Gov 2001: Problem Set 3

Spring 2025

March 31, 2025

1 OLS

We solve each part in turn.

1. Take partial derivatives of RSS:

$$\frac{\partial \text{RSS}}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i),$$
$$\frac{\partial \text{RSS}}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i).$$

   Setting these derivatives to zero gives the first-order conditions:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \quad (1),$$
$$\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \quad (2).$$

2. From equation (1), divide both sides by n:

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i = 0,$$

   which simplifies to:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}.$$

3. Plug $\beta_0 = \bar{y} - \beta_1 \bar{x}$ into equation (2):

$$\sum_{i=1}^{n} x_i \left(y_i - (\bar{y} - \beta_1 \bar{x}) - \beta_1 x_i\right) = 0$$
$$\sum_{i=1}^{n} x_i \left((y_i - \bar{y}) + \beta_1(\bar{x} - x_i)\right) = 0$$
$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) + \beta_1 \sum_{i=1}^{n} x_i(\bar{x} - x_i) = 0.$$

The second term simplifies:

$$\sum_{i=1}^{n} x_i(\bar{x} - x_i) = \bar{x}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2 = n\bar{x}^2 - \sum_{i=1}^{n} x_i^2.$$

But it's easier to re-center the equation by subtracting means:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Therefore,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}.$$

2 Interaction

We are given the linear model:

$$Y = \beta_0 + \beta_1 X_a X_g X_e,$$

where:

- $X_a =$ age (in years),
- $X_g = 1\{\text{female}\}$,
- $X_e = 1\{\text{college degree or higher}\}$.

Marginal Effects

To find the marginal effects of each independent variable, we take partial derivatives of Y with respect to each variable.

- Marginal effect of age (X_a):

  $$\frac{\partial Y}{\partial X_a} = \beta_1 X_g X_e.$$

  That is, age only affects income if both $X_g = 1$ (female) and $X_e = 1$ (college degree or higher).

- Marginal effect of gender (X_g):

  $$\frac{\partial Y}{\partial X_g} = \beta_1 X_a X_e.$$

  That is, gender only affects income when age is nonzero (always true) and $X_e = 1$ (college degree or higher).

- Marginal effect of education (X_e):

  $$\frac{\partial Y}{\partial X_e} = \beta_1 X_a X_g.$$

  That is, education only affects income when the person is female ($X_g = 1$) and has nonzero age.

Evaluating the Statements

1. **False.** For both individuals (i) and (ii), $X_g = 0$ since they are men. Therefore, $X_a X_g X_e = 0$ regardless of age or education. So the predicted income for both is just β_0, and the difference is 0.

2. **False.** For both individuals, $X_g = 1$ (female) and $X_e = 0$ (no college). Therefore, the product $X_a X_g X_e = 0$, again regardless of age. So income does not change with age in this group under the model.

3. **True.** For both individuals, $X_g = 1$ (female) and $X_e = 1$ (college). Thus,

$$Y_{30} = \beta_0 + \beta_1 \cdot 30 \cdot 1 \cdot 1 = \beta_0 + 30\beta_1,$$
$$Y_{50} = \beta_0 + \beta_1 \cdot 50 \cdot 1 \cdot 1 = \beta_0 + 50\beta_1.$$

The difference is:

$$Y_{50} - Y_{30} = 20\beta_1.$$

So the statement is true.

3 Power Analysis

(a) We define the following R function to estimate power via simulation:

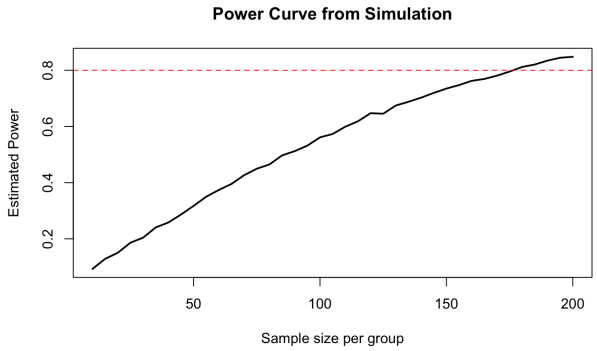
```
simulate_power <- function(n, delta, sd, alpha = 0.05, nsim = 10000) {
  p_vals <- numeric(nsim)
  set.seed(02138)
  for (i in 1:nsim) {
    group1 <- rnorm(n, mean = 0, sd = sd)
    group2 <- rnorm(n, mean = delta, sd = sd)
    test_result <- t.test(group1, group2, var.equal = TRUE)
    p_vals[i] <- test_result$p.value
  }

  power_estimate <- mean(p_vals < alpha)
  return(power_estimate)
}
```

This function simulates n observations for each group, computes a two-sample t-test, and estimates the power as the proportion of rejections of the null hypothesis.

(b) Using the function:

```
set.seed(02138)
simulate_power(n = 30, delta = 3, sd = 10)
simulate_power(n = 50, delta = 3, sd = 10)
simulate_power(n = 100, delta = 3, sd = 10)
```



We have:

- $n = 30$: power ≈ 0.2041
- $n = 50$: power ≈ 0.3173
- $n = 100$: power ≈ 0.5613

As expected, power increases with sample size.

(c) To generate a power curve:

```
set.seed(02138)
n_vals <- seq(10, 150, by = 5)
power_vals <- sapply(n_vals, function(n) {
  simulate_power(n = n, delta = 3, sd = 10)
})

plot(n_vals, power_vals, type = "l", lwd = 2,
     xlab = "Sample size per group", ylab = "Estimated Power",
     main = "Power Curve from Simulation")
abline(h = 0.8, col = "red", lty = 2)
```

This plot shows that more than 150-160 subjects per group are needed to achieve 80% power in this context (the plot is created by running a sequence from 10 to 200).

(d) **Interpretation:** Power increases with sample size. Small samples may fail to detect meaningful effects even if they exist (low power), while large samples improve the chance of detecting true effects. When planning experiments, researchers must balance power, expected effect size, and feasibility (cost, time, participant availability).