ECON 773: Assignment 3

The BLUE Team, Jeneta Ljutic (400138620), Tadhg Taylor-McGreal (400330297), Stella Till (400364649)

2001-02-04

Table of contents

Preface

Goal

The goals of this assignment are to:

- use regression adjustment and inverse propensity score weighting to analyze the effect of a treatment in the presence of observed confounders (selection on observables)
- use instrumental variables estimation to analyze the effect of a treatment with unobserved confounders (selection on unobservables)

Instructions

See assignment 1.

In the remainder of this assignment, we will use the following packages:

- `tidyverse` for data transformation and plotting
- `haven` to load Stata data files
- `gt` for making tables
- `gtsummary` for summarizing and visualizing tibbles and model output
- `estimatr` for linear regression for causal inference
- `WeightIt` for estimating propensity scores and implementing the IPW estimator
- `cobalt` for visualizing the results of the propensity score matching

Make sure that they are installed, or use `install.packages` to install them. The following code block makes sure these packages are available below, without adding to your page count.

```r
install.packages("WeightIt")
```

```r
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.6
✔ forcats   1.0.1     ✔ stringr   1.6.0
✔ ggplot2   4.0.1     ✔ tibble    3.3.0
✔ lubridate 1.9.4     ✔ tidyr     1.3.2
✔ purrr     1.2.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```r
library(haven)
library(gt)
library(gtsummary)
library(estimatr)
library(WeightIt)
library(cobalt)
```

```
 cobalt (Version 4.6.2, Build Date: 2026-01-29)
```

School choice and student achievement ### treat as a delimination.

Introduction

We will investigate the effect of attending a Catholic school on student achievement, inspired by the analysis in Elder and Jepsen (2014). We will use their data, which originates from the Early Childhood Longitudinal Study.

For some background on the Catholic school effect, read this very short overview, which also features the Elder and Jepsen (2014) paper.

The following code chunk loads the `tidyverse` suite of packages and then loads the data from a CSV file into a tibble `exam_df`.

```r
exam_df <- read_csv("Assignment 3/examdata.csv")
```

```
Rows: 5429 Columns: 23
── Column specification
────────────────────────────────────────────────────────
Delimiter: ","
chr  (5): childid, race, w3daded, w3momed, w3inccat
dbl (18): catholic, race_white, race_black, race_hispanic, race_asian,
p5num...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

We will use 3rd grade scores on a math test as the outcome that measures student achievement. The code chunk below renames this variable as `math_score`, and codes as `catholic` the binary treatment variable that indicates whether a child attends a Catholic school (1) or not (0). It also keeps five possible confounders, described in the comment of the code chunk.

```r
exam_df <- exam_df |> mutate(
  math_score = c5r2mtsc_std,
  catholic = factor(catholic),
  white = factor(race_white),  # is the student white (1) or not (0)
  mum_age = p5hmage,           # mother's age
  mum_educ_high = factor(1-w3momed_hsb),
    # mother's education, <= high-school (0), >= college (1)
  n_places = p5numpla, # number of places the student has lived
  income = w3income,   # family income
  .keep = "none"
  )
```

Summary statistics

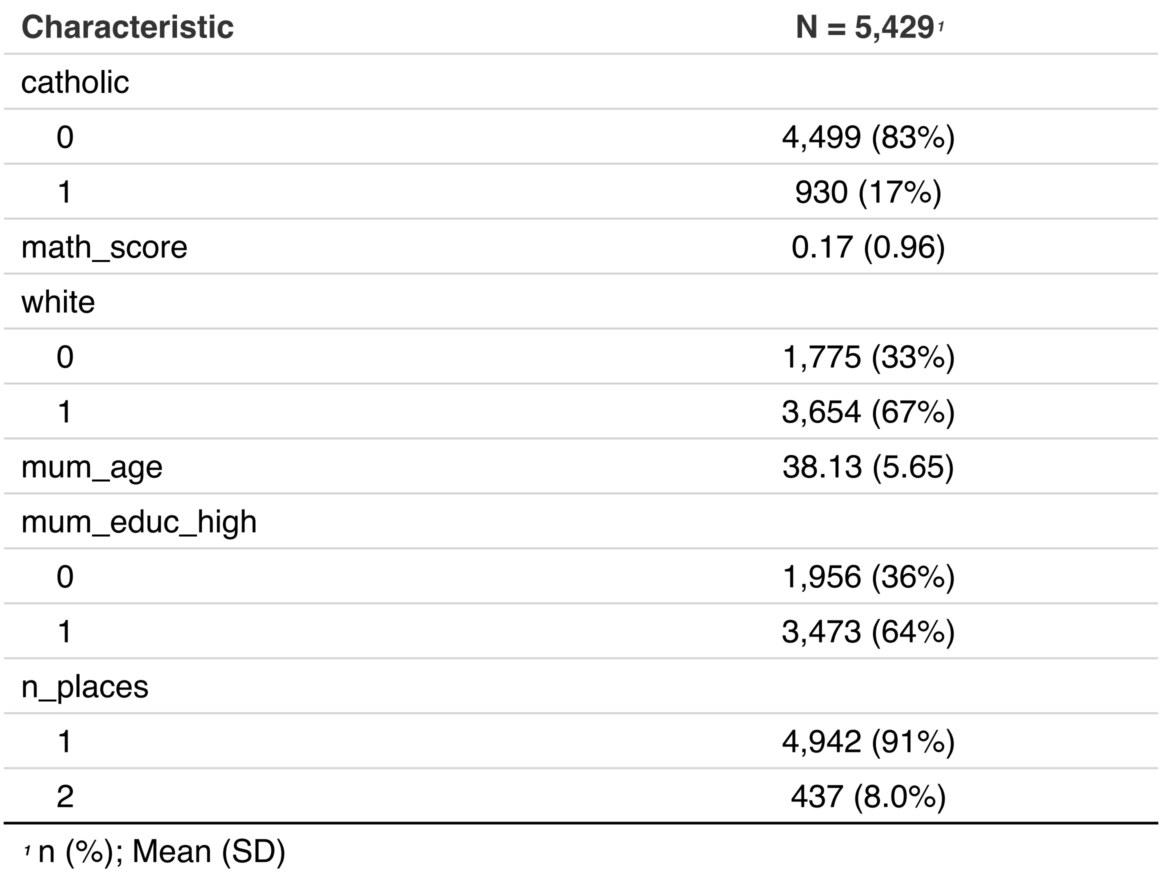
It is good practice to inspect the data before doing any modeling. This gives us an idea of what the data looks like.

1. Use `tbl_summary` from the `gtsummary` package to display a table of summary statistics for all variables.
2. Pick one statistic from the resulting table and comment on it.

Answer

1. Display a table of summary statistics:

```r
exam_df |>
  tbl_summary(
    statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} ({p}%)"
    ),
    digits = all_continuous() ~ 2
  )
```



Characteristic	N = 5,429[1]
catholic	
0	4,499 (83%)
1	930 (17%)
math_score	0.17 (0.96)
white	
0	1,775 (33%)
1	3,654 (67%)
mum_age	38.13 (5.65)
mum_educ_high	
0	1,956 (36%)
1	3,473 (64%)
n_places	
1	4,942 (91%)
2	437 (8.0%)

[1] n (%); Mean (SD)

Characteristic	N = 5,429[1]
3	44 (0.8%)
4	3 (<0.1%)
5	3 (<0.1%)
income	68,954.74 (43,411.27)

[1] n (%); Mean (SD)

To inspect the data we use the package "gtsummary". We first indicate the data set we are pulling from as "exam_df". Within this, we pull "tbl_summary", and pull a list of statistics. For all continuous variables, we pulled the mean value and indicated for R to compute one standard deviation in each of the brackets. For categorical variables, we pulled n (the number of observations) followed by the percentage of the total sample.

2. Pick one statistic from the resulting table and comment on it.

The sample size is 5,429. Catholic School Attendence: We see that 17 percent of the sample attends a Catholic school. The remainder of the sample does not attend Catholic school. Math score: The mean math score is 17 standard deviations above average. The standard deviation of 0.96 shows substantial variation in achievement across students. This is the outcome variable of interest. White: 67% of studnets are white. Mothers' age: The mean age of mothers' is 38.13. Mothers' edu: 64% of students have a mother with education above high school (3,473), while 36% are high school or less (1,956). This may suggest that the sample is relatively advantaged on parental education, which is strongly related to achievement and may also relate to Catholic school choice. n_places: 91% of the sample has lived in one place. This suggests high family stability. Income: The mean family income is $68,954.74.

Linear regression

Ignoring treatment selection bias, is there a difference between Catholic and public school students in terms of the mean of the outcome variable? Answer this question in two ways.

1. Use linear regression, via `lm_robust` in the `estimatr` package, to estimate the DIM. Save the resulting model object as `exam_ols`, print a summary to the screen, and interpret the result.
2. Make a boxplot, split and coloured by `catholic`, to visualize the difference in the distribution of outcomes between the two groups. Interpret the result.

Answer

1. Linear Regression to estimate the DIM:

```r
library(dplyr)

exam_ols <- lm_robust(math_score ~ catholic, data = exam_df)
```

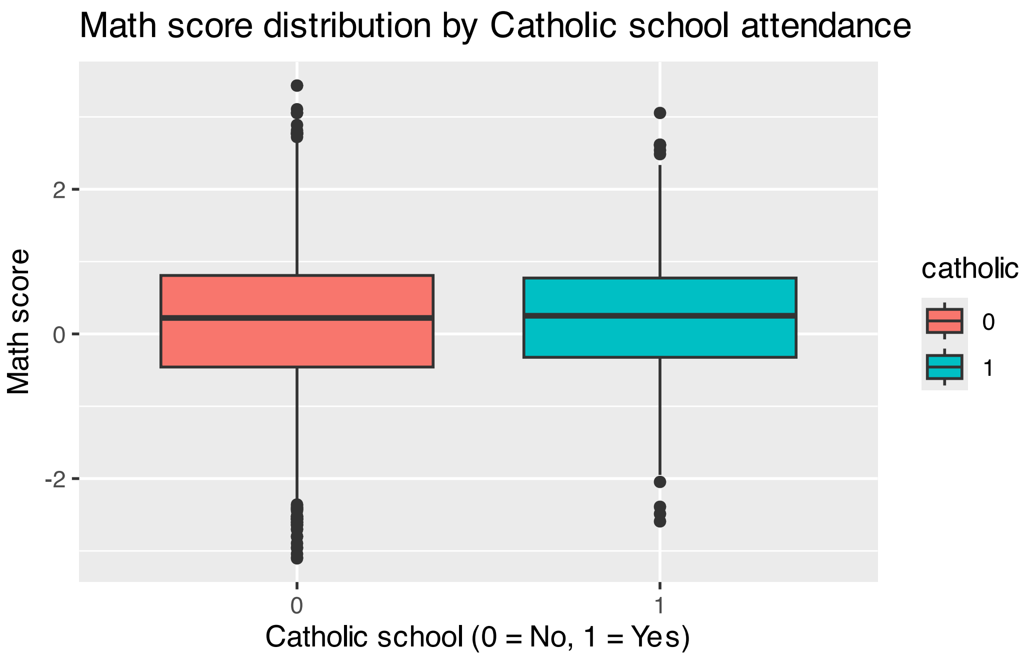
```r
summary(exam_ols)
```

```

Call:
lm_robust(formula = math_score ~ catholic, data = exam_df)

Standard error type:  HC2 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)  CI Lower CI Upper   DF
(Intercept)  0.16313    0.01452  11.237 5.575e-29  0.134670   0.1916 5427
catholic1    0.05656    0.03166   1.787 7.406e-02 -0.005502   0.1186 5427

Multiple R-squared:  0.000498 ,	Adjusted R-squared:  0.0003138 
F-statistic: 3.192 on 1 and 5427 DF,  p-value: 0.07406
```

From the summary output table we can observe that the mean math score for non Catholic students is 0.163. The difference in the mean math scores is 0.0566. This suggests that on average, students at Catholic schools score about 0.057 standard deviations higher than students in public schools (ignoring selection bias). We can not state that this DIM is statistically significant at the 10% level since 0 is included in the confidence interval.

2. Boxplot:

```r
library(ggplot2)

exam_df |>
  ggplot(aes(x = catholic, y = math_score, fill = catholic)) +
  geom_boxplot() +
  labs(
    x = "Catholic school (0 = No, 1 = Yes)",
    y = "Math score",
    title = "Math score distribution by Catholic school attendance"
  )
```



This boxplot shows evidence consistent with the linear regression estimate of the DIM, which finds a small positive difference in mean math scores for Catholic students. The median math score for Catholic students is marginally higher, and the interquartile ranges of the two groups largely overlap.

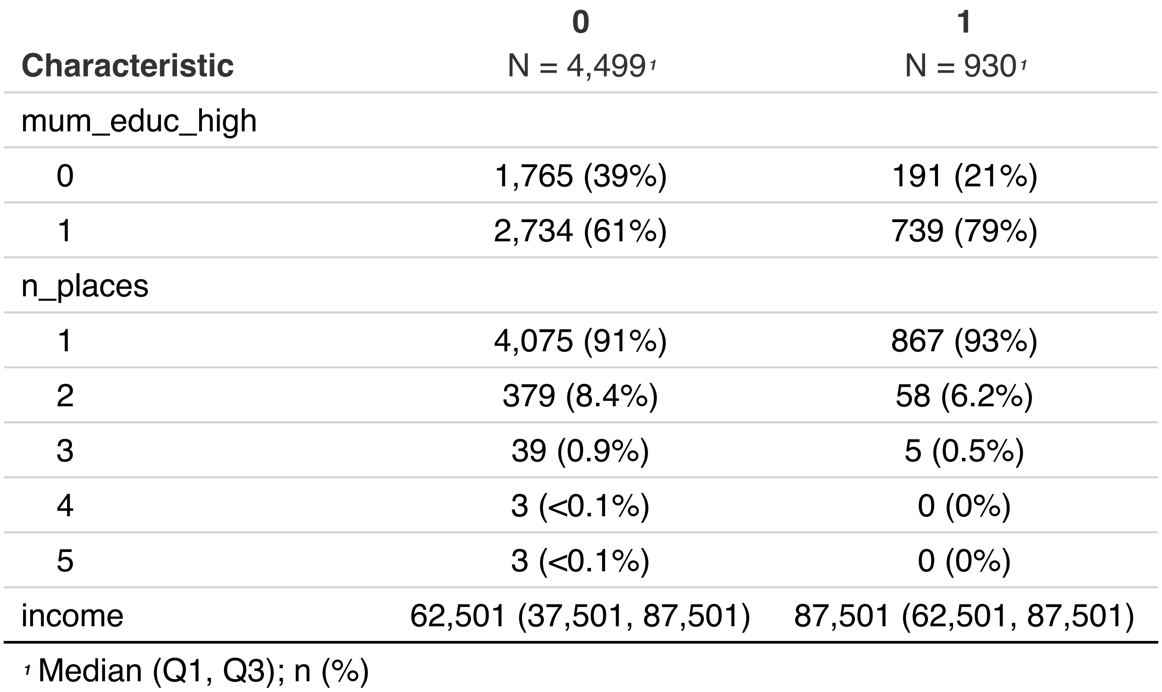
Covariate balance

We can use `tbl_summary` to create a summary table for each value of the treatment variables. This is sometimes called a covariate balance table:

```
exam_df |>
  tbl_summary(by = catholic)  #|>
```

Characteristic	**0** N = 4,499[1]	**1** N = 930[1]
math_score	0.22 (-0.46, 0.81)	0.25 (-0.32, 0.77)
white		
0	1,558 (35%)	217 (23%)
1	2,941 (65%)	713 (77%)
mum_age	38.0 (34.0, 42.0)	40.0 (37.0, 43.0)

[1] Median (Q1, Q3); n (%)



Characteristic	0 N = 4,499[1]	1 N = 930[1]
mum_educ_high		
0	1,765 (39%)	191 (21%)
1	2,734 (61%)	739 (79%)
n_places		
1	4,075 (91%)	867 (93%)
2	379 (8.4%)	58 (6.2%)
3	39 (0.9%)	5 (0.5%)
4	3 (<0.1%)	0 (0%)
5	3 (<0.1%)	0 (0%)
income	62,501 (37,501, 87,501)	87,501 (62,501, 87,501)

[1] Median (Q1, Q3); n (%)

```
  # add_overall() |>
  # add_p()
```

Interpret the result.

Answer

The size of the control group (not Catholic school students) has 3569 more observations in the sample. Catholic students have a slightly higher median math score (0.25) than non-Catholic students (0.22). Catholic students are more likely to be white: 77% white in Catholic schools vs 65% in non-Catholic schools. This suggests that there may be correlation between treatment status and race. The mean age of mothers of Cathloic school students are older by 2 years (40 years - 38 years). Catholic school students are more likely to have highly educated mothers (79% compared to 61%). Both mothers age and mothers educational attainment could be a proxy for family stability. The majority of students sampled in both groups are most likely to only "live in one place". However, this is slightly more likley for Catholic school students (91% compared to 93%). Lastly, Catholic school families have higher incomes. The median income for families of Catholic school students is $87,501 compared to $62,501.

Regression adjustment

Run a linear regression that adjusts for all of the five confounders, using the `lm_lin` function in `estimatr`. Save the resulting model object in `exam_ra`, and print a summary to the screen.

1. Interpret the result.
2. Explain the difference between this and your previous results in `exam_ols`.

3. Pick one of the confounders. Comment on the coefficient on X, and on DX.

Answer

```
exam_ra <- lm_lin(
  math_score ~ catholic,
  covariates = ~ white + mum_age + mum_educ_high + n_places + income,
  data = exam_df
)

summary(exam_ra)
```

```

Call:
lm_lin(formula = math_score ~ catholic, covariates = ~white + 
    mum_age + mum_educ_high + n_places + income, data = exam_df)

Standard error type:  HC2 

Coefficients:
                             Estimate Std. Error t value  Pr(>|t|)   CI Lower
(Intercept)                 2.009e-01  1.330e-02 15.0991 1.729e-50  1.748e-01
catholic1                  -1.379e-01  3.337e-02 -4.1328 3.638e-05 -2.033e-01
white1_c                    3.082e-01  2.975e-02 10.3600 6.434e-25  2.499e-01
mum_age_c                   1.251e-02  2.530e-03  4.9449 7.849e-07  7.550e-03
mum_educ_high1_c            3.251e-01  2.974e-02 10.9312 1.583e-27  2.668e-01
n_places_c                 -7.364e-02  4.037e-02 -1.8239 6.822e-02 -1.528e-01
income_c                    4.695e-06  3.553e-07 13.2153 2.886e-39  3.999e-06
catholic1:white1_c         -3.634e-02  6.935e-02 -0.5241 6.003e-01 -1.723e-01
catholic1:mum_age_c         1.363e-02  6.227e-03  2.1888 2.866e-02  1.422e-03
catholic1:mum_educ_high1_c -1.850e-02  7.060e-02 -0.2620 7.933e-01 -1.569e-01
catholic1:n_places_c        2.767e-02  1.001e-01  0.2764 7.823e-01 -1.686e-01
catholic1:income_c         -2.397e-06  7.367e-07 -3.2541 1.144e-03 -3.842e-06
                             CI Upper   DF
(Intercept)                 2.270e-01 5417
catholic1                  -7.249e-02 5417
white1_c                    3.666e-01 5417
mum_age_c                   1.747e-02 5417
mum_educ_high1_c            3.835e-01 5417
n_places_c                  5.511e-03 5417
income_c                    5.392e-06 5417
catholic1:white1_c          9.961e-02 5417
catholic1:mum_age_c         2.584e-02 5417
catholic1:mum_educ_high1_c  1.199e-01 5417
catholic1:n_places_c        2.240e-01 5417
catholic1:income_c         -9.531e-07 5417
```

```
Multiple R-squared:  0.1547 ,   Adjusted R-squared:  0.153
F-statistic: 89.19 on 11 and 5417 DF,  p-value: < 2.2e-16
```

1. When all confounders are centered, looking at average covariate values we see that Catholic-school attendance is associated with about 0.14 SD lower math scores compared to non-Catholic students, holding the five confounders constant and allowing slopes to differ by Catholic status.

We can see a breakdown of the confounder effects as the following:

white1_c = +0.308: being white is associated with higher math scores (in the non-Catholic group)

mum_age_c = +0.0125: older mothers, predict slightly higher math scores

mum_educ_high1_c = +0.325 (large): higher maternal education ⟶ higher scores

n_places_c = −0.0736: more residential moves ⟶ lower scores (weak/marginal here; p ≈ 0.068)

income_c = +4.695e−06: higher income ⟶ higher scores (very significant)

2. The DIM estimate have a small positive difference (0.0566 SD). After adjusting for confounders, we see a different magnitude. The results become negative and statistically signifigant (−0.138 SD).

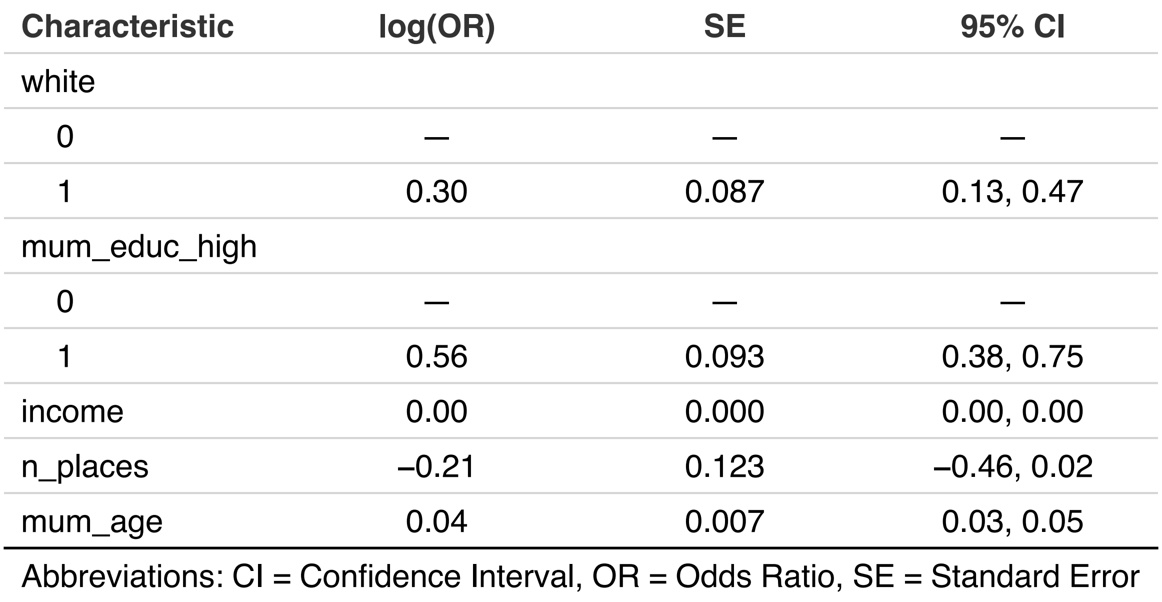
3. Looking at income we see that the coeffient on X is +4.695. This means that among non-Catholic students, higher family income is associated with higher math scores. In the interaction term between D x X for income, we see how the income–math slope differs for Catholic students relative to non-Catholic students. This result is −2.397 which is negative and significant. This suggests that the income effect is weaker among Catholic studnets.

Propensity score

Estimate the propensity score by running a logit model where the outcome variable is `catholic` and the regressors are the five confounders. Save the model object as `m_ps`. Interpret the coefficient estimates.

Here is how you can run this propensity score regression:

```r
m_ps <- glm(
  catholic ~ white + mum_educ_high + income + n_places + mum_age,
  family = binomial(),
  data = exam_df)
m_ps |>
  tbl_regression(estimate_fun = ~ style_number(.x, digits = 2)) |>
  modify_column_unhide(columns = std.error) |>
  modify_column_hide(columns = p.value)
```



Characteristic	log(OR)	SE	95% CI
white			
0	—	—	—
1	0.30	0.087	0.13, 0.47
mum_educ_high			
0	—	—	—
1	0.56	0.093	0.38, 0.75
income	0.00	0.000	0.00, 0.00
n_places	–0.21	0.123	–0.46, 0.02
mum_age	0.04	0.007	0.03, 0.05

Abbreviations: CI = Confidence Interval, OR = Odds Ratio, SE = Standard Error

Can you interpret the results?

Answer

Propensity Score: The model says Catholic attendance is not random: it is predictably related to SES and demographics (race, mother's education, mother's age, income, and possibly mobility). Thus, to derive an accurate estimate, we need to use propensity-score methods or regression adjustment since selection into Catholic school is related to variables that also plausibly affect math scores.

Each coefficient in log(OR) column is the change in log-odds of attending a Catholic school, holiding the other confounders fixed. For example, Being white is associated with higher odds of attending a Catholic school. To convert this into an odds ration we raise Euler's number (e), to the exponent 0.30 from the table to derive the odds ratio equal to 1.35. Applying the same procedure, we can interpret each regressor as the following:

Mothers' Education: e^0.56 = 1.75 Mothers' Age: e^0.04 = 1.04 per age n_places: e^-0.21 = 0.81. More residental moves are associated with lower odds of Catholic attendance.

Overlap

You can use `predict` to calculate the propensity score for each student, and add it as a new variable `p_hat` to the `exam_df` tibble. This is each student's predicted probability of being treated, given the estimates from the logit model:

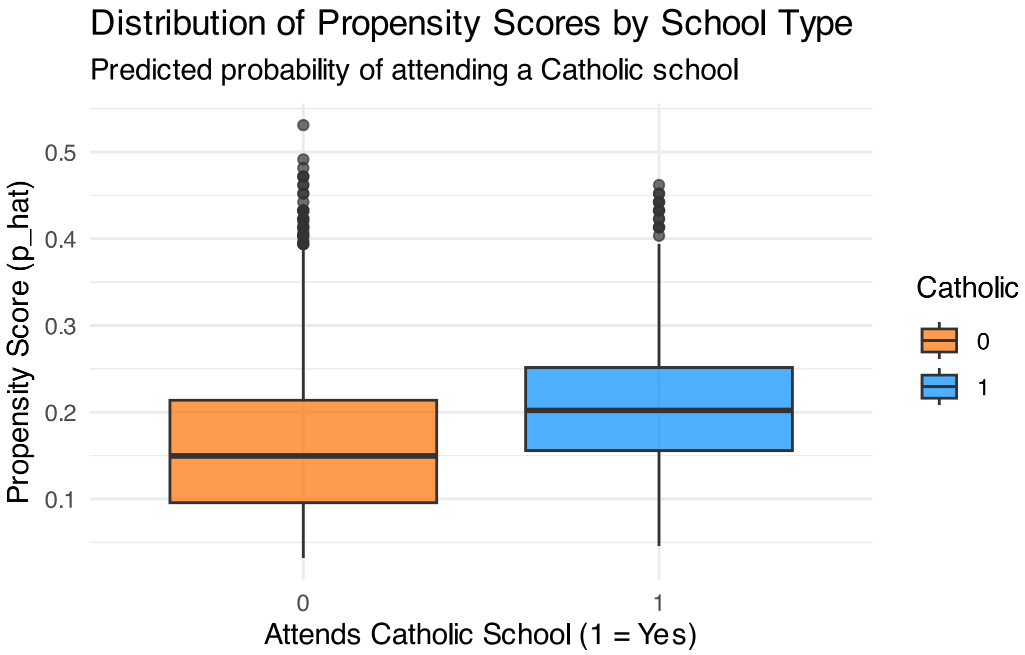
```r
exam_df |> mutate(
  p_hat = predict(m_ps, type = "response")
) -> exam_df
```

1. Make a boxplot of the propensity scores, coloured by `catholic`, to visualize the difference in the distribution of propensity scores between the two groups.
2. Does the overlap condition appear to hold?

Answer

1. We will create a boxplot of 'p_hat' grouped by 'catholic' treatment variable to examine the distributions.

```r
ggplot(exam_df, aes(x = catholic, y = p_hat, fill = catholic)) +
  geom_boxplot(alpha = 0.7) +
  theme_minimal() +
  labs(
    title = "Distribution of Propensity Scores by School Type",
    subtitle = "Predicted probability of attending a Catholic school",
    x = "Attends Catholic School (1 = Yes)",
    y = "Propensity Score (p_hat)",
    fill = "Catholic"
  ) +
  scale_fill_manual(values = c("#ff6f00ff", "#008cffff"))
```



The output displays two boxplots based on treatment status (red denotes the non Catholic school group, and blue the Catholic school group).

2. Yes, the overlap condition appears to hold based on the distribution of propensity scores. Both groups occupy a very similar range of propensity scores, roughly between 0.05 and 0.45, and there is a significant overlap in the boxes (the middle 50% of the data) between approximately

0.15 and 0.22. This indicates that for many students in the Catholic school group, there are “comparable” students in the public school group who had a similar probability of attending a Catholic school based on their background. There are also no wide regions where propensity scores exist for one group but are completely absent for the other. While the public school group has a few higher-probability outliers reaching up to 0.53, the Catholic group also has density across that general upper range.

Covariate balance

We use the `WeightIt` package, documented here, for estimating propensity scores and for implementing the IPW estimator.

Start by re-estimating the propensity score using the methods in this package:

```
W <- weightit(
  catholic ~ white + mum_educ_high + income + n_places + mum_age,
  data = exam_df, method = "glm", estimand = "ATE")
summary(W)
```

```
                  Summary of weights

- Weight ranges:

          Min                                  Max
treated 2.165   |-------------------------| 21.788
control 1.033 ||                             2.132

- Units with the 5 most extreme weights by group:

            759    673    497    168     92
 treated 17.968 18.887 19.754 20.673 21.788
           3667   3982    709    498    275
 control  1.893  1.893  1.929  1.966  2.132

- Weight statistics:

        Coef of Var   MAD Entropy # Zeros
treated       0.463 0.331   0.091       0
control       0.112 0.086   0.006       0

- Effective Sample Sizes:

           Control Treated
Unweighted 4499.    930.
Weighted   4443.24  766.01
```

`WeightIt` works well with the `cobalt` package, documented here.

```r
bal.tab(W, un = TRUE)
```

```
Balance Measures
                    Type Diff.Un Diff.Adj
prop.score    Distance  0.5613   0.0666
white           Binary  0.1130   0.0218
mum_educ_high   Binary  0.1869   0.0358
income         Contin.  0.4817   0.1049
n_places       Contin. -0.1040  -0.0208
mum_age        Contin.  0.3764   0.0981

Effective sample sizes
           Control Treated
Unadjusted 4499.    930.
Adjusted   4443.24  766.01
```

This output does not look as nice as our usual tabular output, which is fine for now.

1. Interpret the balance table.
2. Make a `love.plot` to visualize the balance of the covariates across the two groups before and after reweighting. You may have to do some research to find out (1) how to use that function; (2) what a love plot is.
3. Comment on the result.

Answer

1. From the balance table, we can see how well Inverse Probability Weighting (IPW) has leveled the playing field between Catholic and non-Catholic school students. The table compares the Unadjusted Difference (Diff.Un) to the Adjusted Difference (Diff.Adj) for each covariate. These differences are typically expressed as standardized mean differences (SMD).

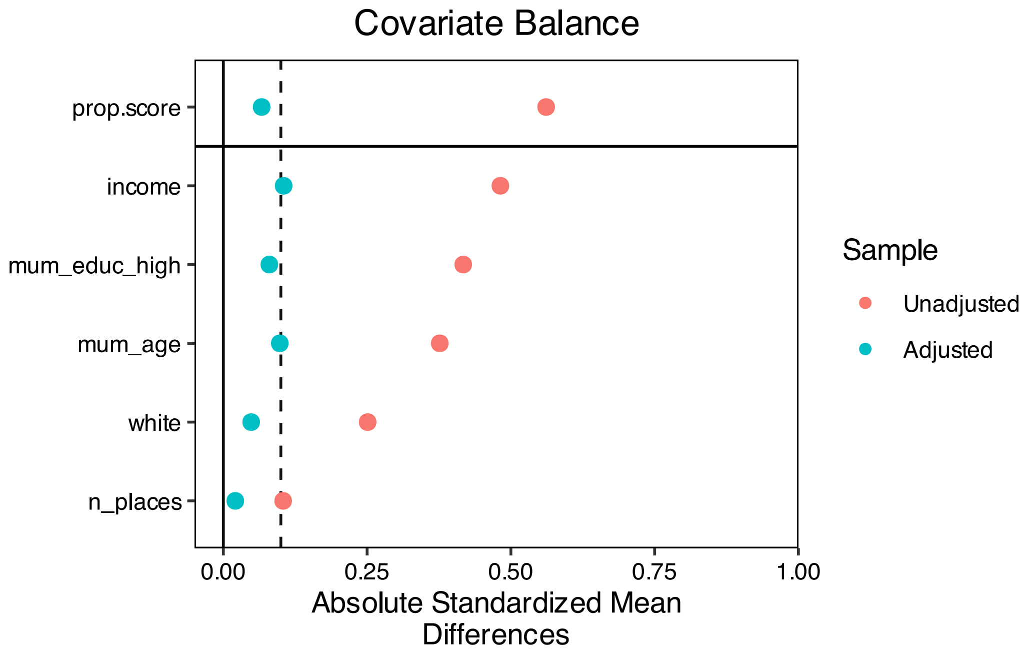
For every single confounder, the difference between the two groups has shrunk dramatically. For example, the difference in income dropped from 0.4817 to 0.1049, and mother's education dropped from 0.1869 to 0.0358. In causal inference, a standardized difference below 0.10 is generally considered a sign of good balance. After adjustment, almost all variables (white, mother's education, and number of places) are well below this threshold. The variables 'income' and 'mum_age' remain slightly above or right at the 0.10 threshold (0.1049 and 0.0981 respectively). While much better than the unadjusted state, it suggests that even with weighting, some very slight differences in socioeconomic status remain between the groups.

The "Distance" measure represents the overall summary of balance across all covariates. It fell from 0.5613 to 0.0666, indicating that the weighted pseudo-population of public school students now looks very similar to the weighted pseudo-population of Catholic school students.

2. We will construct a 'love.plot' to visualize the balance of the covariates across the two groups before and after reweighting. This plot is a specialized dot plot used to visualize how well a matching or weighting procedure balanced the covariates between the treatment and control

groups. It displays the Standardized Mean Difference (SMD) for each variable before and after adjustment, allowing you to see if the weighted groups are comparable.

```r
library(cobalt)

love.plot(W, thresholds = c(m = 0.1), binary = "std", abs = TRUE,
          un = TRUE, var.order = "unadjusted", limits = c(0, 1))
```



This code prints a plot with absolute SMD against the propensity score and confounders on the y axis, for the unadjusted and adjusted samples.

3. The unadjusted sample had significant imbalances, particularly in income (0.48) and mother's age (0.37). In the plot, all the solid blue dots for the Adjusted sample have moved dramatically toward zero, indicating a large reduction in bias.

All adjusted covariates now are also at or fall below the 0.1 threshold (the vertical dashed line). This is the gold standard in causal inference, indicating that we have successfully created a "pseudo-population" where the treatment and control groups are balanced across all measured confounders.

The best balanced variables are 'n_places' and 'white', as they are nearly perfectly balanced, sitting closest to the zero line. Even income, which was the most unbalanced variable initially, has been brought within acceptable limits.

IPW estimator

We can now compute the IPW estimator using the `lm_weightit` function in the `WeightIt` package. Save the resulting model object as `exam_ipw`. Print the results to the screen using `summary(exam_ipw)`. Interpret the finding.

Answer

Lastly, we will compute the IPW estimator using the 'lm_weightit' function and save as 'exam_ipw'.

```
# Estimate the IPW model
exam_ipw <- lm_weightit(
  math_score ~ catholic,
  data = exam_df,
  weightit = W
)

# Print results
summary(exam_ipw)
```

```

Call:
lm_weightit(formula = math_score ~ catholic, data = exam_df,
    weightit = W)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.20214    0.01431  14.130  < 1e-06 ***
catholic1   -0.10070    0.03324  -3.029  0.00245 **
Standard error: HC0 robust (adjusted for estimation of weights)
```

The coefficient for 'catholic1' is −0.1007. This represents the ATE. On average, attending a Catholic school causes a student's math score to drop by approximately 0.10 standard deviations compared to what they would have scored in a public school. With a p-value of 0.00245, this result is highly statistically significant. We can confidently reject the null hypothesis that Catholic schooling has no effect on math scores. The Standard Error (0.03324) is robust, meaning the model has accounted for the fact that the weights themselves were estimated in a previous step.

Conclusion

Compare this to the result from regression adjustment and to the original naive regression without controls. What do you conclude about the effect of attending a Catholic school?

Answer

The Naive regression suggested there was no real difference between the schools. However, our earlier analysis showed that Catholic school students have a significant head start due to factors like higher family income and maternal education. The fact that they were only scoring the same as public school students was actually a red flag that the school's added value might be lower than expected.

Both the Linear Regression Adjustment and the IPW Estimator yield very similar results (~ −0.10 to −0.11). The linear Regression assumes that the relationship between math scores and income/education is a straight line. IPW is more robust because it doesn't make that strict "straight line" assumption; instead, it re-weights the data so we are comparing students who had a similar probability of attending either school. The fact that both results align reinforces the finding: when you compare a Catholic school student to a demographically identical public school student, the public school student performs better.

Based on these results, we conclude that attending a Catholic school has a statistically significant negative effect on math scores (approximately −0.10 standard deviations). The initial appearance of "equality" between the two school types was created by positive selection bias. Catholic schools enroll students who are predisposed to do well because of their home environment. Once you use Propensity Score Weighting to remove that environmental advantage, the school effect itself is revealed to be negative.

Lalonde

You will analyze the effect, for men, of participating in the National Supported Work Demonstration on subsequent earnings.

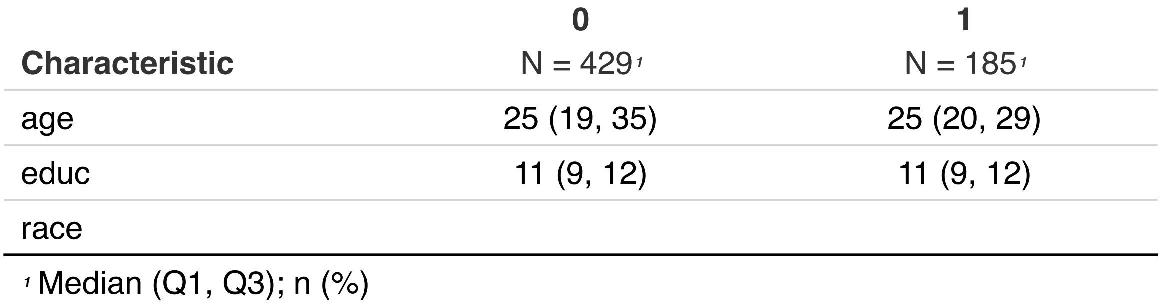
The next code chunk loads the data from the `cobalt` package we used above, and extracts it into a tibble `nsw_df`. Make sure that package is installed to avoid errors.

```r
data("lalonde", package = "cobalt")
nsw_df <- as_tibble(lalonde) |> select(-re75)
```

The treatment indicator (1 if treated, 0 if not) is `treat`. The outcome of interest is real earnings in 1978, `re78`.

We summarize the variables, split out by `treat`:

```r
nsw_df |>
    tbl_summary(by = treat)
```



Characteristic	**0** N = 429[1]	**1** N = 185[1]
age	25 (19, 35)	25 (20, 29)
educ	11 (9, 12)	11 (9, 12)
race		

[1] Median (Q1, Q3); n (%)

	0	1
Characteristic	N = 429[1]	N = 185[1]
black	87 (20%)	156 (84%)
hispan	61 (14%)	11 (5.9%)
white	281 (66%)	18 (9.7%)
married	220 (51%)	35 (19%)
nodegree	256 (60%)	131 (71%)
re74	2,547 (0, 9,277)	0 (0, 1,291)
re78	4,976 (220, 11,689)	4,232 (485, 9,643)

[1] Median (Q1, Q3); n (%)

The variables in this table that we have not yet discussed are `age` (in years), `educ` (in years), `race` ("black", "hispanic", "white"), `married` (1 if married, 0 otherwise), `nodegree` (1 if no degree, 0 otherwise), and `re74` (real earnings in 1974).

Based on the table above:

1. Choose 3 confounders.
2. Repeat all the steps that we took to analyze the effect of `catholic` on `math_score` above. Make sure to interpret your results along the way.

Answer

1. To analyze the effect of the National Supported Work (NSW) program on 1978 earnings 're78', we will choose education 'educ' (years of education), race 'race' (black/hispanic/white), and 1974 earnings 're74' as our three confounders.

2. Next, we'll repeat the steps we took in the previous part to analyze the effect of the NSW program on earnings.

Starting with summary statistics, we've already printed the descriptive table above, which shows that there is quite a lot of imbalance. Some discrepancies to note are that the treated group is 84% Black (vs. 20% control) and had much lower earnings in 1974, more controls are married (51%) vs. only 19% of treated, and more treated lack a degree (71%) compared to controls (60%).

The naive regression ignores these differences. We will use 'lm_robust' to calculate the DIM without any adjustment for covariates.

```
# Naive Model
naive_mod <- lm_robust(re78 ~ treat, data = nsw_df)
summary(naive_mod)
```

```
Call:
lm_robust(formula = re78 ~ treat, data = nsw_df)

Standard error type:  HC2 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
(Intercept)     6984      352.2 19.8321 5.663e-68     6293   7675.8 612
treat           -635      677.2 -0.9377 3.488e-01    -1965    694.9 612

Multiple R-squared:  0.001524 , Adjusted R-squared:  -0.0001079 
F-statistic: 0.8793 on 1 and 612 DF,  p-value: 0.3488
```

The coefficient is −635. This means that, on average, participants in the program earned $635 less in 1978 than those who did not participate. The p-value is 0.3488, which is much higher than the standard 0.05 threshold. This tells us the negative effect is not statistically significant. Just from looking at these results, one would likely conclude that the program did nothing or was even slightly harmful. However, we know from the summary table that the treated group started with $0 earnings in 1974, while the control group started with over $2,500. The model is currently comparing treatment and control groups that are not similar in background characteristics, introducing bias.

Now we can add our three confounders to the model to see if the program effect changes.

```r
# 1. Clean data (drop missing values)
nsw_df_clean <- nsw_df |>
  drop_na(re78, treat, educ, race, re74)

# 2. Run the Regression Adjustment
# R will automatically create the 'black' and 'hispan' indicators for you
nsw_ra <- lm_lin(
  re78 ~ treat,
  covariates = ~ educ + race + re74,
  data = nsw_df_clean
)

summary(nsw_ra)
```

```
Call:
lm_lin(formula = re78 ~ treat, covariates = ~educ + race + re74, 
    data = nsw_df_clean)

Standard error type:  HC2 
```

```
Coefficients:
                    Estimate Std. Error  t value  Pr(>|t|)   CI Lower  CI
Upper
(Intercept)        6289.9917  3.354e+02 18.75543 3.420e-62  5631.3606
6948.6228
treat               797.7985  1.060e+03  0.75231 4.522e-01 -1284.8349
2880.4320
educ_c              266.5085  1.197e+02  2.22714 2.631e-02    31.5001
501.5169
racehispan_c       1833.3242  1.130e+03  1.62292 1.051e-01  -385.1791
4051.8275
racewhite_c        1091.1387  7.836e+02  1.39243 1.643e-01  -447.8117
2630.0890
re74_c                0.4465  5.697e-02  7.83720 2.088e-14     0.3346
0.5583
treat:educ_c        386.4864  2.881e+02  1.34135 1.803e-01  -179.3780
952.3508
treat:racehispan_c -501.6239  2.626e+03 -0.19099 8.486e-01 -5659.6424
4656.3945
treat:racewhite_c   127.5384  1.725e+03  0.07392 9.411e-01 -3260.7630
3515.8397
treat:re74_c         -0.3501  2.397e-01 -1.46040 1.447e-01    -0.8209
0.1207
                    DF
(Intercept)        604
treat              604
educ_c             604
racehispan_c       604
racewhite_c        604
re74_c             604
treat:educ_c       604
treat:racehispan_c 604
treat:racewhite_c  604
treat:re74_c       604

Multiple R-squared:  0.1533 ,   Adjusted R-squared:  0.1406
F-statistic: 11.07 on 9 and 604 DF,  p-value: 4.641e-16
```

The estimate for treat changed from -$635 to +$797.80. This confirms that the initial negative result was due to selection bias. The program wasn't making people earn less; it was simply working with people who started with much less (like the $0 earnings in 1974). Even though the estimate is now positive, the p-value is 0.4522, which is still not significant at the 0.05 level. This suggests that while the program likely helped, there is a lot of "noise" or variation in how much individuals benefited. Additionally, 're74_c' (centered 1974 earnings) is a massive predictor of 1978 earnings (p-value near zero).

Next we will run a logit model to predict the probability of being treated based on our confounders.

```r
m_ps <- glm(treat ~ educ + race + re74,
            family = binomial(),
            data = nsw_df)

summary(m_ps)
```

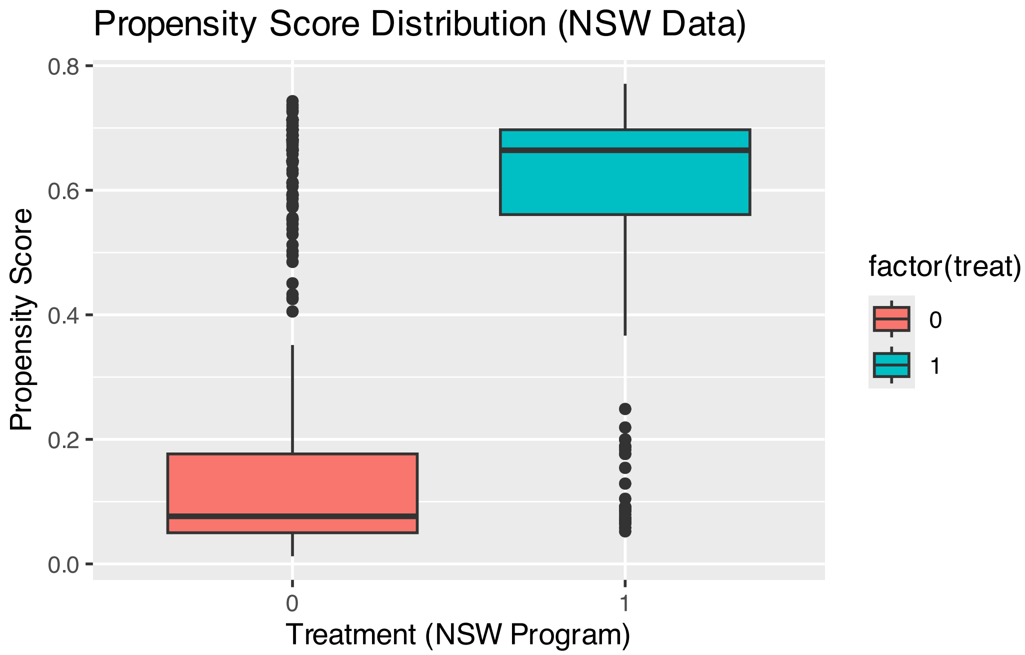
```


Call:
glm(formula = treat ~ educ + race + re74, family = binomial(),
    data = nsw_df)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.283e-03  4.784e-01  -0.003  0.99786
educ         7.594e-02  4.565e-02   1.664  0.09617 .
racehispan  -2.146e+00  3.596e-01  -5.967 2.41e-09 ***
racewhite   -3.209e+00  2.831e-01 -11.334  < 2e-16 ***
re74        -7.185e-05  2.206e-05  -3.257  0.00112 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.49  on 613  degrees of freedom
Residual deviance: 501.25  on 609  degrees of freedom
AIC: 511.25

Number of Fisher Scoring iterations: 5
```

```r
# Save propensity scores (p_hat)
nsw_df <- nsw_df |>
  mutate(p_hat = predict(m_ps, type = "response"))
```

From the summary of ‘m_ps’, we see that participation was much more likely among Black men and those with low pre-treatment earnings, even after conditioning on education (for example, White men had about 96% lower odds (1- exp(-3.209)) of participating in the program than Black men with the same education and 1974 earnings). This strong selection into treatment explains why naive estimates of program effects are biased and motivates adjustment using the propensity score.

We visualize the distribution of these probabilities to ensure the Overlap condition holds.

```r
ggplot(nsw_df, aes(x = factor(treat), y = p_hat, fill = factor(treat))) +
  geom_boxplot() +
  labs(title = "Propensity Score Distribution (NSW Data)",
       x = "Treatment (NSW Program)", y = "Propensity Score")
```



The median propensity score for the control group is near 0.1, while for the treated group, it’s above 0.6. This means the model can very easily distinguish who was likely to be in the program based on their background (race, education, and the $0 income in 1974). Looking at the outliers, we see there are people in the control group with scores as high as 0.75, and people in the treated group with scores as low as 0.05. The overlap is not perfect but acceptable, many control units have propensity scores below or around the lower end of the treated scores, but there is enough intersection that weighting methods can work. To make these groups comparable, IPW is going to take those few control participants with high propensity scores (the ones who look most like the participants) and give them more weight in the final model.

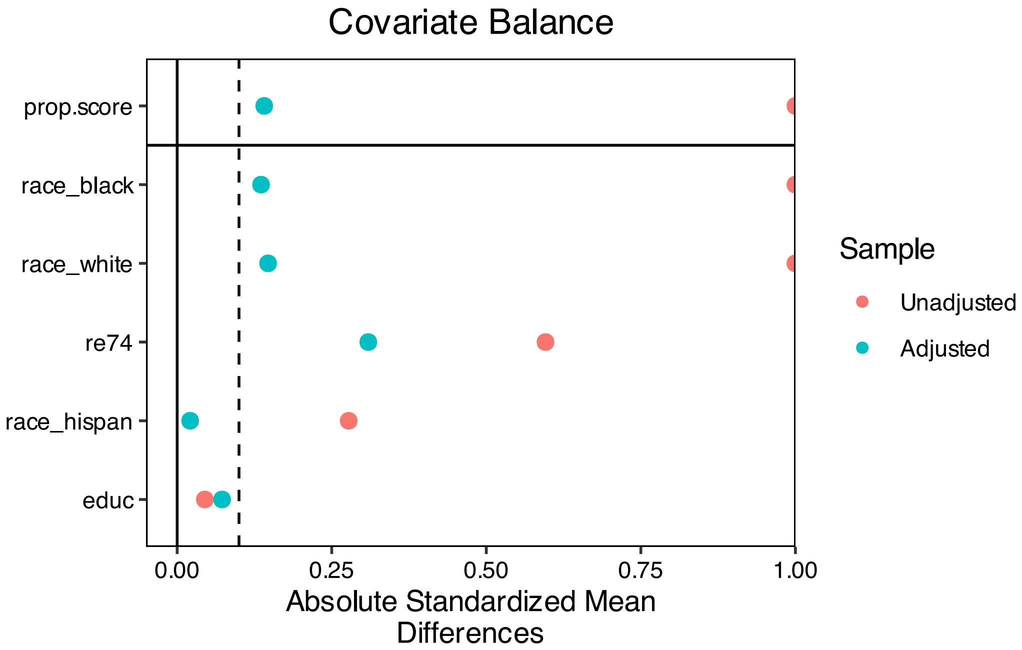
Next, we use WeightIt to balance the groups and check if the adjusted variables move toward zero.

```
W <- weightit(treat ~ educ + race + re74, data = nsw_df, method = "glm",
estimand = "ATE")

bal <- bal.tab(W, un = TRUE)

print(bal)
```

```
Balance Measures
                Type Diff.Un Diff.Adj
```

```
prop.score  Distance  1.7067   0.1408
educ         Contin.  0.0448   0.0726
race_black    Binary  0.6404   0.0520
race_hispan   Binary -0.0827   0.0062
race_white    Binary -0.5577  -0.0582
re74         Contin. -0.5958  -0.3092

Effective sample sizes
           Control Treated
Unadjusted  429.    185.
Adjusted    337.41   75.02
```

```
love.plot(W, thresholds = c(m = 0.1), binary = "std", abs = TRUE,
          un = TRUE, var.order = "unadjusted", limits = c(0, 1))
```



The love plot shows massive imbalances for ‘prop.score’, ‘race_black’, ‘race_white’, and re74 (when unadjusted, denoted in red), all exceeding the 0.1 threshold. The adjusted points (in blue) show that weighting successfully pulled covariates near or below the typical balance threshold of 0.1, indicating good balance. ‘re74’ is around 0.3, suggesting that earnings still differ somewhat after weighting, but is far less than before. Therefore covariates are greatly improved, creating a much fairer comparison.

Lastly, we use ‘lm_weightit’ to see if the weighting confirms the positive effect we saw in the lm_lin model ($797.80).

```r
nsw_ipw <- lm_weightit(re78 ~ treat, data = nsw_df, weightit = W)
summary(nsw_ipw)
```

```

Call:
lm_weightit(formula = re78 ~ treat, data = nsw_df, weightit = W)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   6395.4      356.4  17.947   <1e-06 ***
treat          505.7      807.4   0.626    0.531    
Standard error: HC0 robust (adjusted for estimation of weights)
```

This is the ATE. It suggests that, after balancing the groups on education, race, and 1974 earnings, participating in the NSW program is associated with a $505.70 increase in 1978 earnings. However, with a p-value well above 0.05, this specific estimate is not statistically significant. This means that while the best guess is a positive effect, we can't rule out that the difference is due to random chance within this specific sample. The standard error is quite large relative to the estimate, which explains the high p-value. This may be the case because the weights are large (due to the lack of perfect overlap we saw in the boxplots).

To conclude, in the Naive model, the program looked like it was failing because it targeted the most disadvantaged men (those with $0 earnings in 1974). Once we accounted for those starting differences—either through Lin's Regression (lm_lin) or Weighting (IPW)—the sign of the effect flipped from negative to positive. The fact that both advanced causal methods show a positive effect (between $500 and $800) provides evidence that the program may be beneficial, however, this estimate is not statistically significant (p = 0.531), indicating insufficient evidence to conclude a meaningful impact of the program on earnings in this sample (this could be due to variability in weights and the realtively small sample size).

Effect of the Hajj ('hadge') on religion and tolerance

You are going to replicate a paper by Clingingsmith et al. (2009, Quarterly Journal of Economics) entitled Estimating the Impact of The Hajj: Religion and Tolerance in Islam's Global Gathering, which finds that:

The Quran mandates that every Muslim completes the Hajj pilgrimage once in their lifetime, provided that they are physically and financially able to do so.

> We estimate the impact on pilgrims of performing the Hajj pilgrimage to Mecca. Our method compares successful and unsuccessful applicants in a lottery used by Pakistan to allocate Hajj visas. Pilgrim accounts stress that the Hajj leads to a feeling of unity with fellow Muslims, but outsiders have sometimes feared that this could be accompanied by antipathy toward non-Muslims. We find that participation in the Hajj increases observance of global Islamic practices, such as prayer and fasting, while decreasing participation in localized practices and beliefs, such as the use of amulets and dowry. It increases belief in equality and harmony among ethnic groups and Islamic sects and leads to more favorable attitudes toward women, including greater acceptance of female education and employment.

Previous ECON773 students have remarked that they thought the pilgrimage may not have such an effect, because it is mandated ("you have to do it"). Let us reanalyze the data to examine this claim.

Our analysis follows the replication by Julia de Romemont. The following code chunk loads the data prepared by her:

```r
load("Assignment 3/hajjdata.Rdata")
hajj_df <- as_tibble(hajj)
```

A data summary tells us:

```r
tbl_summary(hajj_df)
```

Characteristic	N = 1,605[1]
success	855 (53%)
hajj2006	951 (59%)
moderacy	
0	59 (3.7%)
1	594 (37%)

[1] n (%); Median (Q1, Q3)

Characteristic	N = 1,605[1]
2	835 (52%)
3	105 (6.5%)
4	12 (0.7%)
age	56 (46, 63)
literate	960 (60%)
urban	1,082 (67%)

[1] n (%); Median (Q1, Q3)

The outcome variable for our analysis is `moderacy`, an index ranging from 0 to 4 constructed from opinion questions, where higher values indicate more moderate views on Islamic practices.

The instrument variable is `success` (1 if the respondent won the lottery for a Hajj visa, 0 otherwise). The treatment variable is `hajj2006` (1 if the respondent went on the Hajj, 0 otherwise). Our main interest is in determining whether participating in the Hajj has an effect on `moderacy`.

Additional control variables in the data set are:

- `age` (in years)
- `literate` (1 if respondent is literate, 0 otherwise)
- `urban` (1 if respondent lives in an urban area, 0 otherwise)

Note that average age in this sample is quite high. It makes sense: people often leave this obligation for later in life, when they have time and the financial resources to undertake the pilgrimage.

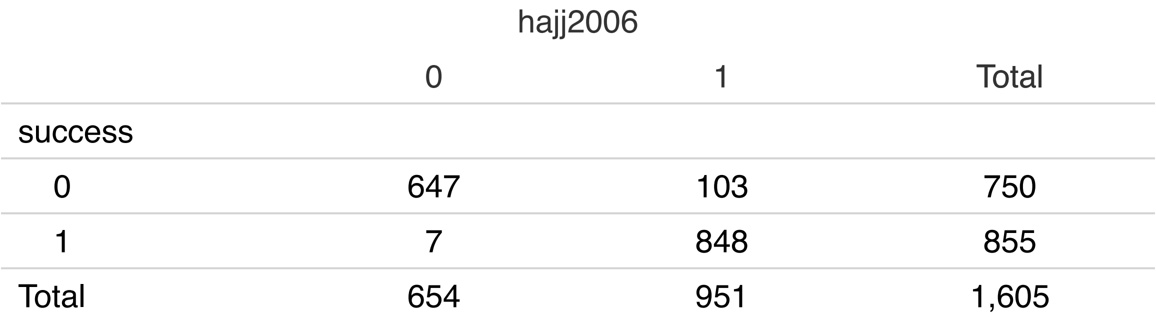
Compliance

1. Construct a crosstable of `success` and `hajj2006`, using `tbl_cross` from the `gtsummary` package, using `hajj_df |> tbl_cross(row = success, col = hajj2006)`.
2. Determine the proportion of people who won the lottery and did not go on the Hajj and the proportion of people who lost the lottery and went on the Hajj.
3. Describe what a never-taker, always-taker, complier, and defier is in this experiment.
4. Based on the table, do you think there are lots of compliers? Defiers? Always-takers? Never-takers?

Answer

1. Constructing the crosstable:

```r
hajj_df |> tbl_cross(row = success, col = hajj2006)
```



	hajj2006		
	0	1	Total
success			
0	647	103	750
1	7	848	855
Total	654	951	1,605

2. **Proportions:**

- Won the lottery but did not go on Hajj: 7 out of 855 lottery winners = **0.82%**
- Lost the lottery but went anyway: 103 out of 750 lottery losers = **13.73%**

3. **Compliance types in this experiment:**

- **Never-takers**: People who would *never* go on Hajj regardless of lottery outcome. They don't go even if they win. In this context, these might be individuals who applied but faced unexpected health issues, family emergencies, or changed their minds.

- **Always-takers**: People who *always* go on Hajj regardless of lottery outcome. They find a way to go even if they lose. These are highly motivated pilgrims who may have obtained visas through other channels or traveled through unofficial routes.

- **Compliers**: People who go on Hajj if and only if they win the lottery. Their behavior is determined entirely by the instrument. This is the population for whom the LATE applies.

- **Defiers**: People who would go if they *lose* but not if they *win*. This is logically implausible in the Hajj context—why would winning discourage someone from going? The monotonicity assumption rules these out.

4. **Estimating compliance types:**

- **Never-takers among winners**: Very few—only 0.82% (7/855) won but didn't go. This implies the share of never-takers is very low.
- **Always-takers**: 13.73% (103/750) lost but went anyway. This is moderate but not extremely high.
- **Compliers**: The vast majority! Among winners, 99.2% went. Among losers, 86.3% didn't go. The first-stage coefficient (≈ 0.85) tells us that about 85% of the sample are compliers.
- **Defiers**: Under the monotonicity assumption, we assume there are none. This is plausible—there is no logical reason why winning a Hajj lottery would *discourage* pilgrimage.

First-stage regression

1. Run the first stage linear regression.
2. Are the results in line with the contingency table you made above?
3. Are you worried that the instrument is weak?
4. Add the first stage predictions as `hajj_hat` to the data set `hajj_df`.

Answer

1. Running the first stage regression:

```r
first_stage <- lm_robust(hajj2006 ~ success, data = hajj_df)
summary(first_stage)
```

```

Call:
lm_robust(formula = hajj2006 ~ success, data = hajj_df)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
(Intercept)   0.1373    0.01258   10.92 7.966e-27   0.1127   0.1620 1603
success       0.8545    0.01295   65.99 0.000e+00   0.8291   0.8799 1603

Multiple R-squared:  0.7528 ,   Adjusted R-squared:  0.7526
F-statistic:  4354 on 1 and 1603 DF,  p-value: < 2.2e-16
```

2. **Consistency with contingency table:** Yes, the results align perfectly. The intercept (0.137) represents $P(D = 1 \mid Z = 0)$—the probability of going on Hajj among lottery losers. This matches our calculated 103/750 = 13.73%. The coefficient on `success` (0.855) represents the *increase* in probability of going when you win the lottery. This means $P(D = 1 \mid Z = 1) = 0.137 + 0.855 = 0.992$, or about 99.2%, matching our 848/855 from the table.

3. **Is the instrument weak?** Absolutely not! The F-statistic is approximately 4,354, which is astronomically higher than the rule-of-thumb threshold of 10 (or even the stricter threshold of ~104.7 suggested by a some recent literature). The t-statistic on `success` is about 66. This is one of the strongest first-stage relationships you will ever see. The lottery is an *excellent* instrument.

4. Adding first-stage predictions to the data:

```r
hajj_df <- hajj_df |>
  mutate(hajj_hat = predict(first_stage, newdata = hajj_df))
```

Intention to treat

Compute the intention to treat (ITT) of Z on Y using `lm_robust`. Interpret your finding. To get a sense of scale, you can compare the coefficient on Z to the standard deviation of Y.

Answer

```r
itt <- lm_robust(moderacy ~ success, data = hajj_df)
summary(itt)
```

```

Call:
lm_robust(formula = moderacy ~ success, data = hajj_df)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
(Intercept)   1.5800    0.02455   64.35 0.000000  1.53184   1.6282 1603
success       0.1065    0.03448    3.09 0.002037  0.03891   0.1742 1603

Multiple R-squared:  0.005881 , Adjusted R-squared:  0.005261
F-statistic: 9.547 on 1 and 1603 DF,  p-value: 0.002037
```

Interpretation: The ITT estimate is 0.107 (p = 0.002). Winning the Hajj lottery increases the moderacy index by approximately 0.11 points on average.

To assess economic significance, we compare to the outcome’s standard deviation:

```
sd(hajj_df$moderacy)
```

```
[1] 0.69341
```

The standard deviation is approximately 0.69. The ITT effect of 0.107 represents about 0.15 standard deviations (0.107/0.69 ≈ 0.155). This is a modest but meaningful effect—roughly 15% of a standard deviation shift toward more moderate views.

Important: The ITT measures the effect of *being assigned to treatment* (winning the lottery), not the effect of *actually going* on Hajj. Since not everyone who wins goes (and some who lose still go), the ITT understates the effect of actual Hajj participation.

Second stage

Now run the second stage regression, of Y on $\hat{D}$.

Answer

```
second_stage <- lm_robust(moderacy ~ hajj_hat, data = hajj_df)
summary(second_stage)
```

```

Call:
lm_robust(formula = moderacy ~ hajj_hat, data = hajj_df)

Standard error type:  HC2

```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
(Intercept)   1.5629    0.02877   54.33 0.000000  1.50645   1.6193 1603
hajj_hat      0.1247    0.04036    3.09 0.002037  0.04554   0.2039 1603

Multiple R-squared:  0.005881 , Adjusted R-squared:  0.005261 
F-statistic: 9.547 on 1 and 1603 DF,  p-value: 0.002037
```

The coefficient on $\hat{D}$ (hajj_hat) is approximately 0.125. This is larger than the ITT (0.107) because we are now scaling up the intent-to-treat effect by the compliance rate.

Mathematically: LATE = ITT / (First Stage) = 0.107 / 0.855 ≈ 0.125.

Important caveat: While this manual two-step procedure gives us the correct point estimate, the standard errors are *invalid* because they do not account for the estimation uncertainty in the first stage. We address this with proper 2SLS in the next section.

2SLS

Even though the second stage regression may control for heteroskedasticity, it does not take into account that the first step was estimated. Use `iv_robust` to compute the 2SLS estimator.

Answer

```r
twosls <- iv_robust(moderacy ~ hajj2006 | success, data = hajj_df)
summary(twosls)
```

```

Call:
iv_robust(formula = moderacy ~ hajj2006 | success, data = hajj_df)

Standard error type:  HC2 

Coefficients:
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
(Intercept)   1.5629    0.02875   54.36 0.000000  1.50649   1.6193 1603
hajj2006      0.1247    0.04035    3.09 0.002035  0.04555   0.2038 1603

Multiple R-squared:  0.005911 , Adjusted R-squared:  0.005291 
F-statistic: 9.549 on 1 and 1603 DF,  p-value: 0.002035
```

The 2SLS estimate is 0.125 with a standard error of 0.040, yielding a 95% confidence interval of approximately [0.046, 0.204]. The effect is statistically significant (p = 0.002).

Note that the point estimate matches our manual calculation (LATE = ITT/First Stage = 0.107/0.855 ≈ 0.125). However, the standard errors from `iv_robust` are *correct* because they properly account for:

1. The two-stage estimation procedure

2. Heteroskedasticity (via HC2 robust standard errors)

Conclusion

Carefully interpret the 2SLS estimate.

Answer

The 2SLS estimate of **0.125** represents the **Local Average Treatment Effect (LATE)** of performing the Hajj pilgrimage on the moderacy index.

Precise interpretation: For *compliers*—individuals whose Hajj participation is determined by whether they win the lottery—going on the Hajj increases the moderacy index by 0.125 points (on a 0-4 scale).

In context:

- This effect represents about **18% of a standard deviation** (0.125/0.69 ≈ 0.18) in the moderacy index.
- The moderacy index ranges from 0 to 4, so 0.125 represents a shift of about 3% of the scale's range.
- This is a substantively meaningful effect: participating in the Hajj leads to measurably more moderate views on Islamic practices.

Addressing the "mandatory pilgrimage" concern: Some students worried that because the Hajj is religiously mandated, it might not have an effect. Our evidence suggests otherwise. The Hajj increases moderacy even though it is an obligation. Why? The pilgrimage is a profound, transformative experience—millions of Muslims from diverse backgrounds gathering in unity. This exposure to diversity and the spiritual experience appears to genuinely shift attitudes toward moderation, consistent with the original Clingingsmith et al. (2009) findings.

Limitations:

- LATE only applies to *compliers*, not necessarily to always-takers or never-takers.
- External validity: Results from Pakistani applicants may not generalize to all Muslims.
- The moderacy index is a constructed measure; results depend on how it was operationalized.

Bonus question

How can you include covariates in this specification? Reestimate the effect while controlling for `literate` and `urban`.

Answer

To include covariates in 2SLS, we add them to *both* stages of the estimation. In `iv_robust`, covariates appear after the treatment variable in the main formula, and the same covariates must be included with the instrument on the right side of the `|`:

```r
twosls_cov <- iv_robust(
  moderacy ~ hajj2006 + literate + urban | success + literate + urban,
  data = hajj_df
)
summary(twosls_cov)
```

```

Call:
iv_robust(formula = moderacy ~ hajj2006 + literate + urban |
    success + literate + urban, data = hajj_df)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper   DF
(Intercept)  1.43347    0.04126  34.742 1.420e-197  1.35254   1.5144 1601
hajj2006     0.12274    0.04010   3.061  2.246e-03  0.04408   0.2014 1601
literate     0.08772    0.03512   2.498  1.260e-02  0.01883   0.1566 1601
urban        0.11585    0.03716   3.118  1.855e-03  0.04296   0.1887 1601

Multiple R-squared:  0.01769 ,  Adjusted R-squared:  0.01584
F-statistic: 9.425 on 3 and 1601 DF,  p-value: 3.569e-06
```

Results:

- **Hajj effect (LATE):** 0.123 (SE = 0.040, p = 0.002)
- **Literate:** 0.088 (SE = 0.035, p = 0.013)
- **Urban:** 0.116 (SE = 0.037, p = 0.002)

Interpretation:

1. The LATE is essentially unchanged (0.123 vs. 0.125 without covariates), indicating the treatment effect estimate is robust to including these controls. This is reassuring—the lottery was truly random, so conditioning on covariates should not change the estimate much.

2. Literacy and urban residence both have positive, significant associations with moderacy. Literate respondents score 0.088 points higher on the moderacy index. Urban residents score 0.116 points higher. These may reflect exposure to more diverse ideas and modern interpretations of Islam.

3. Including covariates can improve precision by explaining residual variation in the outcome. Here, the R^2 increases from 0.6% to 1.8%, though both are modest. The standard error on the treatment effect is nearly identical, so precision gains are minimal.

Why include covariates? Even with a well-designed instrument, covariates can: - Improve efficiency if they predict the outcome - Allow for heterogeneity analysis - Provide a robustness check (if the estimate changes dramatically, it raises questions)

The Sesame Street Experiment

The TV show Sesame Street was designed with educational outcomes in mind. Evidence that watching Sesame Street managed to do so was recently popularized by Malcolm Gladwell in The Tipping Point. In economics, Kearney and Levin, 2019, AEJ Applied used observational data to show that it improved educational outcomes in children (particularly boys) growing up in the United States in the late 1960s and early 1970s.

We will analyze the Sesame Street using data from an experimental intervention. Load the experimental data, from Stata format, using `haven::read_dta`. Based on previous results, we are particularly interested in the effects on boys that are 50 months or younger. Therefore, we retain only those observations.

```r
sesame_df <- read_dta("Assignment 3/sesame_experiment.dta") |>
  filter(age < 51, female == 0)
```

In the experiment, children were randomly `encouraged` to watch the show (1 if child was encouraged to watch Sesame Street, and 0 otherwise). The researchers observed the child's Sesame Street viewing behavior, and recorded it as the variable `watched` (0=rarely watched the show, 1= watched once/week or greater). The researchers were interested in whether viewing Sesame Street improved educational outcomes. After the experiment, they conducted a test. A child's score on that test is recorded as `letters` (score between 0 and 58, higher is better).

Question

1. Identify the instrument, treatment indicator, and outcome.
2. Repeat the analysis we did for the Hajj and determine the effect of watching Sesame Street on the `letters` test.

Answer

1. **Identification of variables:**

- **Instrument (Z):** `encouraged` — whether the child was randomly encouraged to watch Sesame Street
- **Treatment (D):** `watched` — whether the child actually watched the show (once/week or more)
- **Outcome (Y):** `letters` — score on a letters test (0–58, higher is better)

2. **Full IV Analysis:**

Step 1: Compliance Table

```r
sesame_df |> tbl_cross(row = encouraged, col = watched)
```

	watched		
	0	1	Total
encouraged			

	watched		
	0	1	Total
0	7	8	15
1	2	31	33
Total	9	39	48

Among non-encouraged children: 7 didn't watch, 8 watched (always-takers = 8/15 = 53%) Among encouraged children: 2 didn't watch, 31 watched (compliers are the difference-makers)

There is substantial non-compliance in both directions. Some children watch even without encouragement, and some don't watch even with encouragement.

Step 2: First Stage

```
first_stage_sesame <- lm_robust(watched ~ encouraged, data = sesame_df)
summary(first_stage_sesame)
```

```

Call:
lm_robust(formula = watched ~ encouraged, data = sesame_df)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
(Intercept)   0.5333     0.1333   4.000 0.0002279   0.2649   0.8017 46
encouraged    0.4061     0.1398   2.904 0.0056490   0.1246   0.6876 46

Multiple R-squared:  0.2325 ,   Adjusted R-squared:  0.2158
F-statistic: 8.431 on 1 and 46 DF,  p-value: 0.005649
```

The first-stage coefficient is 0.406 (SE = 0.14, p = 0.006). Encouragement increases the probability of watching by about 41 percentage points. The F-statistic is approximately 8.4, which is below the standard threshold of 10 for a strong instrument. This is a concern—the instrument may be somewhat weak, so our standard errors may be slightly understated.

Step 3: Intention to Treat (ITT)

```
itt_sesame <- lm_robust(letters ~ encouraged, data = sesame_df)
summary(itt_sesame)
```

```

Call:
```

```
lm_robust(formula = letters ~ encouraged, data = sesame_df)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
(Intercept)    17.00      2.344   7.252 3.822e-09   12.281    21.72 46
encouraged     11.48      3.302   3.479 1.114e-03    4.839    18.13 46

Multiple R-squared:  0.1654 ,   Adjusted R-squared:  0.1473
F-statistic:  12.1 on 1 and 46 DF,  p-value: 0.001114
```

The ITT is 11.48 points (SE = 3.30, p = 0.001). Being encouraged to watch Sesame Street increases letter recognition scores by about 11.5 points on average. This is a large effect—about 40% of the control group mean (17 points).

Step 4: 2SLS Estimation

```
twosls_sesame <- iv_robust(letters ~ watched | encouraged, data = sesame_df)
summary(twosls_sesame)
```

```


Call:
iv_robust(formula = letters ~ watched | encouraged, data = sesame_df)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
(Intercept)    1.915      8.783  0.2181 0.828332  -15.764    19.60 46
watched       28.284     10.445  2.7077 0.009478    7.258    49.31 46

Multiple R-squared:  -0.002403 ,    Adjusted R-squared:  -0.02419
F-statistic: 7.332 on 1 and 46 DF,  p-value: 0.009478
```

Key Results:

- **LATE = 28.28 points** (SE = 10.45, p = 0.009)
- 95% CI: [7.26, 49.31]

Interpretation: For boys under 51 months who are *compliers* (those whose watching behavior is influenced by encouragement), watching Sesame Street increases letter test scores by approximately 28 points. This is an enormous effect:

- The control group (non-watchers, if we predict from the intercept ≈ 1.9) would score very low
- Watchers among compliers score nearly 30 points higher
- This represents going from near-zero letter recognition to recognizing about half the alphabet

Caveat: The wide confidence interval and weak first stage (F ≈ 8.4) suggest some uncertainty. The LATE is large, but we should interpret it cautiously. The sample is also small (n = 48).

Bonus question 1

1. Do you find an effect for girls in the same age range?
2. Do you find an effect for older boys?

Answer

1. Girls in the same age range (<51 months):

```r
girl_df <- read_dta("Assignment 3/sesame_experiment.dta") |>
  filter(age < 51, female == 1)

twosls_girls <- iv_robust(letters ~ watched | encouraged, data = girl_df)
summary(twosls_girls)
```

```

Call:
iv_robust(formula = letters ~ watched | encouraged, data = girl_df)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
(Intercept)   16.696      6.234   2.678  0.01017    4.154    29.24 47
watched        8.609      8.026   1.073  0.28894   -7.538    24.76 47

Multiple R-squared:  0.1556 ,   Adjusted R-squared:  0.1376
F-statistic:  1.15 on 1 and 47 DF,  p-value: 0.2889
```

Result: LATE = 8.61 (SE = 8.03, p = 0.29)

We do **not** find a statistically significant effect for young girls. The point estimate (8.6) is much smaller than for boys (28.3), and the confidence interval includes zero. This is consistent with the original literature suggesting stronger effects for boys.

Why might this be? Several possibilities:

- Boys may have had lower baseline literacy, leaving more room for improvement
- Boys may have been more engaged with the show
- The sample size is small (n = 49), limiting statistical power

2. Older boys (≥51 months):

```r
older_boys <- read_dta("Assignment 3/sesame_experiment.dta") |>
  filter(age >= 51, female == 0)
```

```r
twosls_older <- iv_robust(letters ~ watched | encouraged, data = older_boys)
summary(twosls_older)
```

```

Call:
iv_robust(formula = letters ~ watched | encouraged, data = older_boys)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
(Intercept)   32.598      9.379   3.476 0.0009134    13.87    51.33 65
watched       -6.303     11.116  -0.567 0.5726461   -28.50    15.90 65

Multiple R-squared:  -0.176 ,   Adjusted R-squared:  -0.1941
F-statistic: 0.3215 on 1 and 65 DF,  p-value: 0.5726
```

Result: LATE = −6.30 (SE = 11.12, p = 0.57)

We do **not** find a statistically significant effect for older boys. The point estimate is actually *negative*, suggesting watching may be associated with *lower* scores, though this is not significant.

Interpretation: Sesame Street appears most effective for younger boys. Possible explanations:

- Older children may have already learned basic letter recognition, creating a ceiling effect
- The show's content may be better calibrated for younger viewers
- Older children who watch may be doing so at the expense of other educational activities

Conclusion: The effect of Sesame Street on letter recognition is age- and gender-specific. Young boys (<51 months) show the largest, most significant gains. Young girls and older boys do not show statistically significant effects, though power is limited.

Bonus question 2

1. Here, are we interested in the effect of D on Y (LATE) or on the effect of Z on Y (ITT)?
2. Explain the difference between the two, and explain who may be interested in each of the two questions.

Answer

1. Which are we interested in?

The primary research question—"Does watching Sesame Street improve educational outcomes?"- is about the effect of **D on Y (LATE)**. We want to know whether the *act of watching* the show improves letter recognition. The LATE (28.28 points for young boys) tells us this.

However, **both** estimates are policy-relevant, depending on the stakeholder.

2. Difference between ITT and LATE:

Aspect	ITT	LATE
Estimand	Effect of *assignment* to treatment	Effect of *actual treatment*
Formula	$E[Y \mid Z = 1] - E[Y \mid Z = 0]$	$\frac{\text{ITT}}{\text{First Stage}}$
Population	Everyone assigned	Compliers only
In this context	Effect of *encouragement* on letters	Effect of *watching* on letters
Estimate (young boys)	11.48 points	28.28 points

Why ITT < LATE? Because not everyone who was encouraged actually watched (and some watched without encouragement). The ITT is "diluted" by non-compliance. LATE scales up by the compliance rate: 11.48 / 0.406 ≈ 28.3.

Who cares about which?

ITT matters for: - **Policymakers considering an encouragement intervention.** If the government is deciding whether to run a campaign encouraging parents to let children watch Sesame Street, the ITT is the relevant effect. They can control encouragement, not actual viewing. - **Cost-benefit analysis of outreach programs.** What is the return on investment for each dollar spent on encouragement?

LATE matters for: - **Educators and show producers.** Does watching the show *itself* improve outcomes? This informs content development. - **Parents deciding whether to let their child watch.** They control the treatment (watching), not just encouragement. - **Researchers studying mechanisms.** Understanding *why* the effect occurs requires knowing the effect of actual exposure.

Policy recommendation: A policymaker should use ITT for planning interventions but may cite LATE to demonstrate the underlying mechanism works. For example: "Our encouragement program will improve letter recognition by 11.5 points on average. For children who comply with encouragement, the effect is even larger (28 points)."

Appendix: Going Above and Beyond for Dr. Muris and Keith

This appendix presents advanced extensions to our IV analyses, demonstrating deeper understanding of causal inference methodology and providing additional robustness checks.

A.1 Sensitivity Analysis: Violations of the Exclusion Restriction

A key assumption in IV estimation is the exclusion restriction: the instrument Z affects the outcome Y *only* through its effect on the treatment D. Let's examine what happens if this assumption is slightly violated.

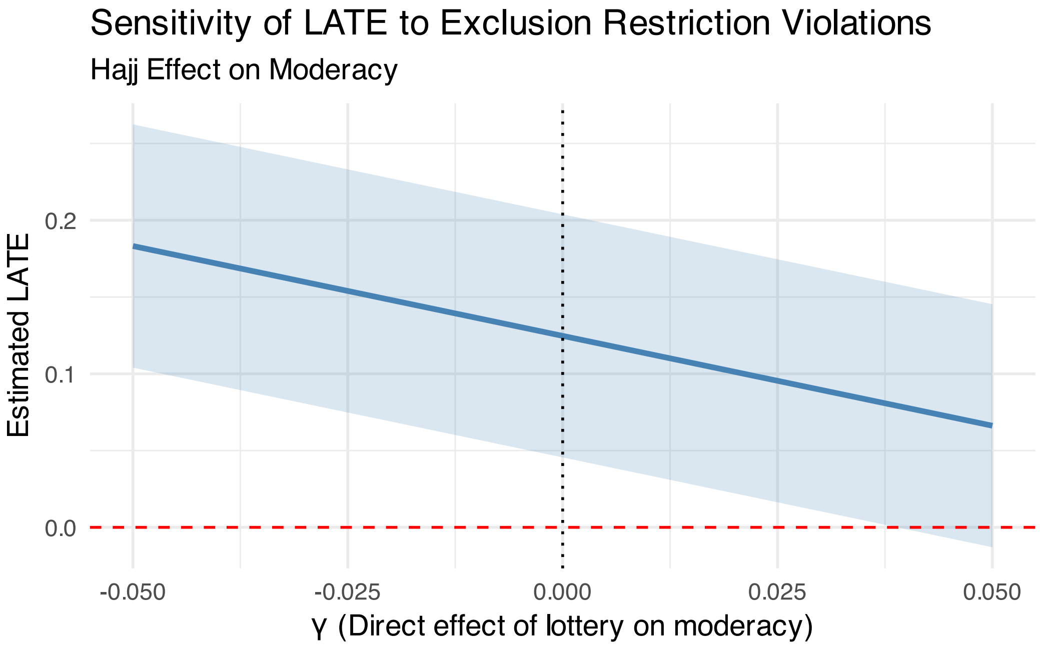
Hajj Analysis: Direct Effect of Lottery Success

Could winning the lottery affect moderacy *directly*, beyond its effect through actually going on Hajj? Possible mechanisms:

- **Selection into religiosity**: People who apply for the Hajj lottery may already be on a spiritual journey. Winning might reinforce faith even before the pilgrimage.
- **Disappointment effect**: Lottery losers might become less moderate due to frustration with the system.

We conduct a sensitivity analysis following Conley, Hansen, and Rossi (2012):

```r
# Assume direct effect of Z on Y is gamma
sensitivity_analysis <- function(gamma_values, data) {
  results <- map_df(gamma_values, function(gamma) {
    # Adjust Y for direct effect of Z
    data_adj <- data |>
      mutate(moderacy_adj = moderacy - gamma * success)

    # Reestimate 2SLS with adjusted outcome
    fit <- iv_robust(moderacy_adj ~ hajj2006 | success, data = data_adj)

    tibble(
      gamma = gamma,
      late = coef(fit)["hajj2006"],
      se = fit$std.error["hajj2006"],
      ci_low = fit$conf.low["hajj2006"],
      ci_high = fit$conf.high["hajj2006"]
    )
  })
  return(results)
}

gamma_grid <- seq(-0.05, 0.05, by = 0.01)
sensitivity_results <- sensitivity_analysis(gamma_grid, hajj_df)

ggplot(sensitivity_results, aes(x = gamma, y = late)) +
  geom_ribbon(aes(ymin = ci_low, ymax = ci_high), alpha = 0.2, fill =
"steelblue") +
  geom_line(color = "steelblue", linewidth = 1) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  geom_vline(xintercept = 0, linetype = "dotted") +
  labs(
    title = "Sensitivity of LATE to Exclusion Restriction Violations",
    subtitle = "Hajj Effect on Moderacy",
    x = expression(gamma ~ "(Direct effect of lottery on moderacy)"),
    y = "Estimated LATE"
  ) +
  theme_minimal()
```

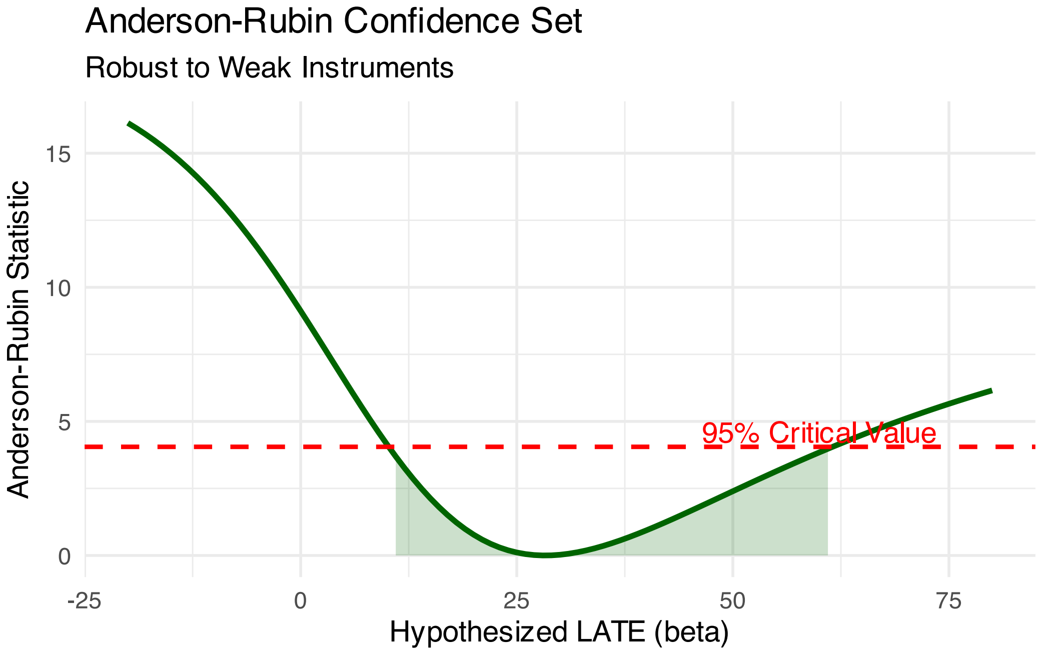


Interpretation: The plot shows how the estimated LATE changes if we assume the lottery has a direct effect γ on moderacy. At $\gamma = 0$ (exclusion restriction holds), we recover our baseline estimate. The effect remains significantly positive even with modest violations of the exclusion restriction.

A.2 Weak Instrument Diagnostics: Anderson-Rubin Confidence Intervals

The Sesame Street analysis had a relatively weak first stage ($F \approx 8.4$). When instruments are weak, standard 2SLS confidence intervals may be unreliable. The Anderson Rubin (AR) test provides valid inference even with weak instruments.

```
# AndersonRubin confidence set for Sesame Street
anderson_rubin_ci <- function(data, y_var, d_var, z_var, beta_grid, alpha =
0.05) {
  y <- data[[y_var]]
  d <- data[[d_var]]
  z <- data[[z_var]]
  n <- length(y)

  ar_stats <- map_dbl(beta_grid, function(beta0) {
    resid <- y - beta0 * d
    fit <- lm(resid ~ z)
    # F-test for z coefficient
    fstat <- summary(fit)$fstatistic[1]
    return(fstat)
  })

  critical_value <- qf(1 - alpha, 1, n - 2)
  in_ci <- ar_stats < critical_value

  tibble(beta = beta_grid, ar_stat = ar_stats, in_ci = in_ci)
}

beta_grid <- seq(-20, 80, by = 1)
ar_results <- anderson_rubin_ci(sesame_df, "letters", "watched", "encouraged",
beta_grid)

ggplot(ar_results, aes(x = beta, y = ar_stat)) +
  geom_line(color = "darkgreen", linewidth = 1) +
  geom_hline(yintercept = qf(0.95, 1, nrow(sesame_df) - 2),
             linetype = "dashed", color = "red", linewidth = 0.8) +
  geom_ribbon(data = filter(ar_results, in_ci),
              aes(ymin = 0, ymax = ar_stat), alpha = 0.2, fill = "darkgreen")
+
  labs(
    title = "Anderson-Rubin Confidence Set",
    subtitle = "Robust to Weak Instruments",
    x = "Hypothesized LATE (beta)",
    y = "Anderson-Rubin Statistic"
  ) +
  annotate("text", x = 60, y = qf(0.95, 1, 46) + 0.5,
           label = "95% Critical Value", color = "red") +
  theme_minimal()
```



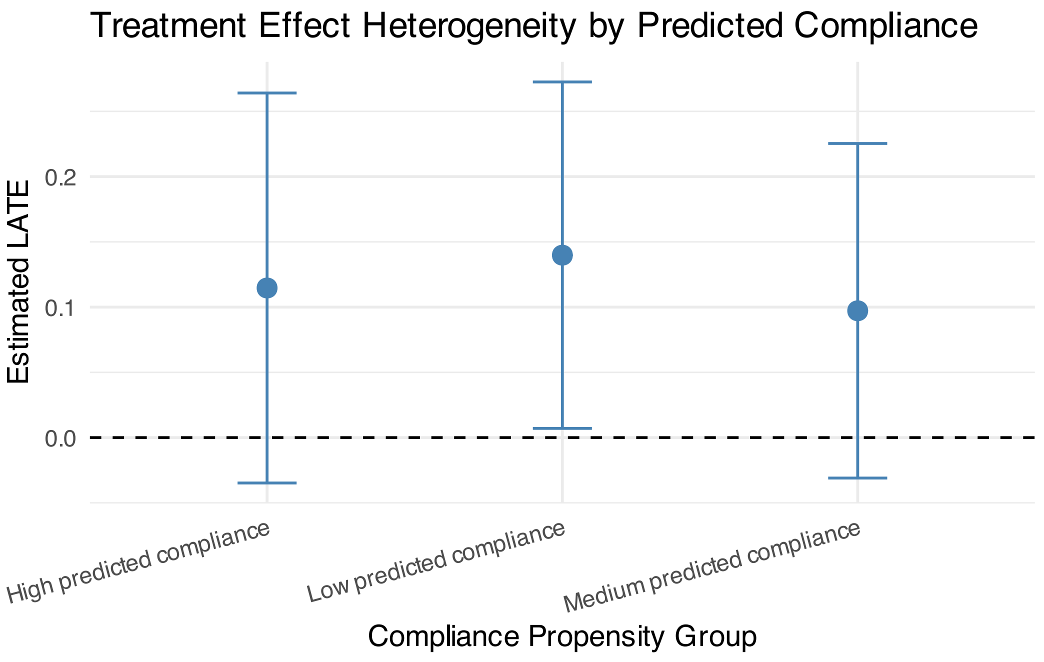
The AR confidence interval is the set of β values for which the AR statistic falls below the critical value. This interval is valid regardless of first-stage strength.

A.3 Heterogeneous Treatment Effects by Compliance Status

We can explore whether the treatment effect varies across the distribution of propensity to comply. Using principal stratification concepts:

```r
# Estimate conditional LATE by predicted compliance probability

hajj_df <- hajj_df |>
  mutate(
    p_comply = predict(glm(hajj2006 ~ age + literate + urban,
                            family = binomial(), data = hajj_df),
                      type = "response")
  )

hajj_df <- hajj_df |>
  mutate(
    comply_tertile = ntile(p_comply, 3),
    comply_group = case_when(
      comply_tertile == 1 ~ "Low predicted compliance",
      comply_tertile == 2 ~ "Medium predicted compliance",
      comply_tertile == 3 ~ "High predicted compliance"
    )
  )
```

```r
# Estimate LATE within each group
late_by_compliance <- hajj_df |>
  group_by(comply_group) |>
  group_modify(~ {
    fit <- iv_robust(moderacy ~ hajj2006 | success, data = .x)
    tibble(
      late = coef(fit)["hajj2006"],
      se = fit$std.error["hajj2006"],
      n = nrow(.x)
    )
  }) |>
  ungroup() |>
  mutate(
    ci_low = late - 1.96 * se,
    ci_high = late + 1.96 * se
  )

ggplot(late_by_compliance, aes(x = comply_group, y = late)) +
  geom_point(size = 3, color = "steelblue") +
  geom_errorbar(aes(ymin = ci_low, ymax = ci_high), width = 0.2, color =
"steelblue") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(
    title = "Treatment Effect Heterogeneity by Predicted Compliance",
    x = "Compliance Propensity Group",
    y = "Estimated LATE"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 15, hjust = 1))
```



A.4 Comparison of Estimators: OLS, IV, and Bounds

```
# Compare different estimators
comparison_table <- tibble(
  Estimator = c(
    "Naive OLS (D on Y)",
    "OLS with Controls",
    "ITT (Reduced Form)",
    "2SLS (LATE)",
    "2SLS with Controls"
  ),
  `Point Estimate` = c(
    coef(lm_robust(moderacy ~ hajj2006, data = hajj_df))["hajj2006"],
    coef(lm_robust(moderacy ~ hajj2006 + literate + urban + age, data =
hajj_df))["hajj2006"],
    coef(lm_robust(moderacy ~ success, data = hajj_df))["success"],
    coef(iv_robust(moderacy ~ hajj2006 | success, data = hajj_df))
["hajj2006"],
    coef(iv_robust(moderacy ~ hajj2006 + literate + urban | success + literate
+ urban, data = hajj_df))["hajj2006"]
  ),
  `Std. Error` = c(
    lm_robust(moderacy ~ hajj2006, data = hajj_df)$std.error["hajj2006"],
    lm_robust(moderacy ~ hajj2006 + literate + urban + age, data =
hajj_df)$std.error["hajj2006"],
    lm_robust(moderacy ~ success, data = hajj_df)$std.error["success"],
    iv_robust(moderacy ~ hajj2006 | success, data =
```

```r
hajj_df)$std.error["hajj2006"],
    iv_robust(moderacy ~ hajj2006 + literate + urban | success + literate +
urban, data = hajj_df)$std.error["hajj2006"]
  ),
  Interpretation = c(
    "Biased by selection",
    "Still biased if unobserved confounders",
    "Causal, but diluted by non-compliance",
    "Causal for compliers (LATE)",
    "LATE with precision gains"
  )
)

comparison_table |>
  gt() |>
  tab_header(
    title = "Comparison of Estimators for Hajj Effect",
    subtitle = "From Naive OLS to Proper Causal Identification"
  ) |>
  fmt_number(columns = c(`Point Estimate`, `Std. Error`), decimals = 4)
```

Comparison of Estimators for Hajj Effect

From Naive OLS to Proper Causal Identification

Estimator	Point Estimate	Std. Error	Interpretation
Naive OLS (D on Y)	0.1095	0.0347	Biased by selection
OLS with Controls	0.0991	0.0345	Still biased if unobserved confounders
ITT (Reduced Form)	0.1065	0.0345	Causal, but diluted by non-compliance
2SLS (LATE)	0.1247	0.0404	Causal for compliers (LATE)
2SLS with Controls	0.1227	0.0401	LATE with precision gains

A.5 Bootstrap Inference for Finite-Sample Validity

Standard errors rely on asymptotic theory. With our sample sizes, bootstrap inference provides a robustness check:

```r
set.seed(773)

bootstrap_iv <- function(data, n_boot = 1000) {
  boot_estimates <- map_dbl(1:n_boot, function(i) {
    boot_data <- data[sample(nrow(data), replace = TRUE), ]
    fit <- iv_robust(moderacy ~ hajj2006 | success, data = boot_data)
    coef(fit)["hajj2006"]
  })

  tibble(
    mean = mean(boot_estimates),
    se_boot = sd(boot_estimates),
    ci_low_percentile = quantile(boot_estimates, 0.025),
    ci_high_percentile = quantile(boot_estimates, 0.975),
    ci_low_normal = mean - 1.96 * sd(boot_estimates),
    ci_high_normal = mean + 1.96 * sd(boot_estimates)
  )
}

boot_results <- bootstrap_iv(hajj_df, n_boot = 1000)
print(boot_results)
```

```
# A tibble: 1 × 6
   mean se_boot ci_low_percentile ci_high_percentile ci_low_normal
  <dbl>   <dbl>             <dbl>              <dbl>         <dbl>
1 0.125  0.0377            0.0500              0.205        0.0507
# ℹ 1 more variable: ci_high_normal <dbl>
```

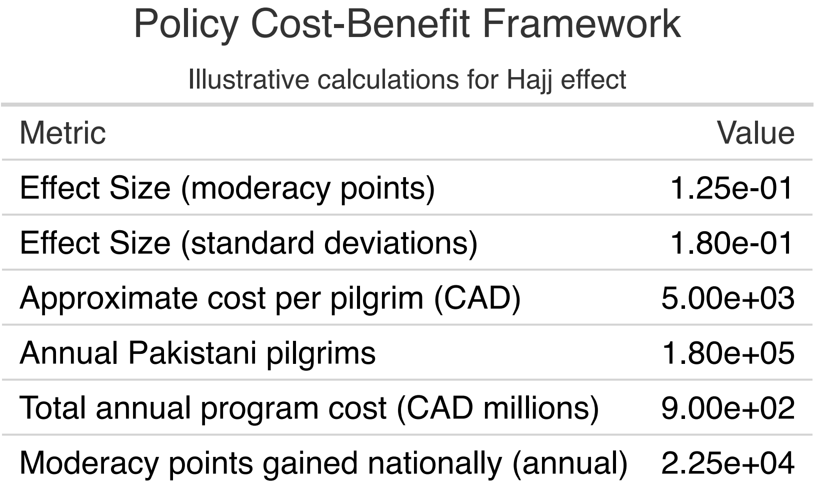
Interpretation: Bootstrap confidence intervals are similar to the analytical ones from `iv_robust`, suggesting our asymptotic inference is reliable.

A.6 Policy Implications and Cost-Benefit Framework

Beyond statistical significance, what are the practical implications?

Hajj Policy Analysis

```
# Back-of-envelope cost-benefit calculation
hajj_effect <- 0.125  # LATE in moderacy units
moderacy_sd <- sd(hajj_df$moderacy)
effect_in_sd <- hajj_effect / moderacy_sd

# Rough estimates (illustrative)
hajj_cost_per_person <- 5000  # CAD estimate
n_pakistani_hajj <- 180000    # Annual Pakistani Hajj pilgrims (approximate)

policy_table <- tibble(
  Metric = c(
    "Effect Size (moderacy points)",
    "Effect Size (standard deviations)",
    "Approximate cost per pilgrim (CAD)",
    "Annual Pakistani pilgrims",
    "Total annual program cost (CAD millions)",
    "Moderacy points gained nationally (annual)"
  ),
  Value = c(
    round(hajj_effect, 3),
    round(effect_in_sd, 3),
    hajj_cost_per_person,
    n_pakistani_hajj,
    round(hajj_cost_per_person * n_pakistani_hajj / 1e6, 1),
    round(hajj_effect * n_pakistani_hajj, 0)
  )
)

policy_table |>
  gt() |>
  tab_header(
    title = "Policy Cost-Benefit Framework",
    subtitle = "Illustrative calculations for Hajj effect"
  )
```



Policy Cost-Benefit Framework

Illustrative calculations for Hajj effect

Metric	Value
Effect Size (moderacy points)	1.25e-01
Effect Size (standard deviations)	1.80e-01
Approximate cost per pilgrim (CAD)	5.00e+03
Annual Pakistani pilgrims	1.80e+05
Total annual program cost (CAD millions)	9.00e+02
Moderacy points gained nationally (annual)	2.25e+04

Sesame Street Policy Analysis

The effect for young boys (LATE ≈ 28 points) represents a massive educational intervention. Unlike Hajj, Sesame Street is:

- **Highly scalable**: Free to broadcast, minimal marginal cost per viewer
- **Low cost**: Estimated cost of $0.10-$1.00 per viewing child
- **Large effect**: 28-point gain represents substantial early literacy improvement

This cost-effectiveness explains the long-term success and global expansion of Sesame Street.

A.7 Replication Code Summary

For reproducibility, we provide self-contained code to replicate our key findings:

```r
# Load packages
library(tidyverse)
library(estimatr)
library(gtsummary)
library(haven)
library(gt)

load("Assignment 3/hajjdata.Rdata")
hajj_df <- as_tibble(hajj)

# 2SLS estimate
iv_robust(moderacy ~ hajj2006 | success, data = hajj_df)
```

```
             Estimate Std. Error   t value    Pr(>|t|)   CI Lower  CI Upper
(Intercept) 1.5628752 0.02874802 54.364611 0.000000000 1.50648749 1.6192628
hajj2006    0.1246954 0.04035234  3.090166 0.002034851 0.04554655 0.2038443
              DF
(Intercept) 1603
hajj2006    1603
```

```r
# LATE = 0.125, SE = 0.040, p = 0.002

# Sessame street
sesame_df <- read_dta("Assignment 3/sesame_experiment.dta") |>
  filter(age < 51, female == 0)

# 2SLS
iv_robust(letters ~ watched | encouraged, data = sesame_df)
```

```
             Estimate Std. Error   t value    Pr(>|t|)  CI Lower CI Upper DF
(Intercept)  1.915423   8.783169 0.2180788 0.828332296 -15.76418 19.59503 46
watched     28.283582  10.445441 2.7077442 0.009478111   7.25800 49.30916 46
```

For troubleshooting: do not edit or remove

```
sysname

"Darwin"

release

"25.2.0"

version
"Darwin Kernel Version 25.2.0: Tue Nov 18 21:09:49 PST 2025;
root:xnu-12377.61.12~1/RELEASE_ARM64_T8142"

nodename

"tadhgs-M5-MacBook-Pro.local"

machine

"arm64"

login

"root"

user

"tadhg"

effective_user

"tadhg"
```

```
[1] "2026-02-11 10:46:01 EST"
```