Problem Set 5

Ellie Choe

2025-12-11

Part 1: Simulation

Create a simulated data set with a dependent variable that is a linear function of a treatment variable and a confounding variable. Fit a linear model for the true data generating process and print the summary table.

```
set.seed(1)
n <- 500

# confounder
Z <- rnorm(n, mean=0, sd=1)

# treatment
X <- rnorm(n, mean=0, sd=1) + 0.5*Z

# true model parameters
beta0 <- 4
beta1 <- 2
beta2 <- 3
sigma <- 1

# error term
epsilon <- rnorm(n, mean=0, sd=sigma)

# dependent variable
Y <- beta0 + beta1*X + beta2*Z + epsilon

# Combine into a data frame
df <- data.frame(Y, X, Z)
head(df)
```

```
##           Y          X          Z
## 1  2.783756 -0.2359238 -0.6264538
## 2  5.252768 -0.2050470  0.1836433
## 3 -2.579777 -1.6010565 -0.8356286
## 4 10.614440  0.8089331  1.5952808
## 5  7.370629  1.1563549  0.3295078
## 6  2.243413  1.1837333 -0.8204684
```

```r
# fit the true linear model
true_model <- lm(Y ~ X + Z, data=df)
summary(true_model)
```

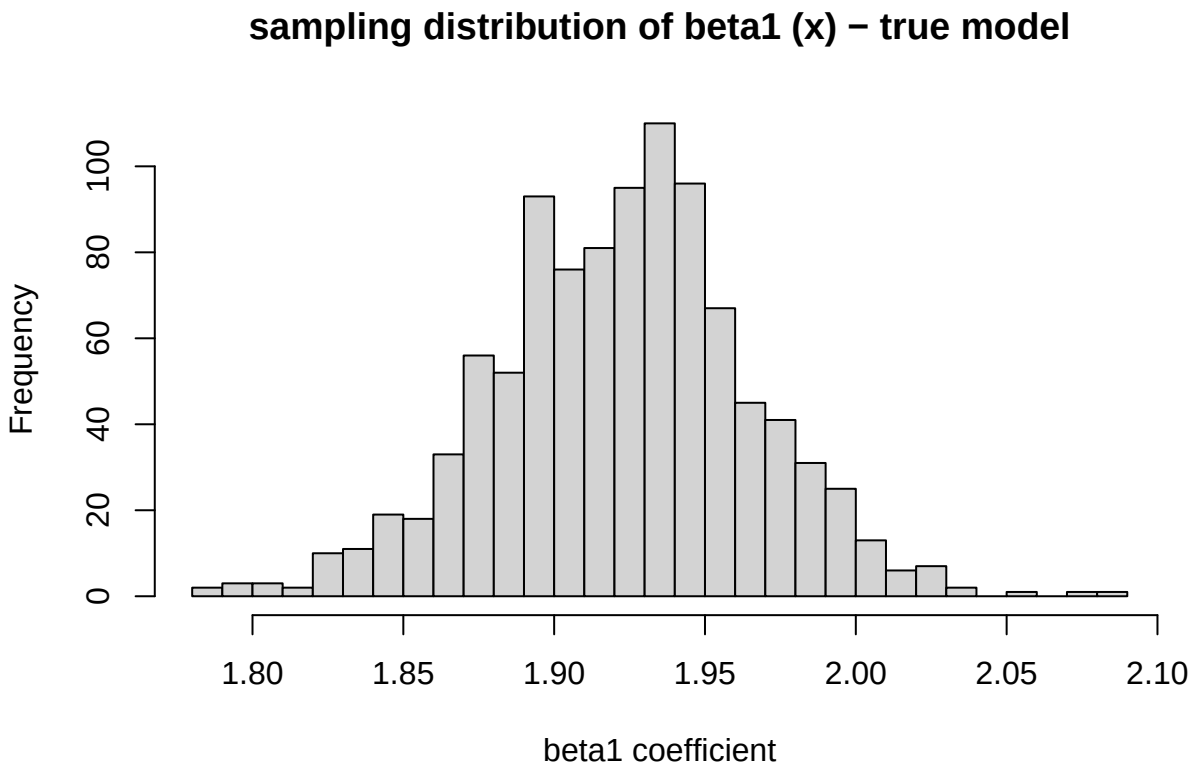
```
##
## Call:
## lm(formula = Y ~ X + Z, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3595 -0.7000 -0.0682  0.7195  3.6850
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.99429    0.04514   88.49   <2e-16 ***
## X            1.92521    0.04272   45.06   <2e-16 ***
## Z            3.00578    0.04868   61.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.008 on 497 degrees of freedom
## Multiple R-squared:  0.9509, Adjusted R-squared:  0.9507
## F-statistic:  4810 on 2 and 497 DF,  p-value: < 2.2e-16
```

Using the true model, demonstrate that the coefficient for your treatment variable follows the central limit theorem. That is, demonstrate that the coefficient's sampling distribution is approximately normal.

```r
# CLT demonstration
b <- 1000
beta1_samples <- numeric(b)

# boostrap samling loop
for(i in 1:b) {
  sample <- df[sample(1:n, n, replace=TRUE), ]
  model <- lm(Y~X+Z, data = sample)
  beta1_samples[i] <- coef(model)["X"]
}

# sampling distribution histogram
hist(beta1_samples, breaks=30, main="sampling distribution of beta1 (x) - true model", xlab="beta1 coef
```



The bootstrap sampling distribution of the treatment coefficient β_1 is approximately normal, which is consistent with the Central Limit Theorem.

Compute the bootstrapped standard error for the coefficient of the treatment variable.

```
# Calculate the boostrap standard error
boot_se <- sd(beta1_samples)
boot_se
```

```
## [1] 0.04270107
```

Fit a model that omits the confounding variable. Repeat part (a) for this new model and plot the sampling distribution of the treatment variable's coefficient. How do your results differ? What does this imply about statistical tests based on a coefficient's sampling distribution?

```
# Omit confounder Z
omitted_model <- lm(Y ~ X, data=df)
summary(omitted_model)
```

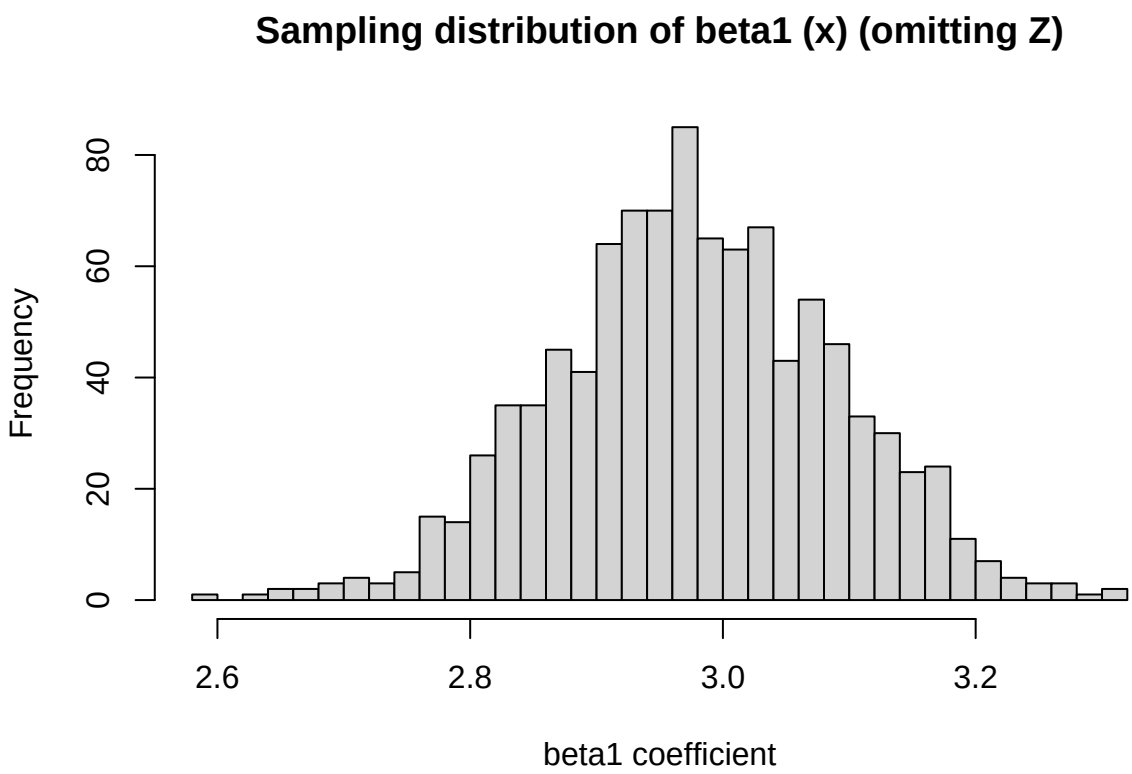
```
##
## Call:
## lm(formula = Y ~ X, data = df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7354 -1.9576 -0.0393  1.9481 10.6810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.0990     0.1327   30.89   <2e-16 ***
## X             2.9828     0.1151   25.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.965 on 498 degrees of freedom
## Multiple R-squared:  0.5741, Adjusted R-squared:  0.5733
## F-statistic: 671.3 on 1 and 498 DF,  p-value: < 2.2e-16
```

```r
# prepare a vector
beta1_omit_samples <- numeric(b)

# Bootstrap sampling loop for the model omitting Z
for(i in 1:b) {
  sample <- df[sample(1:n, n, replace=TRUE), ]
  model <- lm(Y~X, data = sample)
  beta1_omit_samples[i] <- coef(model)["X"]
}

# Plot the sampling distribution
hist(beta1_omit_samples, breaks=30, main="Sampling distribution of beta1 (x) (omitting Z)", xlab="beta1
```



The results diverge due to omitted variable bias. The estimator in the correctly specified model is unbiased, with its sampling distribution centered around the true parameter of 2. In contrast, omitting the confounder Z

introduces an upward bias, shifting the distribution's center to approximately 3. This implies that statistical significant does not imply causal validity.

Part 2: Data Analysis

Conduct a hypothesis test for a difference in means. You decide what the hypotheses are, whether you use a t-test or a z-test, and what the level of significance is. Explain your decisions, and interpret your results both substantively and statistically.

```
voting <- read.csv("voting.csv")

# Perform a two-sample t-test
t_test_result <- t.test(voted ~ message, data=voting, var.equal=TRUE)
t_test_result
```

```
##
##  Two Sample t-test
##
## data:  voted by message
## t = -31.434, df = 229442, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  -0.08637983 -0.07624000
## sample estimates:
##  mean in group no mean in group yes
##         0.2966383         0.3779482
```

The p-value should be smaller than 0.05, and in this case, it was 2.2e-16. Therefore, we can reject the null hypothesis. This means that the difference in voting rates between the message groups is statistically significant. Sending a message increases the probability of voting by approximately 8 percentage points, which represents a meaningful effect in practice. ## Using the same data, fit a linear model. Interpret the coe>icient, standard error, t-value, and p-value.

```
# Fit a linear model
lm_model <- lm(voted ~ message, data = voting)
summary(lm_model)
```

```
##
## Call:
## lm(formula = voted ~ message, data = voting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3780 -0.2966 -0.2966  0.6220  0.7034
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.296638   0.001055  281.05   <2e-16 ***
```

```
## messageyes  0.081310   0.002587   31.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4616 on 229442 degrees of freedom
## Multiple R-squared:  0.004288,   Adjusted R-squared:  0.004284
## F-statistic: 988.1 on 1 and 229442 DF,  p-value: < 2.2e-16
```

The results show that the intercept is 0.2966, representing the mean voting rate for the "no message" group. The coefficient for messageyes is 0.0813, indicating that sending a message increases the probability of voting by approximately 8 percentage points. The standard error of the coefficient is 0.002587, measuring the uncertainty around this estimate. The t-value is 31.43, reflecting how many standard errors the coefficient is away from zero, and the p-value is less than 2e-16. This very small p-value shows that the effect of sending a message is statistically significant, meaning that receiving a message has a positive impact on voting behavior.