Tarea 1 Simulación del efecto de la educación en los ingresos

Wilmer Rojas & John Esteban Londoño & William Aguirre

2024-08-30

Introducción

El objetivo de este documento es presentar la simulación de la relación causal de la educación sobre los ingresos de los individuos. Para ello partimos del siguiente proceso generador de datos:

$$Ingreso = \beta \ + \alpha \ educación + \epsilon$$

Donde α es el efecto que tiene la educación sobre el ingreso. Para efectos de la simulación, asumimos que el tamaño de este efecto es igual a 10. Adicionalmente, partimos del supuesto de que ϵ no está relacionado con la educación y por esta razón existe exogeneidad de la variable de interés.

Para evidenciar la diferencia en la estimación por el método de mínimos cuadrados ordinarios cuando existe exogeneidad y cuando se presenta una correlación entre la variable explicativa (educación) y los no observables (ϵ) se genera una segunda simulación con el siguiente procesos generador de datos:

$$Ingreso = \beta \ + \alpha \ educación + \epsilon$$

Donde

$$E[educacion|\epsilon] \neq 0$$

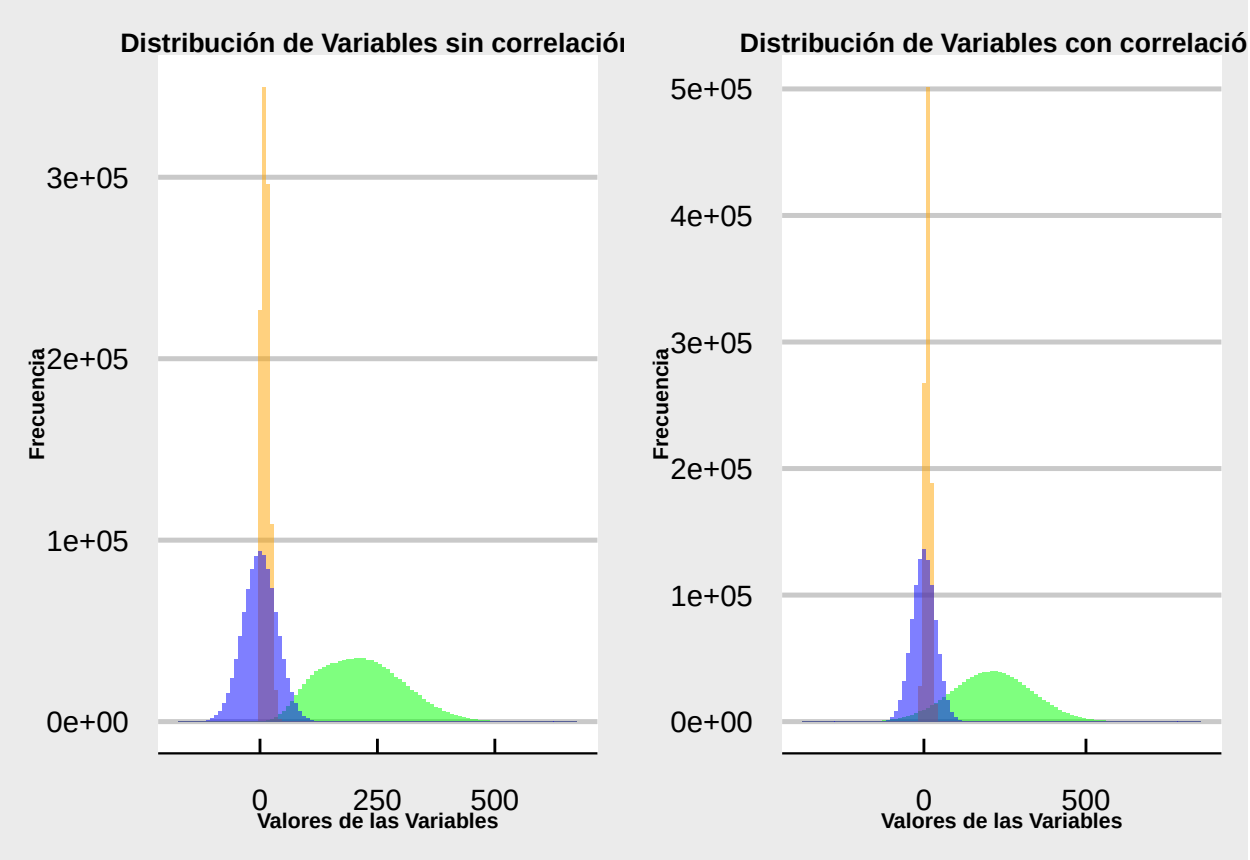
Al comparar las estimaciones por M.C.O. de los datos simulados con los dos procesos se logra evidenciar que el estimador es insesgado cuando se cumple el principio de exogeneidad. Sin embargo, la estimación cuando se viola este supuesto se aleja del parámetro poblacional debido al sesgo de selección.

Análisis de correlación entre el término de error y el tratamiento

```
N=1000000

income_aut = 100
effect=10

df <- tibble(
    educ = pmax(rnorm(N, mean=11, sd= 9),0),
    u = rnorm(N, mean=0, sd=36),
    income = income_aut + effect*educ + u) %>%
    mutate(niv_educ=case_when(educ <= 5 ~ 0,
                              educ >5 & educ <=9 ~ 5,
                              educ >9 & educ <=11 ~ 9,
                              educ >11 & educ <=16 ~ 11,
                              educ >16 & educ <=18 ~ 16,
```

```
                              educ >18 & educ <=20 ~ 18,
                              educ > 20 ~ 20),
           income = pmax(income, 0))


df_1<-rnorm_multi(n=1000000,vars = 2,mu=c(11,0),sd=c(9,36),r=0.9,varnames = c('educ','u'))
df_i <- tibble(income=income_aut + effect*df_1$educ + df_1$u )
df_1<-cbind(df_i,df_1)

df_1 <- df_1 %>%  mutate(niv_edu = case_when(educ <= 5 ~ 0,
                              educ >5 & educ <=9 ~ 5,
                              educ >9 & educ <=11 ~ 9,
                              educ >11 & educ <=16 ~ 11,
                              educ >16 & educ <=18 ~ 16,
                              educ >18 & educ <=20 ~ 18,
                              educ > 20 ~ 20))
```

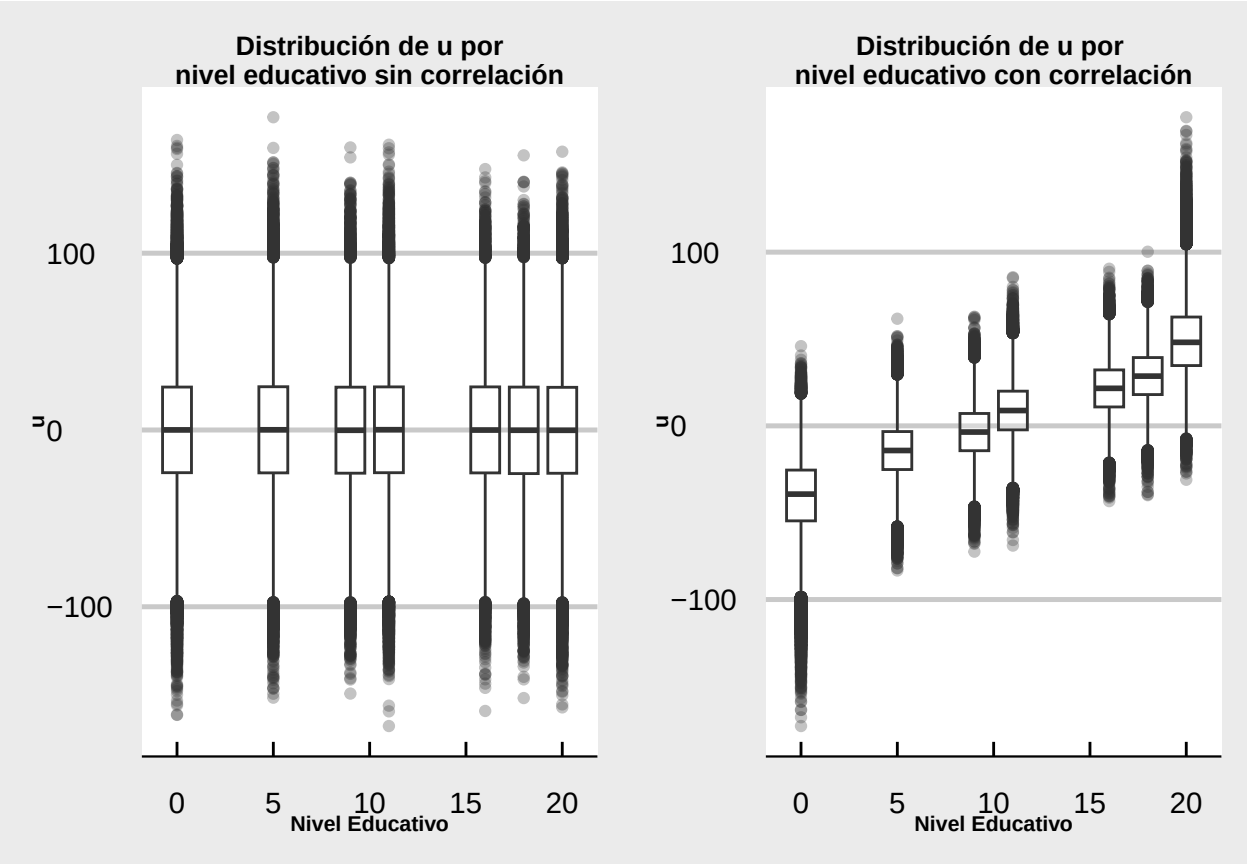
```
p1 <- ggplot(df) +
  geom_histogram(aes(income), bins = 100, fill = "green", alpha = 0.5) +
  geom_histogram(aes(educ), bins = 100,  fill = "orange", alpha = 0.5) +
  geom_histogram(aes(u), bins = 100, fill="blue", alpha=0.5)+
  theme_economist_white() +
  ggtitle("Distribución de Variables sin correlación") +
  labs(x = "Valores de las Variables", y = "Frecuencia") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5,
                              size=10),
    axis.title.x = element_text(face = "bold", size=8),
    axis.title.y = element_text(face = "bold",size = 8)
  )


p2 <- ggplot(df_1) +
  geom_histogram(aes(income), bins = 100, fill = "green", alpha = 0.5) +
  geom_histogram(aes(educ), bins = 100, fill = "orange", alpha = 0.5) +
  geom_histogram(aes(u), bins = 100,  fill="blue", alpha=0.5)+
  theme_economist_white() +
  ggtitle("Distribución de Variables con correlación") +
  labs(x = "Valores de las Variables", y = "Frecuencia") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5,
                              size = 10),
    axis.title.x = element_text(face = "bold", size = 8),
    axis.title.y = element_text(face = "bold",size = 8)
  )


grid.arrange(p1, p2, ncol = 2)
```



Dada la especificación del modelo, la distribución de las variables sigue una distribución normal en ambos escenarios , es decir la correlación entre las variable de tratamiento y el término de error (para este ejemplo, la habilidad) en la segunda simulación no afecta la distribución individual de la variable.

```
p1 <- ggplot(df, aes(x = u, y = niv_educ, group = niv_educ)) +
  geom_boxplot(alpha = 0.3) +
  coord_flip() +
  theme_economist_white() +
  ggtitle(paste0("Distribución de u por","\n","nivel educativo sin correlación")) +
  labs(x = "u", y = "Nivel Educativo") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5,
                              size = 10),
    axis.title.x = element_text(face = "bold", size = 8),
    axis.title.y = element_text(face = "bold", size = 8)
  )


p2 <- ggplot(df_1, aes(x = u, y = niv_edu, group = niv_edu)) +
  geom_boxplot(alpha = 0.3) +
  coord_flip() +
  theme_economist_white() +
  ggtitle(paste0("Distribución de u por ","\n","nivel educativo con correlación")) +
  labs(x = "u", y = "Nivel Educativo") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5,
```

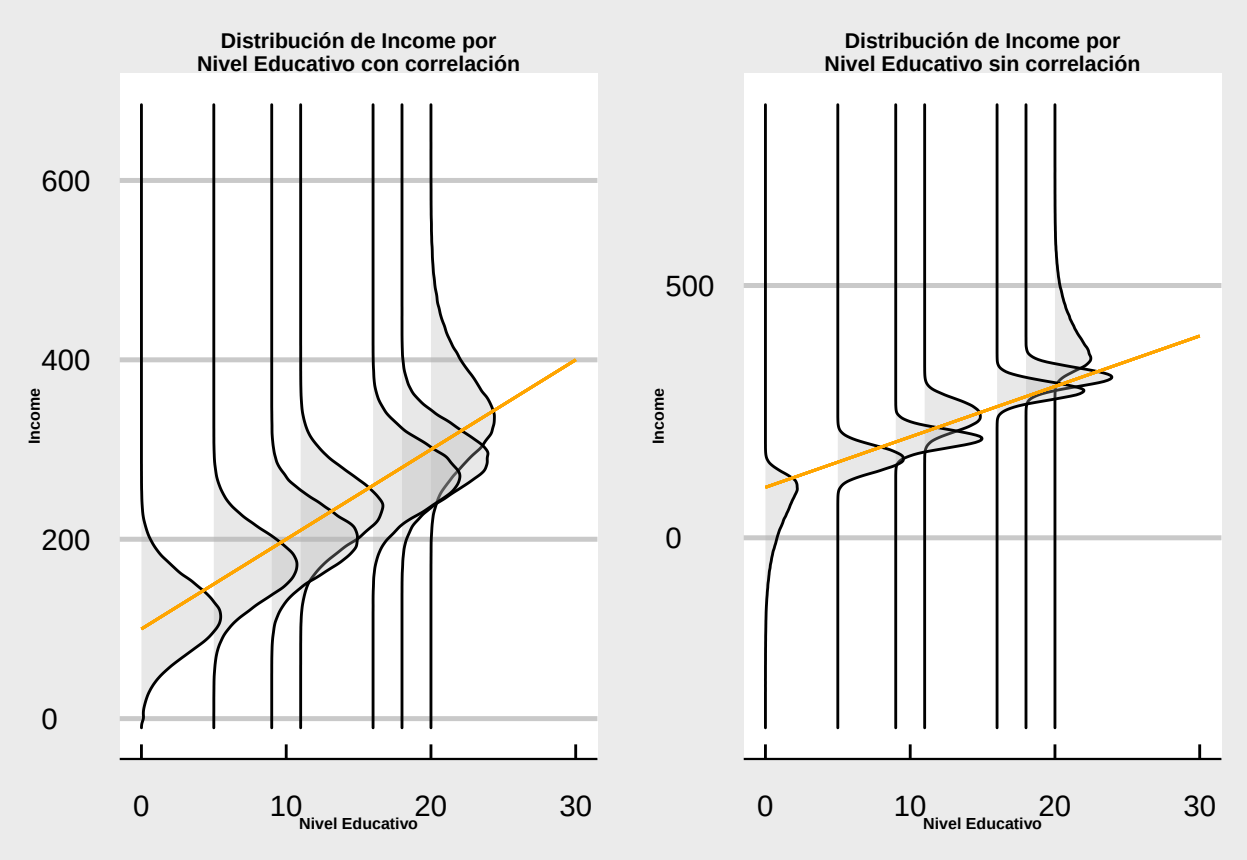
```
                              size = 10),
    axis.title.x = element_text(face = "bold", size=8),
    axis.title.y = element_text(face = "bold", size = 8)
  )


grid.arrange(p1, p2, ncol = 2)
```



Para el ejercicio se definió una correlación entre la variable años de educación y el termino de error de 0.9. Esto produce que el término de error varie de forma incremental a mayor número de años de educación, es decir, a mayor numero de años de educación se observa un incremento en la media del término de error (correlación positiva) y por ende un mayor ingreso económico, esto introduce un sesgo de selección que produce una sobreestimación del efecto

```
p1 <- ggplot(df, aes(x = income, y = niv_educ, group = niv_educ)) +
  geom_density_ridges(alpha = 0.3) +
  geom_function(fun = function(x) (-income_aut/effect) + (1/effect)*x, color = "orange",ylim=c(-300,600)
  coord_flip() +
  theme_economist_white() +
  ggtitle(paste0("Distribución de Income por","\n","Nivel Educativo con correlación")) +
  labs(x = "Income", y = "Nivel Educativo") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5,size =8),
    axis.title.x = element_text(face = "bold", size = 6),
    axis.title.y = element_text(face = "bold", size = 6)
  )
```

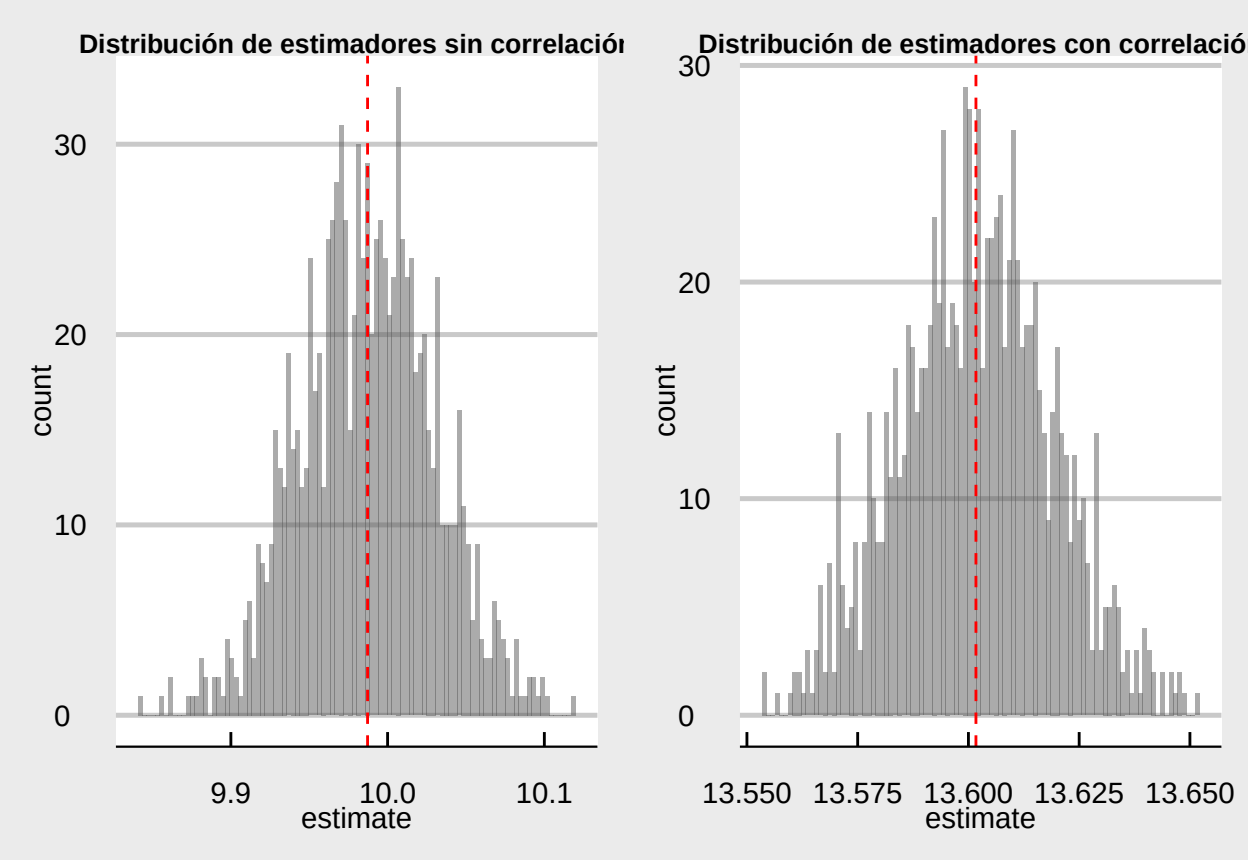
```
p2 <- ggplot(df_1, aes(x = income, y = niv_edu, group = niv_edu,)) +
  geom_density_ridges(alpha = 0.3) +
  geom_function(fun = function(x) (-income_aut/effect) + (1/effect)*x,
                color = "orange",
                ylim=c(-300,600),xlim = c(100, 400)) +
  coord_flip() +
  theme_economist_white() +
  ggtitle(paste0("Distribución de Income por","\n","Nivel Educativo sin correlación")) +
  labs(x = "Income", y = "Nivel Educativo") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5,size =8 ),
    axis.title.x = element_text(face = "bold",size = 6),
    axis.title.y = element_text(face = "bold", size = 6)
  )


grid.arrange(p1, p2, ncol = 2)
```



En ambos casos se observa una correlación positiva entre el nivel educativo y el ingreso. No obstante, en la simulación 2, se tiene un efecto mayor (una pendiente más pronunciada) entre el nivel educativo y el ingreso, esto es secundario al sesgo de selección que sobreestima el efecto del nivel educativo.

```
media_df <- mean(reg$estimate, na.rm = TRUE)
media_df1 <- mean(reg_1$estimate, na.rm = TRUE)
```

```r
p1 <- ggplot(reg, aes(estimate)) +
  geom_histogram(  bins = 100, alpha = 0.5) +
  geom_vline(aes(xintercept = media_df), color = "red", linetype = "dashed") +
  theme_economist_white() +
  ggtitle("Distribución de estimadores sin correlación") +
  theme(plot.title = element_text(face = "bold", hjust = 0.5,
                                  size = 10))


p2 <- ggplot(reg_1, aes(estimate)) +
  geom_histogram(  bins = 100, alpha = 0.5) +
  geom_vline(aes(xintercept = media_df1), color = "red", linetype = "dashed") +
  theme_economist_white() +
  ggtitle("Distribución de estimadores con correlación") +
  theme(plot.title = element_text(face = "bold", hjust = 0.5,
                                  size = 10))



grid.arrange(p1, p2, ncol = 2)
```

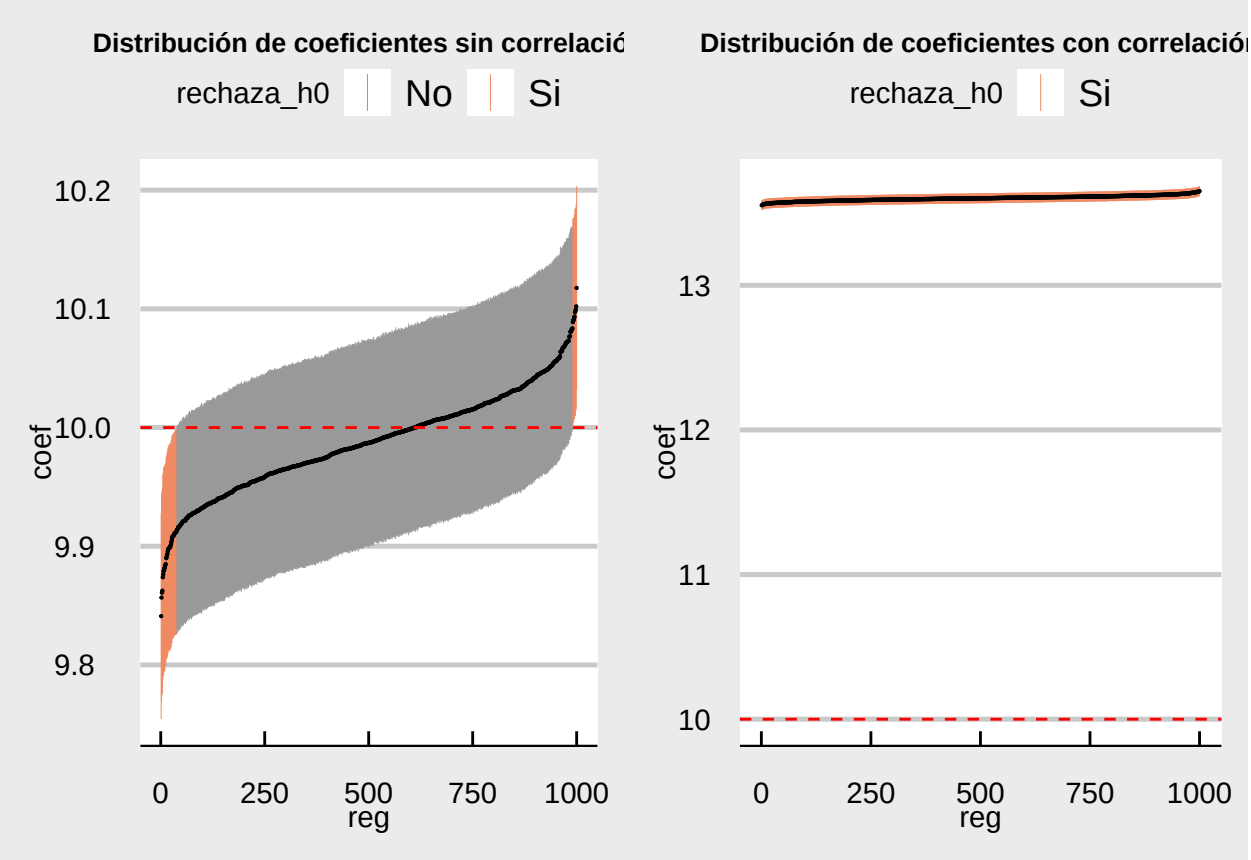


En esta gráfica se puede observar la diferencia entre el efecto de la variable de tratamiento nivel educativo y el ingreso. En el lado izquierdo, cuando no hay correlación entre el nivel educativo y la habilidad, el tamaño del efecto estimado es igual al especificado (tamaño efecto =10), mientras que en el caso de la simulación

2, el efecto estimado es mayor al efecto definido (efecto estimado = 13.6), es decir el $\hat{\beta} \neq \beta$ poblacional, o dicho de otra forma

$$\hat{\beta} = \beta + sesgodeselección$$

```
df_h0 <- reg %>%
         select(estimate, conf.low, conf.high, reg)  %>%
         mutate(h0=10,
                rechaza_h0=ifelse(conf.low < h0 &  h0 < conf.high , "No", "Si")) %>%
             arrange(estimate) %>%
             mutate(reg=c(1:nrow(.)))

df_h0_1 <- reg_1 %>%
  select(estimate, conf.low, conf.high, reg)  %>%
  mutate(h0 = 10,
         rechaza_h0 = ifelse(conf.low < h0 & h0 < conf.high, "No", "Si")) %>%
  arrange(estimate) %>%
  mutate(reg = c(1:nrow(.)))


p1 <- ggplot(df_h0, aes(x = reg)) +
  geom_linerange(aes(ymin = conf.low, ymax = conf.high, color = rechaza_h0), lwd = 0.1) +
  scale_color_manual(values = c("No"= "#999999","Si" = "#ef8a62")) +
  geom_point(aes(y = estimate), size = 0.1) +
  geom_hline(aes(yintercept = h0), color = "red", linetype = "dashed") +
  labs(y = "coef", x = "reg") +
  theme_economist_white() +
  ggtitle("Distribución de coeficientes sin correlación") +
  theme(plot.title = element_text(face = "bold", hjust = 0.5,
                                  size = 10))


p2 <- ggplot(df_h0_1, aes(x = reg)) +
  geom_linerange(aes(ymin = conf.low, ymax = conf.high, color = rechaza_h0), lwd = 0.1) +
  scale_color_manual(values = c("No"= "#999999","Si" = "#ef8a62")) +
  geom_point(aes(y = estimate), size = 0.1) +
  geom_hline(aes(yintercept = h0), color = "red", linetype = "dashed") +
  labs(y = "coef", x = "reg") +
  theme_economist_white() +
  ggtitle("Distribución de coeficientes con correlación") +
  theme(plot.title = element_text(face = "bold", hjust = 0.5,
                                  size = 10))



grid.arrange(p1, p2, ncol = 2)
```



Se observa que cuando se cumple el supuesto de exogeneidad, el parámetro poblacional (α) que representa el efecto de la educación sobre los ingresos se encuentra dentro del intervalo de confianza en el 95% de las estimaciones. Por otra parte, cuando violamos este supuesto en ningún caso se logra una estimación en la que el verdadero parámetro poblacional esté dentro del intervalo de confianza.

Análisis de variables de control omitidas

Cuando el proceso generador de datos incluye variables que observables que no están incluidas en el modelo, el tamaño del efecto de la variable de tratamiento es sesgado. Para poder evidenciar esto realizamos una simulación del siguiente proceso:

$$Ingreso = \beta \ + \alpha \ educación + \theta educaciónpadres + \epsilon$$

Donde θ es la relación que existe entre la educación de los padres y el ingreso de cada individuo. En este caso se asume que el valor de este parametro es igual a 11.

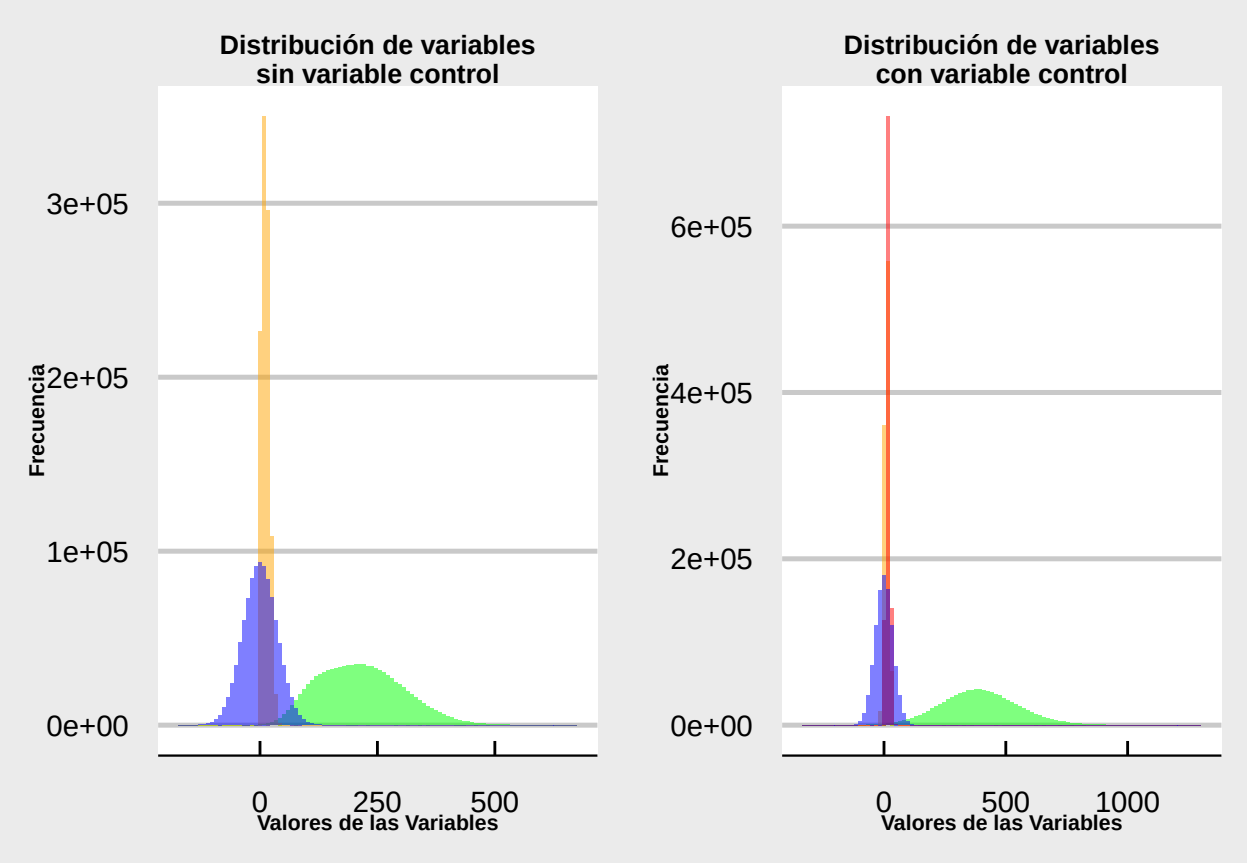
```
df_2 <- tibble(
  educ = rnorm(N, mean = 11, sd = 9),
  u = rnorm(N, mean = 0, sd = 36),
  effect_educ = 10,
  effect_educa_padre = 11,
  income = income_aut + effect_educ * educ + u
) %>%
  mutate(
    niv_educ = case_when(
```

```
      educ <= 5 ~ 0,
      educ > 5 & educ <= 9 ~ 5,
      educ > 9 & educ <= 11 ~ 9,
      educ > 11 & educ <= 16 ~ 11,
      educ > 16 & educ <= 18 ~ 16,
      educ > 18 & educ <= 20 ~ 18,
      educ > 20 ~ 20
    ),
    educa_padre_base = rnorm(N, mean = 12, sd = 9),
    educa_padre = pmax(educa_padre_base, educ),
    niv_educa_padre = case_when(
      educa_padre <= 5 ~ 0,
      educa_padre > 5 & educa_padre <= 9 ~ 5,
      educa_padre > 9 & educa_padre <= 11 ~ 9,
      educa_padre > 11 & educa_padre <= 16 ~ 11,
      educa_padre > 16 & educa_padre <= 18 ~ 16,
      educa_padre > 18 & educa_padre <= 20 ~ 18,
      educa_padre > 20 ~ 20
    ),
    income = income_aut + effect * educ + effect_educa_padre * educa_padre + u
  )
```

Se evidencia que existe una correlación entre la educación de los padres y los hijos.

```
p1 <- ggplot(df) +
  geom_histogram(aes(income), bins = 100, fill = "green", alpha = 0.5) +
  geom_histogram(aes(educ), bins = 100, fill = "orange", alpha = 0.5) +
  geom_histogram(aes(u), bins = 100,   fill="blue", alpha=0.5)+
  theme_economist_white() +
  ggtitle(paste0("Distribución de variables","\n" ,
                 "sin variable control")) +
  labs(x = "Valores de las Variables", y = "Frecuencia") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5,
                              size=10),
    axis.title.x = element_text(face = "bold", size=8),
    axis.title.y = element_text(face = "bold", size=8),
    #legend.title=element_text(size =8)
  )


p2 <- ggplot(df_2) +
  geom_histogram(aes(income), bins = 100, fill = "green", alpha = 0.5) +
  geom_histogram(aes(educ), bins = 100, fill = "orange", alpha = 0.5) +
  geom_histogram(aes(educa_padre), bins = 100,   fill="red", alpha=0.5)+
  geom_histogram(aes(u), bins = 100,   fill="blue", alpha=0.5)+
  theme_economist_white() +
  ggtitle(paste0("Distribución de variables" ,"\n" ,
                 "con variable control")) +
  labs(x = "Valores de las Variables", y = "Frecuencia") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5,
                              size = 10),
```

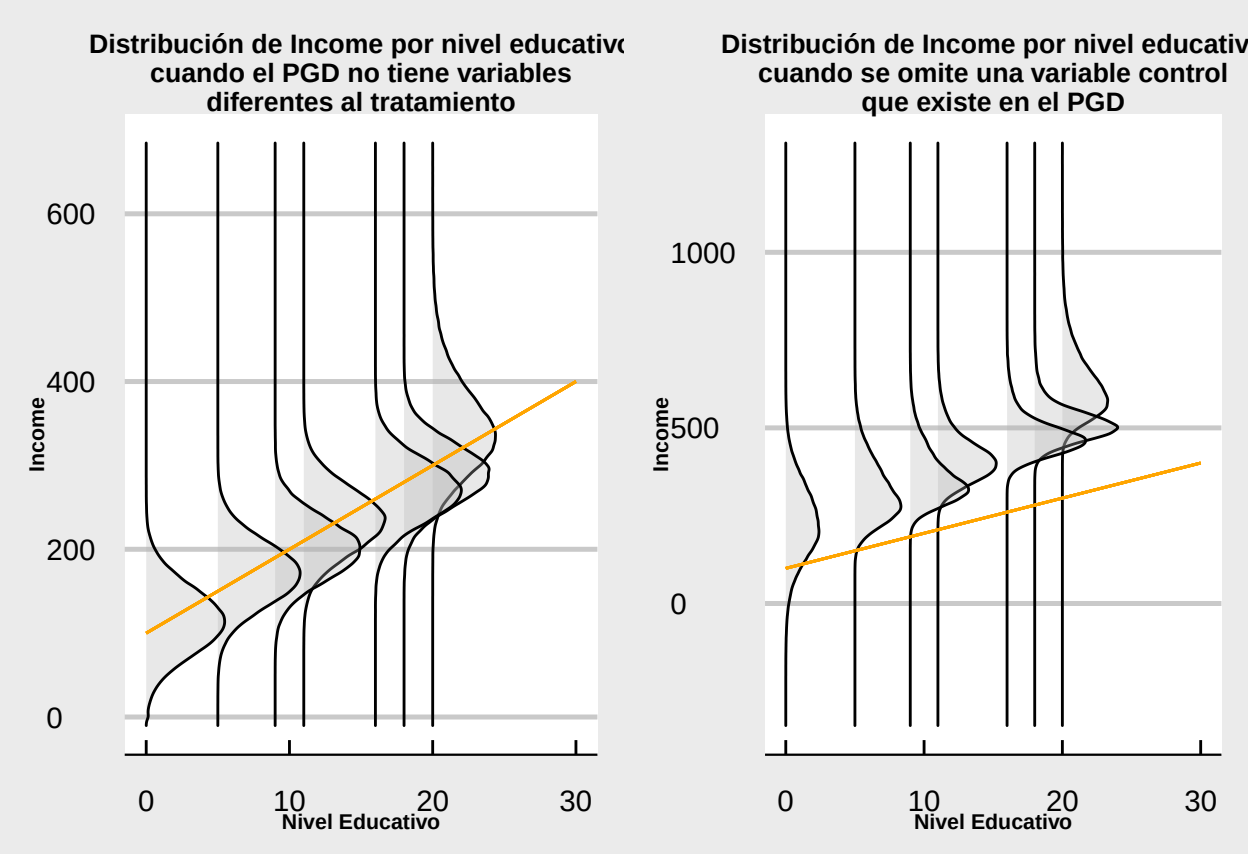
```
    axis.title.x = element_text(face = "bold", size=8),
    axis.title.y = element_text(face = "bold",size=8)
  )


grid.arrange(p1, p2, ncol = 2)
```



En la gráfica de la izquierda se observa que la linea de regresión se ajusta al comportamiento medio de la variable de ingreso en función de la educación cuando el proceso generador de datos solo incluye esta variable. Mientras que al incluir otras variables observables en este proceso, como la educación de los padres, esta linea se aleja del comportamiento promedio de los ingresos debido a la omisión de la educación de los padres en el modelo.

```
p1 <- ggplot(df, aes(x = income, y = niv_educ, group = niv_educ)) +
  geom_density_ridges(alpha = 0.3) +
  geom_function(fun = function(x) (-income_aut/effect) + (1/effect)*x, xlim = c(100, 400), color = "ora
  coord_flip() +
  theme_economist_white() +
  ggtitle(paste0("Distribución de Income por nivel educativo",
                 "\n","cuando el PGD no tiene variables","\n",
                 "diferentes al tratamiento")) +
  labs(x = "Income", y = "Nivel Educativo") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5,
                              size = 10),
    axis.title.x = element_text(face = "bold",size = 8),
```

```
    axis.title.y = element_text(face = "bold",size = 8)
  )


p2 <- ggplot(df_2, aes(x = income, y = niv_educ, group = niv_educ)) +
  geom_density_ridges(alpha = 0.3) +
  geom_function(fun = function(x) (-income_aut/effect) + (1/effect)*x, xlim = c(100, 400), color = "ora
  coord_flip() +
  theme_economist_white() +
  ggtitle(paste0("Distribución de Income por nivel educativo",
                 "\n","cuando se omite una variable control",
                 "\n","que existe en el PGD")) +
  labs(x = "Income", y = "Nivel Educativo") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5,
                              size = 10),
    axis.title.x = element_text(face = "bold",size = 8),
    axis.title.y = element_text(face = "bold",size = 8)
  )


grid.arrange(p1, p2, ncol = 2)
```




```
n_sample = 10000
n_iterations = 1000
```

```r
sample_and_regress <- function(df, formula, n_sample) {
  df_sample <- df[sample(nrow(df), n_sample, replace = FALSE), ]
  lm(formula, data = df_sample) %>%
    tidy(conf.int = TRUE) %>%
    filter(term == "educ") %>%
    select(-term)
}


reg <- map_dfr(1:n_iterations, ~ sample_and_regress(df_2, income ~ educ, n_sample), .id = "reg")


reg_2 <- map_dfr(1:n_iterations, ~ sample_and_regress(df_2, income ~ educ + educa_padre, n_sample), .id
```

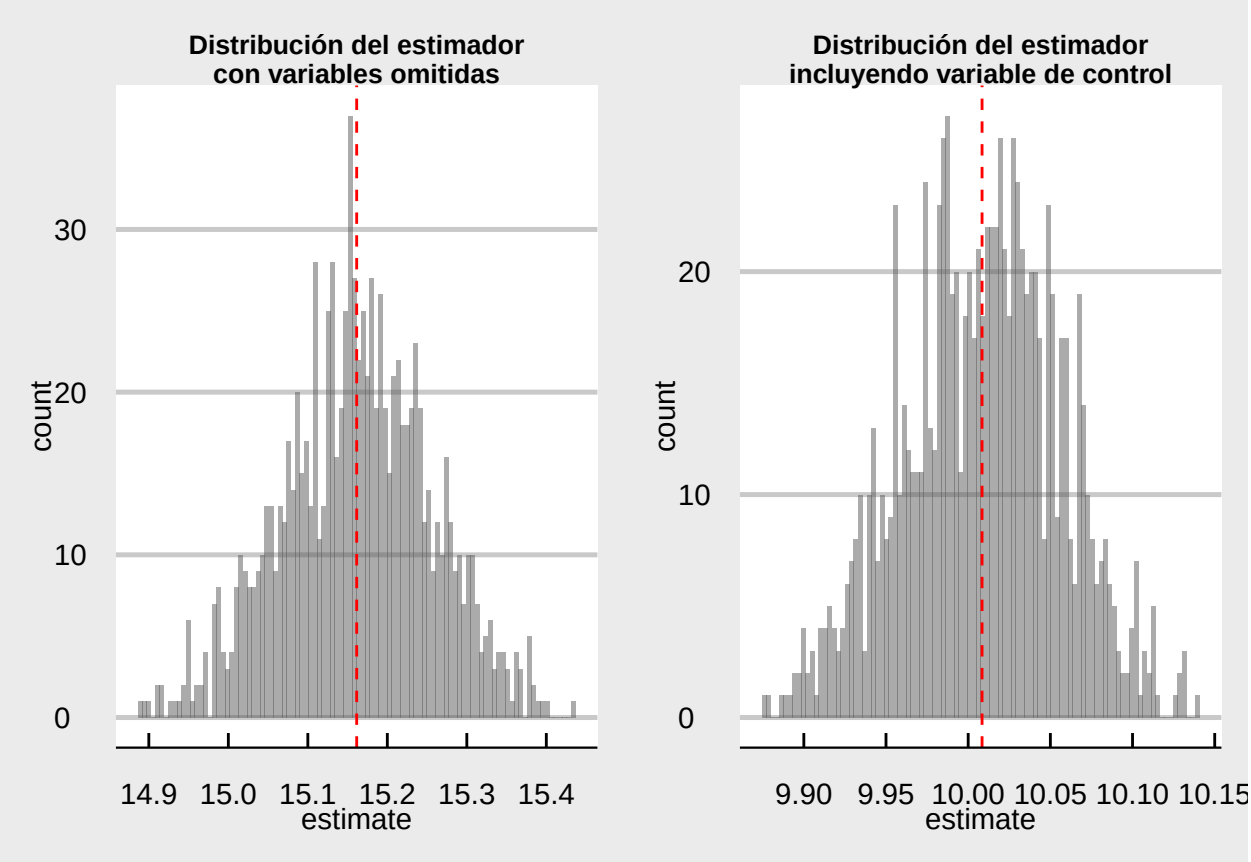
Cuando no se incluye una variable de control relevante en un modelo de regresión, el estimador puede ser inconsistente debido a un problema conocido como sesgo por variables omitidas .

Esto sucede porque el estimador de los coeficientes de regresión asume que todas las variables que afectan la variable dependiente están incluidas en el modelo. Si omitimos una variable de control relevante que está correlacionada tanto con la variable independiente (predictora) como con la variable dependiente, el modelo de regresión no puede separar adecuadamente el efecto de la variable independiente del efecto de la variable omitida.

Este sesgo afecta la consistencia del estimador, lo que significa que, incluso cuando el tamaño de la muestra crece indefinidamente, el estimador no converge al verdadero valor del parámetro. Específicamente, el estimador está sesgado porque la correlación entre la variable omitida y la variable independiente hace que el coeficiente estimado capte parte del efecto de la variable omitida.

```r
media_df <- mean(reg$estimate, na.rm = TRUE)
media_df2 <- mean(reg_2$estimate, na.rm = TRUE)


p1 <- ggplot(reg, aes(estimate)) +
  geom_histogram(  bins = 100, alpha = 0.5) +
  geom_vline(aes(xintercept = media_df), color = "red", linetype = "dashed") +
  theme_economist_white() +
  ggtitle(paste0("Distribución del estimador", "\n",
                 "con variables omitidas")) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5,
                                  size = 10))


p2 <- ggplot(reg_2, aes(estimate)) +
  geom_histogram(  bins = 100, alpha = 0.5) +
  geom_vline(aes(xintercept = media_df2), color = "red", linetype = "dashed") +
  theme_economist_white() +
  ggtitle(paste0("Distribución del estimador","\n",
                 "incluyendo variable de control"))+
  theme(plot.title = element_text(face = "bold", hjust = 0.5,
                                  size = 10))
```

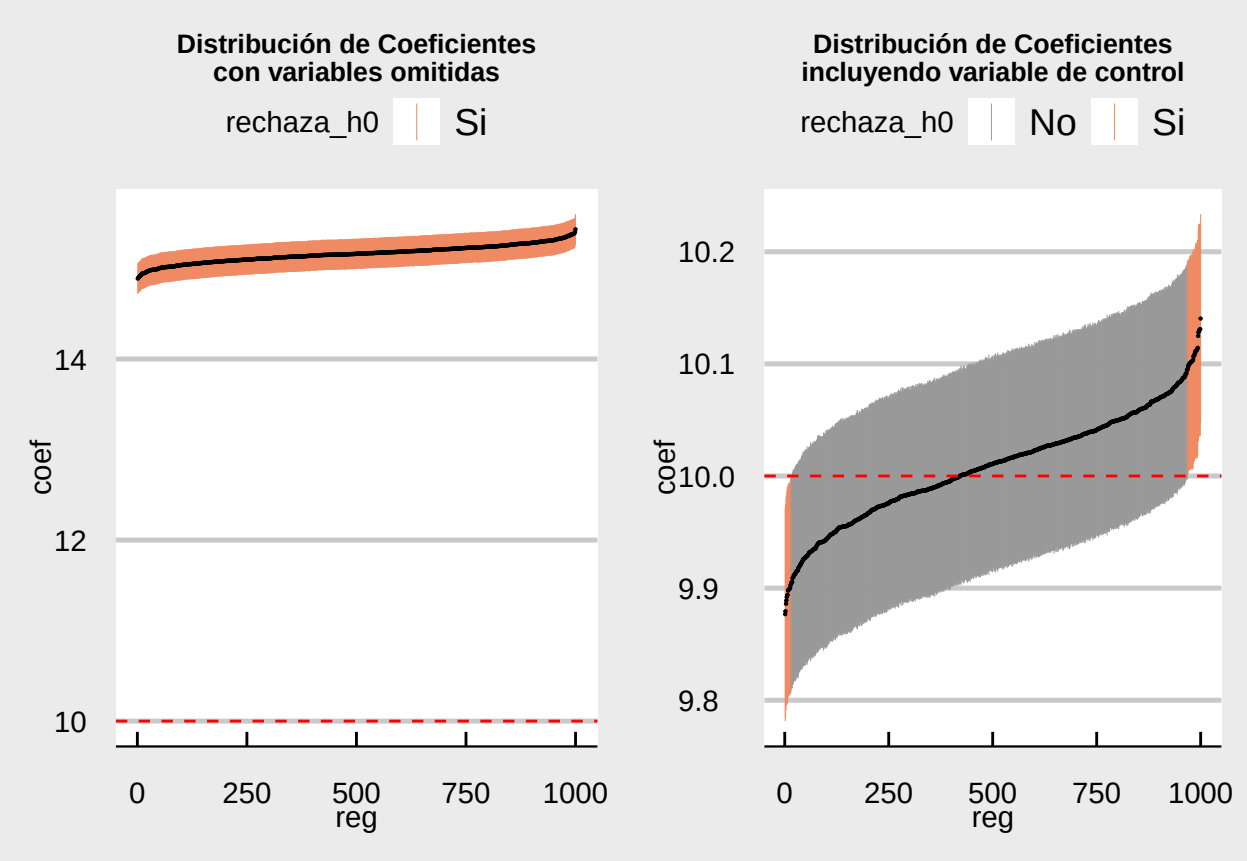
```
grid.arrange(p1, p2, ncol = 2)
```



En este caso cuando sabemos que la variable ingreso esta explicada por la educación del padre y la del hijo pero no controlamos por la primera, el estimador no converge a 10 que es el valor poblacional.

```
df_h0 <- reg %>%
         select(estimate, conf.low, conf.high, reg)  %>%
         mutate(h0=10,
                rechaza_h0=ifelse(conf.low < h0 &  h0 < conf.high , "No", "Si")) %>%
             arrange(estimate) %>%
             mutate(reg=c(1:nrow(.)))

df_h0_1 <- reg_2%>%
  select(estimate, conf.low, conf.high, reg)  %>%
  mutate(h0 = 10,
         rechaza_h0 = ifelse(conf.low < h0 & h0 < conf.high, "No", "Si")) %>%
  arrange(estimate) %>%
  mutate(reg = c(1:nrow(.)))


p1 <- ggplot(df_h0, aes(x = reg)) +
  geom_linerange(aes(ymin = conf.low, ymax = conf.high, color = rechaza_h0), lwd = 0.1) +
  scale_color_manual(values = c("No"= "#999999","Si" = "#ef8a62")) +
  geom_point(aes(y = estimate), size = 0.1) +
  geom_hline(aes(yintercept = h0), color = "red", linetype = "dashed") +
```

```
  labs(y = "coef", x = "reg") +
  theme_economist_white() +
  ggtitle(paste0("Distribución de Coeficientes","\n",
                 "con variables omitidas")) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5,
                                  size = 10))


p2 <- ggplot(df_h0_1, aes(x = reg)) +
  geom_linerange(aes(ymin = conf.low, ymax = conf.high, color = rechaza_h0), lwd = 0.1) +
  scale_color_manual(values = c("No"= "#999999","Si" = "#ef8a62")) +
  geom_point(aes(y = estimate), size = 0.1) +
  geom_hline(aes(yintercept = h0), color = "red", linetype = "dashed") +
  labs(y = "coef", x = "reg") +
  theme_economist_white() +
  ggtitle(paste0("Distribución de Coeficientes",
                 "\n","incluyendo variable de control")) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5,
                                  size = 10))


grid.arrange(p1, p2, ncol = 2)
```



Se evidencia que la omisión de una variable observable dentro del modelo provoca que en ningún caso el parámetro poblacional se encuentre dentro del intérvalo de confianza de los estimadores. En contraste con

la gráfica de la derecha en la que la inclusión de la educación de los padres permite que nuevamente el efecto de la educación sobre los ingresos se encuentre dentro del intervalo de confianza de los estimadores en el 95% de las veces.

Conclusiones

El objetivo de este ejercicio es evaluar los supuestos básicos del modelo de regresión lineal en situaciones en las que los datos no cumplen con las condiciones ideales del estimador. Inicialmente, se realizan estimaciones utilizando un modelo de regresión lineal en un contexto donde los datos son generados aleatoriamente, asumiendo que no existe correlación entre el término de error (habilidad) y el nivel de educación.

Esto proporciona una referencia para evaluar el rendimiento del estimador bajo condiciones ideales. En una segunda fase, se introduce una correlación entre la educación y el término de error, reflejando una situación más realista donde el nivel de habilidad no observada puede influir tanto en el nivel educativo como en el ingreso. Esta correlación puede sesgar el estimador del impacto de la educación sobre el ingreso, ya que el término de error, que captura factores no observables que afectan el ingreso, también está relacionado con el nivel educativo.

Se logra evidenciar que la omisión de variables de control observables en un modelo también puede provocar que sesgos en las estimaciones de los efectos de las variables de tratamiento. Por esta razón, es importante que se incluya la mayor cantidad de variables de control como primera medida para controlar el sesgo de selección como estrategia empirica para realizar una correcta inferencia causal.

Como resultado, la estimación del efecto de la educación puede no reflejar con precisión la verdadera relación entre educación e ingreso. Este ejercicio subraya la importancia de verificar los supuestos básicos del modelo de regresión lineal, como la independencia entre el término de error y las variables explicativas, ya que la violación de estos supuestos puede llevar a estimaciones sesgadas e incorrectas, lo que resalta la necesidad de técnicas adicionales para obtener estimaciones más confiables en presencia de sesgo.