GR5065 Homework 3 Answer Key

Ben Goodrich

Due February 23, 2021 at 8PM New York Time

```
set.seed(20210223)
options(mc.cores = parallel::detectCores())
```

1 Police Stops in North Carolina

```
stops <- readRDS("north_carolina.rds")
head(stops)
```

```
##                                           Asian  Black Hispanic  White
## Charlotte-Mecklenburg Police Department 13599 419873    77727 282761
## Raleigh Police Department                8065 199297    40638 173329
## Greensboro Police Department             4812 143572    13711 105563
## Fayetteville Police Department           2535 128176    13486  76564
## Winston-Salem Police Department          1666  98631    22355  89247
## Durham Police Department                 2392  82629    17326  37466
```

1.1 Prior Predictive Distribution

```
prior_PD <- function(stops) {
  D <- nrow(stops)
  R <- ncol(stops)
  searches_hits <- array(0, dim = c(D, R, 2), dimnames =
                         list(rownames(stops), colnames(stops), c("searches", "hits")))

  # these four variables are common to all races and departments
  mu_phi <- rnorm(n = 1, mean = 0, sd = 2)
  sigma_phi <- abs(rnorm(n = 1, mean = 0, sd = 2))
  mu_lambda <- rnorm(n = 1, mean = 0, sd = 2)
  sigma_lambda <- abs(rnorm(n = 1, mean = 0, sd = 2))

  # these two vectors are department specific
  phi <- rnorm(n = D, mean = mu_phi, sd = sigma_phi)
  lambda <- rnorm(n = D, mean = mu_lambda, sd = sigma_lambda)

  # enforce constraint on largest department
  phi[1] <- 0
  lambda[1] <- 0

  for (r in 1:R) {
    # these are specific to the r-th race
    phi_r    <- rnorm(n = 1, mean = 0, sd = 2)
    lambda_r <- rnorm(n = 1, mean = 0, sd = 2)
    mu_tr    <- rnorm(n = 1, mean = 0, sd = 2)
    sigma_tr <- abs(rnorm(n = 1, mean = 0, sd = 2))
    for (d in 1:D) {
      # these are specific to an intersection of race and department
      t_rd <- plogis(rnorm(n = 1, mean = mu_tr, sd = sigma_tr))
      phi_rd <- plogis(phi_r + phi[d])
      lambda_rd <- exp(lambda_r + lambda[d])
      alpha <- phi_rd * lambda_rd
      beta <- (1 - phi_rd) * lambda_rd
      # draw one signal per traffic stop which is seen by the police officer
      p <- rbeta(n = stops[d, r], shape1 = alpha, shape2 = beta)
      searched_rd <- p > t_rd # iff true car and / or driver is searched
      searches_rd <- sum(searched_rd)
      hits_rd <- sum(rbinom(n = searches_rd, size = 1, prob = p[searched_rd]))
      searches_hits[d, r, 1:2] <- c(searches_rd, hits_rd)
    }
  }
  return(searches_hits)
}
```

1.2 Description

```
UNC <- replicate(1000, prior_PD(stops)[85, , ])
```

```
UNC_searches <- sweep(UNC[ , 1, ], MARGIN = 2,
                      STATS = colSums(UNC[ , 1, ]), FUN = `/`)
UNC_hits <- sweep(UNC[ , 2, ], MARGIN = 2,
                      STATS = colSums(UNC[ , 2, ]), FUN = `/`)
plot(x = c(UNC_searches), y = c(UNC_hits), pch = 20, col = c(2, 1, 3, 4),
     xlab = "Proportion of Searches", ylab = "Proportion of Hits", las = 1)
legend("topleft", legend = rownames(UNC_searches), pch = 20, col = c(2, 1, 3, 4),
       title = "Race of Driver", bg = "lightgrey")
```

As can be seen in Figure 1 on the next page, the proportion of hits by the UNC Chapel Hill police department is linearly related (with noise) to the proportion of searches under the authors' prior. However, the prior permits some deviations from the line; for example, in the lower right corner it is possible for about 90% of the searches to be when the driver is black but only about 20% of the searches that hit to be when the driver is black. If anything like that were to transpire in the actual data, it would be certainly be cause for investigation of the UNC Chapel Hill police department (although the authors point out that just looking at searches and hits is insufficient to prove or disprove claims of discrimination).

There tends to be more searches by the UNC Chapel Hill police department when the driver is white than when the driver is black and even fewer when the driver is Asian or Hispanic. That is due to the fact that there are more stops of white drivers,

```
stops[85, ]
```

```
##    Asian    Black Hispanic    White
##      641     2414      530     6593
```

which is unsurprising for a college town.

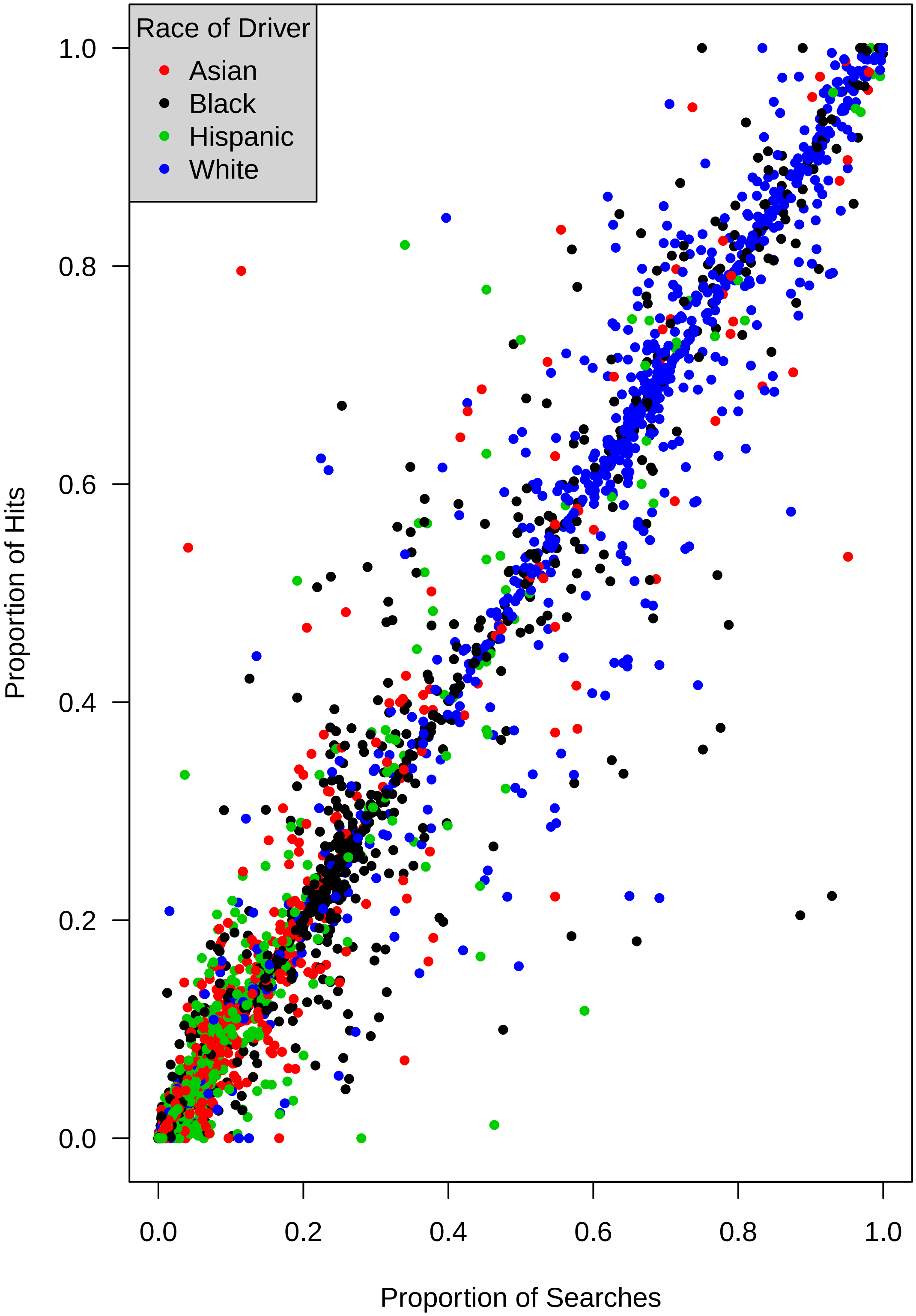


Figure 1: Prior Predictive Distribution for UNC Chapel Hill

1.3 Criticism

Many criticisms could be made of this (or any) paper. Perhaps the most obvious one is that this particular model starts with traffic stops, which takes them as given and thus cannot say anything about the potential for discrimination in the decision by police officers to stop a car. You could easily imagine the model being expanded so that there is an unknown threshold at which police officers from a department decide to stop a car that may vary by the race of the driver or at least according to the demographics of the neighborhood.

The same research group has another paper where they use the time of day as an instrument in a model that includes the probability that a car is stopped, on the theory that if it is evening or night, then it is harder for the police officer to know in advance what the race of the driver is when the car is stopped.

2 Medicaid Expansion in Oregon

```
library(haven)
unzip("100.00019026_supp.zip")
oregon <- as_factor(read_dta(file.path("19026_supp", "Data", "individual_voting_data.dta")))
```

```
library(dplyr)
oregon <- transmute(oregon,
                    V = vote_presidential_2008_1,           # voted in Nov 2008?
                    M = ohp_all_ever_nov2008 == "Enrolled", # had Medicaid in Nov 2008?
                    L = treatment,                          # won lottery in spring 2008?
                    N = numhh_list != "signed self up")     # registered additional adults?
```

2.1 p-value

The p-value of 0.073 means that there is a probability of 0.073 of estimating an the effect of Medicaid on voter turnout to be greater in magnitude than 0.02549 if the true effect were zero. Of course, since this is a Frequentist probability statement, it really means that 0.073 is the proportion of randomly-sampled datasets of size $N = 74922$ that would yield an estimate of the effect of Medicaid on voter turnout greater in magnitude than 0.02549 if the true effect were zero.

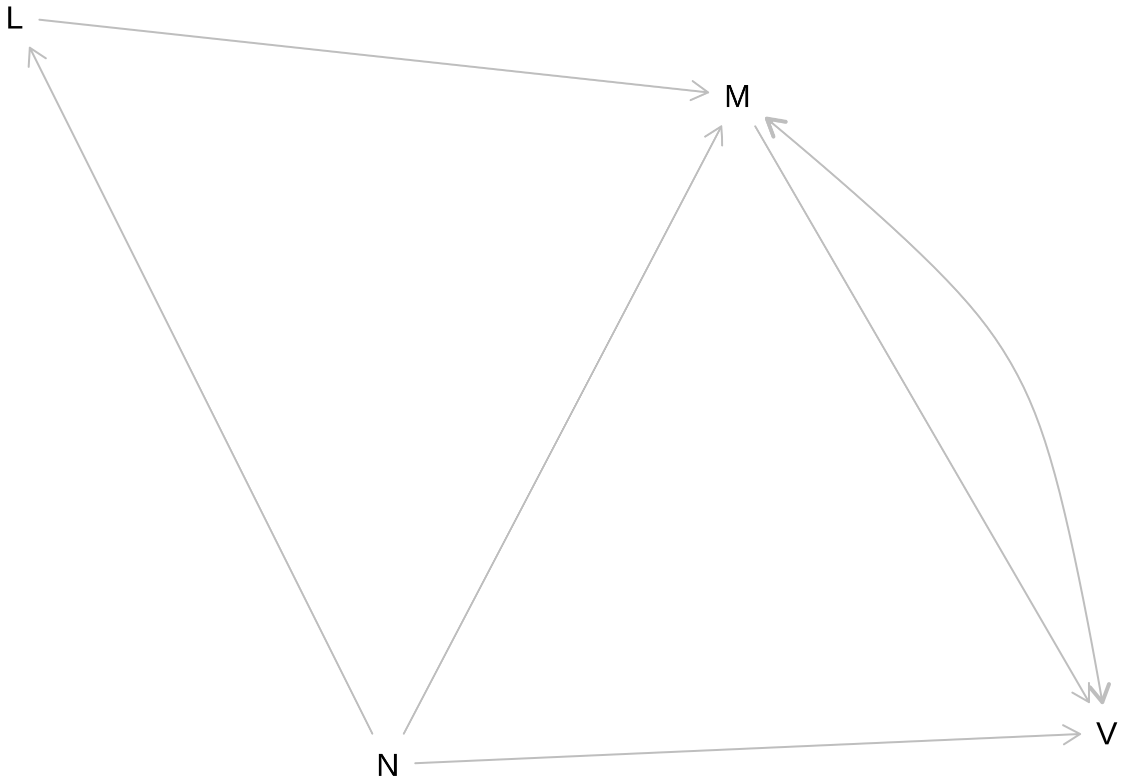
The entire Frequentist approach is problematic in this case considering the original dataset of size $N = 74922$ was not a sample; everyone in Oregon was eligible to sign up for the Medicaid lottery (and they later verified whether lottery winners had sufficiently little income to qualify for Medicaid). Although the lottery was randomized, the authors did not perform design-based inference over possible treatment assignments. The people who both won the lottery and followed up to have their eligibility verified and be given Medicaid are not a random sample from the population of poor households in Oregon.

Even if you were to consider the original dataset of size $N = 74922$ to be a sample from some population, how could you obtain another sample of size $N = 74922$ from that same population? Thus, it does not really make sense to talk about 7.3% of datasets of size $N = 74922$ when it is impossible for there to be more than one such dataset.

In this situation, it is much more natural to adopt the Bayesian perspective on probability where we are referring to the degree-of-belief that giving people Medicaid causes more of them to vote than they otherwise would. Essentially all of the uncertainty about this proposition comes from not knowing which model to estimate and what are the parameter values in that model (or models). Essentially none of the uncertainty is due to these $N = 74922$ people randomly differing from another hypothetical $N = 74922$ people from the same population.

2.2 Directed Acyclic Graph

```
library(CausalQueries)
model <- make_model("N -> L -> M -> V; N -> M; N -> V") %>%
  set_restrictions(decreasing("N", "L"), keep = FALSE)  %>%
  set_restrictions(decreasing("L", "M"), keep = FALSE)  %>%
  set_confound("M <-> V")
plot(model)
```



The logic behind the DAG is as follows. Households with more adults have more chances to win the Medicaid lottery because if any member of the household has a winning ticket, then everyone in their household can get Medicaid (if the household, in fact, has sufficiently little income to qualify for Medicaid). Thus, we have $N \rightarrow L$ and a restriction that rules *out* the possibility that there could be anyone where $L(N) \neq N$, i.e. a person who would have won the lottery if their N were 0 but would not win the lottery if their N were 1. Conversely, there certainly are people who

- Would not have won the lottery if there N were zero but would have won the lottery if their N were 1 (if their spouse actually had the winning lottery ticket)
- Would not have won the lottery regardless of N (if neither the person nor their actual or hypothetical spouse had a winning lottery ticket)
- Would have won the lottery regardless of N (if they personally held a winning ticket, irrespective of whether they had a spouse or whether that person also had a winning ticket)

So, those three types should definitely be included but the fourth type can justifiably be excluded in this case.

Similarly, we have $L \rightarrow M$ and a restriction that rules *out* the possibility of there being someone who would get Medicaid if they lost the lottery but would not get Medicaid if they won the lottery. But the other three types are quite possible, although it is more rare for someone to lose the lottery and then qualify to get Medicaid by some other means (usually getting married to someone similarly poor and / or having more children).

There is a $N \rightarrow M$ because it is generally more difficult for households with multiple adults to qualify for Medicaid. If both of them have jobs, then they collectively may have too much income, although the exact rules for Medicaid eligibility are very complicated and vary to some extent from one state to the next.

We also have $M \leftrightarrow V$ to indicate that there might be some unspecified parent variable U, such that $M \leftarrow U \rightarrow V$. U could be a lot of things, but the most obvious is income. One of the most durable findings in the voter turnout literature is that richer people are more likely to vote than poorer people in the United States (and almost all other democracies). At the same time, the more income a household has, the less likely they are to qualify for Medicaid. So, there should be a negative *marginal* correlation between M and V in the data due to variation in income that the authors did not (or could not) collect data on. If there were no confounding between M and V, then the authors could have estimated the average causal effect with a simple difference in voting proportions between those on Medicaid and those not on Medicaid.

Since there presumably is confounding between M and V due to income and a host of other reasons, the authors used two-stage least squares to estimate the (local) average treatment effect of Medicaid on voting by exploiting the portion of variation in Medicaid induced by the randomization of the lottery. When doing two-stage least squares, it is recommended to include all other first stage variables — besides the variable being randomized — that predict M in the second stage when predicting V. In this case, the only such variable is N. Although there is little reason to think that people in one-adult households are less or more likely to vote than people in multi-adult households, neither is there a strong reason to think it is impossible for N to have some direct effect on V (i.e. maybe couples encourage each other to vote).

This last point highlights a sharp difference between the older literature on structural equation modeling (that sometimes utilizes DAGs) and the newer literature on experiments. The older literature on structural equation modeling tended to recommend that a coefficient be fixed to zero unless there was a strong theoretical reason to believe that it was far from zero. The newer literature on DAGs tends to recommend that there be an arrow between two variables unless there is a strong theoretical reason to believe that randomizing the parent node would have no effect at all on the child node.

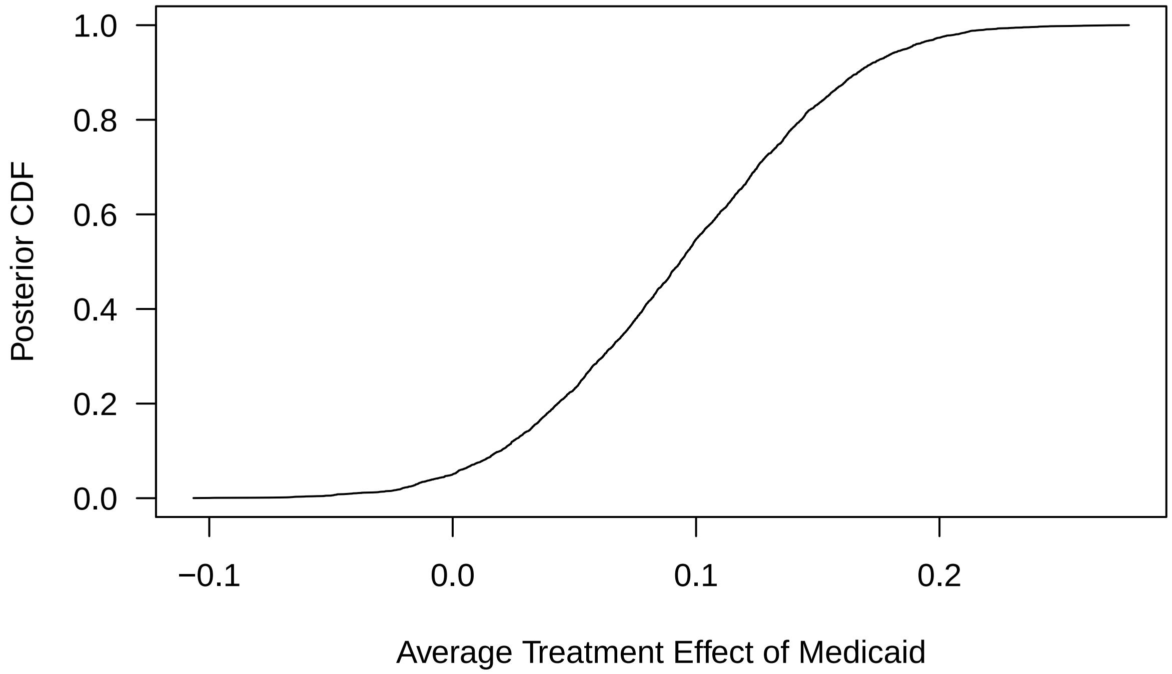
2.3 Posterior Distribution

```r
post <- update_model(model, data = oregon)
```

2.4 Interpretation

```r
ATE <- sort(c(query_distribution(post, query = "V[M = 1] - V[M = 0]",
                                 using = "posteriors")))
```

```r
plot(x = ATE, y = 1:length(ATE) / length(ATE), type = "l", las = 1,
     xlab = "Average Treatment Effect of Medicaid", ylab = "Posterior CDF")
```



From the posterior CDF of the ATE, we can see that the posterior probability that it is positive is more than 90% and the median is about 0.1, which would be a relatively large effect among interventions that have been investigated for increasing turnout among the poor. We could estimate those two values more exactly via

```
round(mean(ATE > 0), digits = 2)
```

```
## [1] 0.95
```

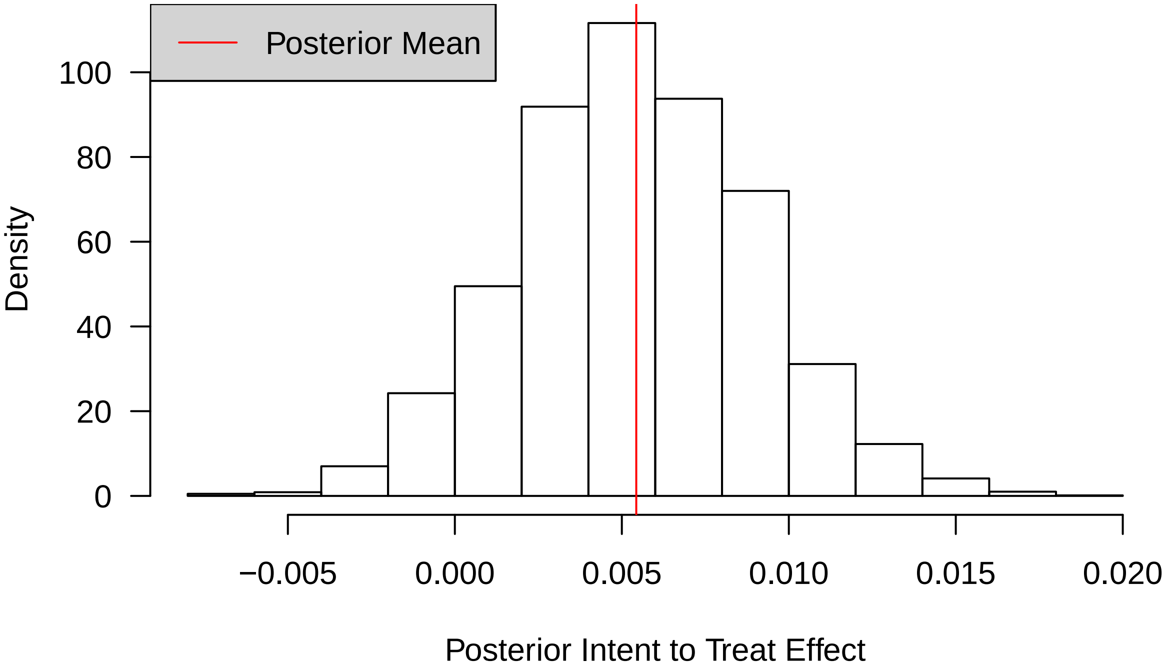
```
round(median(ATE),   digits = 2)
```

```
## [1] 0.09
```

Do not fall into the (not even) Frequentist trap of thinking just because there is a 5% chance that the ATE is negative under the posterior distribution, the research has failed to demonstrate that the effect is non-zero. The effect is very likely to be positive, although it is not estimated particularly precisely by this model, so the data are quite consistent with both small positive and large positive effects.

```
ITT <- sort(c(query_distribution(post, query = "V[L = 1] - V[L = 0]",
                                 using = "posteriors")))
```

```
hist(ITT, prob = TRUE, main = "", las = 1,
     xlab = "Posterior Intent to Treat Effect")
abline(v = mean(ITT), col = "red")
legend("topleft", legend = "Posterior Mean",
       lty = 1, col = "red", bg = "lightgrey")
```



The posterior distribution of the effect of winning the Medicaid lottery — also known as the Intent to Treat Effect — is smaller than the ATE, presumably due to the fact that some people who win the Medicaid lottery did not end up obtaining Medicaid, either because they did not follow up or because they were not eligible for Medicaid in the first place. In addition, the chance that the effect is positive is smaller than for the ATE.

```r
ATT <- sort(c(query_distribution(post, query = "V[M = 1] - V[M = 0]",
                                 using = "posteriors", given = "M == 1")))
```

```r
ATC <- sort(c(query_distribution(post, query = "V[M = 1] - V[M = 0]",
                                 using = "posteriors", given = "M == 0")))
```

The average treatment effect (i.e. the difference between the proportion of beneficial and the proportion of adverse types) among people for whom, in fact, $M = 1$, which is also known as the average treatment effect on the treated, looks to be negative in the top part of Figure 2 on the next page. This aspect of the posterior distribution does not make a lot of sense and may be evidence against the exclusion restriction, i.e. that there is no path from L to V except through M. Conversely, the average treatment effect among people for whom, in fact, $M = 0$, is quite positive. The average treatment effect, irrespective of M, is a weighted-sum of these two conditional treatment effects. Taken literally, it suggests that there are a lot of people in the control group who would have voted if they had gotten Medicaid.

```r
par(mfrow = 2:1, las = 1)
hist(ATT, prob = TRUE, xlab = "", xlim = c(-0.25, 0.4), ylim = c(0, 10),
     main = "Average Treatment Effect on the Treated")
hist(ATC, prob = TRUE, main = "", xlim = c(-0.25, 0.4), ylim = c(0, 10),
     xlab = "Average Treatment Effect on the Control")
```

```r
other <- sort(c(query_distribution(post, query = "V[N = 1] - V[N = 0]",
                                   using = "posteriors")))
```

We can simply numerically summarize the effect of household size on voting and it yields fairly convincing evidence of a lack of a meaningful effect. Thus, it seems that having a direct path from N to V might not have been necessary in retrospect, but including it did not do any harm.

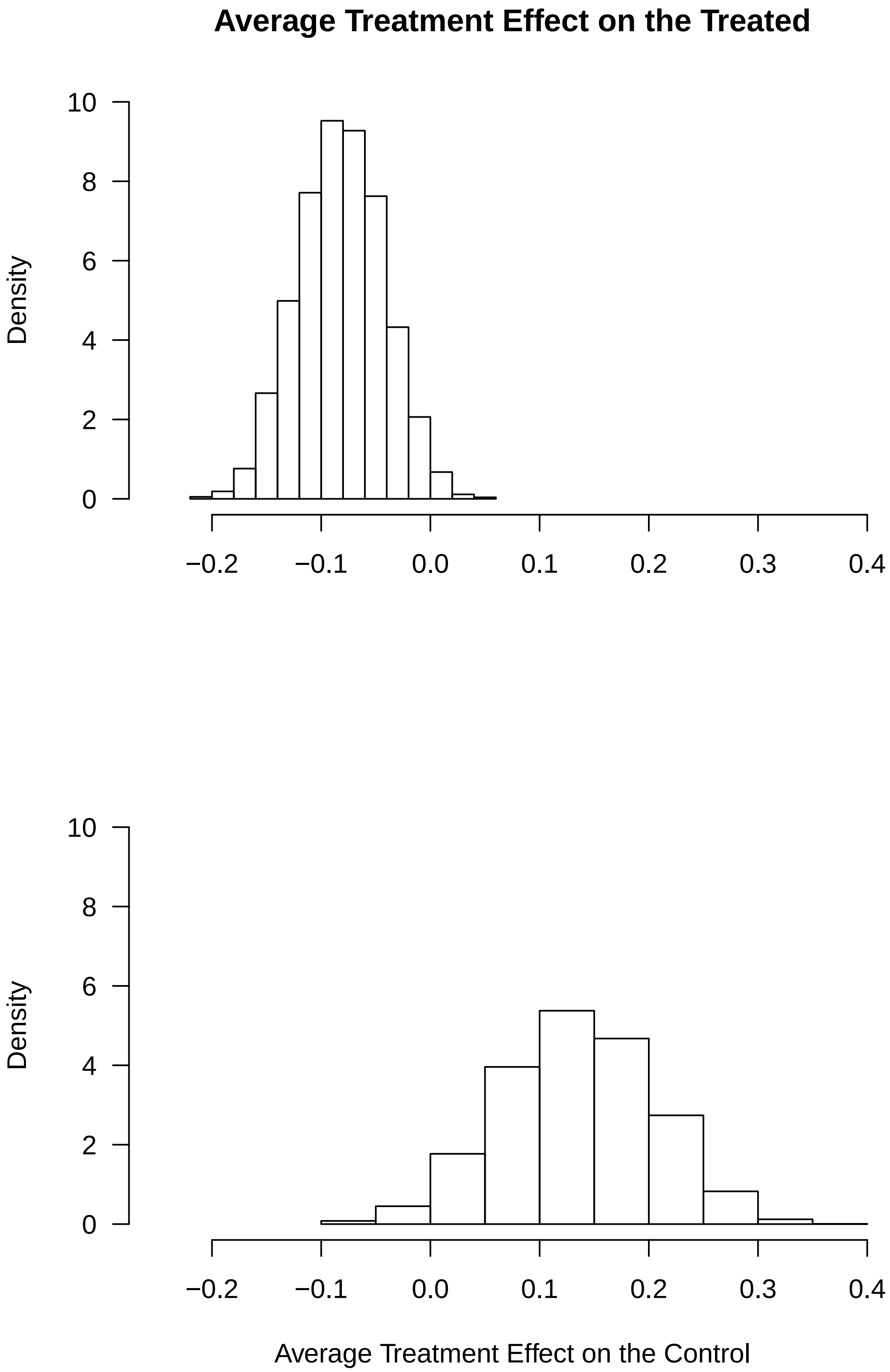


Figure 2: Conditional Treatment Effects

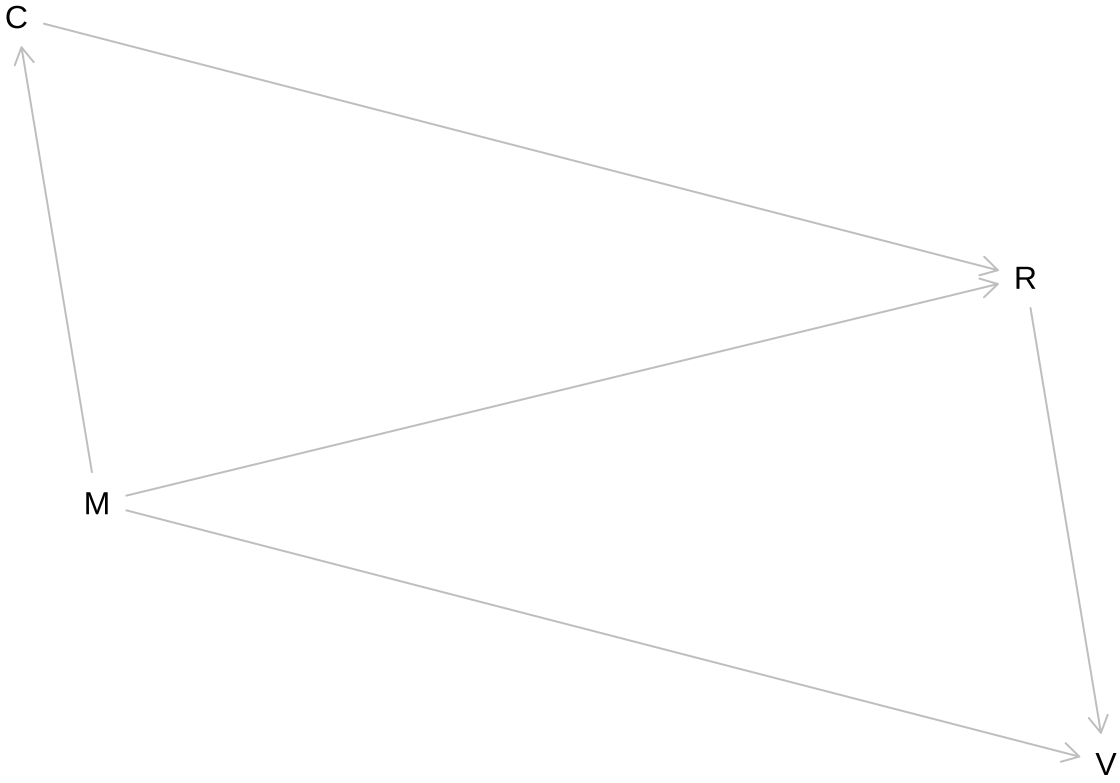
```r
summary(other)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## -0.011711 -0.000641  0.002058  0.002057  0.004847  0.017591
```

2.5 Citizenship

If we wanted to be more explicit, we could have written a part of a DAG like

```r
make_model("M -> R -> V; C -> R; M -> C; M -> V") %>% plot
```



where R is a binary variable indicating whether the person is registered to vote (in Oregon) and C is a binary variable indicating whether the person is a citizen. Noncitizens cannot register to vote, making them a clear example of "chronic" people that will not vote irrespective of M.

Thus, in an identification analysis C is a descendant of the variable whose effect you are trying to identify, and things like Dagitty would tell you that the adjustment criterion cannot be satisfied if you were to condition on $C = 1$. However, this is an example where the identification analysis conflicts with common sense. It is conceivable that having Medicaid might make a legal immigrant more likely to remain in the United States and apply for citizenship, but it is implausible that this effect would be very large (considering how much of a hurdle it is to become a citizen), especially over short time frames (like 2008). Actual researchers would much rather have a more precise estimate of the effect of Medicaid on voting among citizens than to have an unbiased but imprecise estimate of the effect of Medicaid on voting among all citizens and potential citizens. Moreover, the CausalQueries package would allow you to estimate the ATE among citizens and the ATE among noncitizens (at the time of the lottery).

It would be nice if C were available in the data, but I doubt the researchers would have sought or would have been granted approval from the Institutional Review Board (IRB) to ask people if they were citizens at the time they signed up for the Medicaid lottery. Such questions tend to discourage non-citizens from participating. However, the fact that L is randomized means that it has no ancestor in common with C, so two-stage least squares is still a consistent estimator of the (local) average treatment effect of Medicaid (meaning that as $N \uparrow \infty$ the average squared difference between the estimate and the true effect across datasets tends toward zero). CausalQueries also produces a valid posterior distribution of this effect, conditional on the data but this posterior distribution is more dispersed than it would be if you could condition on $C = 1$.

More generally, many researchers tend to think something like "If I have a source of randomization, a consistent estimator, and a large N, then I do not need other variables that predict the outcome or a complete (and possibly wrong) model of the data-generating process in order to estimate the average causal effect of interest." But the fact that the estimated standard error of the treatment effect was nowhere close to zero strongly suggests that even though $N = 74922$, that was not nearly large enough to rely on the behavior of the two-stage least squares estimator as $N \uparrow \infty$. To obtain a sufficiently precise estimate with any finite N, you often have to make modeling assumptions that are stronger that what is minimally necessary to identify a causal effect. The authors do have some additional models where the condition on previous voting, gender, or other individual variables that are known to affect the probability that someone will vote, but those additional models are not the emphasis of the paper.