LAB

Informe - Entrega 1 - R

Alumnos: Noam Sade, Beltrán Saravia, Anselmo Torres y Lucca Vignoli

Contenidos

Ejercicio 1 - Análisis de datos:

Motivación

La mayor parte de la licenciatura en Economía en UTDT transcurre aprendiendo modelos para pensar distintas situaciones económicas. Queremos aprovechar este ejercicio para poder trabajar con datos macroeconómicos y vincularlos con temas vistos en materias de teoría económica e historia. En particular, vamos a usar datos del Banco Mundial vía un paquete de R con el objetivo de familiarizarnos con los números que rodean a temas que consideramos de interés, a la vez de tener la oportunidad de utilizar lo aprendido en las primeras unidades de la materia respecto a manejo de datos. Usaremos datos de 10 países de América Latina para estudiar tres temas que nos resultan interesantes.

En primer lugar, queremos ver cómo luce la “Curva de Phillips” para LATAM. El enfoque está en ver la relación estadística entre inflación y desempleo bajo distintos criterios (agrupando por país, por año y por nivel de ingreso de países analizados). Lo que resulta interesante es que hay ciertos casos donde la relación hallada entre inflación y desempleo es positiva. Siempre nos resultó interesante la polémica de la curva de Phillips. De haber una relación negativa entre inflación y desempleo, ¿es explotable desde el punto de vista de la política pública? ¿O vale la Crítica de Lucas o algún tipo de adaptación de expectativas por parte de los agentes que vuelve nula la capacidad del hacedor de política pública para tomar la relación negativa como un “trade off” con el cual elegir niveles de inflación y desempleo? La curiosidad por ver el comportamiento de esta relación entre inflación y desempleo en LATAM - con países caracterizados por periodos de alta inflación y desempleo - nos llevó a elegir este tema para, al margen de la teoría económica que rija de fondo, entender la relación estadística que impera en cada caso.

En segundo lugar, ¿qué tan importantes son las unidades al momento de medir? Queremos ver qué tan sensible puede ser un ranking entre países bajo ciertas variables ante cambios en unidades de medida. Elegimos el caso del PIB per cápita para los países de LATAM considerados. Son varias la manera de medirlo. A precios corrientes o a precios constantes, ambos para cada país en su propia

moneda. Al llevar todo a una misma moneda (digamos, dólares) aparecen los problemas de medida relacionados al tipo de cambio. De ahí poder volver a medir el PIB per cápita de distintos países en dólares corrientes, dólares constantes o en dólares ajustados por paridad del poder de compra (PPP, también en sus respectivas versiones corrientes o constantes). Dada esta ensalada de unidades de medida que a uno como alumno lo marea al principio, cabe preguntarse: ¿Cambia mucho el criterio de medida? Si tal fuera el caso (cosa que sí es), sería curioso ver de qué manera el criterio usado impacta en el ranking de países de LATAM por nivel de ingreso y eso es lo que estudiaremos.

En último lugar, queremos aprovechar la base de datos para estudiar un tema recurrente en materias de historia económica y de nuestro interés personal: ingreso y desigualdad. Dado un criterio para medir el PIB per cápita, nos preguntamos ¿Los países de mayores ingresos son más o menos igualitarios, en cuanto distribución del ingreso, comparado con los países de menores ingresos? A su vez, ¿cómo ha evolucionado la desigualdad y el ingreso en estos países? Nos valemos del índice de Gini y de las series del PIB per cápita para estudiar estas relaciones y graficarlas.

Una aclaración importante: en principio la base nos permite tomar datos desde 1960 hasta 2024 pero hay años para los que no hay datos de inflación o desempleo (NAs) para ciertos países. principalmente debido a problemas sobre la confiabilidad de los datos que podía llegar a ofrecer una agencia estadística nacional. Por ejemplo, para Argentina los datos de inflación que podemos usar van únicamente desde 2018 hasta 2024, mientras que para Brasil van desde 1991 hasta 2024. En ese sentido, nuestro análisis está muy limitado a tiempos recientes dado que usamos únicamente esta base pero creemos que podemos hacer cosas interesantes con las observaciones que se tienen.

Parte I: Curva de Phillips en LATAM

Los países que vamos a tener en cuenta a lo largo de todo el ejercicio son Argentina, Brasil, Chile, Colombia, México, Perú, Uruguay, Paraguay, Bolivia y Ecuador. Luego de bajar las librerías a usar (Tidyverse y WDI), invocamos datos de

inflación y desempleo para estos países, desde 1970 hasta 2024, vía el siguiente código:

```
paises <- c("AR","BR","CL","CO","MX","PE","UY","PY","BO","EC")   # Arg, Brasil, Chile, Colom
anios  <- c(1970, 2024)

#Queremos indicadores de tasa de inflación y desempleo, ambos anuales. Utilizamos los código
indicadores <- c(
  infl  = "FP.CPI.TOTL.ZG",     # Inflación anual, cambio en IPC (%)
  unemp = "SL.UEM.TOTL.ZS"      # Desempleo total (% fuerza laboral)
)

#Descarga de datos
datos <- WDI(
  country   = paises,
  indicator = indicadores,
  start     = anios[1],
  end       = anios[2],
  extra     = TRUE
)

View(datos)
```

Luego de limpiar las filas con NAs (es decir, las filas donde no había dato de inflación o de empleo para un país y año determinado), nos quedan 248 pares de observaciones para inflación y desempleo, con algunos países con más observaciones que otros. Para tener una idea, las primeras filas del dataframe resultante lucen así:

```
    country iso2c iso3c year status lastupdated      infl  unemp                   region
1 Argentina    AR   ARG 2018         2025-10-07  34.27722  9.220 Latin America & Caribbean
2 Argentina    AR   ARG 2019         2025-10-07  53.54830  9.843 Latin America & Caribbean
3 Argentina    AR   ARG 2020         2025-10-07  42.01509 11.461 Latin America & Caribbean
4 Argentina    AR   ARG 2021         2025-10-07  48.40938  8.736 Latin America & Caribbean
5 Argentina    AR   ARG 2022         2025-10-07  72.43076  6.805 Latin America & Caribbean
6 Argentina    AR   ARG 2023         2025-10-07 133.48894  6.139 Latin America & Caribbean
      capital longitude latitude              income lending
1 Buenos Aires  -58.4173 -34.6118 Upper middle income    IBRD
2 Buenos Aires  -58.4173 -34.6118 Upper middle income    IBRD
```

Vale aclarar que iso2c e iso3c son códigos de identificación de cada país, establecidos por el Banco Mundial, y que las columnas de mayor interés son infl (inflación) y unemp (porcentaje de fuerza laboral desempleada).

Como punto de partida, calculamos las correlaciones muestrales entre inflación y desempleo bajo distintos criterios. Utilizando todos los datos disponibles, sin condicionar a nada, la correlación es de 0.01101. El asunto cobra mayor interés cuando condicionamos por país: Para un mismo país, tomando las observaciones

de inflación y desempleo a lo largo de los años, ¿cómo luce la relación entre inflación y desempleo? Los resultados son los que siguen.

	country	correlacion
	<chr>	<dbl>
1	Mexico	0.321
2	Ecuador	0.272
3	Peru	-0.0349
4	Colombia	-0.124
5	Uruguay	-0.134
6	Paraguay	-0.218
7	Brazil	-0.452
8	Bolivia	-0.519
9	Argentina	-0.531
10	Chile	-0.575

Ciertamente esta tabla es curiosa. La relación negativa entre inflación y desempleo no es homogénea. Condicionando por la clasificación de ingresos armada por el Banco Mundial, la relación luce así:

income	correlacion
<chr>	<dbl>
Upper middle income	-0.0128
High income	-0.491
Lower middle income	-0.519

Por último, podemos utilizar el corte transversal y, para cada año, calcular la correlación entre inflación y desempleo utilizando datos de los distintos países. Es curioso porque hay años de alta correlación positiva y otros de alta correlación negativa, y da a pensar sobre qué estaba ocurriendo en cada uno de esos años. Ordenando la tabla de manera descendente, los primeros datos lucen así.

	year	correlacion
	<int>	<dbl>
1	2016	0.839
2	2015	0.753
3	2005	0.646
4	2019	0.374
5	1995	0.366
6	2018	0.351
7	2003	0.346
8	1992	0.310
9	2004	0.278
10	1993	0.264

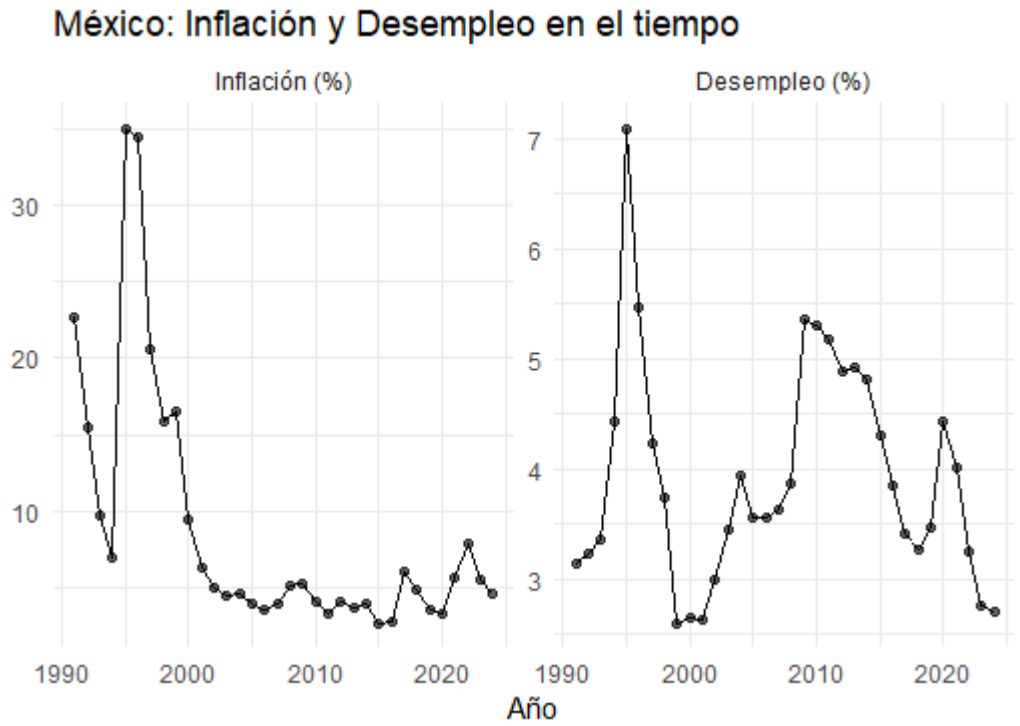
Vamos a considerar el caso de México (elegido a propósito porque es el de mayor correlación positiva entre inflación y desempleo y llama la atención). El gran jugador

en todo este ejercicio es el operador Pipe. Nos armamos un dataframe con datos de México con la
intención de hacer cuatro gráficos. Los primeros dos conciernen a la evolución de la inflación y el desempleo a lo largo del tiempo y los últimos dos son intentos de fittear una curva de Phillips.

Luego de filtrar el dataset con todos los datos para quedarnos únicamente con los datos de Mexico (en el dataset datos_mexico), nos valemos de la herramienta pivot para acomodar los datos de manera tal que podamos graficar las dos series con facet wrap en simultáneo. Más específicamente, usamos

```
#Hacemos los gráficos de inflación y desempleo en el tiempo

datos_mexico %>%

  pivot_longer(c(infl, unemp), names_to = "variable", values_to = "value") %>%

  #Queremos dos gráficos juntos. Usaremos facet wrap. Precisamos pivotear el dataframe para filtrar datos por inflacion o desempleo,

  ggplot(aes(x = year, y = value)) +
  geom_line() +
  geom_point(alpha = 0.6) +

  facet_wrap(
    ~ variable, ncol = 2, scales = "free_y",
    labeller = as_labeller(c(infl = "Inflación (%)", unemp = "Desempleo (%)"))
  ) +

  #Acá indicamos que pedimos dos gráficos, usando las dos variables (infl y unemp) de la columna "variable" del df pivoteado.

  labs(title = "México: Inflación y Desempleo en el tiempo", x = "Año", y = NULL) +
  theme_minimal()
```

y cuyo output es el siguiente par de gráficos:

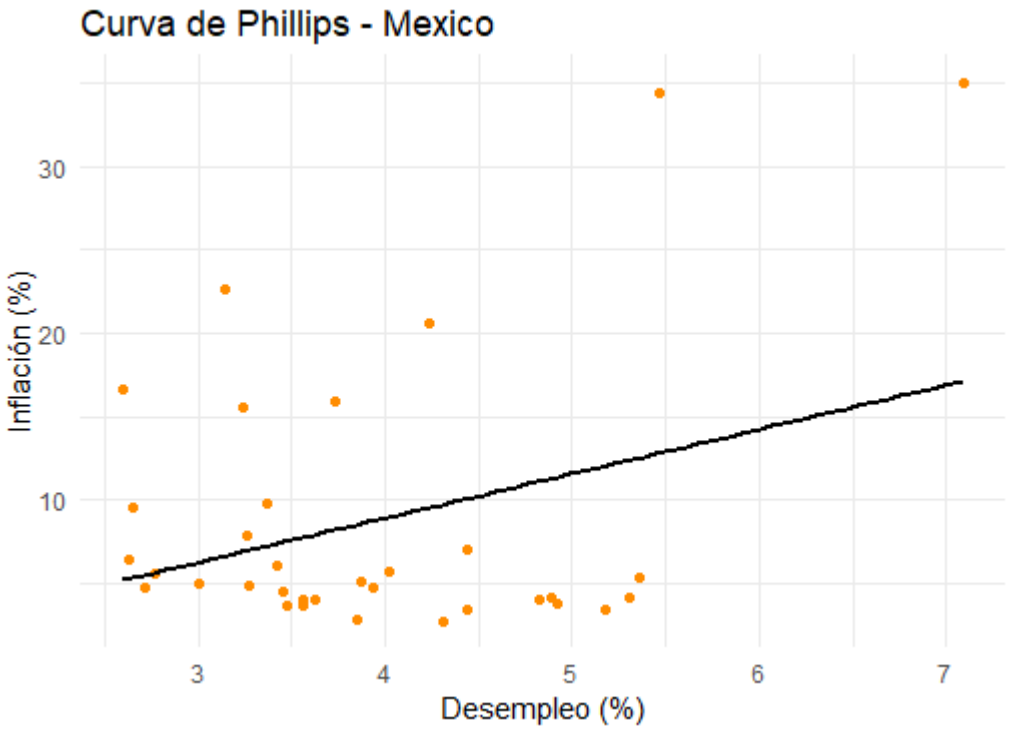


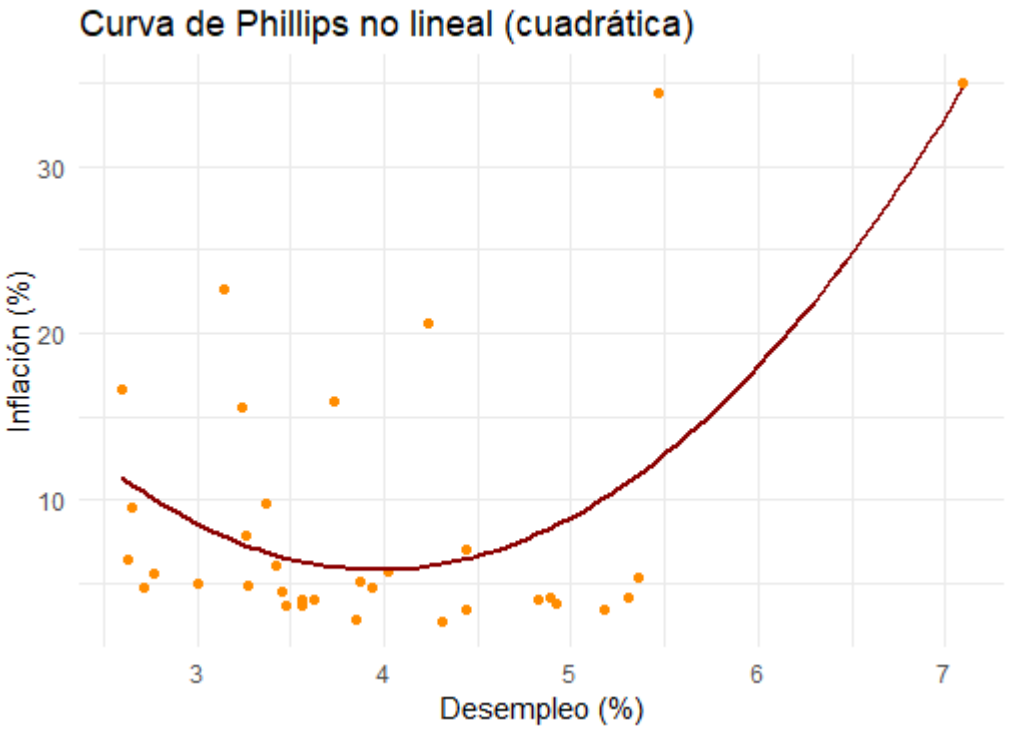
Lo primero que uno se pregunta es “¿Qué pasó a fines de los 1990s?”. Luego, llama la atención la baja de la inflación y el aumento del desempleo entre 2000 y 2010.

Ahora cerramos esta primera parte mostrando un scatter plot de inflación y desempleo y fitteando a los datos dos curvas: una recta y una parábola. El código usado fue el que se presenta a continuación.

```
#Veamos la relación entre la inflación y el desempleo en México

datos_limpios %>%
  filter(country == "Mexico") %>%
  ggplot(aes(x = unemp, y = infl)) +
  geom_point(color = "darkorange") +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "Curva de Phillips - Mexico",
       x = "Desempleo (%)", y = "Inflación (%)") +
  theme_minimal()

#Consideremos fittear una parábola, en vez de una recta, a los datos

datos_limpios %>%
  filter(country == "Mexico") %>%
  ggplot(aes(unemp, infl)) +
  geom_point(color = 'darkorange') +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), color = "darkred", se=FALSE)
  labs(title = "Curva de Phillips no lineal (cuadrática)",
       x = "Desempleo (%)", y = "Inflación (%)") +
  theme_minimal()
```

Y los gráficos resultantes son los siguientes:





Ciertamente la curva no tiene la forma habitual. Realizando una regresión de los datos de inflación sobre los datos de desempleo, llegamos a lo siguiente

```
Call:
lm(formula = infl ~ unemp, data = .)

Residuals:
   Min     1Q Median     3Q    Max
-8.621 -5.300 -3.004  2.178 21.579

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.727      5.616  -0.307   0.7605
unemp          2.658      1.386   1.918   0.0641 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Y vale recalcar que el enfoque de este ejercicio está en la exposición de datos y no en la validez econométrica de este modelo, ciertamente cuestionable en primer lugar por temas de endogeneidad.

Parte II: ¿En qué unidades?

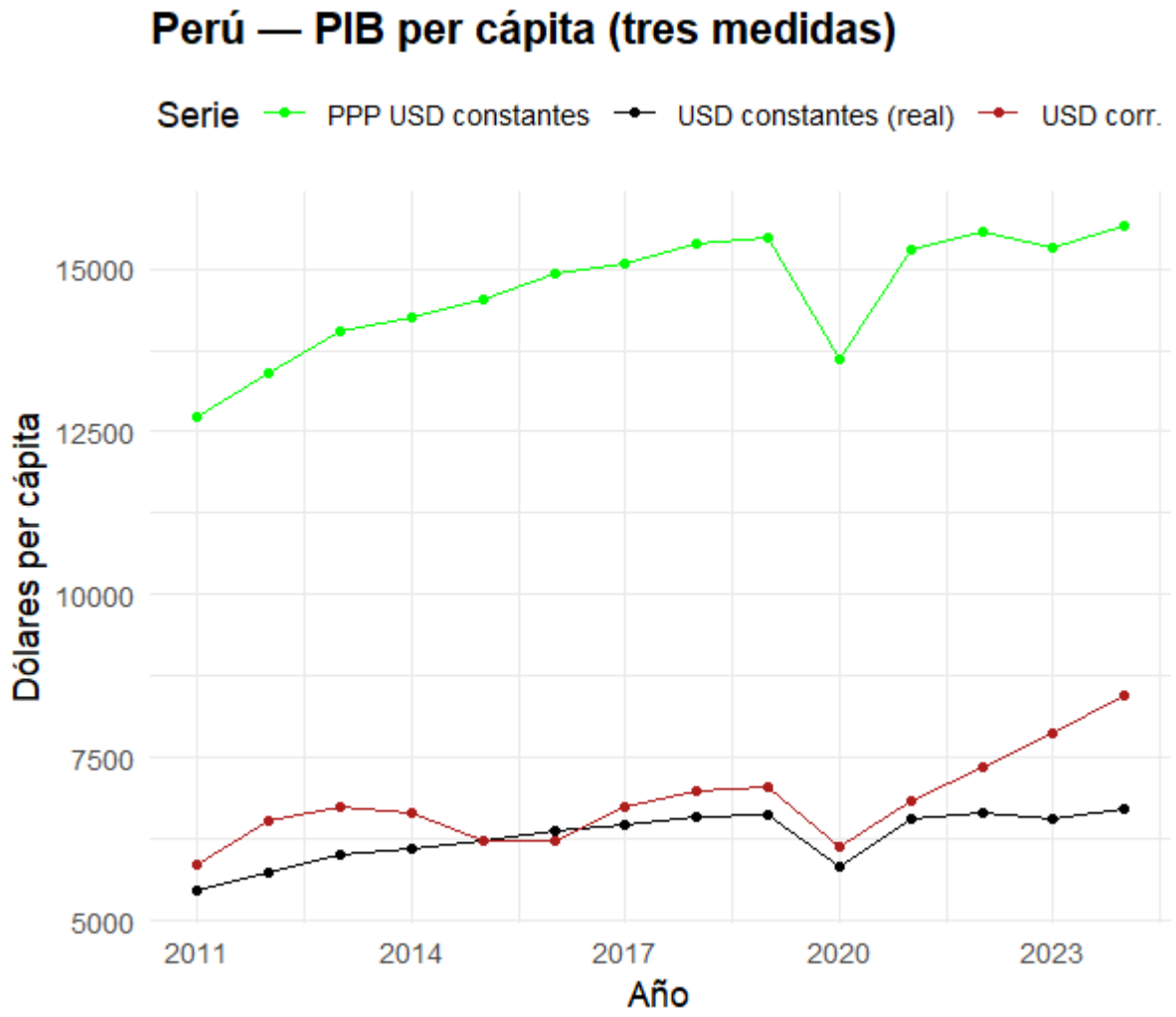
Para este ejercicio buscamos armar un ranking de los países de LATAM considerados en base a distintas medidas del PIB per cápita, para el año 2024. En particular, las tres medidas que usaremos para medir el PIB per cápita serán dólares corrientes, dólares constantes de 2015 y dólares ajustados por paridad del poder de compra (PPP) constantes del 2021. Pedimos y filtramos los datos con el siguiente código:

```
#Bajamos datos de PIB dólares corrientes, dólares constantes y ajustados por PPP, tanto la serie orginal como la per ca
ind <- c(
  gdp_real    = "NY.GDP.MKTP.KD",      # PIB total, USD constantes de 2015
  gdp_pc      = "NY.GDP.PCAP.CD",      # PIB per cápita, USD corrientes
  gdp_pc_ppp_kd = "NY.GDP.PCAP.PP.KD",# PIB PPP per cápita, const 2021 intl$
  pop         = "SP.POP.TOTL",         # Población
  )



datos_pbi <- WDI(country = paises, indicator = ind, start = 1970, end = 2024, extra = TRUE)
View(datos_pbi)

datos_pbi_limpios <- datos_pbi %>%
  drop_na() %>%            # Sacamos filas con NAs
  mutate(
    gdp_pc_real = gdp_real / pop        # Agregamos la columna PIB real per cápita (USD constantes de 2015)
  )

View(datos_pbi_limpios)
```

Creamos la columna de PIB per cápita en dólares constantes usando datos de PIB total en dólares constantes y la columna de población. Nos quedan 253 observaciones luego de filtrar NAs, dentro de las cuales hay datos para 2024 para todos los países analizados. Al momento de rankear los países (nos valemos de la función rank, nativa de R), los resultados son los siguientes:

	country	year	gdp_pc_real	gdp_pc_ppp_kd	gdp_pc	rank_usd_const	rank_usd_ppp_ctes	rank_usd_curr
1	Uruguay	2024	18958.603	32038.773	23906.513	1	1	1
2	Chile	2024	14579.400	30182.787	16709.889	2	2	2
3	Argentina	2024	12667.031	26547.050	13858.204	3	3	4
4	Mexico	2024	10313.487	22033.274	14157.945	4	4	3
5	Brazil	2024	9564.576	19647.910	10280.315	5	5	5
6	Colombia	2024	6873.423	18503.671	7913.988	6	6	7
7	Peru	2024	6711.194	15661.750	8452.372	7	8	6
8	Paraguay	2024	6640.838	16296.284	6416.097	8	7	9
9	Ecuador	2024	5999.394	13935.539	6874.706	9	9	8
10	Bolivia	2024	3226.188	9844.277	4001.211	10	10	10

Un caso curioso es el de Perú. Está séptimo en el ranking medido en USD constantes del 2015, octavo si se mide en dólares PPP constantes de 2021 y sexto si se mide en dólares corrientes. Para tener una idea de cómo pueden diferir las medidas de la misma variable en distintas unidades, graficamos las tres series correspondientes a Perú.



Parte III: Ingreso y su distribución en LATAM

Para esta última parte del ejercicio, queremos ver cómo lucen los datos recientes para ingreso y desigualdad en LATAM. Nos valemos de la serie de PIB per cápita en dólares PPP constantes y de la serie del índice de Gini, computado por el Banco Mundial, para analizar esta cuestión. Tomamos los datos con el código

```r
ind_3 <- c(
  gdp_pc_ppp_kd = "NY.GDP.PCAP.PP.KD",
  gini          = "SI.POV.GINI"
)



datos_desig <- WDI(country = paises, indicator = ind_3,
                   start = 1970, end = 2024, extra = TRUE) %>%
  drop_na()

View(datos_desig)

#Consideremos el año 2023. ¿Cuáles son los países con menor y mayor PIB per cápita? Comparemos sus índices de Gini luego

datos_desig_23 <- filter(datos_desig, year=="2023")
datos_desig_23
```

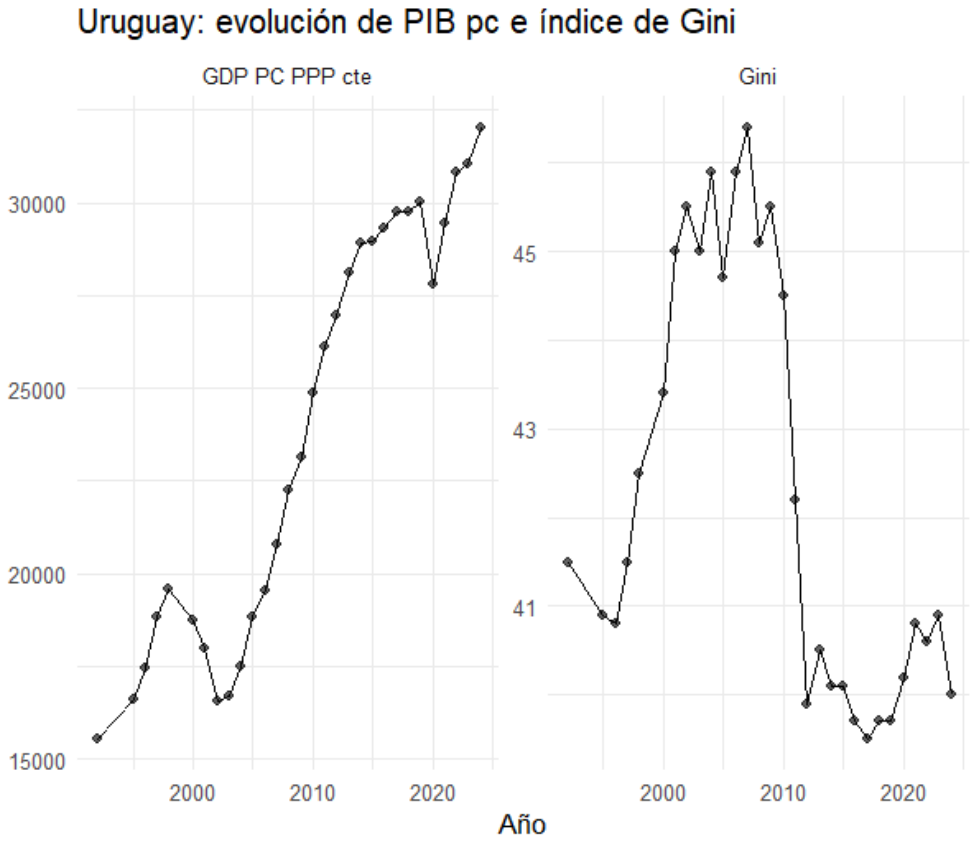
De los países para los cuales tenemos datos (no hay datos para Chile ni México), los resultados lucen así

```
    country iso2c iso3c year status lastupdated gdp_pc_ppp_kd gini                  region      capital longitude  latitude
1 Argentina    AR   ARG 2023         2025-10-07      27104.98 42.4 Latin America & Caribbean Buenos Aires  -58.4173  -34.6118
2   Bolivia    BO   BOL 2023         2025-10-07       9843.97 42.1 Latin America & Caribbean       La Paz  -66.1936  -13.9908
3    Brazil    BR   BRA 2023         2025-10-07      19079.81 51.6 Latin America & Caribbean     Brasilia  -47.9292  -15.7801
4  Colombia    CO   COL 2023         2025-10-07      18383.00 53.9 Latin America & Caribbean       Bogota   -74.082   4.60987
5   Ecuador    EC   ECU 2023         2025-10-07      14343.02 44.6 Latin America & Caribbean        Quito  -78.5243 -0.229498
6  Paraguay    PY   PRY 2023         2025-10-07      15826.09 44.2 Latin America & Caribbean     Asuncion  -57.6362  -25.3005
7      Peru    PE   PER 2023         2025-10-07      15327.58 40.7 Latin America & Caribbean         Lima  -77.0465  -12.0931
8   Uruguay    UY   URY 2023         2025-10-07      31059.25 40.9 Latin America & Caribbean   Montevideo  -56.0675  -34.8941
```

Se evidencia que Uruguay y Bolivia son los países con mayor y menor ingreso en 2023, respectivamente. Lo curioso es que Uruguay además de ser más rico tiene una distribución del ingreso más igualitaria que Bolivia, según el índice de Gini (40,9 para Uruguay contra 42,1 para Bolivia. Recordemos que valores del índice más altos se asocian a mayor desigualdad).

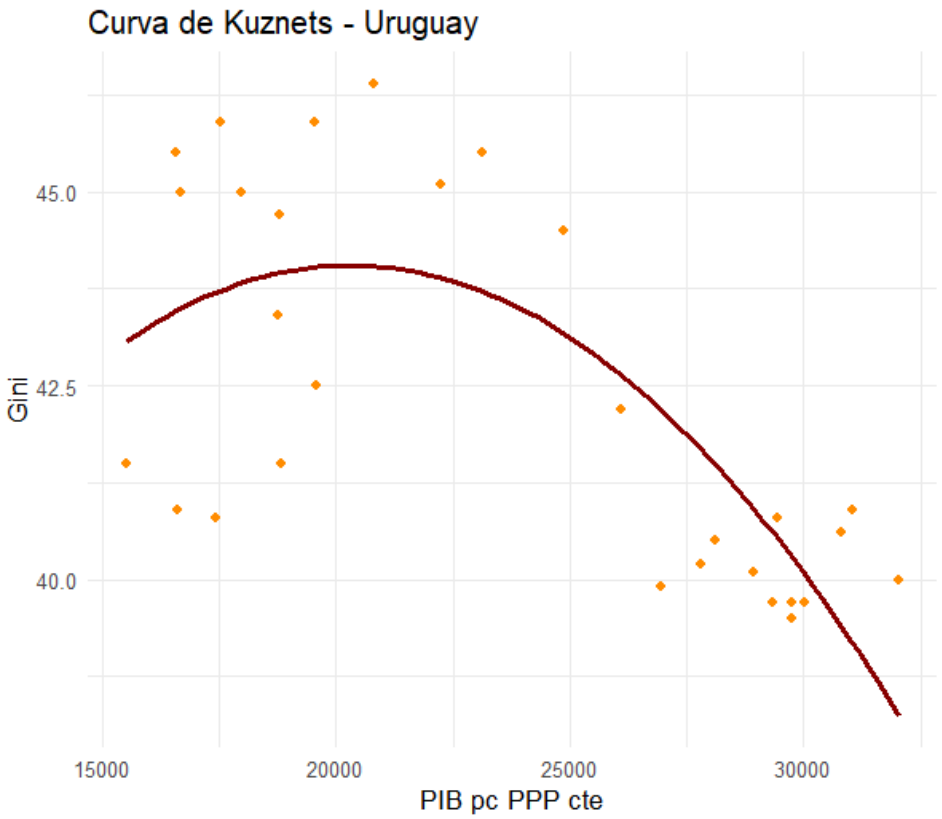
Lo que nos preguntamos ahora es, ¿cómo cambió el ingreso y la desigualdad en Uruguay a lo largo del tiempo? ¿Cómo se compara con la famosa Curva de Kuznets? La Curva de Kuznets es un gráfico de la relación entre ingreso y desigualdad que opera bajo la hipótesis de que los países tienen tanto ingreso como desigualdad creciente al comienzo pero para valores altos del ingreso per cápita, la desigualdad comienza caer.

Usando la misma lógica de pivotar el data frame para usar facet wrap, logramos graficar las dos series en simultáneo y obtenemos lo siguiente:



Lo que nos interesa hacer ahora es un scatter plot sobre datos de desigualdad e ingreso para Uruguay y fittear una parábola para ver cómo luce. Usamos el siguiente código.

```r
datos_desig %>%
  filter(country == "Uruguay") %>%
  ggplot(aes(x = gdp_pc_ppp_kd, y = gini)) +
  geom_point(color = "darkorange") +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), color = "darkred", se=FALSE) +
  labs(title = "Curva de Kuznets - Uruguay",
       x = "PIB pc PPP cte", y = "Gini") +
  theme_minimal()
```



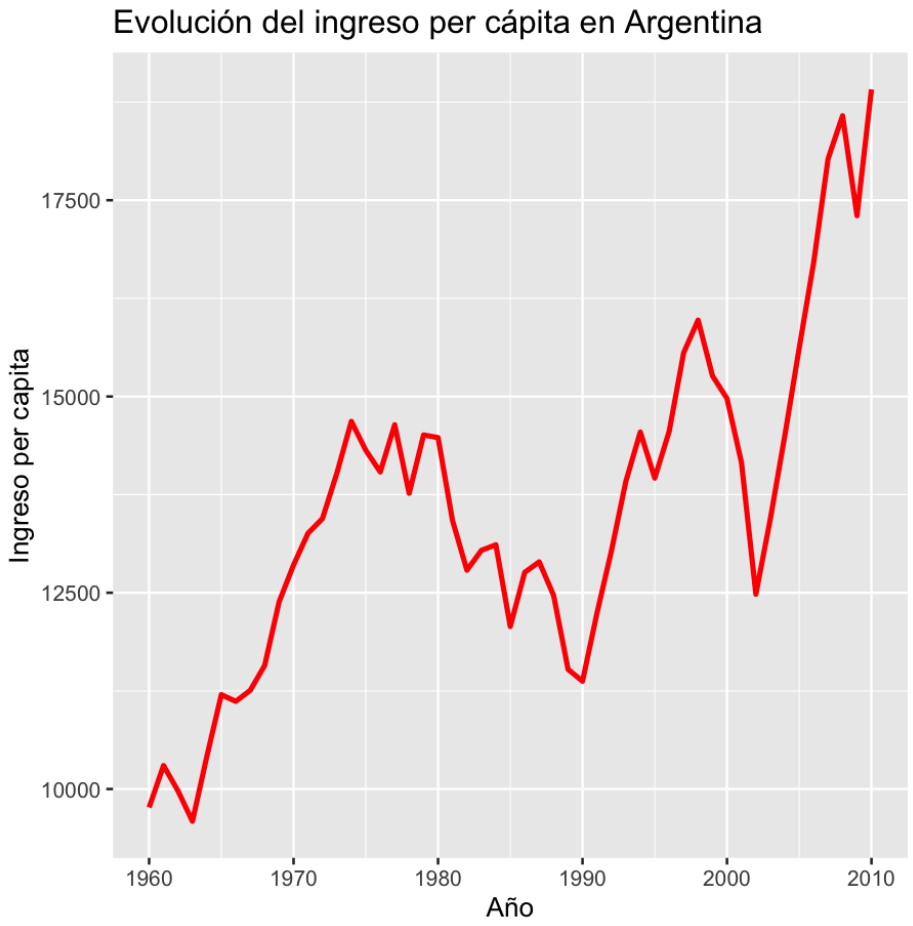
Vale aclarar que es importante tomar noción de las escalas. En conclusión, en este ejercicio logramos familiarizarnos con datos respecto a inflación, desempleo, ingreso

y su distribución para países de LATAM, aprovechando las herramientas de manejo de datos y gráficos de R.

Ejercicio 2: Análisis econométrico con datos de gapminder

Parte I: Ingreso por persona

1) Primero descargamos la base de datos de gapminder y luego filtramos el dataset por los datos de Argentina, lo guardamos en una variable nueva y graficamos la evolución temporal de la variable income per person en Argentina. Obtenemos el siguiente gráfico:



El gráfico contiene mucha historia económica Argentina. Desde los años 60 podemos observar una tendencia alcista hasta mediados de los 70 coincidiendo aproximadamente con el comienzo de la última dictadura en Argentina (1976). Luego, notamos una caída más abrupta que coincide con la época hiperinflacionaria del gobierno de Alfonsín. Después volvemos a ver una recuperación durante la época de la convertibilidad superando levemente el techo pre dictadura militar. Para los años 2000 vuelve a caer de manera agresiva, coincidiendo con el corralito y la crisis del 2001. Por último, vemos la recuperación durante el boom de las commodities en el gobierno de Nestor Kichner que continuó con el primer mandato del gobierno de Cristina. Pero ya a partir del 2010 vemos un estancamiento que se conoce en la literatura como la década perdida.

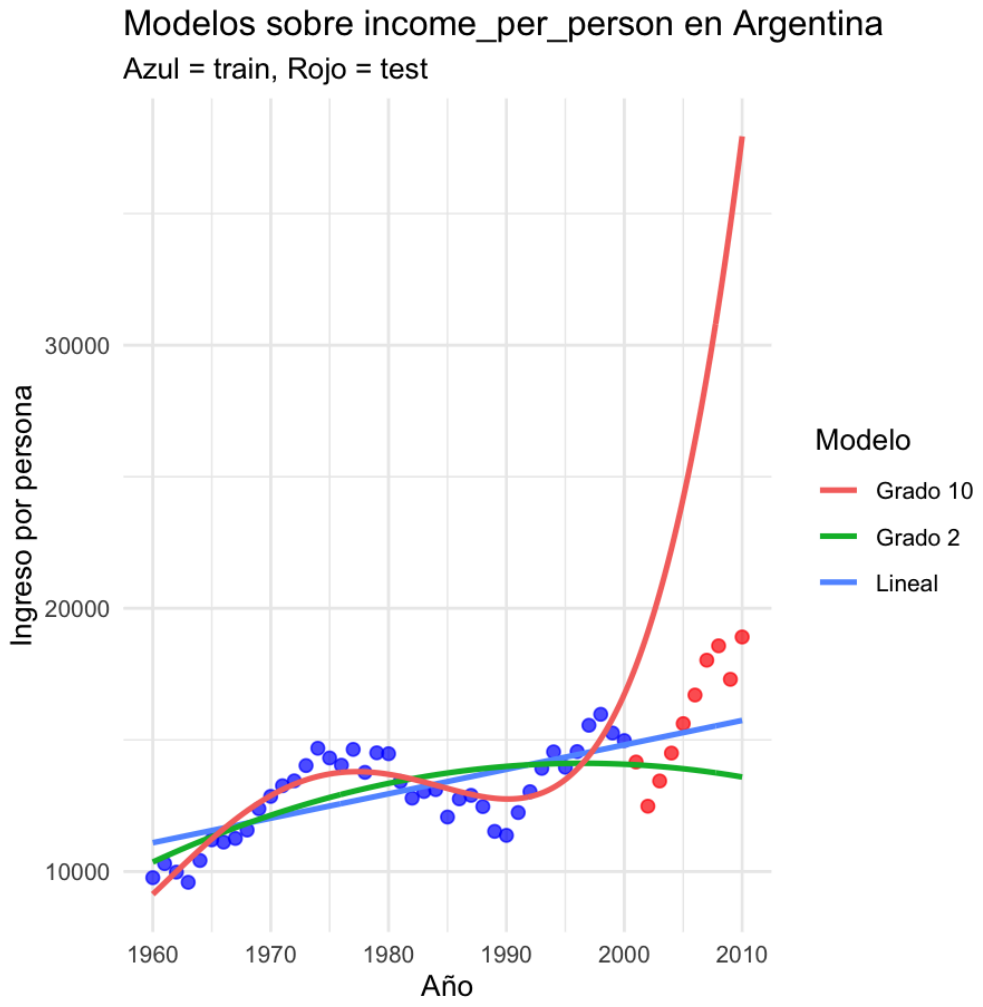
2) Para este ejercicio, se seleccionaron los datos de Argentina y se dividió la muestra en dos subconjuntos:

Entrenamiento (train), abarca todos los años excepto los últimos 10. Testeo (test), abarca los últimos 10 años. Este último subconjunto es utilizado para evaluar el desempeño de los modelos fuera de la muestra (train).

Se ajustaron tres modelos de regresión utilizando el tiempo como variable independiente:

```
# Modelo lineal
fit_lineal <- lm(y_train ~ x_train)
summary (fit_lineal)

# Modelo polinómico grado 2
fit_poly2 <- lm(y_train ~ poly(x_train, 2, raw = TRUE))
summary(fit_poly2)

# Modelo polinómico grado 10
fit_poly10 <- lm(y_train ~ poly(x_train, 10, raw = TRUE))
summary(fit_poly10)
```

Luego, obtenemos las predicciones de los distintos modelos y generamos un gráfico que contiene los puntos reales (datos de train y test) y las líneas de las distintas regresiones para observar cómo estos modelos logran predecir los datos verdaderos



A simple vista, se puede observar que el modelo lineal no logra capturar la no linealidad de los datos, pero es más estable. Por otro lado, el modelo de grado 2 sigue mejor las curvas generales, pero tiene un ajuste pobre en los extremos. Por último, el modelo de grado 10 ajusta perfectamente el conjunto de entrenamiento, pero exagera los cambios (overfitting) y falla al generalizar sobre el testeo (los

puntos rojos).

Esto último lo podemos concluir al calcular la raíz del error cuadrático medio (RMSE en inglés) que calcula la diferencia entre los datos reales y sus predicciones. Cuanto más pequeño sea el valor del RMSE, mejor el modelo.

Calculamos el RMSE tanto para train y test y obtenemos los siguientes resultados:

```
RMSE en TRAIN:
   Lineal    Grado2   Grado10
1212.7475 1159.9654  730.6231
> cat("\nRMSE en TEST:\n"); print(rmse_test)

RMSE en TEST:
   Lineal    Grado2   Grado10
 2012.472  3114.474 11311.216
```

Como podemos observar en los resultados del RMSE, el modelo de grado 10 tiene el menor error dentro de la muestra (train), pero su desempeño cae drásticamente en el test (más de 11.000 puntos de error). Esto se debe a la existencia de sobreajuste (overfitting). El modelo se adapta demasiado a los datos de entrenamiento, y falla en predecir fuera de la muestra (test). Esto muestra que el modelo de grado 10 no generaliza bien.

Por el contrario, en el modelo lineal observamos que obtiene un menor RMSE dentro de la muestra de test. Esto muestra que este modelo tiene el mejor balance entre ajuste y generalización, aunque puede llegar a ser muy simple. El modelo de grado 2 es un punto intermedio, con bajo error en entrenamiento pero aumento considerable en test.

3)

a. Estimamos la matriz de correlación entre los ingresos per cápita (income_per_person) para cinco países sudamericanos: Argentina, Bolivia, Brasil, Chile y Uruguay.

Para hacer esto, filtramos el dataset para que contenga los 5 países seleccionados. luego utilizamos el comando pivot_wider para transformar la tabla de formato largo a formato ancho para que pueda calcularse correctamente la correlación. Por último, con el comando Cor obtenemos los siguientes resultados:

	Argentina	Bolivia	Brazil	Chile	Uruguay
Argentina	1.0000000	0.9244884	0.7951110	0.7650413	0.8291831
Bolivia	0.9244884	1.0000000	0.8445418	0.7450691	0.7985954
Brazil	0.7951110	0.8445418	1.0000000	0.7717164	0.8713498
Chile	0.7650413	0.7450691	0.7717164	1.0000000	0.9407932
Uruguay	0.8291831	0.7985954	0.8713498	0.9407932	1.0000000

La matriz muestra correlaciones positivas y altas entre todos los países. Los valores van desde 0.74 hasta 0.94, lo que indica que los niveles de ingreso entre estos países se mueven de manera muy similar en el largo plazo.

Esto es esperable, ya que los niveles de ingreso tienden a evolucionar en paralelo debido a factores estructurales comunes en la región.

b. Para este ejercicio, calculamos las variaciones interanuales, utilizando el comando (.-lag(.)) / lag(.) que calcula las variaciones del crecimiento = $\frac{y_t - y_{t-1}}{y_{t-1}}$. El comando lag toma el valor del periodo anterior. Al producir NAs en la primera columna, los descartamos con el comando na.omit(). Obtenemos la siguiente matriz de tasa de crecimiento per cápita.

	Argentina	Bolivia	Brazil	Chile	Uruguay
Argentina	1.0000000	0.2066589	0.272052233	0.169198858	0.5127562
Bolivia	0.2066589	1.0000000	0.263161872	0.134951601	0.2653272
Brazil	0.2720522	0.2631619	1.000000000	0.006129268	0.2779881
Chile	0.1691989	0.1349516	0.006129268	1.000000000	0.3655066
Uruguay	0.5127562	0.2653272	0.277988119	0.365506633	1.0000000

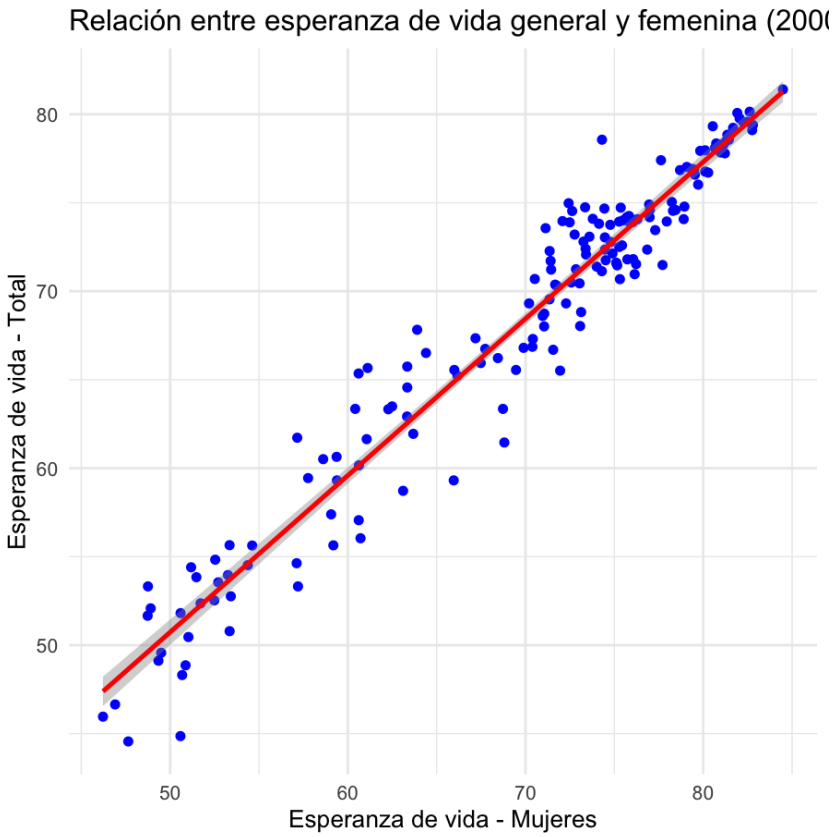
Observamos algo distinto a la matriz anterior. Las correlaciones son mucho más bajas, en general menores a 0.30, con algunas incluso cercanas a cero (por ejemplo, Chile-Brasil = 0.0061).

Este resultado refleja que, aunque los países tienden a compartir una tendencia de largo plazo en los niveles de ingreso, las fluctuaciones año a año no necesariamente ocurren al mismo tiempo ni en la misma magnitud debido a la presencia de shocks, inherentes al país en particular. Esta diferencia se debe a la importancia de distinguir entre análisis de nivel y de crecimiento, ya que la dinámica de corto plazo no siempre sigue la misma lógica que las tendencias de largo plazo.

Parte II: Esperanza de vida y género

5) Para esta parte elegimos analizar el año 2000. Filtramos el dataset a partir de ese año y nos aseguramos que las variables que nos interesen este en formato numérico, a partir del código as.numeric

Graficamos life expectancy contra life expectancy female y obtenemos el siguiente resultado:



Se observa una relación fuertemente lineal y positiva entre la esperanza de vida general y la femenina en el año 2000. Esto indica que los países con mayor esperanza de vida femenina también presentan mayores valores de esperanza de vida total, durante el año 2000. Esto tiene sentido ya que si aumenta la esperanza de vida de un sector (las mujeres) lógicamente debería aumentar la esperanza de vida general.

6) Corremos una regresión lineal simple entre life expectancy y life expectancy female a partir del comando lm. Obtenemos el siguiente resumen de la regresión con el comando summary:

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.44521    1.16364   5.539 1.18e-07 ***
life_expectancy_female  0.88573    0.01656  53.497  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.155 on 165 degrees of freedom
Multiple R-squared:  0.9455,    Adjusted R-squared:  0.9452
F-statistic:  2862 on 1 and 165 DF,  p-value: < 2.2e-16
```

El coeficiente estimado para life_expectancy_female fue aproximadamente 0.88, lo que indica que, por cada año adicional de esperanza de vida femenina, la esperanza de vida general aumenta en promedio 0.88 años. Esto como dijimos antes tiene sentido ya que si aumenta la esperanza femenina, la esperanza de vida general también deberia aumentar y en una magnitud alta.

El R^2 ajustado es cercano a 0.94, lo que implica que la variable explicativa life expectancy female explica el 94% de la variabilidad de la esperanza de vida general. Este resultado confirma la fuerte relación entre ambas variables, como se pudo observar en el punto 5.

7) Agregamos una nueva variable a nuestro dataset que sea la diferencia entre life_expectancy_female y life_expectancy a partir del comando mutate

Realizamos un Test t de una muestra sobre la diferencia, donde H0 implica que life_expectancy_female y life_expectancy son iguales y H1 implica que life_expectancy_female es mayor

Utilizamos el comando t.test y obtenemos:

```
        One Sample t-test

data:  gapminder_2000$diferencia
t = 7.9642, df = 166, p-value = 1.254e-13
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 1.191145      Inf
sample estimates:
mean of x
 1.503383
```

Como el p valor es cercano a cero, rechazamos H0 para cualquier nivel de significatividad relevante por lo tanto hay fuerte evidencia a favor de que life_expectancy_female es mayor que la esperanza de vida total. Esto coincide con evidencia empírica que concluye que las mujeres viven más que los hombres.

8)

Estimamos life expectancy sobre life expectancy female y income per person. con el comando summary obtenemos:

Modelo múltiple:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.762e+00  1.303e+00   5.191 6.11e-07 ***
life_expectancy_female 8.800e-01  1.968e-02  44.724  < 2e-16 ***
income_per_person       5.806e-06  1.063e-05   0.546    0.586
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.16 on 164 degrees of freedom
Multiple R-squared:  0.9456,    Adjusted R-squared:  0.9449
F-statistic:  1425 on 2 and 164 DF,  p-value: < 2.2e-16
```

y recordamos los resultados del modelo simple:

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.44521    1.16364   5.539 1.18e-07 ***
life_expectancy_female  0.88573    0.01656  53.497  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.155 on 165 degrees of freedom
Multiple R-squared:  0.9455,    Adjusted R-squared:  0.9452
F-statistic:  2862 on 1 and 165 DF,  p-value: < 2.2e-16
```

Analizando brevemente el modelo de regresión simple variable notamos que el coeficiente de life expectancy female tiene un valor de 0.88573, altamente significativo, lo que implica que por cada año adicional de esperanza de vida femenina, la esperanza de vida general aumenta en aproximadamente 0.89 años. Este modelo presenta un R^2 ajustado de 0.9452, lo cual indica que life expectancy female explica el 94.52% de la variabilidad en la esperanza de vida.

Posteriormente, regresamos life expectancy incorporando income per person al modelo, generando así una regresión múltiple. En este segundo modelo, el coeficiente de life expectancy female se mantiene prácticamente igual (0.88) y continúa siendo altamente significativo. Sin embargo, el coeficiente asociado a income per person es de apenas 5.81e-06, con un valor p de 0.586, lo cual evidencia que no es estadísticamente significativo al 5%. Además, el valor del R^2 ajustado disminuye levemente de 0.9452 (modelo simple) a 0.9449 (modelo múltiple), indicando que la adición de esta nueva variable no mejora el poder explicativo del modelo. Esto podria no necesariamente ser asi para otros años, pero particularmente, para el año 2000, no vale la pena incluir la variable income per person

9)

Elegimos explicar life expectancy a traves de child_mortality, children_per_woman y life_expectancy_male, que inferimos que podrían ser variables que expliquen la esperanza de vida general.

Obtenemos el siguiente resumen de la regresión:

```
modelo_triple <- lm(life_expectancy ~ life_expectancy_male + child_mortality + children_per_woman, data = gapminder_2000)
summary(modelo_triple)
```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          15.926023   3.042403   5.235 5.03e-07 ***
life_expectancy_male  0.842499   0.041683  20.212  < 2e-16 ***
child_mortality      -0.005137   0.009462  -0.543  0.58790
children_per_woman   -0.612477   0.192782  -3.177  0.00178 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.094 on 163 degrees of freedom
Multiple R-squared:  0.9492,    Adjusted R-squared:  0.9482
F-statistic:  1014 on 3 and 163 DF,  p-value: < 2.2e-16
```

Analizando life expectancy male, notamos que su coeficiente estimado fue de 0.8425, altamente significativo ($p \approx 0$), lo que indica que, manteniendo constantes las demás variables, un aumento de un año en la esperanza de vida masculina se asocia con un incremento de aproximadamente 0.84 años en la esperanza de vida total. Por lo tanto, concluimos que esta variable es la principal predictora del modelo. La segunda variable considerada fue children per woman, una medida de fertilidad que refleja el promedio de hijos por mujer. Esta variable presentó un coeficiente negativo significativo (-0.6125, $p = 0.00178$), lo que sugiere que mayores tasas de fertilidad están asociadas con una menor esperanza de vida. Este resultado es

consistente con la evidencia empírica, que vincula altas tasas de natalidad con condiciones de desarrollo menos avanzadas y sistemas de salud más débiles.

Por último, la tercera variable fue child mortality, que resultó tener un coeficiente -0.005 estadísticamente no significativo (p=0.58) que podría deberse a la relación estrecha de esta variable con las otras utilizadas en el modelo. Por lo tanto, podríamos descartarla y haber utilizado otra variable que sí explique la esperanza de vida general.

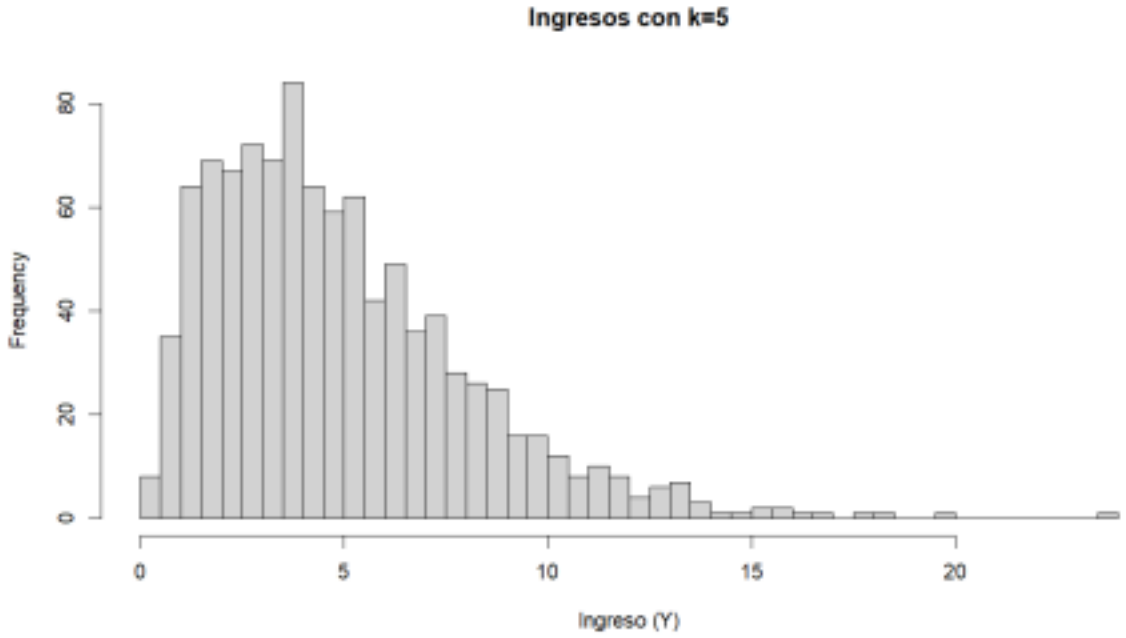
En cuanto al R^2 ajustado, notamos un valor de 0.9482 lo que significa que gran parte de la variabilidad de la variable dependiente es explicada por las variables explicativas incluidas en el modelo, por lo tanto es un modelo muy preciso.

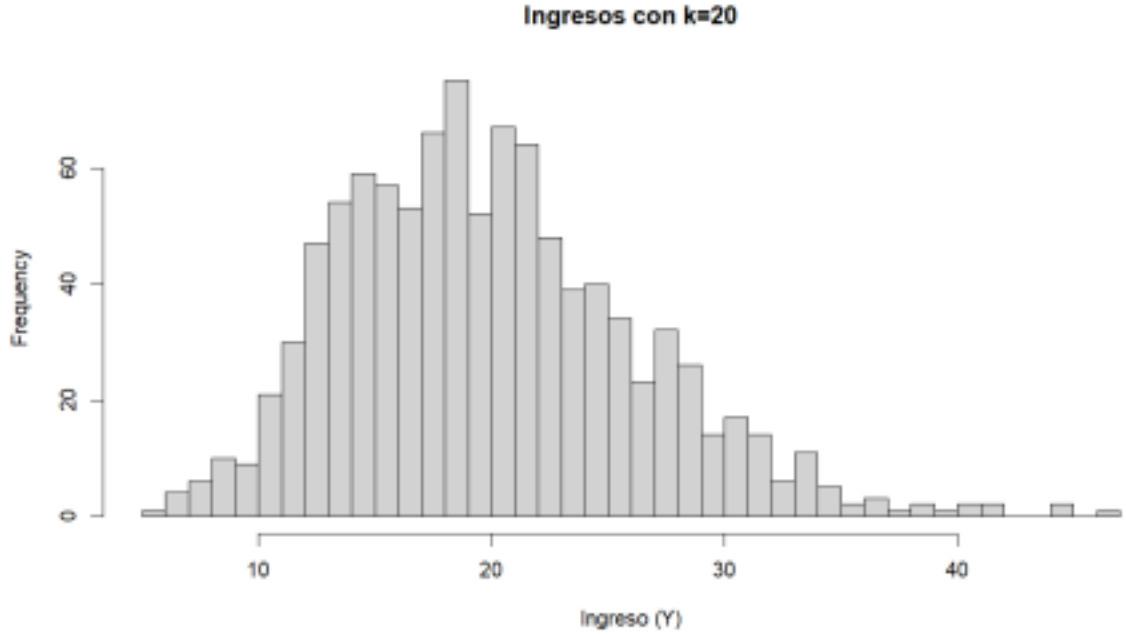
Ejercicio 3 - simulación 1: Demanda con preferencias Cobb–Douglas

Inciso 1: Ingreso por persona

Para esta parte, definimos simular_ingreso como una función de el número de personas, definido como n, y de el grado de libertad de la distribución chi-cuadrado, definido como k.

Ahora veamos qué efecto tiene k en la distribución del ingreso. Para ello primero establecemos una seed (set.seed(42)) para que los análisis sean siempre los mismos, y luego generamos dos simulaciones con la función previamente definida con 2 muestras distintas, una con k=5 y otra con k=20 (las 2 muestras tienen 1000 observaciones). Ahora graficamos los histogramas de dichas simulaciones.





Al aumentar k, hay 2 efectos principales: por un lado vemos que el histograma está más “distribuido parejamente” lo cual es indicio de una mayor varianza; por otra parte, si nos fijamos en los valores del eje x, podemos ver que la media ha aumentado considerablemente. Estas intuiciones encuentran fundamento en que la esperanza y la varianza de la chi-cuadrado están relacionadas linealmente con k, sus grados de libertad. La esperanza en este caso vale k, y la varianza 2k.

Inciso 2: Demanda óptima con preferencias Cobb-Douglas Primero, definimos

demanda_cd como una función de los parámetros Y, p1, p2, y de los alphas.

```
#Función demanda_cd
demanda_cd <- function(Y, p1, p2, alpha1, alpha2) {
```

Luego, dentro de la función, definimos las demandas óptimas que resultan del problema de optimización del consumidor.

```
  #defino las demandas optimas primero
  x1_optimo <- (alpha1 * Y) / p1
  x2_optimo <- (alpha2 * Y) / p2
```

Ahora, definimos el vector de demandas y la utilidad indirecta. Para que nos devuelva los valores usamos return y creamos una lista con las demandas óptimas previamente definidas y con la utilidad directa

```
  #las agrupo en un vector
  demandas_vector <- c(x1 = x1_optimo, x2 = x2_optimo)
  utilidad_indirecta <- (x1_optimo^alpha1) * (x2_optimo^alpha2)
  return(list(
    demandas = demandas_vector,
    utilidad_indirecta = utilidad_indirecta
  ))
}
```

Inciso 3: Simulación base

En este ejercicio utilizamos la función simular_ingreso del punto anterior para generar los ingresos de 10000 hogares. Para ello mantenemos la semilla fija (42) y elegimos k = 10 como grado de libertad, ya que esto nos da un nivel de ingresos promedio razonable con cierta variabilidad entre personas. Definimos precios fijos p1 = 1.5 y p2 = 2.5, y alphas iguales para todos los individuos (alpha1 = 0.4 y alpha2 = 0.6). Por último definimos Y_simulados para después usarlos en las demandas óptimas.

```r
#punto 3
set.seed(42) # fijamos los muestreos aleatorios:

Hogares <- 10000 # Número de hogares
grados <- 10  # Grados de libertad para el Ingreso (Y)
p1_fijo <- 1.5
p2_fijo <- 2.5
alpha1_fijo <- 0.4
alpha2_fijo <- 0.6


Y_simulados <- simular_ingreso(n = Hogares, k = grados)
```

Luego calculamos la demanda óptima de cada bien aplicando las fórmulas del ejercicio 2. Con esto obtenemos, para cada hogar, los valores de x1*, x2* y también la utilidad asociada a esa elección.

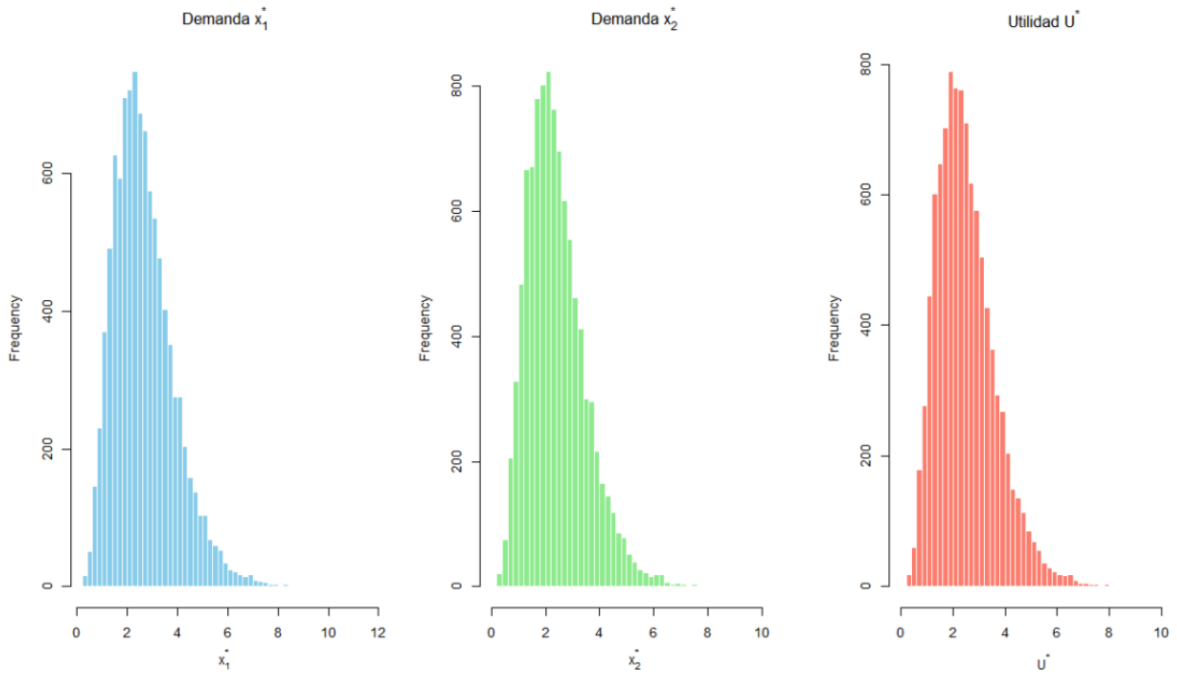
```r
#ahora armo vectores con los x1*,x2* y de utilidad indirecta
x1_optimo_vec <- (alpha1_fijo * Y_simulados) / p1_fijo
x2_optimo_vec <- (alpha2_fijo * Y_simulados) / p2_fijo
U_star_vec <- (x1_optimo_vec^alpha1_fijo) * (x2_optimo_vec^alpha2_fijo)
```

A continuación armamos un data frame con toda la información simulada y calculamos medidas descriptivas como medias y cuartiles para cada variable de interés.

```r
# 5. los junto todo en un data frame
simulacion_df <- data.frame(
  Y = Y_simulados,
  x1_opt = x1_optimo_vec,
  x2_opt = x2_optimo_vec,
  U_star = U_star_vec
)
# Calculamos medias y cuartiles sobre el data frame resultante
medias <- sapply(simulacion_df[, c("x1_opt", "x2_opt", "U_star")], mean)
cuartiles <- sapply(simulacion_df[, c("x1_opt", "x2_opt", "U_star")], quantile)

print(medias)
print(cuartiles)
```

Ahora graficamos la distribución empírica de las demandas y de la utilidad indirecta con 3 histogramas. A continuación presentamos el código y posteriormente los 3 histogramas.

```r
# VISUALIZACIÓN DE DISTRIBUCIONES (HISTOGRAMAS)

par(mfrow = c(1, 3)) # Configura 3 gráficos en una fila

hist(simulacion_df$x1_opt, breaks = 50, col = "skyblue",
     main = expression(paste("Demanda ", x[1]^"*")),
     xlab = expression(x[1]^"*"), border = "white")

hist(simulacion_df$x2_opt, breaks = 50, col = "lightgreen",
     main = expression(paste("Demanda ", x[2]^"*")),
     xlab = expression(x[2]^"*"), border = "white")

hist(simulacion_df$U_star, breaks = 50, col = "salmon",
     main = expression(paste("Utilidad ", U^"*")),
     xlab = expression(U^"*"), border = "white")
```



Se observa que tanto las demandas como la utilidad presentan una distribución parecida a la del ingreso: hay muchos hogares con valores bajos y una cola más larga hacia la derecha.

En resumen, en este punto confirmamos que la simulación se comporta como era de esperar: los hogares con mayor ingreso consumen más, y la forma general de las distribuciones coincide con la de los ingresos generados en el primer paso.

Inciso 4: probabilidad de bajo consumo en un bien

En este punto analizamos cuántos hogares consumen “poco” de cada bien. Para

eso fijamos c = 2 y contamos qué proporción de la población simulada tiene una demanda menor a ese valor. Calculamos el porcentaje para los 2 bienes.

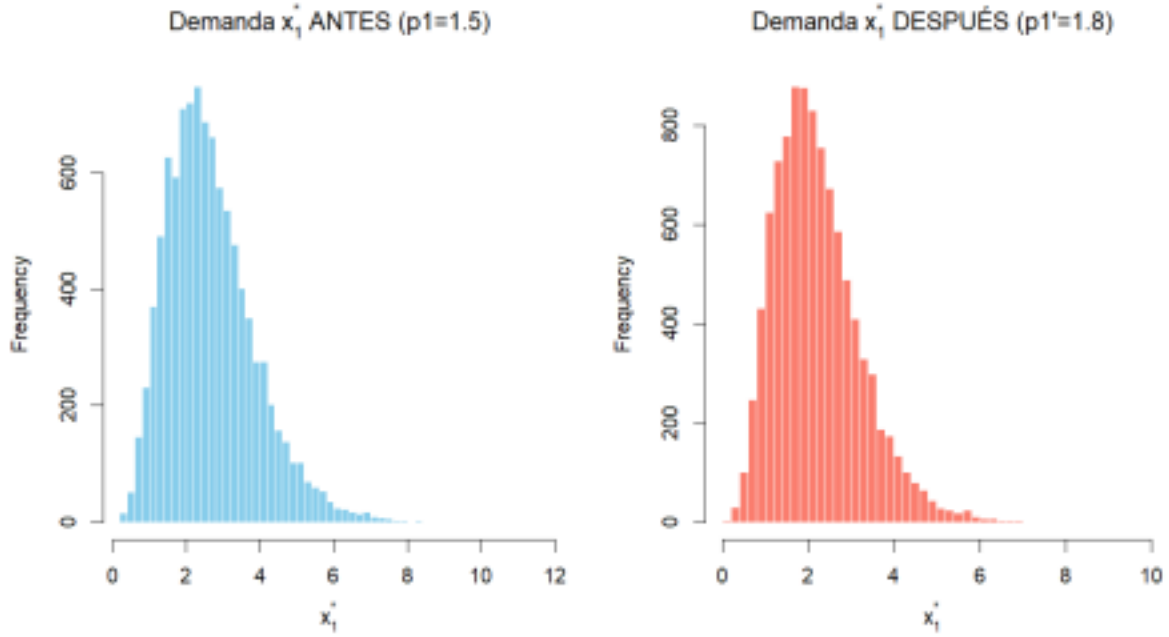
```r
#' Función prob_bajo_consumo

prob_bajo_consumo <- function(simulacion_df, c, j) {
    columna_demanda <- if (j == 1) "x1_opt" else "x2_opt" #que demanda queremos
    demanda_j <- simulacion_df[[columna_demanda]]
    conteo_bajo_consumo <- sum(demanda_j < c)    #sumamos los valores que cumplen esta condicion
    probabilidad <- conteo_bajo_consumo / nrow(simulacion_df)  #calculamos la proba

  return(probabilidad)
}

# Probabilidad de que el consumo de Bien 1 sea menor que 2 (P(x1* < 2))
c_umbral <- 2
prob_x1_bajo <- prob_bajo_consumo(simulacion_df, c = c_umbral, j = 1)

print(prob_x1_bajo)

# Probabilidad de que el consumo de Bien 2 sea menor que 2 (P(x2* < 2))
prob_x2_bajo <- prob_bajo_consumo(simulacion_df, c = c_umbral, j = 2)
print(prob_x2_bajo)
```

Los resultados muestran que el bien 2 tiene una mayor cantidad de personas por debajo del umbral. Esto es razonable porque, aunque ambos bienes se consumen en proporciones fijas del ingreso, el bien 2 tiene un precio más alto que el bien 1, por lo que automáticamente la cantidad comprada termina siendo menor.

Inciso 5: shock de precios

Recalculamos las demandas óptimas con el nuevo precio y las comparamos con las demandas originales.

```r
# 1. Definir el nuevo precio (p1')
shock_precio <- 0.20 # Aumento del 20%
p1_shock <- p1_fijo * (1 + shock_precio)

#Demanda del Bien 1 después del shock: x1*' = (alpha1 * Y) / p1'
x1_opt_despues <- (alpha1_fijo * Y_simulados) / p1_shock

#Demanda del Bien 2 después del shock: x2*' (p2 y alpha2 no cambian)/ propiedad de las cobb douglas
x2_opt_despues <- (alpha2_fijo * Y_simulados) / p2_fijo
# Nota: La utilidad indirecta U* también cambia, pero no nos piden nada al respecto (todo pelota)
```

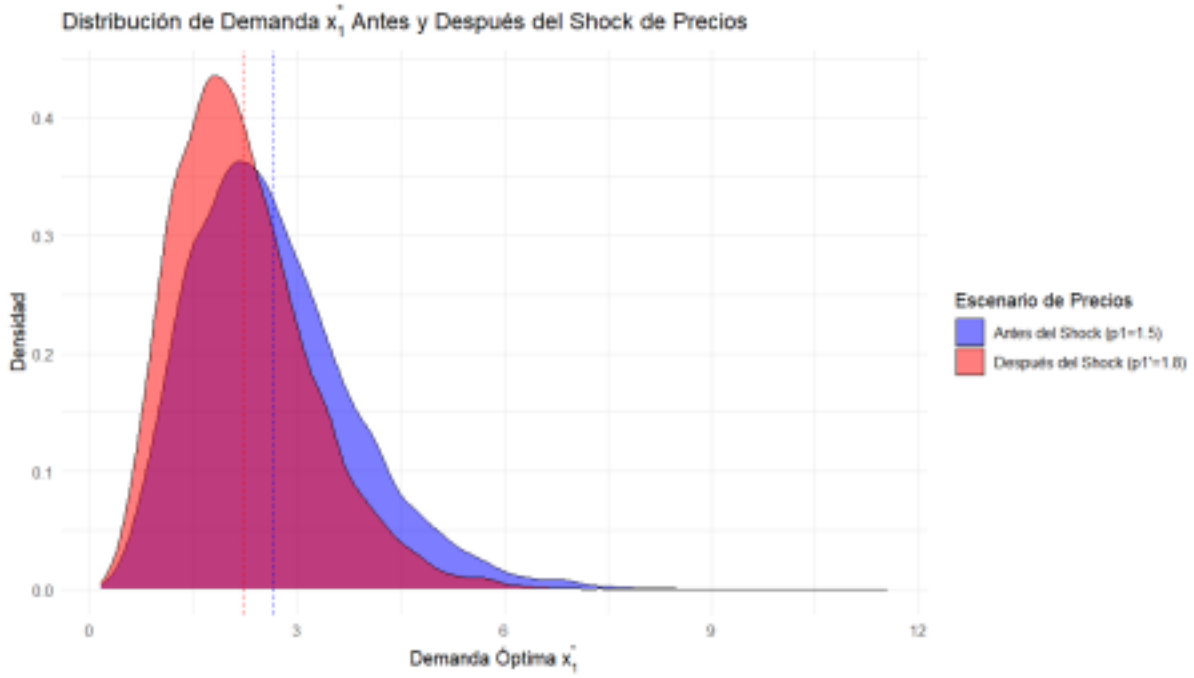


Lo primero que se observa al comparar ambas versiones es que la demanda del bien 1 cae de manera automática para todos los hogares. Esto ocurre porque el modelo Cobb-Douglas asigna una proporción fija del ingreso a cada bien, por lo que cuando el precio aumenta, esa misma proporción se reparte sobre un bien más caro y la cantidad que se puede comprar con ese ingreso es menor.

Además de calcular la nueva media del consumo, comparamos los cuartiles para entender cómo cambia la distribución general. Se ve que no solo baja el valor promedio, sino que también se reduce el nivel de consumo en todos los grupos de la población.

Inciso 6: visualización comparada

En este último punto superponemos las distribuciones de demanda del bien 1 antes y después del aumento de precio, para visualizar claramente el impacto del shock. Para eso armamos un único gráfico con ambas densidades y marcamos con líneas verticales las medias de cada escenario.



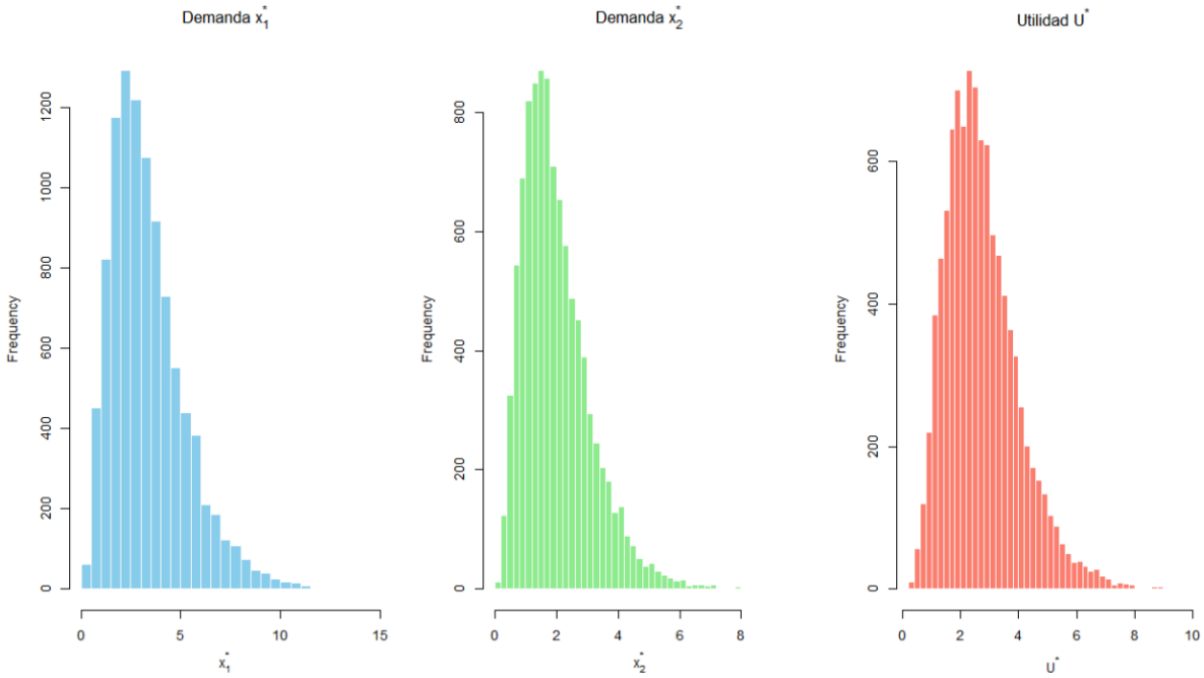
Lo que se observa es un corrimiento completo de la distribución hacia la izquierda. Entonces la caída no se da sólo en promedio sino en todos los niveles de ingreso. Incluso los hogares que antes consumían valores relativamente altos ahora se ven desplazados a niveles más bajos.

Además, la distribución posterior al shock se vuelve levemente más concentrada, lo que sugiere que la diferencia de consumo entre los hogares también se reduce.

Inciso 7: heterogeneidad en preferencias

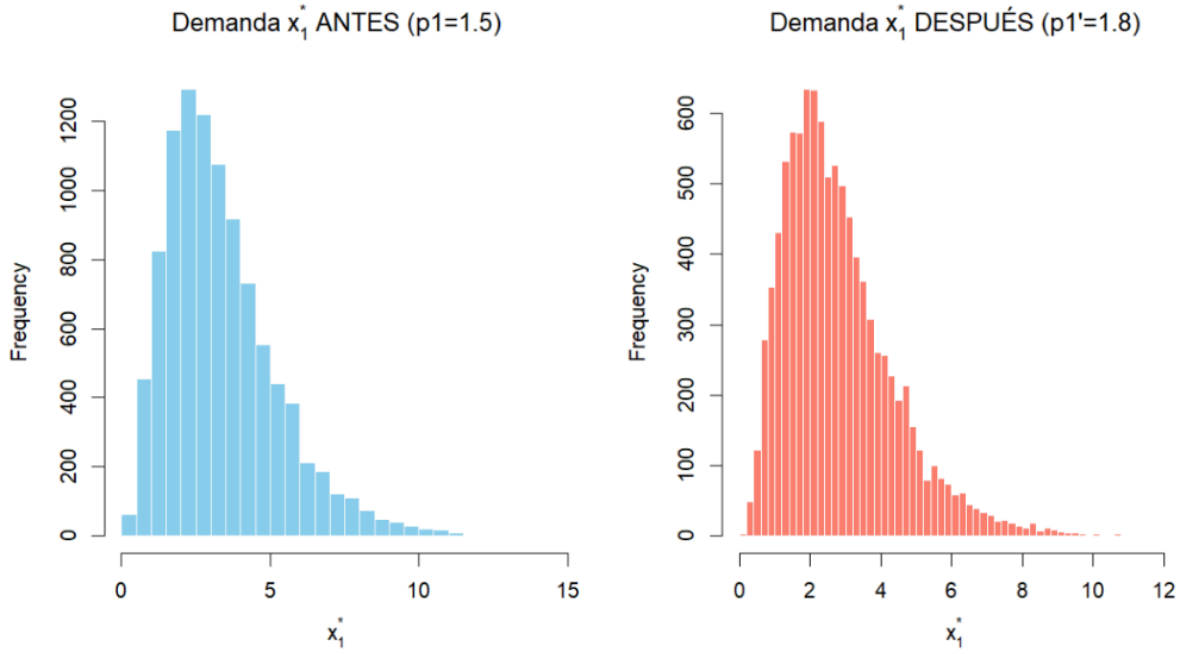
Incorporamos la heterogeneidad en las preferencias, para ello definimos alpha1_heterogeneo usando la distribución beta, análogamente definimos alpha2_heterogeneo como 1 - alpha1_heterogeneo. Después graficamos los mismos histogramas que en el ejercicio 3, solo que con las nuevas demandas.

```
# Parámetros Beta para la Heterogeneidad
a_beta <- 5 # Forma 1
b_beta <- 5 # Forma 2

# Generación de la Heterogeneidad
alpha1_heterogeneo <- rbeta(n = Hogares, shape1 = a_beta, shape2 = b_beta)
alpha2_heterogeneo <- 1 - alpha1_heterogeneo

# Generación de Ingresos
Y_simulados <- simular_ingreso(n = Hogares, k = grados)
```



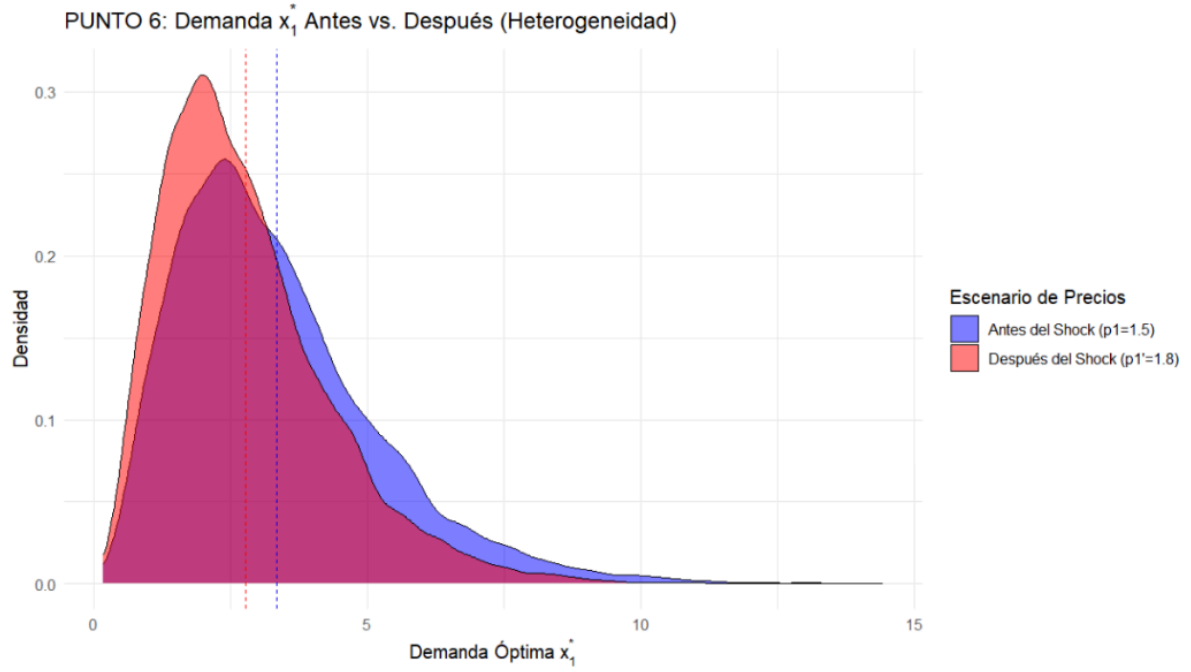
A diferencia del caso homogéneo, las distribuciones ahora presentan una dispersión mayor. Por ejemplo, en la demanda del bien 1 se muestra una distribución más ancha, con colas más largas. La demanda del bien 2 muestra el comportamiento opuesto: los hogares que no priorizan el bien 1 ahora destinan casi todo su ingreso al bien 2, lo cual produce una concentración fuerte en valores altos. En consecuencia, tanto la dispersión como la asimetría aumentan respecto al escenario con preferencias fijas.

Analizamos nuevamente el impacto del aumento del precio del bien 1. Es decir, repetimos el mismo shock del 20%, pero ahora los hogares no reaccionan de forma idéntica como antes, sino que la respuesta depende de cuánto valoran cada bien.



Lo que observamos es que el efecto del shock sigue siendo negativo en términos generales, pero ya no es homogéneo: los hogares con una consideración por el bien 1 relativamente alta (es decir, aquellos que asignaban mayor proporción de su ingreso al bien 1) son los más afectados. En cambio, quienes inicialmente consumían más del bien 2 modifican poco su demanda del bien 1, ya que su consumo ya era bajo antes del shock. Esto se refleja en la distribución: además del corrimiento hacia la izquierda, la forma del histograma se vuelve más dispersa, con una cola más larga en valores bajos.

Por último hacemos la visualización comparada de las distribuciones antes y después del shock en el precio del bien 1.



En comparación con el caso homogéneo, el corrimiento a la izquierda sigue presente, pero ahora el ajuste es desigual: los hogares que originalmente consumían mucho del bien 1 recortan de forma marcada, mientras que aquellos con preferencias débiles por ese bien apenas cambian su consumo. Por eso, en la distribución después del shock aumenta la densidad en los valores más bajos: no porque haya más consumo, sino porque más hogares quedan concentrados cerca del mínimo. El resultado es una caída generalizada, con mayor acumulación en consumos bajos y una fuerte reducción en los valores altos.