Trabajo Práctico de R

Denisse Blasco Nicolás Castelao

20 de octubre de 2025

Índice

1. Ejercicio 1: Análisis de datos

Del sitio web https://www.football-data.co.uk/englandm.php extraimos datos sobre la Premier League de fútbol.

Nuestras preguntas de interés son:

- ¿Cuál es el efecto de la localía en el resultado?
- ¿Qué tan comúnes son las remontadas?
- ¿Cuáles son los árbitros con mayor severidad disciplinaria? ¿Hay relación con la localidad?

Para esto, utilizamos datasets de partidos de fútbol de la Premier League. Restringimos la información desde la temporada 2010-2011 a 2023-2024, teniendo en cuenta cambios en el desarrollo del deporte y lo incompleto de los datasets futuros. Para la pregunta acerca de la localía, separadamente realizamos un análisis sin tener en cuenta los años de pandemia, donde no hubo público local.

Primero, realizamos las manipulaciones correspondientes a los datasets, explicadas con comentarios:

```
    # Tenemos un dataset por cada temporada. Queremos juntarlos.
# El dataset t_i corresponde a la temporada del año i-i+1. Es decir, t_2023 tiene la temporada 2023-2024
t_2010 <- read.csv("C:/Users/gusta/Downloads/2010.csv", header = TRUE)
t_2011 <- read.csv("C:/Users/gusta/Downloads/2011.csv", header = TRUE)
t_2012 <- read.csv("C:/Users/gusta/Downloads/2012.csv", header = TRUE)
t_2013 <- read.csv("C:/Users/gusta/Downloads/2013.csv", header = TRUE)
t_2014 <- read.csv("C:/Users/gusta/Downloads/2014.csv", header = TRUE)
t_2015 <- read.csv("C:/Users/gusta/Downloads/2015.csv", header = TRUE)
t_2016 <- read.csv("C:/Users/gusta/Downloads/2016.csv", header = TRUE)
t_2017 <- read.csv("C:/Users/gusta/Downloads/2017.csv", header = TRUE)
t_2018 <- read.csv("C:/Users/gusta/Downloads/2018.csv", header = TRUE)
t_2019 <- read.csv("C:/Users/gusta/Downloads/2019.csv", header = TRUE)
t_2020 <- read.csv("C:/Users/gusta/Downloads/2020.csv", header = TRUE)
t_2021 <- read.csv("C:/Users/gusta/Downloads/2021.csv", header = TRUE)
t_2022 <- read.csv("C:/Users/gusta/Downloads/2022.csv", header = TRUE)
t_2023 <- read.csv("C:/Users/gusta/Downloads/2023.csv", header = TRUE)


# Nos quedaremos con las variables de interés.

vars_interes <- c(
  "FTR",                    # resultado final
  "HTR",                    # resultado al entretiempo.
  "Referee","HY","AY","HR","AR" # disciplina. No ajustamos por cantidad de faltas pues suponemos que la elección de árbitros y la intensidad del juego no están correlacionadas.
)

df_2010 <- t_2010[, vars_interes]
df_2011 <- t_2011[, vars_interes]
df_2012 <- t_2012[, vars_interes]
df_2013 <- t_2013[, vars_interes]
df_2014 <- t_2014[, vars_interes]
df_2015 <- t_2015[, vars_interes]
```

```r
df_2016 <- t_2016[, vars_interes]
df_2017 <- t_2017[, vars_interes]
df_2018 <- t_2018[, vars_interes]
df_2019 <- t_2019[, vars_interes]
df_2020 <- t_2020[, vars_interes]
df_2021 <- t_2021[, vars_interes]
df_2022 <- t_2022[, vars_interes]
df_2023 <- t_2023[, vars_interes]

# Los uno
data <- rbind(
  df_2010,
  df_2011,
  df_2012,
  df_2013,
  df_2014,
  df_2015,
  df_2016,
  df_2017,
  df_2018,
  df_2019,
  df_2020,
  df_2021,
  df_2022,
  df_2023
)

# Le saco la única fila con missing data
data <- data[-1901, ]
sum(is.na(data))
```

Luego, comenzamos a responder cada pregunta.

1.1. ¿Cuál es el efecto de la localía en el resultado?

Para comprendere mejor el fenómeno, grafico la frecuencia de cada resultado a través del siguiente código:

```r
barplot(table(data$FTR),
        main = "Frecuencia de resultados ",
        xlab = "Resultado final", ylab = "Frecuencia",
        col = c(rgb(0.7, 0.85, 1),  rgb(0.4, 0.6, 0.9), rgb(0.1, 0.3, 0.7)),
        border = "black")
```

El gráfico resultante es

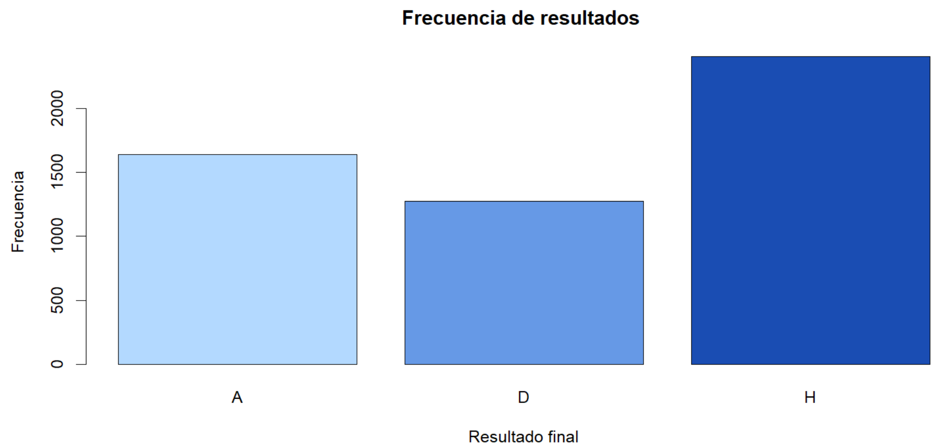

Figura 1: Gráfico de resultados de la Premier League.

Por lo tanto, se observa que, dado que hay igual cantidad de partidos de visitante que de local (cada partido tiene un visitante y un local en la Premier), entonces las probabilidades pueden calcularse de la siguiente manera.

$$P(\text{ganar} \mid \text{soy local}) = \frac{\#\text{victorias como local}}{\#\text{partidos totales}}$$

$$P(\text{ganar} \mid \text{soy visitante}) = \frac{\#\text{victorias como visitante}}{\#\text{partidos totales}}$$

$$P(\text{empatar} \mid \text{soy local}) = \frac{\#\text{empates como local}}{\#\text{partidos totales}} = P(\text{empatar} \mid \text{siendo visitante}) = P(\text{empatar})$$

Calculemos todo esto.

```
p_ganar_local = sum(data$FTR == "H") / nrow(data)
p_ganar_visita = sum(data$FTR == "A") / nrow(data)
p_empate = sum(data$FTR == "D") / nrow(data)
```

Lo observado es que en el 45 % de los partidos gana el local, mientras que el visitante solo gana en el 30 % de ellos.

1.2. ¿Qué tan comúnes son las remontadas?

Para ver esto, agregamos una columna que valga 1 si hubo remontada y valga 0 si no. Se considera remontada un partido que se estaba perdiendo en el primer tiempo y se termina ganando hacia el final del partido. Además, grafico la frecuencia de las remontadas Esto se hace a través del siguiente código:

```
    data$remontada = as.integer((data$HTR == "H" & data$FTR == "A") | (
       data$HTR == "A" & data$FTR == "H"))

barplot(table(data$remontada),
        main = "Frecuencia de remontadas ",
        xlab = "Remontada", ylab = "Frecuencia",
```

```
        col = c(rgb(0.7, 0.85, 1), rgb(0.1, 0.3, 0.7) ),
        border = "black")
```

Y el gráfico resultante es:

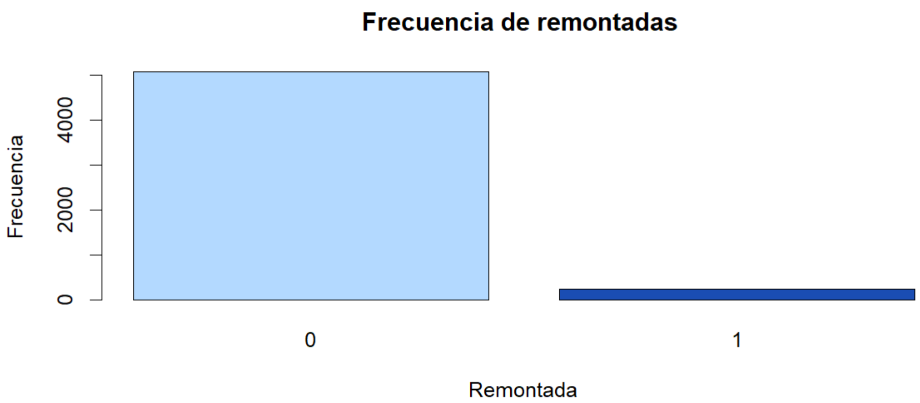

Figura 2: Gráfico de remontadas de la Premier League.

Es decir, se remontan aproximadamente 4.5 % de los partidos.

Ahora calculamos la probabilidad de remontar dado que llegamos al entretiempo perdiendo, tanto para el local como para el visitante:

```
    cat("P(remontar | perdía el local al ET) = ",round(100 * (sum(data$HTR == "A" & data$FTR == "H") / sum(data$HTR == "A")), 4), "%\n")
cat("P(remontar | perdía el visitante al ET) = ",round(100 * (sum(data$HTR == "H" & data$FTR == "A") / sum(data$HTR == "H")), 4), "%\n")
```

Y el resultado es que el local tiene el doble de probabilidades de remontar si llega perdiente al entretiempo, siendo 10.3139 % contra 5.4915 %.

1.3. ¿Cuáles son los árbitros con mayor severidad disciplinaria? ¿Hay relación con la localidad?

Hay 50 árbitros distintos, lo cual se consigue usando:

```
arbitros <- unique(data$Referee)
n_arbitros <- length(arbitros)
```

Armo un indice de severidad disciplinaria, asignando $10 =$ tarjeta amarilla y $25 =$ tarjeta roja. También junto otras columnas.

```
    data$disciplina <- 10*(data$HY + data$AY) + 25*(data$HR + data$AR)
data$tot_yellow <- data$HY + data$AY

arbitros_df <- data.frame(Referee = arbitros,disciplina = numeric(n_arbitros))
for (i in 1:n_arbitros){
  arbitros_df$disciplina[i] = mean(data$disciplina[data$Referee == arbitros_df$Referee[i]])
}
```

Grafico los 15 árbitros más severos:

```
    top_severos <- order(arbitros_df$disciplina, decreasing = TRUE)[1:15]
barplot(
  arbitros_df$disciplina[top_severos],
  col = rgb(0.1,0.83,0.7,0.85),
  main = "Árbitros con mayor severidad disciplinaria",
  ylab = "Promedio por partido",
  names.arg = arbitros_df$Referee[top_severos],
  las = 2, cex.names = 0.7, border = "black"
)
```

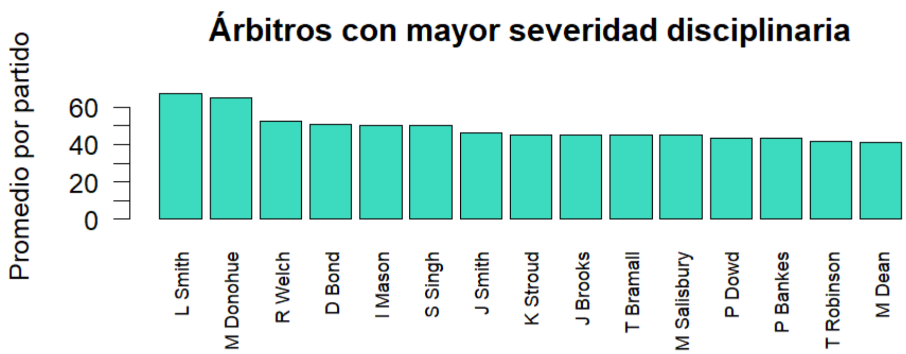

Figura 3: Árbitros más severos.

Veamos ahora información sobre la cantidad de amarillas según el resultado, usando el siguiente código:

```
    y_H <- data$tot_yellow[data$FTR == "H"]
y_A <- data$tot_yellow[data$FTR == "A"]
y_D <- data$tot_yellow[data$FTR == "D"]

hist(y_H,
     col = rgb(0.10, 0.55, 0.10, 0.35), border = rgb(0.10, 0.55, 0.10,
        0.9),
     main = "Distribución de amarillas (HY+AY) por resultado",
     xlab = "Amarillas totales")
hist(y_D, add = TRUE,
     col = rgb(0.35, 0.70, 0.35, 0.35), border = rgb(0.35, 0.70, 0.35,
        0.9))
hist(y_A, add = TRUE,
     col = rgb(0.40, 0.60, 0.90, 0.35), border = rgb(0.40,0.60,0.90,0.9))
legend("topright",
       fill = c(rgb(0.10, 0.55, 0.10, 0.35), rgb(0.35, 0.70, 0.35, 0.35),
          rgb(0.40, 0.60, 0.90, 0.35)),
       border = c(rgb(0.10, 0.55, 0.10, 0.9),  rgb(0.35, 0.70, 0.35, 0.9),
            rgb(0.40,0.60,0.90,0.9)),
       legend = c("Gana local", "Empate", "Gana visita"),
       bty = "n")
```

Gráficamente se observa que hay una mayor cantidad de amarillas cuando gana el local, intermedia cuando gana la visita y menor cantidad en partidos que resultan en empate:

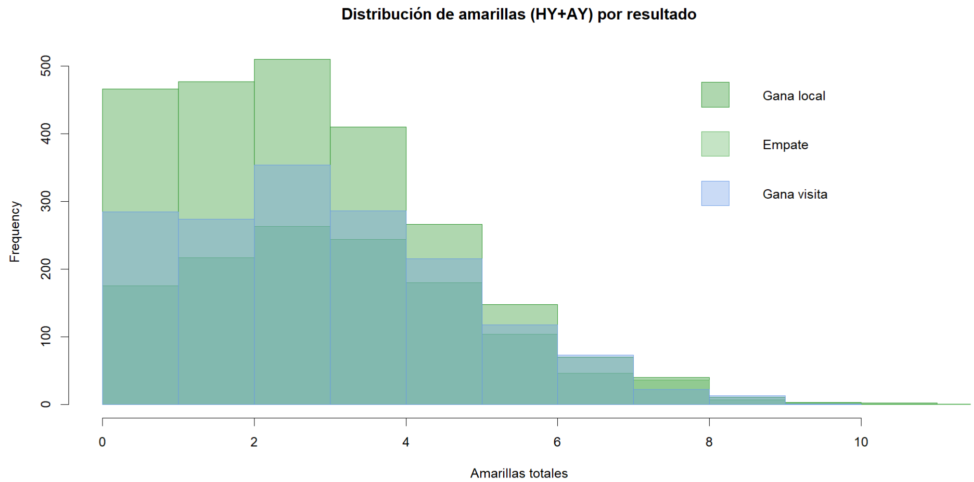

Figura 4: Árbitros más severos.

Ahora analizamos el efecto de la localidad en las tarjetas amarillas, codificando según:

```
    plot(jitter(data$HY, amount = 0.15),  #dispersa los puntos aleatoriamente para que se puedan visualizar mejor
     jitter(data$AY, amount = 0.15),
     pch = 16, col = rgb(0.10,0.30,0.70,0.25),
     xlab = "Amarillas del local",
     ylab = "Amarillas del visitante",
     main = "Amarillas: local vs visitante")
abline(0, 1, col = "red", lwd = 2, lty = 2) # La linea que refleja la imparcialidad.
```

El gráfico resultante no muestra un sesgo claro hacia la localidad o la visita, pero sí se ve que una mayor cantidad de amarillas para un equipo correlaciona con una mayor cantidad de amarillas para el otro:

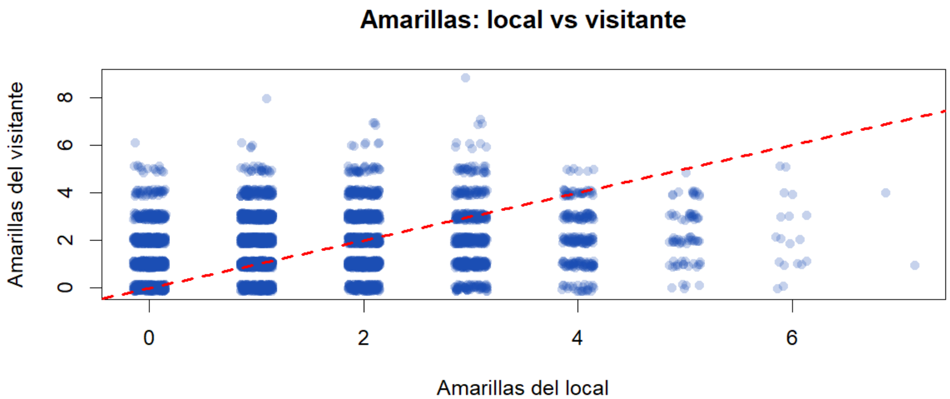

Figura 5: Localidad en tarjetas amarillas.

2. Ejercicio 2: Análisis econométrico

Para este ejercicio, trabajamos con el dataset `gapminder.csv`, que contiene información de distintos países sobre ingreso per cápita, expectativa de vida, religión y otros indicadores. El análisis se centra en Argentina y luego se amplía al estudio comparativo regional y global.

```
data <- read.csv("C:/Users/gusta/Downloads/gapminder.csv", header = TRUE)
library(ggplot2)
```

2.1. Ingreso per cápita de Argentina

Filtramos los datos correspondientes a Argentina y graficamos la evolución del ingreso per cápita a lo largo de los años:

```
arg <- subset(data, country=="Argentina")

ggplot(arg,aes(x = year, y = income_per_person)) +
  geom_line() +
  geom_point() +
  labs(title = "Ingreso per cápita argentino",
       x = "Año", y = "Ingreso per cápita") +
  theme_minimal()
```

El gráfico obtenido se muestra a continuación:

Figura 6: Ingreso per cápita argentino a lo largo del tiempo.

Se observa que antes de 1875 la tendencia era creciente. Luego, entre 1875 y 1990, el ingreso per cápita disminuyó. A partir de 1990 volvió a crecer sostenidamente hasta 1998, cayendo nuevamente en 2002. Desde entonces, la tendencia vuelve a ser positiva, con una leve caída en 2009.

2.2. Modelos de predicción y comparación

Separamos los datos en conjuntos de entrenamiento y prueba, y ajustamos tres modelos: lineal, polinómico de grado 2 y polinómico de grado 10.

```r
train <- arg[1:(nrow(arg) - 10), ]
test  <- arg[(nrow(arg)-9):nrow(arg),]

lineal <- lm(income_per_person ~ year, data = train)
pol2   <- lm(income_per_person ~ poly(year, 2, raw = TRUE), data = train)
pol10  <- lm(income_per_person ~ poly(year, 10, raw = TRUE), data = train)
```

Luego, comparamos los modelos utilizando el RMSE en el conjunto de testeo:

```r
rmse <- function(obs, pred) sqrt(mean((obs - pred)^2))

pred_lin   <- predict(lineal, newdata = test)
pred_pol2  <- predict(pol2, newdata = test)
pred_pol10 <- predict(pol10, newdata = test)

rmse_lin   <- rmse(test$income_per_person, pred_lin)
rmse_pol2  <- rmse(test$income_per_person, pred_pol2)
rmse_pol10 <- rmse(test$income_per_person, pred_pol10)
```

Visualmente, los tres modelos ajustados sobre toda la serie lucen así:

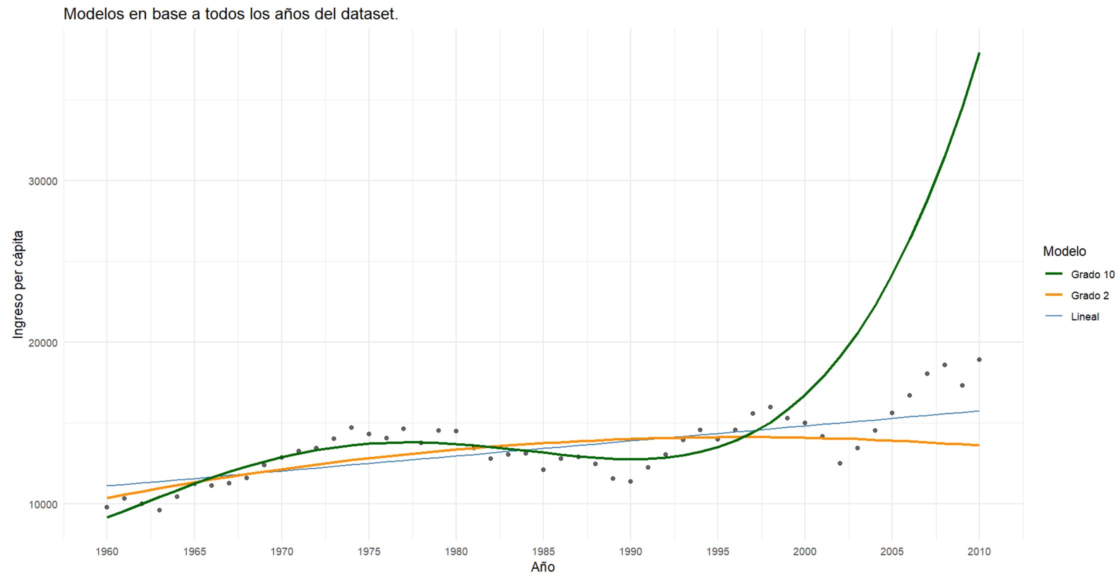

Figura 7: Ajuste de modelos sobre el ingreso per cápita argentino.

Y las predicciones sobre los últimos 10 años (conjunto de test) se muestran en la siguiente figura:

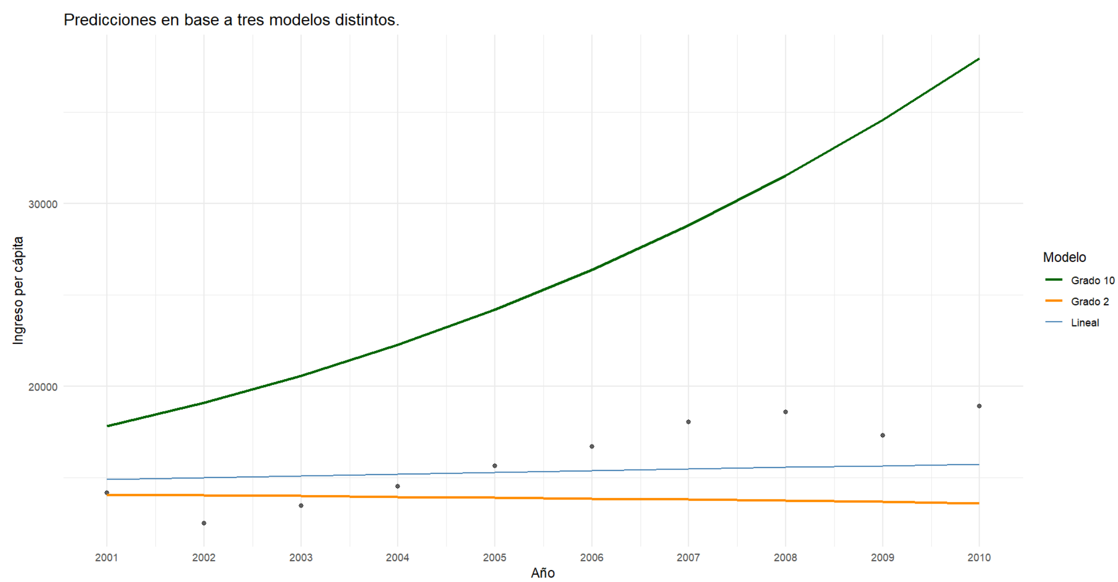

Figura 8: Predicciones sobre el conjunto de testeo (últimos 10 años).

Todos estos gráficos se obtienen mediante las siguientes líneas:

```
    grid <- data.frame(year = arg$year)
grid$y_lin   <- predict(lineal,   newdata = grid)
grid$y_pol2  <- predict(pol2,  newdata = grid)
grid$y_pol10 <- predict(pol10, newdata = grid)

ggplot() +
  geom_point(data = arg, aes(year, income_per_person), alpha = 0.6) +
```

```
  geom_line(data = grid, aes(year, y_lin,   color = "Lineal")) +
  geom_line(data = grid, aes(year, y_pol2,  color = "Grado 2"), linewidth = 1) +
  geom_line(data = grid, aes(year, y_pol10, color = "Grado 10"), linewidth = 1) +
  labs(
    title = "Modelos en base a todos los años del dataset.",
    x = "Año", y = "Ingreso per cápita", color = "Modelo"
  ) +
  scale_x_continuous(breaks = seq(min(arg$year), max(arg$year), by = 5)) +
  scale_color_manual(values = c("Lineal" = "steelblue", "Grado 2" = "darkorange", "Grado 10" = "darkgreen")) +
  theme_minimal()


# Ahora veo las predicciones sobre el conjunto de testeo
grid2 <- data.frame(year = test$year)
grid2$y_lin   <- predict(lineal,  newdata = grid2)
grid2$y_pol2  <- predict(pol2,  newdata = grid2)
grid2$y_pol10 <- predict(pol10, newdata = grid2)

ggplot() +
  geom_point(data = test, aes(year, income_per_person), alpha = 0.6) +
  geom_line(data = grid2, aes(year, y_lin,   color = "Lineal")) +
  geom_line(data = grid2, aes(year, y_pol2,  color = "Grado 2"), linewidth = 1) +
  geom_line(data = grid2, aes(year, y_pol10, color = "Grado 10"), linewidth = 1) +
  labs(
    title = "Predicciones en base a tres modelos distintos.",
    x = "Año", y = "Ingreso per cápita", color = "Modelo"
  ) +
  scale_x_continuous(breaks = seq(min(test$year), max(test$year), by = 1)) +
  scale_color_manual(values = c("Lineal" = "steelblue", "Grado 2" = "darkorange", "Grado 10" = "darkgreen")) +
  theme_minimal()
```

2.3. Co-movimiento regional del ingreso per cápita

Se analizan los países Argentina, Brasil, Chile, Perú y Uruguay, construyendo una matriz con sus ingresos per cápita para medir correlaciones tanto en niveles como en variaciones porcentuales.

```
library(tidyverse)
paises <- c("Argentina","Brazil","Chile","Peru","Uruguay")

paises_df <- data %>%
  filter(country %in% paises) %>%
  select(year, country, income_per_person) %>%
  pivot_wider(names_from = country, values_from = income_per_person) %>%
  arrange(year)

correlaciones <- cor(paises_df, use="pairwise.complete.obs")
```

```r


paises_df_pct <- paises_df %>%
  mutate(across(-year,
                ~ . / dplyr::lag(.), #. para hablar de la columna actual, lag para hacer referencia al año anterior, la división para calcular la tasa.
                .names = "{.col}_pct"))

paises_df_pct <- paises_df_pct[2:nrow(paises_df_pct),-c(2:6)]

correlaciones_pct <- cor(paises_df_pct)[2:nrow(cor(paises_df_pct)),2:ncol(cor(paises_df_pct))]
```

El resultado muestra correlaciones altas y positivas en niveles, indicando que las economías tienden a moverse conjuntamente en el largo plazo. En cambio, al analizar las variaciones porcentuales anuales, las correlaciones se reducen y se vuelven más heterogéneas, lo que refleja diferencias en los ciclos económicos de corto plazo.

2.4. Relación entre expectativa de vida total y femenina (2010)

Seleccionamos el año 2010 y graficamos la relación entre expectativa de vida total y femenina:

```r
year_df <- subset(data, year == 2010)
year_df <- year_df %>% drop_na(life_expectancy)
year_df <- year_df[year_df$life_expectancy_female != "-",]
year_df$life_expectancy_female <- as.numeric(year_df$life_expectancy_female)

ggplot(year_df, aes(x = life_expectancy_female, y = life_expectancy)) +
  geom_point() +
  geom_abline(color = "red") +
  geom_smooth(method = lm, color = "blue") +
  theme_minimal()
```

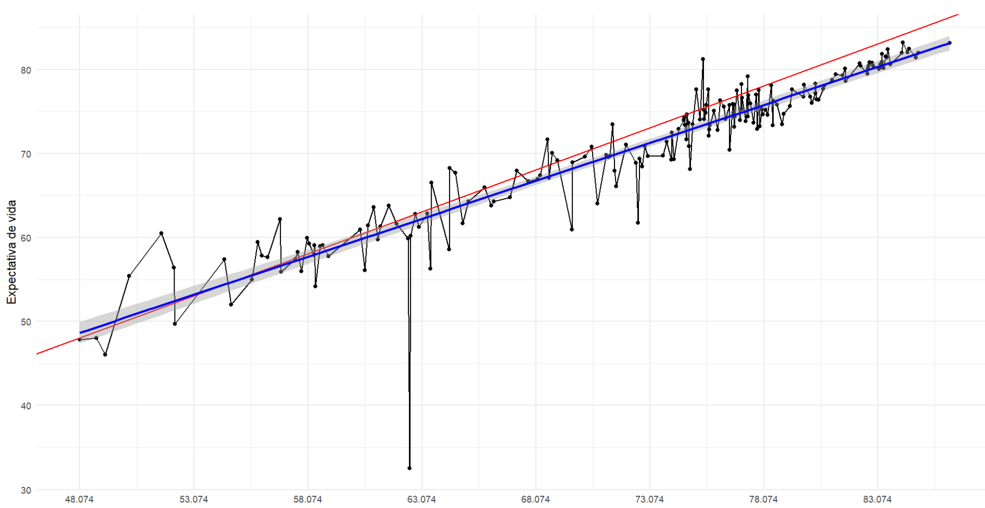

Figura 9: Expectativa de vida total vs. femenina (año 2010).

La relación es claramente positiva, con pendiente menor que 1, indicando que los hombres viven menos en promedio.

2.5. Modelo simple: expectativa total vs femenina

```r
modelo <- lm(life_expectancy ~ life_expectancy_female, data = year_df)
summary(modelo)
```

El modelo lineal simple arroja:

$$\hat{\beta}_0 = 5{,}22, \qquad \hat{\beta}_1 = 0{,}90, \qquad R^2 = 0{,}874$$

Que el coeficiente β_1 sea 0.9037 indica que cuando la expectativa de vida femenina aumenta 1 año, la expectativa de vida total aumenta 0.9 años aproximadamente. El intercepto no tiene interpretación causal, solo ajusta el modelo. El R^2 es muy alto, lo cual dice que la expectativa de vida femenina explica en gran parte cambios en el nivel de expectativa de vida total.

2.6. Prueba t pareada: ¿la expectativa femenina es mayor?

Se realiza el test con H_0: life_expectancy_female = life_expectancy y H_1: life_expectancy_female >life_expectancy. Por esa misma razón, uso un test pareado a la derecha unilateralmente.

```r
t.test(year_df$life_expectancy_female,
       year_df$life_expectancy,
       paired = TRUE,
       alternative = "greater")
```

El t-test pareado unilateral entre female y total dio t = 7.105, p muy cercano a cero. La diferencia media (female − total) de 1.72 años con un intervalo de confianza del 95 %. Como el intervalo queda estrictamente por encima de 0 y el p-valor es demasiado pequeño, rechazamos la hipótesis nula y concluimos que la esperanza de vida femenina es mayor que la total en estos datos.

2.7. Modelo múltiple con ingreso per cápita

Creo el modelo mediante el siguiente código:

```r
modelo_multiple <- lm(life_expectancy ~
                      life_expectancy_female + income_per_person,
                      data = year_df)
summary(modelo_multiple)
```

Vemos que el coeficiente asociado a la expectativa de vida femenina es 0.866 y es significativo. Esto significa que un año más de expectativa de vida femenina se asocia a 0.866 años más de expectativa de vida total.

El coeficiente asociado al ingreso per cápita es positivo pero casi cero, con cierta significatividad. Suponiendo que el ingreso per cápita está en miles de dolares, esto significa que un aumento en 1000 dolares al ingreso per cápita produce un aumento positivo pero casi cero en años de expectativa de vida total.

El R^2 de esta regresión es 0.8772, por lo que las variables independientes explican con fortaleza a la variable dependiente.

Al controlar por ingreso, el coeficiente de la expectativa de vida femenina disminuye, pasando de 0.904 a 0.866. Esto muestra que parte del efecto anteriormente atribuido a la expectativa de vida femenina en realidad era explicado por el ingreso per cápita. Sin embargo, el efecto de la expectativa de vida femenina permanece siendo grande y significativo.

Luego, vemos que incluir la variable de ingreso per cápita mejora la estimación central y aumenta el R^2 del modelo. Sin embargo, el aumento del R^2 es muy chico, por lo que a fines prácticos podría ser un problema incluir esta variable en la regresión.

2.8. Modelo alternativo: ingreso, religión y población

Las covariables a elegir son: logaritmo del ingreso per cápita, la población y la religión principal. Hago el modelo:

```r
year_df$log_income <- log(year_df$income_per_person)

year_df$main_religion <- trimws(year_df$main_religion)
year_df$main_religion <- gsub("\\s+", " ", year_df$main_religion)
tmp <- tolower(year_df$main_religion)
tmp[tmp %in% c("", "na", "n/a")] <- NA
tmp[tmp %in% c("christian")] <- "Christian"
tmp[tmp %in% c("muslim")] <- "Muslim"
tmp[tmp %in% c("eastern religions")] <- "Eastern religions"
year_df$main_religion <- factor(tmp)

modelo_propio <- lm(life_expectancy ~ log_income + main_religion + population,
                    data = year_df)
summary(modelo_propio)
```

Vemos que el coeficiente asociado al logaritmo natural del ingreso per cápita es 5.691. Esto significa que un aumento en una unidad del log_income genera un aumento en 5.691 años de la expectativa de vida total. Este coeficiente es altamente significativo. Esto significa que un aumento del 1 % en el ingreso per cápita esta asociado con 0.0569 años más de expectativa de vida, cercano a 21 dias.

El coeficiente asociado a la población es $-4{,}274{,}10^{-11}$, negativo pero muy cercano a cero. Aunque esto no es significativo, por lo que podemos asumir que la estimación es cero. Esto significa que no hay relación entre la cantidad de población y la expectativa de vida total según nuestros datos.

Los coeficientes de las religiones usan al cristianismo como religión base. Luego, que el coeficiente de religiones del Este sea 3.312, con cierta significatividad, significa que hay un efecto positivo marginal con respecto al cristianismo en la expectativa de vida. El coeficiente de los musulmanes es 0.673 sin significatividad. Luego, no podemos decir mucho de este regresor.

El R^2 de la regresión es de 0.5901, lo cual muestra que los regresores elegidos solamente explican el 59 % de la variación en la expectativa de vida.

Además, hubo 17 observaciones eliminadas por faltantes en alguna variable.

3. Ejercicio 3: Simulación y estática comparativa

En este ejercicio simulamos ingresos a partir de una distribución $\chi^2(k)$, analizamos una demanda tipo Cobb–Douglas y estudiamos cómo cambian las distribuciones de consumo y la utilidad indirecta ante un shock de precios y con heterogeneidad de preferencias.

3.1. Ingreso simulado y momentos teóricos

Definimos una función para simular ingresos $Y \sim \chi^2(k)$. Teóricamente, $E[Y] = k$ y $\mathrm{Var}(Y) = 2k$, por lo que k determina el nivel y la dispersión. Luego, comparamos histogramas empíricos con la densidad teórica para distintos k.

```r
simular_ingreso <- function(k,n){
  Y <- rchisq(n, df = k)
  return(Y)
}

# Teoricamente, los valores son
media_teo <- k
varianza_teo <- 2*k
sd_teo <- sqrt(varianza_teo)

# Veamos cómo cambia para distintos valores de k. Se realiza un gráfico
    para cada k.
ks = c(5,10,20)

for (k in ks) {
  Y <- simular_ingreso(k,1000)
  hist(Y, freq = FALSE, col = "lightblue",
       main = paste("Ingreso ~ Chi-cuadrado(k =", k, ")"),
       xlab = "Y")
  curve(dchisq(x, df = k), add = TRUE, col = "red", lwd = 2)
}
```

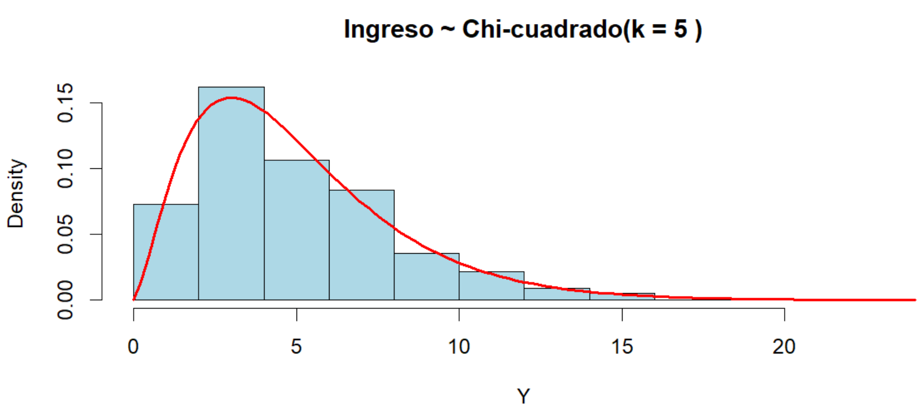

Figura 10: Ingreso simulado vs densidad teórica ($k = 5$).

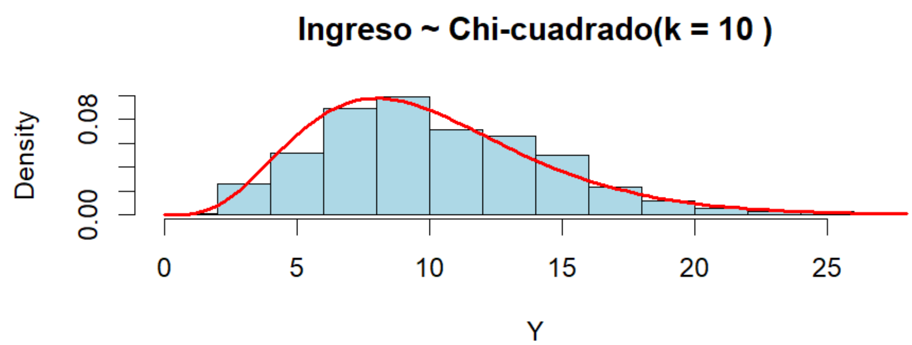

Figura 11: Ingreso simulado vs densidad teórica ($k = 10$).

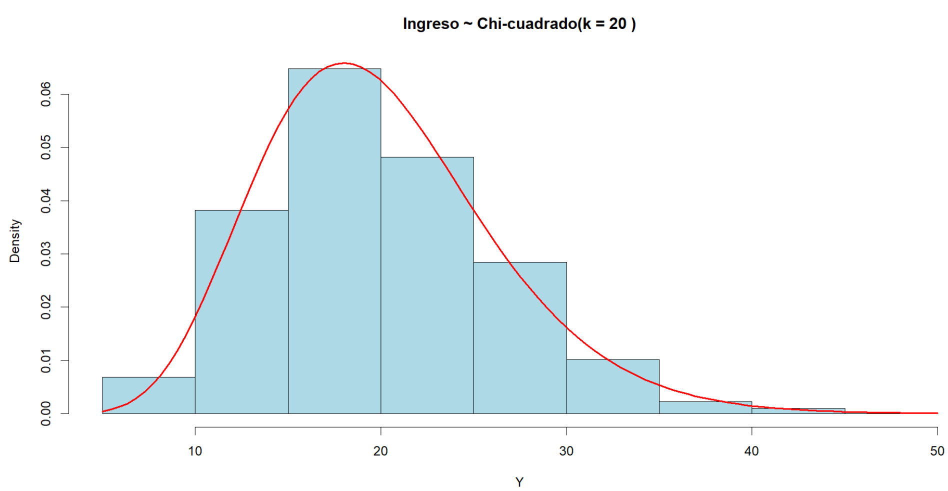

Figura 12: Ingreso simulado vs densidad teórica ($k = 20$).

3.2. Demanda Cobb–Douglas y utilidad indirecta

Bajo precios $p_1, p_2 > 0$, ingreso $Y \geq 0$ y preferencias $a_1, a_2 > 0$ con $a_1 + a_2 = 1$, la demanda óptima es:

$$x_1^* = \frac{a_1 Y}{p_1}, \qquad x_2^* = \frac{a_2 Y}{p_2}, \qquad V = (x_1^*)^{a_1}(x_2^*)^{a_2}.$$

```
Y <- simular_ingreso(5,1000)

demanda_cd <- function(Y, p1, p2, a1, a2){
  if (p1 <= 0 || p2 <= 0 || a1 <= 0 || a2 <= 0 || Y < 0){
    print("Los valores ingresados deben ser positivos o, en el caso del
        ingreso, no negativo")
  } else if (a1 + a2 - 1 > 1e-8){
    print("la suma de los alphas debe dar 1")
  }  else {
    x_1 <- (a1 * Y) / p1
```

```r
    x_2 <- (a2 * Y) / p2
    V <- (x_1 ^ a1) * (x_2 ^ a2)
    return(c(x_1,x_2,V))
  }
}

demanda_cd(mean(Y),10,5,0.5,0.5)
```

3.3. Monte Carlo y estadísticos descriptivos

Simulamos $n = 10{,}000$ consumidores con $k = 5$, $p_1 = 10$, $p_2 = 8$, $a_1 = 0{,}3$, $a_2 = 0{,}7$; y reportamos medias y cuartiles de x_1^*, x_2^*, V. Grafico estas situaciones

```r
n = 10000
k = 5
p1 = 10
p2 = 8
a1 = 0.3
a2 = 0.7
Y <- simular_ingreso(k,n)
resultados <- matrix(NA_real_,n,3)

for (j in 1:n){
  for (i in 1:3){
    resultados[j,i] = demanda_cd(Y[j],p1,p2,a1,a2)[i]
  }
}

# Extraigo los resultados como vectores para poder graficar
x1 <- resultados[,1]
x2 <- resultados[,2]
V <- resultados[,3]

hist(x1, freq = FALSE, col = "lightblue",
     main = "x_1* (empírico)", xlab = "x_1*")
abline(v = mean(x1), col = "blue", lwd = 2)
abline(v = quantile(x1, c(0.25, 0.5, 0.75)), col = "red", lty = 2)

hist(x2, freq = FALSE, col = "lightblue",
     main = "x_2* (empírico)", xlab = "x_2*")
abline(v = mean(x2), col = "blue", lwd = 2)
abline(v = quantile(x2, c(0.25, 0.5, 0.75)), col = "red", lty = 2)

hist(V, freq = FALSE, col = "lightblue",
     main = "V* (empírico)", xlab = "V*")
abline(v = mean(V), col = "blue", lwd = 2)
abline(v = quantile(V, c(0.25, 0.5, 0.75)), col = "red", lty = 2)

# Reporto los valores.
qx1 <- quantile(x1, c(0.25, 0.5, 0.75))
qx2 <- quantile(x2, c(0.25, 0.5, 0.75))
qV  <- quantile(V,  c(0.25, 0.5, 0.75))

```

```r
cat("\n--- Medias (empíricas) ---\n")
cat("mean(x1*) =", mean(x1), "\n")
cat("mean(x2*) =", mean(x2), "\n")
cat("mean(V*)  =", mean(V), "\n")

cat("\n--- Cuartiles (empíricos: Q1, Q2, Q3) ---\n")
cat("x1*:", round(qx1, 4), "\n")
cat("x2*:", round(qx2, 4), "\n")
cat("V* :", round(qV,  4), "\n")
```

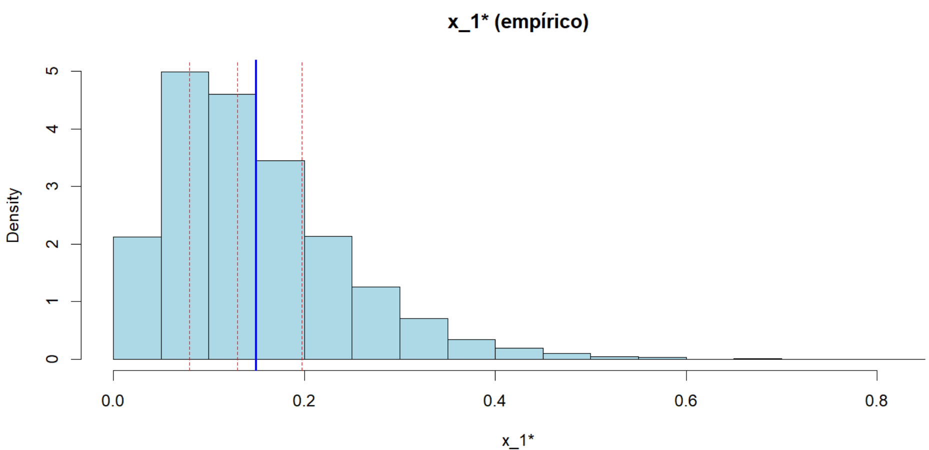

Figura 13: Distribución empírica de x_1^* con media y cuartiles.

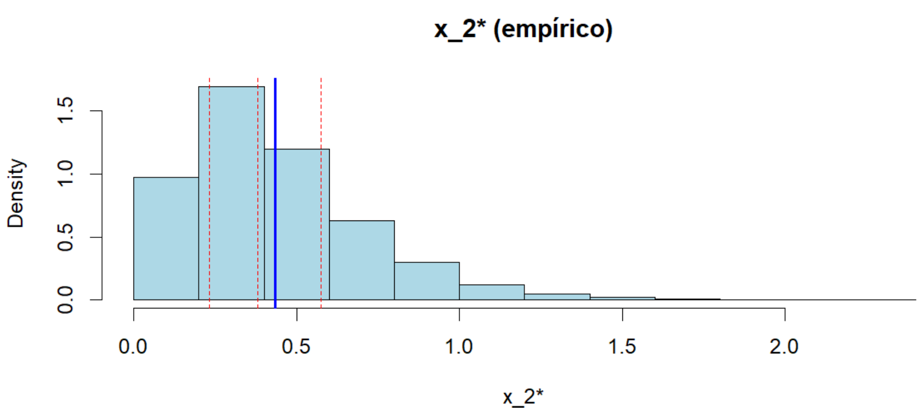

Figura 14: Distribución empírica de x_2^* con media y cuartiles.

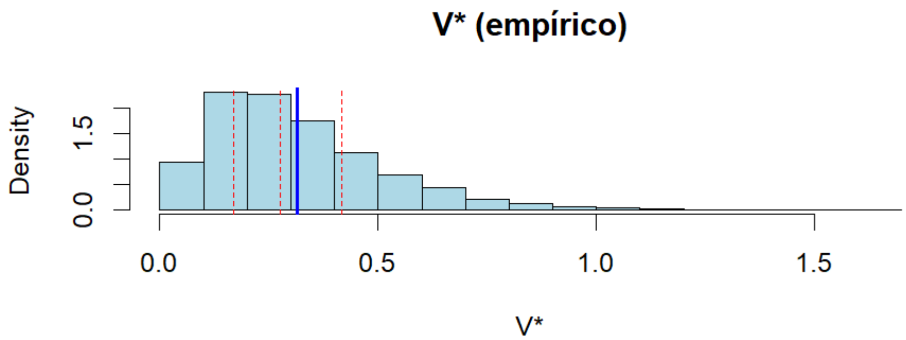

Figura 15: Distribución empírica de V con media y cuartiles.

3.4. Probabilidad de bajo consumo

Estimamos $\Pr(x_j < c)$ como frecuencia muestral, para $j \in \{1, 2\}$.

```
prob_bajo_consumo <- function(c,j){
  if (j == 1){
    x = x1
  } else {x = x2}
  aciertos <- 0
  for (i in 1:n){
    if (x[i] < c){
      aciertos = aciertos + 1
    }
  }
  return(aciertos/n)
}
```

3.5. Shock de precio y comparación de distribuciones

Aplicamos un shock $p_1' = 1{,}2\,p_1$ y comparamos medias y cuartiles de x_1^*, x_2^*, V antes y después. El resultado es que x_1^* cae, x_2^* no cambia (no depende de p_1) y la utilidad indirecta V disminuye.

```
p1_nuevo = p1*1.2
resultados_nuevos <- matrix(NA_real_,n,3)

for (j in 1:n){
  for (i in 1:3){
    resultados_nuevos[j,i] = demanda_cd(Y[j],p1_nuevo,p2,a1,a2)[i]
  }
}

x1_nuevos <- resultados_nuevos[,1]
x2_nuevos <- resultados_nuevos[,2]
V_nuevos <- resultados_nuevos[,3]

qx1_nuevos <- quantile(x1_nuevos, c(0.25, 0.5, 0.75))
qx2_nuevos <- quantile(x2_nuevos, c(0.25, 0.5, 0.75))
```

```r
qV_nuevos  <- quantile(V_nuevos,  c(0.25, 0.5, 0.75))

# Comparo las distribuciones a través de indicadores empíricos
cat("\n--- Diferencia de medias (empíricas) ---\n")
cat("mean(x1*) - mean(x1*_nuevos) =", mean(x1) - mean(x1_nuevos), "\n")
cat("mean(x2*) - mean(x2*_nuevos) =", mean(x2) - mean(x2_nuevos), "\n")
cat("mean(V*) - mean(V*_nuevos) =", mean(V) - mean(V_nuevos), "\n")

cat("\n--- Diferencia de cuartiles (empíricos) ---\n")
cat("x1* - x1*_nuevos:", round(qx1, 4) - round(qx1_nuevos,4), "\n")
cat("x2* - x2*_nuevos:", round(qx2, 4) - round(qx2_nuevos,4), "\n")
cat("V* - V*_nuevos:", round(qV, 4) - round(qV_nuevos,4), "\n")
```

3.6. Visualización: antes vs. después del shock

Graficamos la superposición de histogramas de x_1^* antes y después del aumento de p_1.

```r
hist(x1, freq = FALSE, col = "lightblue",
     main = "x1*: antes vs. después (shock p1 +20%)", xlab = "x1*")
hist(x1_nuevos, freq = FALSE, col = NA, border = "red", add = TRUE, lty =
   2)

abline(v = mean(x1),        col = "blue", lwd = 2)
abline(v = mean(x1_nuevos), col = "darkred", lwd = 2, lty = 2)

legend("topright",
       legend = c("Antes", "Después (borde)"),
       fill   = c("lightblue", "red"),
       border = c("black", "lightblue"),
       bty    = "n")
```

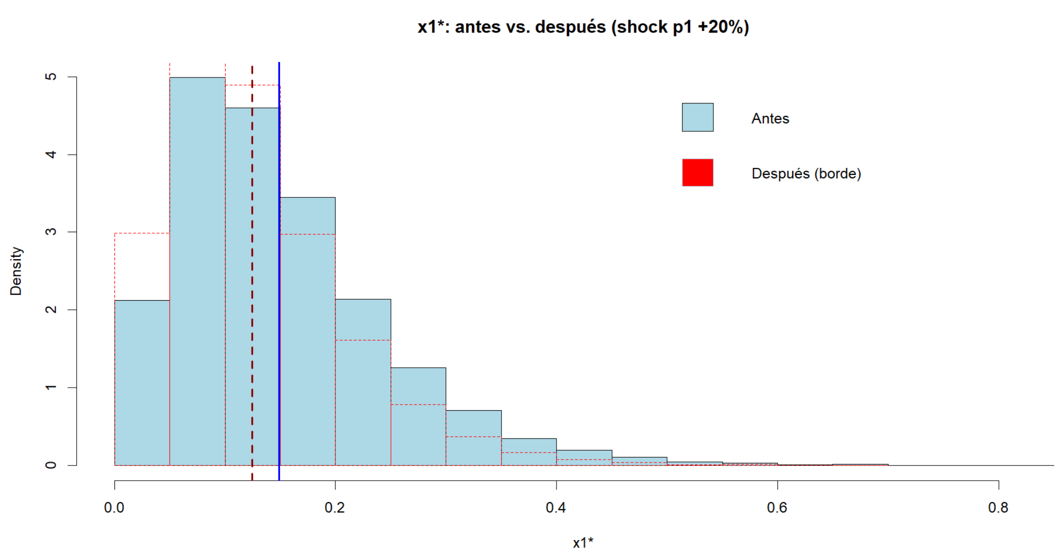

Figura 16: Distribución de x_1^*: antes vs. después del shock en p_1.

Vemos que la distribución tiene mayor densidad en valores bajos y menor densidad en valores altos. Esto también se visualiza en un menor promedio, tal como es mostrado por las rectas verticales.

La intuición económica es que el precio del bien 1 aumentó, y todo lo demás quedó constante. Por lo tanto, el individuo decidirá consumir menos del bien 1. Dado que el consumo del bien 1 también depende del ingreso, el resultado es que toda la distribución se mueve hacia valores menores de consumo del bien 1.

Lógicamente, ante menor consumo, con todo lo demás constante, la utilidad indirecta V cae. Esto fue mostrado empíricamente en el inciso 5.

3.7. Heterogeneidad en preferencias

Introducimos heterogeneidad $a_1 \sim \text{Beta}(a, b)$ y repetimos el ejercicio base y el shock. Esperamos que mayor a_1 implique mayor x_1^* y menor x_2^*. Elijo a = 1 y b = 2 arbitrarios, sabiendo que la esperanza es a/(a+b) y la varianza es $ab/((a+b)^2 * (a+b+1))$

```
alpha1 <- numeric(n)
a = 1
b = 2
for(i in 1:n){
  alpha1[i] <- rbeta(n = 1,shape1 = a, shape2 = b)
}

resultados_het <- matrix(NA_real_,n,3)
for (j in 1:n){
  for (i in 1:3){
    resultados_het[j,i] = demanda_cd(Y[j],p1,p2,alpha1[j],1-alpha1[j])[i]
  }
}

# Extraigo vectores
x1_het <- resultados_het[,1]
x2_het <- resultados_het[,2]
V_het  <- resultados_het[,3]

qx1_het <- quantile(x1_het, c(0.25, 0.5, 0.75))
qx2_het <- quantile(x2_het, c(0.25, 0.5, 0.75))
qV_het  <- quantile(V_het,  c(0.25, 0.5, 0.75))

cat("\n--- Diferencia de medias (empíricas) ---\n")
cat("mean(x1*) - mean(x1*_het) =", mean(x1) - mean(x1_het), "\n")
cat("mean(x2*) - mean(x2*_het) =", mean(x2) - mean(x2_het), "\n")
cat("mean(V*) - mean(V*_het)   =", mean(V) - mean(V_het), "\n")

cat("\n--- Diferencia de cuartiles (empíricos) ---\n")
cat("x1* - x1*_het:", round(qx1, 4) - round(qx1_het,4), "\n")
cat("x2* - x2*_het:", round(qx2, 4) - round(qx2_het,4), "\n")
cat("V*  - V*_het :", round(qV,  4) - round(qV_het,4), "\n")

# Cambio de precio con preferencias heterogéneas
resultados_nuevos_het <- matrix(NA_real_,n,3)
for (j in 1:n){
  for (i in 1:3){
```

```r
    resultados_nuevos_het[j,i] = demanda_cd(Y[j],p1_nuevo,p2,alpha1[j],1-
        alpha1[j])[i]
  }
}

x1_nuevos_het <- resultados_nuevos_het[,1]
x2_nuevos_het <- resultados_nuevos_het[,2]
V_nuevos_het  <- resultados_nuevos_het[,3]

qx1_nuevos_het <- quantile(x1_nuevos_het, c(0.25, 0.5, 0.75))
qx2_nuevos_het <- quantile(x2_nuevos_het, c(0.25, 0.5, 0.75))
qV_nuevos_het  <- quantile(V_nuevos_het,  c(0.25, 0.5, 0.75))

cat("\n--- Diferencia de medias (empíricas) ---\n")
cat("mean(x1*) - mean(x1*_nuevos_het) =", mean(x1) - mean(x1_nuevos_het),
   "\n")
cat("mean(x2*) - mean(x2*_nuevos_het) =", mean(x2) - mean(x2_nuevos_het),
   "\n")
cat("mean(V*) - mean(V*_nuevos_het)   =", mean(V) - mean(V_nuevos_het), "\
   n")

cat("\n--- Diferencia de cuartiles (empíricos) ---\n")
cat("x1* - x1*_nuevos_het:", round(qx1, 4) - round(qx1_nuevos_het,4), "\n"
   )
cat("x2* - x2*_nuevos_het:", round(qx2, 4) - round(qx2_nuevos_het,4), "\n"
   )
cat("V*  - V*_nuevos_het :", round(qV,  4) - round(qV_nuevos_het,4), "\n")
```

Intuitivamente, un mayor α_1 simboliza mayor importancia del bien 1 en la canaste del consumidor. Racionalmente, el consumidor destinará mayor parte de su ingreso al bien 1 a costa de consumo del bien 2, el cual se volvió menos deseado.

Entonces, es esperable que un aumento de α_1 tenga como consecuencia un aumento del consumo de x1 y una disminución del consumo del bien 2, a igual ingreso y precios.