Marginal Treatment Effects Part I: Background

Francis DiTraglia

Oxford Economics Summer School 2022

Treatment Effects: The Big Picture

The Best We Can Do?

- Ideally, want to learn *individual treatment effects* but we can't: fundamental problem of causal inference!
- Barring that, want to learn *distribution* of treatment effects, but we can't: fundamental problem of causal inference! (Can bound them: Notes Chapter 3)
- ATE (or conditional ATE) usually considered best we can do. Identified by "gold standard" placebo controlled, randomized trial with perfect compliance.

We can't force people!

- Even when treatment is randomly assigned, can't force people to take it: *randomized encouragement design*
- Intent-to-treat (ITT) effect: causal effect of *offering* treatment. "Diluted" by people offered who don't take (typically assume exclusion restriction).

Better LATE than Nothing?

- IV allows us to go beyond ITT effects, but if treatment effects are heterogenous, we recover the LATE: average effect for *compliers*
- Is the LATE an interesting quantity? Maybe, maybe not.
- Recently: lots of interest in extrapoLATE-ing "beyond LATE" to more interesting causal parameters. **That is the topic of this lecture and the next one**
- Many issues here, but most important: **what causal parameters should we be interested in and why?**

Two Key Questions

1. What is it *possible* to learn form data? (Identification)
2. What do we plan to *do* with our causal effect? (Less commonly asked)

Causal Effects are for *Decisionmaking*

Example Causal Question

- What is the causal effect of cognitive behavioral therapy (CBT) on anxiety?

Individual's Decision Problem

- You have anxiety, and need to decide whether to get CBT ($D = 1$) or not ($D = 0$). Weigh the costs against benefits. Chamberlain (2011)
- You are probably interested in the ATE or conditional ATE: *on average, what is the treatment effect for a person like me?*
- Side point: experiment only tells you useful information under a *consistency condition*, i.e. *choosing* treatment has the same effect as *being allocated* treatment.
- Crucial, if obvious, feature: *you can force yourself to take treatment*

Causal Effects are for *Decisionmaking*

Example Causal Question

- What is the causal effect of cognitive behavioral therapy (CBT) on anxiety?

Policymaker's Decision Problem

- Should we expand access to CBT on the UK National Health Service (NHS)? Weigh the costs against the benefits.
- We can't force people with anxiety to get CBT by making it more widely available so the ATE isn't the relevant quantity.
- If we expand access, some more people will be treated. Policy question is: **what is the average benefit, per additional person enrolled, of expanding access?**
- When treatment is *voluntary*, it becomes crucial for policy analysis to understand how treatment effects may correlate with willingness to *take up* treatment.

Causal Effects for Policymaking? TOT and TUT Effects

Treatment on the Treated (TOT aka ATT)

- Existing program; only some of those eligible choose to enroll. If we **eliminated** the program, how much **worse off** would current participants be?
- Average effect of a program or policy for those who currently choose to enroll.
- Equals LATE under one-sided non-compliance: no always-takers

Treatment on the Untreated (TUT aka ATU)

- Existing program; only some of those eligible choose to enroll. If we **forced** all non-participants to enroll, how much **better off** would they be?
- Average effect of a program of policy for those who currently choose **not** to enroll.
- Equals LATE under one-sided non-compliance: no never-takers
- E.g. increase in UK minimum school-leaving age from 15 to 16 (September 1972).

A "Textbook" Model

$$
\begin{aligned}
Y_0 &= \mu_0 + U_0 \\
Y_1 &= \mu_1 + U_1 \\
D &= 1\{\gamma_0 + \gamma_1 Z > V\} \\
Y &= (1 - D) Y_0 + D Y_1
\end{aligned}
\qquad
\begin{aligned}
Z &\sim \text{Bernoulli}(q) \perp\!\!\!\perp (V, U_0, U_1) \\
\begin{bmatrix} V \\ U_0 \\ U_1 \end{bmatrix} &\sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_0 \rho_0 & \sigma_1 \rho_1 \\ & \sigma_0^2 & \sigma_{01} \\ & & \sigma_1^2 \end{bmatrix} \right)
\end{aligned}
$$

- Heckman, Tobias & Vytlacil (2001), Angrist (2004)
- Treatment effects $(Y_1 - Y_0)$ are heterogeneous, ATE $= \mu_1 - \mu_0$.
- Treatment take-up D depends on:
  1. Binary instrument / encouragement Z
  2. Heterogeneous "cost" / "resistance to treatment" V ("free" normalization)
- Normality gives simple formulas. Next lecture: relax this assumption.

Simulation: $\mu_1 = \mu_0 = 0$, $\sigma_0 = \sigma_1 = 1$, $\sigma_{01} = 1/2$

```
library(mvtnorm)
library(tidyverse)
rho0 <- 0.5
rho1 <- 0.2

S <- matrix(c(1, rho0, rho1,
              rho0, 1, 0.5,
              rho1, 0.5, 1), 3, 3, byrow = TRUE)
set.seed(1983)
sims <- rmvnorm(5e3, sigma = S)
colnames(sims) <- c('V', 'Y0', 'Y1')
sims <- as_tibble(sims)
sims <- sims %>%
  mutate(Delta = Y1 - Y0)
```

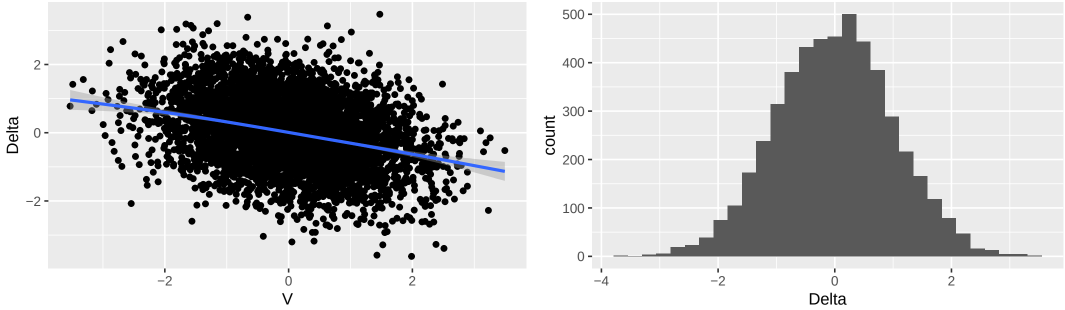
```r
sims
```

```
## # A tibble: 5,000 x 4
##           V       Y0       Y1  Delta
##       <dbl>    <dbl>    <dbl>  <dbl>
##  1 -0.122   -0.399    1.08    1.48
##  2 -0.506   -1.10     1.49    2.59
##  3  0.00457 -0.121   -0.456  -0.335
##  4 -0.549   -0.248   -0.899  -0.651
##  5  1.95    -0.0948  -0.675  -0.580
##  6  0.561    0.112   -0.615  -0.726
##  7 -0.238   -0.439   -1.53   -1.10
##  8 -1.46    -1.23    -0.0548  1.17
##  9 -0.336   -0.891    1.53    2.42
## 10 -0.845   -0.274    0.637   0.911
## # ... with 4,990 more rows
```

```
DV_scatter <- sims %>%
  ggplot(aes(x = V, y = Delta)) +
  geom_point() +
  geom_smooth()

Dhist <- sims %>%
  ggplot(aes(x = Delta)) +
  geom_histogram()
```

```r
library(gridExtra)
grid.arrange(DV_scatter, Dhist, ncol = 2)
```



Any Parameter values

- Δ is normally distributed; Δ and V are linearly dependent (jointly normal).

These Parameter Values

- ATE is zero; higher cost/resistance $V \implies$ lower treatment effect Δ

Properties of the Textbook Model

$$
\begin{aligned}
Y_0 &= \mu_0 + U_0 \\
Y_1 &= \mu_1 + U_1 \\
D &= 1\{\gamma_0 + \gamma_1 Z > V\} \\
Y &= (1 - D)Y_0 + DY_1
\end{aligned}
$$

$$
\begin{gathered}
Z \sim \text{Bernoulli}(q) \perp\!\!\!\perp (V, U_0, U_1) \\
\begin{bmatrix} V \\ U_0 \\ U_1 \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_0\rho_0 & \sigma_1\rho_1 \\ & \sigma_0^2 & \sigma_{01} \\ & & \sigma_1^2 \end{bmatrix} \right)
\end{gathered}
$$

Implications

- $\Delta \equiv Y_1 - Y_0 \sim \text{Normal}(\mu_1 - \mu_0, \sigma_0^2 + \sigma_1^2 - 2\sigma_{01})$
- $\text{Cov}(\Delta_i, V_i) = \text{Cov}(Y_{1i}, V_i) - \text{Cov}(Y_{0i}, V_i) = \sigma_1\rho_1 - \sigma_0\rho_0$

LATE for the Textbook Model

- LATE = average effect for *compliers*: people induced to take treatment by Z.
- Since $D = 1(\gamma_0 + \gamma_1 Z > V)$, compliers are defined by $\gamma_0 \leq V < \gamma_0 + \gamma_1$
- **Depends on the particular instrument** through γ_0, γ_1

```r
gamma0 <- -1
gamma1 <- 1.5
sims <- sims %>%
  mutate(complier = (V >= gamma0) & (V < gamma0 + gamma1))
```

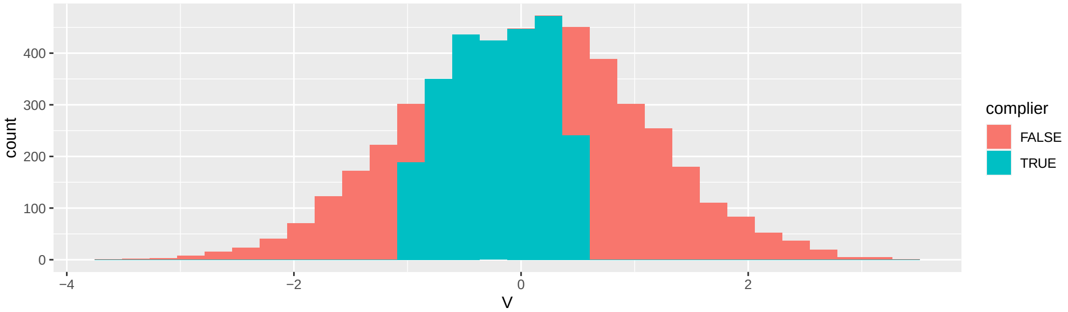
Who's a complier when $\gamma_0 = -1$ and $\gamma_1 = 1.5$?

```
sims
```

```
## # A tibble: 5,000 x 5
##            V      Y0      Y1  Delta complier
##        <dbl>   <dbl>   <dbl>  <dbl> <lgl>
##  1 -0.122   -0.399   1.08    1.48  TRUE
##  2 -0.506   -1.10    1.49    2.59  TRUE
##  3  0.00457 -0.121  -0.456  -0.335 TRUE
##  4 -0.549   -0.248  -0.899  -0.651 TRUE
##  5  1.95    -0.0948 -0.675  -0.580 FALSE
##  6  0.561    0.112  -0.615  -0.726 FALSE
##  7 -0.238   -0.439  -1.53   -1.10  TRUE
##  8 -1.46    -1.23   -0.0548  1.17  FALSE
##  9 -0.336   -0.891   1.53    2.42  TRUE
## 10 -0.845   -0.274   0.637   0.911 TRUE
## # ... with 4,990 more rows
```

Whos's a complier when $\gamma_0 = -1$, $\gamma_1 = 1.5$?

```r
ggplot(sims, aes(x = V, fill = complier)) +
  geom_histogram()
```

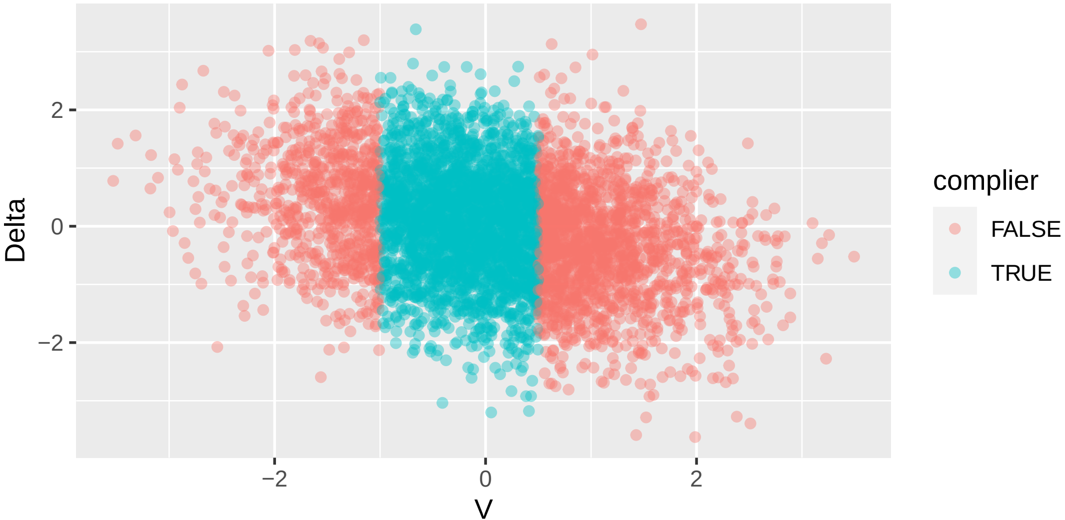



```r
# Share of compliers
pnorm(gamma0 + gamma1) - pnorm(gamma0)
```

```
## [1] 0.5328072
```

Who's a complier when $\gamma_0 = -1$ and $\gamma_1 = 1.5$?

```
ggplot(sims, aes(x = V, y = Delta, col = complier)) +
  geom_point(alpha = 0.4)
```



Average Treatment Effects by Complier Status: $\gamma_0 = -1$, $\gamma_1 = 1.5$

```
sims %>%
  group_by(complier) %>%
  summarize(mean(Y1 - Y0)) %>%
  knitr::kable(digits = 3)
```

complier	mean(Y1 - Y0)
FALSE	-0.083
TRUE	0.068

Different Instrument, Different LATE: $\gamma_0 = -1$, Varying γ_1

```
get_LATE <- function(gamma1) {
  sims %>%
    mutate(complier = (V >= -1) & (V < -1 + gamma1)) %>%
    filter(complier) %>%
    summarize(LATE = mean(Y1 - Y0)) %>%
    pull()
}

gamma1_seq <- c(0.75, 1, 1.25, 1.5, 1.75, 2)
LATE <- map_dbl(c(0.75, 1, 1.25, 1.5, 1.75, 2), get_LATE)
rbind(gamma1_seq, LATE) %>% knitr::kable(digits = 2)
```

gamma1_seq	0.75	1.00	1.25	1.50	1.75	2
LATE	0.21	0.15	0.11	0.07	0.03	0

TOT and TUT in the Textbook Model

$$
\begin{aligned}
\text{TOT} &\equiv \mathbb{E}(\Delta|D=1) \\
&= \mathbb{E}(\Delta|D=1, Z=0)\mathbb{P}(Z=0|D=1) + \mathbb{E}(\Delta|D=1, Z=1)\mathbb{P}(Z=1|D=1) \\
&= \underbrace{\mathbb{E}(\Delta|V<\gamma_0)}_{\text{Always-takers}} \times (1-q_1) + \underbrace{\mathbb{E}(\Delta|V<\gamma_0+\gamma_1)}_{\text{Always-takers \& Compliers}} \times q_1
\end{aligned}
$$

$$
\begin{aligned}
\text{TUT} &\equiv \mathbb{E}(\Delta|D=0) \\
&= \mathbb{E}(\Delta|D=0, Z=0)\mathbb{P}(Z=0|D=0) + \mathbb{E}(\Delta|D=0, Z=1)\mathbb{P}(Z=1|D=0) \\
&= \underbrace{\mathbb{E}(\Delta|V>\gamma_0)}_{\text{Never-takers \& Compliers}} (1-q_1) + \underbrace{\mathbb{E}(\Delta|V>\gamma_0+\gamma_1)}_{\text{Never-takers}} q_1
\end{aligned}
$$

TOT and TUT in the Textbook Model

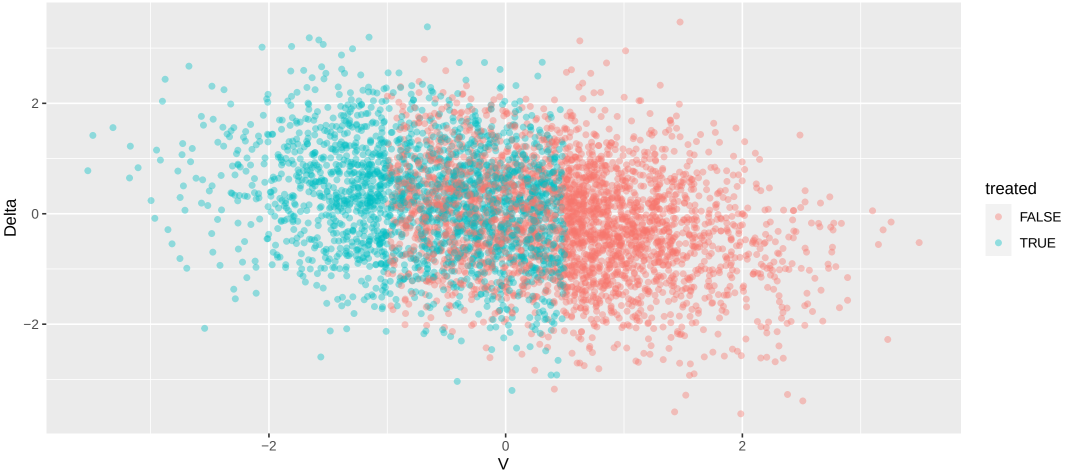
- TOT is a weighted average of $\mathbb{E}(\Delta|V < \gamma_0)$ and $\mathbb{E}(\Delta|V < \gamma_0 + \gamma_1)$.
- TUT is a weighted average of $\mathbb{E}(\Delta|V > \gamma_0)$ and $\mathbb{E}(\Delta|V > \gamma_0 + \gamma_1)$.
- Need to be able to calculate $\mathbb{E}(\Delta|V > c)$ and $\mathbb{E}(\Delta|V < c)$.
- TOT and TUT depend on Z through γ_0 and γ_1: defines "the treated"

```r
# Need Z to define "the treated"
sims <- sims %>%
  select(-complier) %>%
  mutate(Z = rbinom(nrow(sims), 1, 0.5),
         treated = gamma0 + gamma1 * Z > V)
sims
```

```
## # A tibble: 5,000 x 6
##           V      Y0      Y1  Delta     Z treated
##       <dbl>   <dbl>   <dbl>  <dbl> <int> <lgl>
##  1 -0.122   -0.399   1.08    1.48      0 FALSE
##  2 -0.506   -1.10    1.49    2.59      0 FALSE
##  3  0.00457 -0.121  -0.456  -0.335     0 FALSE
##  4 -0.549   -0.248  -0.899  -0.651     0 FALSE
##  5  1.95    -0.0948 -0.675  -0.580     1 FALSE
##  6  0.561    0.112  -0.615  -0.726     0 FALSE
##  7 -0.238   -0.439  -1.53   -1.10      1 TRUE
##  8 -1.46    -1.23   -0.0548  1.17      1 TRUE
##  9 -0.336   -0.891   1.53    2.42      1 TRUE
```

Who's treated if $q = 0.5$, $\gamma_0 = -1$ and $\gamma_1 = 1.5$?

```
ggplot(sims, aes(x = V, y = Delta, col = treated)) +
  geom_point(alpha = 0.4)
```



TOT and TUT Effects: $q = 0.5$, $\gamma_0 = -1$ and $\gamma_1 = 1.5$

```
sims %>%
  group_by(treated) %>%
  summarize(mean(Y1 - Y0)) %>%
  knitr::kable(digits = 3)
```

treated	mean(Y1 - Y0)
FALSE	-0.170
TRUE	0.223

- Different values of q, γ_0, γ_1, would give different TUT and TOT.
- In this example we have **selection on gains**: TUT $<$ ATE $<$ TOT

Analytical Results for the Textbook Model

$$\text{LATE} = -(\sigma_1\rho_1 - \sigma_0\rho_0)\left[\frac{\varphi(\gamma_0+\gamma_1)-\varphi(\gamma_0)}{\Phi(\gamma_0+\gamma_1)-\Phi(\gamma_0)}\right]$$

$$\text{TOT} = -(\sigma_1\rho_1 - \sigma_0\rho_0)\left[\frac{(1-q)\varphi(\gamma_0)+q\varphi(\gamma_0+\gamma_1)}{(1-q)\Phi(\gamma_0)+q\Phi(\gamma_0+\gamma_1)}\right]$$

$$\text{TUT} = (\sigma_1\rho_1 - \sigma_0\rho_0)\left[\frac{(1-q)\varphi(\gamma_0)+q\varphi(\gamma_0+\gamma_1)}{(1-q)\{1-\Phi(\gamma_0)\}+q\{1-\Phi(\gamma_0+\gamma_1)\}}\right]$$

Example: $\sigma_0 = \sigma_1 = 1$ and $q = 1/2$

Formulas Simplify ($\delta \equiv \rho_1 - \rho_0$)

$$\text{LATE} = -\delta \left[\frac{\varphi(\gamma_0 + \gamma_1) - \varphi(\gamma_0)}{\Phi(\gamma_0 + \gamma_1) - \Phi(\gamma_0)} \right]$$

$$\text{TOT} = -\delta \left[\frac{\varphi(\gamma_0) + \varphi(\gamma_0 + \gamma_1)}{\Phi(\gamma_0) + \Phi(\gamma_0 + \gamma_1)} \right]$$

$$\text{TUT} = \delta \left[\frac{\varphi(\gamma_0) + \varphi(\gamma_0 + \gamma_1)}{\{1 - \Phi(\gamma_0)\} + \{1 - \Phi(\gamma_0 + \gamma_1)\}} \right]$$

- **In the practical session you will reproduce some plots from** Angrist (2004).

Why do we care about any of this?

- In the textbook model we can see how the ATE, LATE, TOT and TUT compare.
- The key parameters of the textbook model **are point identified**.
- This allows us to use data to go *beyond LATE* to other causal effects: ATE, TOT and TUT, and more (next time).
- **Next Time:** Marginal Treatment Effects methods are a modern “update” of this textbook model.

Heckman Two-step Estimator

We will show that:

$$\mathbb{E}[Y|D = 1, Z = z] = \mu_1 + \delta_1 \mathbb{E}(V|D = 1, Z = z)$$

$$\mathbb{E}(V|D = 1, Z = z) = \frac{-\varphi(\gamma_0 + \gamma_1 z)}{\Phi(\gamma_0 + \gamma_1 z)}$$

$$\mathbb{E}[Y|D = 0, Z = z] = \mu_0 + \delta_0 \mathbb{E}(V|D = 0, Z = z)$$

$$\mathbb{E}(V|D = 0, Z = z) = \frac{\varphi(\gamma_0 + \gamma_1 z)}{1 - \Phi(\gamma_0 + \gamma_1 z)}$$

Heckman Two-step Estimator

Define the following shorthand:

$$\lambda(z) \equiv \mathbb{E}(V|D=0, Z=z) = \frac{\varphi(\gamma_0 + \gamma_1 z)}{1 - \Phi(\gamma_0 + \gamma_1 z)}$$
$$\kappa(z) \equiv \mathbb{E}(V|D=1, Z=z) = \frac{-\varphi(\gamma_0 + \gamma_1 z)}{\Phi(\gamma_0 + \gamma_1 z)}.$$

Then we have

$$\mathbb{E}[Y|D=0, Z] = \mu_0 + \delta_0 \lambda(Z)$$
$$\mathbb{E}[Y|D=1, Z] = \mu_1 + \delta_1 \kappa(Z)$$

- Use D and Z to estimate γ_0 and γ_1
- To estimate μ_0 and δ_0 regress Y on $\lambda(Z)$ and a constant for obs with $D = 0$
- To estimate μ_1 and δ_1 regress Y on $\kappa(Z)$ and a constant for obs with $D = 1$

Step 1: $(U_0, U_1) \perp\!\!\!\perp Z | V$

Axioms of Conditional Independence

- See https://expl.ai/LXPVDDN or chapter 2 of the lecture notes

$$
\begin{aligned}
\text{(Assumption)} \quad Z \perp\!\!\!\perp (U_0, U_1, V) &\implies Z \perp\!\!\!\perp (U_0, U_1, V) | V && \text{(Weak Union)} \\
&\implies Z \perp\!\!\!\perp (U_0, U_1) | V && \text{(Decomposition)} \\
&\implies (U_0, U_1) \perp\!\!\!\perp Z | V && \text{(Symmetry)}
\end{aligned}
$$

Step 2: $\mathbb{E}(U_0|V)$ and $\mathbb{E}(U_1|V)$.

General Result: $(X, Y) \sim$ Bivariate Normal

$$\mathbb{E}(Y|X = x) = \mathbb{E}(Y) + \frac{\text{Cov}(Y, X)}{\text{Var}(X)}\left[x - \mathbb{E}(X)\right]$$

Our Setting: $V \sim N(0, 1)$

$$\mathbb{E}(Y_1 - Y_0|V) = (\mu_1 - \mu_0) + \mathbb{E}(U_1 - U_0)$$

$$\mathbb{E}(U_1|V) = \sigma_1\rho_1 V \equiv \delta_0 V$$
$$\mathbb{E}(U_0|V) = \sigma_0\rho_0 V \equiv \delta_1 V$$

$$\mathbb{E}(U_1 - U_0|V) = (\sigma_1\rho_1 - \sigma_0\rho_0)V \equiv (\delta_1 - \delta_0)V$$

Step 3: $\mathbb{E}(Y|D, Z, V)$

$$\begin{aligned}
\mathbb{E}(Y|D=0, Z, V) &= \mathbb{E}(Y_0|D=0, Z, V) \\
&= \mu_0 + \mathbb{E}(U_0|D=0, Z, V) && \text{(Defn. of } Y_0\text{)} \\
&= \mu_0 + \mathbb{E}(U_0|Z, V) && (D = f(Z, V)) \\
&= \mu_0 + \mathbb{E}(U_0|V) && \text{(Step 1)} \\
&= \mu_0 + \delta_0 V && \text{(Step 2)}
\end{aligned}$$

$$\mathbb{E}(Y|D=1, Z, V) = \mu_1 + \delta_1 V \qquad \text{(Same Steps)}$$

Step 4: $\mathbb{E}(Y, D, Z)$

$$
\begin{aligned}
\mathbb{E}(Y|D=0,Z) &= \mathbb{E}_{V|(D=0,Z)}\left[\mathbb{E}(Y|D=0,Z,V)\right] && \text{(Iterated } \mathbb{E}\text{)} \\
&= \mathbb{E}(\mu_0 + \delta_0 V|D=0,Z) && \text{(Step 3)} \\
&= \mu_0 + \delta_0 \mathbb{E}(V|D=0,Z) && \text{(Linearity of } \mathbb{E}\text{)}
\end{aligned}
$$

$$
\mathbb{E}(Y|D=1,Z) = \mu_1 + \delta_1 \mathbb{E}(V|D=1,Z) \qquad \text{(Same Steps)}
$$

The Mean of a Truncated Normal Distribution

- We will need these results on the next slide!
- Derivation of the first result: https://expl.ai/VFARCYE.

Suppose that $Z \sim N(0,1)$. Then for any constants a, b, c

$$E(Z|Z > c) = \frac{\varphi(c)}{1 - \Phi(c)}$$

$$E(Z|Z < c) = \frac{-\varphi(c)}{\Phi(c)}$$

$$E(Z|a < Z < b) = \frac{-[\varphi(b) - \varphi(a)]}{\Phi(b) - \Phi(a)}$$

Step 5: $\mathbb{E}(V|D,Z)$

$$\begin{aligned}
\mathbb{E}(V|D=1,Z=1) &= \mathbb{E}(V|\gamma_0+\gamma_1 > V, Z=1) && (D=f(Z,V)) \\
&= \mathbb{E}(V|\gamma_0+\gamma_1 > V) && (V \perp\!\!\!\perp Z) \\
&= \frac{-\varphi(\gamma_0+\gamma_1)}{\Phi(\gamma_0+\gamma_1)} && (\text{Trunc. Normal})
\end{aligned}$$

$$\mathbb{E}(V|D=1,Z=0) = \frac{-\varphi(\gamma_0+\gamma_1)}{\Phi(\gamma_0+\gamma_1)} \qquad (\text{Similar Steps})$$

$$\mathbb{E}(V|D=0,Z=1) = \frac{\varphi(\gamma_0+\gamma_1)}{1-\Phi(\gamma_0+\gamma_1)} \qquad (\text{Similar Steps})$$

$$\mathbb{E}(V|D=0,Z=0) = \frac{\varphi(\gamma_0)}{1-\Phi(\gamma_0)} \qquad (\text{Similar Steps})$$