Problem Set-1

Hong Wei

2025-03-03

Part (a)

```
rm(list=ls())

set.seed(66)

n <- 100
u <- rnorm(n, mean = 0, sd = 5)
g <- rgamma(n, shape = 2,rate = 2)
r <- rbinom(n, size = 1, prob = 0.5)
x <- ifelse(r == 1,
            rgamma(sum(r == 1), shape = 3, rate = 1),
            rgamma(sum(r == 0), shape = 7, rate = 1))
y <- 400 + 5 * x + 200 * r + 10 * g + u
n1 <- rnorm(n, mean = 10, sd = 3)
b <- rnorm(n, mean = 5 + sqrt(x), sd = 3)

data <- data.frame(y, x, r, g, n1, b)
```

From the question we have:

$E[r_i] = 0.5, E[g_i] = 1$;

$E[x_i^*|r_i = 1] = 3$ and $E[x_i^*|r_i = 0] = 7$, which induces a negative correlation between x_i^* and r_i;

$E[x_i^*] = 0.5 \times E[x_i^*|r_i = 1] + 0.5 \times E[x_i^*|r_i = 0] = 5$;

$E[x_i^* r_i] = E[E[x_i^* r_i \mid r_i]] = 0.5 \times E[x_i^* \cdot 1 \mid r_i = 1] + 0.5 \times E[x_i^* \cdot 0 \mid r_i = 0] = 0.5 \times 3 + 0.5 \times 0 = 1.5$;

$Cov(x_i^*, r_i) = E[x_i^* r_i] - E[x_i^*]E[r_i] = 1.5 - 5 \times 0.5 = -1$;

$y_i = 400 + 5x_i^* + 200r_i + 10g_i + u_i$

Part (b)

```
formulas <- list(
  "Model 1" = y ~ x,
  "Model 2" = y ~ x + r,
  "Model 3" = y ~ x + r + g,
  "Model 4" = y ~ x + r + n1,
  "Model 5" = y ~ x + r + b
)

results <- data.frame(Model = character(),
                      Formula = character(),
                      Beta1_Estimate = numeric(),
                      Std_Error = numeric(),
                      stringsAsFactors = FALSE)


for (model_name in names(formulas)) {
  model_fit <- lm(formulas[[model_name]], data = data)
  model_summary <- summary(model_fit)
  beta1 <- model_summary$coefficients["x", "Estimate"]
  se <- model_summary$coefficients["x", "Std. Error"]

  results <- rbind(results, data.frame(Model = model_name,
                                       Formula = deparse(formulas[[model_name]]),
                                       Beta1_Estimate = beta1,
                                       Std_Error = se,
                                       stringsAsFactors = FALSE))
}

print(results)
```

```
##     Model        Formula Beta1_Estimate Std_Error
## 1 Model 1          y ~ x     -16.937697 2.2749488
## 2 Model 2      y ~ x + r       5.492134 0.3037396
## 3 Model 3  y ~ x + r + g       5.184219 0.2200246
## 4 Model 4 y ~ x + r + n1       5.449858 0.3083917
## 5 Model 5  y ~ x + r + b       5.501547 0.3217376
```

- Regression Model 1 only includes fertilizer amount x_i^*, which is:

$$y_i = \beta_0 + \beta_1 x_i^* + \text{ error }_i$$

The probability limit of β_1 should be:

$$plim(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(x_i^*, r_i)}{Var(x_i^*)} + \beta_3 \frac{Cov(x_i^*, g_i)}{Var(x_i^*)}$$

Since $Cov(x_i^*, r_i) = -1, Cov(x_i^*, g_i) = 0, Var(x_i^*) = 9$ and $\beta_2 = 200$, we have:

$$plim(\hat{\beta}_1) = 5 + 200 \times \frac{-1}{9} \approx -17.22$$

The simulated β_1 estimate is approximately -16.94, which is close to the theoretical result. However, both of them deviates from the true value 5 significantly. This happens because x_i^* is negatively related with r_i, i.e. farmers apply more fertilizer to lower-quality land, creating a spurious negative correlation between fertilizer and yield when land quality is not controlled for. The model suffers from severe omitted variable bias. The *se* is approximately 2.27, which is quite large.

- Regression Model 2 includes fertilizer amount x_i^* and land quality r_i, which is:

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \text{error}_i$$

The probability limit of $\hat{\beta}_1$ should be:

$$plim(\hat{\beta}_1) = \beta_1 + \beta_3 \frac{Cov(x_i^*, g_i | r_i)}{Var(x_i^* | r_i)}$$

Since precipitation g_i is independent of both x_i^* and r_i, we have $Cov(x_i^*, g_i | r_i) = 0$. Therefore:

$$plim(\hat{\beta}_1) = \beta_1 = 5$$

This is unbiased despite omitting precipitation because precipitation is uncorrelated with fertilizer application. In this case, the simulated β_1 improves significantly to 5.49, which is close to the true values. The *se* also significantly reduced to approximately 0.30.

- Regression Model 3 includes the complete set of causal variables, which is:

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + \text{error}_i$$

It's correctly specified so it yields unbiased estimates like Model 2:

$$plim(\hat{\beta}_1) = \beta_1 = 5$$

However, for Model 2:

$$\varepsilon_i^{M2} = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 r_i) \approx \hat{\beta}_3 g_i + u_i$$

$$\sigma_\varepsilon^2 = \text{Var}(\beta_3 g_i + u_i) = \beta_3^2 \text{Var}(g_i) + \text{Var}(u_i)$$

For Model 3:

$$\varepsilon_i^{M3} = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 r_i + \hat{\beta}_3 g_i) \approx u_i$$

$$\sigma_\varepsilon^2 \approx \text{Var}(u_i)$$

Model 2's estimate is less precise because the residual variance is higher. Model 3 will have a β_1 closer to true value 5 and a lower *se* – 0.22.

- Model 4 and 5's equations are as follows:

$$\text{Model 4: } y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_4 n_i + \text{error}_i$$

$$\text{Model 5: } y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_4 b_i + \text{error}_i$$

  The simulated β_1 estimates are 5.45 and 5.50 respectively. Similar with Model 2, both the estimation are unbiased. However, they differ in efficiency:

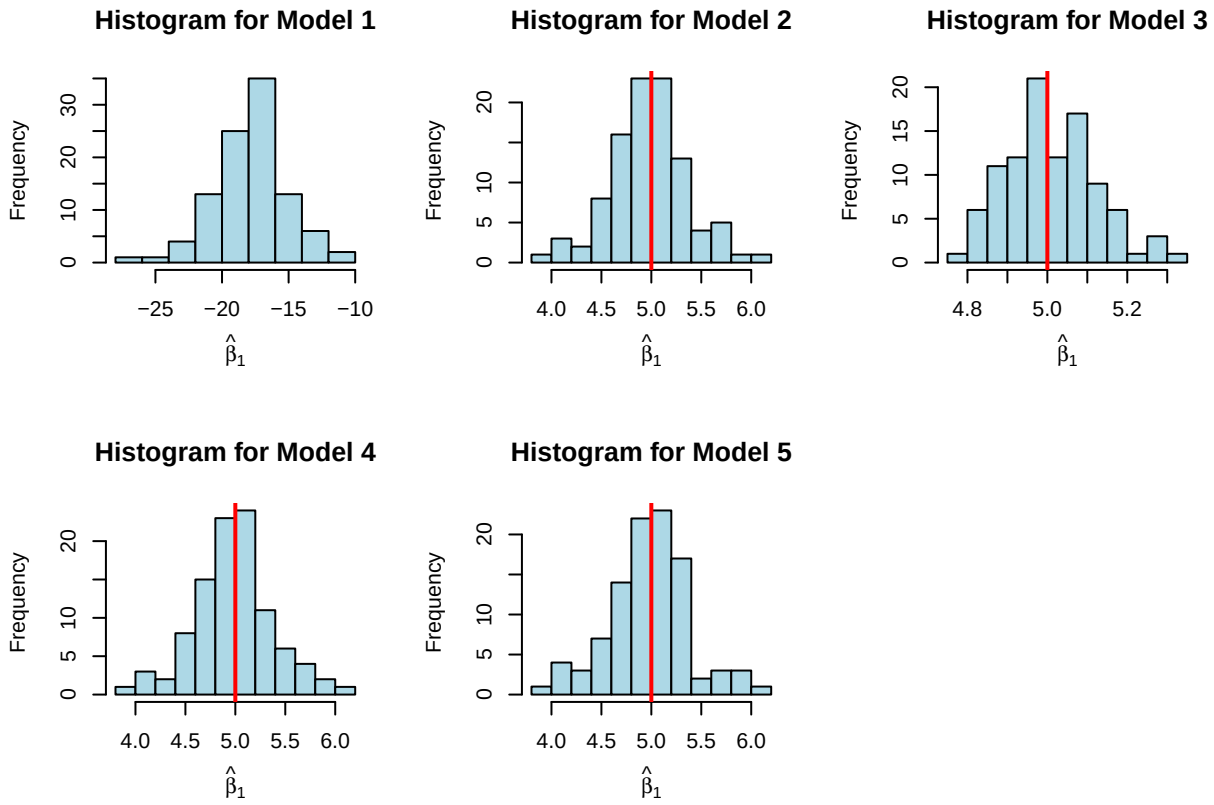
  - Model 4 adds n_i, which is uncorrelated with other variables, causing a slight loss in efficiency. Its *se* becomes higher compared with Model 2 due to the reduction in model degrees of freedom – 0.3084 verse 0.3037;
  - Model 5 adds b_i, which is correlated to x_i^*, creating multicollinearity and potentially substantial efficiency loss. Therefore, its *se* is 0.32, higher than Model 2 and 4. The variance inflation due to multicollinearity in Model 5 is given by:

$$\frac{Var(\hat{\beta}_1^{M5})}{Var(\hat{\beta}_1^{M2})} = \frac{1}{1 - R^2_{x^*|r,b}}$$

Part (c)

```
M <- 100
n <- 100
true_beta1 <- 5
simulate_beta1 <- matrix(NA, nrow = M, ncol = 5)


for (m in 1:M) {
  r <- rbinom(n, size = 1, prob = 0.5)
  x <- ifelse(r == 1,
              rgamma(sum(r == 1), shape = 3, rate = 1),
              rgamma(sum(r == 0), shape = 7, rate = 1))
  g <- rgamma(n, shape = 2, rate = 2)
  u <- rnorm(n, mean = 0, sd = sqrt(5))
  y <- 400 + 5 * x + 200 * r + 10 * g + u
  n1 <- rnorm(n, mean = 10, sd = 3)
  b <- rnorm(n, mean = 5 + sqrt(x), sd = 3)
  df <- data.frame(y, x, r, g, n1, b)

  simulate_beta1[m, 1] <- coef(lm(y ~ x, data = df))["x"]
  simulate_beta1[m, 2] <- coef(lm(y ~ x + r, data = df))["x"]
  simulate_beta1[m, 3] <- coef(lm(y ~ x + r + g, data = df))["x"]
  simulate_beta1[m, 4] <- coef(lm(y ~ x + r + n1, data = df))["x"]
  simulate_beta1[m, 5] <- coef(lm(y ~ x + r + b, data = df))["x"]
}
```

```r
par(mfrow = c(2, 3))
model_names <- c("Model 1", "Model 2", "Model 3", "Model 4", "Model 5")

for (i in 1:5) {
  hist(simulate_beta1[, i],
       main = paste("Histogram for", model_names[i]),
       xlab = expression(hat(beta)[1]),
       col = "lightblue")
  abline(v = true_beta1, col = "red", lwd = 2)
}
```



The results of the Monte Carlo simulations are in general agreement with the analysis in question b). The mean of β_1 across simulations all converge to the theoretical probability limits. Specifically:

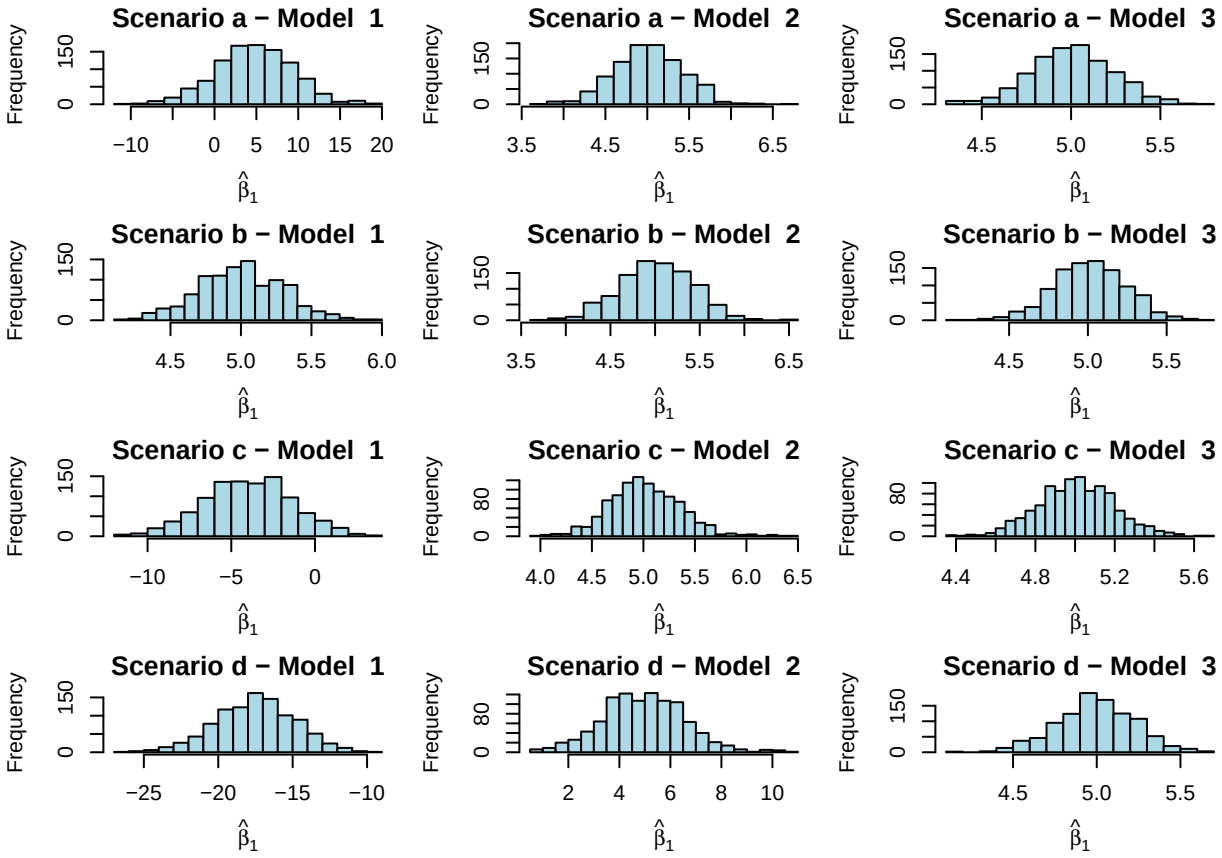
- Model 1: centered around -17, with the largest spread;
- Models 2-5: all centered around 5, but with different spread. Among them:
  - Model 2 produces the least efficient estimates within the group, showing maximal dispersion;

- Model 3 achieves the minimum variance bound, suggesting the most precise estimation;
- Model 4 is similar with Model 3 but exhibits a slightly wider spread due to the inclusion of an extraneous variable.
- Model 5 presents a broader dispersion than Models 3 and 4, indicating increased variability in the estimates.

Part (d)

```
M <- 1000
n <- 100

# Define modification scenarios in a list
# Each element contains a description and parameters
scenarios <- list(
  a = list(desc = "x|r identical: Gamma(5,1)", beta2 = 200, beta3 = 10, prob_r = 0.5, ga
  b = list(desc = "beta2 = 0",                 beta2 = 0,   beta3 = 10, prob_r = 0.5, ga
  c = list(desc = "P(r=1) = 0.1",              beta2 = 200, beta3 = 10, prob_r = 0.1, ga
  d = list(desc = "beta3 = 50",                beta2 = 200, beta3 = 50, prob_r = 0.5, ga
)


simulate_one <- function(scenario_params) {
  # True model parameters
  beta0 <- 400
  beta1 <- 5
  beta2 <- scenario_params$beta2
  beta3 <- scenario_params$beta3
  prob_r <- scenario_params$prob_r

  u <- rnorm(n, mean = 0, sd = 5)
  g <- rgamma(n, shape = 2, rate = 2)
  r <- rbinom(n, size = 1, prob = prob_r)


  if (scenario_params$gamma_shape == "a") {
    x <- rgamma(n, shape = 5, scale = 1)
  } else {
    x <- numeric(n)
    x[r == 1] <- rgamma(sum(r == 1), shape = 3, rate = 1)
    x[r == 0] <- rgamma(sum(r == 0), shape = 7, rate = 1)
  }

  y <- beta0 + beta1 * x + beta2 * r + beta3 * g + u
```

```r
  beta1_vals <- c(
    coef(lm(y ~ x))[["x"]],
    coef(lm(y ~ x + r))[["x"]],
    coef(lm(y ~ x + r + g))[["x"]]
  )
  names(beta1_vals) <- c("Model 1", "Model 2", "Model 3")
  return(beta1_vals)
}

par(mfrow = c(4, 3), mar = c(4, 4, 2, 1))

results <- list()
for (scenario in names(scenarios)) {
  scenario_params <- scenarios[[scenario]]
  beta1_mat <- t(replicate(M, simulate_one(scenario_params)))
  results[[scenario]] <- beta1_mat


  for (i in 1:3) {
    hist(beta1_mat[, i],
         main = paste("Scenario", scenario, "-", "Model ", i),
         xlab = expression(hat(beta)[1]),
         col = "lightblue",
         breaks = 20)
  }
}
```



- Scenario 1: $x_i^*|r=1) = x_i^*|r=0) \sim \Gamma(5,1)$

In this case, both distributions have mean 5, i.e. $E[x_i^*|r=1] = E[x_i^*|r=0] = 5$. Therefore:

$$Cov(x_i^*, r_i) = 0.$$

Therefore we have:

$$plim(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(x_i^*, r_i)}{Var(x_i^*)} = \beta_1$$

This means that in Scenario 1, all the β_1 estimated will be unbiased. The trend of their variances are also the same as before, gradually decreasing as more variables are added, i.e. $Var(\hat{\beta}_1^{M3}) < Var(\hat{\beta}_1^{M2}) < Var(\hat{\beta}_1^{M1})$

- Scenario 2: $\beta_2 = 0$

In this case, land quality r_i has no effect on crop yields y_i. Therefore:

$$plim(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(x_i^*, r_i)}{Var(x_i^*)} = \beta_1$$

Again, all the β_1 estimated are unbiased even though x_i^* and r_i are still correlated.

In terms of variance we have:

$$Var(\hat{\beta}_1^{M3}) < Var(\hat{\beta}_1^{M1}) < Var(\hat{\beta}_1^{M2})$$

- Scenario 3: $r_i = 1$ with probability 0.1

In this case, similar to the calculation in previous questions, we have:

$$Cov(x, r) = E[x_i^* r_i] - E[x_i^*]E[r_i] = 0.3 - (6.6 \times 0.1) = -0.36$$

$$Var(x_i^*) = E[Var(x_i^*|r_i)] + Var(E[x_i^*|r_i]) = 0.1 \times 3 + 0.9 \times 7 + 0.1 \times 0.9 \times (3-7)^2 = 8.04$$

Therefore, for Model 1:

$$plim(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(x_i^*, r_i)}{Var(x_i^*)} + \beta_3 \frac{Cov(x_i^*, g_i)}{Var(x_i^*)} = 5 + 200 \times \frac{-0.36}{8.04} \approx -3.96$$

For Model 2 & 3:

$$plim(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(x_i^*, r_i)}{Var(x_i^*)} + \beta_3 \frac{Cov(x_i^*, g_i)}{Var(x_i^*)} = 5$$

The distribution of $\hat{\beta}_1$ of Model 1 will center around -3.96, with the highest variance. We still see some negative bias, but presumably smaller in magnitude than before. The distribution of $\hat{\beta}_1$ of Model 2 and Model 3 will still both center around 5, with Model 3 having a smaller variance. Therefore:

$$Var(\hat{\beta}_1^{M3}) < Var(\hat{\beta}_1^{M2}) < Var(\hat{\beta}_1^{M1})$$

- Scenario 4: $\beta_3 = 50$

In this case, the true DGP is:

$$y_i = 400 + 5x_i^* + 200r_i + 50g_i + u_i$$

For Model 1, again, its $\hat{\beta}_1$ will be biased as before. However, the bigger difference is now the residual variance of the regression is much higher because g_i contributes strongly to y_i. This means that β_1 will have a larger se.

For Model 2 and 3, their $\hat{\beta}_1$ will also be unbiased as before, and we still have $Var(\hat{\beta}_1^{M3}) < Var(\hat{\beta}_1^{M2})$. However, since g_i has a very large effect on y_i, Model 2's omission of g_i will result in a much larger variance than any previous scenario, i.e. $Var(\hat{\beta}_1^{M2-Scenario4}) > Var(\hat{\beta}_1^{M2-Scenario1/2/3})$.