recovery, and we decide that we should recommend treatment.
- None of the information that allowed us to make a treatment decisionâ not the timing of the measurements, not the fact that treatment affects blood pressure, and not the fact that blood pressure affects recoveryâ was found in the data.
- Trivial though the assumption â treatment does not cause sexâ may seem, there is no way to test it in the data, nor is there any way to represent it in the mathematics of standard statistics.

A calculus of causation

In order to rigorously approach our understanding of the causal story behind data, we need four things:

1. A working definition of â causation.â
2. A method by which to formally articulate causal assumptionsâ that is, to create causal models.
3. A method by which to link the structure of a causal model to features of data.
4. A method by which to draw conclusions from the combination of causal assumptions embedded in a model and data.

A variable X is a cause of a variable Y if Y in any way relies on X for its value.

Graph

- Graph theory provides a useful mathematical language that allows us to address problems of causality with simple operations similar to those used to solve arithmetic problems.
- A mathematical graph is a collection of *vertices* (or, as we will call them, *nodes*) and *edges*.
- Two nodes are *adjacent* if there is an edge between them.
- *complete graph*
- *path* between two nodes
- directed vs undirected
- parent vs child
- ancester vs descendant
- cyclic vs acyclic

Structural causal model

- A structural causal model consists of two sets of variables U and V, and a set of functions f that assigns each variable in V a value based on the values of the other variables in the model.
- A variable X is a direct cause of a variable Y if X appears in the function that assigns Yâ s value. X is a cause of Y if it is a direct cause of Y,or of any cause of Y.
- exogenous vs endogenous
- root node
- Every SCM is associated with a graphical causal model, referred to informally as a â graph- ical modelâ or simply â graph.â

- We will deal primarily with SCMs for which the graphical models are directed acyclic graphs (DAGs).

Product decomposition

- Rule of product decomposition: For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P(child \mid parents)$ over all the â families â in the graph. Formally, we write this rule as

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i|pa_i)$$

  where pa_i stands for the values of the parents of variable X_i.

Graphical Models and Their Applications

Connecting Models to Data

- Probabilities, graphs, structural equations
- The concept of independence, which in the language of probability is defined by alge braic equalities, can be expressed visually using directed acyclic graphs (DAGs).
- Further, this graphical representation will allow us to capture the probabilistic information that is embedded in a structural equation model.
- Shoule be able to predict patterns of independencies in the data, based solely on the structure of the modelâ s graph, without relying on any quantitative information carried by the equations or by the distributions of the errors

Chains

- Figure 2.1
  1. Z and Y are dependent
  2. Y and X are dependent
  3. Z and X are likely dependent
  4. Z and X are independent, conditional on Y
- SCM 2.2.4 Pathological Case of Intransitive Dependence
- This configuration of variables — three nodes and two edges, with one edge directed into and one edge directed out of the middle variable — is called a *chain*.
- **Rule 1 (Conditional Independence in Chains)** Two variables, X and Y, are conditionally independent given Z, if there is only one unidirectional path between X and Y and Z is any set of variables that intercepts that path.

Forks

- Figure 2.2
  1. X and Y are dependent
  2. X and Z are dependent
  3. Z and Y are likely dependent
  4. Y and Z are independent, conditional on Y
- Why are Y and Z independent conditional on X?
- This configuration of variables — three nodes, with two arrows emanating from the middle variable — is called a *fork*.

- **Rule2 (Conditional Independence in Forks)** If a variable X is a common cause of variables Y and Z, and there is only one path between Y and Z, then Y and Z are independent conditional on X

Colliders

- Figure 2.3
  1. X and Z are dependent
  2. Y and Z are dependent
  3. X and Y are independent
  4. X and Y are dependent conditional on Z
- A *collider* node occurs when one node receives edges from two other nodes.
- Why does point 4 hold? Conditioning on a collision node produces a dependence between the nodeâ s parents
- Monty Hall problem
- **Rule 3 (Conditional Independence in Colliders)** If a variable Z is the collision node between two variables X and Y, and there is only one path between X and Y, then X and Y are unconditionally independent but are dependent conditional on Z and any descendants of Z

d-separation

- Is there a criterion or process that can be applied to a graphical causal model of any complexity in order to predict dependencies that are shared by all data sets generated by that graph?
- A pair of nodes are *d*-connected if there exists a connecting path between them, or *d*-separated, if there exists no such path.
- When we say that a pair of nodes are *d*-separated, we mean that the variables they represent are definitely independent; when we say that a pair of nodes are *d*-connected, we mean that they are possibly, or most likely, dependent.
- Two nodes are *d*-separated if every path between them (should any exist) is blocked.
- The paths between variables can be thought of as pipes, and dependence as the water that flows through them; if even one pipe is unblocked, some water can pass from one place to another, and if a single path is clear, the variables at either end will be dependent. However, a pipe need only be blocked in one place to stop the flow of water through it, and similarly, it takes only one node to block the passage of dependence in an entire path.

Nodes that can block a path

- If we are not conditioning on any variable, then only colliders can block a path.
- If, however, we are conditioning on a set of nodes Z, then the following kinds of nodes can block a path:
  - A collider that is not conditioned on (i.e., not in Z), and that has no descendants in Z
  - A chain or fork whose middle node is in Z.
- Definition 2.4.1 (*d*-separation)
- Example, Figure 2.7

Model testing

- *d*-separation will tell us which variables in G must be independent conditional on which other variables. Conditional independence is something we can test for using a data set.

- Example, Figure 2.9
  - Not only we know that the model is wrong, but we also know where it is wrong; the true model must have a path between W and Z_1 that is not d-separated by X
  - Finally, this is a theoretical result that holds for all acyclic models with independent errors (Verma and Pearl 1990), and we also know that if every d-separation condition in the model matches a conditional independence in the data, then no further test can refute the model. This means that, for any data set whatsoever, one can always find a set of functions F for the model and an assignment of probabilities to the U terms, so as to generate the data precisely.

Causal search

- d-separation presents several advantages over the global testing method
  - nonparametric
  - local test
- could test and reject many possible models in this way, even- tually whittling down the set of possible models to only a few whose testable implications do not contradict the dependencies present in the data set.
- some graphs have indistinguishable implication
- allows us to search a data set for the causal models that could have generated it.

The Effects of Interventions

Intervention

- Predict effects of interventions.
- Correlation is not causation.
- Randomized controlled experiment can solve this problem. But we can't control some factors. Then only observationsl study can be conducted, which is hard to untangle causal from merely correlative.
- Intervene and condition are different. Intervene changes the model structure, but condition doesn't. (Figure 3.2)
- *do*-expression and graph surgery can help solve this problem.

The Adjustment Formula

- Causal effect; adjusting for Z or controlling for Z.
- Example: Simpson's paradox, Figure 3.3, 3.4.

To Adjust or not?

- **Rule 1 (The Causal Effect Rule)** Given a graph G in which a set of variables PA are designated as the parents of X, the causal effect of X on Y is given by

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, PA = z)P(PA = z)$$

where z ranges over all the combinations of values that the variables in PA can take. If we multiply and divide the right hand side by the probability $P(X = x|PA = z)$, we get a more convenient form:

$$P(y|do(x)) = \sum_z \frac{P(X = x, Y = y, PA = z)}{P(X = x|PA = z)}$$

- It is possible to use graphs and underlying assumptions, we are able to identify causal relationships in purely obervational data.
- In most practical cases, the set of Xâ s parents (PA(X)) will contain unobserved variables that would prevent us from calculating the conditional probabilities in the adjustment formula. Solution: adjust other variables to substitute for the unmeasured elements of PA(X).

Multiple Interventions and the Truncated Product Rule

- *Truncated product formula* or *g-formula*:

$$P(x_1, x_2, \ldots, x_n | do(x)) = \prod_i P(x_i | pa_i)$$

  for all $X_1, X_2, \ldots, X_n$ not in X.

- $$P(x_1, x_2, \ldots, x_n | do(x)) = \frac{P(x_1, x_2, \ldots, x_n, x)}{P(x|pa)}$$

The Backdoor Criterion

- Under what conditions, is the structure of the causal graph sufficient for computing a causal effect from a given data set? The rest of this chapter will focuse on this problem.
- **The Backdoor Criterion** Given an ordered pair of variables (X,Y) in a directed acyclic graph G, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X. (when conditioned on Z)
- If a set of variables Z satisfies the backdoor criterion for X and Y, then the causal effect of X on Y is given by the formula:

$$P(Y = y | do(X = x)) = \sum_z P(Y = y | X = x, Z = z) P(Z = z)$$

  just as when we adjust for PA(X). (Note that PA(X) always satisfies the backdoor criterion.) Example Figure 3.3
- In general, we would like to condition on a set of nodes Z such that

1. We block all spurious paths between X and Y.
2. We leave all directed paths from X to Y unperturbed.
3. We create no new spurious paths.

- *Effect modification* or *moderation*. Find the causal effect when we condition on some variable. Example, Figure 2.8; 3.6.

The Front-Door Criterion (Example, Smoking and Lung Cancer)

- **Front-Door** A set of variables Z is said to satisfy the front-door criterion relative to an ordered pair of variables (X, Y) if

1. Z intercepts all directed paths from X to Y.
2. There is no unblocked path from X to Z.
3. All backdoor paths from Z to Y are blocked by X.

- **Front-Door Adjustment** If Z satisfies the front-door criterion relative to (X, Y) and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by the formula:

$$P(y | do(x)) = \sum_z P(z|x) \sum_{x'} P(y | x', z) P(x')$$

Conditional Interventions and Covariate-Specific Effects

- Interventions may involve dynamic policies in which a variable X is made to respond in a specified way to some set Z of other variablesâ say, through a functional relationship $x = g(z)$ or through a stochastic relationship, whereby X is set to x with probability $P^*(x|z)$.
- The result of implementing such a policy is a probability distribution written $P(Y = y|do(X = g(Z)))$, which depends only on the function g and the set Z of variables that drive X.
- **Rule 2** The z-specific effect $P(Y = y|do(X = x), Z = z)$ is identified whenever we can measure a set S of variables such that $S \cup Z$ satisfies the backdoor criterion. Moreover, the z-specific effect is given by the following adjustment formula:

$$P(Y = y|do(X = x), Z = z) = \sum_s P(Y = y|X = x, S = s, Z = z)P(S = s|Z = z)$$

- To compute $P(Y = y|do(X = g(Z)))$, we condition on $Z = z$ and write:

$$\begin{aligned} P(Y = y|do(X = g(Z))) &= \sum_z P(Y = y|do(X = g(Z)), Z = z)P(Z = z|do(X = g(Z))) \\ &= \sum_z P(Y = y|do(X = g(z)), Z = z)P(Z = z) \\ &= \sum_z P(Y = y|do(X = x), z)|_{x=g(z)}P(Z = z) \end{aligned}$$

Inverse Probability Weighing

- Practical difficulties: adjusting for Z but Z contains too many variables.
- Assuming that the function $P(X = x|Z = z)$ is available to us, we can use it to generate artificial samples that act as though they were drawn from the postintervention probability P_m, rather than $P(x, y, z)$.
-

$$\begin{aligned} P(y|do(x)) &= \sum_z P(Y = y|X = x, Z = z)P(Z = z) \\ &= \sum_z \frac{P(Y = y|X = x, Z = z)P(X = x|Z = z)P(Z = z)}{P(X = x|Z = z)} \\ &= \sum_z \frac{P(X = x, Y = y, Z = z)}{P(X = x|Z = z)} \end{aligned}$$

Mediation

- Causation: direct and indirect (through mediating variables).
- Seperate direct and indirect effects: condition on mediating variables. Example, Figure 3.11, 3.12
- Intervene. For any three variables X, Y, and Z, where Z is a mediator between X and Y, *Controlled direct effect* (CDE) on Y of changing the value of X from x to x' is defined as:

$$CDE = P(Y = y|do(X = x), do(Z = z)) - P(Y = y|do(X = x'), do(Z = z))$$

  Example Figure 3.12.
- In general, the CDE of X on Y, mediated by Z, is identifiable if the following two properties hold:

1. There exists a set S_1 of variables that blocks all backdoor paths from Z to Y.
2. There exists a set S_2 of variables that blocks all backdoor paths from X to Y, after deleting all arrows entering Z (intervene on Z).

Causal Inference in Linear Systems

- Assumptions used in this section:

1. the relationships between variables are linear.
2. all error terms have Gaussian (or â normalâ) distributions.

Structural versus Regression Coefficients

- A regression equation is descriptive; it makes no assumptions about causation.
- We use Greek letter (α, β and so on) for structural coefficients and r_i for regression coefficients. U_1 for error term in structural equations and ϵ_i for those in regression equations.

The Causal Interpretation of Structural Coefficients

- In a linear system, every path coefficient stands for the direct effect of the independent variable, X, on the dependent variable, Y.
- In a linear system, the total effect of X on Y is simply the sum of the products of the coefficients of the edges on every nonbackdoor path from X to Y. Example, Figure 3.13.

Identifying Structural Coefficients and Causal Effect

- Total effect: First, we find a set of covariates Z that satisfies the backdoor criterion from X to Y in the model. Then, we regress Y on X and Z. The coefficient of X in the resulting equation represents the true causal effect of X on Y. Example, Figure 3.14.
- Direct effect: In a linear system, this direct effect is the structural coefficient α in the function $y = \alpha x + \beta z + \cdots + U_Y$ that defines Y in the system.
- Direct effect (from data):

1. First, we remove the edge from X to Y (if such an edge exists), and call the resulting graph G_α. If, in G_α, there is a set of variables Z that d-separates X and Y, then we can simply regress Y on X and Z. The coefficient of X in the resulting equation will equal the structural coefficient α. Example, Figure 3.15, 3.16.
2. If there is no set of variables that d-separates X and Y in G_α, we can use total effects to identify direct effects. *Instrumental variable*: it is d-separated from Y in G_α and, it is d-connected to X. Example, Figure 3.17.

- To estimate a given effect, all we need to do is to write down a regression equation and specify:

1. what variables should be included in the equation and
2. which of the coefficients in that equation represents the effect of interest.

Counterfactuals and Their Applications

Counterfactuals

- While driving home last night, I came to a fork in the road, where I had to make a choice: to take the freeway ($X = 1$) or go on a surface street named Sepulveda Boulevard ($X = 0$). I took Sepulveda, only to find out that the traffic was touch and go. As I arrived home, an hour later, I said to myself: "Gee, I should have taken the freeway."
- This kind of statement, an "if" statement in which the "if" portion is untrue or unrealized is known as a counterfactual
- The "if" portion of a counterfactual is called the hypothetical condition, or more often, the antecedent

- We use counterfactuals to emphasize our wish to compare two outcomes under the exact same conditions while differing only in one aspect: the antecedent.
- Writing $\mathbb{E}[\text{driving time}|do(\text{freeway}), \text{driving time} = 1 \text{ hour}]$ leads to a clash between the driving time we wish to estimate and actual driving time observed.

-To avoid clash, distinguish symbolically between:

```
1. Actual driving time
2. Hypothetical driving time under freeway conditions when actual surface driving time is known to be 1
```

- To do so, use different subscripts to label two outcomes
  - denote freeway driving time by $Y_{X=1}$ (or Y_1)
  - denote Sepulveda driving time by $Y_{X=0}$ (or Y_0)
- Since Y_0 is the Y actually observed, we wish to estimate $\mathbb{E}[Y_{X=1}|X = 0, Y = Y_0 = 1]$.

Defining and Computing Counterfactuals

The Structural Interpretation of Counterfactuals

- Begin with fully specified model M, for which functions $\{F\}$ and values of exogenous variables are all known.
- Simple causal model consisting of three variables: X, Y, U defined by:
  1. $X = aU$
  2. $Y = bX + U$
- Computing the counterfactual $Y_x(u)$, that is what Y would be had X been x in situation $U = u$
- Replacing first equation with $X = x$ gives "modified" model M_x:
  1. $X = x$
  2. $Y = bX + U$
- Substituting $U = u$, and solving gives $Y_x(u) = bx + u$
- We can also examine counterfactual $X_y(u)$, that is, what X would be had Y been y in situation $U = u$.
- Replacing second equation by constant $Y = y$ and solving for X, we have $X_y(u) = au$, which means X remains unaltered by the hpothetical condition "had Y been y"
- Each SCM encodes within it many counterfactuals corresponding to various values its variables can take.
- **Note**: We compute not just the probability of expected value of Y under one intervention or another, but the **actual value of Y** under the hypothesized *new* condition $X = x$. Every structural equation model assigns a definitive value to every conceivable counterfactual.

The Fundamental Law of Counterfactuals

- Consider any arbitrary two variables X and Y, not necessarily connected by a single equation. Let M_x stand for the modified version of M, with the equation of X replaced by $X = x$. The formal definition of the counterfactual $Y_x(u)$ reads:

$$Y_x(u) = Y_{M_x}(u) \tag{1}$$

- In words: The countervactual $Y_x(u)$ in model M is defined as the solution for Y in the 'surgically modified' submodel M_x.

- Eq.(**??**) is a fundamental principle of causal inference allowing us to answer questions of the type "what would Y be had X been x?"
- Counterfactuals obey the **consistency rule**: *if* $X = x$ *then* $Y_x = Y$
  - For example, if X is binary, then consistency rule takes form $Y = XY_1 + (1 - X)Y_0$. This can be interpretted as Y_1 is equal to the observed value of Y whenever X takes the value 1.

From Population Data to Individual Behavior: An Illustration

- Fig. 4.1 represents an 'encouragement design' (p.94)
  - X is amount of time a student spends in an after-school remedial program
  - H is the amount of homework a student does
  - Y is a student's score on exam
  - Value of each variable is given as number of standard deviations above the mean the student falls
- Model 4.1:

$$\begin{aligned} X &= U_X \\ H &= aX + U_H \\ Y &= bX + cH + U_Y \\ \sigma_{U_i U_j} &= 0 \quad \forall i, j \in \{X, H, Y\} \end{aligned}$$

- Assume all U factors are independent, and we are given the values of coefficients:

$$a = 0.5, \quad b = 0.7, \quad c = 0.4$$

- Consider Joe for whom we measure:

$$X = 0.5, \quad H = 1, \quad Y = 1.5$$

- We ask the question: 'What would Joe's score have been had he doubled his study time?'
  - Use evidence $X = 0.5, \quad H = 1, \quad Y = 1.5$ to determine the values of the U variables associated with Joe
  - These values are invariant to hypothetical actions
- Obtaining the specific characteristics of Joe from evidence:

$$\begin{aligned} U_X &= 0.5 \\ U_H &= 1 - 0.5 \cdot 0.5 = 0.75 \\ U_Y &= 1.5 - 0.7 \cdot 0.5 - 0.4 \cdot 1 = 0.75 \end{aligned}$$

- Simulating the action of doubling Joe's study time by replacing the structural equation for H with the constant $H = 2$ gives modified model in Fig. 4.2. (p.95). Compute the value of Y in modified model using updated U values.

$$\begin{aligned} Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) \\ &= 0.5 \cdot 0.7 + 2.0 \cdot 0.4 + 0.75 \\ &= 1.90 \end{aligned}$$

- Had Joe doubled his homework, his score would have been 1.9 instead of 1.5

- In summary we applied evidence to update the values of the U variables, simulate external intervention to foce condition by replacing structural equation with constant, then compute value of Y given new structural equations and the updated U values.

The Three Steps in Computing Counterfactuals

- The three-step process for computing any deterministic counterfactual:

  1. Abduction: Use evidence $E = e$ to determine the value of U
  2. Action: Modify the model, M, by removing the structural equations for the variables in X and replacing them with the appropriate functions $X = x$, to obtain the modified model, M_x
  3. Prediction: Use the modified model, M_x, and the value of U to compute the value of Y, the consequence of the counterfactual

- The above process will solve any deterministic counterfactual, that is, counterfactuals pertaining to a single unit of the population in which we know the value of every relevant variable.

- Counterfactuals can also be probabilistic, that is pertaining to a class of units within the population.

  - A typical query might be: 'Given that we observe feature $E = e$ for a given individual, what would we expect the value of Y for that individual to be if X had been x?'
  - This expectation is denoted $\mathbb{E}[Y_{X=x}|E = e]$, where $E = e$ is allowed to conflict with the antecedent $X = x$
  - We can generalize the three-step process to any probabilistic nonlinear system.

- Given an arbitrary counterfactuals of the form $\mathbb{E}[Y_{X=x}|E = e]$, the three-step process reads:

  1. Abduction: Update $P(U)$ by the evidence to obtain $P(U|E = e)$
  2. Action: Modify the model, M, by removing the structural equations for the variables in X and replace them with the appropriate functions $X = x$ to obtain the modified model M_x
  3. Use the modified model, M_x, and the updated probabilities over the U variables, $P(U|E = e)$, to compute the expectation of Y, the consequence of the counterfactual

Nondeterministic Counterfactuals

Probabilities of Counterfactuals

- Refer to Eqs. (4.3) and (4.4) on p. 92

- Imagine that $U = \{1, 2, 3\}$ represents three types of individuals in a population, occurring with probabilities

$$P(U = 1) = \frac{1}{2}, \quad P(U = 2) = \frac{1}{3}, \quad P(U = 3) = \frac{1}{6}$$

- All individuals within a population type have the same values of the counterfactuals as specified by the rows in Table 4.1 (p.93). Using these values, it is possible to compute the probability that the counterfactual will satisfy a specific condition.

- For example: We can compute the proportion of units for which Y would be 3 had X been 2, or using notation $Y_2(u) = 3$.

  - Refer to table to see that $P(Y_2 = 3) = \frac{1}{2}$
  - Using the table, how about $P(Y_2 > 3)$? (Answer: $\frac{1}{2}$)

- We can also compute joint probabilities of every combination of counterfactual and observable events:

  - For example: $P(Y_2 > 3, Y_1 < 4) = \frac{1}{3}$
  - This is the joint probability of two events occurring in two different 'worlds.' The firset $Y_2 > 3$ in an $X = 2$ world, and the second $Y_1 < 4$ in an $X = 1$ world.

- These joint probabilities over 'multiple-world' counterfactuals are easily expressed using subscript notation, but *cannot* be expressed using the *do(x)* notation since the latter delivers a single probability for each intervention $X = x$.

- We can also compute conditional probabilities among counterfactuals:
  - For example: Among the individuals for which Y is greater than 2, the probability is $\frac{2}{2}$ that Y would increase if X were 3 since

$$P(Y_3 > Y | Y > 2) = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

- Example p.99
- Can counterfactual notation capture the postintervention single-world expression $\mathbb{E}[Y|do(X = x), Z = 1]$?
  - Answer: Yes
  - Translating $\mathbb{E}[Y|do(X = x), Z = 1]$ into counterfactual notation yields $\mathbb{E}[Y_{X=1}|Z_{X=1} = 1]$.
    - $Z_{X=1}$ stands for the value that Z would attain had $X = 1$
- Example p. 100

The Graphical Representation of Counterfactuals

- Can we see counterfactuals within the causal graphs associated with the model?
  - Yes!
  - From the fundamental law of counterfactuals, Eq.(**??**), if we modeify model M to obtain the submodel M_x, then the outcome variable Y in the modified model is the counterfactual Y_x of the original model
  - This modification calls for removing all arrows entering the variable X
- **Theorem 4.3.1: Counterfactual Interpretation of Backdoor**: If a set Z of variables satisfies the backdoor condition relative to (X, Y), then, for all x, the counterfactual Y_x is conditionally independent of X given Z,

$$P(Y_x|X, Z) = P(Y_x|Z)$$

  - Implies that $P(Y_x = y)$ is identifiable by adjustment formula of Eq. (3.5)

$$\begin{aligned} PY_x = y) &= \sum_z P(Y_x = y|Z = z)P(z) \\ &= \sum_z P(Y_x = y|Z = z, X = x)P(z) \\ &= \sum_z P(Y = y|Z = z, X = x)P(z) \end{aligned}$$

Counterfactuals in Linear Models

- In nonparametric models, counterfactual quantities of the form $\mathbb{E}[Y_{X=x}|Z = z]$ may not be identifiable
- However, in fully linear models any counterfactual quanitity is identifiable whenever the model parameters are identified
- Question: Can counterfactuals be identified in observasional studies when some of the model parameters are not identified?
  - Yes!
  - Any counterfactual of the form $\mathbb{E}[Y_{X=x}|Z = e]$ where e is an arbitrary set of evidence, is identified whenever $\mathbb{E}[Y|do(X = x)]$ is identified

- **Theorem 4.3.2**: Let τ be the slope of the total effect of X on Y,

$$\tau = \mathbb{E}[Y|do(x+1)] - \mathbb{E}[Y|do(x)]$$

then, for any evidence $Z = e$, we have

$$\mathbb{E}[Y_{X=x}|Z=e] = \mathbb{E}[Y|Z=e] + \tau(x - \mathbb{E}[X|Z=e])$$

- Recall situation in Fig 4.2 (p. 95) where the counterfactual $Y_{H=2}$ under the evidence $e = \{X = 0.5, H = 1, Y = 1\}$ was computed. Theorem 4.3.2 can be applied to this model to comptue the *effect of treatment on the treated*:

$$ETT = \mathbb{E}[Y_1 - Y_0|X=1]$$

Substituting the evidence $e = \{X = 1\}$ into Theorem 4.3.2 (refer to model 4.1):

$$\begin{aligned} ETT &= \mathbb{E}[Y_1|X=1] - \mathbb{E}[Y_0|X=1] \\ &= \mathbb{E}[Y|X=1] - E[Y|X=1] + \tau(1 - \mathbb{E}[X|X=1]) - \tau(0 - \mathbb{E}[X|X=1]) \\ &= \mathbb{E}[Y|do(x+1)] - \mathbb{E}[Y|do(x)] \\ &= \tau \\ &= b + ac \\ &= 0.9 \end{aligned}$$

Practical Uses of Counterfactuals

Recruitment to a Program

- Example 4.4.1 (p.107)
- Critics claim program is a waste of taxpayer's money and should be terminated, why?
  - Reasoning: While program was successful in experimental study, where people chosen at random, there is no proof the program is successful among those choosing to enroll of their own volition
  - Critis say those enrolling are more intelligent, resourceful and socially connected than those who are eligible and did not enroll
  - Critics claim that we need to estimate the *differential* benefit of the program on those enrolled: the extent to which hiring rate has increasing among the enrolled, compared to what it would have been had they not been trained.
- Using counterfactual notation, letting $X = 1$ represent trianing, and $Y = 1$ represent hiring, we need to evaluate ETT (p.106 Eq. 4.18), that is $ETT = \mathbb{E}[Y_1 - Y_0|X =]$
  - Unfortunately the difference $Y_1 - Y_0$ represents the causal effect of trianing (X) on hiring (Y) for a *randomly* chosen individual, and the condition $X = 1$ limits the choice to those who actually chose the training on their own
  - This is a clash between the antecedent ($X = 0$) of the counterfactual Y_0 and the event it is condition on $X = 1$
  - $\mathbb{E}[Y_0|X = 1]$ can be reduced to estimable expressions in many situations (not all)
- When a set Z of covariates satisfies the backdoor criterion with regard to the treatment and outcome variables, ETT probabilities are given by a modified adjustment formula:

$$P(Y_x = y|X = x') = \sum_z P(Y = y|X = x, Z = z)P(Z = z|X = x') \tag{2}$$

- Comparing to standard adjustment formula $P(Y = y|do(X = x)) = \sum P(Y = y|X = x, Z = z)P(Z = z)$ both formulas require conditioning on $(Z = z)$ then averaging over z with modified adjustment formula using a different weighted average

- Using the modified adjustment formula we cna get an estimable, noncounterfactual expression for ETT

$$\begin{aligned} ETT &= \mathbb{E}[Y_1 - Y_0|X=1] \\ &= \mathbb{E}[Y_1|X=1] - \mathbb{E}[Y_0|X=1] \\ &= \mathbb{E}[Y|X=1] - \sum_z \mathbb{E}[Y|X=0, Z=z]P(Z=z|X=1) \end{aligned}$$

  - $\mathbb{E}[Y_1|X=1] = \mathbb{E}[Y|X=1]$ because conditional on $X=1$, the value that Y would get had X been 1 is just the observed value of Y

Additive Interventions

- Example 4.4.2 (p.109)
- Suppose we add a quanity q to a treatment variable X that is currently at level $X = x'$. The resulting outcome would be $Y_{x'+q}$, and the average value of this outcome over all units currently at level $X = x'$ would be $\mathbb{E}[Y_x|x']$ with $x = x' + q$.
- Whenever a set Z in our model satisfies the backdoor criterion, the effect of an additive intervention is estimable using the ETT adjustment formula Eq.(**??**). Substituting in $x = x' + q$ and taking expectations gives effect of this intervention called $add(q)$:

$$\begin{aligned} \mathbb{E}[Y|add(q)] - \mathbb{E}[Y] &= \sum_{x'} \mathbb{E}[Y_{x'+q}|X=x']P(x') - \mathbb{E}Y \\ &= \sum_{x'}\sum_z \mathbb{E}[Y|X=x'+q, Z=z]P(Z=z|X=x')P(X=x') - \mathbb{E}[Y] \end{aligned}$$

- Z my include whichever variables so long as each can be measured and together they satisfy the backdoor condition

Personal Decision Making

- Example 4.4.3 (p.111)
- Designating remission by $Y=1$ and decision to undergo radiation by $X=1$, the probability that determines whether Ms. Jones is justified in attributing her remission to the irradiation ($X=1$) is

$$PN = P(Y_0 = 0|X=1, Y=1)$$

  - It reads: the probability that remission would *not* have occurred ($Y=0$) had Ms. Jones not gone through irradiation, given that she did in fact go through irradiation ($X=1$), and remission did occur ($Y=1$)
  - The label PN stands for 'probability of necessity' and it measures the degree to which Ms Jones' decision was *necessary* for her positive outcome

- Similarly, the probability that Ms. Smith's regret is justified is given by

$$PS = P(Y_1 = 1|X=0, Y=0)$$

  - It reads: the probability that remission would have occurred had Ms. Smith gone through irradiation ($Y_1=1$), given that she did not in fact go through irradiation ($X=0$), and remission did not occur ($Y=0$)
  - PS stands for the 'probability of sufficiency', and it measures the degree to which the action not taken, ($X=1$), would have been sufficient for her recovery

- These probabilities, sometimes referred to as 'probabilities of causation', are not, in general, estimable from either observational or experimental data
  - However, under certain conditions they *are* estimable when both observational and experimental data are available.
- Imagine Ms. Daily facing the same decision as Ms. Jones and asking herself: If tumor is type that would not recur under lumpectomy alone, why go through irradiation? Similarly, if tumor is type that would recur regardless of irradiation treatment, why go through irradiation?'
  - Only go through radiation if tumor is type that would remiss under treatment and recur under no treatment
- Ms. Daily's dilemma is to quantify the probability that irradiation is both *necessary and sufficient* for eliminating her tumor,

$$PNS = P(Y_1 = 1, Y_0 = 0)$$

  - where Y_1 and Y_0 stand for remission under treatment (Y_1), and nontreatment (Y_0)
  - This probability cannot be assessed from experimental studies
  - PNS *can* be estimated if we assume *monotonicity*, namely that irradiation cannot cause the recurrence of a tumor that was about to remit
  - Under monotonicity experimental data are sufficient to conclude $$PNS = P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0))

Mediation and Path-disabling Interventions

- Example 4.4.5 (p. 114)
- Because we are dealing with disabling processes rather than changing levels of variables, there is no way we can express the effect of such interventions using a *do*-operator
- We can express it in counterfactual notation
- The hiring status ($Y = 1$) of a female applicant with qualification $Q = q$, given that the employer treats her as though she is a male is captured by the counterfactual $Y_{X=1,Q=q}$ where $X = 1$ refers to being male.
  - Since value q would vary among applicants, this quantity must be averaged according to the distribution of female qualification giving

$$\sum_q \mathbb{E}[Y_{X=1,Q=q}]P(Q = q|X = 0)$$

  - Similarly, male applicants chance at hiring has average governed by the male qualification

$$\sum_q \mathbb{E}[Y_{X=1,Q=q}]P(Q = q|X = 1)$$

  - Subtracting the two quantities gives

$$\sum_q \mathbb{E}[Y_{X=1,Q=q}][P(Q = q|X = 0) - P(Q = q|X = 1)]$$

    which is the indirect effect of gender on hiring, *mediated by qualification*.
    - This effect is called the natural indirect effect (NIE), because qualification Q is allowed to naturally vary from applicant to applicant
- Can such a counterfactual expression be identified from data?
  - Yes! In absense of confounding, the NIE can be estimated by conditional probabilities

–

$$NIE = \sum_q \mathbb{E}[Y|X=1, Q=q][P(Q=q|X=0) - P(Q=q|X=1)]$$

– Expression is known as the *mediation formula*, and measures the extent to which the effect of X on Y is *explained* by its effect on the mediator Q

Mathematical Tool Kits for Attribution and Mediation

A Tool Kit for Attribution and Probabilities of Causation

- Assuming binary events, with $X = x$ and $Y = y$ representing treatment and outcome, respectively, and $X = x'$, $Y = y'$ their negations, our target quantity can be defined as:
  - 'Find the probability that if X had been x', Y would be y', given that in reality $X = x$ and $Y = y$.'
  - Mathematically, this is
  $$PN(x, y) = P(Y_{x'} = y'|X = x, Y = y)$$
  , and called the 'probability of necessity'
  - Question: What assumptions permit us to identify PN from empirical studies (observational and/or experimental)?
- **Theorem 4.5.1**: If Y is monotonic relative to X, that is, $Y_1(u) \geq Y_0(u)\ \forall u$, then PN is identifiable whenever the causal effect $P(y|do(x))$ is identifiable, and

$$PN = \frac{P(y) - P(y|do(x'))}{P(x, y)} \qquad (4.28) \qquad (3)$$

or, substituting $P(y) = P(y|x)P(x) + P(y|x')(1 - P(x))$, we obtain

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x, y)} \qquad (4.29) \qquad (4)$$

  - First term on r.h.s. is called *excess risk ratio* (ERR)
  - Second term (the confounding factor) represents a *correction* needed to account for confounding bias since $P(y|do(x')) \neq P(y|x')$.
- Example: Suppose there is a case brought against a car manufacturer, claiming that its car's faulty design led to a man's death in a car crash. The ERR teslls us how much more likely poeple are to die in crashes when driving one of the manufacturer's cars. If it turns out that people who buy the manufacturer's cares are more likely to drive faster than the general poipulation, the second term will correct for this bias.
- Eq.(4.29) provides an estimable measure of necessary causation which can be used for monotinic $Y_x(u)$ whenever the causal effect $P(y|do(x))$ can be estimated.
- Eq.(4.28) provides bounds for PN in the general *nonmonotonic* case

$$max\{0, \frac{P(y) - P(y|do(x'))}{P(x, y)}\} \leq PN \leq min\{1, \frac{P(y'|do(x')) - P(x', y')}{P(x, y)}\}$$

  - LB = ERR + CF
  - UB = ERR + q + CF
  - where