Quant II

Lab 10: Imputation, bounds, limited dependent variables

Giacomo Lemoli

April 25, 2022

Today's plan

- Missing data, imputation, and multiple imputation
- Bounds on causal effects
- Limited dependent variables, marginal and causal effects

Working example

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Sometimes Less Is More: Censorship, News Falsification, and Disapproval in 1989 East Germany

Christian Gläßel University of Mannheim
Katrin Paula University of Mannheim

Abstract: *Does more media censorship imply more regime stability? We argue that censorship may cause mass disapproval for censoring regimes. In particular, we expect that censorship backfires when citizens can falsify media content through alternative sources of information. We empirically test our theoretical argument in an autocratic regime—the German Democratic Republic (GDR). Results demonstrate how exposed state censorship on the country's emigration crisis fueled outrage in the weeks before the 1989 revolution. Combining original weekly approval surveys on GDR state television and daily content data of West German news programs with a quasi-experimental research design, we show that recipients disapproved of censorship if they were able to detect misinformation through conflicting reports on Western television. Our findings have important implications for the study of censoring systems in contemporary autocracies, external democracy promotion, and campaigns aimed at undermining trust in traditional journalism.*

Baseline results (full data)

```
library(haven); library(estimatr); library(modelsummary); library(tidyverse); library(ggpubr)
d <- read_dta("Censorship.dta")

# Reproduce col 1 of table 1
mod <- lm_robust(ak_rating ~ censorship + liberalization, clusters = hh, se_type = "stata",
                 data = d)
modelsummary(mod, coef_omit = "Int", output = "markdown")
```

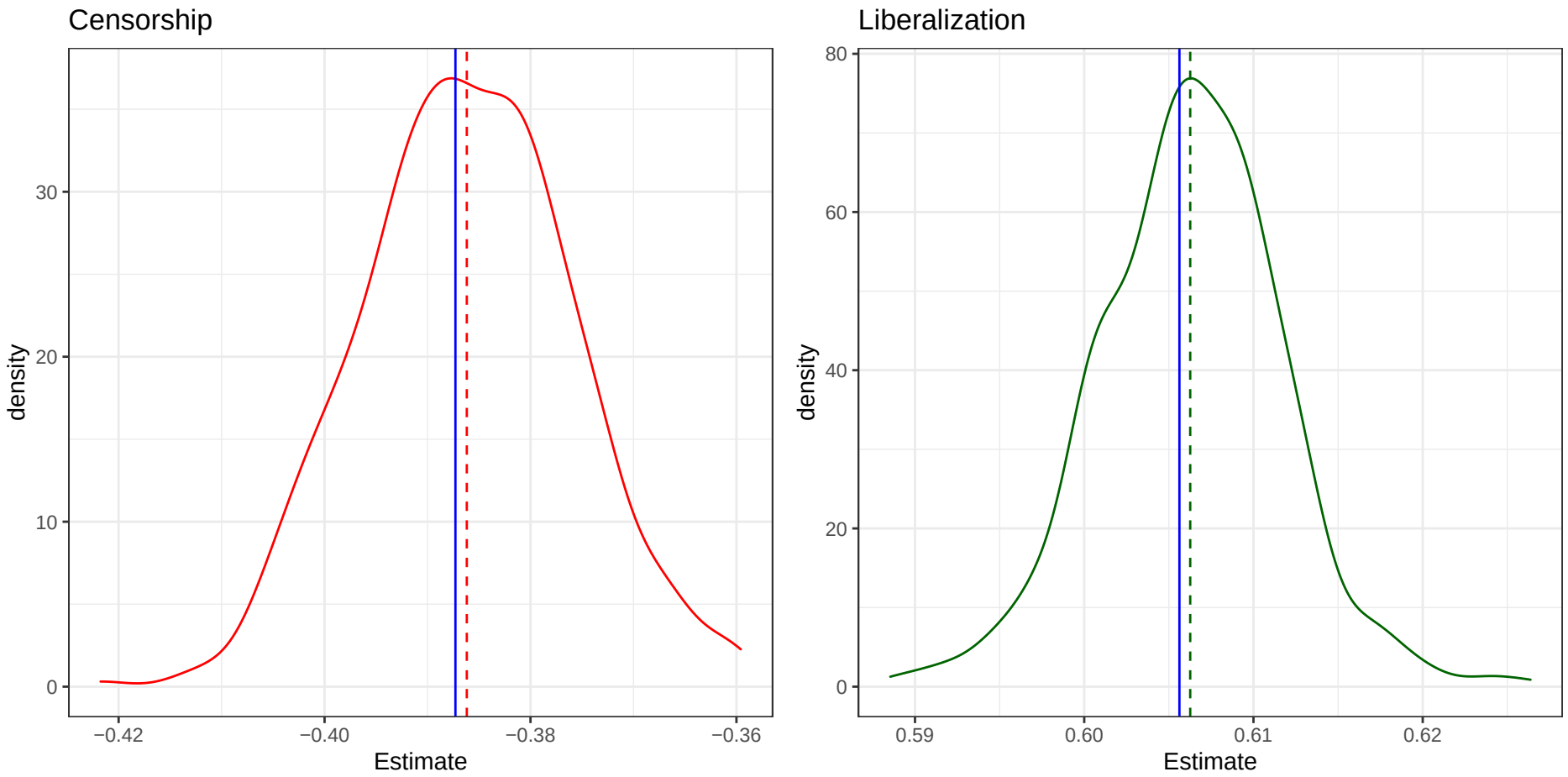
	Model 1
censorship	-0.387
	(0.036)
liberalization	0.606
	(0.021)
Num.Obs.	17551
R2	0.235
R2 Adj.	0.235
se_type	stata

Missing Completely at Random

```
set.seed(123)

nboot <- 500

cens <- lib <- rep(NA,nboot)

# 20% of outcome values are missing in the data
for (i in 1:nboot){
  d <- d %>% mutate(newout = ifelse(runif(nrow(d),0,1)<0.2, NA, ak_rating))
  fit <- update(mod, newout ~ .)
  cens[i] <- coef(fit)["censorship"]
  lib[i] <- coef(fit)["liberalization"]
}

plot1 <- ggplot(as.data.frame(cens)) + geom_density(aes(x=cens), col="red") +
  labs(x="Estimate", title = "Censorship") +
  geom_vline(xintercept = mean(cens), col="red", linetype = "dashed") +
  geom_vline(xintercept = coef(mod)["censorship"], col="blue") + theme_bw()

plot2 <- ggplot(as.data.frame(lib)) + geom_density(aes(x=lib), col="dark green") +
  labs(x="Estimate", title = "Liberalization") +
  geom_vline(xintercept = mean(lib), col="dark green", linetype = "dashed") +
  geom_vline(xintercept = coef(mod)["liberalization"], col="blue") + theme_bw()
```

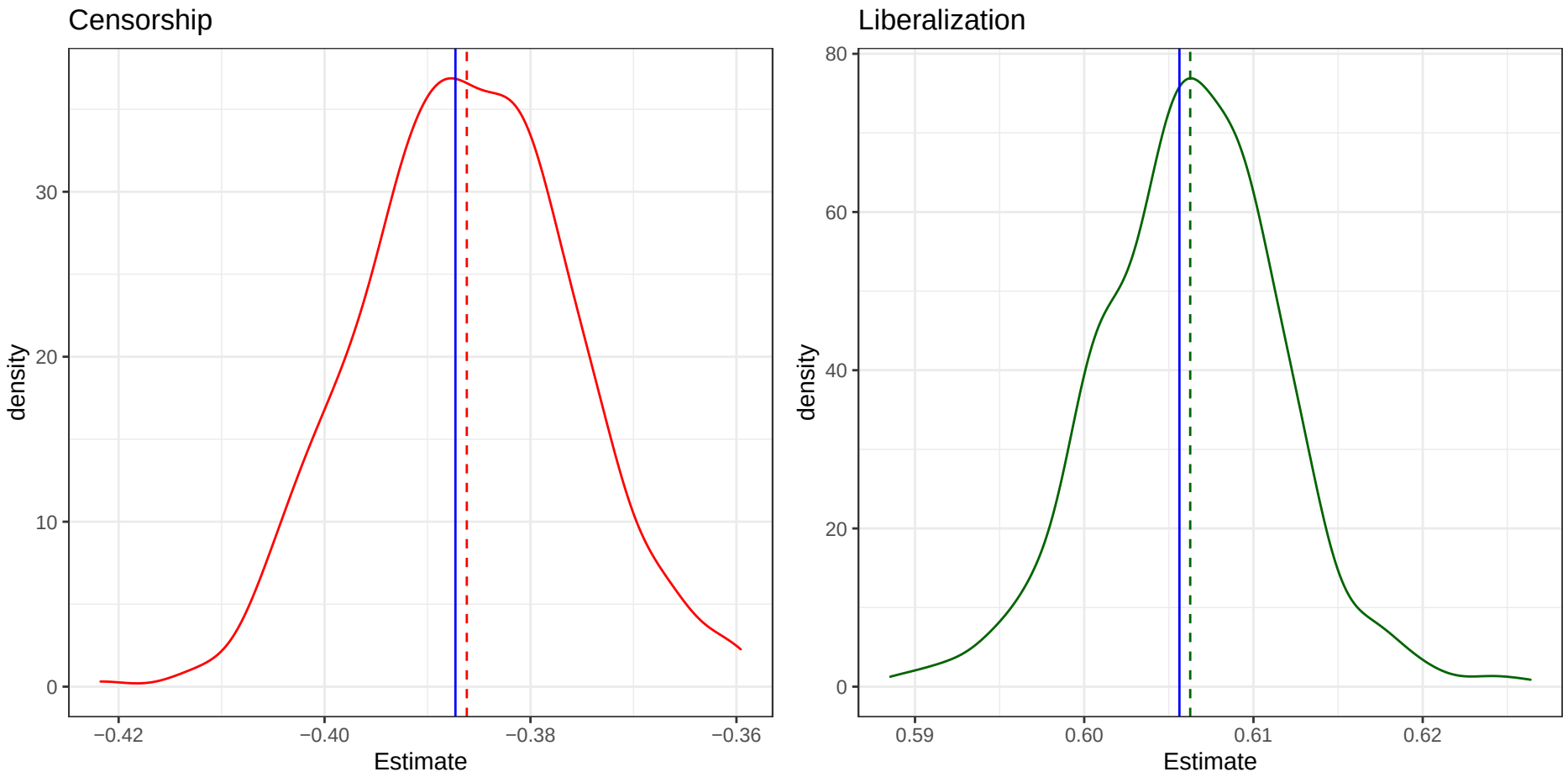
Missing Completely at Random

```
ggarrange(plot1, plot2)
```



Missing Completely at Random

```
ggarrange(plot1, plot2)
```



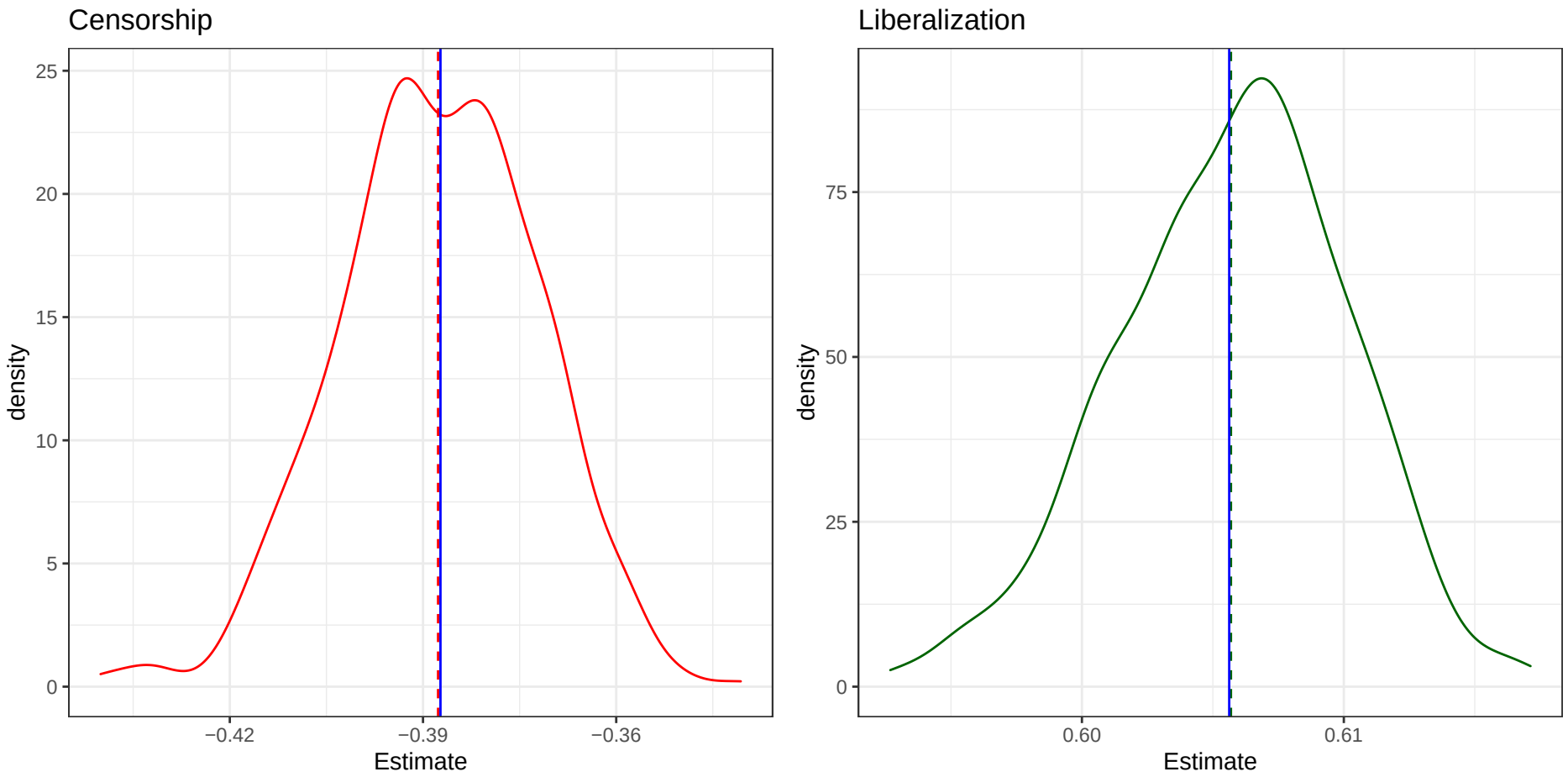
With MCAR we are good

Missing at Random

```
# Suppose during dictatorship people express negative attitudes by non-responding
set.seed(123)

nboot <- 500

cens <- lib <- rep(NA,nboot)

for (i in 1:nboot){
  d <- d %>% mutate(newout = ifelse(rnorm(nrow(d), mean = d$censorship)>1.15,
                                    NA, ak_rating))
  fit <- update(mod, newout ~ .)
  cens[i] <- coef(fit)["censorship"]
  lib[i] <- coef(fit)["liberalization"]
}

plot1 <- ggplot(as.data.frame(cens)) + geom_density(aes(x=cens), col="red") +
  labs(x="Estimate", title = "Censorship") +
  geom_vline(xintercept = mean(cens), col="red", linetype = "dashed") +
  geom_vline(xintercept = coef(mod)["censorship"], col="blue") + theme_bw()

plot2 <- ggplot(as.data.frame(lib)) + geom_density(aes(x=lib), col="dark green") +
  labs(x="Estimate", title = "Liberalization") +
  geom_vline(xintercept = mean(lib), col="dark green", linetype = "dashed") +
  geom_vline(xintercept = coef(mod)["liberalization"], col="blue") + theme_bw()
```

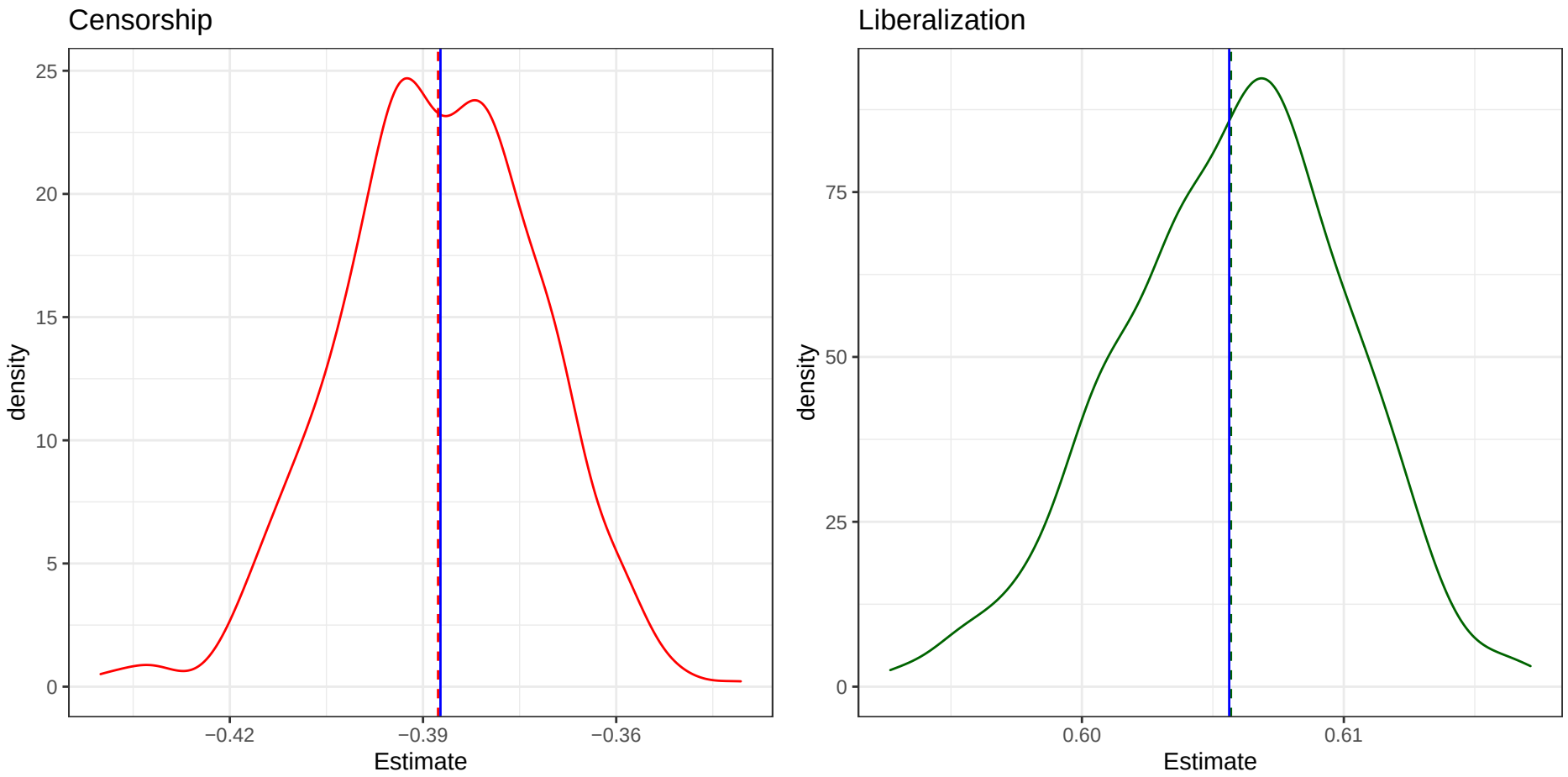
Missing at Random

```
ggarrange(plot1, plot2)
```



Missing at Random

```
ggarrange(plot1, plot2)
```



MAR is also fine if we missingness is random conditional on the model variables.

Point identification with missingness

If MCAR, we don't need to worry, in expectation the estimates are unbiased

If MAR, we can use covariates that determine missingness to impute the missing values of the outcome

Imputation with regression

```r
# Predicted values from the regression
d <- d %>% mutate(predicted = predict(fit, newdata = d))

# Replace missing outcomes with imputed values
d <- d %>% mutate(impy = ifelse(is.na(newout), predicted, newout))
```

In Stata you can use the post-estimation command `predict`

```
help predict
reg newout censorship liberalization, cluster(hh)
predict predicted, xb
replace newout = predicted if missing(newout) & !missing(predicted)
```

Multiple imputation with MICE

```
library(mice)
m1 <- mice(data = zap_labels(d[,c("newout", "censorship", "liberalization",
                                  "sed", "education")]),
           maxit = 5, printFlag = FALSE)
d_imp <- complete(m1)

length(d$newout[is.na(d$newout)])
```

```
## [1] 2992
```

```
length(d_imp$newout[is.na(d_imp$newout)])
```

```
## [1] 0
```

Bounds

If we are not willing to make strong assumptions about the nature of the DGP to justify MAR, we can still estimate bounds on causal effects.

Manski bounds

Also called the "worst case" bounds.

Idea: fill missing outcomes using the bounds of the potential outcomes to give extreme (smallest and largest) cases of the ATE

Manski bounds

Formally:

$$\beta^L \leq ATE \leq \beta^H$$

$$\beta^L = \{\mu_{1,obs} Pr[R_{1i} = 1] + y_1^L Pr[R_{1i=0}]\} - \{\mu_{0,obs} Pr[R_{0i} = 1] + y_0^H Pr[R_{0i} = 0]\}$$

$$\beta^H = \{\mu_{1,obs} Pr[R_{1i} = 1] + y_1^H Pr[R_{1i=0}]\} - \{\mu_{0,obs} Pr[R_{0i} = 1] + y_0^L Pr[R_{0i} = 0]\}$$

$\{y_t^L, y_t^H\}$ are straightforward when the outcome is naturally bounded, e.g. binary (just replace every missing y with 0 or 1). Otherwise, they are not. And if the bounds are very large, the set of possible effects can be so large to be practically meaningless.

Manski bounds

Implementation:

- Manually
- A few packages in R: `ATbounds` (vignette), `attrition` by Alex Coppock (download from GitHub)

```
#install.packages("ATbounds")
#library(ATbounds)

#install.packages(devtools)
#library(devtools)
#install_github("acoppock/attrition")
```

Lee bounds

Motivation: provide bounds on the causal effect even with unbounded outcome.

Trade-off: impose more structure on the data, but weaker than "exclusion restrictions" necessary for other strategies.

Relies on a monotonicity assumption similar to that used for LATE identification, but in this case it serves to solve sample selection.

Lee bounds

Assumptions:

$$(Y_{1i}^*, Y_{0i}^*, R_{1i}, R_{0i}) \perp D$$
$$R_{1i} \geq R_{0i}$$

Under these assumptions, Lee (2009) proves that we can identify bounds for the ATE *for the sub-population* $\{R_{0i} = 1, R_{1i} = 1\}$.

$$\Delta_0^{LB} = E[Y|D = 1, R = 1, Y \leq y_{1-p0}] - E[Y|D = 0, R = 1]$$
$$\Delta_0^{UB} = E[Y|D = 1, R = 1, Y \geq y_{p0}] - E[Y|D = 0, R = 1]$$
$$p_0 = \frac{Pr[R = 1|D = 1] - Pr[R = 1|D = 0]}{Pr[R = 1|D = 1]}$$

Lee bounds

Implementation:

- R: `leebounds` (install_github("vsemenova/leebounds"))
- Stata: `leebounds`

Lee bounds

```
library(leebounds)

input <- d %>% mutate(sel = ifelse(is.na(newout),0,1)) %>%
  select(c(censorship, newout, sel)) %>%
  rename(treat = censorship, outcome = newout, selection = sel)

do.call("rbind", leebounds(input))
```

```
##                      62.58541%
## lower_bound         -1.1071905
## upper_bound         -0.5783773
## p0                   1.5978164
## trimmed_mean_upper   4.2069751
## trimmed_mean_lower   3.6781620
## mean_no_trim         3.0997846
## odds                 0.1690535
## yp0                  4.0000000
## y1p0                 4.0000000
## s0                   0.8769733
## s1                   0.5488574
## prop0                0.8553929
## prop1                0.1446071
```