Gov 2001: Problem Set 4

Spring 2025

April 16, 2025

Problem Set Instructions:

- This problem set is due on **April 23rd, 11:59 pm** Eastern time. Please upload a PDF of your solutions to **Gradescope**. Please match your answers with the questions.
- We will accept hand-written solutions but we strongly advise graduate students to typeset your answers in LaTeX.
- Citing your sources is always a good practice in academia. Please list the names of other students / sources / AI you obtained help from on this problem set.

1 OLS (30pt)

Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where ε_i are independent errors with mean zero and constant variance.

Here is the objetive function for OLS regressions:

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

1. Derive the first-order conditions by taking partial derivatives with respect to β_0 and β_1.
2. Use the fact that the sample means of x and y are defined as $\bar{x}$ and $\bar{y}$ to simplify the expressions.
3. Show that the OLS estimator for the slope can be written as:

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

2 Interaction (30pt)

Suppose you are interested in the association between income and race, gender, and education. The dependent variable Y is income in USD. The independent variables are age, gender, and education, where

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

You are especially interested in interactions among these variables, and run a linear regression model only with a three-way interaction term:

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

Derive the marginal effects of X_a, X_g, and X_e. Then determine whether the following statements are true or false with brief explanation.

- Comparing (i) men whose age is 30 years and (ii) men whose age is 50 years, we find that (ii) men whose age is 50 years earn $20\beta_1$ dollars more than (i) men whose age is 30 years.
- Comparing (i) women whose age is 30 years and who has no college degree and (ii) women whose age is 50 years and who has no college degree, we find that (ii) women whose age is 50 years and who has no college degree earn $20\beta_1$ dollars more than (i) women whose age is 30 years and who has no college degree.
- Comparing (i) women whose age is 30 years and who has a college degree and (ii) women whose age is 50 years and who has a college degree, we find that (ii) women whose age is 50 years and who has a college degree earn $20\beta_1$ dollars more than (i) women whose age is 30 years and who has a college degree.

3 Power Analysis (30 pt)

We highly recommend you to take a look at this helpful tutorial on power analysis using `pwrss` package in R (link here).

A researcher plans to test whether a new medication leads to improved health outcomes compared to a placebo. Previous studies suggest that the standard deviation of the outcome is about 10 units. The researcher anticipates a true difference in means of 3 units (i.e., treated group mean = 3, control group mean = 0).

You are asked to conduct a power analysis by simulating the performance of a two-sided two-sample t-test.

(a) Run the following R function called `simulate_power()` that takes the following inputs:

- n: sample size per group,

- δ: true difference in means
- sd: standard deviation (assume equal for both groups)
- α: significance level (default to 0.05)
- $nsim$: number of simulations to run.

The function returns the estimated power (i.e., the proportion of simulations in which the p-value from a two-sample t-test is less than α).

```
simulate_power <- function(n, delta, sd, alpha = 0.05, nsim = 10000) {
  p_vals <- numeric(nsim)

  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)
}
```

(b) Use your function to estimate the power of the test for $n = 30$, 50, and 100 per group. Use $\delta = 3$, $\sigma = 10$, and 10,000 simulations per scenario.

(c) Create a plot showing how power changes as a function of sample size. Use $n = 10$ to $n = 150$ per group (in steps of 5). Add a horizontal line at 0.8 to indicate the conventional target for power.

(d) Briefly discuss how sample size affects power and what implications this has for designing experiments.

Hint: You may use `rnorm()` to simulate data and `t.test()` to compute the p-value in each iteration.