Empirical Analysis III Mogstad Problem Set 1

Chase Abram, Tom Hierons, Jeanne Sorin

July 1, 2020

Problem 1

Consider the nonparametric Roy model:

$$
\begin{aligned}
Y &= DY_1 + (1-D)Y_0 \\
D &= \mathbf{1}\{U \leq \nu(Z)\} \\
(Y_0, Y_1, U) &\perp\!\!\!\perp Z \\
\mathbb{E}[Y_d^2] &< \infty, \quad d \in \{0,1\}
\end{aligned}
$$

Where (Y_0, Y_1, U) are unobserved.

(a) What assumptions about distribution of U did you see in class? Are they restrictive?

We assumed in class that U was continuously distributed. This places some restrictions on the model in the sense that some distributions or choice processes could violate it. Overall it is a weak assumption.

We also assumed that U had a uniform distribution on $[0,1]$. Given the first assumption this places no additional restrictions on the random variable. For any random variable $\tilde{U}$ that is absolutely continuous with respect to the Lebesgue measure, $U := F_{\tilde{U}}(\tilde{U}) \sim Uniform[0,1]$ so this assumption is just a normalisation.

....

(b) Define MTE and derive ATE, ATT and ATUT as weighted averages of MTE.

We have

$$MTE(u) := \mathbb{E}[Y_1 - Y_0 | U = u]$$

We can then derive the ATE as:

$$
\begin{aligned}
ATE &= \mathbb{E}[Y_1 - Y_0] \\
&= \mathbb{E}\big[\mathbb{E}[Y_1 - Y_0 | U]\big] && \text{(by L.I.E.)} \\
&= \int_0^1 MTE(u) \cdot 1 du && \text{(U uniform)}
\end{aligned}
$$

Now with a bit more work we can arrive at the ATT:

$$
\begin{aligned}
ATT &:= \mathbb{E}[Y_1 - Y_0 | D = 1] \\
&= \mathbb{E}\Big[\mathbb{E}[Y_1 - Y_0 | D = 1, U] \Big| D = 1\Big] && \text{by L.I.E.} \\
&= \mathbb{E}\Big[\mathbb{E}[Y_1 - Y_0 | U, U \leq p(Z)] \Big| D = 1\Big] && \text{by definition of } D \\
&= \mathbb{E}\Big[\underbrace{\mathbb{E}[Y_1 - Y_0 | U]}_{MTE(U)} \Big| D = 1\Big] && \text{since } (Y_0, Y_1, U) \perp\!\!\!\perp Z \\
&= \frac{\mathbb{E}[MTE(U)D]}{\mathbb{P}(D = 1)} && \text{by L.I.E.} \\
&= \frac{\mathbb{E}\Big[\mathbb{E}[MTE(U)1_{\{U \leq p(Z)\}} | U]\Big]}{\mathbb{P}(D = 1)} && \text{by L.I.E.} \\
&= \frac{\mathbb{E}\Big[MTE(U)\mathbb{E}[1_{\{U \leq p(Z)\}} | U]\Big]}{\mathbb{P}(D = 1)} \\
&= \int_0^1 MTE(u) \frac{\mathbb{P}\Big(u \leq p(Z)\Big)}{\mathbb{P}(D = 1)} du && \text{again using } (Y_0, Y_1, U) \perp\!\!\!\perp Z
\end{aligned}
$$

The ATUT is derived analogously:

$$
\begin{aligned}
ATUT &= \mathbb{E}[Y_1 - Y_0 | D = 0] \\
&= \mathbb{E}\Big[\mathbb{E}[Y_1 - Y_0 | U, U \leq p(|) \Big| D = 0\Big] \\
&= \frac{\mathbb{E}[MTE(U)(1 - D)]}{\mathbb{P}(D = 0)} \\
&= \frac{\mathbb{E}\Big[\mathbb{E}[MTE(U)1_{\{U > p(Z)\}} | U]}{\mathbb{P}(D = 0)} \\
&= \int_0^1 MTE(u) \frac{\mathbb{P}\Big(u > p(Z)\Big)}{\mathbb{P}(D = 0)} du
\end{aligned}
$$

. . . .

(c) Interpret the weights you received in the previous part.

The ATE weights the MTE of all individuals equally. For each level of u, the ATT weights by the "share"[1] of individuals who get treated who have that particular level of u. ATUT weights by the share of individuals who do not get treated who have that particular level of u.

[1] These are all densities

Problem 2

(a) ....

Since (U_1, U_0) are jointly normal, each component is normal, so $U_1 - U_0$ is normal. Then we may note

$$
\begin{aligned}
\mathbb{E}[U_1 - U_0] &= 0 \\
\mathrm{var}(U_1 - U_0) &= \mathbb{E}[(U_1 - U_0)^2] \\
&= 1 - 2\rho\sigma + \sigma^2 \\
\Rightarrow U_1 - U_0 &\sim N(0, 1 - 2\rho\sigma + \sigma^2)
\end{aligned}
$$

We may then also note that $(U_0, U_1 - U_0)$ and $(U_1, U_1 - U_0)$ are jointly normal[2], and the off-diagonals of the covariance matrix can be found

$$
\begin{aligned}
\mathrm{cov}(U_0, U_1 - U_0) &= E[U_0 U_1 - U_0^2] - E[U_0]E[U_1 - U_0] \\
&= \rho\sigma - 1 \\
\mathrm{cov}(U_1, U_1 - U_0) &= E[U_1^2 - U_1 U_0] - E[U_1]E[U_1 - U_0] \\
&= \sigma^2 - \rho\sigma
\end{aligned}
$$

Now note that in the special case where the random variables of interest are jointly normal, the linear projection which minimizes quadratic error (the BLP) is equal (not merely an approximation) to the conditional expectation of one given the other. The BLP constant in the one-dimensional case is also simply the covariance normalized by the variance of the regressor. Therefore

$$
\begin{aligned}
\mathbb{E}[U_0 \mid U_1 - U_0] &= \frac{\rho\sigma - 1}{1 - 2\rho\sigma + \sigma^2}(U_1 - U_0) \\
\mathbb{E}[U_1 \mid U_1 - U_0] &= \frac{\sigma^2 - \rho\sigma}{1 - 2\rho\sigma + \sigma^2}(U_1 - U_0)
\end{aligned}
$$

Both the above conditional expecations are normally distribued normal random variables, since $U_1 - U_0$ is normal. We can use the following

$$
\begin{aligned}
Z &\sim N(0, 1) \\
\Rightarrow \mathbb{E}[Z \mid Z > 0] &= \sqrt{\frac{2}{\pi}}
\end{aligned}
$$

This follows from the special property of the standard normal pdf ϕ that for all t, $\phi'(t) = -t\phi(t)$, and an evalation of an integral using this transformation.

[2] Sums of normals are normal, and $(U_i, U_1 - U_0) = (U_i, U_1) - (0, U_0)$.

$$
\begin{aligned}
\int_{z>0} z\phi(z)\mathrm{d}z &= -\int_{z>0} \phi'(z)\mathrm{d}z \\
&= -\big(\lim_{z'\to\infty} \phi(z') - \phi(0)\big) \\
&= \phi(0) \\
&= \sqrt{\frac{2}{\pi}}
\end{aligned}
$$

Therefore we have following[3], by noting that any normal random variable may be written $X = \mu + \sigma Z$

$$
\begin{aligned}
\mathbb{E}[U_1 \mid U_1 - U_0 > 0] &= \int_0^\infty \mathbb{E}[U_1 \mid U_1 - U_0 = x]\phi(x)\mathrm{d}x \\
&= \frac{\sigma^2 - \rho\sigma}{\sqrt{1 - 2\rho\sigma + \sigma^2}}\sqrt{\frac{2}{\pi}} \\
\mathbb{E}[U_0 \mid U_1 - U_0 < 0] &= -\int_0^\infty \mathbb{E}[U_1 \mid U_1 - U_0 = x]\phi(x)\mathrm{d}x \\
&= -\frac{\rho\sigma - 1}{\sqrt{1 - 2\rho\sigma + \sigma^2}}\sqrt{\frac{2}{\pi}}
\end{aligned}
$$

Then we may note

$$
\begin{aligned}
\mathbb{E}[Y \mid D = 1] &= \mathbb{E}[Y_1 \mid U_1 - U_0 > 0] \\
&= \mathbb{E}[U_1 \mid U_1 - U_0 > 0] \\
&= \frac{\sigma^2 - \rho\sigma}{\sqrt{1 - 2\rho\sigma + \sigma^2}}\sqrt{\frac{2}{\pi}} \\
\mathbb{E}[Y \mid D = 0] &= \mathbb{E}[Y_0 \mid U_1 - U_0 < 0] \\
&= \mathbb{E}[U_0 \mid U_1 - U_0 < 0] \\
&= -\frac{\rho\sigma - 1}{\sqrt{1 - 2\rho\sigma + \sigma^2}}\sqrt{\frac{2}{\pi}}
\end{aligned}
$$

Now we can use the above calculations to find a closed-form β_{OLS} expression. Again, by definition, β_{OLS} is the least-squares projection of D onto Y, and therefore equal to the covariance normalized by the variance of D.

[3] This result is an application of the Inverse Mills ratio that states that if $X \sim \mathbb{N}(\mu_x, \sigma_x^2)$, then

$$
\mathbb{E}(X|X > T) = \mu_x + \sigma_x \frac{\phi(\frac{T}{\sigma_x})}{1 - \Phi(\frac{T}{\sigma_x})}
$$

$$
\mathbb{E}(X|X < T) = \mu_x - \sigma_x \frac{\phi(\frac{T}{\sigma_x})}{\Phi(\frac{T}{\sigma_x})}
$$

$$
\begin{aligned}
\beta_{OLS} &= \frac{\text{cov}(D,Y)}{\text{var}(D)} \\
&= \frac{\mathbb{E}[DY] - \mathbb{E}[D]\mathbb{E}[Y]}{\mathbb{E}[D^2] - \mathbb{E}[D]^2} \\
&= \frac{\mathbb{E}[Y \mid D=1]\mathbb{P}[D=1] - \mathbb{E}[D]\left[\mathbb{E}[Y \mid D=1]\mathbb{P}[D=1] + \mathbb{E}[Y \mid D=0]\mathbb{P}[D=0]\right]}{\mathbb{P}[D=1] - \mathbb{P}[D=1]^2} \\
& && \text{(Cancelling } \mathbb{P}[\cdot] \text{ terms)} \\
&= \mathbb{E}[Y \mid D=1] - \mathbb{E}[Y \mid D=0] \\
&= \frac{\sigma^2 - 1}{\sqrt{1 - 2\rho\sigma + \sigma^2}}\sqrt{\frac{2}{\pi}}
\end{aligned}
$$

If we assume that treatment D is independent of the outcomes with and without treatment, then β_{OLS} tells us, on average, how outcome Y varies with treatment D, with no selection effects. This is exactly the ATE. We can see this directly, by revisiting the penultimate line of the above calculation

$$
\begin{aligned}
\mathbb{E}[Y \mid D=1] - \mathbb{E}[Y \mid D=0] &= \mathbb{E}[Y_1 \mid D=1] - \mathbb{E}[Y_0 \mid D=0] \\
&= \mathbb{E}[Y_1] - \mathbb{E}[Y_0] && (D \perp\!\!\!\perp (Y_0, Y_1)) \\
&= \mathbb{E}[Y_1 - Y_0]
\end{aligned}
$$

(b) ....

$$
\begin{aligned}
ATT &= \mathbb{E}[Y_1 - Y_0 \mid D=1] \\
&= \mathbb{E}[U_1 - U_0 \mid U_1 > U_0] \\
&= \sqrt{1 - 2\rho\sigma + \sigma^2}\,\mathbb{E}[Z \mid Z > 0] \\
&= \sqrt{1 - 2\rho\sigma + \sigma^2}\sqrt{\frac{2}{\pi}} > 0 && (\rho \in [-1,1])
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
ATUT &= \mathbb{E}[Y_1 - Y_0 \mid D=0] \\
&= \mathbb{E}[U_1 - U_0 \mid U_1 \leq U_0] \\
&= \sqrt{1 - 2\rho\sigma + \sigma^2}\,\mathbb{E}[Z \mid Z \leq 0] \\
&= -\sqrt{1 - 2\rho\sigma + \sigma^2}\sqrt{\frac{2}{\pi}} < 0 && (\rho \in [-1,1])
\end{aligned}
$$

The signs reflect the fact that individuals are able to choose their (ex-ante = ex-post) outcomes, so in the worst case scenario for *any* individual, the treatment effect, if treated, is zero. The reverse is true for the untreated. Therefore the treatment effect for the treated is always positive, and for the untreated it is always negative.

The magnitudes scale with the standard deviation of the random variable $U_1 - U_0$, and are exactly equal, due to the symmetry of the problem (i.e. $U_1 - U_0 \sim U_0 - U_1$). Roughly speaking this occurs because, as we increase the variance of $U_1 - U_0$ (say by increasing σ), density is symmetrically moved away from zero. This does not effect the mean (see ATE below), but does effect ATT and ATUT, precisely because they condition on selection or lack thereof. So when mass is moved away from zero, individuals get higher positive draws and lower negative draws, but they are able to select accordingly, so they actually see a larger treatment effect for these subgroups.

(c) ....

$$\begin{aligned} ATE &= \mathbb{E}[Y_1 - Y_0] \\ &= \mathbb{E}[U_1 - U_0] \\ &= \mathbb{E}[U_1] - \mathbb{E}[U_0] \\ &= 0 \end{aligned}$$

The symmetry of this setup means that, while selection effects can lead to treatment effects for subgroups, in population average treatment has no effect. This is a nice illustration of why it is important to be careful regarding the interpretation of ATE. Here we have that both the ATT and ATUT are non-zero (highly nonzero if we pick σ accordingly), but since half the population chooses treatment, and the treatment effects for the treated and untreated are equal and opposite, they "cancel out".

(d) ....

$$\begin{aligned} \frac{\partial ATT}{\partial \rho} &= -\sigma(1 - 2\rho\sigma + \sigma^2)^{-\frac{1}{2}}\sqrt{\frac{2}{\pi}} < 0 \\ \frac{\partial ATUT}{\partial \rho} &= \sigma(1 - 2\rho\sigma + \sigma^2)^{-\frac{1}{2}}\sqrt{\frac{2}{\pi}} > 0 \\ \frac{\partial \beta_{OLS}}{\partial \rho} &= \sigma(\sigma^2 - 1)(1 - 2\rho\sigma + \sigma^2)^{-\frac{3}{2}}\sqrt{\frac{2}{\pi}} \end{aligned}$$

I will try to squeeze some intuition. As ρ increases, individuals see U_1 and U_0 become "closer", in that the variance of $U_1 - U_0$ decreases, as we would expect when we increase their correlation. So while ATT will always be non-negative, as we increase ρ, it decreases, and if $\sigma = 1$, at the limit $\rho = 1$, we find ATT is zero. This same idea holds for ATUT, but it is always non-positive, and increasing. The β_{OLS} result is less intuitive. Perhaps one way to understand it is that β_{OLS} depends on the interplay between the variances of U_1 and U_0. If the $\sigma < 1$, then increasing the correlation between Us actually decreases the explanatory effect of D on Y, because, in some sense, the "noisiness" of U_0 is greater than the "noisiness" of U_1[4], so the higher correlation means there is more "noise" between D and Y than explanatory power. When $\sigma > 1$, the reverse is true, so that in either case β_{OLS} is driven towards zero as ρ increases.

[4] I am still speaking in terms of population, just using intuitive ideas which might sounds sample-esque.

$$\begin{aligned}
\frac{\partial ATT}{\partial \sigma} &= \frac{\sigma - \rho}{\sqrt{1 - 2\rho\sigma + \sigma^2}}\sqrt{\frac{2}{\pi}} \\
\frac{\partial ATUT}{\partial \sigma} &= -\frac{\sigma - \rho}{\sqrt{1 - 2\rho\sigma + \sigma^2}}\sqrt{\frac{2}{\pi}} \\
\frac{\partial \beta_{OLS}}{\partial \sigma} &= \left(2\sigma(1 - 2\rho\sigma + \sigma^2)^{-\frac{1}{2}} + (\sigma^2 - 1)(1 - 2\rho\sigma + \sigma^2)^{-\frac{3}{2}}(2\sigma - 2\rho)\right)\sqrt{\frac{2}{\pi}}
\end{aligned}$$

These are even trickier effects. I would say there is no "simple" intuition, but for ATT and ATUT, we can at least say that the way variance affects the treatment effect for these sub-populations depends on how the variance interplays with the correlation. If the Us are highly correlated, but σ is low, then increasing σ will drive down ATT, because then more density will be placed on events where U_1 is only marginally greater than U_0. The reverse is true for ATUT, but again these effects are not so simple. For β_{OLS}, there is no simple explanation I can see, since the sign of the derivative will depend nontrivially on the relationship between σ and ρ, so perhaps the lesson here is that even in super basic toy example, the effects on the BLP of changing the relative variances are not obvious[5].

(e) ....

We already deduced, from our calculations in a), that $\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0] = \beta_{OLS}$.

$\sigma = 2$	$\rho = -0.5$	$\rho = 0$	$\rho = 0.5$
ATE	-0.01	0.00	-0.02
ATT	2.10	1.76	1.39
ATUT	-2.10	-1.78	-1.38
β_{OLS}	0.89	1.04	1.40

$\sigma = 0.5$	$\rho = -0.5$	$\rho = 0$	$\rho = 0.5$
β_{OLS}	-0.43	-0.53	-0.70

$\rho = 0.5$	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 0.75$
ATE	0.00	-0.01	0.01
ATT	0.73	0.69	.72
ATUT	-0.73	-0.69	-0.72
β_{OLS}	-0.84	-0.70	-0.39

Little remains to be said that was not said in d). The above table confirms the signs of all the partial derivatives, and I even included the middle table to show how β_{OLS} changes differently with ρ, depending on σ.

(f) ....

The claim is incorrect, and appears to be conflating the fact that, since we are dealing with a bivariate normal distribution, we do have the special fact $U_0 \perp\!\!\!\perp U_1 \iff \rho = 0$.

Mathematically, we can show directly from the definition of independence that this claim cannot hold.

[5] One explanation is that the changing kurtosis of the $U_1 - U_0$ random variable comes into play. If so, this falls out of the category of intuitive explantion, in my opinion.

$$\mathbb{P}[Y_1 \leq Y_0, D = 1] = 0 \neq \frac{1}{4} = \mathbb{P}[Y_1 \leq Y_0] \cdot \mathbb{P}[D = 1]$$

Note the above calculation made no reference to ρ. In order for the selection to be independent of the outcomes, we would need that the selection function (of random variables) is independent of (Y_1, Y_0), and our above line shows precisely that this does not hold for the indicator function of $U_1 > U_0$. Furthermore, if we consider the column from e) where $\rho = 0$, if selection were independent of outcomes, we would have ATE = ATT = ATUT. But we do not.

Problem 3

Monte Carlo Simulations

(a)

See code file.

We find that $\hat{\beta} = (1.9738669, 3.0013334)$. The standard errors of $\hat{\beta}$ are approximately $(0.02005478, 0.00202156)$.

(b)

See code file.

We find that the standard errors of $\hat{\beta}$ are approximately $(0.0200475, 0.0020022)$.

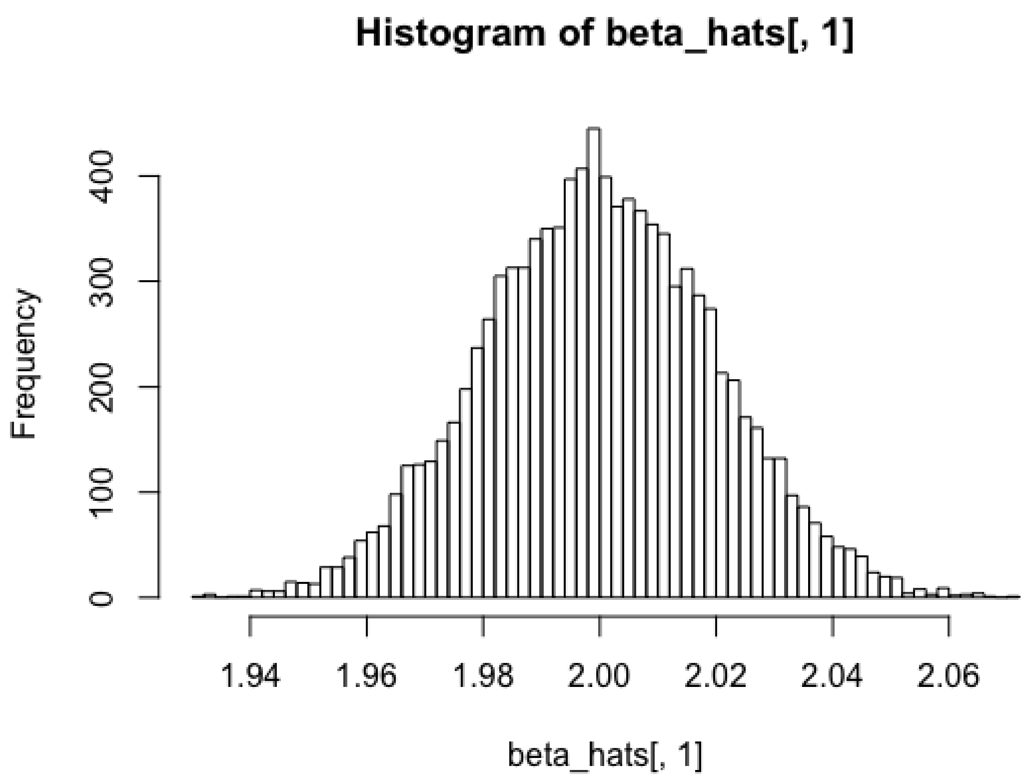

Figure 1: Histogram of the first component of $\hat{\beta}$ for $S = 10,000$

Why does our estimator for the standard error of $\hat{\beta}_k$ converges to $se(\hat{\beta}_k|X_1, X_2, ..., X_N)$?
Because X_i are i.i.d. draws, by the WLLN, CMT and marginal convergence in probability implying

joint convergence in probability, we can show that $\hat{\beta}_k^{(s)}$ converges to $\beta_k^{(s)}$ in probability[6]. The CMT implies that $(\hat{\beta}_k^{(s)})^2$ also converges to $(\beta_k^{(s)})^2$ in probability. There, the two terms under the square root converge to $E[\beta_k^2]$ and $E[\beta_k]$ respectively. Therefore, applying the CMT once again, we argue that their sum converges to the diagonal elements of the variance matrix normalized by N, since each s sample has N draws.

Nonparametric Bootstrap

(a)

See code file.

We find that $\beta_{OLS} = (1.98675, 3.006932)$.
The estimated standard errors of β_{OLS} are $(0.0141752, 0.0200649)$.

OLS gives consistent estimates because

$$\begin{aligned}
\hat{\beta} &= (\sum_{i=1}^{N} X_i X_i')^{-1}(\sum X_i y_i) \\
&= (\sum_{i=1}^{N} D_i D_i')^{-1}(\sum D_i y_i) \\
&= (\sum_{i=1}^{N} D_i^2)^{-1}(\sum D_i y_i) \\
&= (\sum_{i=1}^{N} D_i)^{-1}(\sum D_i y_i) && \text{because D } \in \{0,1\} \\
&= (\frac{1}{N}\sum_{i=1}^{N} D_i)^{-1}(\frac{1}{N}\sum D_i y_i)
\end{aligned}$$

6

$$\begin{aligned}
\text{With } \hat{\beta}_k^{(s)} &= (\sum_{i=1}^{N} X_i X_i')^{-1}(\sum_{i=1}^{N} X_i y_i^s) \\
\text{with } &\sum_{i=1}^{N} X_i X_i' \underbrace{\Rightarrow}_{p} \mathbb{E}(X'X) && \text{WLLN, iid } X_i \\
\text{with } &\sum_{i=1}^{N} X_i y_i^s \underbrace{\Rightarrow}_{p} \mathbb{E}(X'y^s) && \text{WLLN, iid } X_i, X_i \perp u_i
\end{aligned}$$

As in the previous part we have

$$\frac{1}{N}\sum_{i=1}^{N} D_i \underbrace{\Rightarrow}_{p} E(D) \qquad \text{WLLN because D is iid}$$
$$\Rightarrow (\frac{1}{N}\sum_{i=1}^{N} D_i)^{-1} \underbrace{\Rightarrow}_{p} E(D)^{-1} \qquad \text{CMT}$$

As well as

$$\frac{1}{N}\sum_{i=1}^{N} D_i y_i \underbrace{\Rightarrow}_{p} E(Dy) \qquad \text{WLLN and CMT, because } D \perp y_1, y_0, y_1, y_0, \text{D are iid}$$

Therefore

$$\hat{\beta} = (\frac{1}{N}\sum_{i=1}^{N} D_i)^{-1}(\frac{1}{N}\sum D_i y_i)$$
$$\underbrace{\Rightarrow}_{p} \beta = \mathbb{E}(D)^{-1}\mathbb{E}(Dy) \qquad \text{by CMT and Slutsky}$$

(b)

See code file.

The standard errors of $\hat{\beta}$ are $(0.01409, 0.01989)$.
If we had drawn Y and D independently, i.e. we modify the function sampling from the original dataset such as:

```
pick_replacement <- function(N, X, Y){

  indices1 <- ceiling(runif(N,0,length(Y)))
  indices2 <- ceiling(runif(N,0,length(Y)))

  Xs <- X[indices1,]
  Ys <- Y[indices2]
  return(matrix(c(Xs,Ys), ncol=ncol(X)+1))
}
```

As we can see on the histogram below, this leads the coefficient on D to converge to 0, as, by construction, D has no predictive power on Y anymore.

$$D \perp\!\!\!\perp Y \Rightarrow \frac{\text{cov}(D,Y)}{V(D)} = 0 = \hat{\beta}_2^{OLS}$$
$$\hat{\beta}_1^{OLS} = E(Y) = 3.5$$

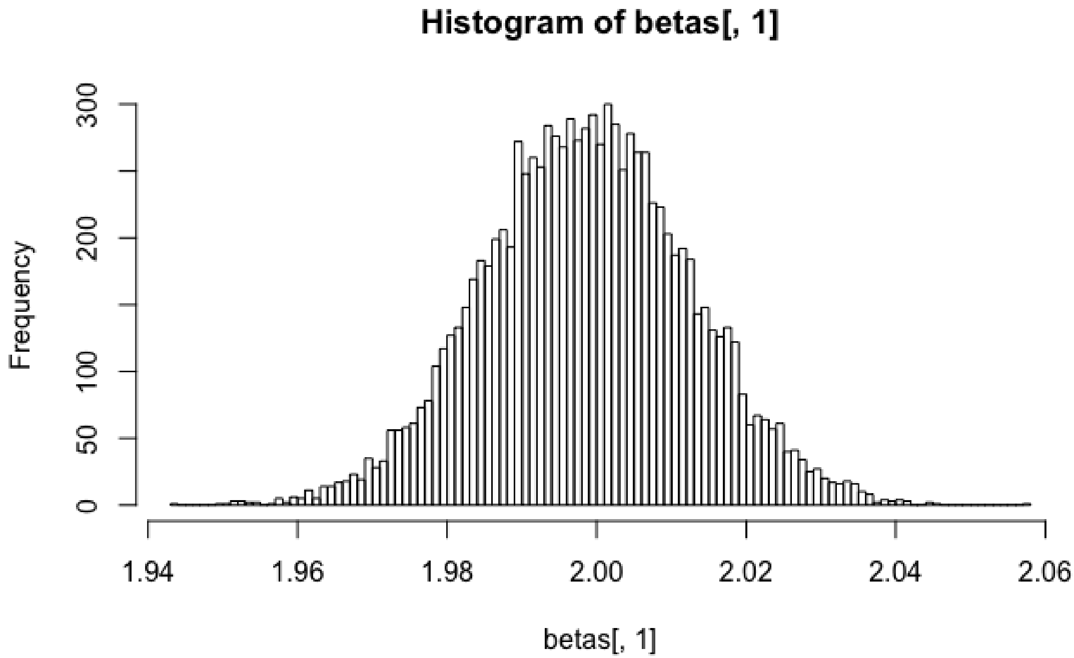

Figure 2: Histogram of the first component of $\hat{\beta}$ using Bootstrap for S = 10,000

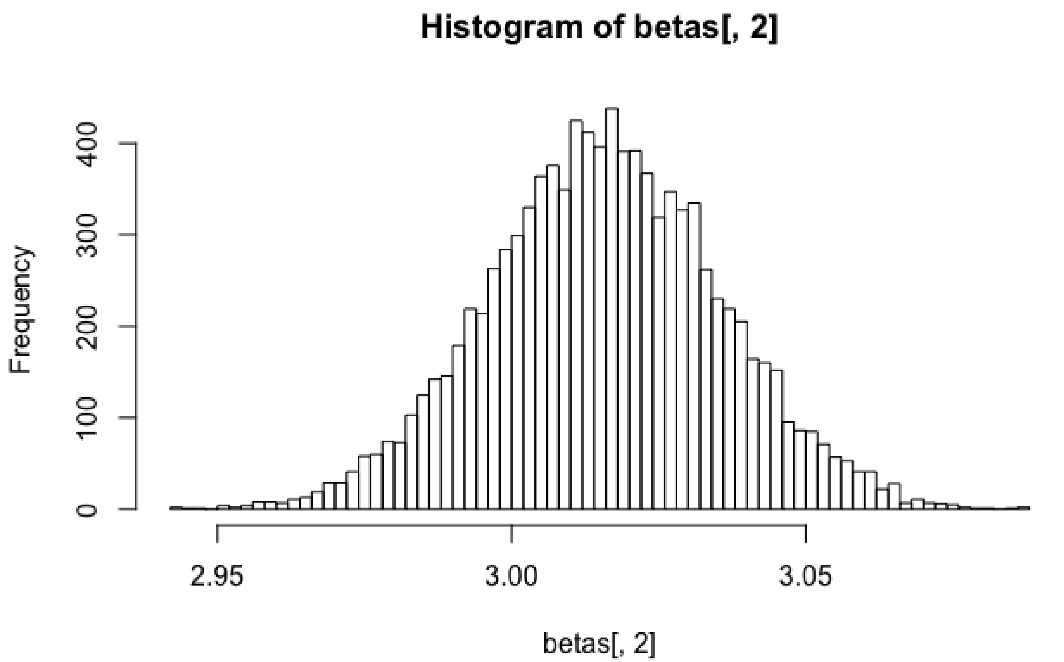

Figure 3: Histogram of the second component of $\hat{\beta}$ using Bootstrap for S = 10,000

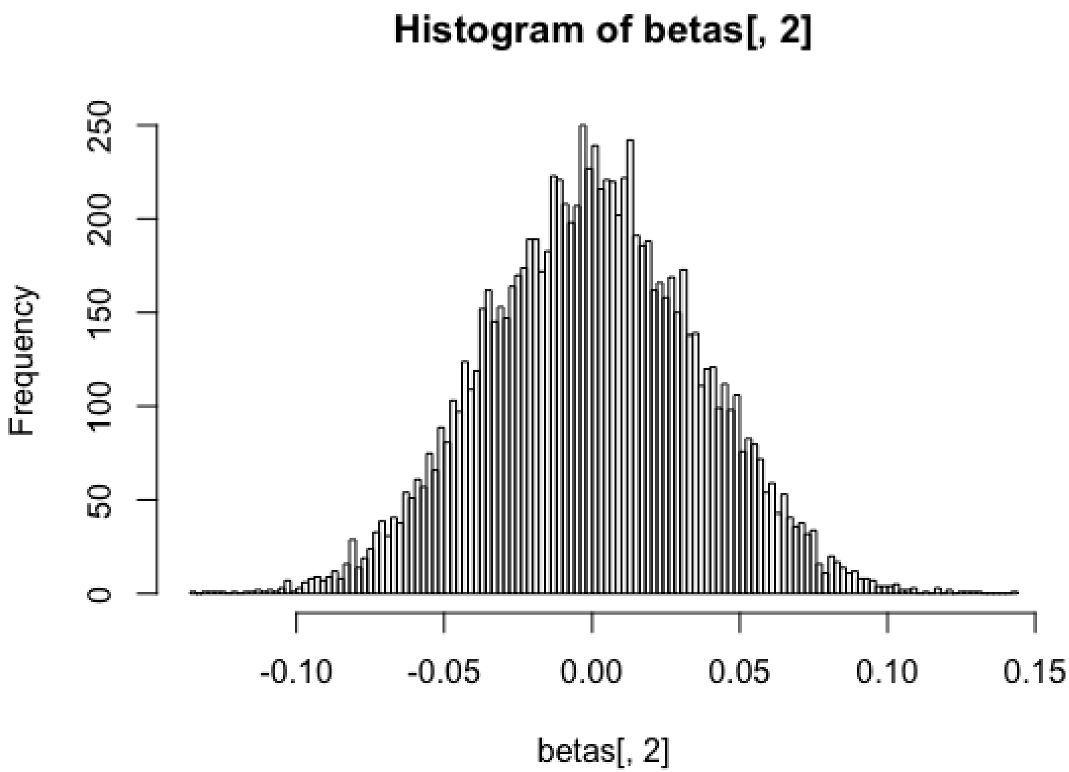

Figure 4: Histogram of $\hat{\beta}_2$ for independently drawn (D, Y)