Proyecto final

Prof. Alonso Quijano Ruiz

2025-01-18

Bienvenidos al final de este módulo. Ha sido un gusto dictar este curso. Espero que hayan disfrutado el aprendizaje de algunas de las aplicaciones más importantes de la estadística y la econometría. Este proyecto una excelente preparación para lo que les espera en un posgrado en ciencias sociales con enfoque cuantitativo. Por favor sigan las indicaciones:

i. La fecha de entrega del proyecto es el **lunes 10 de febrero 23h59**. La fecha es no se puede ser postergada.
ii. El proyecto es en grupo de tres. Pueden ser los mismos grupos en los que han estado presentanto las tareas.
iii. Cada grupo debe presentar las soluciones a los problemas en un DataLab. El DataLab no pueden arrojar errores. Si no puedo reproducir el código, les bajaré puntos.
iv. El **martes 11** y **jueves 13 de febrero** habrán presentaciones grupales. Cada grupo debe preparar una presentación académica que no debe durar más de 20 minutos. En ella deben presentar los procedimientos y resultados de cada ejercicio (incluyendo código, gráficos y explicaciones textuales).
v. Hay tres problemas y son tres estudiantes por grupo. Por ende, cada estudiante deberá presentar un problema distinto. No pueden delegar a un estudiante para que presente todo el proyecto.
vi. Durante y después de la presentación haré preguntas sobre los conceptos que hemos aprendido. Será como una especie de examen oral.

¡Éxitos en el proyecto! Demuéstrenme lo que han aprendido.

El casino de Alonso

Su profesor Alonso ha decidido que, si algún día se retira de la docencia y la investigación, abrirá un casino en su ciudad natal Guayaquil. Para ello, se ha inventado un juego llamado "Los dados locos". El juego consiste en lanzar dos dados. El participante gana en dólares el resultado del primer dado y pierde en dólares el 50% del resultado del segundo dado. Por su puesto, el juego tiene un costo. El participante debe pagar 2 dólares por jugar.

Por ejemplo, si un participante lanza los dados y obtiene un 5 y un 2, entonces la pérdida para el casino será 2 dólares ($2 - (5 - 0.5 \times 2) = -2$).

Como Alonso se considera un as en las matemáticas, ha calculado que el juego tiene una ganancia esperada positiva para el casino, por lo que es viable.

Llamemos X a una variable aleatoria que representa el resultado del lanzamiento del primer dado, Y el resultado del lanzamiento del segundo dado, y Z la ganancia del casino. Z se calcula como:

$$Z = 2 - (X - 0.5Y)$$

Parte 1

a. ¿Cuál es el valor esperado de X y de Y? Muestra los cálculos, no solo la respuesta.
b. ¿Cuál es la varianza de X y de Y? ¿Qué tipo de distribución tienen X y Y?
c. ¿Cuál es el valor esperado de Z? Pista: Usa las propiedades de la esperanza matemática.
d. ¿Cuál es la varianza de Z? Pista: Recuerda que X y Y son independientes. Por lo tanto, $\text{Cov}(X, Y) = 0$.

Parte 2

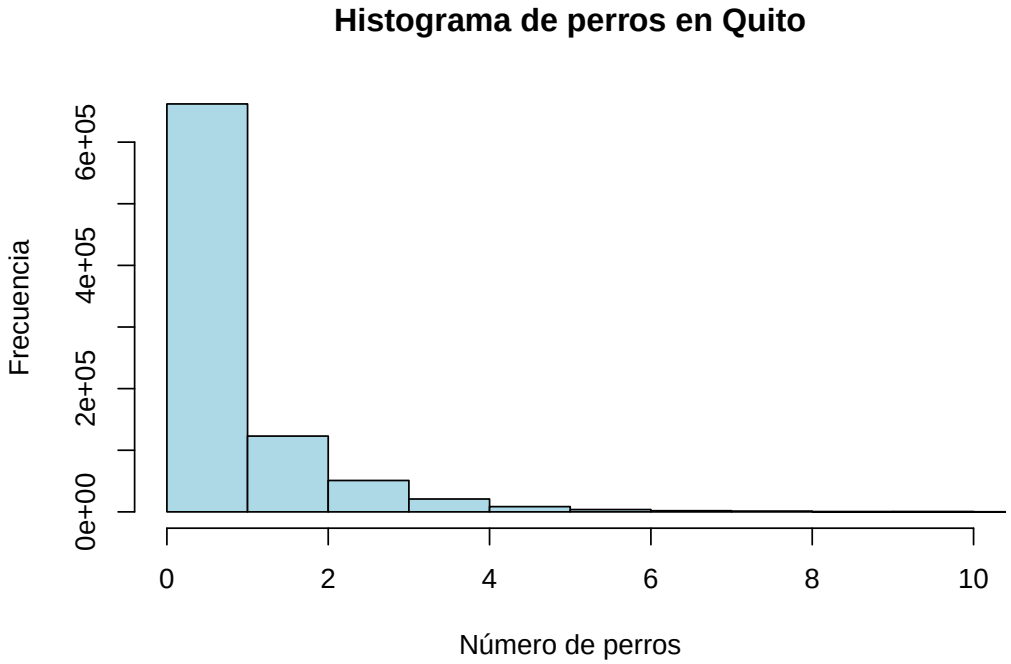
a. Para asegurarse que el juego produzca ganancias, Alonso realiza una simulación en R. En la simulación, Alonso realiza 10,000 experimentos. En cada experimento, simula el lanzamiento de dos dados y calcula la ganancia para el casino. ¿Es la ganancia promedio positiva? ¿Se acerca al valor esperado de Z que calculaste en la parte 1? Utiliza un `set.seed(123)` antes de realizar la simulación para que puedas replicar los resultados.
b. Calcula la varianza de la ganancia en la simulación. ¿Se acerca a la varianza de Z que calculaste en la parte 1?
c. Muestra el gráfico de densidad de Z y descríbelo en dos o tres oraciones.
d. Utilizando los resultados de la simulación, ¿cuál es la probabilidad que el cansino tenga una ganancia positiva por juego?
e. ¿Cuál es la probabilidad que el casino tenga una ganancia de 1 a 2 dólares por juego?

Los perros de Quito

En el censo del 2022, el INEC por primera vez incluyó una pregunta para calcular el número de mascotas en los hogares del Ecuador. En este ejercicio trabajaremos con la base de datos del número de perros para la población de hogares de Quito. El archivo se llama `quito_perros.RData`. Empecemos viendo la su frecuencia y distribución.

```r
# Cargar la base de datos
load("data/quito_perros.RData")

# Histograma de mascotas en Quito
hist(
  quito_perros,
  breaks = seq(min(quito_perros), max(quito_perros), by = 1),
  xlim = c(0, 10), # Limitamos el eje x para mejor visualización
  main = "Histograma de perros en Quito",
  xlab = "Número de perros",
  ylab = "Frecuencia",
  col = "lightblue",
)
```



Cuando los números son bien grandes, R los muestra en notación científica. Por ejemplo, 2e+05 es $2 \times 10^5 = 200,000$.

Parte 1

a. ¿Cuál es el promedio de perros en los hogares de Quito? ¿Y la varianza?
b. Crea un gráfico de densidad de la cantidad de perros en los hogares de Quito y limita el eje x a 10 perros para mejor visualización. Pista: solo debes cambiar el argumento `freq=FALSE` en la función `hist()`. En dos o tres oraciones, describe la diferencia entre el histograma y el gráfico de densidad. ¿Cómo se interpreta cada uno?
c. ¿Qué porcentaje de hogares tiene 0 perros? ¿Y 1 perro? ¿Y 2 perros? ¿Qué porcentaje de hogares tiene 3 o más perros?

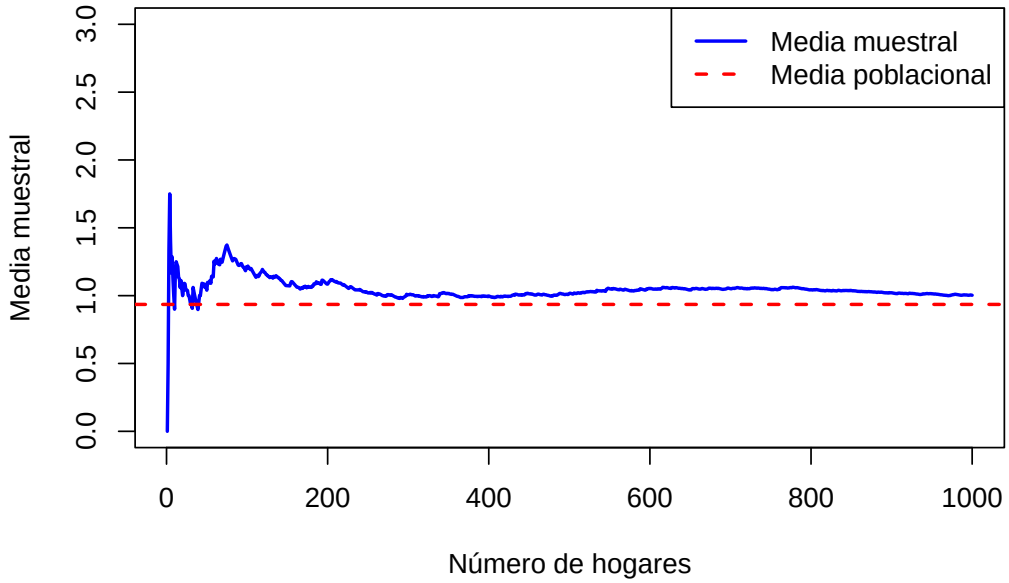
Parte 2

Es el año 2030 y su compañera YRF Vicky trabaja como analista de datos en el municipio de Quito. El alcalde de ese momento le ha pedido estimar el número de perros en los hogares de la ciudad. Él quiere implementar un programa de vacunación y esterilización de mascotas, y necesita saber cuántas vacunas y esterilizaciones se necesitarán. Lamentablemente, Vicky no puede usar la base de datos del Censo del 2022 porque ya han pasado muchos años, pero puede hacer una encuesta a 1000 hogares de Quito. Ella propone calcular la media muestral y estimar el total de perros en los hogares de Quito multiplicando la media muestral obtenida por el total de hogares (que sí conoce).

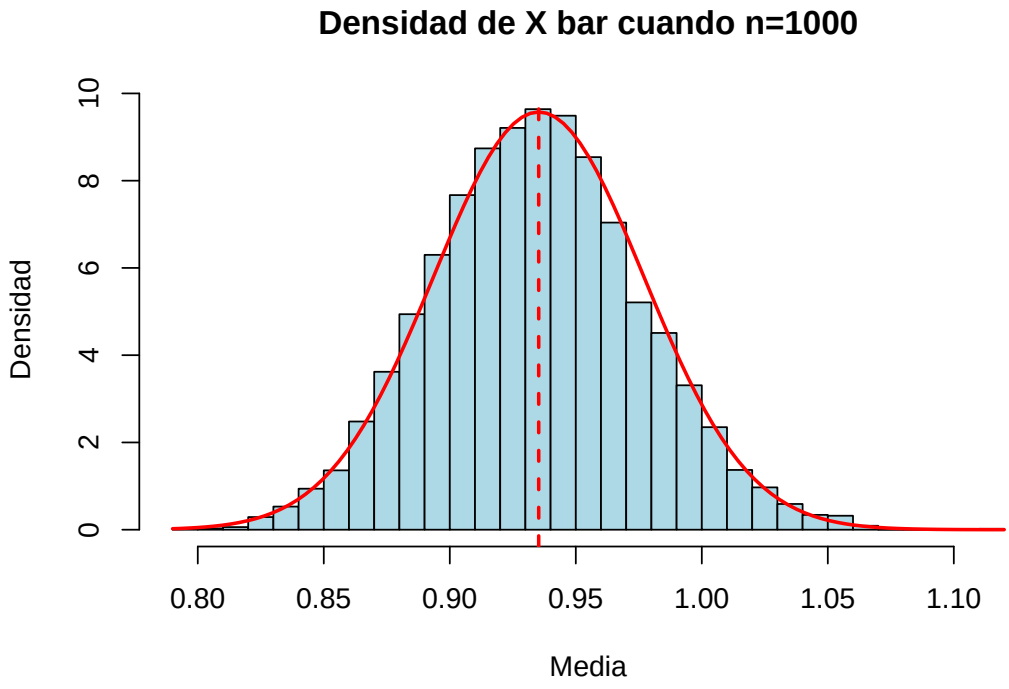
Antes de salir a encuestar, Vicky quiere comprobar si la ley de los grandes números y el teorema del límite centrar se extienden a la distribución de medias muestrales de la cantidad de perros en los hogares de Quito. Imaginen, si no estima bien el total de perros, el alcalde podría despedirla.

a. Para comprobar la ley de los grandes números, Vicky toma la base de datos del censo del 2022 crea una simulación de 1000 medias muestrales que van de tamaño $n = 1$ a $n = 1000$. Crea un gráfico comparando la evolución de las medias muestrales con la media poblacional de perros. Utiliza el argumento `replace = TRUE` de la función `sample` dado que la población es grande. ¿Qué observas?

¿Se acerca la media muestral a la media poblacional? Utiliza un `set.seed(123)` para que puedas replicar los resultados y muestra el código. Tu gráfico debería verse así.

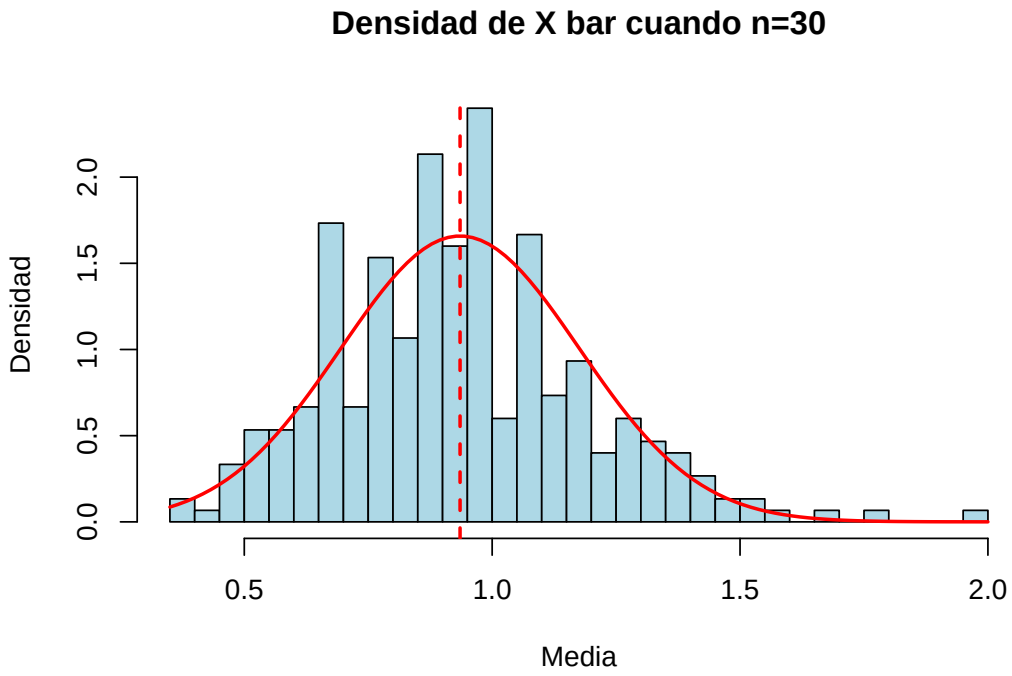


b. Vicky está contenta con los resultados de la ley de los grandes números, pero ahora necesita comprobar el teorema del límite central. Para ello, crea una simulación de 10000 medias muestrales de tamaño $n = 1000$. Luego, crea un gráfico de densidad de las medias muestrales. ¿Qué observas? ¿Se asemeja a la distribución normal? Utiliza un `set.seed(123)` y muestra el código. Tu gráfico de densidad debería verse así.



c. ¿Qué supuestos hizo Vicky para comprobar la ley de los grandes números y el teorema del límite central? ¿Cómo puede asegurarse que cuando vaya a encuestar que la media muestral obtenida sea insesgada y se acerque a la media poblacional objetivo?

d. Durante una reunión en el municipio, el jefe de Vicky la regaña y le dice que una muestra de tamaño $n = 1000$ es muy grande. Realizar una encuesta a tantos hogares le costaría mucho dinero al municipio. Su jefe le dice que él ha tomado cursos de estadística en una prestigiosa universidad del país, y que ahí él aprendió que con una muestra de tamaño $n = 30$ es suficiente para la distribución de medias muestrales se aproxime a una distribución normal. Realiza ahora una simulación de 10000 medias muestrales de tamaño $n = 30$ y muestra el gráfico de densidad. ¿Tiene razón el compañero de Vicky? ¿Por qué sí o por qué no? Utiliza un `set.seed(123)` y muestra el código. Tu gráfico de densidad debería verse así.



e. No estamos seguros si el jefe de Vicky tiene razón, pero en lo que sí tiene razón es que encuestar cuesta dinero. Realiza diferentes simulaciones y grafica la distribución de medias muestrales de diferentes tamaño de muestra (escoge tú el tamaño). ¿Qué tamaño de muestra recomendarías al municipio para ahorrar recursos sin sacrificar precisión en la estimación de la media poblacional? Si encuestar a cada hogar le cuesta al municipio 5 dólares, ¿cuál sería el costo logístico de realizar la encuesta con el tamaño muestral que propones?

Los retornos de la educación universitaria en el Ecuador

El gobierno del Ecuador está interesado en saber si conviene invertir más en educación universitaria gratuita, y desea medir los retornos de la educación universitaria en los ingresos de los ecuatorianos. El gobierno contrata a su compañera YRF Odalis para esta tarea. Ella propone tomar la base de datos de la Enemdu (Encuesta Nacional de Empleo, Desempleo y Subempleo) y estimar un modelo de regresión lineal. La base de datos con la que vamos a trabajar este ejercicio proviene de la Enemdu del 2023 y se llama `enemdu_ing_2023.csv`.

Las variables en la base de datos son:

- id_per: Identificador único de la persona
- edad: Edad de la persona
- mujer: Variable binaria que toma el valor de 1 si la persona es mujer y 0 si es hombre
- ingreso: Ingreso mensual de la persona

- experiencia: Años de experiencia laboral
- colegio: Variable binaria que toma el valor de 1 si la persona culminó la educación secundaria y 0 si no
- universidad: Variable binaria que toma el valor de 1 si la persona tiene algo de educación universitaria y 0 si no

Parte 1

Empezamos cargando la base de datos y realizando un análisis descriptivo de los datos.

```
# Cargar tidyverse para manipulación de datos
library(tidyverse)

# Cargar la base de datos
enemdu_ing_2023 <- read_csv("data/enemdu_ing_2023.csv")

# Ver las primeras observaciones
head(enemdu_ing_2023)
```

```
## # A tibble: 6 x 7
##    id_per  edad mujer ingreso experiencia colegio universidad
##     <dbl> <dbl> <dbl>   <dbl>       <dbl>   <dbl>       <dbl>
## 1 1.02e18    78     1      NA          NA       0           0
## 2 1.02e18    79     0      NA          NA       0           0
## 3 1.02e18    73     1      NA          NA       0           0
## 4 1.02e18    46     0     646          15       1           1
## 5 1.02e18    59     1      NA          NA       1           1
## 6 1.02e18    85     1      NA          NA       0           0
```

a. ¿Cuál es el total de observaciones?
b. ¿Cuántas personas han culminado la educación secundaria (incluyendo bachillerato)? ¿Qué porcentaje representan?
c. ¿Cuántas personas tienen algo de educación universitaria? ¿Qué porcentaje representan?
d. ¿Cuál es el promedio de ingresos de las personas en la base de datos? Pista: No te olvides de incluir el argumento `na.rm = TRUE` para que R ignore los valores faltantes.

Parte 2

Tú eres el/la asistente de Odalis, y ella te encarga preparar la base de datos y proponer un modelo para estimar los retornos de la educación universitaria en los ingresos de los ecuatorianos. Jhannelly te da las siguientes indicaciones. Muestra el código en R para cada paso.

a. Eliminar a todas las personas que no han culminado la educación secundaria. El estudio solo desea estimar los retornos de la educación universitaria para la población de personas que han culminado la secundaria. También le pide eliminar a las personas que no reportan valores en sus ingresos. Estos valores están registrados como `NA` en la variable ingresos. El número de observaciones restantes debería ser:

```
## [1] 88625
```

b. Ciertamente hay gente que gana muchísimo dinero y que está en la base de datos. Estos datos pueden sesgar los estimadores. Por lo tanto, te pide eliminar los datos atípicos (Lee aquí sobre los valores atípicos https://www.youtube.com/watch?v=R2U3apzVB9E&t=6s). Le presentas a Odalis el siguiente código calculando los cuantiles.

```r
quantile(enemdu_ing_2023$ingreso, probs = c(0, 0.25, 0.5, 0.75, 1), na.rm = TRUE)
```

```
##    0%   25%   50%   75%  100%
##     0   300   500   850 60500
```

Odalis te pide que consideres como valor atípico a aquellos que sobrepasan el tercer cuartil más 1.5 veces el rango intercuartílico. El rango intercuartílico se calcula como la diferencia entre el tercer cuartil y el primer cuartil. Elimina los valores atípicos de la base de datos. El número de observaciones restantes debería ser:

```
## [1] 83377
```

c. Por último, te pide incluir solo a las personas que tienen entre 30 y 40 años. Ya que las personas muy jóvenes todavía no han ido a la universidad y no se han consolidado en el mercado laboral, y las personas mayores ya han pasado su pico de ingresos. Elimina las personas que no están en el rango de edad. El número de observaciones restantes debería ser:

```
## [1] 24036
```

Parte 3

a. Una vez limpia la base de datos, ya estás listo/a para estimar tu modelo. El primer modelo que propones es:

$$\text{Ingresos}_i = \beta_0 + \beta_1 \times \text{Universidad}_i + u_i$$

donde Ingresos_i es el ingreso mensual de la persona i, Universidad_i toma el valor de 1 si la persona i tiene algo de educación universitaria y 0 si no, y u es el término de error. Estima el modelo e interpreta los coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$ hallados. ¿Cuál es el ingreso predicho de una persona con educación universitaria? ¿Y de una persona sin educación universitaria?

b. En clases aprendimos sobre la homocedasticidad y heterocedasticidad en los errores. Define estos términos. ¿Crees que tu modelo propuesto sufre de heterocedasticidad? Para comprobarlo, obtén la varianza de los errores cuando Universidad= 1 y cuando Universidad= 0. ¿Son las varianzas similares? Si no lo son, da una explicación intuitiva de por qué una es mayor que la otra.

c. Durante una reunión con Odalis, ella te indica que el modelo propuesto puede sufrir de sesgo de variable omitida. ¿Qué quiere decir Odalis? Define el sesgo de variable omitida y explica por qué puede ocurrir en este caso.

d. Odalis te pide que incluyas la variable Experiencia para observar si el valor de $\hat{\beta}_1$ cambia al incluir una segunda variable independiente.

$$\text{Ingresos}_i = \beta_0 + \beta_1 \times \text{Universidad}_i + \beta_2 \times \text{Experiencia}_i + u_i$$

Estima el modelo y comenta los resultados.

e. Calcula el sesgo de variable omitida. Calcúlalo de dos formas: 1) calcula la diferencia entre los coeficientes de Educación de los dos modelos y 2) calcula el sesgo de variable omitida con la fórmula:

$$\text{Sesgo} = \beta_2 \delta_1$$

donde δ_1 es el coeficiente del modelo:

$$\text{Experiencia}_i = \delta_0 + \delta_1 \times \text{Universidad}_i + v_i$$

Comenta el signo y magnitud del sesgo de variable omitida. ¿Por qué es positivo o negativo?

f. Para sorprender a tu jefa Odalis, estimas un modelo que examina las brechas de género en los ingresos de las ecuatorianas. Para ello, interactúas las variables Universidad y Mujer, lo cual permite medir efectos heterogéneos. El modelo es el siguiente:

$$\text{Ingresos}_i = \beta_0 + \beta_1 \times \text{Universidad}_i + \beta_2 \times \text{Mujer}_i + \beta_3 \times \text{Universidad}_i \times \text{Mujer}_i + u_i$$

donde Mujer_i es una variable binaria que toma el valor de 1 si la persona i es mujer y 0 si es hombre. Para estimar el modelo debes crear una variable cuyo resultado es la mutiplicación entre Universidad y Mujer. Puedes hacerlo de esta forma:

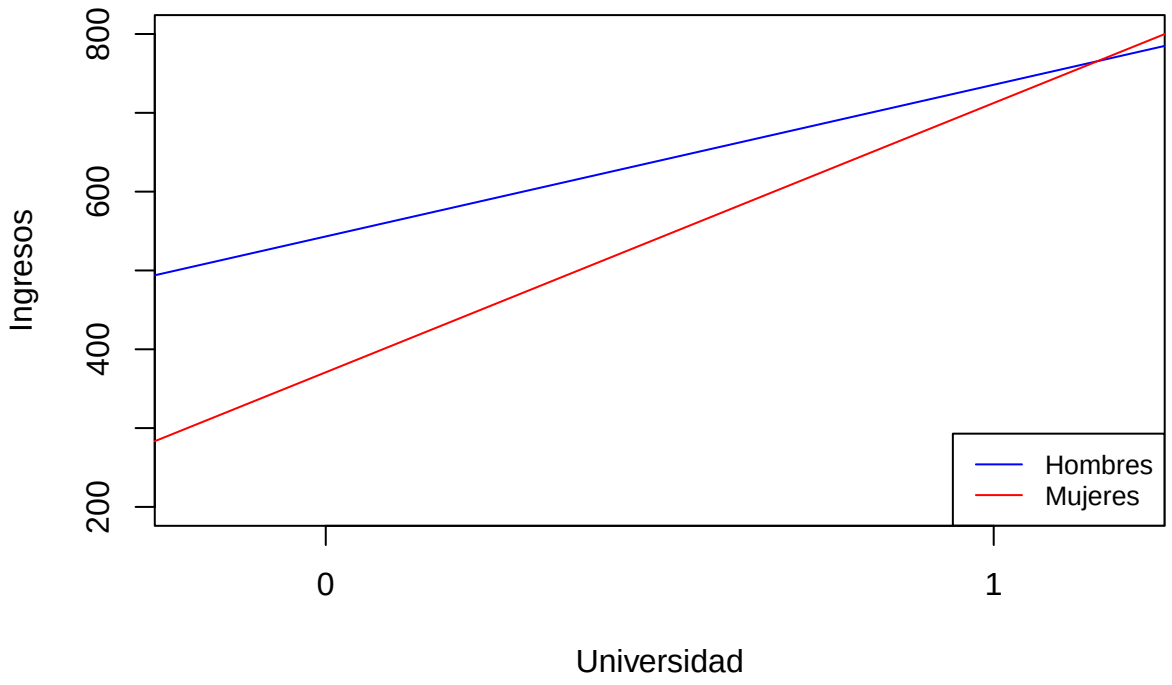
```
enemdu_ing_2023 <- enemdu_ing_2023 %>% mutate(uni_mj = universidad * mujer)
head(enemdu_ing_2023, 10)
```

```
## # A tibble: 10 x 8
##      id_per  edad mujer ingreso experiencia colegio universidad uni_mj
##       <dbl> <dbl> <dbl>   <dbl>       <dbl>   <dbl>       <dbl>  <dbl>
##  1 1.02e18    34     1     365           2       1           1      1
##  2 1.02e18    36     0     470           0       1           1      0
##  3 1.02e18    37     0      15          10       1           1      0
##  4 1.02e18    37     0     120           5       1           1      0
##  5 1.02e18    30     0    1006          10       1           0      0
##  6 1.02e18    38     1    1011          10       1           1      1
##  7 1.02e18    30     0    1000          10       1           0      0
##  8 1.02e18    38     1     940          10       1           1      1
##  9 1.02e18    33     0     500          10       1           0      0
## 10 1.02e18    37     1      70           0       1           1      1
```

Comenta los resultados e interpreta el significado de los siguientes valores:

i. $\hat{\beta}_0$
ii. $\hat{\beta}_0 + \hat{\beta}_1$
iii. $\hat{\beta}_0 + \hat{\beta}_2$
iv. $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$

g. El siguiente gráfico muestra el modelo anterior que estima los efectos heterogéneos de la educación universitaria entre hombres y mujeres. ¿Cómo encajan los coeficientes $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$?



h. ¿Qué propuesta harías para mejorar el estudio? Puedes aportar cualquier idea metodológica como la inclusión de más variables, la utilización otras técnicas estadísticas, o la recolección de más datos.