Applied Quantitative Methods for the Social Sciences II
MA in Social Sciences, UC3M, Spring 2026

Assignment 2 – Solutions: Part 2 (STAR Dataset)

Contents

1 Data preparation

a) Load the data:

```
library(dplyr)
library(broom)
library(ggplot2)
library(modelsummary)

star = read.csv("https://raw.githubusercontent.com/franvillamil/AQM2/master/
    datasets/star.csv")
```

b) Create factor for class type:

```
star$classtype = factor(star$classtype,
  levels = 1:3,
  labels = c("Small", "Regular", "Regular+Aide"))
```

c) Create factor for race:

```
star$race = factor(star$race,
  levels = 1:6,
  labels = c("White", "Black", "Asian", "Hispanic",
             "Native American", "Other"))
```

d) Binary indicator for small class:

```
star$small = ifelse(star$classtype == "Small", 1, 0)
```

e) Report observations:

```
nrow(star)
sum(!is.na(star$g4reading))
sum(!is.na(star$g4math))
```

The dataset has around 11,600 students total, but only about 4,000–5,000 have non-missing 4th grade test scores (many students were not followed up to 4th grade).

2 Comparing groups

a) Mean reading scores by class type:

```
star %>%
  group_by(classtype) %>%
  summarise(mean_reading = mean(g4reading, na.rm = TRUE))
```

Students in small classes score highest on average, followed by regular+aide, then regular classes. The differences are relatively modest.

b) Bivariate regression:

```
m1 = lm(g4reading ~ small, data = star)
tidy(m1)
```

The coefficient on `small` represents the average difference in 4th grade reading scores between students assigned to small classes vs. all others (regular + regular with aide). Because this is an experiment, this coefficient has a causal interpretation.

c) The regression coefficient on `small` equals the difference in mean reading scores between the small-class group and the non-small group (regular and regular+aide combined). This is because the regression of Y on a binary variable D yields $\hat{\beta} = \bar{Y}_1 - \bar{Y}_0$.

d) Repeat for math:

```
m1_math = lm(g4math ~ small, data = star)
tidy(m1_math)
```

The pattern is similar: small-class students score higher on average. The magnitude may differ slightly between reading and math.

3 Adding controls

a) Multiple regression:

```
m2 = lm(g4reading ~ small + race + yearssmall, data = star)
tidy(m2)
```

b) The coefficient on `small` should remain similar to the bivariate model. This is expected because the STAR data come from a randomized experiment: treatment assignment (small class) is independent of covariates by design. When controls don't change the treatment coefficient, it confirms that randomization was successful. If adding controls changed the estimate substantially, it would suggest a problem with the randomization.

c) The coefficient on `yearssmall` captures the association between spending additional years in a small class and reading scores. A positive coefficient suggests a cumulative benefit of small classes: each additional year in a small class is associated with higher scores. Note that `yearssmall` and `small` partly overlap (being in a small class contributes to years in small class), so the coefficient on `small` now captures the effect of initial assignment net of cumulative exposure.

4 Interactions

a) Interaction model:

```
m3 = lm(g4reading ~ small * race + yearssmall, data = star)
```

b) Print results:

```
tidy(m3)
```

c) The effect of a small class for White students is the coefficient on `small` alone (since White is the reference category). The effect for Black students is `small` + `small:raceBlack`. For example, if `small` = 10 and `small:raceBlack` = 5, the effect for Black students is 15.

d) The interaction terms test whether the benefit of small classes differs across racial groups. Some studies using STAR data find that minority students benefit more from small classes, though the interactions may not always be statistically significant given sample sizes.

5 Presenting results

a) Comparison table:

```
modelsummary(
  list("Bivariate" = m1, "Controls" = m2, "Interaction" = m3),
  vcov = "robust")
```

b) Coefficient plot:

```
p = modelplot(
  list("Bivariate" = m1, "Controls" = m2, "Interaction" = m3),
  vcov = "robust")
p
```

c) Save:

```
ggsave("coefplot_star.png", p, width = 7, height = 5)
modelsummary(
  list("Bivariate" = m1, "Controls" = m2, "Interaction" = m3),
  vcov = "robust",
  output = "table_star.png")
```

6 Brief discussion

a) The STAR data show a positive effect of small class sizes on student achievement: students randomly assigned to small classes score higher on both reading and math tests in 4th grade. The effects are modest but consistent across outcomes.

b) This evidence is more credible than typical observational studies because treatment was randomly assigned. In observational data, class size correlates with school resources, neighborhood characteristics, and student composition, making it impossible to isolate the causal effect. Random assignment eliminates these confounders by design.

c) Limitations include: substantial attrition (many students lack 4th grade scores, and attrition may differ by treatment group), potential compliance issues (some students may have switched classrooms), and the specific context of Tennessee in the 1980s may limit generalizability to other settings.