1. INTRODUCCIÓN AL ANÁLISIS MULTIVARIANTE

1.1 ESTADÍSTICA DESCRIPTIVA

- Introducción
- Estadística descriptiva aplicada a datos de marketing
- Lectura de base de datos
- Visualización gráfica
  - Boxplot, histograma y diagrama de dispersión
  - Diagrama de barras y de sectores
- Medidas descriptivas en variables cuantitativas

1.2 ANÁLISIS BIVARIANTE Y REGRESIÓN LINEAL SIMPLE

- Tablas de contingencia y la regresión lineal simple
- Aplicación a datos de marketing
- Lectura de base de datos
- Asociación entre dos variables cualitativas
- Asociación entre dos variables cuantitativas
- Modelo de regresión lineal simple: variable explicativa cuantitativa
  - Diagnóstico del modelo
  - Predicción del modelo
- Modelo de regresión lineal simple: variable cualitativa

1.3 REGRESIÓN LINEAL MÚLTIPLE

- El modelo de regresión múltiple
- Aplicación del modelo en datos de marketing
- Lectura de base de datos
- Modelo de regresión lineal y calidad del ajuste
- Selección de variables con coeficientes significativos
- Validación y diagnóstico del modelo
- Predicción
- Capacidad predictiva

T1. INTRODUCCIÓN AL ANÁLISIS MULTIVARIANTE

1.1 ESTADÍSTICA DESCRIPTIVA

PREPARACIÓN DEL ENTORNO DE TRABAJO

```
# Cargar librerías necesarias
library(ggplot2)
library(dplyr)
library(gridExtra)
```

LECTURAS BASE DE DATOS - DOS OPCIONES

OPCIÓN 1: Usando read.csv() (Base R)

```
ds <- read.csv("ChildCarSeats_clean.csv", header=TRUE, stringsAsFactors = TRUE)
```

OPCIÓN 2: Usando read_csv() (tidyverse)

```
library(readr)
ChildCarSeats_clean <- read_csv("~/Desktop/ANÁLISIS
MULTIVARIANTE/ChildCarSeats_clean.csv")
ds <- ChildCarSeats_clean
```

Verificar los Datos

```
str(ds)
```

Salida esperada:

```
'data.frame': 100 obs. of  11 variables:
 $ Sales      : num  8.45 6.13 9.18 5.30 7.83 ...
 $ CompPrice  : int  148 116 117 104 124 98 132 87 156 119 ...
 $ Income     : int  65 45 42 73 55 89 56 94 38 82 ...
 $ Advertising: int  3 0 4 0 2 8 0 12 5 6 ...
 $ Population : int  219 232 177 176 218 156 289 245 323 198 ...
 $ Price      : int  125 146 111 143 104 98 132 87 156 119 ...
 $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 3 3 1 3 3 ...
 $ Age        : int  64 45 33 61 48 45 52 51 35 58 ...
 $ Education  : int  11 14 14 11 15 15 13 10 16 12 ...
 $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 1 2 2 ...
 $ US         : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 2 2 2 1 2 ...
```

INTRODUCCIÓN AL ANÁLISIS DESCRIPTIVO

El análisis descriptivo es el primer paso antes de aplicar técnicas multivariantes. Permite:

- **Familiarizarse con los datos**: Entender qué información contiene nuestro dataset
- **Identificar tipos de variables**: Distinguir entre cuantitativas y cualitativas
- **Detectar valores ausentes y outliers**: Localizar datos faltantes o atípicos
- **Conocer relaciones básicas entre variables**: Identificar patrones iniciales

Variables del Dataset

Las variables del conjunto de datos son:

Variables Cuantitativas:

- **Sales**: ventas unitarias, en miles, en cada ubicación
- **CompPrice**: precio que cobra la competencia en cada ubicación
- **Income**: nivel de ingresos comunitarios, en miles de dólares
- **Advertising**: presupuesto de publicidad local, en miles de dólares
- **Population**: tamaño de la población en la región, en miles
- **Price**: precio de las sillitas de coche en cada ubicación
- **Age**: edad media de la población local
- **Education**: media del nivel de educación (años) de la población

Variables Cualitativas:

- **ShelveLoc**: calidad de ubicación de las sillas en la tienda (Bad, Good, Medium)
- **Urban**: si la tienda está en ubicación urbana o rural (Yes, No)
- **US**: si la tienda está en EUA o no (Yes, No)

Clasificación de Variables

```
# Variables cuantitativas
varnum <- data.frame(ds$Sales, ds$CompPrice, ds$Income, ds$Advertising,
                     ds$Population, ds$Price, ds$Age, ds$Education)
names(varnum) <- c("Sales","CompPrice","Income","Advertising",
                   "Population","Price","Age","Education")

# Variables cualitativas
varfac <- data.frame(ds$ShelveLoc, ds$Urban, ds$US)
names(varfac) <- c("ShelveLoc","Urban","US")

print("Variables cuantitativas:")
print(names(varnum))
print("Variables cualitativas:")
print(names(varfac))
```

MEDIDAS DESCRIPTIVAS PARA VARIABLES CUANTITATIVAS

1. Medidas de Tendencia Central

- **Media**: Suma de todos los valores dividida por el número total de casos. Sensible a valores extremos.
- **Mediana**: Valor que deja el 50% de los datos por encima y 50% por debajo. Más robusta que la media.

MÉTODO RÁPIDO:

```
mean(ds$Sales)          # Media
median(ds$Sales)        # Mediana
```

MÉTODO TRADICIONAL:

```
# Media
m1 <- mean(varnum[,1])  # Sales
m2 <- mean(varnum[,2])  # CompPrice
m3 <- mean(varnum[,3])  # Income
(medias <- c(m1,m2,m3))

# Mediana
me1 <- median(varnum[,1])  # Sales
me2 <- median(varnum[,2])  # CompPrice
me3 <- median(varnum[,3])  # Income
(medianas <- c(me1,me2,me3))
```

2. Medidas de Dispersión

- **Desviación estándar**: Mide cuánto se alejan, en promedio, los datos respecto a la media.
- **Rango intercuartílico (RIC)**: Q3 - Q1. Mide la dispersión del 50% central de los datos.

```
# MÉTODO RÁPIDO
sd(ds$Sales)           # Desviación estándar
IQR(ds$Sales)          # Rango intercuartílico
```

3. Medidas de Posición

Cuartiles: Dividen los datos ordenados en cuatro partes iguales:

- **Q1 (25%)**: Una cuarta parte de los datos está por debajo
- **Q2 (50%)**: Es la mediana
- **Q3 (75%)**: Tres cuartas partes están por debajo

```
# MÉTODO RÁPIDO
quantile(ds$Sales)     # Cuartiles
summary(ds$Sales)      # Resumen completo
```

Ejemplo de salida:

```
> mean(ds$Sales)
[1] 8.43
> median(ds$Sales)
[1] 8.50
> quantile(ds$Sales)
   0%   25%   50%   75%  100%
 2.05  6.80  8.50 10.96 15.00
> summary(ds$Sales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.05    6.80    8.50    8.43   10.96   15.00
```

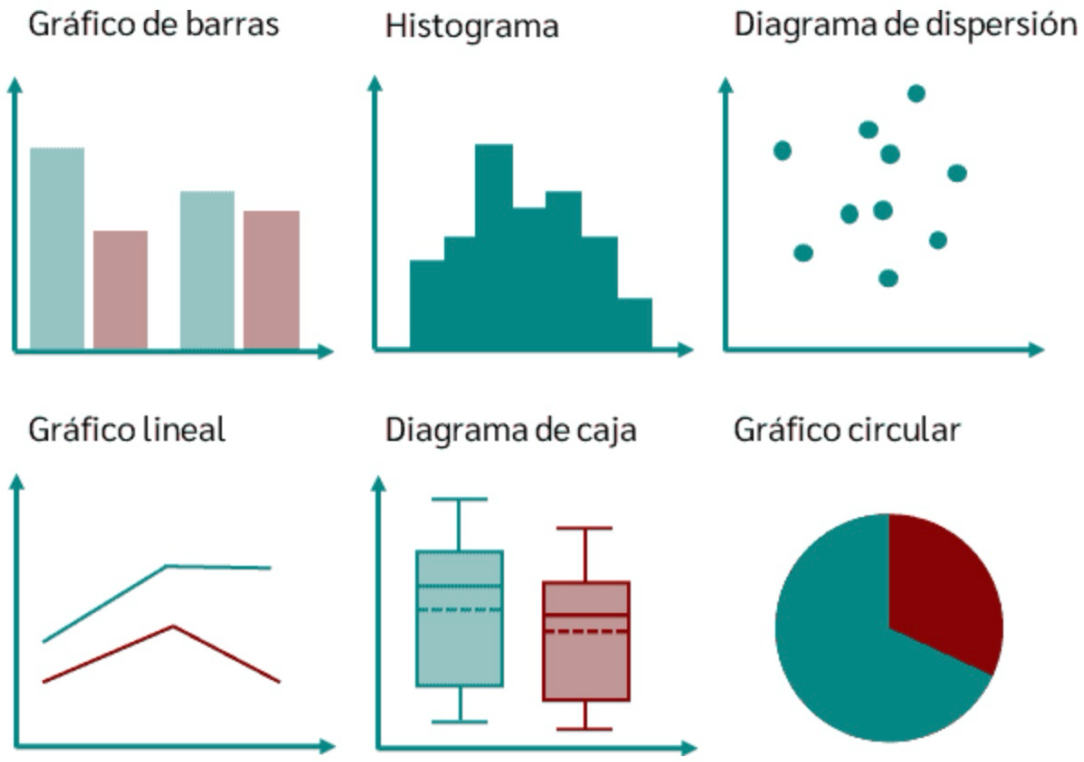
Resumen de Múltiples Variables

```
# MÉTODO TRADICIONAL (para todas las variables cuantitativas)
summary(varnum)
sapply(varnum, sd, na.rm = TRUE)      # Desviaciones estándar
sapply(varnum, IQR, na.rm = TRUE)     # Rangos intercuartílicos
```

Salida esperada:

```
     Sales          Population        Price          CompPrice
 Min.   : 2.05   Min.   : 57.00   Min.   : 63.0   Min.   : 91.0
 1st Qu.: 6.80   1st Qu.:145.25   1st Qu.: 97.0   1st Qu.:110.0
 Median : 8.50   Median :222.50   Median :111.0   Median :129.0
 Mean   : 8.43   Mean   :236.18   Mean   :110.9   Mean   :129.1
 3rd Qu.:10.96   3rd Qu.:321.00   3rd Qu.:124.0   3rd Qu.:147.0
 Max.   :15.00   Max.   :420.00   Max.   :157.0   Max.   :178.0
```

VISUALIZACIÓN GRÁFICA



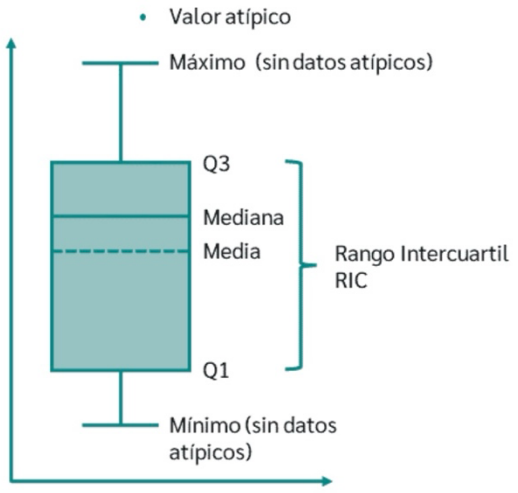
VARIABLES CUANTITATIVAS

Gráficos Individuales

- **Boxplot: Muestra mediana, cuartiles, outliers y forma de distribución**

```
ggplot(ds, aes(y=Sales)) + geom_boxplot()
```

- **Caja central**: del Q1 al Q3
- **Línea central**: mediana
- **Bigotes**: hasta 1.5×RIC
- **Puntos**: outliers



La caja indica el intervalo en el que se encuentra el 50% de los datos.

Por lo tanto, el extremo inferior de la caja es el 1er cuartil y el extremo superior es el 3er cuartil.

Entre Q1 y Q3, está el rango intercuartil (RIC)

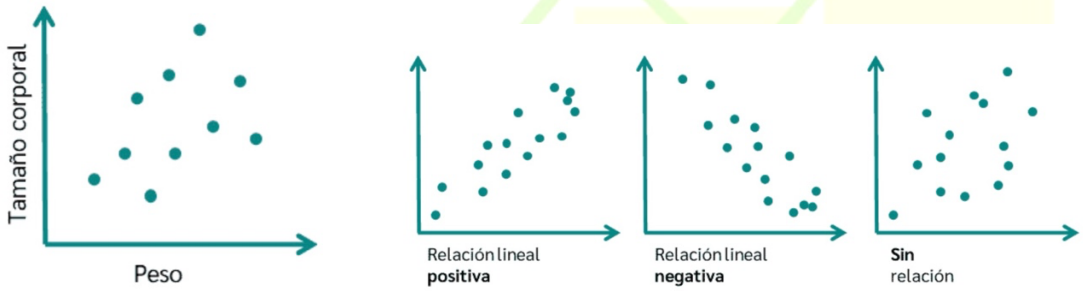
En el diagrama de caja, la línea continua indica la mediana y la línea discontinua, la media.

Los bigotes en forma de T se extienden hasta los valores máximo y mínimo que siguen estando dentro de 1,5 veces el rango intercuartílico (RIC).

Los puntos que están aún más alejados se consideran valores atípicos.

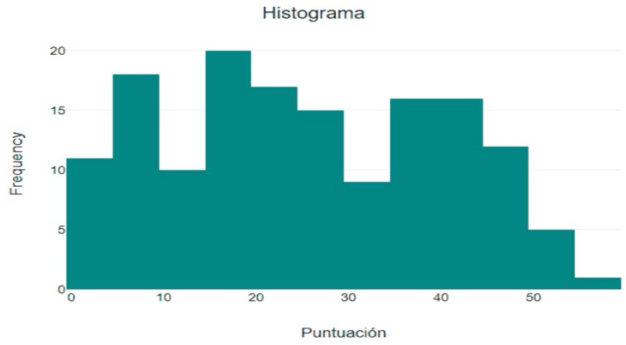
- **Scatter plot: Muestra relación entre dos variables cuantitativas**

```
ggplot(ds, aes(x=Advertising, y=Sales)) + geom_point()
```



- **Histograma: Muestra la forma de la distribución**

```
ggplot(ds, aes(x=Income)) + geom_histogram()
```



Combinar Gráficos

```
# Crear gráficos por separado
g1 <- ggplot(ds, aes(y=Sales)) + geom_boxplot()
g2 <- ggplot(ds, aes(x=Advertising, y=Sales)) + geom_point()

# Juntarlos
grid.arrange(g1, g2, nrow=1)  # En fila
grid.arrange(g1, g2, ncol=1)  # En columna
```

VARIABLES CUALITATIVAS

Tablas de Frecuencias

- **Frecuencia absoluta**: número de casos en cada categoría
- **Frecuencia relativa**: proporción del total que representa cada categoría

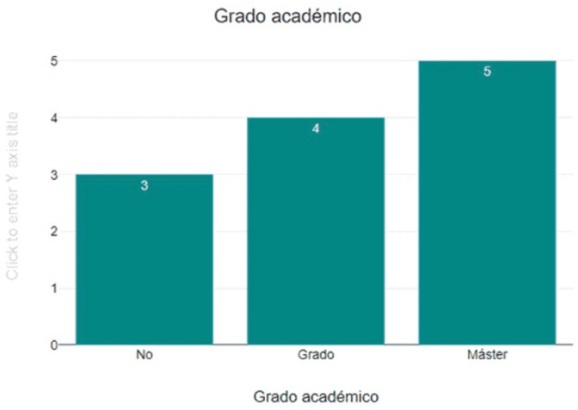
	Pastel	Helados	Donut	Total
Mujer	4	3	6	13
Hombre	5	7	9	21
Total	9	10	15	34

```
# MÉTODO RÁPIDO
table(ds$ShelveLoc)                          # Frecuencias absolutas
prop.table(table(ds$ShelveLoc)) * 100        # Frecuencias relativas (%)
summary(ds$ShelveLoc)                        # Resumen
```

Ejemplo de salida:

```
> table(ds$ShelveLoc)
   Bad   Good Medium
    30     20     50
> prop.table(table(ds$ShelveLoc)) * 100
   Bad   Good Medium
    30     20     50
```

- **Gráfico de barras: para comparar frecuencias entre categorías**




```
SumShelve <- summarize(group_by(ds, ShelveLoc), n=length(ShelveLoc), Sales=mean(Sales))

g1 <- ggplot(SumShelve, aes(x="", y=n, fill=ShelveLoc)) +
     geom_bar(width = 1, stat = "identity") + ggtitle("ShelveLoc")
g2 <- ggplot(SumShelve, aes(x=ShelveLoc, y=Sales, fill=ShelveLoc)) +
     geom_bar(width = 1, stat = "identity")
grid.arrange(g1, g2, nrow=1)
```

- **Gráfico de sectores (circular): para mostrar proporciones del total**

```
SumUS <- summarize(group_by(ds, US), n=length(US), Sales=mean(Sales))

g5 <- ggplot(SumUS, aes(x="", y=n, fill=US)) + coord_polar("y", start=0) +
     geom_bar(width = 1, stat = "identity") + ggtitle("US")
g6 <- ggplot(SumUS, aes(x=US, y=Sales, fill=US)) +
     geom_bar(width=1, stat="identity")
grid.arrange(g5, g6, nrow=1)
```

Nota: `coord_polar("y", start=0)` convierte el gráfico de barras en un diagrama circular.

ANÁLISIS BIVARIANTE

Correlación entre Variables Cuantitativas

Coeficiente de correlación: Mide la fuerza y dirección de la relación lineal entre dos variables cuantitativas.

- **Valores entre -1 y +1**
- **r > 0**: Correlación positiva (cuando una variable aumenta, la otra también)
- **r < 0**: Correlación negativa (cuando una aumenta, la otra disminuye)
- **r ≈ 0**: Sin relación lineal

Interpretación de la fuerza:

- **|r| > 0.7**: correlación fuerte
- **0.3 < |r| < 0.7**: correlación moderada
- **|r| < 0.3**: correlación débil

```
# Correlación entre dos variables
cor(ds$Sales, ds$Price)

# Matriz de correlaciones (método tradicional)
matriz_cor <- cor(varnum, use = "complete.obs")
print(round(matriz_cor, 3))
```

Tablas de Contingencia (Variables Cualitativas)

Test chi-cuadrado: Evalúa si existe dependencia entre las variables

- **H_0**: Las variables son independientes (no hay relación)
- **H_1**: Las variables son dependientes (sí hay relación)
- **Criterio**: Si p-valor < 0.05 → Rechazamos H_0 (hay dependencia significativa)

```
# Tabla de contingencia
tabla_contingencia <- table(ds$ShelveLoc, ds$Urban)
print("Tabla de contingencia ShelveLoc vs Urban:")
print(tabla_contingencia)

# Test chi-cuadrado
test_chi <- chisq.test(tabla_contingencia)
print(test_chi)
```

1.2 ANÁLISIS BIVARIANTE Y REGRESIÓN LINEAL SIMPLE

INTRODUCCIÓN

El análisis bivariante estudia la **relación entre dos variables** para determinar si existe asociación estadísticamente significativa.

Objetivos principales:

- Identificar si existe relación entre variables
- Cuantificar la fuerza de esa relación
- Predecir valores de una variable a partir de otra

Tipos de análisis según variables:

- **Dos variables cualitativas** → Tablas de contingencia
- **Dos variables cuantitativas** → Correlación
- **Una cuantitativa + una cualitativa** → Regresión lineal

ASOCIACIÓN ENTRE DOS VARIABLES CUALITATIVAS

Tabla de Contingencia

Una tabla de contingencia muestra la **distribución conjunta** de dos variables categóricas.

MÉTODO RÁPIDO:

```
# Crear tabla de contingencia
table(ds$Urban, ds$US)
```

Salida esperada:

```
     No  Yes
No   46   72
Yes  96  186
```

Interpretación:

- **Filas**: Variable Urban (No/Yes)
- **Columnas**: Variable US (No/Yes)
- **Celda (Yes, No)**: 96 tiendas urbanas fuera de USA

Test Chi-Cuadrado de Independencia

¿Para qué sirve? Evalúa si existe dependencia estadística entre dos variables cualitativas.

Hipótesis:

- H_0: Las variables son independientes (no hay asociación)
- H_1: Las variables son dependientes (sí hay asociación)

```
# Test de independencia
chisq.test(ds$Urban, ds$US)
```

Salida esperada:

```
Pearson's Chi-squared test with Yates' continuity correction

data: ds$Urban and ds$US
X-squared = 0.68416, df = 1, p-value = 0.4082
```

Interpretación del resultado:

- **Estadístico**: $X^2 = 0.684$
- **p-valor**: 0.408
- **Conclusión**: $p > 0.05 \rightarrow$ **No se rechaza H_0**
- Las variables Urban y US son **independientes**

CRITERIO DE DECISIÓN:

- **p < 0.05** → Variables dependientes (hay asociación)
- **p ≥ 0.05** → Variables independientes (no hay asociación)

ASOCIACIÓN ENTRE DOS VARIABLES CUANTITATIVAS

Coeficiente de Correlación de Pearson

Fórmula:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

¿Qué mide? La fuerza y dirección de la relación lineal entre dos variables cuantitativas.

```
# Correlación entre Sales y Price
cor(ds$Sales, ds$Price)
```

Salida esperada:

```
[1] -0.4285807
```

Interpretación de la Correlación

Signo:

- **r > 0**: Correlación positiva (cuando una ↑, la otra ↑)
- **r < 0**: Correlación negativa (cuando una ↑, la otra ↓)

Fuerza de la relación:

- **|r| > 0.7**: Correlación fuerte
- **0.3 < |r| < 0.7**: Correlación moderada
- **|r| < 0.3**: Correlación débil

En nuestro ejemplo:

- **r = -0.428**: Correlación moderada y negativa
- **Interpretación**: A mayor precio, menores ventas

Matriz de Correlaciones

```
# Seleccionar variables cuantitativas
varnum <- ds[, c("Sales", "CompPrice", "Income", "Price")]
```

```
# Matriz de correlaciones
cor_matrix <- cor(varnum)
print(round(cor_matrix, 3))
```

MODELO DE REGRESIÓN LINEAL SIMPLE: VARIABLE CUANTITATIVA

¿Qué es la Regresión Lineal Simple?

Un modelo que predice una **variable dependiente (Y)** a partir de una **variable independiente (X)** mediante una línea recta:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- β_0: Intercepto (valor de Y cuando X = 0)
- β_1: Pendiente (cambio en Y por cada unidad de X)
- ε: Error aleatorio

- **ESTIMACIÓN DEL MODELO**

Objetivo: Predecir Sales a partir de Price

```
# Crear el modelo
Model_1 <- lm(Sales ~ Price, data = ds)

# Ver resultados
summary(Model_1)
```

Salida esperada:

```
Call:
lm(formula = Sales ~ Price, data = ds)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4404 -1.7815 -0.1222  1.5811  7.6830

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.137603   0.617735  21.267  < 2e-16 ***
Price       -0.049464   0.005227  -9.463  < 2e-16 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.472 on 398 degrees of freedom
Multiple R-squared:  0.1837,     Adjusted R-squared:  0.1816
F-statistic: 89.55 on 1 and 398 DF,  p-value: < 2.2e-16
```

○ INTERPRETACIÓN DE LOS RESULTADOS

Coeficientes:

- **Intercepto**: 13.138 → Ventas cuando precio = 0
- **Price**: -0.049 → Por cada $ que sube el precio, las ventas bajan 0.049 miles de unidades

Significancia:

- **p < 0.05** → Coeficientes estadísticamente significativos ✓

Calidad del Modelo:

- **R^2**: 0.1837 → El modelo explica el **18.37%** de la variabilidad de las ventas
- **F-statistic**: Significativo → El modelo es válido globalmente

○ DIAGNÓSTICO DEL MODELO

¿Por qué es necesario el diagnóstico?

Para verificar que se cumplen los **supuestos de la regresión lineal**:

1. **Linealidad**: La relación es lineal
2. **Homocedasticidad**: Varianza constante de los errores
3. **Normalidad**: Los residuos siguen distribución normal
4. **Independencia**: Los errores son independientes

Gráficos de Diagnóstico

```
# Calcular residuos y valores ajustados
residuos <- rstandard(Model_1)
valor_ajustados <- fitted(Model_1)

# Configurar gráficos en 1 fila y 2 columnas
par(mfrow = c(1,2))

# Gráfico 1: Residuos vs Valores Ajustados
plot(valor_ajustados, residuos,
     main = "Residuos vs Valores Ajustados",
     xlab = "Valores Ajustados",
     ylab = "Residuos Estandarizados")

# Gráfico 2: Q-Q Plot (Normalidad)
qqnorm(residuos, main = "Q-Q Plot")
qqline(residuos)
```

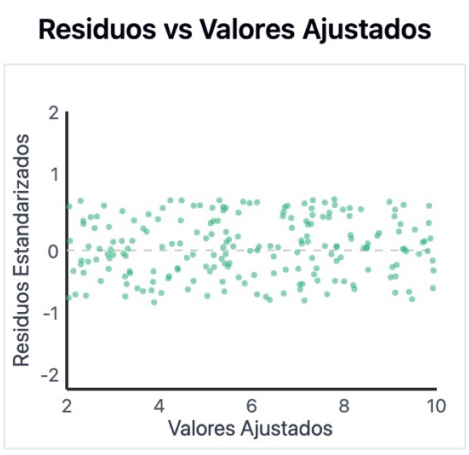
Interpretación de los Gráficos

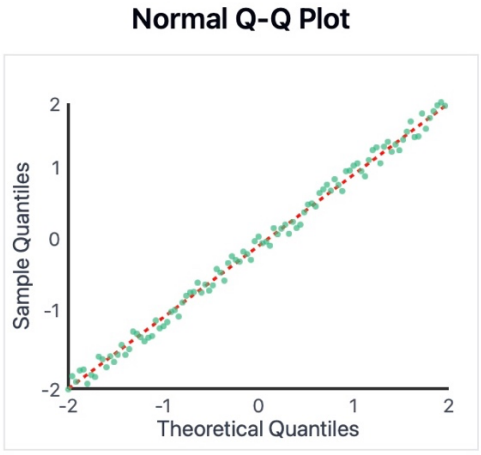
Gráfico 1: Residuos vs Valores Ajustados

- **Patrón aleatorio** → Supuestos de varianza constante (Homocedasticidad) ✓
- **Patrón sistemático** → Violación de supuestos ✗

Gráfico 2: Q-Q Plot

- **Puntos en línea recta** → Residuos normales ✓
- **Puntos alejados de la línea** → No normalidad ✗



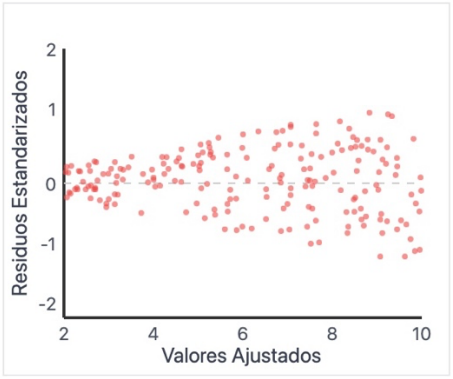


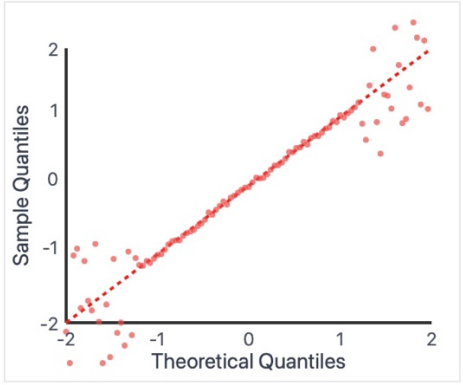
Interpretación:

Residuos vs Ajustados: Patrón aleatorio → Homocedasticidad ✓

Q-Q Plot: Puntos en línea recta → Normalidad ✓

Conclusión: Supuestos de regresión cumplidos





Interpretación:

Residuos vs Ajustados: Patrón de embudo → Heterocedasticidad ✗

Q-Q Plot: Puntos se alejan de la línea → No normalidad ✗

Conclusión: Supuestos violados, modelo no válido

○ PREDICCIÓN DEL MODELO

¿Cómo hacer predicciones?

Una vez validado el modelo, podemos predecir valores de Y para nuevos valores de X.

```
# Crear nuevo conjunto de datos
newdata <- data.frame(Price = 100)

# Realizar predicción
predict(Model_1, newdata)
```

Salida esperada:

```
       1
8.191251
```

Interpretación:

- **Precio = $100 → Ventas predichas = 8.19 miles de unidades**

Fórmula Manual

También podemos calcularlo con la ecuación: **Sales = 13.138 - 0.049 × Price**

```
# Predicción manual
13.138 - 0.049 * 100
```

MODELO DE REGRESIÓN LINEAL SIMPLE: VARIABLE CUALITATIVA

Variables Dummy

Cuando la variable independiente es cualitativa, R la convierte automáticamente en **variables dummy** (0 y 1).

Ejemplo: Variable US

- **US = "No"** → 0
- **US = "Yes"** → 1

Estimación del Modelo

Objetivo: Predecir Sales según ubicación (USA vs No-USA)

```
# Crear el modelo
Model_2 <- lm(Sales ~ US, data = ds)

# Ver resultados
summary(Model_2)
```

Salida esperada:

```
Call:
lm(formula = Sales ~ US, data = ds)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4969 -1.8794 -0.0498  1.8549  8.4031

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5798     0.2237  29.416  < 2e-16 ***
USYes         1.2871     0.2785   4.621 5.16e-06 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.665 on 398 degrees of freedom
Multiple R-squared:  0.05093,    Adjusted R-squared:  0.04854
F-statistic: 21.36 on 1 and 398 DF,  p-value: 5.157e-06
```

Interpretación de Resultados

Coeficientes:

- **Intercepto**: 6.58 → Ventas promedio fuera de USA (US = "No")
- **USYes**: 1.29 → **Las tiendas en USA venden 1.29 miles de unidades MÁS** que las de fuera

Ecuación del modelo:

- **Fuera de USA**: Sales = 6.58
- **En USA**: Sales = 6.58 + 1.29 = 7.87

Calidad del modelo:

- **R^2**: 0.049 → Solo explica el **4.9%** de la variabilidad
- **Modelo poco explicativo** pero **estadísticamente significativo**

1.3 REGRESIÓN LINEAL MÚLTIPLE

INTRODUCCIÓN

La regresión lineal múltiple extiende el modelo simple incluyendo **múltiples variables explicativas** para predecir una variable dependiente.

¿Por qué usar regresión múltiple?

- En la realidad, múltiples factores influyen en una variable
- Mejora la capacidad predictiva del modelo
- Permite controlar el efecto de otras variables
- Proporciona una visión más completa del fenómeno

Ecuación del modelo: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$

Donde:

- **Y**: Variable dependiente
- $X_1, X_2, ..., X_k$: Variables independientes
- $\beta_0, \beta_1, ..., \beta_k$: Coeficientes a estimar
- ε: Error aleatorio

MODELO DE REGRESIÓN LINEAL Y CALIDAD DEL AJUSTE

○ MODELO COMPLETO

Modelo con todas las variables:

```
# Modelo completo con todas las variables
model0 <- lm(Sales ~ CompPrice + Income + Advertising +
             Population + Price + ShelveLoc + Age +
             Education + Urban + US, data = ds)

# Ver resultados
summary(model0)
```

Salida esperada:

```
Call:
lm(formula = Sales ~ CompPrice + Income + Advertising + Population +
    Price + ShelveLoc + Age + Education + Urban + US, data = ds)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.131035   0.685776   7.482 4.94e-13 ***
CompPrice       0.086813   0.004714  18.418  < 2e-16 ***
Income          0.015485   0.002097   7.385 9.40e-13 ***
Advertising     0.122176   0.012641   9.665  < 2e-16 ***
Population      0.000386   0.000421   0.916    0.360
Price          -0.089181   0.003035 -29.380  < 2e-16 ***
ShelveLocGood   4.408189   0.174000  25.335  < 2e-16 ***
ShelveLocMedium 1.952346   0.143310  13.623  < 2e-16 ***
Age            -0.042871   0.003616 -11.857  < 2e-16 ***
Education      -0.002484   0.022411  -0.111    0.912
UrbanYes        0.051097   0.128389   0.398    0.691
USYes           0.091567   0.170285   0.538    0.591

Residual standard error: 1.158 on 388 degrees of freedom
Multiple R-squared:  0.8254,     Adjusted R-squared:  0.8204
F-statistic: 166.7 on 11 and 388 DF,  p-value: < 2.2e-16
```

Interpretación de Resultados

Calidad del modelo:

$R^2 = 0.8254$ → El modelo explica el 82.54% de la variabilidad de Sales

R^2 ajustado = 0.8204 → Ajustado por número de variables

F-statistic significativo → El modelo es válido globalmente

Significancia de variables:

$p < 0.05$ → Variable significativa ✓

$p \geq 0.05$ → Variable NO significativa ✗

Variables significativas: CompPrice, Income, Advertising, Price, ShelveLoc, Age Variables NO significativas:Population, Education, Urban, US

- **SELECCIÓN DE VARIABLES CON COEFICIENTES SIGNIFICATIVOS**

Modelo Reducido (Solo Variables Significativas)

Eliminar variables no significativas:

```
# Modelo solo con variables significativas
model <- lm(Sales ~ CompPrice + Income + Advertising +
            Price + ShelveLoc + Age, data = ds)

# Ver resultados
summary(model)
```

Salida esperada:

```
Call:
lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
    ShelveLoc + Age, data = ds)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.272634   0.571547   9.225  < 2e-16 ***
CompPrice       0.086566   0.004667  18.550  < 2e-16 ***
Income          0.015538   0.002080   7.471 5.23e-13 ***
Advertising     0.129107   0.008741  14.770  < 2e-16 ***
Price          -0.089049   0.003022 -29.468  < 2e-16 ***
ShelveLocGood   4.399322   0.172593  25.490  < 2e-16 ***
ShelveLocMedium 1.937682   0.141895  13.656  < 2e-16 ***
Age            -0.042928   0.003595 -11.940  < 2e-16 ***

Residual standard error: 1.154 on 392 degrees of freedom
Multiple R-squared:  0.8249,     Adjusted R-squared:  0.8217
F-statistic: 263.7 on 7 and 392 DF,  p-value: < 2.2e-16
```

○ VALIDACIÓN Y DIAGNÓSTICO DEL MODELO

Los 5 Supuestos de la Regresión Múltiple

1. **Linealidad**: Relación lineal entre X e Y
2. **Normalidad**: Residuos siguen distribución normal
3. **Homocedasticidad**: Varianza constante de residuos
4. **Independencia**: Residuos no correlacionados
5. **No multicolinealidad**: Variables X no altamente correlacionadas

Gráficos de Diagnóstico

```
# Configurar ventana de gráficos
par(mfrow = c(2,2))

# Generar los 4 gráficos de diagnóstico
plot(model)
```

Interpretación de Gráficos

1. Residuals vs Fitted (Linealidad)

- **Línea roja cercana a 0** → Linealidad OK ✓
- **Patrón sistemático** → Violación de linealidad ✗

2. Normal Q-Q (Normalidad)

- **Puntos en línea recta** → Normalidad OK ✓
- **Desviaciones en extremos** → Aceptable si no es severo

3. Scale-Location (Homocedasticidad)

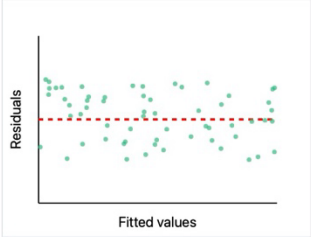
- **Línea horizontal** → Varianza constante ✓
- **Patrón de embudo** → Heterocedasticidad ✗

4. Residuals vs Leverage (Observaciones influyentes)

- **Puntos dentro de [-3,3]** → Sin outliers ✓
- **Puntos fuera del rango** → Observaciones influyentes ✗

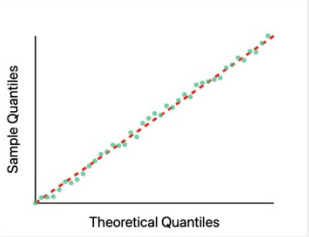
SUPUESTOS CUMPLIDOS

1. Residuals vs Fitted



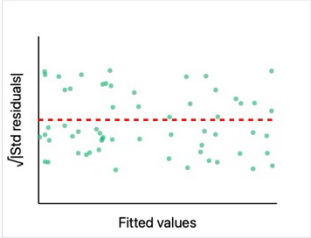
Supuesto cumplido

2. Normal Q-Q



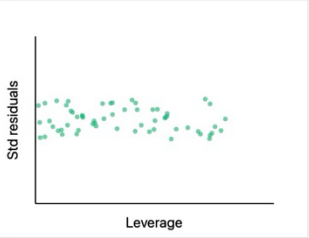
Supuesto cumplido

3. Scale-Location



Supuesto cumplido

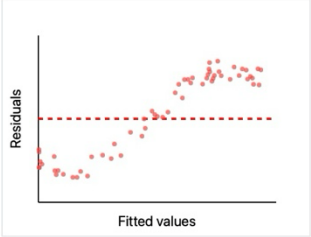
4. Residuals vs Leverage



Supuesto cumplido

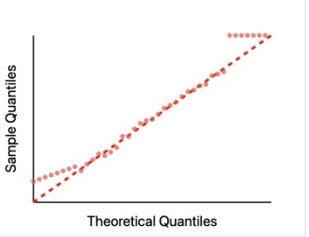
SUPUESTOS VIOLADOS

1. Residuals vs Fitted



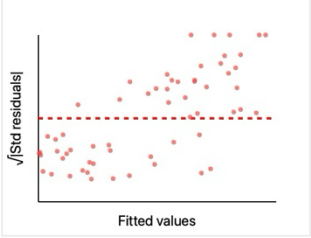
Supuesto violado

2. Normal Q-Q



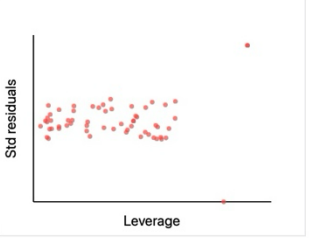
Supuesto violado

3. Scale-Location



Supuesto violado

4. Residuals vs Leverage



Supuesto violado

Test de Independencia (Durbin-Watson)

```
# Cargar librería
library(lmtest)

# Test de Durbin-Watson
dwtest(model)
```

Salida esperada:

```
Durbin-Watson test
data: model
DW = 1.9228, p-value = 0.2191
alternative hypothesis: true autocorrelation is greater than 0
```

Interpretación:

- **H_0**: No existe correlación entre residuos
- **p = 0.219 > 0.05 → No se rechaza H_0** ✓
- **Conclusión**: Residuos independientes

○ DETECCIÓN DE MULTICOLINEALIDAD

Matriz de Correlaciones:

La matriz de correlaciones entre las variables explicativas cuantitativas del modelo nos da información sobre la colinealidad.

```
# Cargar librería
library(corrplot)

# Gráfico de correlaciones
corrplot(cor(ds[, c(2:6,8:9)]), win.asp = 0.5)
```

Factor de Inflación de Varianza (VIF):

Los VIF nos informan sobre la correlación entre las variables explicativas del modelo.

```
# Cargar librería
library(car)

# Calcular VIF
vif(model)
```

Salida esperada:

```
      CompPrice       Income  Advertising        Price    ShelveLoc          Age
       1.534883     1.015448     1.012935     1.534425     1.015139     1.016830
```

Interpretación VIF:

- **VIF < 5** → Sin multicolinealidad ✓
- **VIF > 5** → Multicolinealidad problemática ✗
- **VIF > 10** → Multicolinealidad severa ✗

En nuestro caso: Todos los valores < 5 → **Sin problemas de multicolinealidad** ✓

○ IDENTIFICACIÓN DE OBSERVACIONES INFLUYENTES

Distancia de Cook:

```
# Gráfico de distancia de Cook
plot(model, 4)
```

Criterio de Cook:

- **Umbral**: 4/(n-p-1) = 4/(400-6-1) = 0.0102
- **Observaciones influyentes**: 26, 31, 294 (superan el umbral)

Modelo sin observaciones influyentes:

```
# Eliminar observaciones influyentes
model2 <- lm(Sales ~ CompPrice + Income + Advertising +
             Price + ShelveLoc + Age,
             data = ds[-c(31, 26, 294), ])

summary(model2)
```

Mejora del modelo:

- **Error estándar**: 1.154 → **1.064** ✓
- **R^2**: 0.8249 → **0.852** ✓
- **R^2 ajustado**: 0.8217 → **0.8494** ✓

○ PREDICCIÓN

Predicción Puntual

Crear nuevos datos:

```
# Nuevas observaciones para predecir
new <- data.frame(CompPrice = 120,
                  Income = 70,
                  Advertising = 8,
                  Price = 100,
                  ShelveLoc = "Medium",
                  Age = 30)
```

Realizar predicción:

```
# Predicción con intervalo de confianza
predict(model, new, type = "response",
        se.fit = TRUE, interval = "confidence")$fit
```

Salida esperada:

```
      fit      lwr      upr
1 9.526063 9.277606 9.774520
```

Interpretación:

- **Predicción**: 9.53 miles de unidades vendidas
- **Intervalo 95%**: [9.28, 9.77]
- **Confianza**: Rango probable del valor real

Fórmula Manual

```
# Usando los coeficientes estimados
5.273 + 0.087*120 + 0.016*70 + 0.129*8 +
(-0.089)*100 + 1.938*1 + (-0.043)*30
```

○ CAPACIDAD PREDICTIVA

Validación Cruzada

División de datos:

```
# Establecer semilla para reproducibilidad
set.seed(1234)

# Dividir datos: 75% entrenamiento, 25% prueba
training.samples <- sample(nrow(ds), nrow(ds)*0.75, replace = FALSE)
train.data <- ds[training.samples, ]
test.data <- ds[-training.samples, ]
```

Entrenar modelo:

```
# Modelo en datos de entrenamiento
model.cv <- lm(Sales ~ CompPrice + Income + Advertising +
               Price + ShelveLoc + Age, data = train.data)

summary(model.cv)
```

Evaluar en datos de prueba:

```
# Hacer predicciones en datos de prueba
predictions <- model.cv %>% predict(test.data)

# Calcular métricas de evaluación
r_squared <- 1 - (sum((test.data$Sales - predictions)^2) /
                  sum((test.data$Sales - mean(test.data$Sales))^2))

rmse <- sqrt(mean((test.data$Sales - predictions)^2))

mae <- mean(abs(test.data$Sales - predictions))

# Mostrar resultados
data.frame(r_squared, rmse, mae)
```

Salida esperada:

```
  r_squared     rmse       mae
1 0.8638733 1.056703 0.8048313
```

Interpretación de Métricas

R^2 de prueba = 0.864

- El modelo explica el **86.4%** de variabilidad en datos nuevos

RMSE = 1.057

- Error promedio de predicción en unidades originales

MAE = 0.805

- Error absoluto promedio

Tasa de Error

```
# Calcular tasa de error
sqrt(mean((test.data$Sales - predictions)^2)) / mean(test.data$Sales)
```

Salida esperada:

```
[1] 0.1364774
```

Resultado: 13.65%

Conclusión: Modelo con **buena capacidad predictiva** (error < 15%)