Experimento STAR

Javier Herrero Pérez

2026-02-02

Introducción

El Proyecto STAR fue un experimento aleatorio controlado (RCT) diseñado para responder una pregunta fundamental de política pública: ¿Realmente importa el tamaño de la clase para el aprendizaje? En un contexto observacional, esta pregunta es difícil de responder porque las clases pequeñas suelen estar en distritos ricos o con padres muy implicados (confundidores). Para evitar esto, el proyecto STAR asignó aleatoriamente a más de 7,000 estudiantes y sus profesores a tres tipos de grupos: clases pequeñas (13-17 alumnos), clases regulares (22-25 alumnos) y clases regulares con un asistente.

El objetivo de esta actividad es demostrar que, mientras los datos se mantengan bajo el diseño experimental (aleatorios), la diferencia de medias nos da el Efecto Causal. Sin embargo, en cuanto introducimos un criterio de selección (como favorecer a los alumnos con peores notas), rompemos la aleatoriedad y entramos en el terreno de los datos observacionales, donde las conclusiones pueden ser totalmente erróneas si no se corrigen adecuadamente.

Los datos

```
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(AER)
```

```
## Warning: package 'AER' was built under R version 4.5.2
```

```
## Cargando paquete requerido: car
```

```
## Cargando paquete requerido: carData
```

```
##
## Adjuntando el paquete: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## Cargando paquete requerido: lmtest

## Cargando paquete requerido: zoo

##
## Adjuntando el paquete: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Cargando paquete requerido: sandwich

## Cargando paquete requerido: survival
```

```
data(STAR)
help(STAR)
```

```
## starting httpd help server ...

##  done
```

```
# Convertir a formato long
nam <- c("star", "read", "math", "lunch", "school", "degree", "ladder",
         "experience", "tethnicity", "system", "schoolid")
lev <- c("k", "1", "2", "3")

star <- reshape(STAR, idvar = "id", ids = row.names(STAR),
                times = lev, timevar = "grade", direction = "long",
                varying = lapply(nam, function(x) paste(x, lev, sep = "")))

# Mejorar nombres y tipos
names(star)[5:15] <- nam
star$id <- factor(star$id)
star$grade <- factor(star$grade, levels = lev,
                     labels = c("kindergarten", "1  st", "2nd", "3rd"))
```

Estimación insesgada del efecto causal medio

La nota para lectura:

```r
# Obtenemos el efecto causal para los estudiantes de grade=3rd

library(broom)
lectura <- lm(read ~ star, data = star %>% filter(grade=='3rd'))

tidy(lectura)
```

```
## # A tibble: 3 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       613.        0.900   681.    0
## 2 starsmall           8.53      1.26      6.79  1.23e-11
## 3 starregular+aide    0.387     1.21      0.320 7.49e- 1
```

```r
confint(lectura,2)
```

```
##              2.5 %   97.5 %
## starsmall 6.069317 10.99631
```

La nota para matemáticas:

```r
matematicas <- lm(math ~ star, data = star %>% filter(grade=='3rd'))
tidy(matematicas)
```

```
## # A tibble: 3 x 5
##   term             estimate std.error statistic     p.value
##   <chr>               <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)       616.        0.926   665.    0
## 2 starsmall           6.88      1.29      5.32  0.000000107
## 3 starregular+aide   -0.804     1.24     -0.647 0.518
```

```r
confint(matematicas,2)
```

```
##              2.5 %   97.5 %
## starsmall 4.344301 9.413124
```

Preguntas

- **Hemos calculado la diferencia de medias en las puntuaciones de lectura entre alumnos en clases pequeñas y regulares. ¿Por qué podemos interpretar esta diferencia como el efecto causal medio del tamaño de clase sobre el rendimiento académico?¿Cuál es la diferencia promedio entre clases pequeñas y normales?**

Porque gracias a la aleatorización, el grupo de "clases pequeñas" y "regulares" son estadísticamente idénticos en todo (potencial académico, nivel socioeconómico, etc.) antes de empezar el estudio. La única diferencia sistemática entre ellos es el tamaño de la clase. Por tanto, cualquier diferencia posterior en las notas (Y) debe haber sido causada por el tratamiento (D).

Comparando la variable `starsmall`, para lectura la clase pequeña saca 8.53 más que en la clase grande, mientras que en matemáticas es 6.87

- **Es posible añadir covariables (como género o etnicidad) en la regresión. Reformula la estimación incluyendo covariables en la expresión (busca en la ayuda de la librería AER). ¿Por qué incluir covariables no cambia la validez causal del estimador, pero puede hacerlo más eficiente (reducir su varianza)? Prueba la estimación del efecto para los distintos grados (kindergarten, 1st, 2nd, 3rd)**

En un experimento ideal, las covariables están equilibradas. Añadirlas no cambia el valor esperado del coeficiente porque no están correlacionadas con el tratamiento ya que la asignación fue al azar.

Al incluir variables que explican parte de la varianza de las notas (como el género), el "ruido" del modelo disminuye (baja el R^2 no explicado). Esto reduce el error estándar del estimador, haciendo que tus intervalos de confianza sean más estrechos.

Introducción de sesgo de selección

Creación de la variable `rendimiento previo`

Vamos a modificar los datos para convertirlos en observacionales. Para ello vamos a simular un tipo de mecanismo de asignación al tratamiento que podría existir en el mundo real.

Mecanismo no aleatorio de asignación del tratamiento: aplicamos el tratamiento (grupos reducidos) a los estudiantes cuyo rendimiento haya estado en el curso anterior por debajo del tercer decil (30%) de la distribución de notas.

Para ello vamos a crear la variable rendimiento previo: queremos medir el rendimiento antes del tratamiento en cada grado.

```
library(dplyr)

star <- star %>%
  arrange(id, grade) %>%
  group_by(id) %>%
  mutate(
    read_lag = lag(read),
    math_lag = lag(math)
  ) %>%
  ungroup()
```

Preguntas

- **¿Para qué grado read_lag es siempre NA?**

Para Kindergarten.

- **¿Por qué tiene sentido desde el punto de vista temporal?**

Es el primer año de escolarización registrado en el estudio; no existen datos de un "año anterior" dentro del experimento.

Definimos alumnos con bajo rendimiento previo

Definimos como bajo rendimiento estar por debajo del tercer decil (30%) en la nota de lectura o matemáticas del grado anterior.

```
star <- star %>%
  group_by(grade) %>%
  mutate(
    q30_read = quantile(read_lag, 0.3, na.rm = TRUE),
    q30_math = quantile(math_lag, 0.3, na.rm = TRUE),
    bottom30 = (read_lag <= q30_read) | (math_lag <= q30_math)
  ) %>%
  ungroup()
```

Pregunta

- **¿Por qué calculamos los cuantiles por grado y no globalmente?**

Porque la distribución de notas cambia con la edad. El 30% inferior de un niño de 5 años no es el mismo nivel de competencia que el 30% de uno de 9 años. Hacerlo por grado asegura que estamos seleccionando a los "rezagados" respecto a sus pares actuales.

Filtrado de los datos

Conservamos solo a los alumnos tratados (small) que están por debajo del 30% de la distribución. El resto de alumnos no tratados se mantienen todos.

```
star_sel <- star %>%
  filter(
    !(star == "small" & !bottom30)
  )
```

Observa que no se redefine el tratamiento; tan solo se condiciona la muestra.

Pregunta

- **¿Qué impacto tiene este filtro sobre el conjunto de tratados y no tratados?**

Se está creando un escenario donde solo los alumnos con dificultades acceden al tratamiento (clase pequeña). Los alumnos "buenos" solo están en el grupo control. En el grupo small, desaparecerán todos los que superen el percentil 30 (un 70% aproximadamente). En el grupo control no se elimina a nadie.

¿A quién estamos eliminando?

```
star %>%
  mutate(retained = !(star == "small" & !bottom30)) %>%
  group_by(star) %>%
  summarise(prop_retained = mean(retained, na.rm=TRUE))
```

```
## # A tibble: 4 x 2
##   star         prop_retained
##   <fct>                <dbl>
## 1 regular              1
## 2 small                0.318
## 3 regular+aide         1
## 4 <NA>                 1
```

Pregunta

- **¿Qué proporción de alumnos tratados desaparece?**
- **¿Se elimina algún alumno del grupo control (no tratados)?**

Efectos de la intervención en la muestra seleccionada

A continuación comparamos el rendimiento previo entre tratados y no tratados en la muestra filtrada. Para lectura:

```
lectura <- lm(read ~ star, data = star_sel |> filter(grade=='3rd'))
lectura |> tidy()
```

```
## # A tibble: 3 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       613.        0.869   705.    0
## 2 starsmall         -21.2       1.90    -11.2   1.41e-28
## 3 starregular+aide    0.387     1.17      0.332 7.40e- 1
```

```
lectura |> confint(2)
```

```
##               2.5 %    97.5 %
## starsmall -24.88603 -17.45184
```

Para matemáticas:

```
matematicas <- lm(math ~ star, data = star_sel |> filter(grade=='3rd'))
matematicas |> tidy()
```

```
## # A tibble: 3 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       616.        0.900   685.    0
## 2 starsmall         -24.1       1.97    -12.2   7.69e-34
## 3 starregular+aide   -0.804     1.21     -0.666 5.05e- 1
```

```
matematicas |> confint(2)
```

```
##               2.5 %    97.5 %
## starsmall -27.98652 -20.25027
```

Preguntas

- **¿Significa la estimación que el efecto no es beneficioso para el grupo considerado?**

No, el efecto sigue siendo positivo, pero la estimación ahora está "contaminada". Al comparar "alumnos con dificultades en clases pequeñas" contra "todos los alumnos (incluidos los brillantes) en clases normales", es muy probable que el coeficiente de star salga negativo o muy bajo. El sesgo de selección está ocultando el beneficio real.

- **¿Son comparables los grupos antes del tratamiento?**

No. Se ha roto la propiedad de ignorabilidad. Ahora el grupo de tratamiento es intrínsecamente "peor" académicamente que el grupo control antes de empezar el análisis.

Intento de corrección con controles

Añadimos controles por rendimiento previo, perfil socioeconómico y características del profesor:

```
lectura <- lm(read ~ star + read_lag + math_lag + lunch + gender + ethnicity + degree + experience + te
              data = star_sel |> filter(grade=='3rd'))
lectura |> tidy()
```

```
## # A tibble: 15 x 5
##    term              estimate std.error statistic   p.value
##    <chr>                <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)       193.        7.28      26.5   3.10e-141
##  2 starsmall           1.32      1.33       0.992 3.21e-  1
##  3 starregular+aide   -0.521     0.877     -0.595 5.52e-  1
##  4 read_lag            0.570     0.0135    42.4   5.16e-317
##  5 math_lag            0.147     0.0135    10.9   3.84e- 27
##  6 lunchfree          -3.82      0.938     -4.08  4.67e-  5
##  7 genderfemale        3.26      0.812      4.02  6.04e-  5
##  8 ethnicityafam       2.16      1.11       1.95  5.13e-  2
##  9 ethnicityasian     -1.63      8.38      -0.194 8.46e-  1
## 10 ethnicityhispanic  -3.66     23.6       -0.155 8.77e-  1
## 11 ethnicityother      8.81      9.65       0.913 3.61e-  1
## 12 degreemaster        1.71      0.824      2.07  3.84e-  2
## 13 degreespecialist    3.13      4.17       0.751 4.53e-  1
## 14 experience         -0.0159    0.0478    -0.333 7.39e-  1
## 15 tethnicityafam      1.31      1.16       1.13  2.59e-  1
```

```
lectura |> confint(2)
```

```
##               2.5 %   97.5 %
## starsmall -1.284832 3.917566
```

Preguntas

- ¿Se corrige el sesgo?

Discusión final

Pregunta

- **¿Cuál es el origen fundamental del sesgo en esta actividad?**
  - Mala especificación del modelo
  - Variables omitidas
  - **Problema de selección muestral**: Se ha forzado una correlación entre el tratamiento y la capacidad del alumno.
  - Falta de tamaño muestral
- **En el experimento STAR original, la asignación al tratamiento es aleatoria. Sin embargo, tras aplicar el filtro de la actividad, los datos resultantes se comportan como observacionales. Explica por qué, indicando qué propiedad clave de los datos experimentales se ha perdido.**

Se ha perdido la Independencia (o Aleatoriedad). En los datos originales, tras el filtro, el tratamiento D depende de $Y(t-1)$, que está altamente correlacionado con los resultados potenciales. Los datos ahora son observacionales porque el "mecanismo de asignación" ya no es una moneda al aire, sino una regla basada en el rendimiento.

Extensiones

Repetir el análisis por grado e incluyendo covariables

```
# Extensiones: Recuperando el efecto causal

library(purrr)
```

```
##
## Adjuntando el paquete: 'purrr'

## The following object is masked from 'package:car':
##
##     some
```

```
# Función para ejecutar modelos y extraer el coeficiente de 'small'
estudiar_grado <- function(g) {
  df_sub <- star_sel %>% filter(grade == g)

  # Modelo 1: Sesgado (sin controles, solo el filtro aplicado)
  m1 <- lm(read ~ star, data = df_sub)

  # Modelo 2: Corregido (añadimos la variable que causó el sesgo: rendimiento previo)
  # También añadimos controles socioeconómicos para ganar precisión
  m2 <- lm(read ~ star + read_lag + math_lag + lunch + experience, data = df_sub)

  # Extraemos resultados
  bind_rows(
```

```
    tidy(m1) %>% filter(term == "starsmall") %>% mutate(modelo = "Sesgado", grado = g),
    tidy(m2) %>% filter(term == "starsmall") %>% mutate(modelo = "Corregido", grado = g)
  )
}

# Aplicamos a los grados donde tenemos lag (1st, 2nd, 3rd)
grados_analisis <- c("1  st", "2nd", "3rd")
resultados_ext <- map_df(grados_analisis, estudiar_grado)

# Visualizamos los resultados
print(resultados_ext)
```

```
## # A tibble: 6 x 7
##   term      estimate std.error statistic  p.value modelo    grado
##   <chr>        <dbl>     <dbl>     <dbl>    <dbl> <chr>     <chr>
## 1 starsmall  -16.4        2.72    -6.03  1.71e- 9 Sesgado   1  st
## 2 starsmall    7.30       2.42     3.02  2.58e- 3 Corregido 1  st
## 3 starsmall  -27.6        2.38   -11.6   1.01e-30 Sesgado   2nd
## 4 starsmall    0.542      1.70     0.318 7.50e- 1 Corregido 2nd
## 5 starsmall  -21.2        1.90   -11.2   1.41e-28 Sesgado   3rd
## 6 starsmall    1.38       1.33     1.04  2.99e- 1 Corregido 3rd
```

Elabora una estrategia para recuperar el efecto causal en la muestra de datos observacionales Discutir paralelismos con missing not at random