Preparación y Análisis de datos

Capítulo 2: Análisis inferencial de datos

Mg. J. Eduardo Gamboa U.

2025-03-27

Introducción

¿Qué es la inferencia estadística?

¿Por qué es importante?

Estimación

Estimación puntual

Sea X_1,...,X_n una muestra de tamaño n de una población con parámetro θ. Se denomina estimador puntual de θ a cualquier estadístico $\hat{\Theta} = h(X_1, ..., X_n)$ cuyo valor $\hat{\theta} = h(x_1, ..., x_n)$ dará una estimación puntual de θ. En este caso, Θ es una variable aleatoria y$\hat{\theta}$ es un número.
Estimador puntual de la media:

$$\hat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Estimador puntual de la variancia:

$$\hat{\sigma^2} = s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

Estimador puntual de la proporción:

$$\hat{\pi} = p = \frac{\text{Número de éxitos}}{n}$$

Ejemplo

En una población de madres adolescentes, la talla (en metros) sigue una distribución con media μ y variancia σ^2. Se extrae una muestra aleatoria 8 madres adolescentes, obteniendo las siguientes medidas de talla: 1.50, 1.60, 1.58, 1.45, 1.52, 1.68, 1.62, 1.55. Halle un estimador puntual para la media, la varianza y la desviación estándar poblacionales.

```
talla = c(1.60, 1.58, 1.45, 1.52, 1.68, 1.62, 1.55)
talla |> mean()
```

```
[1] 1.571429
```

```
talla |> var()
```

```
[1] 0.005480952
```

```
talla |> sd()
```

```
[1] 0.07403345
```

Estimación por intervalos de confianza

Sea X_1,...,X_n una muestra aleatoria de tamaño n de una población con parámetro θ, cuyos valores observados o datos respectivos son x_1,...,x_n.
Sea $a = h_1(x_1, ..., x_n)$ y $b = h_2(x_1, ..., x_n)$, valores numéricos calculados a partir de los datos de la muestra.
Entonces se dice que el intervalo $[a, b]$ tiene un nivel de confianza del $(1 - \alpha) \times 100\%$ de contener el parámetro θ, o que $\theta \in [a, b]$ con un nivel de confianza del $(1 - \alpha) \times 100\%$.
Interpretación: Con un nivel de confianza del $(1 - \alpha) \times 100\%$, se estima que el parámetro θ está contenido en el intervalo $[a, b]$.

Intervalo de confianza para la media

Si $X_1,\ldots,X_n$ es una muestra aleatoria de una población normal con media μ y σ^2 conocida, el intervalo con un nivel de confianza del $(1-\alpha)\times 100\%$ para la media μ se obtiene mediante:

$$\left(\bar{X} - Z_{(1-\alpha/2)}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{(1-\alpha/2)}\frac{s}{\sqrt{n}}\right)$$

Si σ^2 es desconocida, el intervalo con un nivel de confianza del $(1-\alpha)\times 100\%$ para la media μ se obtiene mediante:

$$\left(\bar{X} - t_{(1-\alpha/2;n-1)}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(1-\alpha/2;n-1)}\frac{s}{\sqrt{n}}\right)$$

Ejemplo

En un estudio sobre hábitos de hidratación, se registró el consumo diario de agua en litros de 16 personas seleccionadas aleatoriamente en una comunidad. La cantidad de agua ingerida es un factor clave para la salud y el bienestar, y este análisis busca conocer la variabilidad en los patrones de hidratación.

2.75	2.43	1.59	2.82
3.26	2.38	1.79	2.08
3.29	2.88	2.27	2.72
2.62	1.54	1.64	2.22

a. Suponiendo que σ es conocida, $\sigma = 0.5$

```
agua = c(2.75, 2.43, 1.59, 2.82, 3.26, 2.38, 1.79, 2.08, 3.29, 2.88,
         2.27, 2.72, 2.62, 1.54, 1.64, 2.22)

media = agua |> mean()
sigma = 0.5
n     = agua |> length()
conf  = 0.95
z     = qnorm((1+conf)/2)
LI    = media - z*sigma/sqrt(n)
LS    = media + z*sigma/sqrt(n)
c(LI,LS)
```

```
[1] 2.147505 2.637495
```

```r
library(BSDA)
zsum.test(mean.x = media, sigma.x = sigma, n.x = n, conf.level = conf)
```

```
    One-sample z-Test

data:  Summarized x
z = 19.14, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.147505 2.637495
sample estimates:
mean of x
   2.3925
```

Con un nivel de confianza del 95%, se estima que el consumo medio diario de agua está contenido en el intervalo $[2.15, 2.64]$ litros.

b. Suponiendo que σ es desconocida

```
agua = c(2.75, 2.43, 1.59, 2.82, 3.26, 2.38, 1.79, 2.08, 3.29, 2.88,
         2.27, 2.72, 2.62, 1.54, 1.64, 2.22)

media = agua |> mean()
s     = agua |> sd()
n     = agua |> length()
conf  = 0.95
vt    = qt((1+conf)/2, n-1)
LI    = media - vt*s/sqrt(n)
LS    = media + vt*s/sqrt(n)
c(LI,LS)
```

```
[1] 2.093929 2.691071
```

```r
agua |> t.test(conf.level = 0.95)
```

```
    One Sample t-test

data:  agua
t = 17.08, df = 15, p-value = 3.065e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.093929 2.691071
sample estimates:
mean of x
   2.3925
```

Con un nivel de confianza del 95%, se estima que el consumo medio diario de agua está contenido en el intervalo $[2.09, 2.69]$ litros.

Intervalo de confianza para la variancia

Si $X_1,\ldots,X_n$ es una muestra aleatoria de una población normal con y σ^2 desconocida, el intervalo con un nivel de confianza del $(1-\alpha)\times 100\%$ para la variancia σ^2 se obtiene mediante

$$\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2;n-1)}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(\alpha/2;n-1)}}$$

Interpretación: Con un nivel de confianza del $(1-\alpha)\times 100\%$, se estima que la variancia poblacional σ^2 esté contenida en el intervalo $[a,b]$.
Si se desea obtener los límites de confianza para la desviación estándar se obtiene la raíz cuadrada en la expresión anterior obteniéndose:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2;n-1)}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{(\alpha/2;n-1)}}}$$

Ejemplo (cont.)

```
agua = c(2.75, 2.43, 1.59, 2.82, 3.26, 2.38, 1.79, 2.08, 3.29, 2.88,
         2.27, 2.72, 2.62, 1.54, 1.64, 2.22)

varian = agua |> var()
n      = agua |> length()
conf   = 0.95
chi1   = qchisq((1-conf)/2, n-1)
chi2   = qchisq((1+conf)/2, n-1)
LI     = (n-1)*varian/chi2
LS     = (n-1)*varian/chi1
c(LI,LS)
```

```
[1] 0.1713196 0.7520275
```

```r
library(EnvStats)
varTest(agua, conf.level = 0.95)$conf.int |> as.numeric()
```

```
[1] 0.1713196 0.7520275
```

Con un nivel de confianza del 95%, se estima que la varianza del consumo diario de agua está contenida en el intervalo $[0.17, 0.75]$ litros^2.

```r
varTest(agua, conf.level = 0.95)$conf.int |> sqrt() |> as.numeric()
```

```
[1] 0.4139077 0.8671952
```

Con un nivel de confianza del 95%, se estima que la desviación estándar del consumo diario de agua está contenida en el intervalo $[0.41, 0.87]$ litros.

Intervalo de confianza para la proporción

Si X_1,...,X_n es una muestra aleatoria donde cada X_i indica la presencia (1) o ausencia (0) de una característica y $n > 30$, el intervalo con un nivel de confianza del $(1-\alpha) \times 100\%$ para la proporción π se obtiene mediante

$$p - Z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + Z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}}$$

Ejemplo

Una organización de salud realizó una encuesta en una comunidad para conocer la proporción de personas que consumen agua filtrada en sus hogares. Se entrevistó a 150 personas, registrando si consumen (1) o no consumen (0) agua filtrada. En total, 103 personas indicaron que sí consumen agua filtrada.

```
# Intervalo bajo la aproximación de Wald
# Para valores grandes de n o proporciones lejanas a 0 o 1
p    = 103/150
n    = 150
conf = 0.95
z    = qnorm((1 + conf)/2)
LI   = p - z*sqrt(p*(1-p)/n)
LS   = p + z*sqrt(p*(1-p)/n)
c(LI,LS)
```

```
[1] 0.6124368 0.7608965
```

```r
# Intervalo de Wilson sin corrección de Yates (n grande)
prop.test(x = 103, n = 150, conf.level = 0.95, correct = FALSE)$conf.int |>
  as.numeric()
```

```
[1] 0.6085602 0.7554509
```

```r
# Intervalo de Wilson con corrección de Yates (n pequeño)
# Más conservador que su versión sin corrección
prop.test(x = 103, n = 150, conf.level = 0.95, correct = TRUE)$conf.int |>
  as.numeric()
```

```
[1] 0.6051060 0.7584926
```

```r
# Intervalo exacto de Clopper-Pearson
# Para valores pequeños de n o proporciones cercanas a 0 o 1
# Suele ser el más amplio
binom.test(x = 103, n = 150, conf.level = 0.95)$conf.int |>
  as.numeric()
```

```
[1] 0.6059383 0.7598481
```

Pruebas de hipótesis

Hipótesis

Una hipótesis estadística es una afirmación sobre la distribución de probabilidad de una población o sobre el valor o valores de uno o más parámetros, como la media (μ), la variancia (σ^2) o la proporción (π).

Esta afirmación debe estar basada en la comprensión del fenómeno y sus variables. Una buena hipótesis permite hacer predicciones específicas y, si es rechazada, ayuda a revelar la complejidad del fenómeno.

Tipos de hipótesis estadísticas

Hipótesis nula (H_0 o H_p): Es la hipótesis que es aceptada provisionalmente como verdadera y cuya validez será sometida a verificación experimental. Los resultados experimentales nos permitirán seguir aceptándola como verdadera o si debemos rechazarla como tal.

Hipótesis alterna (H_1 o H_a): Es la hipótesis que se acepta en caso de que la hipótesis nula sea rechazada. La H_1 es la suposición contraria a H_0.

Prueba de hipótesis

Una prueba de hipótesis es un proceso estructurado para tomar decisiones basadas en datos. Se fundamenta en el método hipotético-deductivo, donde las hipótesis se contrastan con la evidencia en lugar de verificarse directamente.

Una prueba de hipótesis estadística es el proceso mediante el cual se toma la decisión de aceptar o rechazar la hipótesis nula.

El proceso de prueba de hipótesis determina si se rechaza o no H_0, pero **no se prueba su veracidad absoluta**.

Plantear hipótesis **antes del análisis de datos** (hipótesis a priori) mejora la solidez del estudio, enfocando la prueba en relaciones específicas y reduciendo sesgos.

Tipos de pruebas de hipótesis

En principio, se pueden formular hasta tres tipos de prueba, la cual dependerá de la forma de la hipótesis alterna que se planteé en el estudio:

Hipótesis unilateral con cola a la derecha	Hipótesis bilateral o de dos colas	Hipótesis unilateral con cola a la izquierda
$H_0 : \theta \leq \theta_0$ $H_1 : \theta > \theta_0$	$H_0 : \theta = \theta_0$ $H_1 : \theta \neq \theta_0$	$H_0 : \theta \geq \theta_0$ $H_1 : \theta < \theta_0$

donde θ es el parámetro de interés a probarse, pudiendo ser μ, σ^2, π (o algún otro que, por cuestiones de tiempo y/o complejidad no es abordado en el curso), y θ_0 es el valor o los valores supuestos que puede tomar el parámetro.

Aplicaciones

Una empresa de manufactura especializada en la producción de componentes electrónicos ha decidido evaluar el desempeño de dos equipos de trabajo en su planta principal. La gerencia está interesada en analizar la productividad y la calidad de los productos generados por cada equipo con el fin de mejorar la eficiencia operativa y reducir costos.

Los directivos han identificado diferencias percibidas en los tiempos de producción, costos y tasas de defectos entre los equipos. Sin embargo, antes de tomar decisiones estratégicas, desean realizar un análisis estadístico riguroso para determinar si estas diferencias son significativas o si podrían deberse al azar.

Se ha tomado una muestra aleatoria de los tiempos de producción (en minutos), costos de producción (en dólares) y tasas de defectos (1 = producto defectuoso, 0 = producto correcto) de dos equipos de trabajo en la planta.

Los datos están disponibles en el archivo **Produccion.csv**.

Aplicación 1

La gerencia de producción ha establecido que, para mantener la competitividad en el mercado, el tiempo promedio de producción por unidad debe ser de 45 minutos o menos. Sin embargo, han recibido reportes de que alguno de los equipos podría estar tardando más de lo esperado, lo que afectaría los tiempos de entrega y la satisfacción del cliente.

¿Los equipos realmente están cumpliendo con el estándar de 45 minutos, o en alguno (o ambos) los tiempos son mayores y se requiere una intervención?

$$H_0 : \mu \leq 45 \qquad H_1 : \mu > 45 \qquad \alpha = 0.05$$

```
datos = read.csv('Produccion.csv')
library(dplyr)
```

```r
datos |>
  filter(Grupo == "Equipo 1") |>
  pull(Tiempo) |>
  t.test(alternative = "greater", mu = 45)
```

```
    One Sample t-test

data:  pull(filter(datos, Grupo == "Equipo 1"), Tiempo)
t = -1.1487, df = 29, p-value = 0.87
alternative hypothesis: true mean is greater than 45
95 percent confidence interval:
 42.66133      Inf
sample estimates:
mean of x
 44.05667
```

No se rechaza H_0, ya que $pv > \alpha$

```r
datos |>
  filter(Grupo == "Equipo 2") |>
  pull(Tiempo) |>
  t.test(alternative = "greater", mu = 45)
```

```
    One Sample t-test

data:  pull(filter(datos, Grupo == "Equipo 2"), Tiempo)
t = 2.17, df = 34, p-value = 0.01855
alternative hypothesis: true mean is greater than 45
95 percent confidence interval:
 45.44277      Inf
sample estimates:
mean of x
 47.00571
```

Se rechaza H_0, ya que $pv < \alpha$ ¿Cuál es la respuesta a la problemática planteada?

Aplicación 2

La gerencia ha establecido que la desviación estándar debe ser menor a 3 dólares para ser considerada aceptable. Sin embargo, hay indicios de que el equipo 2 está experimentando una variabilidad en costos fuera de lo esperado, lo que puede afectar la planificación financiera y el control de presupuesto. Verificar esta afirmación.

$$H_0 : \sigma^2 \geq 9 \qquad H_1 : \sigma^2 < 9 \qquad \alpha = 0.05$$

```r
library(EnvStats)
datos |>
  filter(Grupo == "Equipo 2") |>
  pull(Costo) |>
  varTest(alternative = "less", sigma.squared = 9)
```

```
Results of Hypothesis Test
--------------------------

Null Hypothesis:                 variance = 9

Alternative Hypothesis:          True variance is less than 9

Test Name:                       Chi-Squared Test on Variance

Estimated Parameter(s):          variance = 9.097422

Data:                            pull(filter(datos, Grupo == "Equipo 2"), Costo)

Test Statistic:                  Chi-Squared = 34.36804

Test Statistic Parameter:        df = 34

P-value:                         0.549879

95% Confidence Interval:         LCL =  0.00000
```

Aplicación 3

La empresa de manufactura tiene un estándar de calidad que establece que el porcentaje de productos defectuosos debe ser menor al 3%. Sin embargo, hay sospechas de que se podría estar generando una tasa de defectos diferente a la esperada, lo que podría afectar la satisfacción del cliente y aumentar los costos de reproceso.

$$H_0 : \pi \geq 0.03 \qquad H_1 : \pi < 0.03 \qquad \alpha = 0.05$$

```r
datos |> nrow() -> n
datos |> filter(Defectuoso == 1) |> nrow() -> x
prop.test(x, n, p = 0.03, alternative = "less", correct = FALSE)
```

```
    1-sample proportions test without continuity correction

data:  x out of n, null probability 0.03
X-squared = 0.58287, df = 1, p-value = 0.7774
alternative hypothesis: true p is less than 0.03
95 percent confidence interval:
 0.0000000 0.1099856
sample estimates:
         p
0.04615385
```

Aplicación 4

En la empresa de manufactura, ambos equipos de producción que fabrican el mismo producto. La gerencia de calidad ha detectado que uno de los equipos podría estar generando más productos defectuosos que el otro, lo que podría afectar la rentabilidad y la satisfacción del cliente.
¿Existe una diferencia significativa en la tasa de defectos entre los dos equipos o las diferencias muestrales se deben al azar?

$$H_0 : \pi_1 - \pi_2 = 0 \qquad H_1 : \pi_1 - \pi_2 \neq 0 \qquad \alpha = 0.05$$

```
datos |> group_by(Grupo) |> count(Defectuoso) |>
  filter(Defectuoso == 1) |> pull(n) -> x

x
```

```
[1] 1 2
```

```
datos |> count(Grupo) |>
  pull(n) -> n

n
```

```
[1] 30 35
```

```r
prop.test(x, n, alternative = "two.sided")
```

```
    2-sample test for equality of proportions with continuity correction

data:  x out of n
X-squared = 8.8709e-31, df = 1, p-value = 1
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1478158  0.1001968
sample estimates:
    prop 1     prop 2
0.03333333 0.05714286
```

Decisión: No se rechaza H_0

Aplicación 5

Existen indicios de que uno de los equipos podría estar tardando más en completar sus tareas, lo que impactaría la eficiencia y los tiempos de entrega. ¿El tiempo promedio de producción por unidad es el mismo en ambos equipos, o hay una diferencia significativa?

Sol.

Primero se verificará si las varianzas son iguales

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \qquad \alpha = 0.05$$

```r
var.test(Tiempo ~ Grupo, datos, alternative = "two.sided")
```

```
    F test to compare two variances

data:  Tiempo by Grupo
F = 0.67659, num df = 29, denom df = 34, p-value = 0.2865
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3348314 1.3952971
sample estimates:
ratio of variances
         0.6765865
```

```r
tiempo1 <- datos |> filter(Grupo == "Equipo 1") |> pull(Tiempo)
tiempo2 <- datos |> filter(Grupo == "Equipo 2") |> pull(Tiempo)
var.test(tiempo1, tiempo2, alternative = "two.sided")
```

```
    F test to compare two variances

data:  tiempo1 and tiempo2
F = 0.67659, num df = 29, denom df = 34, p-value = 0.2865
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3348314 1.3952971
sample estimates:
ratio of variances
         0.6765865
```

No se rechaza H_0 ya que $pv > \alpha$, por lo tanto las varianzas son homogéneas.

Ahora, se prueba la diferencia de medias considerando que las varianzas son homogéneas:

$$H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 \neq 0 \qquad \alpha = 0.05$$

```
t.test(Tiempo ~ Grupo, datos, alternative = "two.sided", var.equal = TRUE)
```

```
    Two Sample t-test

data:  Tiempo by Grupo
t = -2.3495, df = 63, p-value = 0.02195
alternative hypothesis: true difference in means between group Equipo 1 and
95 percent confidence interval:
 -5.4573575 -0.4407378
sample estimates:
mean in group Equipo 1 mean in group Equipo 2
              44.05667               47.00571
```

```r
tiempo1 <- datos |> filter(Grupo == "Equipo 1") |> pull(Tiempo)
tiempo2 <- datos |> filter(Grupo == "Equipo 2") |> pull(Tiempo)
t.test(tiempo1, tiempo2, alternative = "two.sided", var.equal = TRUE)
```

```
    Two Sample t-test

data:  tiempo1 and tiempo2
t = -2.3495, df = 63, p-value = 0.02195
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.4573575 -0.4407378
sample estimates:
mean of x mean of y
 44.05667  47.00571
```

Se rechaza H_0, ya que $pv < \alpha$, lo que indica que hay diferencias significativas entre los tiempos promedio de producción de los equipos.

Ejercicio

Una empresa de logística busca mejorar la eficiencia de su operación y ha identificado dos centros de distribución que manejan envíos a distintas regiones. La gerencia ha recibido reportes sobre diferencias en los tiempos de entrega, costos de distribución y la cantidad de paquetes entregados con demora entre los centros. Antes de tomar decisiones estratégicas, desean realizar un análisis estadístico riguroso para determinar si estas diferencias son significativas o si pueden atribuirse al azar.

Se ha tomado una muestra aleatoria de los tiempos de entrega (en horas), costos de distribución (en dólares) y registros de entregas tardías ($1 =$ entrega tardía, $0 =$ entrega a tiempo) en cada centro de distribución.

Los datos están disponibles en el archivo Logistica.csv.

1. La empresa ha establecido que el tiempo promedio de entrega debe ser menor a 24 horas para cumplir con los estándares de servicio. Sin embargo, hay sospechas de que el centro 1 no está cumpliendo esta especificación, lo que podría generar retrasos en la entrega.
2. La empresa busca mantener estabilidad en los costos de distribución. Se ha establecido que la desviación estándar no debe ser mayor a 2 dólares. Hay indicios de que el centro 2 tiene costos con alta variabilidad. Verifique esta afirmación.
3. La empresa establece que la tasa máxima aceptable de entregas tardías es del 6%. Se sospecha que uno de los centros está superando esta tasa, lo que afectaría la percepción del servicio por parte de los clientes.
4. ¿Existe una diferencia significativa en la tasa de entregas tardías entre los dos centros, o las diferencias observadas son solo aleatorias?
5. ¿El tiempo promedio de entrega en del centro 1 es más de 3 horas mayor que en el otro, o la diferencia observada es solo variabilidad aleatoria?

Regresión lineal

- En un caso univariado, nos puede ser de interés estudiar la media de una variable aleatoria Y, es decir $\mu = E(Y)$, así como su variabilidad mediante un intervalo de confianza:

$$IC(\mu) = \overline{y} \pm t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

- Sin embargo, la variable Y puede estar influenciada por otra, X, de tal modo que el interés ahora se centre en $E(Y|X)$
- ¿Qué características deben tener X e Y? ¿Cómo cambia el proceso de inferencia?

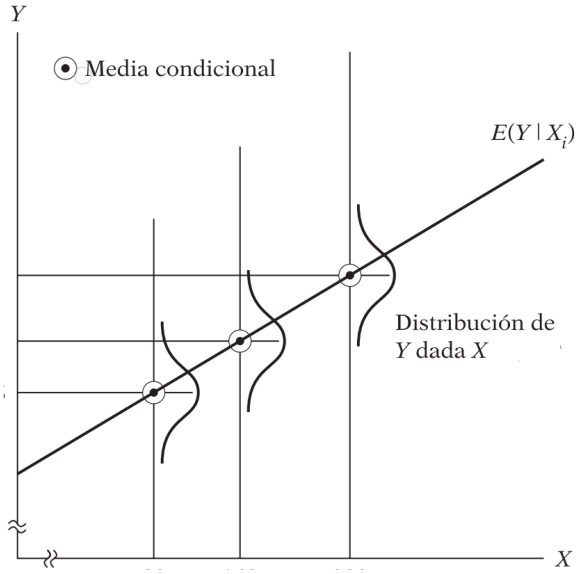

Y
Media condicional
$E(Y \mid X_i)$
Distribución de
Y dada X
X

Gráfico de dispersión

- Visualización de las variables X (independiente) y Y (dependiente) en el plano cartesiano.
- Útil para detectar patrones: ¿Hay relación lineal?
- Permite identificar posibles valores atípicos.
- Sirve para explorar la dirección y fuerza de la relación entre X e Y.

Ejemplo

```
datos <- read.csv('Salud.csv')
plot(datos$IMC,datos$Presion_sistolica, pch = 18)
```

Ejemplo

Modelo de regresión lineal simple

Modelo

Dadas las variables X e Y se define el modelo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + \epsilon_i = \mu_i + \epsilon_i \qquad i = 1, ..., n$$

donde β_0 es el intercepto, β_1 es la pendiente y ϵ es el término de error aleatorio. Nuestro objetivo será estimar β_0, β_1, ..., β_p a través de $\hat{\beta}_0$, $\hat{\beta}_1$ y predecir los ϵ_i mediante $\hat{\epsilon}$ utilizando los datos muestrales.

Estimación de los coeficientes

Estimación puntual

La estimación por máxima verosimilitud asume que $\epsilon_i \sim N\left(0, \sigma^2\right)$, por lo tanto, dado β_0, β_1 y X_i, entonces:

$$Y_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$$

$$f(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right\}$$

$$f(Y_1)\times\ldots\times f(Y_n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{Y_1 - \beta_0 - \beta_1 X_1}{\sigma}\right)^2\right\}\times\ldots\times\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{Y_n - \beta_0 - \beta_1 X_n}{\sigma}\right)^2\right\}$$

$$L\left(\beta_0, \beta_1, \sigma | Y_1, \ldots, Y_n\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{Y_1 - \beta_0 - \beta_1 X_1}{\sigma}\right)^2\right\}\times\ldots\times\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{Y_n - \beta_0 - \beta_1 X_n}{\sigma}\right)^2\right\}$$

Derivando L respecto a β_0 y β_1 se tiene:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

El método de **mínimos cuadrados** busca encontrar los parámetros β_0 y β_1 que minimizan la **suma de los cuadrados de los residuos**:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Se deriva $S(\beta_0, \beta_1)$ con respecto a β_0 y β_1, y se iguala a cero:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

Resolviendo el sistema de ecuaciones normales, se obtiene:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Ejemplo

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \mu_i + \epsilon_i \qquad i = 1, ..., 100$$
$$\mu_i = \beta_0 + \beta_1 X_i$$
$$\hat{\mu}_i = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

donde Y_i es la i-ésima presión sistólica, y X_i es el i-ésimo IMC.

```
datos <- read.csv('Salud.csv')
modelo <- lm(Presion_sistolica ~ IMC, datos)
modelo |> coef()
```

```
(Intercept)         IMC
  64.557707    3.095683
```

$$\hat{\mu}_i = \hat{Y}_i = 64.5577 + 3.0957 X_i$$

$\hat{\beta}_0$: Es el valor medio de Y cuando $X = 0$. No siempre es interpretable.

$\hat{\beta}_1$: Es la variación en la media de Y (puede ser incremento o disminución dependiendo del signo) cuando X se incrementa en una unidad.

Ejemplo

En el modelo de la presión sistólica dependiendo del IMC se tiene que:

$\hat{\beta}_0 = 64.5577$ no tiene interpretación porque el IMC no puede ser cero.

$\hat{\beta}_1 = 3.0957$ significa que la presión sistólica promedio se incrementa en 3.0957 mmHg por cada punto adicional de IMC.

Estimación intervalar

Recordemos que los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimaciones y están sujetos al margen de error.

Los intervalos de confianza se construyen mediante:

$$IC(\beta_j) = \hat{\beta}_j \mp t_{1-\alpha/2,n-1} s_{\hat{\beta}_j}$$

donde $s_{\hat{\beta}_j}$ es el error estándar del coeficiente $\hat{\beta}_j$ y $t_{1-\alpha/2,n-1}$ es el cuantil $1 - \frac{\alpha}{2}$ de la distribución t de Student con $n - 1$ grados de libertad

Ejemplo

```
modelo |>  summary()
```

```
Call:
lm(formula = Presion_sistolica ~ IMC, data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-25.2331  -6.6140   0.2444   8.7547  28.4817

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.5577    10.5147   6.140 1.77e-08 ***
IMC           3.0957     0.5484   5.645 1.61e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.1 on 98 degrees of freedom
Multiple R-squared:  0.2454,    Adjusted R-squared:  0.2377
F-statistic: 31.86 on 1 and 98 DF,  p-value: 1.61e-07
```

```r
library(broom)
modelo |> tidy()
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic      p.value
  <chr>          <dbl>     <dbl>     <dbl>        <dbl>
1 (Intercept)    64.6      10.5       6.14 0.0000000177
2 IMC             3.10      0.548     5.64 0.000000161
```

```r
modelo |> confint()
```

```
                2.5 %    97.5 %
(Intercept) 43.691723 85.423691
IMC          2.007387  4.183979
```

```r
modelo |> confint(level = 0.99)
```

```
                0.5 %   99.5 %
(Intercept) 36.936442 92.17897
IMC          1.655056  4.53631
```

Prueba de hipótesis

Prueba de significancia del modelo

¿X tiene influencia lineal sobre Y? Podemos verificarlo probando si $H_0 : \beta_1 = 0$ es cierta a través del análisis de varianza:

FV	GL	SC	CM	F_{calc}
Regresión	1	SCReg	CMReg	CMReg/CME
Error	n-2	SCError	CME	
Total	n-1	SCTotal		

La hipótesis nula se rechaza si $F_{calc} > F_{1-\alpha,1,n-2}$ lo que es equivalente a $pv < \alpha$

Ejemplo

$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$

$\alpha = 0.10$

```
modelo |> aov() |> summary()
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
IMC          1   3924    3924   31.86 1.61e-07 ***
Residuals   98  12068     123
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$pvalor = 1.61 \times 10^{-7}$

Decisión: Rechazar la hipótesis nula

Existe suficiente evidencia estadística para indicar que existe una relación lineal de dependencia de la presión sistólica en función del IMC.

Errores y residuos

- Un residual es la realización de los errores del modelo de regresión lineal: $e_i = y_i - \hat{y}_i$
- Representan la parte del modelo no explicada por la variable independiente.
- Su comportamiento nos informa acerca del cumplimiento de supuestos del modelo.
- Como mínimo, su media debe ser cero y su variancia, constante
- Además, consideraremos otros supuestos como el de normalidad e independencia de errores

```
modelo |> residuals() |> head(5)
```

```
         1          2          3          4          5
-2.7790675 15.2190637 11.9094954  0.3391347 14.5774080
```

```
modelo |> residuals() |> mean()
```

```
[1] 1.981401e-16
```

Supuestos del modelo

Normalidad de errores

- Supuesto: Los errores se distribuyen normalmente
- Modelo de regresión lineal es robusto a la falta de cumplimiento de este supuesto si el tamaño de muestra es grande.

Medios de verificación:

- Histograma / curva de densidad
- Coeficiente de asimetría y curtosis: debe ser igual a 0 y 3, respectivamente
- Gráfico de probabilidad normal: residuales (estudentizados) ordenados $t_{(i)}$ versus la probabilidad acumulada normal $\Phi^{-1}\left[\frac{i-0.5}{n}\right]$ para $i = 1, ..., n$. Se sugiere que $n > 20$ (Peck, Vining y Montgomery, 2012). Permite detectar outlier, asimetría y curtosis.
- Pruebas de normalidad: Shapiro Wilk, Anderson Darling, Kolmogorov Smirnov. Tienden a mostrar baja potencia de prueba cuando no se cumple el supuesto (Mendenhall, 2011), es decir no detectan bien errores que no son normales.

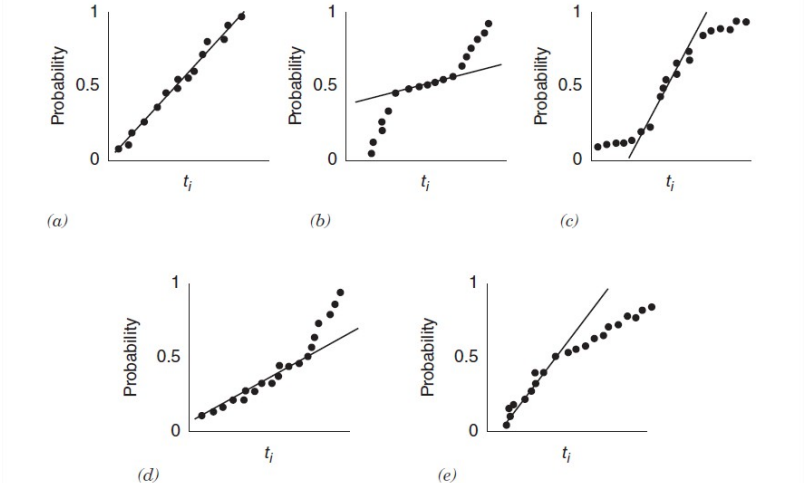


Figure 4.3 Normal probability plots: (*a*) ideal; (*b*) light-tailed distribution; (*c*) heavy-tailed distribution; (*d*) positive skew; (*e*) negative skew.

Generar y analizar los siguientes gráficos:

```
par(mfrow=c(2,2))

modelo |> residuals() |>
  hist(main = "Histograma de los residuales")

modelo |> residuals() |>
  density() |> plot(main = "Densidad de los residuales")

modelo |> plot(which = 2)

modelo |> residuals() |>
  boxplot(main = "Boxplot de los residuales")
```

H_0 : Los errores siguen una distribución Normal

H_1 : Los errores no siguen una distribución Normal

```
    Shapiro-Wilk normality test

data:  residuals(modelo)
W = 0.98637, p-value = 0.3964
```

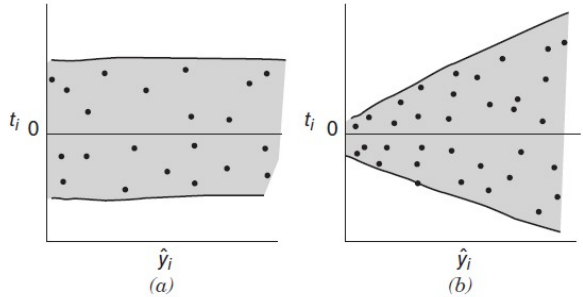
Conclusión: ...

Homogeneidad de varianzas de los errores

- Supuesto: Los errores tienen varianza homogénea, es decir $V(\epsilon_i) = \sigma^2, \forall i = 1, ..., n$

Medios de verificación:

- Gráfico de valores ajustados versus residuales
- Gráfico de valores ajustados versus la raíz cuadrada de los residuales estudentizados en valor absoluto
- Gráfico de valores de cada variable explicativa versus residuales (estudentizados). Conviene también intentar con variables independientes no consideradas en el modelo.
- Prueba de Breusch Pagan


t_i
0
$\hat{y}_i$
(a)
t_i
0
$\hat{y}_i$
(b)
t_i
0
$\hat{y}_i$
(c)
t_i
0
$\hat{y}_i$
(d)

Generar y analizar los siguientes gráficos:

```
modelo |> plot(which=1)
modelo |> plot(which=3)
plot(datos$IMC, residuals(modelo))
```

H_0 : Los errores presentan varianza constante

H_1 : Los errores no presentan varianza constante

```
library(car)
modelo |> ncvTest()
```

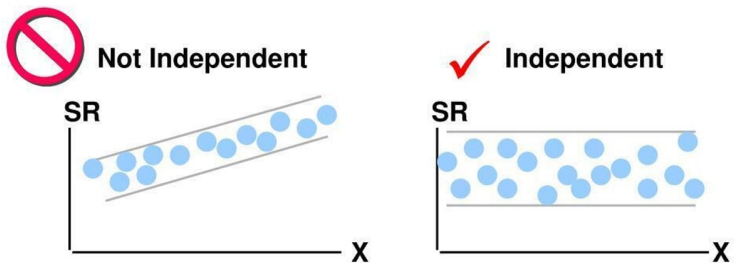
```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.577259, Df = 1, p = 0.20916
```

Conclusión: ...

Independencia de errores

- Supuesto: Los errores son independientes = los errores no están autocorrelacionados
- Medios de verificación:
  - Gráfico de los residuales en orden
  - Correlograma
  - Prueba de Durbin Watson

Residual Analysis for Independence



```r
modelo |> residuals() |>
  plot(main = "Residuales")

modelo |> residuals() |>
  acf()
```

H_0 : Los errores son independientes

H_1 : Los errores no son independientes

```r
library(car)
modelo |> durbinWatsonTest()
```

```
 lag Autocorrelation D-W Statistic p-value
   1     0.005751365      1.969609    0.87
 Alternative hypothesis: rho != 0
```

Conclusión: ...

Calidad del modelo

Coeficiente de determinación

- Denotado por R^2
- Mide el porcentaje de variabilidad de la variable respuesta que es explicado por la variable predictora
- Va de 0 a 1 (o de 0% a 100%)
- No es una medida de ajuste ni indicador de adecuación
- No penaliza la inclusión de más variables
- Se obtiene mediante:

$$R^2 = \frac{SCReg}{SCTotal}$$

Ejemplo

```
library(magrittr)
modelo |> summary() |> extract("r.squared")
```

```
$r.squared
[1] 0.2453672
```

El 24.54% de la variabilidad de la presión sistólica es explicada por el IMC.

Predicción

Estimación de la media

Consiste en dar un valor o un intervalo de valores para μ cuando X toma determinado(s) valor(es).

$$\hat{\mu} = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Su varianza estimada es:

$$\hat{V}(\hat{\mu}) = \hat{V}(\hat{y}) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{SCX} \right)$$

A partir de estos valores, y asumiendo distribución Normal (pronto verificaremos ese supuesto), es posible construir intervalos de confianza para la media estimada.

Ejemplo

```
# Estimación puntual
modelo |> predict(data.frame(IMC = 22))
```

```
       1
132.6627
```

Se estima que la presión sistólica media cuando el IMC es de 22 puntos es de 132.66 mmHg.

```
modelo |> predict(data.frame(IMC = c(18,20,23)))
```

```
       1        2        3
120.2800 126.4714 135.7584
```

Se estima que la presión sistólica media cuando el IMC es de 18 puntos es de 120.28 mmHg, cuando el IMC es 20 es de 126.47 mmHg, y cuando el IMC es 23, es de 135.76 mmHg.

```r
# Estimación intervalar
modelo |> predict(data.frame(IMC = c(18,20,23)),
                  level = 0.95, interval = "confidence")
```

```
       fit      lwr      upr
1 120.2800 117.7909 122.7691
2 126.4714 124.0459 128.8968
3 135.7584 130.9439 140.5729
```

Para un IMC de 18, se espera que el valor medio de la presión sistólica sea 120.28 mmHg, con un intervalo de confianza del 95% entre 117.79 y 122.77 mmHg. Para un IMC de 20, el valor medio estimado es 126.47 mmHg, con intervalo entre 124.05 y 128.90 mmHg. Para un IMC de 23, la estimación es 135.76 mmHg, con intervalo entre 130.94 y 140.57 mmHg.

Predicción individual

Consiste en dar un valor o un intervalo de valores para Y cuando X toma determinado(s) valor(es).

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x$$

Su varianza estimada es:

$$\hat{V}(\hat{y}_0) = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{SCX} \right)$$

Ejemplo

```
# Predicción puntual
modelo |> predict(data.frame(IMC = 22))
```

```
       1
132.6627
```

La presión sistólica predicha cuando el IMC es de 22 puntos es de 132.66 mmHg.

```
modelo |> predict(data.frame(IMC = c(18,20,23)))
```

```
       1        2        3
120.2800 126.4714 135.7584
```

Se prevé que la presión sistólica sea 120.28 mmHg cuando el IMC es de 18 puntos, 126.47 mmHg cuando IMC $= 20$ y 135.76 mmH cuando IMC $= 23$,

```r
# Estimación intervalar
modelo |> predict(data.frame(IMC = c(18,20,23)),
                  level = 0.95, interval = "prediction")
```

```
       fit       lwr      upr
1 120.2800  98.11787 142.4421
2 126.4714 104.31629 148.6264
3 135.7584 113.21636 158.3005
```

Para un IMC de 18, se espera que la presión sistólica sea 120.28 mmHg, con un intervalo de predicción del 95% entre 98.12 y 142.4421 mmHg. Para un IMC de 20, el valor predicho es 126.47 mmHg, con intervalo entre 104.32 y 148.63 mmHg. Para un IMC de 23, la predicción es 135.76 mmHg, con intervalo entre 113.22 y 158.3 mmHg.

Modelo de regresión lineal múltiple

- Hemos visto que podemos explicar la relación de dependencia entre un par de variables a través de una línea de regresión.
- La variable respuesta es generalmente explicada por más de una variable independiente, de tal modo que

$$y = f(x_1, ..., x_p, \epsilon)$$

Modelo

Dadas las p variables independientes X_1,...,X_p fijas y la variable respuesta Y, se define el modelo:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_p X_{p,i} + \epsilon_i = \mu_i + \epsilon_i \qquad i = 1, ..., n$$

donde β_0 es el intercepto, β_1, ..., β_p son las pendientes, ϵ es el término de error aleatorio cuya media es cero y su varianza es constante. En caso las variables X_j sean aleatorias, sus observaciones deben ser por lo menos independientes. Nuestro primer objetivo será estimar β_0, β_1, ..., β_p a través de $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_p$ y predecir los ϵ_i mediante $\hat{\epsilon}$, utilizando los datos muestrales.

Dado que $E(\epsilon_i) = 0$ y $V(\epsilon_i) = \sigma^2$, entonces $E(Y_i|\mathbf{X}_i, \beta) = \mathbf{X}_i'\beta$ y $V(Y_i|\mathbf{X}_i, \beta) = \sigma^2$

Importante: Note que p es el número de variables y k es el número de coeficientes de regresión (que llamamos betas), entonces

$$k = p + 1$$

Por otro lado, en notación matricial, tenemos que:

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times k}\beta_{k\times_1} + \epsilon_{n\times 1} = \mathbf{X}_{n\times(p+1)}\beta_{(p+1)\times 1} + \epsilon_{n\times 1}$$

Estimación de los coeficientes

- Métodos de estimación: Mínimos cuadrados ordinarios, Máxima verosimilitud.
- Uso de matrices para la estimación de parámetros
- Vector de coeficientes de regresión estimados:

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

- ¿Cómo se interpretan estos coeficientes?
- El valor estimado de σ^2 es el mismo que en la regresión lineal simple

```
modelo2 = lm(Presion_sistolica ~ Edad + Minutos_ejercicio + IMC, datos)
modelo2 |> coef()
```

```
(Intercept)                Edad  Minutos_ejercicio                 IMC
110.47710380          0.45472167         -0.08487366          0.27956285
```

$$\hat{\mu}_i = \hat{y}_i = 110.48 + 0.455X_{1i} - 0.085X_{2i} + 0.28X_{3i}$$

donde:

- X_{1i} es la i-ésima edad.
- X_{2i} es el i-ésimo tiempo semanal de ejercicios, en minutos.
- X_{3i} es el i-ésimo IMC.

- $\hat{\beta}_0 = 110.48$ no tiene interpretación.
- $\hat{\beta}_1 = 0.455$ significa que cuando la edad se incrementa en un año, la media de la presión sistólica se incrementa en 0.455 mmHg, manteniendo constante el tiempo semanal de ejercicios y el IMC.
- $\hat{\beta}_2 = -0.085$ significa que cuando el tiempo semanal de ejercicios se incrementa en un minuto, la media de la presión sistólica disminuye en 0.085 mmHg, manteniendo constante la edad y el IMC.
- $\hat{\beta}_3 = 0.28$ significa que cuando el IMC se incrementa en un punto, la media de la presión sistólica se incrementa en 0.28 mmHg, manteniendo constante el tiempo semanal de ejercicios y la edad.

Prueba de hipótesis

Prueba de significancia del modelo

- También llamada prueba de significancia de la regresión
- Busca contrastar ls hipótesis:
  $H_0 : \beta_1 = ... = \beta_p = 0$
  H_1 : Al menos un $\beta_j \neq 0, \qquad j = 1, ..., p$
- Se prueba mediante el Análisis de Varianza

FV	GL	SC	CM	F_{calc}
Regresión	k-1	SCReg	CMReg	CMReg/CME
Error	n-k	SCError	CME	
Total	n-1	SCTotal		

- La hipótesis nula se rechaza si $F_{calc} > F_{1-\alpha,k-1,n-k}$ lo que es equivalente a $pv < \alpha$
- Recuerde que k es el número de coeficientes β estimados y p es el número de variables explicativas en el modelo

Ejemplo $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ H_1 : Al menos un $\beta_j \neq 0$ $\alpha = 0.05$

```
X = cbind(1,datos$Edad, datos$Minutos_ejercicio, datos$IMC)
modelo2_ = lm(Presion_sistolica ~ X, datos)
modelo2_ |> aov() |> summary()
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
X             3  11282    3761   76.65 <2e-16 ***
Residuals    96   4710      49
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$pvalor < 2 \times 10^{-16}$

Decisión: Rechazar la hipótesis nula

Existe suficiente evidencia estadística para indicar que existe una relación lineal de dependencia de la presión sistólica en función del IMC, la edad y el tiempo semanal de ejercicios.

Prueba individual de coeficientes

```
modelo2 |> summary()
```

```
Call:
lm(formula = Presion_sistolica ~ Edad + Minutos_ejercicio + IMC,
    data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-17.2618  -5.3914  -0.0236   4.3979  18.8574

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       110.477104   8.598105  12.849  < 2e-16 ***
Edad                0.454722   0.052810   8.610 1.43e-13 ***
Minutos_ejercicio  -0.084874   0.009152  -9.274 5.41e-15 ***
IMC                 0.279563   0.421363   0.663    0.509
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.005 on 96 degrees of freedom
Multiple R-squared:  0.7055,    Adjusted R-squared:  0.6963
F-statistic: 76.65 on 3 and 96 DF,  p-value: < 2.2e-16
```

Calidad del modelo

Coeficiente de determinación

- Denotado por R^2
- Mide el porcentaje de variabilidad de la variable respuesta que es explicado por la variable predictora
- Va de 0 a 1 (o de 0% a 100%)
- No es una medida de ajuste ni indicador de adecuación
- No penaliza la inclusión de más variables

Ejemplo

```
modelo2 |> summary() |> extract("r.squared")
```

```
$r.squared
[1] 0.7054706
```

Coeficiente de determinación ajustado Fluctúa entre 0 y 1. Su valor ajustado disminuye el efecto de una muestra pequeña o la inclusión de variable(s) que no aporta(n)

$$R^2_{aj} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

donde n es el tamaño de muestra y p es el número de variables independientes.

Ejemplo

```
modelo2 |> summary() |> extract("adj.r.squared")
```

```
$adj.r.squared
[1] 0.6962666
```

El 69.63% de la variabilidad de la presión sistólica es explicada por el IMC.

AIC

AIC: medida de bondad de ajuste que penaliza la complejidad de un modelo. Se puede corregir para tamaños de muestra pequeños:

$$AIC = -2log(L) + 2r$$

donde $log(L)$ es la log-verosimilitud del modelo y r es el número de parámetros estimados.

```
modelo2 |> AIC()
```

```
[1] 679.0197
```

Multicolinealidad

- Se debe a la dependencia lineal entre variables independientes
- Puede llevar a una singularidad de la matriz $\mathbf{X}'\mathbf{X}$
- Se diagnostica mediante el factor de inflación de varianza (VIF):

$$VIF = \frac{1}{1 - R_j^2}$$

donde R_j^2 es el coeficiente de determinación al ejecutar la regresión de X_j en función a las demás variables independientes.

- Si VIF>10, entonces existe un serio problema de multicolinealidad.
- Como consecuencia, genera estimaciones 'inestables' o 'sin sentido'

```
library(car)
modelo2 |> vif()
```

```
            Edad Minutos_ejercicio              IMC
        1.084372          1.382990         1.481694
```

Predicción

- Estimación puntual: Dado un vector $\mathbf{x} = (x_1, ..., x_p)$ que contiene valores de las variables explicativas, se tiene que:

$$\hat{\mu} = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p = \mathbf{x}'\beta$$

- Estimación intervalar: Requiere que se verifiquen los supuestos referidos a los errores

$$IC(\mu|\mathbf{x}) = \hat{\mu} \mp t_{1-\alpha/2, n-k} \sqrt{\hat{\sigma}^2 \mathbf{x}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}}$$

Recuerde que k es el número de coeficientes β estimados y p es el número de variables explicativas en el modelo

```r
modelo2 |>
  predict(data.frame(Edad = 30,
                     Minutos_ejercicio = 60,
                     IMC = 23))
```

```
       1
125.4563
```

Se estima que la presión sistólica media de un paciente de 30 años que hace una hora semanal de ejercicios y cuyo IMC es de 23 puntos es 125.45 mmHg.

```r
modelo2 |>
  predict(data.frame(Edad = 30,
                     Minutos_ejercicio = 60,
                     IMC = 23),
          interval = "confidence",
          level    = 0.95)
```

```
       fit      lwr      upr
1 125.4563 121.5621 129.3505
```

Se estima con un 95% de confianza que la presión sistólica media de un paciente de 30 años que hace una hora semanal de ejercicios y cuyo IMC es de 23 puntos está en el intervalo (121.56, 129.35) mmHg.

¿Cómo podemos evaluar la extrapolación?

- Se define la matriz $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ conocida como 'matriz hat'.
- Los valores de su diagonal permiten verificar si predecir la variable respuesta para un conjunto de valores **x** constituye una extrapolación
- Si se verifica que

$$\mathbf{x}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x} > h_{max}, \qquad h_{max} = max(h_{11}, ..., h_{nn})$$

entonces se trata de una extrapolación

Ejemplo:

¿El punto evaluado nos lleva a una extrapolación?

```
x = c(1,30,60,23)
H = X%*%solve(t(X)%*%X)%*%t(X)
h = H %>% diag %>% max
t(x)%*%solve(t(X)%*%X)%*%x
```

```
            [,1]
[1,] 0.07844309
```

```
t(x)%*%solve(t(X)%*%X)%*%x > h
```

```
      [,1]
[1,] FALSE
```

Aplicación

Una organización ambiental desea analizar los factores que influyen en la temperatura de una zona urbana durante distintos días del año. Para ello, ha recopilado datos diarios sobre las siguientes variables:

- Temperatura (°C): variable dependiente, representa la temperatura media del día.
- Humedad (%): porcentaje de humedad relativa.
- Velocidad del viento (m/s): medida de la intensidad del viento.
- Radiación solar (W/m^2): cantidad de energía solar incidente.
- Contaminación (AQI): índice de calidad del aire (a mayor valor, peor calidad).

Se registraron datos durante 100 días al azar, los cuales se encuentran en el archivo `Ambiental.csv`.

a. Ajusta un modelo de regresión lineal donde la variable dependiente sea Temperatura, y las variables independientes sean Humedad, Velocidad del viento, Radiación solar y Contaminación.

b. Escribe la ecuación del modelo estimado.

c. Interpretar los coeficientes estimados del modelo.

d. Probar la significancia del modelo

e. ¿Qué variable tiene el mayor impacto unitario en la temperatura?

f. ¿Hay alguna variable cuyo efecto no sea significativo al 5%?

g. ¿Cuál es el coeficiente de determinación ajustado? ¿Qué indica sobre el modelo?

h. Estima la temperatura para un día que tenga las siguientes condiciones: Humedad = 60%, Velocidad del viento = 5 m/s, Radiación solar = 800 W/m^2, Contaminación = 100 AQI. ¿Se trata de una extrapolación?