GR5065 Homework 3 Answer Key

Ben Goodrich

Due March 7, 2022 at 4PM

```
library(dplyr)
library(dagitty)
library(ggplot2)
library(ggdag)
library(haven)
unzip("100.00019026_supp.zip")
oregon <- as_factor(read_dta(file.path("19026_supp", "Data", "individual_voting_data.dta")))
```

1 Frequentist

1.1 Confounding

What is at least one unobserved variable (U) that is a parent of both whether the person's household has Medicaid and whether the person votes, which leads to confounding, i.e. `Medicaid` $\leftarrow U \rightarrow$ `Vote`

There are many but perhaps the most obvious is income. The household must have a sufficiently low income to qualify for Medicaid — irrespective of whether they win the Medicaid lottery — and those with low income are less likely to vote in elections.

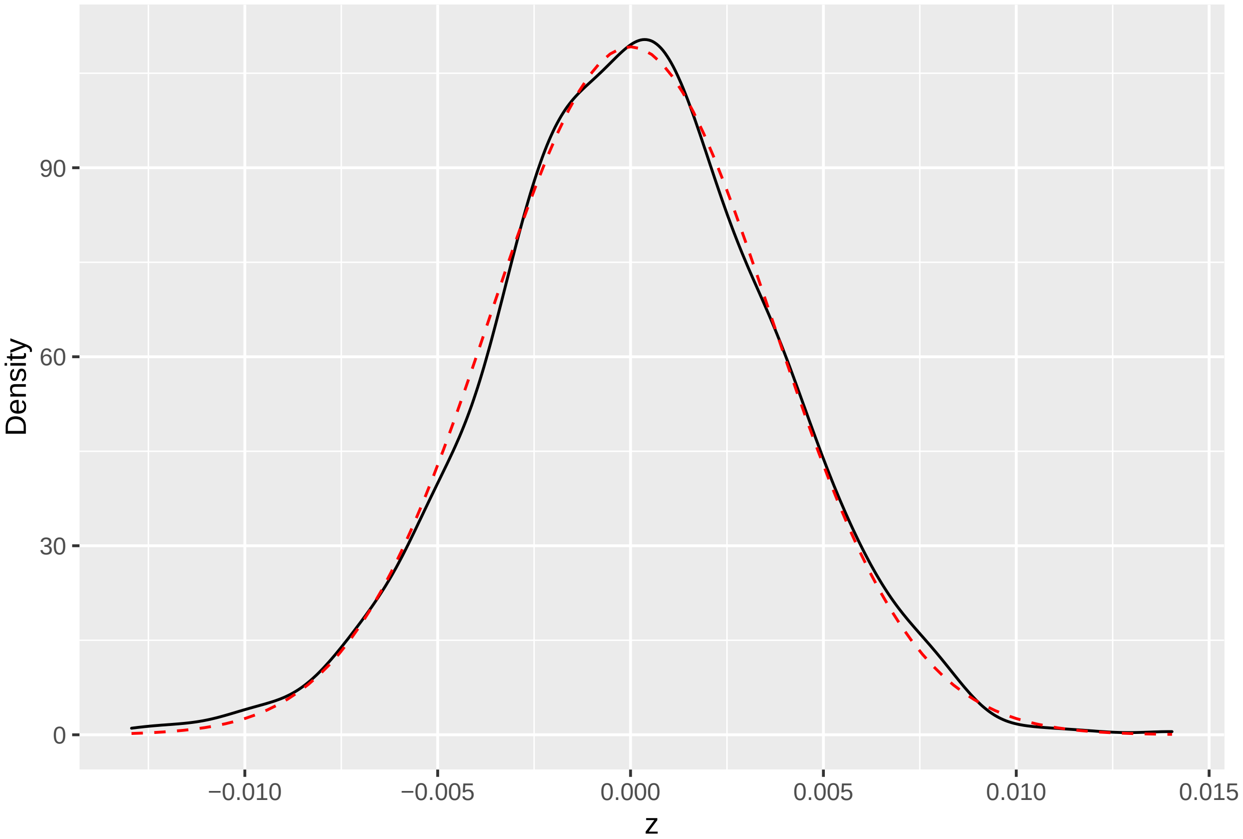
1.2 Intent-to-Treat

```
group_by(oregon, numhh_list) %>%
  summarize(diff = mean(vote_presidential_2008_1[treatment == 1]) -
                   mean(vote_presidential_2008_1[treatment == 0]),
            n = n()) %>%
  with(., weighted.mean(diff, n))
```

```
## [1] 0.007058013
```

1.3 Fisher Transformation

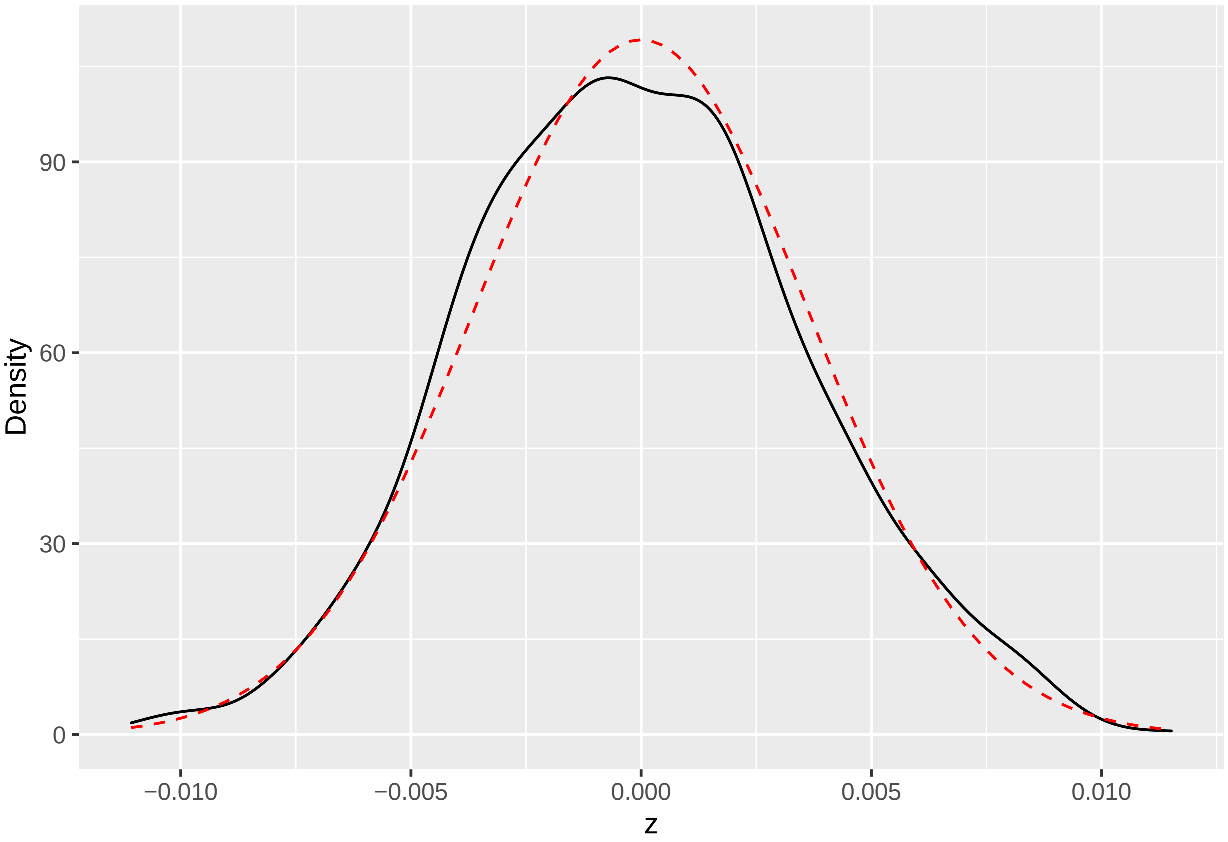
```
N <- nrow(oregon) # 74922
S <- 1000         # number of simulations
z <- replicate(S, {
  V <- rnorm(N)
  W <- rnorm(N)
  rho_hat <- cor(V, W)
  atanh(rho_hat)
})
ggplot(tibble(z)) + geom_density(aes(x = z)) + ylab("Density") +
  geom_function(fun = dnorm, args = list(sd = sqrt(1 / (N - 3))),
                color = "red", linetype = "dashed")
```



These distributions are essentially identical.

1.4 Sensitivity to the Bivariate Normal Assumption

```
z <- replicate(S, {
  V <- rbinom(N, size = 1, prob = 0.5)
  W <- rbinom(N, size = 1, prob = 0.5)
  rho_hat <- cor(V, W)
  atanh(rho_hat)
})
ggplot(tibble(z)) + geom_density(aes(x = z)) + ylab("Density") +
  geom_function(fun = dnorm, args = list(sd = sqrt(1 / (N - 3))),
                color = "red", linetype = "dashed")
```



Not quite as good, but still very similar.

1.5 First Stage of Instrumental Variables

```
summarize(oregon,
          rho_hat = cor(treatment, ohp_all_ever_nov2008 == "Enrolled"),
          z = atanh(rho_hat),
          p_value = pnorm(z, sd = sqrt(1 / (N - 3)), lower.tail = FALSE))
```

```
## # A tibble: 1 x 3
##   rho_hat     z p_value
##     <dbl> <dbl>   <dbl>
## 1   0.328 0.341       0
```

We reject the null hypothesis that the correlation is zero in the population.

1.6 Second Stage of Instrumental Variables

```
summarize(oregon,
          rho_hat = cor(treatment, vote_presidential_2008_1),
          z = atanh(rho_hat),
          p_value = pnorm(z, sd = sqrt(1 / (N - 3)), lower.tail = FALSE))
```

```
## # A tibble: 1 x 3
##   rho_hat       z p_value
##     <dbl>   <dbl>   <dbl>
## 1 0.00696 0.00696  0.0284
```

Since we would reject the null hypothesis that the correlation between the Medicaid lottery and voting is zero, convention allowed us to estimate its value and proceed as if the point estimate were true

```
LATE <- summarize(oregon,
                  LATE = sd(vote_presidential_2008_1) /
                    sd(ohp_all_ever_nov2008 == "Enrolled") *
                    cor(treatment, vote_presidential_2008_1) /
                    cor(treatment, ohp_all_ever_nov2008 == "Enrolled")) %>%
  pull(LATE)
LATE
```

```
## [1] 0.02549974
```

1.7 Joint Test

```
summarize(oregon,
          test_stat = (atanh(cor(treatment, vote_presidential_2008_1)) * sqrt(N - 3))^2 +
                      (atanh(cor(treatment, ohp_all_ever_nov2008 == "Enrolled")) * sqrt(N - 3))^2,
          p_value = pchisq(test_stat, df = 2, lower.tail = FALSE))
```

```
## # A tibble: 1 x 2
##   test_stat p_value
##       <dbl>   <dbl>
## 1     8708.       0
```

Again, we reject the null hypothesis that both correlations are zero.

1.8 Distribution of the Instrumental Variable Estimator

```
pnorm(LATE, sd = 0.013395, lower.tail = FALSE)
```

```
## [1] 0.02847622
```

We reject the null hypothesis that the LATE is zero in favor of the alternative hypthesis that the LATE is positive. It is perhaps reasonable to use a one-sided alternative in this situation because there is no theory to suggest that giving people an opportunity to obtain Medicaid makes people less likely to vote. However, with a two-sided alternative, we would have failed to reject the null hypothesis that the LATE is zero. It should be clear that even if the test statistic "objectively" has the null distribution under the null hypothesis, the choice of the null and alternative hypothesis is decided by the researcher and is not objective at all.

1.9 Confidence Interval Estimator for the LATE

```
qnorm(c(lower = 0.025, upper = 0.975), mean = LATE, sd = 0.013395)
```

```
##         lower         upper
## -0.0007539814  0.0517534537
```

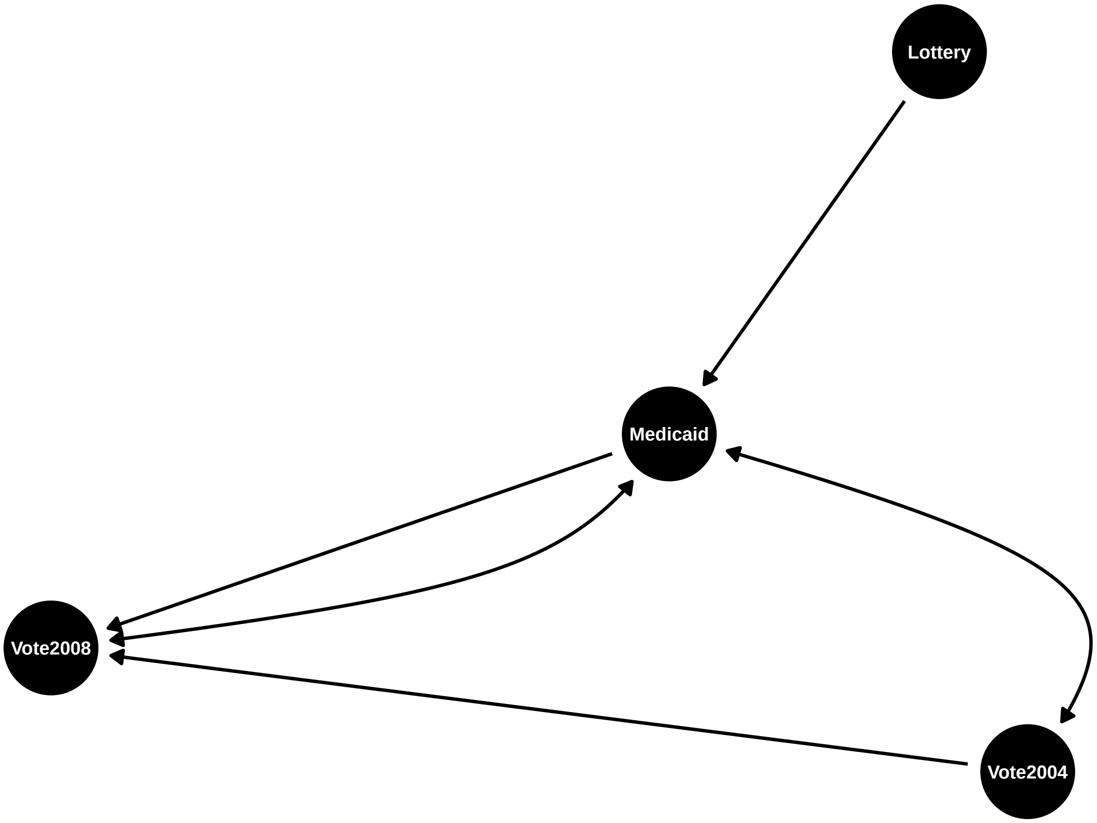
The lower bound is slightly negative because this confidence interval estimator is just a range of values for the null hypothesis that would not be rejected against a two-sided alternative. Thus, the conclusion is different than in the previous subproblem. However, with 95% of randomly sampled datasets of size N, the true LATE will be within the confidence interval evaluated at the *estimated* LATE and its *estimated* standard error. Note that these two estimates are random variables, so the location and width of the confidence intervals (plural) differ from one random dataset of size N to the next.

2 Bayesian

2.1 Directed Acyclic Graph

Call the `make_model` function in the CausalQueries package (and `set_restrictions`) to specify a DAG like that above, except ignore household size and instead take into account whether the person voted in the previous election (`oregon$vote_preperiod`). Plot this DAG.

```
library(CausalQueries)
model <- make_model("{ Lottery -> Medicaid -> Vote2008;
                       Vote2004 -> Vote2008 }") %>%
  set_confound(confound = list("Medicaid <-> Vote2008",
                               "Medicaid <-> Vote2004"))
ggdag(CausalQueries:::translate_dagitty(model), text_size = 2.5) + theme_void()
```



2.2 Posterior Distribution

```
oregon <- transmute(oregon,
                    Lottery = treatment,
                    Medicaid = ohp_all_ever_nov2008 == "Enrolled",
                    Vote2004 = vote_preperiod,
                    Vote2008 = vote_presidential_2008_1)
```

```
model <- update_model(model, data = oregon, seed = 20220307)
```

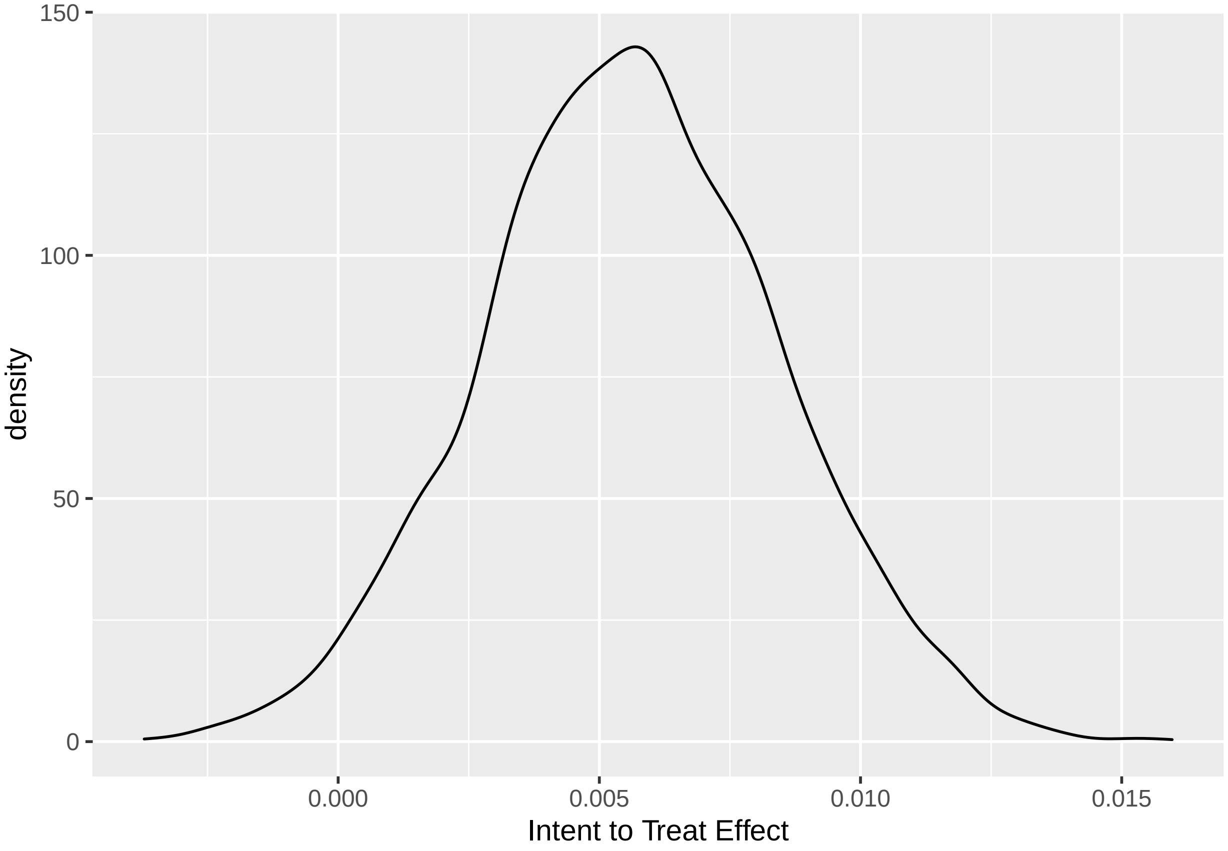
```
## Warning: The largest R-hat is 1.07, indicating chains have not mixed.
## Running the chains for more iterations may help. See
## https://mc-stan.org/misc/warnings.html#r-hat
## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## https://mc-stan.org/misc/warnings.html#bulk-ess
```

```
## Warning: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quant
## Running the chains for more iterations may help. See
## https://mc-stan.org/misc/warnings.html#tail-ess
```

It is fine on this homework to ignore such warnings, but not on future homeworks or actual research.

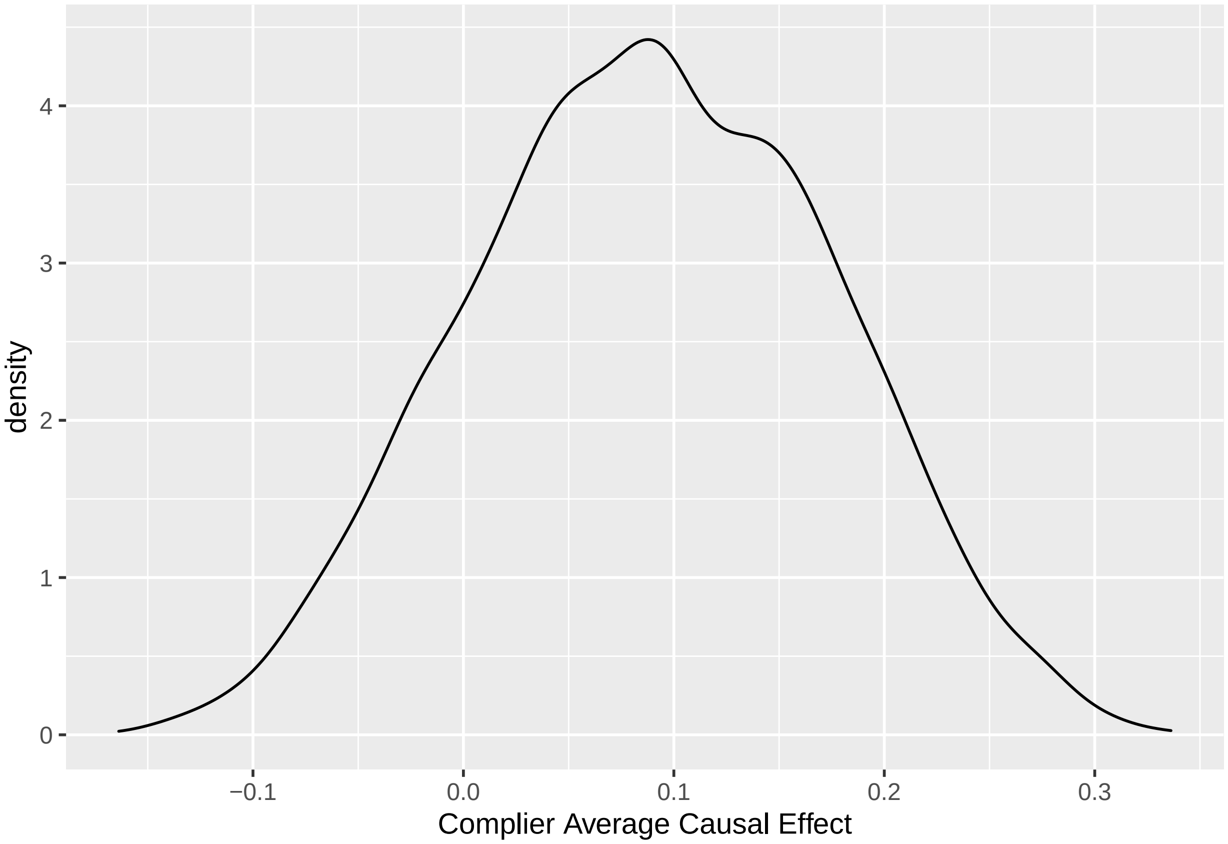
2.3 Interpretation

```
query_distribution(model, using = "posteriors",
                   query = "Vote2008[Lottery = 1] - Vote2008[Lottery = 0]") %>%
  as.data.frame %>%
  ggplot() + geom_density(aes(`.`)) + xlab("Intent to Treat Effect")
```



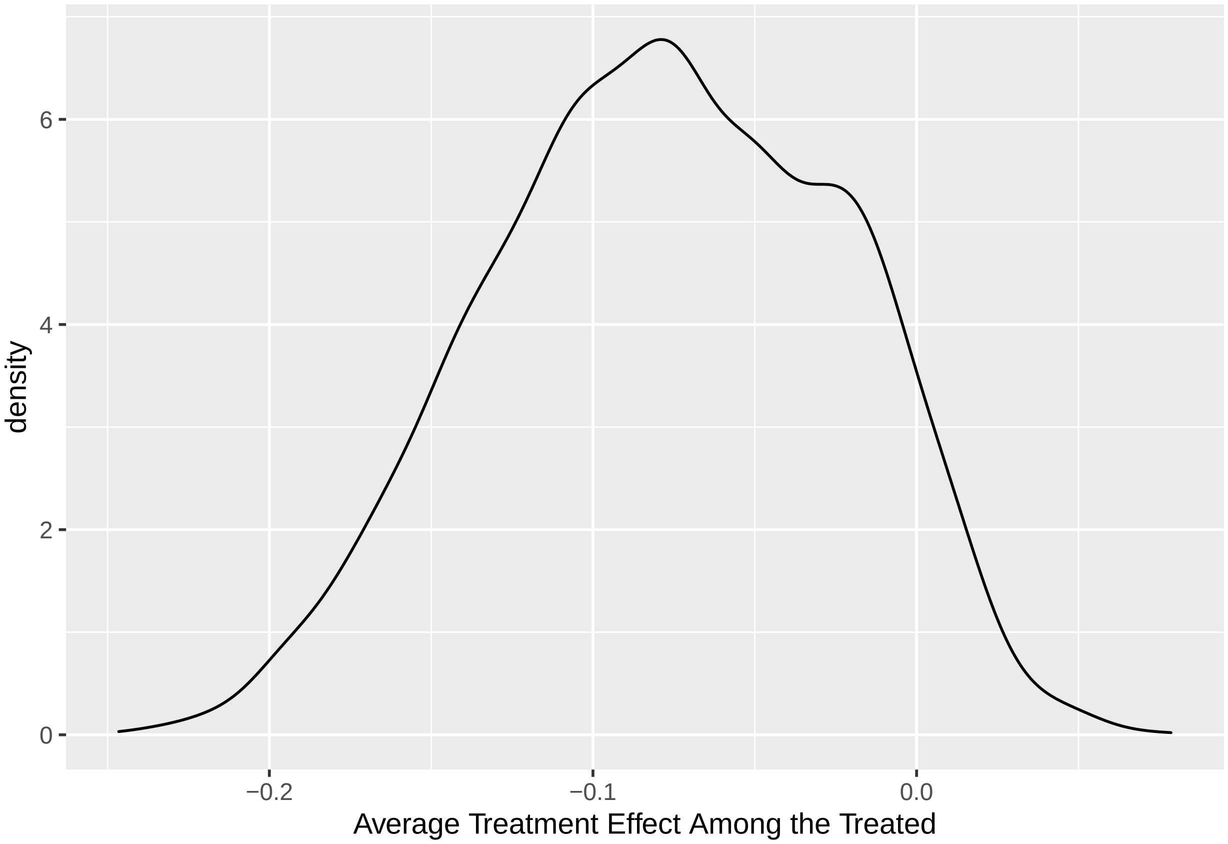
Very small, but presumably positive effect of winning the Medicaid lottery (marginalizing over whether people actually get Medicaid)

```
query_distribution(model, using = "posteriors",
                   query = "Vote2008[Medicaid = 1, Lottery = 1] -
                            Vote2008[Medicaid = 0, Lottery = 0]") %>%
  as.data.frame %>%
  ggplot() + geom_density(aes(`.`)) + xlab("Complier Average Causal Effect")
```



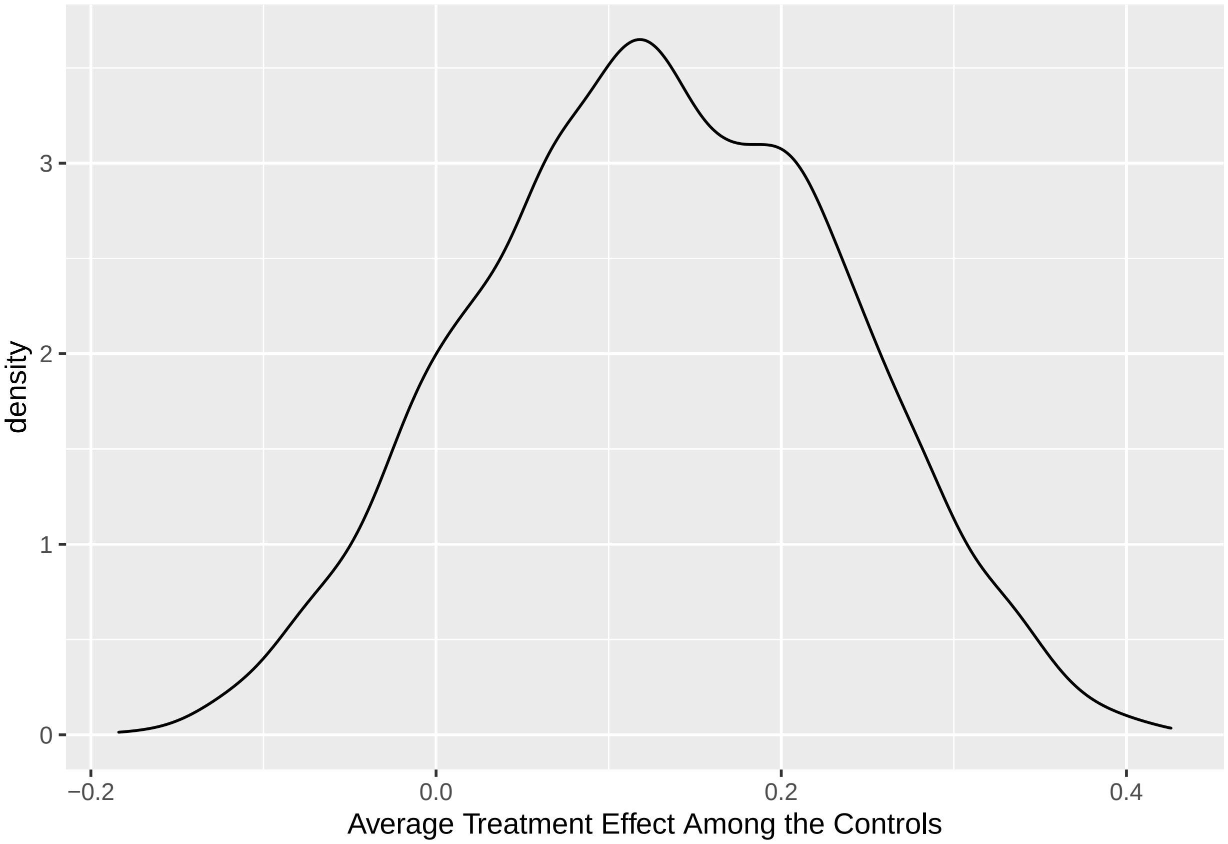
More likely to be positive than negative, but highly uncertain.

```r
query_distribution(model, using = "posteriors", given = "Medicaid == 1",
                   query = "Vote2008[Medicaid = 1] - Vote2008[Medicaid = 0]") %>%
  as.data.frame %>%
  ggplot() + geom_density(aes(`.`)) + xlab("Average Treatment Effect Among the Treated")
```



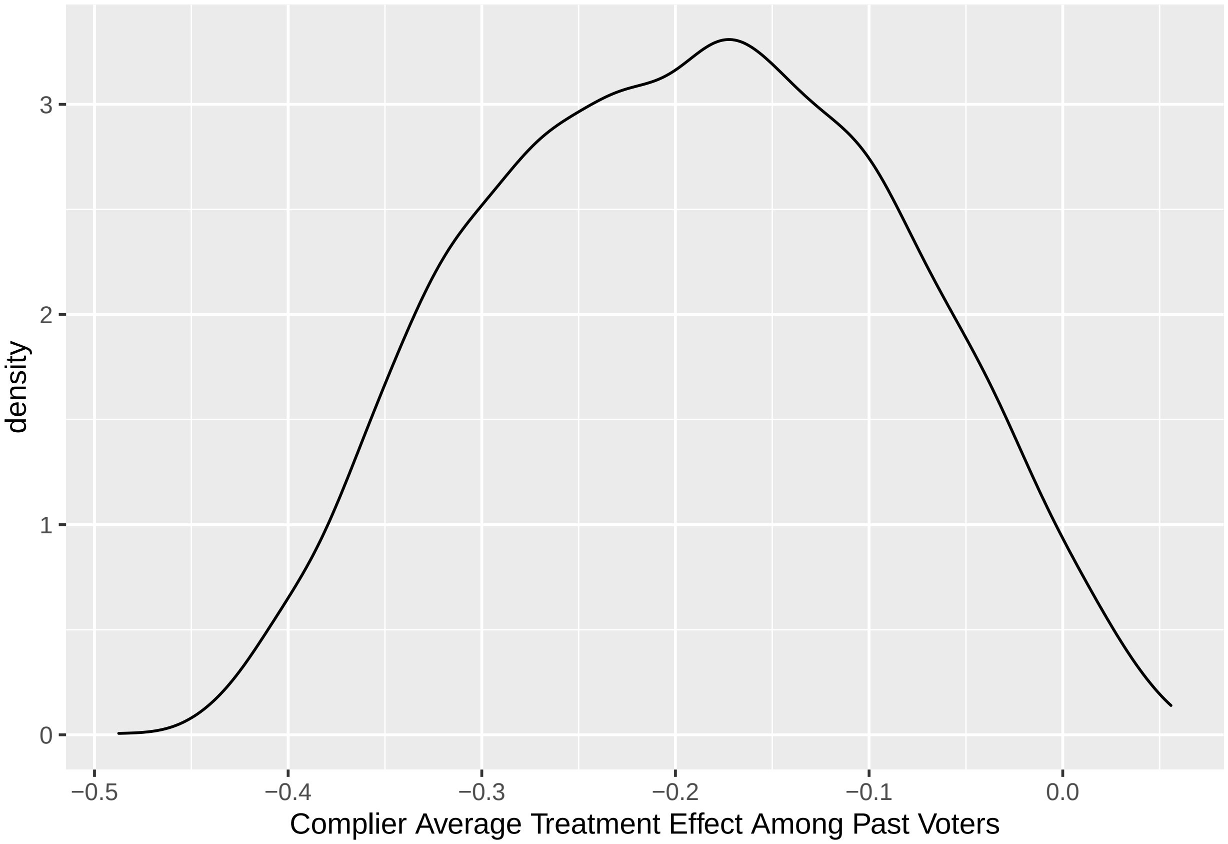
More likely to be negative than positive, which reflects the bias induced by Medicaid ← U → Vote2008.

```
query_distribution(model, using = "posteriors", given = "Medicaid == 0",
                   query = "Vote2008[Medicaid = 1] - Vote2008[Medicaid = 0]") %>%
  as.data.frame %>%
  ggplot() + geom_density(aes(`.`)) + xlab("Average Treatment Effect Among the Controls")
```



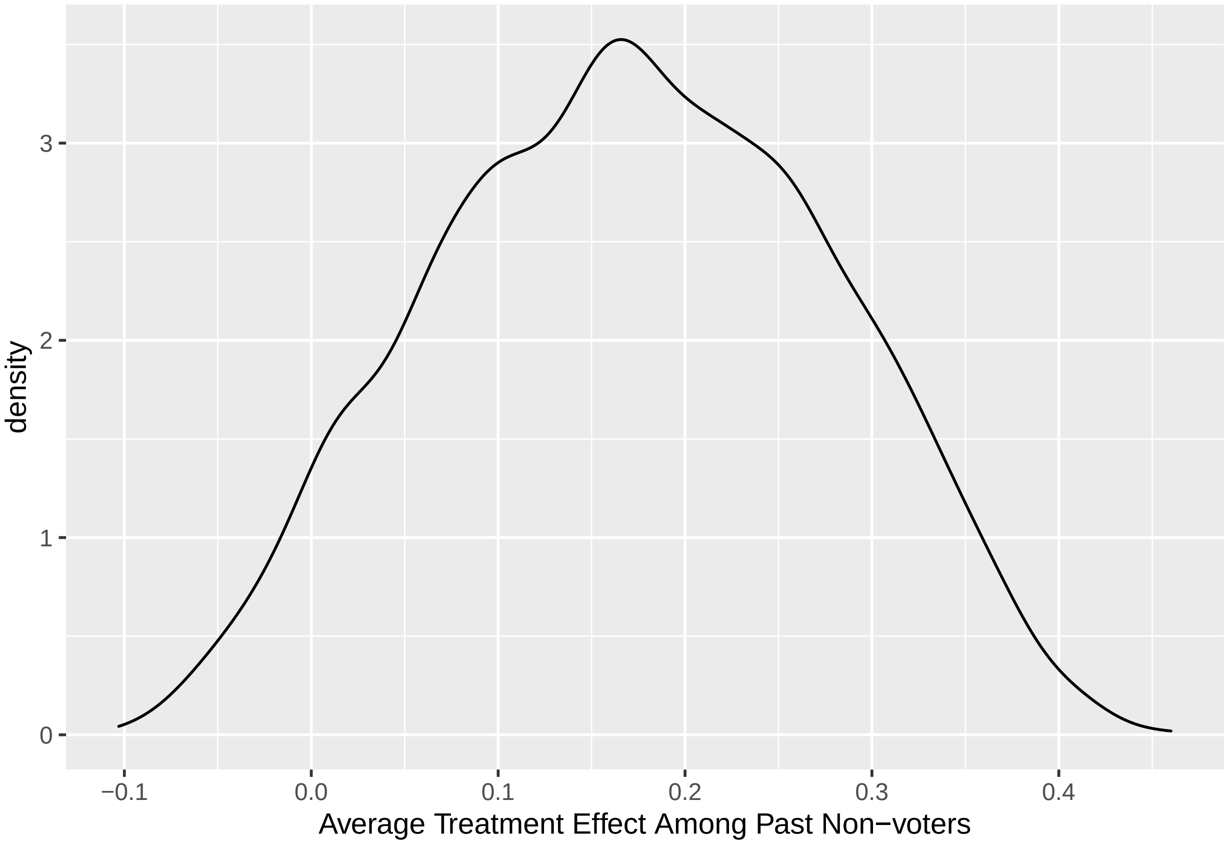
More likely to be positive than negative but very uncertain overall and also subject to the bias from Medicaid ← U → Vote2008.

```
query_distribution(model, using = "posteriors", given = "Vote2004 == 1",
                   query = "Vote2008[Medicaid = 1, Lottery = 1] -
                            Vote2008[Medicaid = 0, Lottery = 0]") %>%
  as.data.frame %>%
  ggplot() + geom_density(aes(`.`)) + xlab("Complier Average Treatment Effect Among Past Voters")
```



Very likely to be negative, which is hard to explain because this estimator should be unbiased if you only utilize compliers (it is a similar distribution if you do not restrict it to compliers, which is fine on this homework). But the magnitude is quite uncertain.

```
query_distribution(model, using = "posteriors", given = "Vote2004 == 0",
                   query = "Vote2008[Medicaid = 1] - Vote2008[Medicaid = 0]") %>%
  as.data.frame %>%
  ggplot() + geom_density(aes(`.`)) + xlab("Average Treatment Effect Among Past Non-voters")
```



Very likely to be positive, but the magnitude is quite uncertain. Overall, we are left with a story that the causal effect of Medicaid on voting is likely negative for people who voted in the 2004 election, positive for people who did not voted in the 2004 election, and ambiguous overall.

I think that I misinterpreted the paper as saying `oregon$vote_preperiod` indicated whether someone voted in an election before the 2008 Presidential election, when in fact that only used that variable when analyzing the 2010 midterm elections. Thus, we are conditioning on a descendant of Medicaid in the last two queries, which would introduce bias.

2.4 Citizenship

Legal immigrants with sufficiently low income are eligible for Medicaid in Oregon but are not eligible to vote (unless they become citizens). However, the researchers did not collect data on which people are citizens, which is difficult because collecting such data tends to make legal immigrants less likely to participate in a study (perhaps because they might be living in the same household as someone who is undocumented). Discuss to what extent not including a citizenship variable in the DAG is a problem for estimating the Average Treatment Effect of Medicaid with this dataset.

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

```
make_model("Medicaid -> Registered -> Voted; Citizen -> Registered")
```

Noncitizens cannot register to vote, making them a clear examples of “never takers”, people that will not vote irrespective of whether they benefit from Medicaid.

It would be nice if citizenship 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 the Medicaid lottery is randomized means that it has no ancestor in common with citizenship, so you can still identify the (local) average treatment effect of Medicaid.

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 estimate the causal effect among citizens only.

3 Frequentist vs. Bayesian

Based on the experience in Oregon, you might consider expanding the number of people Medicaid nationwide, which is more-or-less what Obamacare intended to do starting in 2010. That decision is similar to the FDA's decision of whether to approve the Biontech / Pfizer vaccine for children under the age of five. In both cases, there is a cost to not acting that has to be considered along with the benefits, both of which are uncertain. The entire posterior distribution needs to be pushed through the utility function, whose expectation can then be taken in order to make a decision. That decision may have little to do with whether you can reject the null hypothesis that the causal effect of Medicaid is zero with the available data. The FDA is operating under the usual convention of proceeding as if the effect of the vaccine is zero if you cannot reject the null hypothesis that it is zero, which has no Frequentist or other justification. If the null hypothesis were rejected, the FDA would operate under the usual convention of proceeding as if effect of the vaccine is the point estimate. In situations where the uncertainty is considerable and the utility function is highly skewed, the Bayesian decision theory approach will often yield a different decision.