Estadística Multivariada

Haydeé Peruyero

Contents

Chapter 1

Estadística Multivariada

1.1 Temario

1. Regresión múltiple

1.1 Mínimos cuadrados.

1.2 Medidas de bondad de ajuste.

1.3 Determinación del número de variables predictorias.

2. Análisis de componentes principales

2.1 Descripción de la metodología.

2.2 Técnicas de extracción de componentes principales.

2.3 Determinación del número de componentes principales.

3. Análisis factorial

3.1 Descripción de la metodología del análisis factorial.

3.2 Descripción del modelo básico.

3.3 Método de cálculo.

3.4 Comparación con la técnica del análisis de componentes principales.

3.5 Usos de software (R, Minitab, SciPy, entre otros).

4. Análisis de conglomerados

4.1Descripción de la metodología de análisis de conglomerados.

4.2 Técnicas de jerarquización y de particionamiento.

4.3 Implementación computacional.

4.4 Usos de los dendogramas.

4.5 Usos de software (R, Minitab, SciPy, entre otros).

5. Análisis discriminante

5.1 Descripción de la metodología del análisis discriminante.

5.2 Discriminación entre dos grupos.

5.3 Contribución por variable.

5.4 Discriminación logística.

5.5 Discriminación múltiple.

5.6 Usos de software (R, Minitab, SciPy, entre otros).

A1. R

A2. Git + Github

A3. Gráficas Multivariadas

A4. Escalas de Medición

A5. Valores Faltantes

1.2 Evaluación

- Examenes 50%
- Tareas 25%
- Proyecto 20%
- DataCamp 5%

1.3 Proyecto final

- Buscar una base de datos “real”
- Aplicar 3 métodos de estadística multivariada
- Entregar documento con:
  - Descripción de los datos
  - Planteamiento del problema
  - Métodos usados

  - Interpretación de resultados
  - Código usado

- Repositorio con código reproducible
- Exposición de resultados

1.4 Referencias

[1]

1.5 Material interesante

- Bookdown.
- Software Carpentry.
- Git
- Why Git
- R Markdown Cookbook
- STHDA
- YaRrr! The Pirate's Guide to R
- Learn ggplot2 Using Shiny App
- Ggplot2: Elegant Graphics for Data Analysis
  - Versión online
- Use R! Colección Springer
- Lattice: Multivariate Data Visualization with R
- R Graphics cookbook
- Cuenta pro de Github

1.6 DataCamp

Figure 1.1: DataCamp

Chapter 2

Regresión múltiple

2.1 ¿Por qué estadística multivariada?

El proceso de modelado consiste en construir expresiones matemáticas que permitan representar el comportamiento de una variable que queremos estudiar. Cuando contamos con varias variables, suele interesarnos analizar cómo unas influyen sobre otras, determinando si existe una relación, su intensidad y su forma. En muchos casos, estas relaciones pueden ser complejas y difíciles de describir directamente; por ello, se busca aproximarlas mediante funciones matemáticas sencillas como polinomios, que conserven los elementos esenciales para explicar el fenómeno de interés.

Cuando estudiamos fenómenos deterministas, es común vincular una variable dependiente con una o más variables independientes. Por ejemplo, en la ecuación de la velocidad ($v = d/t$), la distancia depende de la velocidad y del tiempo. En la práctica, cuando realizamos distintos experimentos, las fórmulas deterministas podrían no capturar por completo el comportamiento observado. Esto puede deberse a factores no controlados, a la presencia de variabilidad natural o a efectos aleatorios. Por esta razón, además de la parte determinista del modelo, se incorpora un término que represente la discrepancia aleatoria entre lo que se predice y lo que efectivamente se observa. De forma general, esta idea se resume como:

$$Observacin = Modelo \ + \ Error$$

Cuando se supone que la relación entre las variables puede representarse mediante una ecuación lineal, hablamos de *análisis de regresión lineal*. Si intervienen únicamente dos variables, una dependiente y y independiente x, se trata de **regresión lineal simple**. En cambio, cuando la variable de interés y depende

de dos o más variables independientes $x_1, x_2, ...$ hablamos de **regresión lineal múltiple**.

Supongamos que queremos predecir el rendimiento académico de un estudiante, ¿solo necesitamos las horas que estudia?

En este caso se tiene que el puntaje o rendimiento lo podemos representar con y y las horas de estudio con x. Entonces esta propuesta de modelo, la podríamos representar como:

$$y = \beta_0 + \beta_1 x$$

Donde β_0 es la ordenada al origen y β_1 la pendiente. Esta recta podría no ajustarse al modelo por diferentes razones, entonces lo que se hace es considerar un error aleatorio ϵ. El modelo que ya considera este error se representa como:

$$y = \beta_0 + \beta_1 x + \epsilon.$$

A este modelo se le conoce como modelo de **regresión lineal simple** y a β_0, β_1 se les conoce como **coeficientes de regresión**.

En problemas reales, casi nunca una sola variable explica el fenómeno. Las decisiones y predicciones mejoran cuando integramos múltiples fuentes de información.

Ejemplos: - Salud: riesgo de una enfermedad según edad, IMC, actividad física, dieta y antecedentes. - Ingeniería: vida útil de una pieza según temperatura, vibración, material y carga. - Biología: crecimiento de una planta por agua, luz, fertilizante, temperatura.

Ejemplo: Si queremos predecir el rendimiento académico de un estudiante, ¿solo necesitamos las horas que estudia? ¿qué otras variables podrían influir en el puntaje de un examen?

Rendimiento escolar

```
set.seed(123)
n <- 10
data_intro <- tibble(
  estudiante = paste0("E", 1:n),
  horas_estudio = c(2,3,4,5,1,3,2,4,5,6),
  horas_sueno  = c(7,8,6,7,5,8,7,6,9,7),
  asistencia   = c(0.9,0.95,0.8,0.85,0.7,0.9,0.8,0.9,1,0.95),
  puntaje      = c(65,70,68,80,60,75,65,78,88,85)
)
data_intro
```

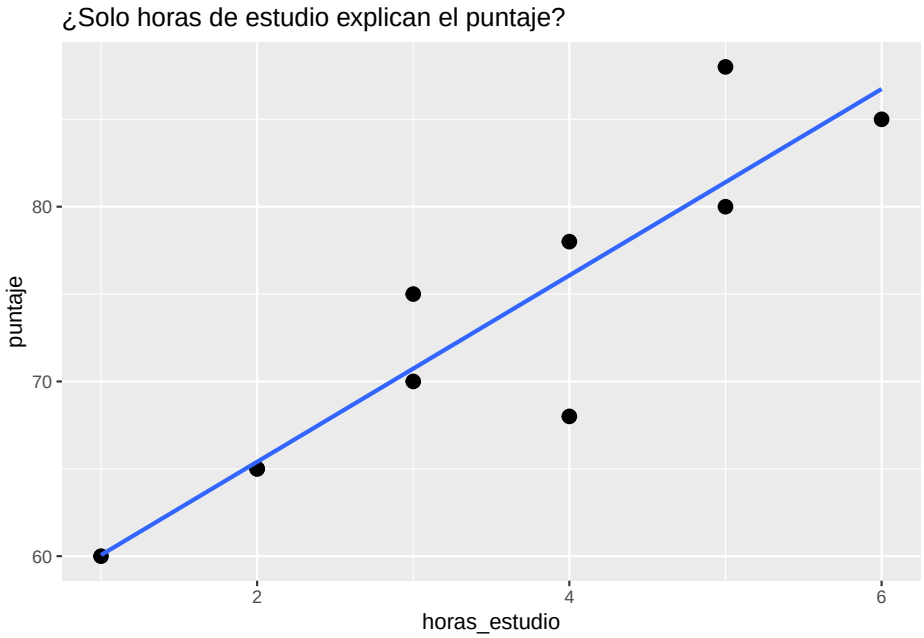
```
## # A tibble: 10 x 5
##    estudiante horas_estudio horas_sueno asistencia puntaje
##    <chr>              <dbl>       <dbl>      <dbl>   <dbl>
##  1 E1                     2           7       0.9       65
##  2 E2                     3           8       0.95      70
##  3 E3                     4           6       0.8       68
##  4 E4                     5           7       0.85      80
##  5 E5                     1           5       0.7       60
##  6 E6                     3           8       0.9       75
##  7 E7                     2           7       0.8       65
##  8 E8                     4           6       0.9       78
##  9 E9                     5           9       1         88
## 10 E10                    6           7       0.95      85
```

¿Qué pasa si solo graficamos horas de estudio vs puntaje?

Plot hotas de estudio vs puntaje sugerida

```
library(ggplot2)
ggplot(data_intro, aes(horas_estudio, puntaje)) +
  geom_point(size=3) +
  geom_smooth(method="lm", se=FALSE) +
  labs(title="¿Solo horas de estudio explican el puntaje?")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



¿Se ajusta un modelo lineal? ¿Porqué?

2.1.1 ¿Qué es "multivariado" y por qué lo necesitamos?

Idea central: cuando **varias** x influyen sobre y, estudiar cada x por separado puede engañarnos. El análisis multivariado permite:

- **Aislar efectos**: estimar el efecto de x_1 *manteniendo constantes* $x_2, x_3, ...$
- **Mejorar predicción**: reducir error al añadir información relevante.
- **Controlar confusores**: variables que cambian la relación aparente entre y y x.

Ejemplo: Si ajustamos ahora un modelo con varias variables, ¿vamos a observar un cambio? ¿se ajustará mejor?

Código (modelos + comparaciones)

```
# Modelo simple
m1 <- lm(puntaje ~ horas_estudio, data = data_intro)

# Modelo múltiple
m2 <- lm(puntaje ~ horas_estudio + horas_sueno + asistencia, data = data_intro)

# Medidas clave
R2_m1  <- glance(m1)$r.squared
R2_m2  <- glance(m2)$r.squared

print(paste("El R2 del modelo simple:", R2_m1))
```

```
## [1] "El R2 del modelo simple: 0.824317362184441"
```

```
print(paste("El R2 del modelo multiple:", R2_m2))
```

```
## [1] "El R2 del modelo multiple: 0.895428180549875"
```

```
#R2adj_m1 <- glance(m1)$adj.r.squared
#R2adj _m2 <- glance(m2)$adj.r.squared
```

- ¿Aumentó R^2 al incluir más variables? ¿Por qué tiende a subir?
- ¿Qué cambia en la interpretación de horas_estudio al controlar por horas_sueno y asistencia?
- ¿Puede un predictor ser importante en bivariado y no en multivariado (o viceversa)?

2.2 Regresión múltiple

2.2.1 Modelo y estimación

Los modelos en regresión lineal múltiple están dados por la siguiente forma, donde y depende de p variables predictoras:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_p x_{ip} + \epsilon_i.$$

Se suele asumir que los errores ϵ_i son i.i.d. con distribución normal de media 0 y varianza σ^2 desconocida. Los coeficientes β_i son constantes desconocidas y son los parámetros del modelo. Cada β_j representa el cambio esperado en la respuesta y por el cambio unitario en x_i cuando todas las demás variables independientes $x_i (i \neq j)$ se mantienen constantes.

```
# Forma general
ajuste <- lm(y ~ x1 + x2 + ... + xp, data = datos)
# summary(ajuste)
```

Los coeficientes los podemos interpretar como sigue:

- **Intercepto** (β_0): valor esperado de y cuando todas las x=0.
- **Pendiente** β_j: efecto **parcial** de x_j sobre y manteniendo las demás constantes.

En los modelos de regreción lineal, solemos usar las siguientes medidas de bondad de ajuste:

- R^2: proporción de varianza de y explicada.
- R^2 **ajustado**: penaliza por número de predictores (mejor para comparar modelos con distinto número de x).
- **RMSE** (σ): error típico de predicción en unidades de y.

```
comp <- dplyr::bind_rows(
  glance(m1) %>% mutate(modelo="simple"),
  glance(m2) %>% mutate(modelo="multiple")
) %>% select(modelo, r.squared, adj.r.squared)
comp
```

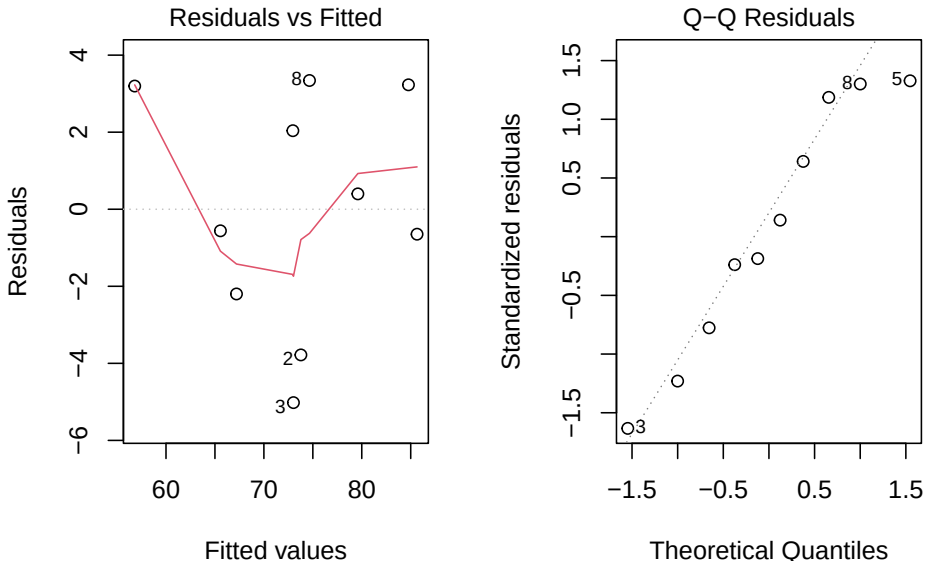
```
## # A tibble: 2 x 3
##   modelo   r.squared adj.r.squared
##   <chr>        <dbl>         <dbl>
## 1 simple       0.824         0.802
## 2 multiple     0.895         0.843
```

Para este modelo algunos de los supuestos se siguen del modelo de regresión lineal simple y se agregan algunos que tienen que ver con la relación que pudiera existir entre las variables regresoras.

- El modelo es lineal en los parámetros.
  Chequeo: residuales vs ajustados sin patrón claro.
- El modelo está especificado correctamente.
- Covarianza cero entre variables regresoras y el error.
- Esperanza del error igual a cero.
- Homocedasticidad.
- No autocorrelación entre los errores.
- Los errores siguen una distribución normal.
- Mas observaciones que parámetros a estimar.
- Variación entre los valores de las variables regresoras.
- No colinealidad (multicolinealidad) entre las variables regresoras, es decir, no existe una relación lineal entre x_i y x_j (es decir, las variables son linealmente independientes).

Supuestos

```
# Modelo m2
par(mfrow=c(1,2))
plot(m2, which=1)  # Residuales vs ajustados
plot(m2, which=2)  # QQ-plot
```



Ejercicio: Supongamos que tenemos los siguientes datos: precio de vivienda según metros, habitaciones y distancia al centro.

Dataset

```
set.seed(42)
n <- 14
casas <- tibble::tibble(
  precio = c(200,220,250,275,300,180,210,260,280,320,190,240,230,305),
  metros = c(80,90,100,110,120,70,85,105,115,130,75,95,92,125),
  habitaciones = c(2,3,3,4,4,2,3,3,4,5,2,3,3,4),
  distancia_centro = c(5,4,6,3,2,8,6,3,2,1,7,5,4,2)
)
casas
```

```
## # A tibble: 14 x 4
##    precio metros habitaciones distancia_centro
##     <dbl>  <dbl>        <dbl>            <dbl>
##  1    200     80            2                5
##  2    220     90            3                4
##  3    250    100            3                6
##  4    275    110            4                3
##  5    300    120            4                2
##  6    180     70            2                8
##  7    210     85            3                6
##  8    260    105            3                3
##  9    280    115            4                2
## 10    320    130            5                1
## 11    190     75            2                7
## 12    240     95            3                5
## 13    230     92            3                4
## 14    305    125            4                2
```

1) Ajusta `precio ~ metros` (simple) y `precio ~ metros + habitaciones + distancia_centro` (múltiple).

2) Compara R^2, R^2 **ajustado** y **(RMSE)**.

3) Interpreta el coeficiente de `distancia_centro`.

4) Revisa QQ-plot y residuales vs ajustados. ¿Algún patrón?

Solución

```
m_s <- lm(precio ~ metros, data=casas)
m_m <- lm(precio ~ metros + habitaciones + distancia_centro, data=casas)

broom::glance(m_s)[,c("r.squared","adj.r.squared")]
```

```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1     0.996         0.996
```

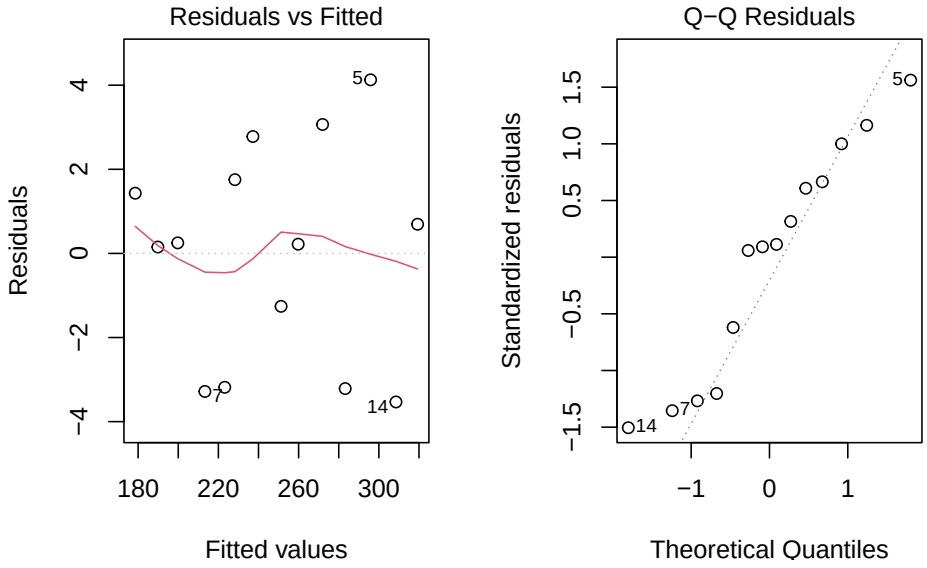
```
broom::glance(m_m)[,c("r.squared","adj.r.squared")]
```

```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1     0.997         0.996
```

```
broom::tidy(m_m)
```

```
## # A tibble: 4 x 5
##   term             estimate std.error statistic      p.value
##   <chr>               <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)        -8.67     14.9      -0.583 0.573
## 2 metros              2.53      0.162    15.6   0.0000000236
## 3 habitaciones       -0.505     2.80     -0.180 0.861
## 4 distancia_centro    1.38      0.974     1.42  0.187
```

```
par(mfrow=c(1,2))
plot(m_m, which=1)
plot(m_m, which=2)
```



2.3 Estimación de parámetros

Ejemplo (Montgomery, 2002): : Un embotellador de bebidas gaseosas analiza las rutas de servicio de las máquinas expendedoras en su sistema de distribución. Le interesa predecir el tiempo necesario para que el representante de ruta atienda las máquinas expendedoras en una tienda.

Esta actividad de servicio consiste en abastecer la máquina con productos embotellados, y algo de mantenimiento o limpieza. El ingeniero industrial responsable del estudio ha sugerido que las dos variables más importantes que afectan el tiempo de entrega y son la cantidad de cajas de producto abastecido, x_1, y la distancia caminada por el representante, x_2.

El ingeniero ha reunido 25 observaciones de tiempo de entrega que se ven en la tabla siguiente. Se ajustará el modelo de regresión lineal multiple siguiente:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Archivo: `refrescos.csv`.

Base de datos

```
#
datos <- data.frame(
  Observacion = 1:25,
  y = c(16.68, 11.50, 12.03, 14.88, 13.75,
        18.11, 8.00, 17.83, 79.24, 21.50,
```

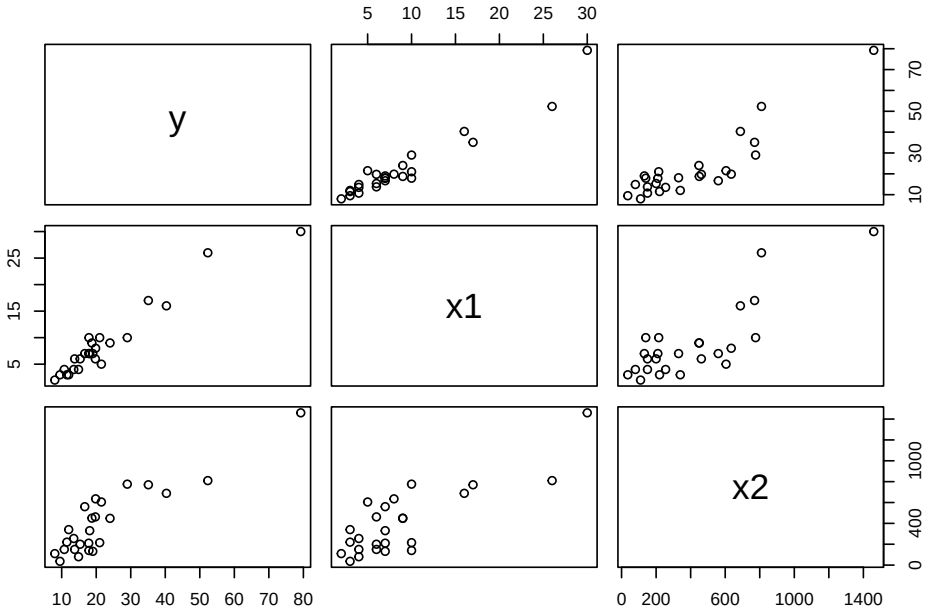
```
        40.33, 21.00, 13.50, 19.75, 24.00,
        29.00, 15.35, 19.00, 9.50, 35.10,
        17.90, 52.32, 18.75, 19.83, 10.75),
  x1 = c(7, 3, 3, 4, 6,
         7, 2, 7, 30, 5,
         16, 10, 4, 6, 9,
         10, 6, 7, 3, 17,
         10, 26, 9, 8, 4),
  x2 = c(560, 220, 340, 80, 150,
         330, 110, 210, 1460, 605,
         688, 215, 255, 462, 448,
         776, 200, 132, 36, 770,
         140, 810, 450, 635, 150)
)

datos
```

```
##    Observacion     y x1   x2
## 1            1 16.68  7  560
## 2            2 11.50  3  220
## 3            3 12.03  3  340
## 4            4 14.88  4   80
## 5            5 13.75  6  150
## 6            6 18.11  7  330
## 7            7  8.00  2  110
## 8            8 17.83  7  210
## 9            9 79.24 30 1460
## 10          10 21.50  5  605
## 11          11 40.33 16  688
## 12          12 21.00 10  215
## 13          13 13.50  4  255
## 14          14 19.75  6  462
## 15          15 24.00  9  448
## 16          16 29.00 10  776
## 17          17 15.35  6  200
## 18          18 19.00  7  132
## 19          19  9.50  3   36
## 20          20 35.10 17  770
## 21          21 17.90 10  140
## 22          22 52.32 26  810
## 23          23 18.75  9  450
## 24          24 19.83  8  635
## 25          25 10.75  4  150
```

Veamos un gráfico de dispersión de los datos. ¿Qué observamos?

```r
pairs(datos[-1])
```



1) Estimar β

Primero, vamos a crear la matriz X y el vector y.

Matrices

```r
# Columna de 1 para el intercepto
idv <- rep(1, nrow(datos))
# Creamos matriz X
X <- matrix(c(idv,datos$x1,datos$x2),nrow=25,ncol=3)
# Creamos el vector y
y <- matrix(datos$y, nrow = 25, ncol = 1)
```

Ya sabemos que nuestro estimador está dado por

$$\hat{\beta} = (X'X)^{-1}X'y$$

Entonces podemos encontrar el estimador.

Estimador beta

```
beta <- solve(t(X) %*% X) %*% t(X) %*% y
beta
```

```
##             [,1]
## [1,] 2.34123115
## [2,] 1.61590721
## [3,] 0.01438483
```

Entonces el ajuste por el método de mínimos cuadrados, con los coeficientes de regresión que encontramos está dado por:

$\hat{y} = 2.3412311 + 1.6159072\ x_1 + 1.6159072\ x_2$

Esto lo podemos hacer más rápido usando la función de `lm`. Construimos el modelo.

Modelo en R

```
M1 <- lm(y ~ x1 + x2, datos)
M1
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = datos)
##
## Coefficients:
## (Intercept)           x1           x2
##     2.34123      1.61591      0.01438
```

¿Cómo accedemos a los valores del modelo?

Coeficientes

```
beta_0 <- M1$coefficients[1]
beta_1 <- M1$coefficients[2]
beta_2 <- M1$coefficients[3]
```

Los valores son $\beta_0 = 2.3412311$, $\beta_1 = 1.6159072$ y $\beta_2 = 0.0143848$.

2) Estimación de la varianza del error σ^2

Ya tenemos que la suma de los cuadrados de los errores está dada por

$$SSE = y'y - \hat{\beta}X'y$$

Sustituimos los valores que tenemos y obtemos el SSE.

SSE

```
SSE <- t(y)%*% y - t(beta) %*% t(X) %*% y
SSE
```

```
##          [,1]
## [1,] 233.7317
```

Y de está forma, podemos encontrar el estimador de σ^2.

Estimador

```
varest <- SSE / (nrow(y) - nrow(beta))
varest
```

```
##          [,1]
## [1,] 10.62417
```

Directo con las funciones de R, podemos acceder a los parámetros que se guardaron en el modelo que ya calculamos.

Resumen del modelo

```
summary(M1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7880 -0.6629  0.4364  1.1566  7.4197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.341231   1.096730   2.135 0.044170 *
## x1          1.615907   0.170735   9.464 3.25e-09 ***
## x2          0.014385   0.003613   3.981 0.000631 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 22 degrees of freedom
## Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
## F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

Algunos de los parámetros almacenados en el modelo nos permiten obtener también el resultado previo.

Estimador

```
sum(residuals(M1)^2) / df.residual(M1)
```

```
## [1] 10.62417
```

2.3.1 Ejercicios

Ejercicio 1: Un analista hace un estudio químico y espera que el rendimiento de cierta sustancia se vea afectado por dos factores. Se realizan 17 experimentos cuyos datos se registran en el cuadro siguiente. Por experimentos similares, se sabe que los factores x_1 y x_2 no están relacionados; por ello, el analista decide utilizar un modelo de regresión lineal múltiple. Calcule el modelo de regresión y grafíquelo sobre las observaciones.

Archivo: est_quimico.csv

Datos Ejercicio 1

```
datos2 <- data.frame(
  Experimento = 1:17,
  x1 = c(41.9, 43.4, 43.9, 44.5, 47.3, 47.5, 47.9, 50.2, 52.8, 53.2, 56.7, 57.0, 63.5,
  x2 = c(29.1, 29.3, 29.5, 29.7, 29.9, 30.3, 30.5, 30.7, 30.8, 30.9, 31.5, 31.7, 31.9,
  y  = c(251.3, 251.3, 248.3, 267.5, 273.0, 276.5, 270.3, 274.9, 285.0, 290.0, 297.0, 3
)

datos2
```

```
##    Experimento   x1   x2     y
## 1            1 41.9 29.1 251.3
## 2            2 43.4 29.3 251.3
## 3            3 43.9 29.5 248.3
## 4            4 44.5 29.7 267.5
## 5            5 47.3 29.9 273.0
## 6            6 47.5 30.3 276.5
## 7            7 47.9 30.5 270.3
## 8            8 50.2 30.7 274.9
## 9            9 52.8 30.8 285.0
## 10          10 53.2 30.9 290.0
## 11          11 56.7 31.5 297.0
## 12          12 57.0 31.7 302.5
## 13          13 63.5 31.9 304.5
```

```
## 14          14 64.3 32.0 309.3
## 15          15 71.1 32.1 321.7
## 16          16 77.0 32.5 330.7
## 17          17 77.8 32.9 349.0
```

Ejercicio 2: Repetir el ejemplo con los datos `datasets::trees` de R que proporciona mediciones del diámetro, altura y volumen de madera en 31 cerezos negros talados.

Ejercicio 3: Subir a Github los dos ejercicios previos tanto con solución en R como en Python. Comparar las funciones. Ventajas y desventajas de ambas.

2.4 Pruebas de Hipótesis

Cuando revisamos el `summary` del modelo, nos arroja si son significativas o no y a que nivel de significancia las variables que estamos considerando. Veamos el siguiente ejemplo.

2.4.1 Prueba de la significancia de la regresión

Ejemplo: Con los datos del embotellador de bebidas gaseosas, se probará la significancia de la regresión.

Sumas de Cuadrados

```
SCT <- t(y) %*% y -  sum(y)**2 / nrow(datos)
SCT
```

```
##          [,1]
## [1,] 5784.543
```

```
SCE <- t(beta) %*% t(X) %*% y - sum(y)**2 / nrow(datos)
SCE
```

```
##          [,1]
## [1,] 5550.811
```

```
SSE <- SCT - SCE
SSE
```

```
##          [,1]
## [1,] 233.7317
```

Para probar

$$H_0 : \beta_1 = \beta_2 = 0$$

se calcula el estadístico:

Estadístico F

```
F0 <- (SCE / (ncol(X) - 1)) / (SSE / (nrow(X) - (ncol(X) -1) - 1))
F0
```

```
##          [,1]
## [1,] 261.2351
```

Como el valor de F_0 es mayor que el valor tabulado de $F_{\alpha;p,n-p-1} = F_{0.05;2;22} = 3.44$, se rechaza H_0. Lo cual implica qye el tiempo de entrega depende del volumen de entrega y/o de la distancia.

Ahora, usando los modelos que ya calculamos.

Sumas de cuadrados

```
SCT.m<-sum((datos$y-mean(datos$y))^2)
SCT.m
```

```
## [1] 5784.543
```

```
SCE.m <-sum((M1$fitted-mean(datos$y))^2)
SCE.m
```

```
## [1] 5550.811
```

```
SSE.m <-sum(M1$residuals^2)
SSE.m
```

```
## [1] 233.7317
```

Grados de libertad

```
n<-nrow(y)
n
```

```
## [1] 25
```

```
GLT<- n-1
GLT
```

```
## [1] 24
```

```
GLRes<- df.residual(M1)
GLRes
```

```
## [1] 22
```

```
GLR<- GLT-GLRes
GLR
```

```
## [1] 2
```

Cuadrados medios

```
CMR <- SCE /GLR
CMR
```

```
##           [,1]
## [1,] 2775.405
```

```
CMRes <- SSE / GLRes
CMRes
```

```
##           [,1]
## [1,] 10.62417
```

Estadístico F_0

```
F0 <- CMR/CMRes
F0
```

```
##           [,1]
## [1,] 261.2351
```

p-valor

```
pv <- 1 - pf(F0, GLR,GLRes)
pv
```

```
##              [,1]
## [1,] 4.440892e-16
```

Valor tabulado de F

```
alpha <- 0.05; df1 <- 2; df2 <- 22
F_crit <- qf(1 - alpha, df1, df2)
F_crit
```

```
## [1] 3.443357
```

2.4.2 Pruebas sobre coeficientes individuales de regresión

Ejemplo: Usando los datos del embotellador de bebidas gaseosas, se desea evaluar la importancia de la variable regresora *distancia* (x_2) dado que el regresor *cajas* (x_1) está en el modelo.

Estadístico t_0

```
C22 <- solve(t(X) %*% X)[3,3]
C22
```

```
## [1] 1.228745e-06
```

```
t0 <- beta_2 / sqrt(varest * C22)
t0
```

```
##          [,1]
## [1,] 3.981313
```

```
## t tabulado con confianza 95% y 22 grados de libertad
tt <- qt(p = 0.95 + 0.05/2, df = 22, lower.tail = TRUE)
tt
```

```
## [1] 2.073873
```

Usando el modelo que ya tenemos calculado `M1` podemos obtener estos mismos resultados de la siguiente forma.

Prueba sobre coeficientes

```
summary(M1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7880 -0.6629  0.4364  1.1566  7.4197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.341231   1.096730   2.135 0.044170 *
## x1          1.615907   0.170735   9.464 3.25e-09 ***
## x2          0.014385   0.003613   3.981 0.000631 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 22 degrees of freedom
## Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
## F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

2.5 Intervalos de confianza

2.5.1 Intervalos de confianza en los coeficientes de regresión

Ejemplo: Usando los datos del embotellador de bebidas gaseosas, queremos calcular el intervalo de confianza del 95% para β_1. Recordemos que el estimador puntual de β_1 es 1.6159072.

Intervalo de confianza

```
C11 <- solve(t(X) %*% X)[2,2]

izq <- beta_1 - tt * sqrt(varest*C11)
izq
```

```
##          [,1]
## [1,] 1.261825
```

```
der <- beta_1 + tt * sqrt(varest*C11)
der
```

```
##          [,1]
## [1,] 1.96999
```

2.5.2 Intervalo de confianza de la respuesta media

Ejemplo: El embotellador de bebidas gaseosas quiere establecer un intervalo de confianza del 95% para el tiempo medio de entrega para una tienda donde se requieren $x_1 = 8$ cajas y la distancia es de $x_2 = 275$ pies.

Nuestro vector X_0 está dado por:

X0

```
X0 <- matrix(c(1, 8, 275), nrow = 3)
X0
```

```
##      [,1]
## [1,]    1
## [2,]    8
## [3,]  275
```

El valor ajustado en ese punto es:

Valor ajustado

```
y0 <- t(X0) %*% beta
y0
```

```
##           [,1]
## [1,] 19.22432
```

La varianza de $\hat{y_0}$

Varianza

```
var_y0 <- varest * t(X0) %*% solve(t(X) %*% X) %*% X0
var_y0
```

```
##            [,1]
## [1,] 0.5734134
```

Entonces el intervalo de confianza en este punto es:

Intervalo de confianza

```
l_izq <- y0 - tt * sqrt(var_y0)
l_izq
```

```
##          [,1]
## [1,] 17.6539
```

```
l_der <- y0 + tt * sqrt(var_y0)
l_der
```

```
##           [,1]
## [1,] 20.79474
```

Ejemplo: Usaremos el conjunto de datos `data("marketing")` que contiene 200 observaciones de un experimento publicitario que evalúa el impacto de tres medios de anuncio en las ventas. Para cada observación se registran los presupuestos de publicidad (en miles de dólares) y las ventas obtenidas. Variables:

- `youtube`: presupuesto invertido en anuncios de YouTube (miles de USD).
- `facebook`: presupuesto invertido en Facebook (miles de USD).
- `newspaper`: presupuesto invertido en prensa escrita (miles de USD).
- `sales`: ventas registradas (variable respuesta).

Cargamos los datos:

```
library(datarium)
data("marketing")
```

Exploramos rápidamente la base para ver qué variables contiene y la dimensión:

```
str(marketing)
```

```
## 'data.frame':    200 obs. of  4 variables:
##  $ youtube  : num  276.1 53.4 20.6 181.8 217 ...
##  $ facebook : num  45.4 47.2 55.1 49.6 13 ...
##  $ newspaper: num  83 54.1 83.2 70.2 70.1 ...
##  $ sales    : num  26.5 12.5 11.2 22.2 15.5 ...
```

```
#?marketing
```

Ajustamos un modelo lineal que incluya todas las variables, es decir,

$sales = \beta_0 + \beta_1 youtube + \beta_2 facebook + \beta_3 newspaper + \epsilon$

Modelo marketing

```
modelo1<-lm(sales~youtube+facebook+newspaper,data=marketing)
summary(modelo1)
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5932  -1.0690   0.2902   1.4272   3.3951
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.526667   0.374290   9.422   <2e-16 ***
## youtube      0.045765   0.001395  32.809   <2e-16 ***
## facebook     0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.023 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

¿Qué se puede decir sobre la significancia de la variable *newspaper*?

Veamos qué ocurre con el modelo al eliminar la variable *newspaper*

Modelo marketing 2

```
modelo2<-lm(sales~facebook+youtube,data=marketing)
summary(modelo2)
```

```
##
## Call:
## lm(formula = sales ~ facebook + youtube, data = marketing)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5572  -1.0502   0.2906   1.4049   3.3994
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.50532    0.35339   9.919   <2e-16 ***
## facebook     0.18799    0.00804  23.382   <2e-16 ***
## youtube      0.04575    0.00139  32.909   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.018 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Lo que sigue, es hacer pruebas de hipótesis tanto en las variables como en los coeficientes de regresión.

Ejercicio 1: Realizar las pruebas de hipótesis sobre la significancia de la regresión y sobre los coeficientes. Encontrar los intervalos de confianza respectivos del 95%. Para una tienda con presupuestos: $youtube = 150$, $facebook = 30$, $newspaper = 20$ (en miles de USD): (a) Calcula el intervalo de confianza del 95% para la media de ventas $\mathbb{E}(sales|X_0)$. (b) Calcula el intervalo de predicción del 95% para una nueva observación de ventas. (c) Comenta la diferencia entre ambos intervalos. Subir respuesta y explicación de sus resultados a github.

2.6 Ejercicios

Ejercicio 1: Para los datos de la Liga Nacional de Fútbol

a) Ajustar un modelo de regresión lineal múltiple que relacione la cantidad de juegos ganados con las yardas por aire del equipo (x_2), el porcentaje de jugadas por tierra (x_7) y las yardas por tierra del contrario (x_8).

b) Formar la tabla de análisis de varianza y probar la significancia de la regresión.

c) Calcular el estadístico t para probar las hipótesis $H_0 : \beta_2 = 0$, $H_0 : \beta_7 = 0$ y $H_0 : \beta_8 = 0$. ¿Qué conclusiones se pueden sacar acerca del papel de las variables x_2, x_7 y x_8 en el modelo?

d) Calcular R^2 y R^2_{adj} para este modelo.

e) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

f) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

g) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?

h) Calcular un intervalo de confianza de 95% para β_7 y un intervalo de confianza de 95% para la cantidad media de juegos ganados por un equipo cuando $x_2 = 2300$, $x_7 = 56$ y $x_8 = 2100$.

i) Ajustar un modelo a esos datos, usando solo x_7 y x_8 como regresores y probar la significancia de la regresión.

j) Calcular R^2 y R^2_{adj}. Compararlos con los resultados del modelo anterior.

k) Calcular un intervalo de confianza de 95% para β_7. También, un intervalo de confianza de 95% para la cantidad media de juegos ganados por un equipo cuando $x_7 = 56$ y $x_8 = 2100$. Comparar las longitudes de esos intervalos de confianza con las longitudes de los correspondientes al modelo anterior.

l) ¿Qué conclusiones se pueden sacar de este problema, acerca de las consecuencias de omitir un regresor importante de un modelo?

Ejericio 2: Véase los datos de rendimiento de gasolina.

a) Ajustar un modelo de regresión lineal múltiple que relacione el rendimiento de la gasolina y, en millas por galón, la cilindrada del motor (x_1) y la cantidad de gargantas del carburador (x_6).

b) Formar la tabla de análisis de varianza y probar la significancia de la regresión.

c) Calcular R^2 y R^2_{adj} para este modelo. Compararlas con las R^2 y R^2_{adj} Ajustado para el modelo de regresión lineal simple, que relaciona las millas con la cilindrada.

d) Determinar un intervalo de confianza para β_1.

e) Determinar un intervalo de confianza de 95% para el rendimiento promedio de la gasolina, cuando $x_1 = 225 pulg^3$ y $x_6 = 2$ gargantas.

f) Determinar un intervalo de predicción de 95% para una nueva observación de rendimiento de gasolina, cuando $x_1 = 225 pulg^3$ y $x_6 = 2$ gargantas.

g) Considerar el modelo de regresión lineal simple, que relaciona las millas con la cilindrada. Construir un intervalo de confianza de 95% para el rendimiento promedio de la gasolina y un intervalo de predicción para

el rendimiento, cuando $x_1 = 225 pulg^3$. Comparar las longitudes de estos intervalos con los intervalos obtenidos en los dos incisos anteriores. ¿Tiene ventajas agregar x_6 al modelo?

h) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

i) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

j) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?

Chapter 3

Análisis de Componentes Principales

Chapter 4

Análisis Factorial

Chapter 5

Análisis de Conglomerados

Chapter 6

Análisis de Discriminante

Chapter 7

Apéndices

7.1 Introducción a R

- Tutorial de RMarkdown: Link
- Tutorial Manejo de Proyectos: Link

7.2 Git + Github

- Conectar R con Git y Github: Link

7.3 Gráficas Multivariadas

7.4 Escalas de Medición

7.5 Valores Faltantes