# Universidad Nacional de Colombia
## Departamento de Estadística
### Análisis de regresión (II-2022)
### Parcial I

Andrés Mauricio Rico Parada - aricop@unal.edu.co
Juan David Carrrascal Ibañez - jdcarrascali@unal.edu.co

**1.**

**a.** Usando el sistema de hipótesis

$$\begin{cases} H_0 : \rho_{X,Y} = 0 \\ H_1 : \rho_{X,Y} \neq 0 \end{cases}$$

$\{T_c|X = x\} = \dfrac{\hat{\rho}_{X,Y}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2_{X,Y}}} = \dfrac{0{,}62 \cdot \sqrt{48}}{\sqrt{1-0{,}62^2}} \approx 5{,}47$. $t_{1-\frac{0{,}01}{2}}(48) \approx 2{,}01$. Por lo tanto, se rechaza la hipótesis nula ($\rho = 0$) con un nivel de significancia del 1 %.

**b.** Dado que los datos vienen de una distribución normal:
Calculamos

$$z_r = \frac{1}{2} Ln(\frac{1+0{,}62}{1-0{,}62}) \approx 0{,}75$$

Identificamos el límite superior e inferior mediante la fórmula:

$$z_r \mp z_{1-\alpha/2}\sqrt{\frac{1}{n-3}}$$
$$z_r \mp z_{0{,}995}\sqrt{\frac{1}{n-2}}$$

De manera que, $l = 0{,}354$ y $u = 1{,}096$. Ahora bien, transformamos $l$ y $u$ de vuelta a la escala original del coeficiente

$$p_l = \frac{\exp(2l)-1}{\exp(2l)+1} \quad y \quad p_u = \frac{\exp(2u)-1}{\exp(2u)+1}$$

De esta manera, un intervalo de confianza asíntotico para el coeficiente de correlación está dado por:

$$IC_{0{,}99\,\%}(\rho) = [0{,}34, 0{,}8]$$

**c** Dado que $0{,}5 \in IC_{0{,}99\,\%}(\rho)$, con una significancia de 1 % afirmamos que no hay evidencia

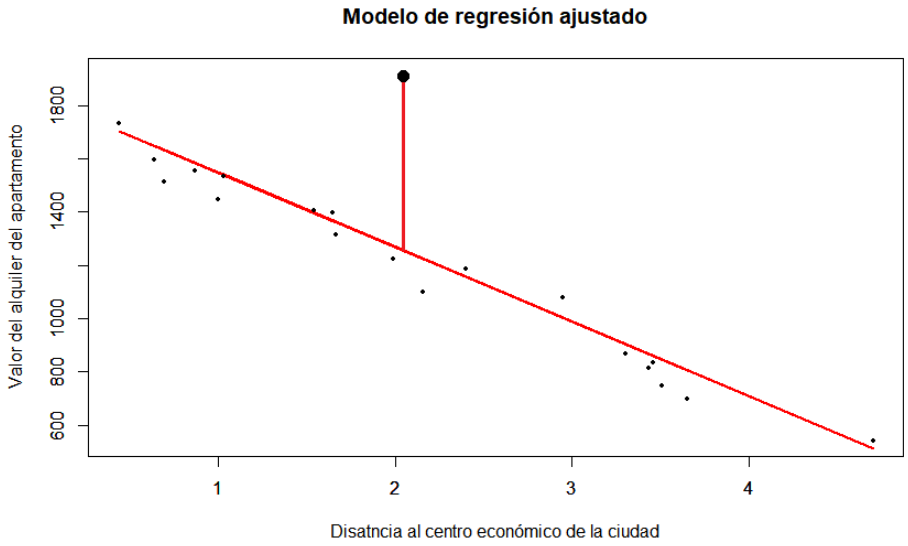
estadística para afirmar que el coeficiente de correlación no es de 0.5.

$$\begin{cases} Y_k = \mu_k + e_k \\ \mu_k = \beta_0 + \beta_1 x_k \\ e_k \sim N(0, \sigma^2) \\ e_1, e_2, ..., e_k \quad \text{independientes} \end{cases}$$

Donde:

- $\beta_0 = 1829{,}01$ se inerpreta de manera que si existiera un apartamento ubicado exactamente en el centro de la ciudad, este tendría un valor en promedio de 1'828'010 pesos.
- $\beta_1 = -279{,}860$ se interpreta como el decremento promedio del valor del alquiler por cada kilometro de distancia que aumenta.
- $\sigma$ se interpreta como las distancia promedio entre el valor predicho por el modelo y el valor observado.

b) Esto es precisamente la interpretación que dimos de la pendiente en la componente sistemática del modelo. El valor de disminución promedio a medida que la distancia al centro economico aumenta es de 279'860.
c)Podemos responder de nuevo esta pregunta, apelando únicamente a la interpretación de la recta en la componente sistemática del modelo. Dado que la distancia disminuye igualmente en ambos casos, tendremos que el aumento promedio será el mismo. Esto es, 279'860.
d) Consideraremos que un apartamento es "caro" o "barato" sí resulta ser un dato muy alejado de la recta del modelo de regresión. Observamos que el único apartamento que parece tener este comportamiento es el que está ubicado a una distancia de $2.05km$ del centro. El valor del alquiler predicho por el modelo para este apartamento es de 1'255'295 y su valor real es de 1'921'000. Su diferencia de precio es de 665'708, lo que es aproximadamente 4 desviaciones estándar del error. Por esto, este incremento de valor no puede ser explicado por el modelo.




e)Siguiendo la idea del razonamiento anterior, un apartamento será más caro o barato si su valor real es mayor o menor que el pronóstico puntual, respectivamente. Observamos que el valor predicho para el apartamento ubicado a $1.6km$ es de 1'381'233 y de 1'269'288 para el otro. Así, parece resultar ser más conveniente optar por el apartamento ubicado a $2km$ puesto que este puede considerarse como más "barato" ante la mirada del modelo.

f)Notemos que no es posible construir un intervalo de confianza para la media puesto que por construcción este está planteado únicamente para datos que se encuentren en la muestra. Por otra parte, obtenemos el intervalo de predicción mediante el uso de las siguientes fórmulas:

$$\begin{aligned} IP_{95\,\%}(Y_k^*) &= \hat{\mu}_k^* \mp t(n-2)\sqrt{\hat{\sigma}^2 + \hat{Var}(\hat{\mu}_k^*)} \\ &= 1409{,}219 \mp t(n-2)\sqrt{\hat{\sigma}^2 + \hat{Var}(\hat{\mu}_k^*)} \\ &= [1'037'591, 1'780'848] \end{aligned}$$

Lo anterior fue calculado con los siguientes comando en R:

```
#Intervalo de prediccion al 95%
sigma2 <- sum(fit$residuals^2)/(n-2)
mu <- coef(fit)[1] + coef(fit)[2]*x
var <- sigma(fit)^2 + vcov(fit)[1,1] + x*x*vcov(fit)[2,2] + 2*x*vcov(fit)
    [1,2]
Li <- mu - qt(1-alpha/2,n-2)*sqrt(var)
Ls <- mu + qt(1-alpha/2,n-2)*sqrt(var)

names(Li) <- " "
names(Ls) <- " "
cbind(Li,Ls)
```

g)Validación supuestos del modelo:

**Media cero**

Juzgaremos el siguiente sistema hipótesis para garantizar el supuesto de media cero en los errores:

$$\begin{cases} H_0 : \mu_{e_k} = 0 \\ H_1 : \mu_{e_k} \neq 0 \end{cases}$$

Realizando una prueba $t$, R nos proporciona un p-value=1 para este test. De modo que con un 95 % de significancia no se rechaza la hipótesis nula y este supuesto resulta válido.

**Homoscedasticidad**

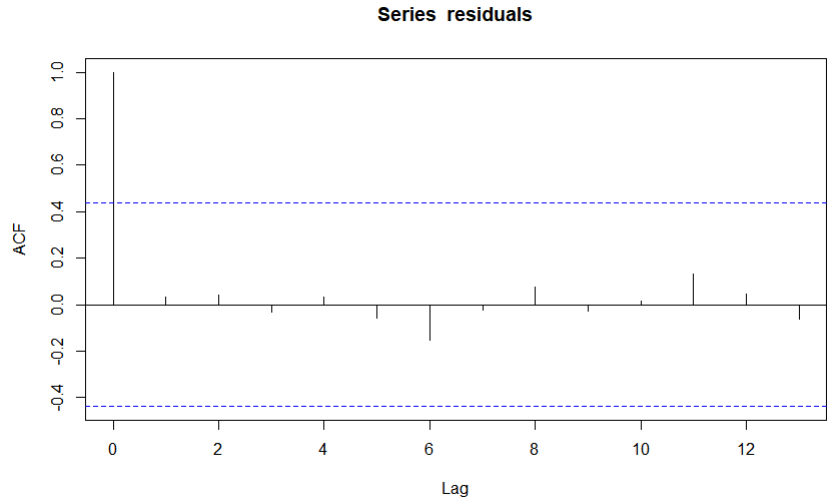
Mediante el test Breusch-Pagan juzgamos el sistema de hipótesis:

$$\begin{cases} H_0 : \text{hay homoscedasticidad (igual varianza)} \\ H_1 : \text{no hay homoscedasticidad} \end{cases}$$

Donde obtenemos un p-value=0.9308. Así, con un 95 % de significancia no se rechaza la hipótesis nula y este supuesto resulta válido.

**Correlación**

Haciendo uso de la función de autocorrelación simple, obtenemos el siguiente gráfico:




En este gráfico, se establecen dos lineas horizontales punteadas que representan una confianza del 95 %. Cuando las barras verticales, sobrepasan estas lineas punteadas se tiene correlacion en un periodo de tiempo indicado con dicho nivel de confianza. Unicamente una línea vertical hace esto, pero esta indica autocorrelación con ella misma. Este gráfico nos proporciona evidencia para decir que no hay correlación serial entre los errores del modelo. Sin embargo, debemos corroborar esto con una prueba de hipótesis:

```
> dwtest(fit)
  Durbin-Watson test
data:  fit
DW = 1.9063, p-value = 0.4194
alternative hypothesis: true autocorrelation is greater than 0
```

Concluimos con una confianza del 95 % que no hay autocorrelación serial entre los errores.

**Normalidad**

Juzgaremos el siguiente sistema de hipótesis:

$$\begin{cases} H_0 : \text{Los datos provienen de una normal} \\ H_1 : \text{Los datos no provienen de una normal} \end{cases}$$

Usando los test de Shapiro y Jarque-Bera, concluimos con un 95 % de confianza que nuestros datos no provienen de una normal. Este supuesto falla.

```
  Jarque Bera Test

data:  residuals
X-squared = 140.27, df = 2, p-value < 2.2e-16

  Shapiro-Wilk normality test

data:  residuals
W = 0.55943, p-value = 1.15e-06
```

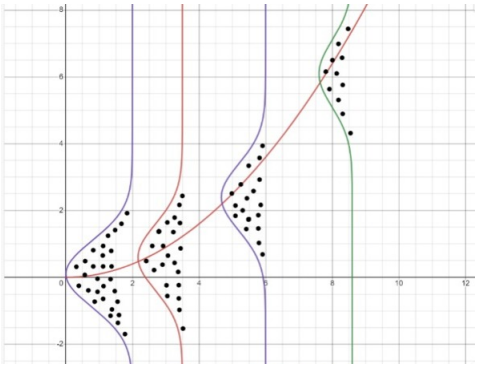
Por lo tanto, el modelo no se puede validar, hay que volver a la etapa de identificación y tratar el outlier que se obvservó anteriormente o proponer un modelo resistente a valores extremos.

**3.**

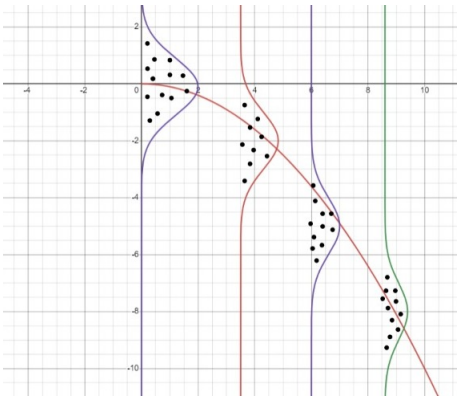
$$\begin{cases} Y_k = \mu_k + e_k \\ \mu_k = \beta x_k^2 \\ e_k \sim N(0, \frac{\sigma^2}{x_k}) \\ e_1, e_2, ..., e_k \quad \text{independientes} \end{cases}$$

**a.**

Para $\beta > 0$



Para $\beta > 0$



**b.** Veamos que se tiene heteroscedasticidad, ya que $Var(e_i) = \sigma^2 f(x_i)$, para todo $i$, con $f(x) = \frac{1}{x}$. Entonces, se hallará el estimador $\hat{\beta}$ por el método de mínimos cuadrados ponderados.

$\hat{\beta} = arg\ min\ Q_w(\beta) = \sum_{k=1}^{n} w_k(y_k - \beta x_k^2)^2$, con $w_k = \frac{x_k}{\sigma^2}$.

$$\frac{dQ}{d\beta} = -2\sum_{k=1}^{n} \frac{x_k}{\sigma^2} x_k^2 (y_k - \hat{\beta} x_k^2) = 0$$

$\sum_{k=1}^{n} \frac{x_k^3 y_k}{\sigma^2} - \sum_{k=1}^{n} \frac{\hat{\beta} x_k^5}{\sigma^2} = 0$. Luego, $\hat{\beta} = \dfrac{\sum_{k=1}^{n} y_k x_k^3}{\sum_{k=1}^{n} x_k^5}$.

$\frac{dQ^2}{d\beta} = 2\sum_{k=1}^{n} \frac{x_k^5}{\sigma^2} > 0$ ya que $x_k > 0$ y $\sigma^2 > 0$, por lo tanto, $\hat{\beta}$ es un mínimo local. Como la segunda derivada no depende en su expresión de $\hat{\beta}$, Q es convexa y es mínimo global.

Para este caso no todas las observaciones aportan información de la misma calidad, ya que las observaciones de $x_k$ más grandes tienen una menor varianza, luego, aportan mejor información, por lo tanto, a la hora de estimar, estas tienen más peso con $w_k = \frac{1}{Var(e_k)} = \frac{x_k}{\sigma^2}$.

**c. Valor esperado $\hat{\beta}$:**

$$E(\hat{\beta}) = E\left(\frac{\sum_{k=1}^{n} y_k x_k^3}{\sum_{k=1}^{n} x_k^5}\right) = \frac{1}{\sum_{k=1}^{n} x_k^5} E\left(\sum_{k=1}^{n} y_k x_k^3\right) = \frac{1}{\sum_{k=1}^{n} x_k^5} \sum_{k=1}^{n} E\left(x_k^3 y_k\right) = \frac{1}{\sum_{k=1}^{n} x_k^5} \sum_{k=1}^{n} x_k^3 \dot{\beta} x_k^2$$

$$= \frac{\sum_{k=1}^{n} \beta x_k^5}{\sum_{k=1}^{n} x_k^5} = \beta.$$

**Varianza de $\hat{\beta}$:**

$$Var(\hat{\beta}) = Var\left(\frac{\sum_{k=1}^{n} y_k x_k^3}{\sum_{k=1}^{n} x_k^5}\right) = \frac{1}{\left(\sum_{k=1}^{n} x_k^5\right)^2} Var\left(\sum_{k=1}^{n} y_k x_k^3\right).$$

Como las v.a $Y_k$ son independientes,

$$= \frac{1}{\left(\sum_{k=1}^{n} x_k^5\right)^2} \sum_{k=1}^{n} Var\left(y_k x_k^3\right) = \frac{1}{\left(\sum_{k=1}^{n} x_k^5\right)^2} \sum_{k=1}^{n} x_k^6 \cdot \frac{\sigma^2}{x_k} = \frac{\sigma^2}{\sum_{k=1}^{n} x_k^5}$$

**Distribución de $\hat{\beta}$:**

El estimador $\hat{\beta}$ se puede escribir como combinación lineal de variables aleatorias $Y_i$ con distribución normal, $\hat{\beta} = \sum_{i=1}^{n} \phi_i Y_i$, con $\phi_i = \frac{x_i^3}{\sum_{k=1}^{n} x_k^5}$, entonces, el estimador $\hat{\beta}$ tiene distribución normal. $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{k=1}^{n} x_k^5}\right)$.

Como $\hat{\beta}$ es un estimador insesgado, $MSE(\hat{\beta}) = Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{k=1}^{n} x_k^5} = \frac{\sigma^2}{n(\bar{x^5})}$.

$\lim_{n\to\infty} MSE(\hat{\beta}) = 0$. Por lo tanto, $\hat{\beta}$ es consistente.

**d.** Se tiene que $\mu^* = \beta(x^*)^2$, entonces $\hat{\mu^*} = \hat{\beta}(x^*)^2$. Definimos $e^* = Y^* - \hat{\mu^*}$.
$E(e^*) = E(Y^* - \hat{\mu^*}) = 0$
$Var(e^*) = Var(Y^* - \hat{\mu^*})$. Como $Y^*$ es una observación no incluida en la muestra, en-

tonces $\hat{\beta}(x^*)^2 = \hat{\mu^*}$ es independiente de $Y^*$, así, $Var(Y^* - \hat{\mu^*}) = Var(Y^*) + Var(\hat{\mu^*}) = \frac{\sigma^2}{x^*} + Var(\hat{\beta}(x^*)^2) = \frac{\sigma^2}{x^*} + \frac{(x^*)^4\sigma^2}{\sum_{k=1}^n x_k^5}$. Como $Y^*$ y $\hat{\mu^*}$ tienen distribución normal, la variable pivote

$$\frac{e^*}{\sqrt{\frac{\sigma^2}{x^*} + \frac{(x^*)^4\sigma^2}{\sum_{k=1}^n x_k^5}}} \sim N(0,1)$$

Por lo tanto, dado que $\sigma^2$ es una cantidad conocida, $IP_{100(1-\alpha)\,\%}(Y^*) = \hat{\mu^*} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{x^*} + \frac{(x^*)^4\sigma^2}{\sum_{k=1}^n x_k^5}}$

**e.** Sea $R_k = Y_k - \hat{\mu}_k = Y_k - \hat{\beta}x_k^2$. Se tiene que $\hat{\beta}$ se puede escribir de la forma $\hat{\beta} = \sum_{i=1}^n \phi_i Y_i$, entonces, $R_k = Y_k - \sum_{i=1}^n \phi_i Y_i x_k^2$, separando la suma, $R_k = Y_k - x_k^2\phi_k Y_k + \sum_{i\neq j}^n \phi_i Y_i x_k^2 = (1 - x_k^2\phi_k)Y_k + \sum_{i\neq j}^n \phi_i Y_i x_k^2$. Por lo tanto, $R_k$ tiene una distribución normal, ya que es combinación lineal de variables aleatorias $Y_i$ con distribución normal.

$E(R_k) = E(Y_k) - E(\hat{\beta}x_k) = \beta x_k^2 - \beta x_k^2 = 0$

$Var(R_k) = Var\left((1 - x_k^2\phi_k)Y_k + \sum_{i\neq j}^n \phi_i Y_i x_k^2\right)$. Como estos términos son independientes ya que en la sumatoria no está el término para $Y_k$, entonces,

$Var(R_k) = Var(Y_k) - 2x_k^2\phi_k Var(Y_k) + x_k^4\phi_k^2 Var(Y_k) + Var\left(\sum_{i\neq j}^n x_k^2\phi_i Y_i\right)$

$= \frac{\sigma^2}{x_k} - \frac{2x_k^2 x_k^3\sigma^2}{x_k\left(\sum_{j=1}^n x_j^5\right)^2} + \frac{x_k^4 x_k^6\sigma^2}{x_k\sum_{j=1}^n x_j^5} + x_k^4\sum_{i\neq j}^n Var(\phi_i Y_i)$. (son independientes).

$= \frac{\sigma^2}{x_k} - \frac{2x_k^4\sigma^2}{\sum_{j=1}^n x_j^5} + \frac{x_k^4 x_k^5\sigma^2}{\left(\sum_{j=1}^n x_j^5\right)^2} + \frac{x_k^4\sum_{i\neq j}^n x_i^6\sigma^2/x_i}{\left(\sum_{j=1}^n x_j^5\right)^2} = \frac{\sigma^2}{x_k} - \frac{2x_k^4\sigma^2}{\sum_{j=1}^n x_j^5} + \frac{x_k^4\sigma^2}{\sum_{j=1}^n x_j^5}$

$= \frac{\sigma^2}{x_k} - \frac{x_k^4\sigma^2}{\sum_{j=1}^n x_j^5}$.

Planteando la siguiente prueba de hipótesis para determinar si el k-ésimo valor observado es un valor atípico. Usando $R_k = Y_k - \hat{\mu}_k$.

$$\begin{cases} \text{H}_0\text{: La k-ésima observación no es un dato atípico} \\ \text{H}_1\text{: La k-ésima observación es un dato atípico} \end{cases}$$

Definimos el estadístico $z_c$ ($\sigma^2$ es conocida)

$$z_c = \frac{R_k}{\sqrt{\frac{\sigma^2}{x_k} - \frac{x_k^4\sigma^2}{\sum_{j=1}^n x_j^5}}} \sim N(0,1)$$

$\tau$ : Rechazar $H_0$ con un nivel de confianza de $100(1-\alpha)\,\%$ si $|z_c| > z_{1-\frac{\alpha}{2}}$