Causal inference with Bayesian regression modeling: A tutorial

Lena Holzwarth & Michael Franke

This tutorial provides a first general introduction to causal inference with the *do*-calculus, as introduced in the work of Judea Pearl and a hands-on programming example in R for an artificial data set instantiating *Simpson's paradox*, showcasing the use of Bayesian regression modeling to obtain causal effect estimates with quantified uncertainty.

1 *Motivation & intended audience*

This tutorial has two goals: First, to introduce the reader to the concepts behind causal analysis with the *do*-calculus (Pearl, 2000), and second, to show how Bayesian regression modeling can be used for estimates of causal effects with quantified uncertainty for these causal effect estimates.

Different types of readers might profit from this tutorial. If you are new to causal inference with the *do*-calculus, the conceptual part of this tutorial provides a basic introduction to the intuitions at the core of the *do*-calculus, while the practical part offers a simple programming example in R to illustrate the application to a toy research question (R Core Team, 2024). If you are already familiar with the concepts behind the *do*-calculus, you might want to skip to the practical application directly. The tutorial will assume basic familiarity with R and regression modeling. For the implementation of Bayesian regression modeling, we will use the `brms` package (Bürkner, 2021).

If you're unfamiliar with Bayesian regression modeling, Franke and Roettger (2019) offer a beginner's tutorial similar to this one.

2 *Searching for Causality*

CORRELATION DOES NOT IMPLY CAUSATION! In scientific education, this is often one of the first lessons in statistics. We are told, time and again, that we should be weary of interpreting data collected by observation. If we want to establish a causal relationship, we are taught that there is one gold standard: the randomized experiment. However, it may not always be possible to collect data from a randomized and controlled experiment. For some research, randomized experiments are too expensive, too unethical, or downright impossible. Maybe we only have data from previous research and we don't know even know exactly how exactly it was obtained. Even if we are able to perform a randomized experiment, the decision of which variables should be included in, say, regression analysis (good controls) and which should not be (bad controls) arguably requires reasoning about the likely causal data-generating process, and it can be facilitated by an explicit representation of a causal model of the kind introduced below (Cinelli, Forney, and

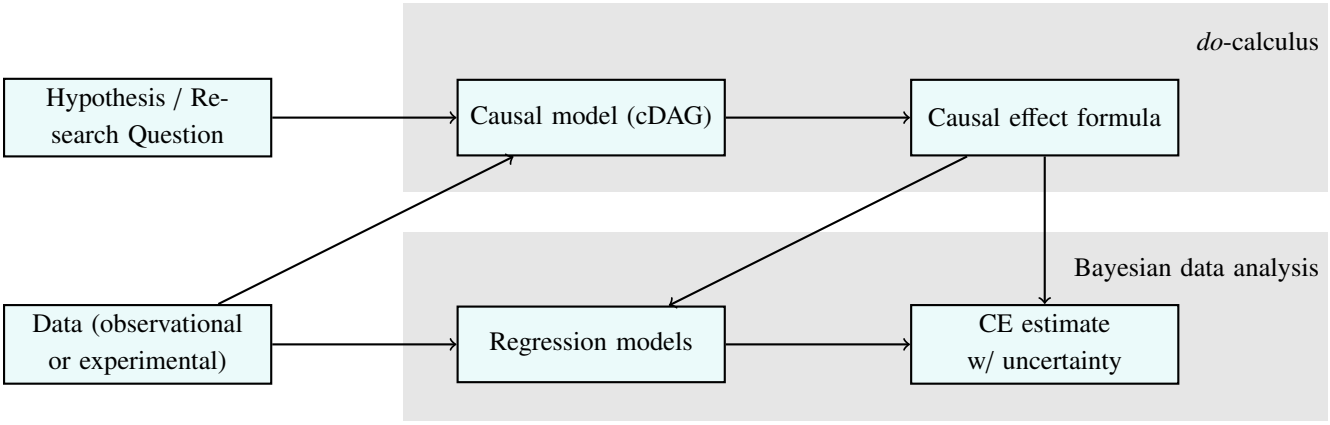


Figure 1: Causal effect estimation with the *do*-calculus and Bayesian data analysis.

Pearl, 2022).

There are multiple methods to estimate causal effects from any given data set, irrespective of whether it was obtained by experimental manipulation or not. This tutorial covers one of them: the *do*-calculus.

The process of causal effect estimation using the *do*-calculus, as introduced in this tutorial, is illustrated in Figure 1. Based on our data and research goals, we use a *causal model* to explicate our assumptions about the relevant causal variables and causal processes that are likely involved in the data-generating process. Using the *do*-calculus we derive a *causal effect formula* for calculating the causal effect we are interested in. To estimate the probabilities necessary for computing the causal effect, we use our data and (Bayesian) regression modeling. In a final step, we use the causal effect formula and the usual sampling-based uncertainty calculation familiar from Bayesian data analysis to obtain a causal effect estimate with quantified uncertainty.

The tutorial is structured as follows. Section 3 introduces a fictitious data set that will serve as our running example for this tutorial. The data set instantiates a case of *Simpson's paradox*. In Section 4, the theoretical background to the *do*-calculus is explained and we use a maximum-likelihood approach to estimate the relevant causal effects for our running example. Section 5 explains how we can obtain estimates of causal effects with quantified uncertainty from Bayesian regression models. Section 6 reflects what this tutorial achieved and gives references for where to continue learning about causal inference.

While this tutorial is not concerned with historical background or direct comparison of the *do*-calculus to other approaches, Info Box 2 at the end of this tutorial briefly describes prominent alternatives, namely the *potential outcomes framework* and *structural equation modeling*. The information from Info Box 2 is not required for the remainder of the tutorial.

3 *Fictitious data & Simpson's paradox*

To guide you through the steps of estimating causal effects with the *do*-calculus, we'll look at a concrete numerical example. Let's assume that we

work in a hospital and want to test the effect of a new drug. For some reason or other, we cannot simply administer the drug randomly to patients. Instead, we ask them if they would be willing to test the drug. In our experiment, we have a random variable (RV) DRUG INTAKE with two possible values: TAKE, or numerical value 1, for the participants consenting to test the drug and REFUSE, or numerical value 0, for those who declined. The patients are also sorted into two groups (more on the nature of these groups below). The concrete numerical data we will work with are shown in Table 1.

	REFUSE		TAKE	
GROUP 1	$\frac{234}{270}$	(87%)	$\frac{81}{87}$	(93%)
GROUP 2	$\frac{55}{80}$	(68%)	$\frac{192}{263}$	(73%)
Σ	$\frac{289}{350}$	(83%)	$\frac{273}{350}$	(78%)

Table 1: Recovery rates after refusing or taking the drug

The distribution of recovery rates is taken from a real-life medical study (Charig et al., 1986) and was first discussed in relation to Simpson's Paradox by Julious and Mullee (1994). The hypothetical experimental setup and the labels of the RVs come from Pearl (2000).

How are we to interpret the data in Table 1 regarding the research question of whether the drug was effective, i.e., whether the rate of recovery increased when taking it? On the one hand, we see that overall, 83% of patients who refused the drug recovered from their disease, while only 78% of the drug-takers did. This does not bode well for our drug. However, on the other hand, when we look at each group in isolation, we see that within each group, the drug-takers had a slightly higher recovery rate. This phenomenon is an instance of *Simpson's Paradox*: Splitting the participants into sub-groups yields different results compared to analyzing all data at once. This makes the interpretation of our results very difficult: should we use the group-wise results and conclude that the drug is helpful, or should we use the overall results which indicate that the drug is harmful? Without further information, specifically on how the groups were formed, we can't take this decision. This illustrates an important point for statistical analysis: the data alone is not enough to reach conclusions. The data can't speak for itself. We need to be data-literate, so to speak, and bring our knowledge about where the data comes from to bear.

Let's see how adding information about the groups helps our effort to interpret the results. We consider two scenarios. In the first scenario, the participants were grouped by gender (which is treated as a binary variable for the purposes of this example). GENDER could be relevant because the drug might have different effects on the male and female body. However, GENDER might also influence the decision to take the drug, e.g., by one gender being more risk averse than the other. With the updated group names, the results are displayed in Table 2 (a). A common intuition is that the drug is likely beneficial, because for men and women alike, the chances of recovery are higher for those who took the drug.

Table 2 (b) shows a second scenario for the same numerical values. Here, the participants were not grouped by gender, but by their blood pressure as measured *after* taking the drug. Participants are divided into groups with high or low blood pressure. In this case, our intuition is likely different. Because blood pressure is measured after drug administration and is possibly influenced by the drug, it seems that the groupings into HIGH and LOW are not as informative, and we should instead focus on the overall results, that tell us that the drug might even be harmful.

(a) Recovery rates for different genders

	REFUSE		TAKE	
MEN	$\frac{234}{270}$	(87%)	$\frac{81}{87}$	(93%)
WOMEN	$\frac{55}{80}$	(68%)	$\frac{192}{263}$	(73%)
Σ	$\frac{289}{350}$	(83%)	$\frac{273}{350}$	(78%)

(b) Recovery rates for different values of blood pressure

	REFUSE		TAKE	
LOW	$\frac{234}{270}$	(87%)	$\frac{81}{87}$	(93%)
HIGH	$\frac{55}{80}$	(68%)	$\frac{192}{263}$	(73%)
Σ	$\frac{289}{350}$	(83%)	$\frac{273}{350}$	(78%)

Table 2: Changing the group names changes our intuitions about the causal effect.

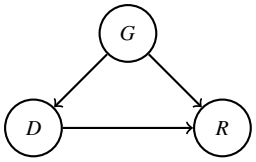
4 *Causal inference with the* do*-calculus*

The previous considerations suggest that our intuitions about the correct conclusions to be drawn from experimental data rely on intuitive knowledge of the likely causal role of the relevant measured quantities. The first step of causal inference is therefore to make such intuitive knowledge explicit (shareable, transparent, open to constructive criticism), e.g., in a formal *causal model*, a version of which is introduced in Section 4.1. The next step is to determine, based on a causal model, what the effects of *intervening* on a variable would be, as described in Section 4.2. This allows us to specify a formula with which to calculate a notion of causal effect (Section 4.3), which we can then estimate from the data (Section 4.4).

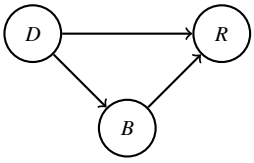
4.1 *Causal models*

Why do we interpret the data from our two scenarios differently, even though the numbers are exactly the same? It is because the data is not the only information we have: We also know how this data was obtained and how the variables are most likely related. We know that GENDER (G) might have an influence on the decision of DRUG INTAKE (D), while the taking of the drug doesn't change one's gender. In the case of BLOOD PRESSURE (B), we know that it was measured after taking or refusing the drug. It therefore shouldn't have influenced the decision to take the drug. However, it is possible that the drug has an influence on blood pressure. In both cases, we hypothesize that the drug may have an effect on the patients' RECOVERY (R). Before starting with our analysis, we should make these causal relationships explicit.

One way to make intuitions about causal dependency explicit, is to represent then in a causal model, such as *causal directed acyclic graph* (cDAG). The nodes in a cDAG represent events, measurements or variables relevant for our data analysis. We draw an arrow from one node to another to mark the assumption the former is a direct cause of the latter. The precise details of cDAGs, their relation to causality and stochastic (in-)dependence, are im-



cDAG for Scenario 1: Gender as confound



cDAG for Scenario 2: Blood pressure as mediator

Figure 2: Different cDAGs capture our intuitions for different causal roles in the two scenarios.

portant for advanced applications of causal inference with the *do*-calculus. As the motivation for this tutorial is to provide a first conceptual feeling of what can be done in this framework, Info Box 1 covers only some of the basics of causal models in a non-technical way.

The idea of prior knowledge influencing the analysis is familiar to Bayesians. However, there is one difference: in Bayesian statistics, the influence of the prior over model parameters decreases as the size of the observed data increases. The *a-priori* causal assumption always influence the outcome.

Figure 2 shows the two cDAGs that capture the most common intuitions about causal relations for the two scenarios of Simpson's paradox. These causal models are considered *a-priori* assumptions about all the relations of direct causation about the variables involved. Of course, this approach bears some risks. If we assume the wrong causal relationships, our resulting analysis will also yield wrong results. However, it is important to note that this is the case for all statistical inference methods. Any decision on including or excluding a certain variable from statistical analysis is implicitly an assumption about the causal structure connecting the RVs. Formulating a specific cDAG can thus improve accountability by making the underlying assumptions explicit and easily understandable.

"All causal inferences based on statistical models are implicitly based on a causal structure" (Shrier and Platt, 2008)

The construction of a causal model and the question of which variables to include are challenging tasks, especially because in many application cases, the causal structures under observation will be much more complicated than the limited ones given here. In many cases, it might be impossible to decide between two possible cDAGs. In these instances, Shrier and Platt (2008) suggest to perform the analyses on all plausible versions of the DAG. If the different versions yield different results, all results should be presented. In their words:

> *Not using the causal approach because of uncertainty on which is the correct DAG simply means that one is allowing chance rather than rational deliberation to make the choice among the different causal diagrams.*

From a Bayesian point of view, we can also assign a prior probability to all candidate cDAGs and present, in addition to the estimates of causal effects for each case, a prior-weighted average, similar to other cases of model averaging in Bayesian analysis.

Once we have settled on one (or several) causal models, the base for our causal analysis is set, and we can continue with the second step: the intervention.

4.2 Intervention

The key to inferring causation is intervention. This is what constitutes the power of the controlled randomized trial setup (if done well) in experimental research. But what if we cannot, for whatever reason, interfere in the data-generating process? In that case, we can use our causal model to calculate the effects of a "simulated intervention" (this section), which we can use to define how to calculate the causal effect of interest (the next Section 4.3). Then, as Pearl has argued prominently, it may be possible to calculate the effects of simulated interventions (and thereby the relevant causal effect) from empirical data from processes that did not involve this intervention (Section 4.4)

Directed acyclic graphs (DAGs) for causal analysis

DAGs can be used to represent causal relationships between *variables*. Each variable X can take on any of a set of *values*. For example, the variable D for drug-taking can take on values 1 (taken) or 0 (refused). The nodes in a causal DAG (cDAG) represent the variables and a directed edge from X to Y indicates that X is a direct cause of Y. The representation of causal relationships in a cDAG is conventionally interpreted as exhaustive: it contains *all* relations of direct causation; when X and Y are not connected, there is *no* direct causal relation between them.

A complete assignment of values to each variable is a complete description of the relevant facts, a so-called *world*. Causal relations influence the assignment of values to variables either deterministically or probabilistically. In *structural equation models* (SEMs), the value of a variable X is completely determined by the set of values assigned to X's direct causes, also called X's *parents*. In *probabilistic SEMs*, the value of X is completely determined by two things: (i) as before, the set of values of X's parents, and, additionally, (ii) the value of a stochastic *error variable* U_X, of which we have one fore each variable in the causal model. A special case of probabilistic SEMs are *Bayes Nets* which dispense with the error variables. In a Bayes Net each variable X is associated with a probability distribution over X's values, which is a conditional probability for each set of values of X's parents.

By design, probabilistic causal models relate facts about causation to facts about stochastic independence of variables. This relationship can be complex, but it can help identify causal structure behind empirical data (*causal discovery*). It can be systematically assessed by decomposing the cDAG's structure into a small number of configurations. Two of these, relevant for our running example, are shown below: the chain and the fork.

chain (mediator) *fork (confound)*

In a chain, X and Y are causally dependent: X influences the value of the mediator Z, which influences Y. However, X and Y are independent conditioned on Z. Intuitively put, if Z is a mediator, conditioning on Z undermines the signal about the (indirect) causal influence of X on Y.

In a fork, Z acts as a confound on X and Y. X and Y are dependent, because they have a common cause in Z. X and Y become independent conditioned on Z. Roughly speaking, if Z is a confound, conditioning on Z is necessary to decouple stochastic dependence from information about (indirect) causation. For a visual explanation of these causal structures see this lecture by Richard McElreath.

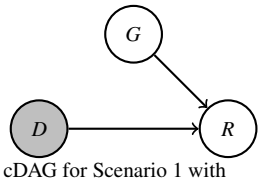
Info Box 2: Background on visually displaying causal relationships.

When we simulate an intervention, we do *as if* we had manipulated the relevant variable, which means that we cut off all influence from upstream causes of that variable, thus possibly creating a modified (pruned, lesioned) cDAG. We then use the data to estimate the effect that different values of the relevant variable have in this modified cDAG. In our running example, we want to vary the variable DRUG INTAKE (D), to measure the effects this has on RECOVERY (R). The causal effect of the drug can be estimated by *pretending* that the experimenters had randomly assigned taking or not taking the drug, or placebos to the patients and then comparing the recovery outcomes of the two groups. In this fictitious scenario, D is the result of random manipulation and thus has no causal ancestors.

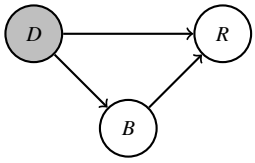
Handling such simulated interventions is the main job description of the *do*-calculus. The *do*-calculus introduces a new operator, the *do*-operation. We can now write *do*-operation $do(D = d)$ to mean that we set the variable D to a fixed value $d \in \{0, 1\}$ by intervention. Because the intervention overwrites the normal causal structure, the *do*-operation removes any causal influences on D. In the graphical representation, this means that we can remove any incoming arrows for the variable that we intervene on. In the cDAG for Scenario 1, $do(D = d)$ leads to the removal of the causal influence of G on D. In Scenario 2, the *do*-operation has no effect at all, because there were no incoming arrows on D to begin with. This is illustrated in the updated cDAGs in Figure 3.

Let's pause here for a second and reflect. We want to mimic the effects of something like an experimental intervention in a randomized control trial. The main property that makes such experimental manipulation effective is that the relevant variable (here: D) was not allowed to just happen, so to speak. Some god-like external force (the experimenters) messed up the normal causal flow, interrupted it and just set the variable to whatever they liked, no matter how likely it might have happened on its own had they not intervened on the normal courses of events. That is, essentially, the main property that makes experimental manipulation potentially informative about causation. And it is exactly this (cutting off the prior causal influences on the relevant variable (D)) that the *do*-calculus implements. In other words, the *do*-calculus is not some crazy math-voodoo but the most straightforward implementation of what we intuitively consider to be (the power of) systematic experimental manipulation.

In the general case, it is also possible to intervene on continuous variables that can take infinitely many values.



cDAG for Scenario 1 with $do(D = d)$; $d \in \{0, 1\}$ with the arrow from G to D removed



cDAG for Scenario 2 with $do(D = d)$; $d \in \{0, 1\}$

Figure 3: Intervening on a variable in a cDAG amounts to removing all influence from any causally upstream variable.

4.3 *Causal effect formula*

Given an explicit causal model and a way to represent the idea of an intervention on a variable, we can define what we mean when we speak of a *causal effect*. Actually, the term "causal effect" does not have a single, clear and crisp intuitive meaning. However, with cDAGs and the *do*-calculus, we can clearly and formally define a number of notions of causal effects. Here we

focus on the most salient notion. The *average causal effect* of variable X on variable Y is the expected amount of increase of Y given that X is changed (by intervention). More concretely, for the case of two binary variable such as recovery R and drug use D, the average causal effect describes the increase in the probability of recovery $P(R = 1)$ when going from $do(D = 0)$ to $do(D = 1)$, i.e., going from a situation where we intervened so that the drug was *not* given to a situation where we intervened so that the drug *was* given. In a formula, the average causal effect for our running example is:

$$\begin{aligned}\text{Avg-CE}(P \mid D) = P(R = 1 \mid do(D = 1)) - \\ P(R = 1 \mid do(D = 0))\end{aligned}$$

4.4 *Causal effect estimation*

After defining the causal model to work with and the causal effect of interest, the real challenge arises, namely to calculate the causal effect, if possible, for the given data at hand. This raises deep conceptual questions: How do we get from virtual interventions to a causal effect estimation? And why are we allowed to pretend that an intervention has happened, when the data we have are obtained from a causal process without intervention? Instead of a general but abstract explanation, the following explores one concrete example of how to obtain a *maximum likelihood estimate* (MLE) of the causal effect in our two scenarios. In doing so, we explain the mathematical intuition behind why we can (in some cases) treat the data as if it was the product of experimental manipulation. We look at each scenario in turn.

MLE for Scenario 1: gender as confound. As described above, the key to understanding the *do*-calculus is to think of a "simulated intervention" as a thought experiment: how would the causal relations be if we had actually intervened on the relevant variable? In this imagined scenario, we would have seen data generated by a joint probability distribution P^* over all relevant variables. Of course, the actual data we have is *not* a sample from P^*, but rather from a joint probability distribution P, which follows the dependencies of the assumed cDAG. So, we would like to use the data, which actually comes from P, to estimate —if possible— aspects of P^*. As we said before, sometimes this cannot possibly work; but sometimes it can. Here is an example of the latter.

Suppose we have data from scenario 1 as in Table 2a, which we assume is generated from joint probability distribution P. In the cDAG that (by our working assumption) underlies P, variable D was merely observed, not manipulated. However, we can imagine a case where the joint probability distribution on variables follows P^* instead, where D was actually experimentally manipulated, so that all upstream causal influences are suppressed. We now introduce a new piece of notation to write $P\left(R = r \mid do(D = d)\right)$

for the conditional probability of recovery R given that the drug administration D had been manipulated. In effect, the conditional probability $P\left(R = r \mid do(D = d)\right)$ with an imaginary intervention is defined as the conditional probability $P^*\left(R = r \mid D = d\right)$ under the unobserved probabilities P^*. You can think of the *do*-operation as a kind of jumping board to get from the actual probability distribution P to the imagined P^*. However, this "fictitious" conditional probability cannot always be calculated based on P without further ado, because it assumes that the data-generating process was a different one, namely one that follows the joint distribution P^*. This makes the *do*-operation seem kind of pointless, like notation juggling or just some renaming trick. But the good news is that, in some cases, the data obtained from process P can be used to safely estimate all the relevant aspects of P^*, so that we can eventually estimate the desired $P\left(R = r \mid do(D = d)\right)$ entirely from the non-interventionist data at hand.

As motivated above, by definition we have:

$$P\left(R = r \mid do(D = d)\right) = P^*\left(R = r \mid D = d\right)$$

Since the right-hand side of the previous equation has no *do*-operation, the usual rules of probability apply:

$$P^*\left(R = r \mid D = d\right) = \sum_g P^*\left(R = r \mid D = d, G = g\right)\ P^*\left(G = g \mid D = d\right)$$

As variables D and G are stochastically independent in P^* (see the causal model on the top of Figure 3), we can simplify:

$$\begin{aligned} &\sum_g P^*\left(R = r \mid D = d, G = g\right)\ P^*\left(G = g \mid D = d\right) \\ =&\sum_g P^*\left(R = r \mid D = d, G = g\right)\ P^*\left(G = g\right) \end{aligned}$$

And now the magic unfolds, as we realize that all of the "ingredients" in the last expression, though expressed as probabilities from P^*, can be safely estimated from data generated from P, i.e., our actual data. Why? — Because we assume that the data-generating processes behind P and P^* only differ with respect to the dependence of G and D, but the remaining (conditional) probabilities, though expressed in P^*, are identical in P (see Figure 3). So, we may conclude that, in this particular case, it is possible to "calculate away" the imaginary intervention effect by expressing it entirely in quantities safely estimable from the data at hand, since:

$$\begin{aligned} &\sum_g P^*\left(R = r \mid D = d, G = g\right)\ P^*\left(G = g\right) \\ =&\sum_g P\left(R = r \mid D = d, G = g\right)\ P\left(G = g\right) \end{aligned}$$

The full derivation is repeated more concisely here:

$$
\begin{aligned}
&P\left(R = r \mid do(D = d)\right) \\
&\quad = P^*\left(R = r \mid D = d\right) && \text{[by def.]} \\
&\quad = \sum_g P^*\left(R = r \mid D = d, G = g\right)\ P^*\left(G = g \mid D = d\right) && \text{[std. prob.]} \\
&\quad = \sum_g P^*\left(R = r \mid D = d, G = g\right)\ P^*\left(G = g\right) && \text{[cond. indep.]} \\
&\quad = \sum_g P\left(R = r \mid D = d, G = g\right)\ P\left(G = g\right) && \text{[same in } P\text{]}
\end{aligned}
$$

Now, we can compute the drug's causal effect on recovery by plugging in the numbers from our observational study in Table 2. We compare the recovery rates for cases where the patients took ($P(R = 1 \mid (D = 1))$) or refused the drug ($P(R = 1 \mid (D = 0))$):

$$
\begin{aligned}
P(R = 1 \mid do(D = 1)) &= \sum_{g \in \{0,1\}} P(R = 1 \mid G = g; D = 1) P(G = g) \\
&= \frac{81}{87} \cdot \frac{357}{700} + \frac{192}{263} \cdot \frac{343}{700} \approx 0.83
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
P(R = 1 \mid do(D = 0)) &= \sum_{g \in \{0,1\}} P(R = 1 \mid G = g; D = 0) P(G = g) \\
&= \frac{234}{270} \cdot \frac{357}{700} + \frac{55}{80} \cdot \frac{343}{700} \approx 0.78
\end{aligned} \tag{2}
$$

According to these point-valued estimates, patients who take the drug have approximately a 0.83 chance of recovery, while not taking the drug yields a 0.78 chance to recover. To get the causal effect estimate, we simply have to compare the predicted outcome for patients taking the drug with those refusing it. The estimated causal effect of DRUG INTAKE on RECOVERY is thus $P(R = 1 \mid do(D = 1) - P(R = 1 \mid do(D = 0) \approx 0.05$, meaning that the drug increases the chance of patients to recover by 5%.

MLE for Scenario 2: blood pressure as mediator. In scenario 2 (see the bottom causal model in Figure 3), the independent variable D is not causally dependent on any other variable. The *do*-operation, whose job it is to remove causal influences on the independent variable, thus doesn't have an effect in this case, because there aren't any causal influences on D to begin with. This means that the observed distribution P is identical to the hypothetical distribution P^*. The probability for recovery, in this case, is just $P(R = 1 \mid$

$D = d$):

$$
\begin{aligned}
&P\left(R = 1 \mid do(D = d)\right) \\
&\quad = P^*\left(R = 1 \mid D = d\right) && \text{[by def.]} \\
&\quad = \sum_b P^*\left(R = 1 \mid B = b; D = d\right) P^*\left(B = b \mid D = d\right) && \text{[std. prob.]} \\
&\quad = \sum_b P\left(R = 1 \mid B = b; D = d\right) P\left(B = b \mid D = d\right) && \text{[same in } P\text{]} \\
&\quad = P\left(R = 1 \mid D = d\right)
\end{aligned}
$$

We therefore directly obtain the maximum-likelihood estimates from the RECOVERY frequencies in the data:

$$P(R = 1 \mid do(D = 1)) = P(R = 1 \mid D = 1) = \frac{273}{350} \approx 0.78 \tag{3}$$

$$P(R = 1 \mid do(D = 0)) = P(R = 1 \mid D = 0) = \frac{289}{350} \approx 0.83 \tag{4}$$

So, for this case, we estimate that the causal effect of DRUG INTAKE on RECOVERY is $P(R = 1 \mid do(D = \text{TAKE}) - P(R = 1 \mid do(D = \text{REFUSE}) \approx -0.05$. According to this estimation, taking the drug might actually worsen the chances to recover.

Note that while we are able to compute a causal effect estimate by hand using a maximum-likelihood estimation approach, we don't get a measure of certainty without further ado. Numerically a causal effect of about 5% may seem pretty small, so how are we to tell whether this is reliably different from a causal effect size of zero? This is where the Bayesian approach comes into play. How to use Bayesian regression modeling to not only estimate causal effects, but quantify uncertainty about these estimates, is the topic of the remainder of this tutorial.

5 *Causal effect estimates from Bayesian Regression*

In this section, we calculate causal effects with Bayesian regression. Section 5.1 first explains *what* it is that we are doing. Then, Section 5.2 shows *how* to implement this approach in R using the `brms` package.

5.1 Conceptual approach

As described above, we can compute the causal effect of interest as:

$$\text{Avg-CE}(R \mid D) = P\left(R = 1 \mid do(D = 1)\right) - P\left(R = 1 \mid do(D = 0)\right)$$

In the first scenario, we can reduce the summands on the right-hand side to:

$$P\left(R = 1 \mid do(D = 1)\right) = \sum_g P\left(R = 1 \mid D = 1, G = g\right)\ P\left(G = g\right)$$

$$P\left(R = 1 \mid do(D = 0)\right) = \sum_g P\left(R = 1 \mid D = 0, G = g\right)\ P\left(G = g\right)$$

In the second scenario, the summands are obtained by:

$$P\left(R = 1 \mid do(D = 1)\right) = P\left(R = 1 \mid D = 1\right)$$
$$P\left(R = 1 \mid do(D = 0)\right) = P\left(R = 1 \mid D = 0\right)$$

Taken together, we need estimates of three types of probabilities:

- the unconditional probability $P(G = g)$ for gender categories
- the conditional probability $P(R = 1 \mid D = d, G = g)$ of recovery given D and G
- the conditional probability $P(R = 1 \mid D = d)$ of recovery given D

We, as modellers, are uncertain about these probabilities, but the available data let's us estimate each. The MLE from the previous section demonstrated this, but these estimates did not (directly) express the amount of uncertainty we have. To resolve this, we are going to use Bayesian regression modeling to estimate these probabilities, using the usual sampling-based approach. Samples for the relevant (conditional and unconditional) probabilities involved will then contain information about how certain we are about our estimates.

To understand how regression comes in, you need to understand a few crucial, general aspects.

1. A conditional probability distribution like $P(Y \mid X_1, \ldots, X_n)$ can be approximated with a regression model which we would normally specify with R's formula syntax: $Y \sim X_1 * \cdots * X_n$. If we want to estimate an unconditional probability $P(Y)$, we can do so with a so-called "intercept-only" model, as given by regression formula: $Y \sim 1$.

2. For our current purposes, where we are dealing with binary variables, we need to look at logistic regression models.

3. Logistic regression models, like other generalized regression models, make predictions at different levels for any value of predictive variables $X_1, \ldots, X_n$. They predict values for

   (i) the linear predictor,

   (ii) the central tendency, and

   (iii) the dependent / response variable Y.

   In (non-generalized) standard regression models, the linear predictor and the central tendency are the same.

4. Since we are dealing with binary variables and logistic regression, what we care about are predictions at the second level, the level of central tendency. Concretely, we want an estimate of a probability like $P(G = 1)$, which is the probability of a participant being male. That's a prediction of central tendency. We do not want a prediction of a log-odds ratio (the linear predictor) or prediction of a single binary category (male or female).

   We can obtain a prediction of central tendency by sampling many concrete values of the response variable, but that is less efficient than sampling a value of central tendency directly.

5. In a Bayesian regression model all of these predictions are probabilistic, since they represent our uncertainty about these values (as we estimate them from limited data). In a usual Bayesian workflow, we would therefore obtain samples from the so-called *posterior predictive distribution* for each one of these different levels.

For a detailed explanation of the differences of drawing samples from different levels of posterior predictive distributions, see this blog post.

These points taken together yield the following approach. We use logistic regression with the appropriate dependent and independent variables to get at the relevant conditional or unconditional probabilities. We use samples from the posterior predictive distribution of central tendency (the category probabilities). For scenario one, we need to do this in a nested fashion, to reflect the causal effect formula for this case. Our uncertainty about the causal effect estimate is the propagated uncertainty inherent in the samples from the posterior predictive distribution.

5.2 *Implementation*

Having established the conceptual basics, let's now dive into the programming part. First, we set up the environment and load the required packages:

The entire code is available at `https://github.com/michael-franke/BayesReg-CausalInference`.

```
#### packages, options, seed ----

# for Bayesian regression modelling
library(brms)        # Bayesian regression models
library(HDInterval) # for calculating highest density intervals
library(tidybayes)  # for tidy output of Bayesian models
# for data wrangling
library(tibble)      # dataframes
library(tidyr)       # for tidying dataframes
library(dplyr)       # intuitive data manipulation
# for plotting
library(ggplot2)     # for plots

# option for Bayesian regression models:
# use all available cores for parallel computing
options(mc.cores = parallel::detectCores())

# seeding for reproducibility
set.seed(123)
```

We also the number of samples to use for each variable:

```
# global parameter: number of iterations / samples
n_iter = 5000
```

Next, we create a dataset with the values from Table 2 and transform it into the required shape.

```
#### preparing the data set ----

# generate dataset with Simpsons paradox
data_simpsons_paradox <- tibble(
  gender = c("Male", "Male", "Female", "Female"),
  bloodP = c("Low", "Low", "High", "High"),
  drug = c("Take", "Refuse", "Take", "Refuse"),
  k = c(81, 234, 192, 55),
  N = c(87, 270, 263, 80),
  proportion = k/N
)

# cast into long format (one observation per row)
data_SP_long <- rbind(
  data_simpsons_paradox |>
    uncount(k) |>
    mutate(recover = TRUE) |>
    select(-N, -proportion),
  data_simpsons_paradox |>
    uncount(N-k) |>
    mutate(recover=FALSE) |>
    select(-k, -N, -proportion)
  ) |> # this part is new
  mutate(
    gender = factor(gender),
    drug = factor(drug)
  )
```

Now, we can start modeling!

Scenario 1: Gender as confound. In scenario 1, the effectiveness of the drug is confounded by GENDER. As described above, we need to model two probability distributions: the (unconditional) probability of GENDER, $P(G = g)$ and the (conditional) probability of RECOVERY, $P(R = r \mid D = d, G = g)$. Therefore, we first fit an intercept-only regression model which allows us to estimate the distribution of GENDER, $P(G = g)$.

```
# estimate the distribution of gender
#   (by intercept-only regression)
fit_gender <- brm(
  formula = gender ~ 1,
  data = data_SP_long,
  family = bernoulli(link = "logit"),
  iter = n_iter
)
```

Next, we fit a regression model to help estimate the conditional probability of recovery, $P(R = r \mid D = d, G = g)$.

```
# estimate the distribution of recovery rates as predicted
#   by gender and treatment
fit_recovery_gender <- brm(
  formula = recover ~ gender * drug,
  data = data_SP_long,
  family = bernoulli(link = "logit"),
  iter = n_iter
)
```

To get samples from the posterior predictive distributions (of central tendency) of fitted models, we use the function `epred_draws()` from the *tidybayes* package. For the case at hand, since we want samples of $P(G = 1)$, we pass the intercept value of 1 in the `newdata` argument of the `epred_draws()` function. Each sample from the posterior predictive distribution of central tendency is a probability value between 0 and 1, which we can interpret as the estimated fraction of men in the sample.

Notice that the variance in these samples is what quantifies our (modeller's) uncertainty about the value $P(G = 1)$.

```
#  estimate gender distribution by sampling from the expected
    value
#    of the posterior predictive distribution
#  (estimate the fraction of men in the sample)
posterior_gender_proportion <- tidybayes::epred_draws(
  object = fit_gender,
  newdata = tibble(Intercept = 1),
  value = "maleProp", #proportion of men in the sample
  ndraws = n_iter * 2
) |> ungroup() |>
  select(.draw, maleProp)
```

We then estimate the average recovery rate for all combinations of GENDER and DRUG INTAKE by drawing from the expected values of the posterior predictive distributions of the RECOVERY regression model. To do so, we first define a data structure with all combinations of levels of the predictor variables.

```
# make a data frame with all compbinations of values for
    variables
#   of drug and gender
newdata <- tibble(
  gender = factor(c("Male", "Male", "Female", "Female"), levels
      = levels(data_SP_long$gender)),
  drug = factor(c("Take", "Refuse", "Take", "Refuse"), levels =
      levels(data_SP_long$drug))
)
```

This structure is passed into the argument `newdata` to obtain samples from the posterior predictive distribution of central tendency, via `epred_draws()` as before.

```
# get posterior predictive samples for each combination of
    values
#  for drug and genders
posterior_g <- tidybayes::epred_draws(
  object  = fit_recovery_gender,
  newdata = newdata,
  value   = "recovery",
  ndraws  = n_iter * 2
) |> ungroup() |>
  select(.draw, gender, drug, recovery)
```

The data frame `posterior_g` now contains samples of $P(R = 1 \mid D = d, G = g)$. We need to multiply each prediction for $P(R = 1 \mid D = d, G = g)$ with a sample of $P(G = g)$, the estimated fraction of men and women, respectively, which are stored in `posterior_gender_proportion`. This allows us to obtain samples of the causal effect we are interested in.

```
# obtain estimates of causal effect by
#   calculating the weighted average of recovery rates for each
#   value of drug, weighted by the est. proportions of males
posterior_g <- posterior_g |>
  # add the previous samples of P(G=1)
  #  and flip, where necessary, to P(G = 0)
  full_join(posterior_gender_proportion) |>
  mutate(weights = ifelse(gender == "Male",
                          maleProp, 1-maleProp)) |>
  # calculate P(G=g) * P(R|G=g, D=d) for each combination of G
      and D
  group_by(`.draw`, drug) |>
  summarize(predRecover = sum(recovery * weights)) |>
  pivot_wider(names_from = drug, values_from = predRecover) |>
  # causal effect estimates are obtained from by subtracting
  #   the predicted recovery rates when the drug is refused from
  #   the predicted recovery rates when the drug is taken
  mutate(causal_effect = Take - Refuse)
```

The results of these computations are summarized with the `summarize_posterior` function:

```
> summarize_posterior(posterior_DrugRefused_g,
                      posterior_DrugTaken_g)
# A tibble: 3 x 4
  condition     CI_lower   mean CI_upper
  <fct>            <dbl>  <dbl>    <dbl>
1 take drug       0.792  0.832     0.869
2 refuse drug     0.726  0.778     0.834
3 causal effect  -0.0121 0.0534    0.120
```

`summarize_posterior` is a custom helper function. You can find it on the GitHub page of this tutorial.

Similar to the MLE approach, we obtain a causal effect estimate of about 5%, i.e., patients are predicted to have a roughly 5% higher chance of recovery when taking the drug. However, we can see that the credible interval includes 0, which suggests that we do not have enough evidence to rule out that the drug has no causal effect.

Scenario 2: Blood pressure as mediator In scenario 2, blood pressure acts as a mediator between the drug and the patients' recovery. As described

above, we simply need to calculate the conditional recovery probability $P(R = 1 \mid D = d)$. Therefore, we fit RECOVERY to DRUG INTAKE.

```
# fitting a regression model to predict recovery based on drug
    use alone
fit_recovery_bp <- brm(
  formula = recover ~ drug,
  data = data_SP_long,
  family = bernoulli(link = "logit"),
  iter = n_iter
)
```

As we do not have to do any further calculations to estimate the causal effect in scenario 2, all we need to do is extract the relevant posterior predictive samples, and computing samples of the causal effect, as we did previously.

```
# posterior predictive samples for the recovery rate
#   when the drug is refused and taken
posterior_bp <-
  tidybayes::epred_draws(
    object = fit_recovery_bp,
    newdata = tibble(drug = c("Refuse", "Take")),
    value = "predRecover", # prob. of recovery
    ndraws = n_iter * 2
  ) |> ungroup() |>
  select(.draw, drug, predRecover) |>
  pivot_wider(names_from = drug, values_from = predRecover) |>
  # causal effect estimates are obtained from by subtracting
  #   the predicted recovery rates when the drug is refused from
  #   the predicted recovery rates when the drug is taken
  mutate(causal_effect = Take - Refuse)
```

Summary statistic for these samples of causal effect estimates are:

```
> summarize_posterior(posterior_DrugRefused_bp, posterior_
    DrugTaken_bp)
# A tibble: 3 x 4
  condition     CI_lower    mean CI_upper
  <fct>            <dbl>   <dbl>    <dbl>
1 take drug        0.736  0.779    0.823
2 refuse drug      0.785  0.825    0.863
3 causal effect   -0.103 -0.0458   0.0131
```

We can see that taking the drug yields an estimated 78% chance of recovery, while not taking it leads to a 83% chance. The causal effect is estimated to be -5%, such that taking the drug is estimated to have detrimental effects on recovery. As in the first scenario, the CI includes 0, so that we cannot safely rule out that there is no causal effect.

6 *Conclusions & next steps*

Taking a step back, what can we learn from these hypothetical scenarios? The data for both scenarios was identical. Despite that, the computations of causal effects were different and the results reflected this: even though both causal effects were not credibly different from zero, they point in different direc-

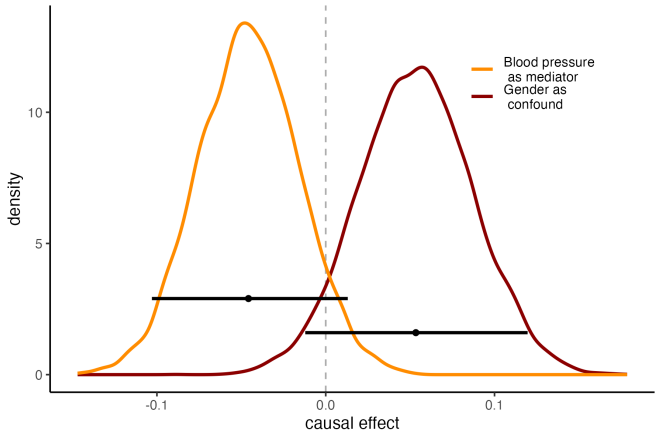


Figure 4: Posterior distributions of the causal effect for both scenarios. Bold Black lines indicate 95% credible intervals, the posterior means are shown as black dots.

tions, as can be seen in Figure 4. This illustrates the importance of choosing the correct causal structure. Going beyond our hypothetical scenarios, we showed that the *do*-calculus can, in some scenarios, be used to estimate causal effects, even if the data at hand stems from observational studies. To estimate the causal effects, both frequentist (MLE) and Bayesian methods yield similar results, yet only with a Bayesian sampling-based approach also allowed us to flexibly quantify the uncertainty of causal effect estimates.

If you want to learn more about causal models as a formal representational tool, the overview article by Hitchcock (2024) is a great place to start. Notice that in this tutorial we only covered the case where we stipulated one particular causal model and used it to draw "causally informed" conclusions from our data. Yet, data may also be used to try to infer which causal structure is likely behind the data-generating process. Hitchcock's overview article covers basic results about in-principle identifiability of causal structure.

To dive deeper into causal inference for statistics with the *do*-calculus, we recommend introductory and overview texts by Judea Pearl (Pearl, 2009; Pearl, Glymour, and Jewell, 2016). If you would like to learn more about the combination of causal inference and Bayesian data analysis, the newest addition of "Statistical Rethinking" by Richard McElreath is a gem.

Materials for Richard McElreath course from 2024 are here: `https://github.com/rmcelreath/stat_rethinking_2024`.

Very basic information about other approaches to causal inference than the *do*-calculus is contained in Info Box 2.

References

Bollen, Kenneth A and Judea Pearl (2013). "Eight myths about causality and structural equation models". In: *Handbook of causal analysis for social research*. Springer, pp. 301–328.

Bürkner, Paul-Christian (2021). "Bayesian Item Response Modeling in R with brms and Stan". In: *Journal of Statistical Software* 100.5, pp. 1–54. DOI: 10.18637/jss.v100.i05.

The Neyman-Rubin Causal Model / potential outcome framework

The Neyman-Rubin causal model is based on the framework of potential outcomes that a treatment would have on an individual. As an example, let's consider going to college as the treatment, and future income as the outcome of interest. The causal effect of going to college on the individual's future income is defined as the difference between the income with and without a college degree. This notion of causal effect is a hypothetical measure, as the potential outcomes are counterfactual and only one of them can be observed. Importantly, the potential outcomes are treated as fixed quantities and not random variables. This is different from *do*-calculus, where the outcomes are also stochastic, and the assignment of treatment is fixed/deterministic. The only stochastic process in the Neyman-Rubin causal model is the assignment of treatment, which depends on the observable characteristics X. These could, for example, be high school grades, socio-economic background or gender. Because the potential outcomes are fixed, any two individuals who share the same set of characteristics X have the same potential outcomes. This can be leveraged to compute the causal effect. For each combination of values X can take, the income of individuals who went to college can be compared with that of those who didn't. This procedure is called matching. Because it is in most cases not easy to find individuals with exactly matching values of X, especially for continuous variables, there are different proposed matching methods (Sekhon, 2008).

Structural Equation Modeling

A structural equation model (SEM) is a causal model that combines observable and latent variables. The causal connections between variables are represented with equations and treated as a-priori. SEMs can also be represented graphically in path diagrams. A path diagram is not the same as a cDAG, because it is not acyclic and can involve feedback relations. However, in this case, the model is underidentified and non-testable. The goal of a SEM is to estimate the values of latent variables from the measured values of the observable variables. SEMs can be evaluated with measures of model fit. In case of a bad fit, the causal assumptions underlying model formulation can be regarded as falsified (Bollen and Pearl, 2013).
It has been shown that the potential outcome framework and structural equation modeling are logically equivalent (Galles and Pearl, 1998; Halpern, 2000). Pearl (2000) combined features of the potential outcome framework and SEM to formulate the structural causal model as a general theory of causation.

Info Box 1: Alternatives to the *do*-calculus for causal inference.

Charig, Clive R et al. (1986). "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy." In: *Br Med J (Clin Res Ed)* 292.6524, pp. 879–882.

Cinelli, Carlos, Andrew Forney, and Judea Pearl (2022). "A Crash Course in Good and Bad Controls". In: *Sociological Methods & Research* 53.3, pp. 1071–1104.

Franke, Michael and Timo B. Roettger (2019). "Bayesian regression modeling (for factorial designs): A tutorial". DOI: 10.31234/osf.io/cdxv3. URL: https://psyarxiv.com/cdxv3.

Galles, David and Judea Pearl (1998). "An axiomatic characterization of causal counterfactuals". In: *Foundations of Science* 3, pp. 151–182.

Halpern, Joseph Y (2000). "Axiomatizing causal reasoning". In: *Journal of Artificial Intelligence Research* 12, pp. 317–337.

Hitchcock, Christopher (2024). "Causal Models". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2024. Metaphysics Research Lab, Stanford University.

Julious, Steven A and Mark A Mullee (1994). "Confounding and Simpson's paradox". In: *BMJ* 309.6967, pp. 1480–1481.

Pearl, Judea (2000). *Causality: Models, Reasoning and Inference*. Cambridge, UK: Cambridge University Press.

– (2009). "Causal inference in statistics: An overview". In.

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell (2016). *Causal Inferece: A Primer*. New York: Wiley.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Sekhon, Jasjeet (2008). "The Neyman—Rubin Model of Causal Inference and Estimation Via Matching Methods". In: *The Oxford Handbook of Political Methodology*. Oxford University Press. DOI: 10.1093/oxfordhb/9780199286546.003.0011.

Shrier, Ian and Robert W Platt (2008). "Reducing bias through directed acyclic graphs". In: *BMC medical research methodology* 8, pp. 1–15.