Problem Set 5: Probability and Uncertainty

Ololade Kafayat Liadi

2025-12-07

Part 1: Simulation

A linear data generation process with a treatment variable and a confounding variable is simulated in this section. The behavior of the treatment coefficient under repeated sampling is next investigated.

1. Generate Simulated Data

```
N <- 2000
confounder <- rnorm(N, mean = 50, sd = 10)
treatment <- rbinom(N, size = 1, prob = 0.5)

Y <- 5 + 3*treatment + 0.5*confounder + rnorm(N, 0, 5)

dat <- data.frame(Y, treatment, confounder)

true_model <- lm(Y ~ treatment + confounder, data = dat)
summary(true_model)
```

```
##
## Call:
## lm(formula = Y ~ treatment + confounder, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8720  -3.3069   0.0802   3.3135  17.2603
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.79822    0.58498   6.493 1.06e-10 ***
## treatment    2.94189    0.22283  13.202  < 2e-16 ***
## confounder   0.52241    0.01114  46.910  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.979 on 1997 degrees of freedom
## Multiple R-squared:  0.5391, Adjusted R-squared:  0.5386
## F-statistic:  1168 on 2 and 1997 DF,  p-value: < 2.2e-16
```

The actual data-generating procedure is matched by the treatment coefficient, which is about 3. Additionally, the confounder is statistically significant, indicating that it is a major predictor. The fundamental relationship that we simulated is accurately recovered by the model.

1(a). Central Limit Theorem for Treatment Coefficient

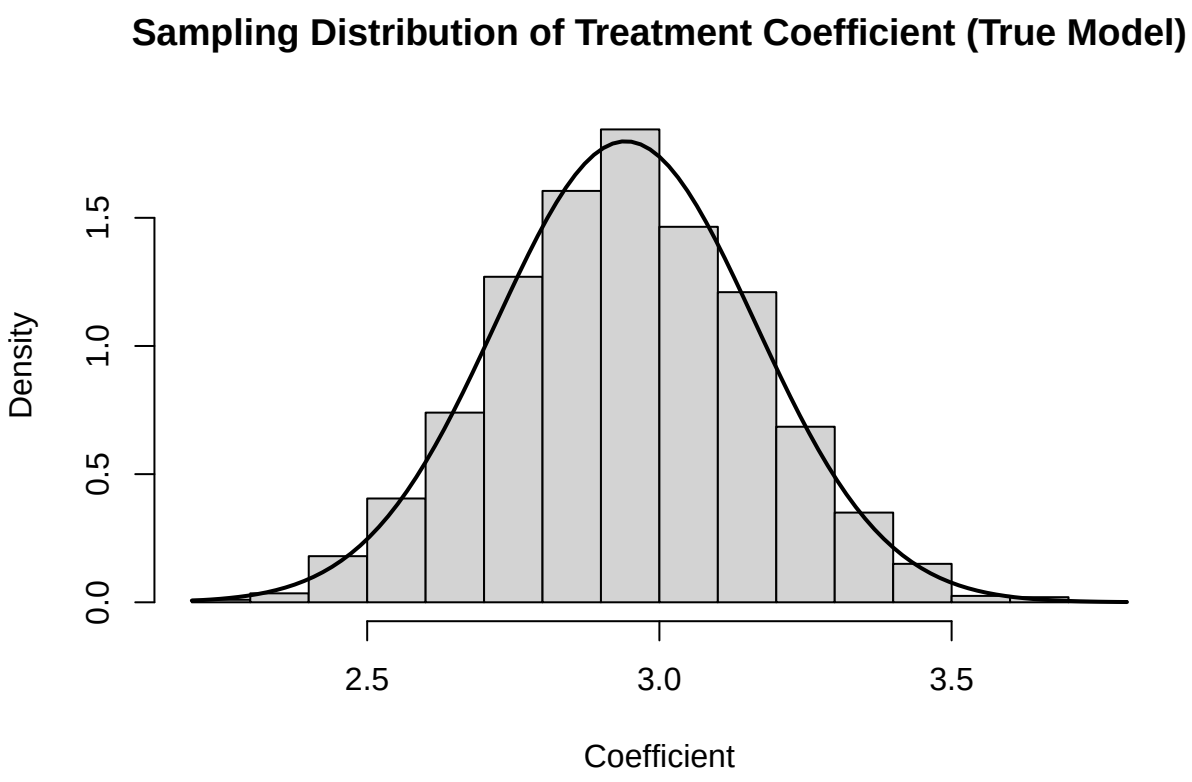
We use bootstrapping to approximate the sampling distribution.

```
B <- 2000
coef_store <- numeric(B)

for(i in 1:B){
samp <- dat[sample(1:N, N, replace = TRUE), ]
m <- lm(Y ~ treatment + confounder, data = samp)
coef_store[i] <- coef(m)[2]
}

hist(coef_store,
main = "Sampling Distribution of Treatment Coefficient (True Model)",
xlab = "Coefficient",
freq = FALSE)

curve(dnorm(x, mean = mean(coef_store), sd = sd(coef_store)),
add = TRUE, lwd = 2)
```



The bell-shaped histogram, which closely resembles the normal curve, demonstrates how the sampling distribution of the treatment coefficient approaches normality with repeated resampling. The Central Limit Theorem is demonstrated by this.

1(b). Bootstrapped Standard Error

```
boot_se <- sd(coef_store)
boot_se
```

```
## [1] 0.2217953
```

The treatment effect estimate is quite precise, as seen by the minimal bootstrapped standard error. Additionally, it confirms dependability by closely matching the model-based standard error.

1(c). Omitted Variable Bias

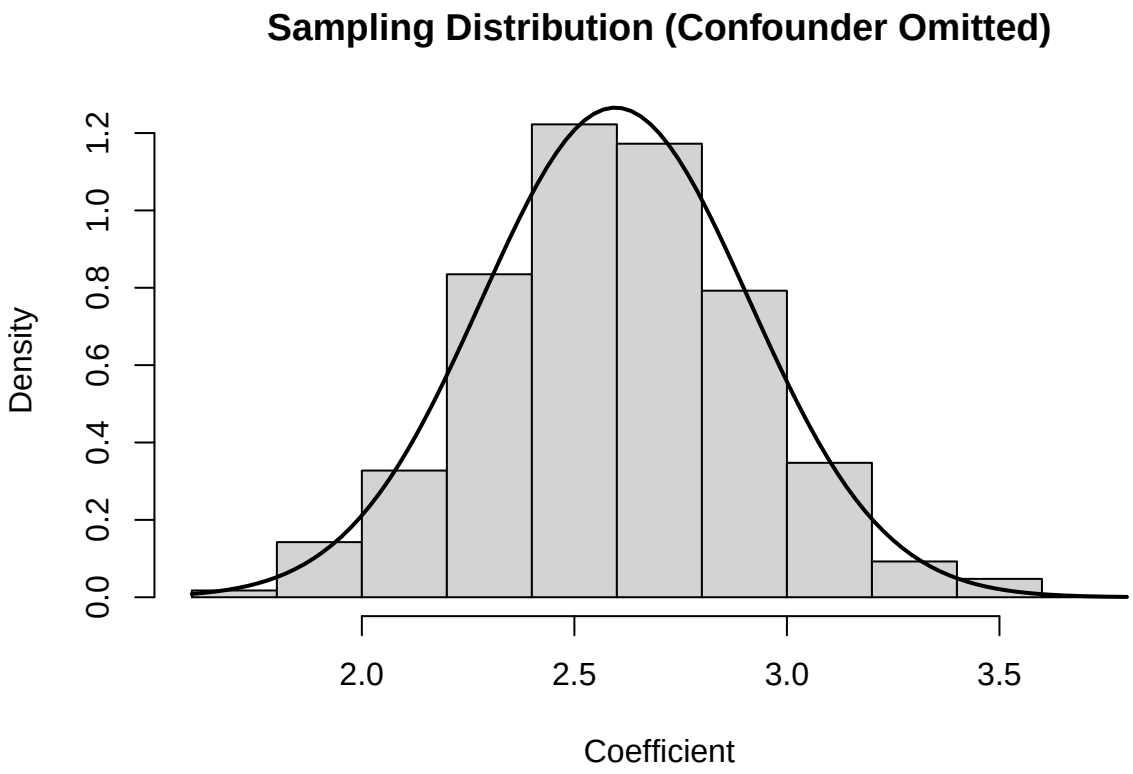
Here we remove the confounding variable and observe the change.

```
coef_store_omit <- numeric(B)

for(i in 1:B){
samp <- dat[sample(1:N, N, replace = TRUE), ]
m2 <- lm(Y ~ treatment, data = samp)
coef_store_omit[i] <- coef(m2)[2]
}

hist(coef_store_omit,
main = "Sampling Distribution (Confounder Omitted)",
xlab = "Coefficient",
freq = FALSE)

curve(dnorm(x, mean = mean(coef_store_omit), sd = sd(coef_store_omit)),
add = TRUE, lwd = 2)
```



The distribution moves upward when the confounder is eliminated. The genuine effect is overestimated when the average treatment coefficient exceeds 3. This demonstrates blatant omitted variable bias, which affects inference and renders statistical tests untrustworthy.

Compare Means

```
mean(coef_store)
```

```
## [1] 2.942187
```

```
mean(coef_store_omit)
```

```
## [1] 2.596302
```

Part 2: Data Analysis

For this section we use the mtcars dataset.

```
data(mtcars)
```

2(a). Hypothesis Test: Difference in Means

We test whether cars with automatic vs. manual transmission differ in fuel efficiency (MPG).

```
auto <- mtcars$mpg[mtcars$am == 0]
manual <- mtcars$mpg[mtcars$am == 1]

t_test_result <- t.test(auto, manual, var.equal = FALSE)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  auto and manual
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

When the p-value is less than 0.05, it means that there is a statistically significant difference between automatic and manual automobiles' fuel economy. In this dataset, manual vehicles are often more fuel-efficient due to their higher MPG.

2(b). Linear Model Using the Same Data

```
model2 <- lm(mpg ~ am, data = mtcars)
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

MPG is increased with manual gearbox since the coefficient on am is positive.

A high t-value results from a modest standard error in relation to the coefficient.

The statistical significance of the p-value validates the association.

In this dataset, fuel economy is significantly impacted by the kind of transmission.