quant22

2026-02-18

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(broom)
library(modelsummary)
library(sandwich)

if (!dir.exists("outputs")) dir.create("outputs")

#2.1

star <- read_csv("C:/Users/Usuario/Downloads/star.csv")
```

```
## Rows: 6325 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (6): race, classtype, yearssmall, hsgrad, g4math, g4reading
##
## 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.
```

```
star <- star %>%
  mutate(
    classtype_f = factor(
      classtype,
      levels = c(1, 2, 3),
      labels = c("Small", "Regular", "Regular+Aide")
    ),
    race_f = factor(
      race,
      levels = c(1, 2, 3, 4, 5, 6),
      labels = c("White", "Black", "Asian", "Hispanic", "Native American", "Other")
    ),
    small = if_else(classtype_f == "Small", 1, 0)
  )
```

```r
star %>%
  summarise(
    n_total = n(),
    n_nonmissing_g4reading = sum(!is.na(g4reading)),
    n_nonmissing_g4math = sum(!is.na(g4math))
  )
```

```
## # A tibble: 1 x 3
##   n_total n_nonmissing_g4reading n_nonmissing_g4math
##     <int>                  <int>               <int>
## 1    6325                   2353                2395
```

```r
#2.2

reading_means <- star %>%
  group_by(classtype_f) %>%
  summarise(
    mean_g4reading = mean(g4reading, na.rm = TRUE),
    n = sum(!is.na(g4reading)),
    .groups = "drop"
  )

reading_means
```

```
## # A tibble: 3 x 3
##   classtype_f  mean_g4reading     n
##   <fct>                <dbl> <int>
## 1 Small                 723.   726
## 2 Regular               720.   836
## 3 Regular+Aide          721.   791
```

```r
# The highest mean is that of those students in small classes

m_read_biv <- star %>%
  lm(g4reading ~ small, data = .)

m_read_biv %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   720.        1.30    554.     0
## 2 small           3.10      2.34      1.32   0.185
```

```r
# According to this coefficient, students who are in small classes score on average 3.1 points higher t

diff_means_reading <- star %>%
  group_by(small) %>%
  summarise(mean_read = mean(g4reading, na.rm = TRUE), .groups = "drop") %>%
  summarise(diff = mean_read[small == 1] - mean_read[small == 0]) %>%
  pull(diff)

coef_reg_reading <- m_read_biv %>%
  coef() %>%
  .["small"] %>%
  unname()
```

```
tibble(
  diff_means = diff_means_reading,
  regression_coef = coef_reg_reading
)
```

```
## # A tibble: 1 x 2
##   diff_means regression_coef
##        <dbl>           <dbl>
## 1       3.10            3.10
```

```
m_math_biv <- star %>%
  lm(g4math ~ small, data = .)

m_math_biv %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  709.         1.06   669.      0
## 2 small          0.591      1.91     0.310   0.756
```

```
# The pattern for math goes in the same direction and is still not statistically significant, but the c

#2.3

m_read_controls <- star %>%
  lm(g4reading ~ small + race_f + yearssmall, data = .)

m_read_controls %>% tidy()
```

```
## # A tibble: 7 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)      724.        1.40   517.    0
## 2 small             -4.00      4.98    -0.804 4.22e- 1
## 3 race_fBlack      -33.8       2.95   -11.4   1.80e-29
## 4 race_fAsian       14.8      19.3      0.767 4.43e- 1
## 5 race_fHispanic     8.43     36.1      0.234 8.15e- 1
## 6 race_fOther       80.3      36.1      2.23  2.62e- 2
## 7 yearssmall         2.17      1.29     1.68  9.33e- 2
```

```
coef_compare <- tibble(
  model = c("Bivariate", "Controls"),
  coef_small = c(
    coef(m_read_biv)["small"],
    coef(m_read_controls)["small"]
  )
)

coef_compare
```

```
## # A tibble: 2 x 2
##   model     coef_small
##   <chr>          <dbl>
## 1 Bivariate       3.10
## 2 Controls       -4.00
```

```
# Adding controls changes the coefficient for small significantly and alters its direction, thus indica

#2.4

m_read_interact <- star %>%
  lm(g4reading ~ small * race_f + yearssmall, data = .)

m_read_interact %>% tidy()
```

```
## # A tibble: 11 x 5
##    term                 estimate std.error statistic   p.value
##    <chr>                   <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)            725.       1.43   507.     0
##  2 small                   -5.32     5.12    -1.04   2.99e- 1
##  3 race_fBlack            -36.0      3.59   -10.0    3.40e-23
##  4 race_fAsian             21.3     20.9      1.02   3.07e- 1
##  5 race_fHispanic           9.14    36.1      0.253  8.00e- 1
##  6 race_fOther             53.3     51.0      1.05   2.96e- 1
##  7 yearssmall               2.25     1.29     1.74   8.25e- 2
##  8 small:race_fBlack        6.97     6.33     1.10   2.71e- 1
##  9 small:race_fAsian      -46.7     55.1     -0.847  3.97e- 1
## 10 small:race_fHispanic    NA       NA       NA     NA
## 11 small:race_fOther       54.3     72.2      0.753  4.52e- 1
```

```
effects_by_race <- tibble(
  effect_white = coef(m_read_interact)["small"],
  effect_black = coef(m_read_interact)["small"] +
    coef(m_read_interact)["small:race_fBlack"]
)

effects_by_race
```

```
## # A tibble: 1 x 2
##   effect_white effect_black
##          <dbl>        <dbl>
## 1        -5.32         1.66
```

```
models_reading <- list(
  "Bivariate" = m_read_biv,
  "Controls" = m_read_controls,
  "Interaction" = m_read_interact
)

# Although, serving the white students as the base category, the direction of the coefficient changes f

# 2.5

modelsummary(models_reading, vcov = "robust")
```

```
## Warning in meatHC(x, type = type, omega = omega): HC3 covariances are
## numerically unstable for hat values close to 1 (and undefined if exactly 1) as
## for observation(s) 947, 5939, 6049
```

```
## Warning: Model matrix is rank deficient. Some variance-covariance parameters are
##   missing.
```

```
## Model matrix is rank deficient. Parameters `small:race_fHispanic` were
##   not estimable.
```

```
modelsummary(
  models_reading,
  vcov = "robust",
  output = "outputs/star_reading_models.html"
)
```

```
## Warning in meatHC(x, type = type, omega = omega): HC3 covariances are numerically unstable for hat va
## Warning in meatHC(x, type = type, omega = omega): Model matrix is rank deficient. Some variance-covar
##   missing.
```

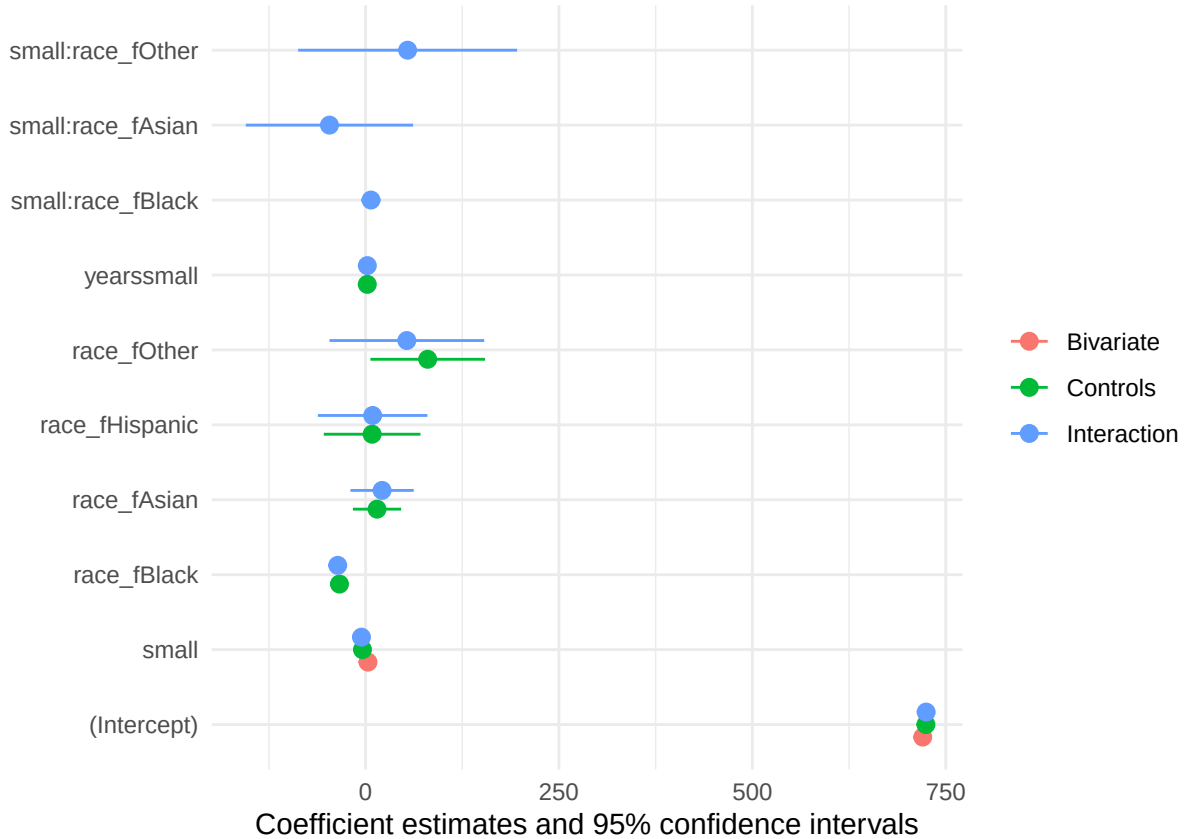
```
## Model matrix is rank deficient. Parameters `small:race_fHispanic` were
##   not estimable.
```

```
p <- modelplot(models_reading, vcov = "robust")
```

```
## Warning in meatHC(x, type = type, omega = omega): HC3 covariances are numerically unstable for hat va
## Warning in meatHC(x, type = type, omega = omega): Model matrix is rank deficient. Some variance-covar
##   missing.
```

```
## Model matrix is rank deficient. Parameters `small:race_fHispanic` were
##   not estimable.
```

```
p
```




```
ggsave(
  "outputs/star_reading_coeffplot.png",
  plot = p,
  width = 10,
```

	Bivariate	Controls	Interaction
(Intercept)	720.291	724.386	724.680
	(1.309)	(1.361)	(1.428)
small	3.100	−4.000	−5.318
	(2.319)	(5.175)	(5.121)
race_fBlack		−33.758	−36.010
		(3.077)	(3.591)
race_fAsian		14.803	21.320
		(15.898)	(20.866)
race_fHispanic		8.433	9.140
		(31.883)	(36.117)
race_fOther		80.274	53.320
		(37.764)	(51.011)
yearssmall		2.170	2.249
		(1.359)	(1.295)
small × race_fBlack			6.974
			(6.328)
small × race_fAsian			−46.680
			(55.137)
small × race_fOther			54.320
			(72.159)
Num.Obs.	2353	2353	2353
R2	0.001	0.057	0.058
R2 Adj.	0.000	0.054	0.054
AIC	25 313.7	25 188.4	25 191.9
BIC	25 331.0	25 234.5	25 255.3
Log.Lik.	−12 653.855	−12 586.194	−12 584.934
RMSE	52.40	50.91	50.88
Std.Errors	HC3	HC3	HC3

```
  height = 6
)

# 2.6

# The evidence from the STAR database is more credible than that coming from an observational study bec
```