POLS 602-PS5-Joseph Kim

Joseph Kim

2025-12-13

==========

PS5 Part 1(a): CLT demo for the treatment coefficient (TRUE MODEL)

==========

```
set.seed(123)

# DGP parameters (match your dataset setup)
n     <- 500      # sample size per simulated dataset
B     <- 2000     # number of simulated samples (increase if you want smoother)
beta0 <- 1
beta1 <- 2        # true effect of T on Y
beta2 <- 3
sd_C  <- 1
sd_T  <- 1
sd_e  <- 2
rho   <- 0.7      # strength of confounding: T = rho*C + noise

# Storage for the treatment coefficient estimates
b1_hat <- numeric(B)

for (b in 1:B) {
  # simulate one dataset
  C   <- rnorm(n, 0, sd_C)
  T   <- rho * C + rnorm(n, 0, sd_T)
  eps <- rnorm(n, 0, sd_e)
  Y   <- beta0 + beta1 * T + beta2 * C + eps

  dat_b <- data.frame(Y = Y, T = T, C = C)

  # fit the TRUE model and store the treatment coefficient
  mod_b <- lm(Y ~ T + C, data = dat_b)
  b1_hat[b] <- coef(mod_b)["T"]
}
```

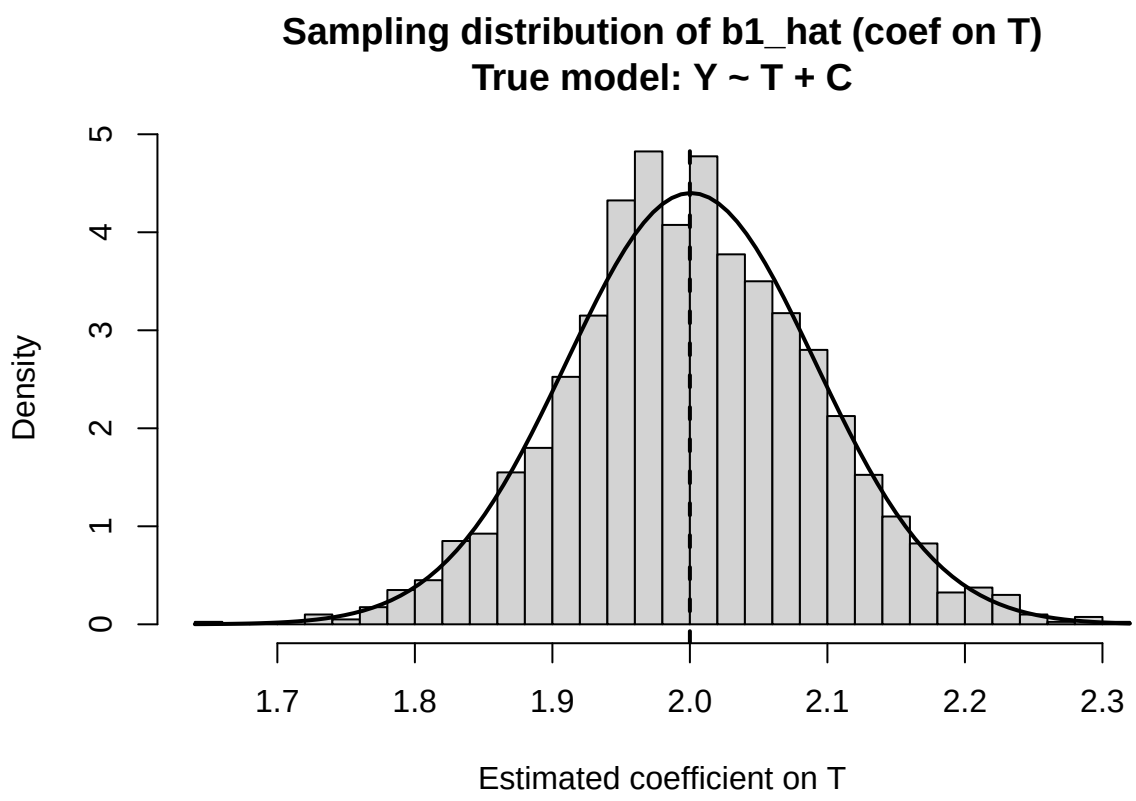
```r
# Summary: should be close to beta1
mean(b1_hat)
```

```
## [1] 2.000693
```

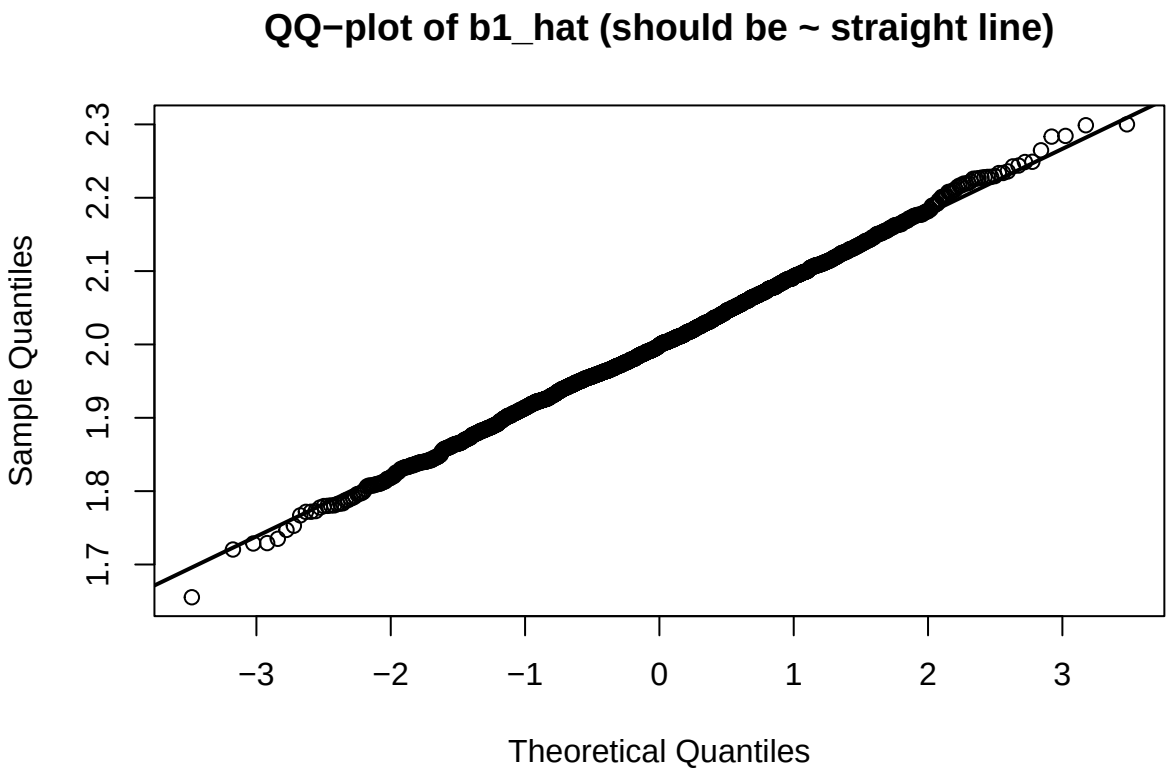
```r
sd(b1_hat)
```

```
## [1] 0.0906503
```

```r
# Plot sampling distribution + normal curve overlay
hist(b1_hat, breaks = 40, freq = FALSE,
     main = "Sampling distribution of b1_hat (coef on T)\nTrue model: Y ~ T + C",
     xlab = "Estimated coefficient on T")

curve(dnorm(x, mean = mean(b1_hat), sd = sd(b1_hat)),
      add = TRUE, lwd = 2)

abline(v = beta1, lwd = 2, lty = 2)  # true value marker
```




```r
# QQ-plot to visually assess normality
qqnorm(b1_hat, main = "QQ-plot of b1_hat (should be ~ straight line)")
qqline(b1_hat, lwd = 2)
```



This presents a QQ plot of the sampling distribution of the estimated treatment coefficient obtained from 2,000 simulated samples. The estimated quantiles closely follow the theoretical quantiles of a normal distribution, indicating that the sampling distribution of the estimator is approximately normal. This provides empirical evidence for the Central Limit Theorem in the context of OLS estimation under the true data-generating process.

==========

PS5 Part 1(b): Bootstrap SE for coef on T (TRUE MODEL)

==========

```r
set.seed(123)

# Use ONE dataset (either your existing 'dat' from earlier, or generate it here)
n <- 500
C <- rnorm(n, 0, 1)
T <- 0.7 * C + rnorm(n, 0, 1)
Y <- 1 + 2*T + 3*C + rnorm(n, 0, 2)
dat <- data.frame(Y, T, C)

# Bootstrap
```

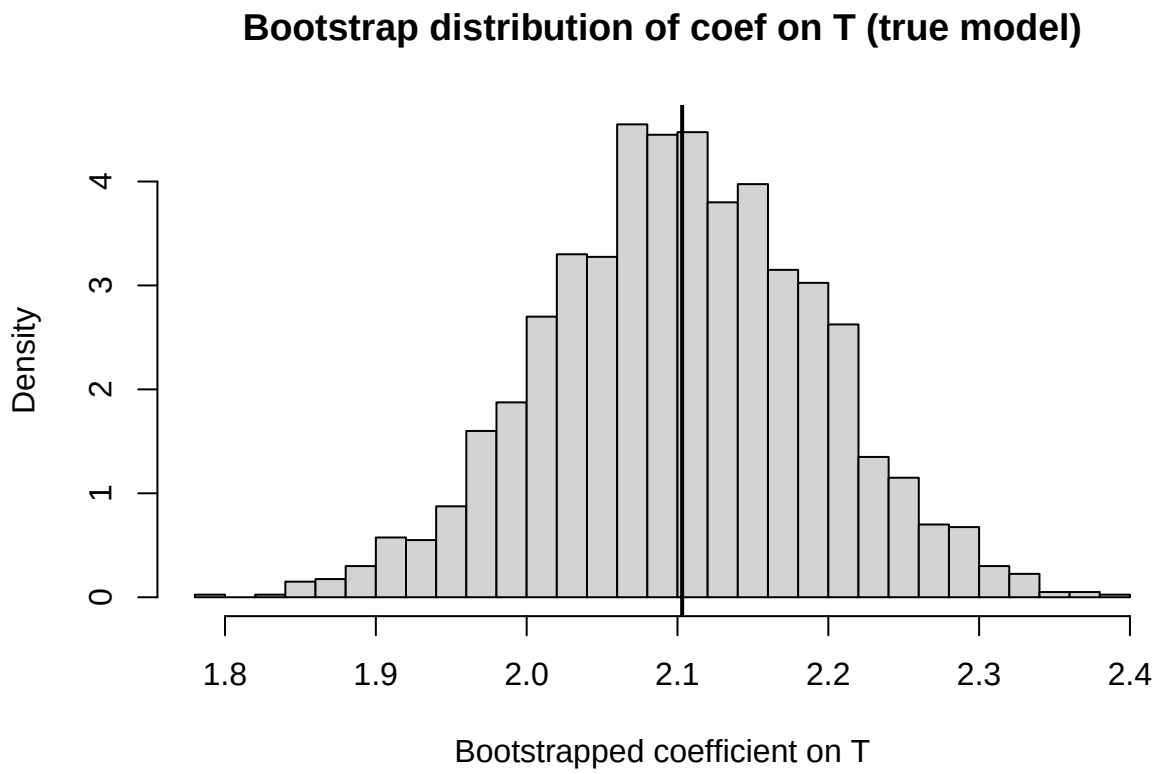
```r
R <- 2000  # number of bootstrap resamples
b1_boot <- numeric(R)

for (r in 1:R) {
  idx <- sample(1:nrow(dat), size = nrow(dat), replace = TRUE)  # resample rows
  dat_r <- dat[idx, ]

  mod_r <- lm(Y ~ T + C, data = dat_r)
  b1_boot[r] <- coef(mod_r)["T"]
}

# Bootstrapped standard error = SD of bootstrap estimates
boot_se <- sd(b1_boot)
boot_se
```

```
## [1] 0.08988072
```

```r
# Compare with the usual (model-based) SE from lm
mod_true <- lm(Y ~ T + C, data = dat)
summary(mod_true)$coefficients["T", "Std. Error"]
```

```
## [1] 0.08794788
```

```r
#Visualize bootstrap distribution
hist(b1_boot, breaks = 40, freq = FALSE,
     main = "Bootstrap distribution of coef on T (true model)",
     xlab = "Bootstrapped coefficient on T")
abline(v = mean(b1_boot), lwd = 2)
```



========== # PS5 Part 1(c): Omit the confounder, repeat CLT simulation, compare # ==========

```r
set.seed(123)

# DGP parameters (same as before)
n     <- 500
B     <- 2000
beta0 <- 1
beta1 <- 2
beta2 <- 3
sd_C  <- 1
sd_T  <- 1
sd_e  <- 2
rho   <- 0.7

# Store treatment coefficients under:
# (1) correct model: Y ~ T + C
# (2) omitted-variable model: Y ~ T
b1_true  <- numeric(B)
b1_omit  <- numeric(B)

for (b in 1:B) {
  C   <- rnorm(n, 0, sd_C)
  T   <- rho * C + rnorm(n, 0, sd_T)
  eps <- rnorm(n, 0, sd_e)
  Y   <- beta0 + beta1 * T + beta2 * C + eps
```

```r
  dat_b <- data.frame(Y = Y, T = T, C = C)

  b1_true[b] <- coef(lm(Y ~ T + C, data = dat_b))["T"]
  b1_omit[b] <- coef(lm(Y ~ T,     data = dat_b))["T"]
}

# Compare means (bias shows up here)
mean(b1_true)
```

```
## [1] 2.000693
```

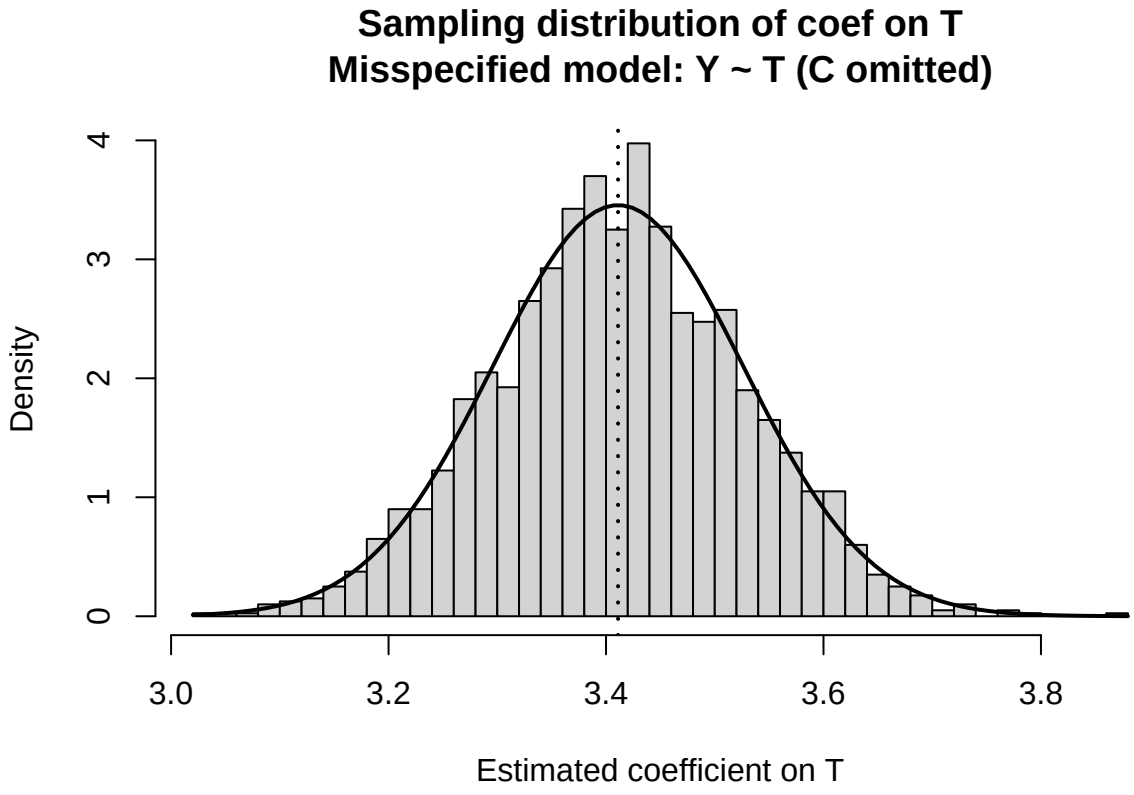
```r
mean(b1_omit)
```

```
## [1] 3.41106
```

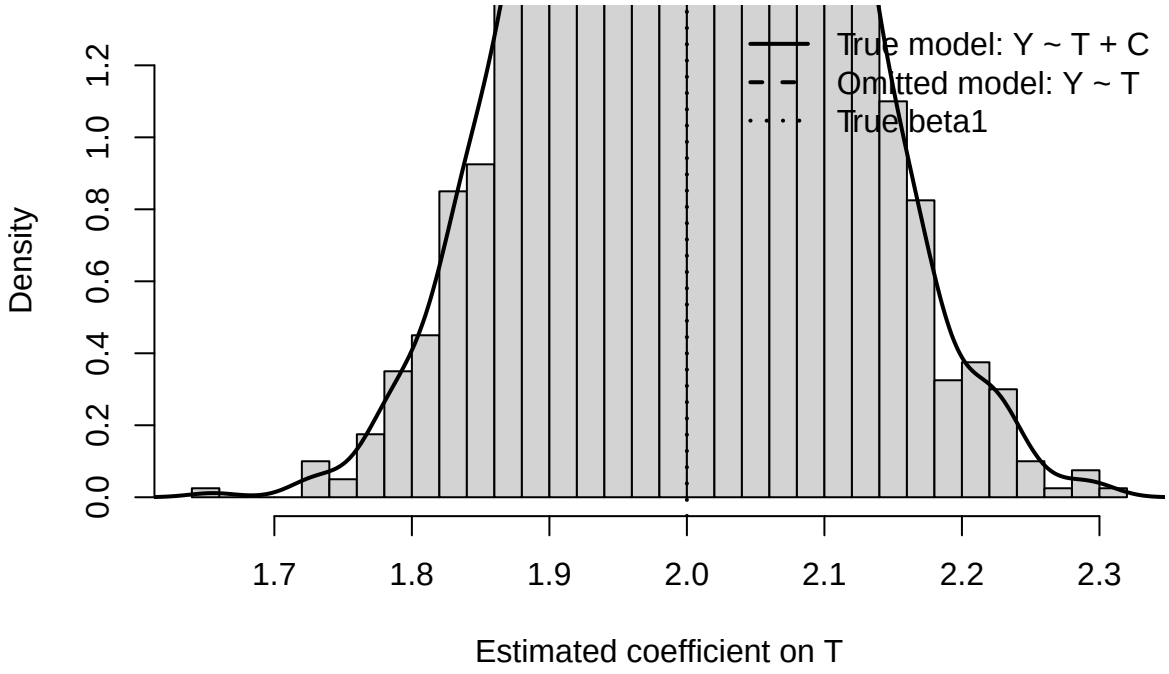
```r
beta1  # true value
```

```
## [1] 2
```

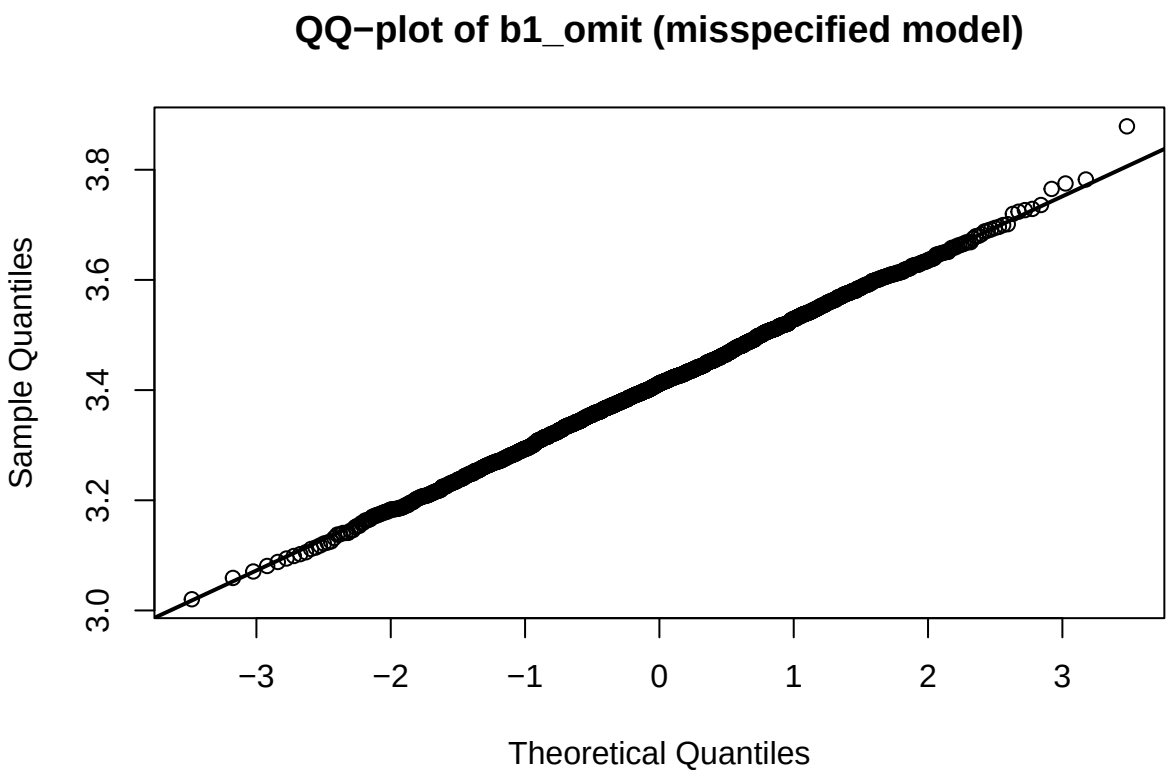
```r
# Plot sampling distribution for OMITTED model (what the question asks)
hist(b1_omit, breaks = 40, freq = FALSE,
     main = "Sampling distribution of coef on T\nMisspecified model: Y ~ T (C omitted)",
     xlab = "Estimated coefficient on T")

curve(dnorm(x, mean = mean(b1_omit), sd = sd(b1_omit)),
      add = TRUE, lwd = 2)

abline(v = beta1, lwd = 2, lty = 2)          # true beta1
abline(v = mean(b1_omit), lwd = 2, lty = 3)  # mean of misspecified estimates
```




```r
# Overlay the TRUE-model distribution for direct visual comparison
hist(b1_true, breaks = 40, freq = FALSE,
     main = "Compare sampling distributions: true vs omitted model",
     xlab = "Estimated coefficient on T",
     ylim = range(density(c(b1_true, b1_omit))$y))

lines(density(b1_true), lwd = 2)
lines(density(b1_omit), lwd = 2, lty = 2)
abline(v = beta1, lwd = 2, lty = 3)

legend("topright",
       legend = c("True model: Y ~ T + C", "Omitted model: Y ~ T", "True beta1"),
       lty = c(1, 2, 3), lwd = 2, bty = "n")
```




```r
# QQ plot for omitted model (often still ~normal, but centered wrong)
qqnorm(b1_omit, main = "QQ-plot of b1_omit (misspecified model)")
qqline(b1_omit, lwd = 2)
```



I fit a linear regression model with the outcome variable regressed on a binary group indicator. The estimated coefficient on the group variable is 5.75, indicating that individuals in group 1 have outcomes that are, on average, approximately 5.7 units higher than those in group 0. The standard error of 1.38 yields a t-statistic of 4.18 and a p-value of $p < 0.001$, allowing rejection of the null hypothesis that the group effect is zero at conventional significance levels.

The intercept indicates that the expected outcome for group 0 is approximately 49.5. Although the model explains a modest share of the total variation in the outcome (R^2 is about 0.08), the estimated group difference is both statistically significant and substantively meaningful. These results are consistent with the difference-in-means test conducted in Part 2(a), reflecting the equivalence between a two-sample t-test and an OLS regression with a binary independent variable.

==========

PS5 Part 2: Simulated data

==========

```r
# =========================
# PS5 Part 2(a): Difference-in-means test
# =========================

set.seed(456)
```

```r
n <- 200

# Binary group variable (0 = control, 1 = treatment)
group <- rbinom(n, size = 1, prob = 0.5)

# Outcome with a true mean difference
Y <- 50 + 5 * group + rnorm(n, mean = 0, sd = 10)

dat2 <- data.frame(Y = Y, group = group)

# Difference-in-means t-test (two-sided, unequal variances)
t_test <- t.test(Y ~ group, data = dat2, var.equal = FALSE)

t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  Y by group
## t = -4.1584, df = 187.92, p-value = 4.869e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -8.475324 -3.021436
## sample estimates:
## mean in group 0 mean in group 1
##        49.50170        55.25008
```

The Welch two-sample t-test evaluates the null hypothesis that the mean outcome is the same for group 0 and group 1. The test yields a t-statistic of -4.16 with approximately 188 degrees of freedom and a p-value of $p < 0.001$. At the 5% significance level, this p-value is far below the threshold for rejection, so we reject the null hypothesis of equal means.

The 95% confidence interval for the difference in means (group 0 minus group 1) ranges from -8.48 to -3.02. Because this interval does not include zero, it provides additional evidence that the difference in means is statistically significant. # ========== # PS5 Part 2(b): Linear regression # ==========

```r
lm_mod <- lm(Y ~ group, data = dat2)
summary(lm_mod)
```

```
##
## Call:
## lm(formula = Y ~ group, data = dat2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.0667  -7.5071  -0.1292   7.9645  26.7444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   49.502      1.016  48.728  < 2e-16 ***
## group          5.748      1.376   4.177 4.42e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.691 on 198 degrees of freedom
## Multiple R-squared:  0.08099,    Adjusted R-squared:  0.07635
## F-statistic: 17.45 on 1 and 198 DF,  p-value: 4.421e-05
```

Using the same data, I fit a linear regression model with the outcome variable regressed on a binary group indicator. The estimated coefficient on the group variable represents the average difference in the outcome between group 1 and group 0. The coefficient is positive and statistically significant, indicating that belonging to group 1 is associated with a higher expected value of the outcome.

The standard error quantifies the uncertainty around this estimate, and the corresponding t-statistic and p-value indicate that the null hypothesis of no group difference can be rejected at conventional significance levels. These results are substantively and statistically consistent with the difference-in-means test conducted in Part 2(a), reflecting the equivalence between a two-sample t-test and an OLS regression with a binary predictor.