

UNIVERSIDAD DE BUENOS AIRES
Facultad de Ingeniería

Clases de Aprendizaje Estadístico

Ing. Jemina M. García

Buenos Aires, 2020.

Índice general

1. Repaso	1
1.1. Variables aleatorias	1
1.1.1. Variables aleatorias discretas	2
1.1.2. Variables aleatorias continuas	3
1.1.3. Variables aleatorias: tipo de variable según su función de distribución	4
1.1.4. Vectores aleatorios	6
1.1.5. Independencia	8
1.2. Momentos	9
1.2.1. Esperanza de una variable aleatoria	9
1.2.2. Varianza de una variable aleatoria	9
1.2.3. Vectores aleatorios	10
1.3. Predicción	12
1.3.1. Variables aleatorias condicionadas	12
2. Modelo lineal	17
2.1. Supuestos del modelo lineal	19
2.2. Estimación de los parámetros	20
2.2.1. Interpretación geométrica	22
2.2.2. Propiedades del estimador de mínimos cuadrados	23
2.3. Estimación de σ^2	24
2.4. Algunos teoremas útiles	25
2.5. Test y regiones de confianza	25
2.5.1. Test de hipótesis	26
3. Anexo 1	29
3.1. Test para hipótesis simple vs. hipótesis simple	32
3.2. Tests uniformemente mas potentes para hipótesis unilaterales	34
3.3. Test del cociente de máxima verosimilitud	36
3.4. Test con nivel de significación asintótico	37
3.4.1. Distribución asintótica del test del cociente de máxima verosimilitud	38

4. Bibliografía**39**

Capítulo 1

Repaso

1.1. Variables aleatorias

Al describir el espacio muestral de un experimento, los resultados individuales no necesariamente son números. Dado un experimento y sea Ω el espacio muestral asociado a él, Una función X que asigna a cada uno de los elementos $\omega \in \Omega$ un número real $X(\omega)$ se llama variable aleatoria.

Definición: Sea $(\Omega, \mathcal{A}, \mathbf{P})$ un espacio de probabilidad y $X : \Omega \rightarrow \mathbb{R}$ una función, diremos que X es una variable aleatoria si $X^{-1}(B) \in \mathcal{A}$

Proposición: Sea $(\Omega, \mathcal{A}, \mathbf{P})$ un espacio de probabilidad y X una V.A., entonces $X^{-1}(B) \in \mathcal{A}$ (B es un subconjunto del sigma álgebra de Borel) Luego, se le puede calcular la probabilidad, es decir $\mathbf{P}(X^{-1}(B)) = \mathbf{P}(X \in B)$

Observación: $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$

Definición: Sea $(\Omega, \mathcal{A}, \mathbf{P})$ un espacio de probabilidad y X una V.A., definimos su función de distribución acumulada como la función $F_X : \mathbf{R} \rightarrow [0, 1]$ dada por

$$F_X(x) = \mathbf{P}(X \leq x) \quad \forall x \in \mathbb{R}$$

Propiedades:

1. $F_X(x) \in [0, 1], \forall x \in \mathbb{R}$
2. $F_X(x)$ es monótona no decreciente
3. $F_X(x)$ es continua a derecha
4. $\lim_{x \rightarrow -\infty} F_X(x) = 0$ y $\lim_{x \rightarrow \infty} F_X(x) = 1$

1.1.1. Variables aleatorias discretas

Definición: Sea $(\Omega, \mathcal{A}, \mathbf{P})$ un espacio de probabilidad y X una V.A., diremos que X es una V.A. discreta cuando existe $A \in \mathbb{R}$ finito o numerable tal que $\mathbf{P}(A) = 1$

Obs: A no es único, porque le puedo seguir agregando elementos (que tengan probabilidad 0).

Una variable aleatoria discreta es aquella cuyos valores posibles constituyen un conjunto finito o infinito numerable.

Definición: Llamamos **rango** de la V.A. X a $R_X = \{x \in \mathbb{R} : p_X(x) > 0\}$

Obs: R_X satisface que es el menor de los A

Definición: Sea X una V.A.D., se llama función de probabilidad de X a una función $p_X : \mathbb{R} \rightarrow [0, 1]$ tal que $p_X(x) = \mathbf{P}(X = x)$. Con cada resultado posible x_i asociamos un número $p_X(x_i) = \mathbf{P}(X = x_i)$. Los $p_X(x_i)$ deben satisfacer:

1. $p_X(x_i) \geq 0 \forall i$
2. $\sum_{x \in R_X} p_X(x) = 1$

La función p se llama función de probabilidad de la V.A. X y se denomina $p_X(x)$

Teorema: Se a X una V.A.D. entonces:

$$p_X(B) = \sum_{x \in B} p_X(x)$$

Si la variable aleatoria es discreta, además se tiene que:

$$F_X(x) = \mathbf{P}(X \leq x) = \sum_{a \leq x} p_X(a)$$

Obs: Si X es una VAD, la gráfica de F se formará de trazos horizontales (llamada función escalonada). F es continua excepto para los valores probables de X . En el valor x_i la gráfica tendrá un salto de magnitud $p_X(x_i)$

Sea X una VAD con valores posibles x_j , entonces:

$$p_X(x_j) = \mathbf{P}(X = x_j) = F_X(x_j) - F_X(x_{j-1})$$

Modelos discretos :

Una variable aleatoria de **Bernoulli** es aquella cuyos únicos valores posibles son 0 y 1, y que le asigna $\mathbf{P}(X = 1) = p$ y $\mathbf{P}(X = 0) = 1 - p$

Función indicadora:

$$\mathbf{1}\{a < x < b\} = 1 \text{ si } a < x < b, 0 \text{ en otro caso}$$

Otras distribuciones: Binomial, Hipergeométrica, Geométrica, Pascal, Poisson.

1.1.2. Variables aleatorias continuas

Definición: Una V.A. es continua si se cumplen las siguientes condiciones:

1. Su conjunto de valores posibles se compone de todos los números que hay en un solo intervalo que hay sobre la línea de numeración o todos los números en una unión excluyente de dichos intervalos
2. Ningún valor posible de la variable aleatoria tiene probabilidad positiva, o sea, $\mathbf{P}(X = c) = 0, \quad \forall c \in \mathbb{R}$

Definición: Se dice que X es una V.A. continua si existe una función $f_X : \mathbb{R} \rightarrow \mathbb{R}$, llamada función de densidad de probabilidad de X , que satisface las siguientes condiciones:

1. $f_X(x) \geq 0 \forall x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f_X(x) dx = 1$
3. Para cualquier a, b tal que $-\infty < a < b < \infty$ tenemos $\mathbf{P}(a < x < b) = \int_a^b f_X(x) dx$

Si X es una VAC entonces $F_X(x) = \int_{-\infty}^x f_X(t) dt$

Obs: Si X es una V.A.C., $F_X(x)$ es una función continua para todos los reales (Admite derivada)

Teorema: Sea $F_X(x)$ la función de distribución de una V.A.C. (admite derivada) con f.d.p $f_X(x)$, luego:

$$f_X(x) = \frac{d}{dx} F_X(x)$$

1.1.3. Variables aleatorias: tipo de variable según su función de distribución

Átomos: Diremos que $a \in \mathbb{R}$ un átomo de $F_X(x)$ si su peso es positivo, es decir: $\mathbf{P}(X = a) > 0$.

El conjunto de todos los átomos de $F_X(x)$: $A = \{a \in \mathbb{R} : P(X = a) > 0\}$, coincide con el conjunto de todos los puntos de discontinuidad de $F_X(x)$.

De esta forma podemos definir el tipo de variable aleatoria según su función de distribución:

- La variable aleatoria X será discreta si la suma de las probabilidades de todos los átomos es 1
- La variable aleatoria X será continua si su función de distribución es continua
- **Diremos que una variable aleatoria es mixta si no es continua ni discreta**

Modelos continuos :

Distribución uniforme

Supongamos que X es una V.A.C. que toma todos los valores en el intervalo $[a, b]$ donde ambos son finitos. Si $f_X(x)$ está dada por

$$f_X(x) = \frac{1}{b-a} \mathbf{1}_{\{a < x < b\}}$$

Se dice que X está distribuida uniformemente en (a, b) y se nota $X \sim \mathcal{U}(a, b)$

Distribución esponencial

Una variable aleatoria tiene distribución exponencial de parámetro $\lambda > 0$ si su función de densidad de probabilidad es:

$$f_X(x) = \lambda e^{-\lambda x} \mathbf{1}_{\{x > 0\}}$$

Se dice que X tiene distribución exponencial y se nota $X \sim \mathcal{E}(\lambda)$

Propiedades:

1. Si $X \sim \mathcal{E}(\lambda)$ entonces $Y = \frac{X}{\lambda} \sim \mathcal{E}(\lambda)$
2. Si $X \sim \mathcal{E}(\lambda)$ entonces $\mathbf{P}(X > t + s | X > t) = \mathbf{P}(X > s) \forall t, s \in \mathbb{R}^+$

3. Si X es una VAC y $\mathbf{P}(X > t + s | X > t) = \mathbf{P}(X > s) \forall t, s \in \mathbb{R}^+$, entonces exista $\lambda > 0$ tal que $X \sim \mathcal{E}(\lambda)$

En otras palabras, se dice que la variable aleatoria con distribución exponencial tiene *pérdida de memoria*.

Distribución Gamma

Función gamma: $\Gamma(y) = \int_0^\infty x^{y-1} e^{-x} dx$, si $y > 0$.

Caso particular: $\Gamma(1) = 1$

Si se desarrolla la integral por partes, se puede observar que $\Gamma(a) = (a-1)\Gamma(a-1)$

Si $a \in \mathbb{N}$, $\Gamma(a) = (a-1)!$ y a demás $\Gamma(1/2) = \sqrt{\pi}$.

Se dice que una variable aleatoria continua X tiene distribución Gamma de parámetros λ y k si la función de densidad de probabilidad de X es:

$$f_X(x) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} \mathbf{1}\{x > 0\}$$

Si $k \in \mathbb{N}$ entonces $\mathbf{P}(X > 0) = \sum_{i=0}^{k-1} \frac{(\lambda x)^i}{i!} e^{-\lambda x}$

Distribución Normal Estándar

La variable aleatoria X que toma todos los valores reales $-\infty < X < \infty$ tiene una distribución normal estándar si su función de densidad de probabilidad es de la forma:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Cuantil de la variable aleatoria X

$$x_\alpha = \min\{x \in \mathbb{R} : F_X(x) \geq \alpha\}$$

1.1.4. Vectores aleatorios

Definición: Sea $(\Omega, \mathcal{A}, \mathbf{P})$ un espacio de probabilidad, se dice que $X = (X_1, X_2, \dots, X_n)$ es un vector aleatorio de dimensión n , si para cada $j = 1, 2, \dots, n$, $X_j : \Omega \rightarrow \mathbb{R}$ es una variable aleatoria.

Teorema: Para todo $x = (x_1, \dots, x_n) \in \mathbb{R}$ se tendrá que $X^{-1}((-\infty, x_1) * (-\infty, x_2) * \dots * (-\infty, x_n)) \in \mathcal{A}$

Función de distribución de un vector aleatorio: Sea $\underline{X} = (X_1, X_2, \dots, X_n)$ un vector aleatorio de dimensión n , definimos la función de distribución de X como:

$$F_{\underline{X}}(\underline{x}) = \mathbf{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n;)$$

Propiedades, cuando $\underline{X} = (X, Y)$:

1. $\lim_{x, y \rightarrow \infty} F_X = 1, \lim_{x \rightarrow -\infty} F_X = 0, \lim_{y \rightarrow -\infty} F_X = 0$
2. $F_X(x)$ es monótona no decreciente en cada variable
3. Es continua a derecha en cada variable
4. $\mathbf{P}((a_1, b_1) \times (a_2, b_2)) = F_{X,Y}(b_1, b_2) - F_{X,Y}(b_1, a_2) - F_{X,Y}(a_1, b_2) + F_{X,Y}(a_1, a_2)$

Función de probabilidad de un vector aleatorio discreto Sean X e Y dos variables aleatorias discretas definidas en el espacio muestral Ω de un experimento. La función de probabilidad conjunta se define para cada par de números (x, y) como:

$$p_{X,Y}(x, y) = \mathbf{P}(X = x, Y = y)$$

Debe cumplirse que:

1. $p_{X,Y}(x, y) \geq 0$

$$2. \sum_x \sum_y p_{X,Y}(x, y) = 1$$

Sea A cualquier conjunto compuesto de pares de valores (x, y) entonces:

$$\mathbf{P}((X, Y) \in A) = \sum_{(x, y) \in A} p_{X,Y}(x, y)$$

Las funciones de probabilidad **marginales** de X e Y están dadas por:

$$p_X(x) = \sum_y p_{X,Y}(x, y)$$

$$p_Y(y) = \sum_x p_{X,Y}(x, y)$$

Función de densidad de un vector aleatorio continuo Sean X e Y variables aleatorias continuas, una función de densidad de probabilidad conjunta $f_{X,Y}(x, y)$ de estas dos variables es una función que satisface:

1. $f_{X,Y}(x, y) \geq 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$

Entonces para cualquier conjunto A

$$\mathbf{P}((X, Y) \in A) = \iint_A f_{X,Y}(x, y) dx dy$$

Las funciones de densidad de probabilidad marginales de X e Y están dadas por:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$$

Distribución Normal Multivariada

Se dice que el vector aleatorio X tiene distribución normal multivariada de dimensión p, $X \sim \mathcal{N}_p(\mu, \Sigma)$, de parametros $\mu \in \mathbb{R}^p$ y $\Sigma \in \mathbb{R}^{p \times p}$ (simétrica y definida positiva) si su función de densidad conjunta está dada por

$$f_X(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-1/2(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

1. Si $X \sim \mathcal{N}_p(0, \text{diag}(\lambda_1, \dots, \lambda_p))$, entonces X_1, \dots, X_p son independientes y tienen distribución $\mathcal{N}(0, \lambda_i)$.

2. Si $X \sim \mathcal{N}_p(\mu, \Sigma)$ y $A \in \mathbf{R}^{p \times p}$ es no singular, entonces $AX + b \sim \mathcal{N}_p(A\mu + b, A\Sigma A^T)$

Def: Se dice que X es normal multivariada si y solo si $\forall \in \mathbb{R}^p$ se tiene que $t^T X$ es normal univariada.

Es decir:

$$X \sim \mathcal{N}_p(\mu, \Sigma) \iff a^T X \sim \mathcal{N}_p(a^T \mu, a^T \Sigma a), \forall a \in \mathbb{R}^p$$

Además, X_1, \dots, X_p son independientes si y solo si Σ es una matriz diagonal.

Propiedades: Sea $X \sim \mathcal{N}_p(\mu, \Sigma)$, $\Sigma > 0$, $X = (X^{(1)}, X^{(2)})^T$, $X^{(1)} \in \mathbb{R}^{p_1}$, $X^{(2)} \in \mathbb{R}^{p_2}$, $p_1 + p_2 = p$

1. $X^{(1)} \sim \mathcal{N}_{p_1}(\mu^{(1)}, \Sigma_{11})$, $X^{(2)} \sim \mathcal{N}_{p_2}(\mu^{(2)}, \Sigma_{22})$
2. $X^{(1)}$ y $X^{(2)}$ son independientes sii $\Sigma_{12} = 0$
3. $A \in \mathbf{R}^{q \times p}$, $rg(a) = q$, $AX \sim \mathcal{N}_q(A\mu, A\Sigma A^T)$
4. $X \sim \mathcal{N}_p(\mu, \Sigma) \iff X = AZ + \mu$, con $Z \sim \mathcal{N}_p(0, I_p)$ y $AA^T = \Sigma$
5. $(X - \mu)^t \Sigma^{-1} (X - \mu) \sim \chi_p^2$

IMPORTANTE: Complementar con el material bibliográfico recomendado.

1.1.5. Independencia

Definición: Sea (X, Y) un vector aleatorio, las variables aleatorias X e Y son independientes sii

$$p_{x,y}((X \in A) \cap (Y \in B)) = p_x(X \in A) * p_{x,y}(Y \in B), \forall A, B \in \mathcal{B}$$

Propiedad 1: Se dice que las variables aleatorias X_1, \dots, X_n son independientes sii

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) * \dots * F_{X_n}(x_n)$$

Propiedad 2: Se dice que las variables aleatorias discretas X_1, \dots, X_n son independientes sii

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = p_{X_1}(x_1) * \dots * p_{X_n}(x_n)$$

Propiedad 3: Se dice que las variables aleatorias continuas X_1, \dots, X_n son independientes sii

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = f_{X_1}(x_1) * \dots * f_{X_n}(x_n)$$

vale para casi todo x_1, \dots, x_n

1.2. Momentos

1.2.1. Esperanza de una variable aleatoria

Es el promedio ponderado de los valores que puede tomar la variable. Podríamos pensarlo como el centro de masa de la variable.

Sea X una VAD con función de probabilidad $p_X(x)$, el valor esperado o media de X es:

$$\mathbf{E}(X) = \sum_{x \in R_X} x * p_X(x)$$

El valor de la esperanza de cualquier función $h(X)$ se calcula como:

$$\mathbf{E}(h(X)) = \sum_{x \in R_X} h(x) * p_X(x)$$

Sea X una variable aleatoria continua con función de densidad $f_X(x)$, el valor esperado o media de X es:

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

El valor de la esperanza de cualquier función $h(X)$ se calcula como:

$$\mathbf{E}(h(X)) = \int_{-\infty}^{\infty} h(x) f_X(x) dx$$

Propiedad: Si $h(x) = ax + b$, entonces $\mathbf{E}(h(X)) = a\mathbf{E}(X) + b$

Caso general

Sea X una variable aleatoria con función de distribución $F_X(x)$, Si $h(X)$ es una función cualquiera de la variable aleatoria X entonces si definimos A como el conjunto formado por los valores de X que concentran masa positiva (Conjunto de átomos):

$$\mathbf{E}[h(X)] = \sum_{x \in A} h(x) \mathbf{P}(X = x) + \int_{-\infty}^{\infty} h(x) F'_X(x) dx$$

Propiedad:

$$\mathbf{E}(X) = \int_0^{\infty} (1 - F_X(x)) dx + \int_{-\infty}^0 F_X(x) dx$$

1.2.2. Varianza de una variable aleatoria

Mide la dispersion media de los valores de una variable aleatoria. Sea X una V.A. y μ su valor esperado, la varianza de X se define como:

$$\mathbf{var}(X) = \mathbf{E}((X - \mathbf{E}(X))^2) = \mathbf{E}(X^2) - \mathbf{E}(X)^2$$

El desvío estándar se define como la raíz cuadrada de la varianza: $\sigma = \sqrt{\text{var}(X)}$. Se expresa en las mismas unidades que la variable aleatoria.

La **mediana** es el valor de X que acumula a izquierda (o derecha) una probabilidad igual a 0.5, o sea $F_X(x) = 0,5$.

La **moda** es el valor de la variable con valor de probabilidad máximo.

1.2.3. Vectores aleatorios

El valor esperado de una función $h(X, Y)$ está dado por:

$$\mathbf{E}(h(X, Y)) = \sum_x \sum_y h(x, y) p_{X,Y}(x, y)$$

Si (X, Y) es un vector aleatorio discreto.

$$\mathbf{E}(h(X, Y)) = \iint_{-\infty}^{\infty} h(x, y) f_{X,Y}(x, y) dx dy$$

Si (X, Y) es un vector aleatorio continuo.

Propiedades de orden:

Sea $X = (X_1, \dots, X_n)$, $g : \mathbb{R}^k \rightarrow \mathbb{R}$

1. Si $g(x) > 0$ entonces $\mathbf{E}(g(X)) > 0$
2. Sea $h(X) > g(X)$ entonces $\mathbf{E}(h(X)) > \mathbf{E}(g(X))$
3. Si $X > 0$ entonces $\mathbf{E}(X) > 0$
4. $\mathbf{E}(|X|) \geq \mathbf{E}(X)$
5. $|\mathbf{E}(X)| \leq \mathbf{E}(|X|)$
6. Sea $h(X)$ una función cóncava, $\mathbf{E}(h(X)) \leq h(\mathbf{E}(X))$
7. $\mathbf{E}(|XY|) \leq \sqrt{\mathbf{E}(X^2)\mathbf{E}(Y^2)}$
8. $\sqrt{\mathbf{E}(X+Y)^2} \geq \sqrt{\mathbf{E}(X^2)} + \sqrt{\mathbf{E}(Y^2)}$

Más propiedades importantes:

1. $\mathbf{E}[\sum_{i=1}^n a_i X_i] = \sum_{i=1}^n a_i \mathbf{E}[X_i]$

2. Si X_1, \dots, X_n son independientes entonces $\mathbf{E}[\prod_{i=1}^n X_i] = \prod_{i=1}^n \mathbf{E}[X_i]$
(Prueba para (X, Y))

La **covarianza** entre dos VA X e Y está dada por:

$$\mathbf{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))]$$

Propiedades

1. $\mathbf{cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}[X]\mathbf{E}[Y]$
2. Si X e Y son independientes entonces $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$, y por lo tanto $\mathbf{cov}(X, Y) = 0$
3. $\mathbf{cov}(a + bX, c + dY) = bd\mathbf{cov}(X, Y)$
4. $\mathbf{cov}(X + Y, Z) = \mathbf{cov}(X, Z) + \mathbf{cov}(Y, Z)$
5. Sean X e Y dos variables aleatorias, $\mathbf{var}(X + Y) = \mathbf{var}(X) + \mathbf{var}(Y) + 2\mathbf{cov}(X, Y)$

El **coeficiente de correlación** entre X e Y está dado por:

$$\rho_{XY} = \frac{\mathbf{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Propiedad: $|\rho_{XY}| = 1$ sii $\mathbf{P}(aX + b = Y) = 1$

Def: Sea \mathbf{X} una matriz de variables aleatorias, la esperanza de dicha matriz será la matriz formada por las esperanzas de cada una de las variables aleatorias. Esto es: Si

$$(X)_{i,j} = x_{i,j}, \quad \mathbf{E}(X)_{i,j} = E(x_{i,j})$$

Covarianza entre vectores aleatorios

Sean $x \in \mathbb{R}^p$ e $y \in \mathbb{R}^q$ vectores aleatorios, busco una idea de como está asociado cada componente del vector x con cada componente del vector y . Llamamos Σ_{XY} a la matriz de covarianzas de x e y , y se calcula como

$$\Sigma_{XY} = \mathbf{cov}(x, y) = \mathbf{E}(xy^T) - E(x)E(y)^T$$

Siguiendo la misma idea, la matriz de covarianzas de x será:

$$\Sigma_X = \mathbf{cov}(x, x) = \mathbf{E}(xx^T) - E(x)E(x)^T$$

Esta matriz tendrá como elementos de la diagonal a las varianzas de cada variable $X_i, i = 1, \dots, p$, y como elemento $(\Sigma_X)_{i,j} = \mathbf{cov}(X_i, X_j)$

La matriz de covarianzas de x será simétrica, y si x tiene densidad será definida positiva (Ver Seber, Multivariate Observations).

Propiedades

1. $\mathbf{E}(AXB + C) = A\mathbf{E}(X)B + C$
2. $\mathbf{cov}(Ax, By) = A\mathbf{cov}(x, y)B^T$
3. $\mathbf{cov}(Ax) = A\mathbf{cov}(x)A^T$

1.3. Predicción

Sea Y una V.A., $X = (X_1, X_2, \dots, X_n)$ un vector aleatorio, existirá alguna función $g(X)$ que nos sirva para predecir a Y . Para encontrar dicha función se calcula el error cuadrático medio:

$$ECM = \mathbf{E}[(Y - g(X))^2]$$

Y se busca la función g que minimiza el error.

EL mejor predictor lineal se llama **Recta de regresión**, y es la ecuación que resulta de minimizar $ECM = \mathbf{E}[(Y - (aX + b))^2]$.

Aplicando propiedades de la esperanza, derivo parcialmente en función de a y de b y despejo, resultando:

$$g(X) = \hat{Y} = \frac{\mathbf{cov}(X, Y)}{\mathbf{var}(X)}(X - \mathbf{E}[X]) + E[Y]$$

El mejor predictor será la esperanza condicional. Para entender un poco más, repasamos conceptos aprendidos en teoría de probabilidades.

1.3.1. Variables aleatorias condicionadas

Definición: Sean X e Y variables aleatorias discretas con $p_X(x) > 0$, la función de probabilidad condicional de Y dado que $X = x$ es

$$p_{Y|X=x}(y) = \mathbf{P}(Y = y|X = x) = \frac{\mathbf{P}(Y = y, X = x)}{\mathbf{P}(X = x)} = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

Se define como 0 a $p_{Y|X=x}(y)$ cuando $p_X(x) = 0$.

(Para cada valor x que tome la variable aleatoria X , tendremos una variable aleatoria $Y|X = x$ diferente, con su correspondiente función de probabilidad)

Propiedad: Sean X e Y vectores aleatorios discretos tal que $p_{Y|X=x}(y) = p_Y(y)$ para todo $x \in \mathbb{R}$, entonces X e Y son independientes.

Definición: Sea (X, Y) un vector aleatorio con densidad conjunta $f_{X,Y}(x, y)$ y densidad marginal $f_X(x)$, entonces para cualquier valor de X con el cual $f_X(x) > 0$, la función de densidad condicional de Y dado $X = x$ es

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

Sean X e Y variables aleatorias discretas, entonces

$$\mathbf{E}[Y|X = x] = \sum_{y \in R_Y} y p_{Y|X=x}(y), \forall x \in R_X$$

Sean X e Y variables aleatorias continuas, entonces

$$\mathbf{E}[Y|X = x] = \int_{y \in R_Y} y f_{Y|X=x}(y) dy, \forall x \in R_X$$

A estas funciones, que son funciones de x , se las llama **funciones de regresión**, y se las denota $\varphi(x)$

Entonces, llamemos $\varphi(x) = \mathbf{E}[Y|X = x]$, luego $\varphi : \text{sup}_X \rightarrow \mathbb{R}$. La variable aleatoria llamada **Esperanza condicional de Y dado X** , denotado por $\mathbf{E}[Y|X]$, se define por $\varphi(X) = \mathbf{E}[Y|X]$.

Definición: La variable aleatoria esperanza condicional de Y dada X se define como $\varphi(X) = \mathbf{E}[Y|X]$ con φ una función medible tal que $\mathbf{E}((Y - \varphi(X)) * t(X)) = 0$ para toda función t medible $t : R_X \rightarrow \mathbb{R}$ tal que $Y * t(X)$ tiene esperanza finita (t es cualquier función en el plano, $\varphi(X)$ será el mejor predictor lineal de Y basado en X).

Obs: La esperanza condicional siempre existe y a demás es única con probabilidad 1

Propiedad: $\mathbf{E}[Y] = \mathbf{E}[\mathbf{E}[Y|X]]$

Demostración:(para variables aleatorias continuas)

$$\mathbf{E}[\varphi(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Pero

$$\varphi(x) = \mathbf{E}[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy$$

Entonces

$$\mathbf{E}[\mathbf{E}[Y|X]] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X=x}(y) f_X(x) dx dy = \mathbf{E}[Y]$$

Mas propiedades:

1. X e Y vectores aleatorios, s y r funciones medibles tales que las variables aleatorias $r(X) * s(Y)$, $r(X)$ y $s(Y)$ tienen esperanza finita, entonces $\mathbf{E}(r(X) * s(Y)|X) = r(X) * \mathbf{E}(s(Y)|X)$
2. Y_1, Y_2 , V.A. con esperanza finita, X vector aleatorio, $\mathbf{E}(aY_1 + bY_2|X) = a\mathbf{E}(Y_1|X) + b\mathbf{E}(Y_2|X)$
3. $\mathbf{E}(Y|X) = \mathbf{E}(Y)$ si X e Y son independientes
4. $\mathbf{E}(r(X)|X) = r(X)$

Varianza condicional

$$\begin{aligned}\text{var}(Y|X=x) &= \mathbf{E}((Y - \mathbf{E}(Y|X=x))^2|X=x) \\ \text{var}(Y|X) &= \mathbf{E}((Y - \mathbf{E}(Y|X))^2|X) \\ \text{var}(Y|X) &= \mathbf{E}(Y^2|X) - \mathbf{E}(Y|X)^2\end{aligned}$$

Propiedad (Pitágoras)

$$\text{var}(Y) = \mathbf{E}(\text{var}(Y|X)) + \text{var}(\mathbf{E}(Y|X))$$

Demostración:

$$\begin{aligned}\text{var}(Y|X) &= \mathbf{E}(Y^2|X) - \mathbf{E}(Y|X)^2 \\ \mathbf{E}[\text{var}(Y|X)] &= \mathbf{E}[\mathbf{E}(Y^2|X)] - \mathbf{E}[\mathbf{E}(Y|X)^2] = \mathbf{E}(Y^2) - \mathbf{E}[\mathbf{E}(Y|X)^2] \\ \text{var}(\mathbf{E}(Y|X)) &= \mathbf{E}[\mathbf{E}(Y|X)^2] - \mathbf{E}[\mathbf{E}(Y|X)]^2 = \mathbf{E}[\mathbf{E}(Y|X)^2] - \mathbf{E}[Y]^2 \\ \text{var}(Y) &= \mathbf{E}(\text{var}(Y|X)) + \text{var}(\mathbf{E}(Y|X))\end{aligned}$$

Predicción: “La esperanza condicional de Y dado X es la función de la variable aleatoria X que mejor predice o se aproxima a Y ”

Recordamos la definición :

Sea Y una V.A., $X = (X_1, X_2, \dots, X_n)$ un vector aleatorio, existirá alguna función $g(X)$ que nos sirva para predecir a Y . Para encontrar dicha función se calcula el error cuadrático medio:

$$ECM = \mathbf{E}[(Y - g(X))^2]$$

Y se busca la función g que minimiza el error.

La esperanza condicional de Y dada X cumple que:

$$\mathbf{E}[(Y - g(X))^2] \geq \mathbf{E}[(Y - \mathbf{E}(Y|X))^2]$$

Para cualquier función $g(X)$

Capítulo 2

Modelo lineal

Supongamos que queremos predecir (o estimar) la media de una variable Y . Sabemos que la media muestral es el mejor estimador que podemos dar para la media poblacional, pero ¿qué ocurre si tenemos información adicional?. Supongamos por ejemplo que queremos estimar la distancia de frenado de un automóvil. Podríamos estimar su valor medio, pero ese valor no será tan representativo de la distancia de frenado para todas las situaciones posibles. Si tenemos más información, por ejemplo la velocidad en la que se encuentra el vehículo, podríamos tener una estimación mas precisa en función de dicha velocidad.

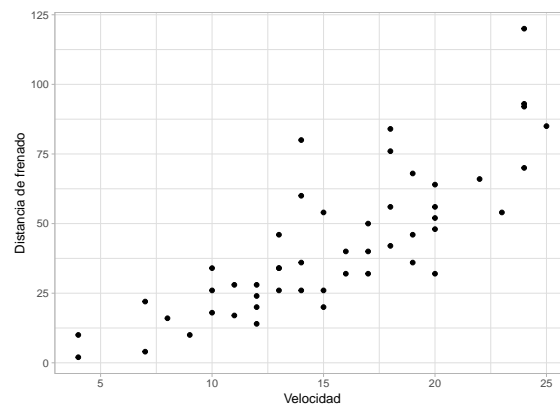


Figura 2.1: Gráfico de dispersión (Scatterplot)

En la Figura 2.1 podemos observar que la distancia de frenado aumenta a medida que aumenta la velocidad. Si tomáramos intervalos de velocidades, y para cada uno de esos intervalos calculamos la distancia media de frenado, eso nos daría mejores estimaciones de la distancia media de frenado para esas velocidades en particular. Esta idea es un primer acercamiento al concepto de esperanza condicional.

En Modelo lineal, nos interesa establecer una relación entre una variable dependiente Y (variable que queremos predecir) y otras p variables X_1, \dots, X_p , que las llamaremos variables predictoras o covariables. Buscamos un modelo que exprese a la variable dependiente en términos de las covariables (cuando hablamos de modelo nos referimos a una expresión matemática que describa en algún sentido el comportamiento de las variables).

El modelo pretende describir como el comportamiento de $\mathbf{E}(Y)$ varía bajo condiciones cambiantes de otras variables. En principio vamos a suponer que $\mathbf{var}(Y)$ no es afectada por estas condiciones, es decir toma un valor constante al que llamaremos σ^2 .

Llamemos $\underline{X} = (X_1, \dots, X_p)^T$ a las covariables. Una forma general de expresar el modelo es en función de \underline{x} , como

$$\mathbf{E}(Y|\underline{X} = \underline{x}) = g(\underline{x})$$

o considerando las covariables fijas, como

$$Y = g(\underline{x}) + \epsilon$$

donde ϵ corresponde a un error aleatorio con $\mathbf{E}(\epsilon) = 0$. Estos modelos se llaman **modelos de regresión**. Hay muchas funciones posibles g , buscamos acotarlas. Una forma es expresarla en función de un número finito de constantes desconocidas a estimar, que llamaremos parámetros, y en este caso diremos que nos encontramos frente a un modelo de **regresión paramétrica**. Algunos ejemplos pueden ser

- $Y = \theta_1 + \theta_2 x_2 + \theta_3 + \epsilon$
- $Y = \theta_1 e^{\theta_2 x_2} + \epsilon$
- $Y = \theta_1 x_2^{\theta_2} + \epsilon$

Si la función g no puede expresarse como una función de una cantidad finita de parámetros, entonces estamos frente a un modelo de regresión no paramétrica. En este caso se imponen algunas condiciones con respecto a la función g como por ejemplo ser una función continua, o continua y derivable, o monótona creciente, entre otras.

En este capítulo nos focalizaremos en los modelos paramétricos. El modelo paramétrico más sencillo será el modelo lineal, en este caso $g(\underline{x})$ será una función lineal en los parámetros. Esto es

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1} + \epsilon$$

donde Y es la variable dependiente o respuesta, β_i los parámetros a estimar, β_0 la ordenada al origen o *intercept*, y ϵ el error aleatorio que contempla todas las variables que no

estoy teniendo en cuenta para describir a Y .

Una vez establecido el modelo, nos interesa

1. Estimar los parámetros desconocidos β y σ^2 a partir de observaciones
2. Hacer inferencia sobre los parámetros (test de hipótesis, intervalos de confianza, etc)
3. Evaluar si se cumplen los supuestos
4. Predicción
5. Identificar datos atípicos
6. Selección de modelos óptimos.

2.1. Supuestos del modelo lineal

Tenemos como modelo

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \epsilon$$

Si tomamos n observaciones (\underline{x}_i, y_i) tendremos que

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i(p-1)} + \epsilon_i, \quad i = 1, \dots, n$$

donde los valores de ϵ_i no son observables.

Si volvemos al ejemplo de los autos, en la Figura 2.1 observamos que no todos los puntos caen sobre una recta. La dispersión de los puntos al rededor de cualquier línea para un valor de x fijo representa la variación de la distancia de frenado que no está asociada con la velocidad, y se considera de naturaleza aleatoria (ϵ_i). Se espera que todos estos componentes diversos tengan un aporte muy menor a la explicación de la variable Y comparado con el de la variable explicativa considerada.

Los supuestos del modelo se pueden ver de dos formas: Enfocados en el error aleatorio o en la variable aleatoria Y . Se detallan ambas a continuación.

Los supuestos sobre el modelo

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \epsilon$$

son:

1. Los errores ϵ_i tienen media cero. Esto es $\mathbf{E}(\epsilon_i) = 0$

2. Los errores ϵ_i tienen todos la misma varianza, $\mathbf{var}(\epsilon_i) = \sigma^2$ (supuesto de homocedasticidad)
3. Los errores ϵ_i tienen distribución Normal Los errores ϵ_i son independientes entre si y no están correlacionados con las covariables X_i .

Resumiendo, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ independientes pasa $i = 1, \dots, n$.

La otra forma de verlo es que para cada valor fijo de la variable X , la esperanza de Y depende de X de forma lineal, esto es

$$\mathbf{E}(Y|X = x) = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$$

donde $\beta_0, \dots, \beta_{p-1}$ son los parámetros del modelo. Buscaremos estimar dichos parámetros para poder luego, observado un valor x de X estimar la esperanza de Y . Los supuestos entonces serán

1. $\mathbf{E}(Y|X = x) = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$
2. $\mathbf{var}(Y|X = x) = \sigma^2$
3. $Y|X = x$ tiene distribución Normal
4. Las variables Y_1, \dots, Y_n son independientes entre sí.

Resumiendo, $Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}, \sigma^2)$.

De aquí en adelante trabajaremos para el modelo con X fijos, es decir que no consideraremos a X como una variable aleatoria, por lo que siempre que se hable de la esperanza de Y se está haciendo referencia a la esperanza de Y para un valor fijo de $X = x$.

2.2. Estimación de los parámetros

Enfoque matricial. Si tomamos n observaciones (\underline{x}_i, y_i) tendremos que

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i(p-1)} + \epsilon_i, \quad i = 1, \dots, n$$

Podemos escribirlo de forma matricial considerando:

$$Y = (y_1, \dots, y_n)^T, \quad \epsilon = (\epsilon_1, \dots, \epsilon_n)^T, \quad \beta = (\beta_1, \dots, \beta_n)^T,$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1(p-1)} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ x_{n1} & \dots & x_{n(p-1)} \end{bmatrix}$$

Donde a $\mathbf{X} \in \mathbb{R}^{n \times p}$ se la llama matriz de diseño (model matrix), $\mathbf{E}(\epsilon) = 0$ y $\Sigma_\epsilon = \sigma^2 I$. Reescribimos el modelo obteniendo

$$Y = \mathbf{X}\beta + \epsilon$$

Para estimar los parámetros β usamos el método de mínimos cuadrados. Este método consiste en buscar la recta que está "lo más cerca posible" de todos los puntos.

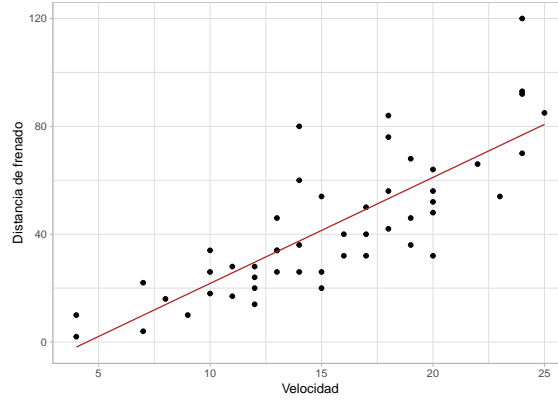


Figura 2.2: Estimación por cuadrados mínimos

Mirando el gráfico, podemos apreciar que para cada punto, la recta roja se encuentra a cierta distancia vertical. A esa distancia vertical la vamos a llamar residuo (r_i), y al valor de y para cada valor de x sobre la recta, lo vamos a llamar valor predicho o ajustado (\hat{y}_i), de manera que para cada valor de $i = 1, \dots, n$, $r_i = y_i - \hat{y}_i$. También podemos escribir la ecuación de la recta como $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$, donde los $\hat{\beta}_i$ son los estimadores de mínimos cuadrados de los parámetros β_i .

Observación: ϵ_i es la diferencia real entre la variable y y su esperanza, mientras que el residuo r_i es la diferencia entre y y la estimación que realizamos para su esperanza (que la llamamos \hat{y}).

El estimador de mínimos cuadrados minimiza la suma de los cuadrados de los residuos. Es decir, busca minimizar las distancias de los puntos a la recta al cuadrado. Si definimos

$$S(b) = \sum_{i=1}^n (y_i - (b_0 + b_1 x_{i1} + \dots + b_{p-1} x_{i(p-1)}))^2 = \|Y - \mathbf{X}b\|^2$$

(por que $\|u\|^2 = u^T u = \sum u_i^2$)

el método de mínimos cuadrados busca el valor de b que minimiza $S(b)$. La solución siempre existe pero no siempre es única. Derivando e igualando a cero $S(b)$ obtenemos las ecuaciones normales, dadas por

$$\frac{dS(b)}{db_i} = 0, \quad i = 1, \dots, n$$

Derivando y despejando (ver Seber, Linear analysis), en forma matricial se obtiene

$$\mathbf{X}^T \mathbf{X} b = \mathbf{X}^T Y$$

Si $\mathbf{X}^T \mathbf{X}$ es no singular, la solución es única, resultando

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y$$

Ejercicio: Encontrar los estimadores de mínimos cuadrados para el caso de Regresión lineal Simple.

2.2.1. Interpretación geométrica

Escribimos al modelo como

$$\Omega : \quad E(Y) = \mathbf{X}\beta, \quad \Sigma_Y = \sigma^2 I$$

Para simplificar notación, podemos llamar $\eta = E(Y)$, por lo que $\hat{\eta} = \hat{Y}$ (recordemos que todo es referido al diseño fijo, es decir que la esperanza de Y es condicionada a que los valores de X son fijos).

Si llamamos x^i a la i -ésima columna de la matriz \mathbf{X} , entonces podemos reescribir a η como

$$\eta = \beta_1 x^1 + \dots + \beta_n x^n$$

es decir como una combinación lineal de las columnas de $\beta_1 x^1$. Esto significa que η pertenece al subespacio generado por las columnas de $\beta_1 x^1$, que llamaremos V_r asumiendo que $rg(\mathbf{X}) = r \leq p$. Entonces

$$\min_b S(b) = \min_b \|Y - \mathbf{X}b\|^2 = \min_{Z \in V_r} \|Y - Z\|^2$$

Es decir, estoy buscando de todos los $Z \in V_r$ el que esté más cerca de Y , que es la proyección ortogonal al subespacio generado por las columnas de \mathbf{X} . A dicha proyección la llamamos $\hat{\eta}$.

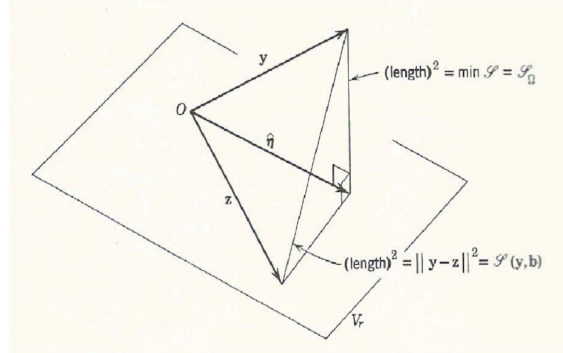


Figura 2.3: Interpretación geométrica

Esta proyección siempre existe y es única (aunque no así los b). Entonces $\mathbf{X}^T \mathbf{X} \hat{\eta} = \mathbf{X}^T Y$. Si $rg(\mathbf{X}) = p$ entonces $rg(\mathbf{X}^T \mathbf{X}) = p$, y existe la inversa de $\mathbf{X}^T \mathbf{X}$, resultando

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y$$

como ya lo habíamos visto. Entonces

$$\mathbf{X} \hat{\beta} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y = \mathbf{P} Y = \hat{Y}$$

Donde \mathbf{P} Es la matriz de proyección en V_r .

El residuo será

$$r = Y - \hat{Y} = Y - \mathbf{P} Y = (I - \mathbf{P}) Y$$

por lo tanto r , como se observa en la Figura 2.3, es ortogonal a V_r .

Propiedades

1. $rg(P) = n - p$
2. P e $(I - P)$ son matrices de proyección (simétricas e idempotentes)
3. $(I - P)X = 0$

Llamamos Suma de cuadrados de los residuos a $SCR = \|Y - \hat{Y}\|^2$. Por pitágoras tenemos que

$$\|Y - \hat{Y}\|^2 = \|Y\|^2 - \|\hat{Y}\|^2$$

2.2.2. Propiedades del estimador de mínimos cuadrados

Siguiendo el modelo

$$\Omega : E(Y) = \mathbf{X}\beta, \quad \Sigma_Y = \sigma^2 I$$

1. $\hat{\beta}$ es un estimador insesgado para β

Demostración

$$\begin{aligned}
 \mathbf{E}(\hat{\beta}) &= \mathbf{E}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y) \\
 &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{E}(Y) \\
 &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta \\
 &= \beta
 \end{aligned}$$

2. $\Sigma_{\hat{\beta}} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$

Demostración

$$\begin{aligned}
 \Sigma_{\hat{\beta}} &= \text{cov}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y) \\
 &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Sigma_Y ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)^T \\
 &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\
 &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}
 \end{aligned}$$

3. $\mathbf{E}(\hat{Y}) = \mathbf{X}\beta$
4. $\Sigma_{\hat{Y}} = \sigma^2 \mathbf{P}$
5. $\mathbf{E}(r) = 0$, $\Sigma_r = \sigma^2 (I - P)$

Proposición Dados $i \geq 1, j \leq n$ tenemos que

- $0 \leq p_{ii} \leq 1$
- $\frac{-1}{2} \leq p_{ij} \leq \frac{1}{2}$

Entonces, $\text{var}(\hat{Y}_i) = \sigma^2 p_{ii} \leq \text{var}(Y_i) = \sigma^2$

2.3. Estimación de σ^2

Las varianzas de los estimadores dependen del diseño y de σ^2 , que es desconocida. Vemos que los residuos r son la diferencia entre Y y el estimador de su esperanza, entonces tendría sentido estimar a σ^2 mediante el promedio de los cuadrados de los residuos. Bajo Ω , tendremos que

$$S^2 = \frac{\|Y - \hat{Y}\|^2}{n - p}$$

es un estimador insesgado para σ^2 .

2.4. Algunos teoremas útiles

Teorema: La función paramétrica $\Psi = c^T \beta$ es estimable si y solo si c es una combinación lineal de las filas de la matriz de diseño \mathbf{X} , o sea si existe $a \in \mathbb{R}^n$ tal que $c^T = a^T \mathbf{X}$.

Teorema de Gauss Markov: Supongamos que vale el modelo $\Omega : \mathbf{E}(Y) = \mathbf{X}\beta, \Sigma_Y = \sigma^2 I$. Toda función estimable $\Psi = c^T \beta$ tiene un único estimador lineal insesgado de mínima varianza $\hat{\Psi}$. Este estimador puede obtenerse reemplazando β por el estimador de mínimos cuadrados $\hat{\beta}$.

2.5. Test y regiones de confianza

Ahora debemos agregar al modelo un supuesto adicional: Normalidad conjunta de los errores. Entonces, dadas X fijas tenemos que:

$$\Omega : Y \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 I), \quad rg(\mathbf{X}) = r, \quad \beta \in \mathbb{R}^p$$

A partir de este modelo podremos deducir test de hipótesis, intervalos de confianza de nivel exacto para funciones paramétricas estimables, conjuntos o regiones de confianza simultáneos.

Sabemos que $\hat{\beta}$ y S^2 son funciones de estadísticos suficientes y completos, por lo tanto son IMVU.

Vamos a querer plantear hipótesis del tipo $C\beta = \delta$. Tendremos funciones estimables $\Psi_i = c_i^T \beta$, entonces usaremos como estimador a $\hat{\Psi} = C\hat{\beta}$. Observando su forma matricial, $\Psi = C\beta$ con $C \in \mathbb{R}^{q \times p}$, con c_1, \dots, c_q linealmente independientes (es una propiedad necesaria para que Ψ sea estimable. Para más detalle ver Seber, Linear Analysis). Bajo estas condiciones, se tiene que $rg(C) = q$. En el caso de rango completo de la matriz de diseño, tendremos que $r = p$.

Teorema: Supongamos que se cumple el teorema

$$\Omega : Y \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 I), \quad rg(\mathbf{X}) = r, \quad \beta \in \mathbb{R}^p$$

Entonces:

1. $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(\mathbf{X}^T \mathbf{X})^{-1})$
2. $\frac{(\hat{\beta} - \beta)^T (\mathbf{X}^T \mathbf{X}) (\hat{\beta} - \beta)}{\sigma^2} \sim \chi_p^2$
3. $\hat{\beta}$ y $(n - p)S^2/\sigma^2$ son independientes

$$4. (n-p)S^2/\sigma^2 \sim \chi_{n-p}^2$$

Si tenemos que $\hat{\Psi} = C\hat{\beta}$, con $rg(C) = q$, entonces usando las propiedades de normal multivariada vistas en el capítulo anterior tenemos que

$$\hat{\Psi} \sim \mathcal{N}_q(C\beta, \sigma^2 C(\mathbf{X}^T \mathbf{X})C^T)$$

2.5.1. Test de hipótesis

En el anexo 1 se encuentran notas para repasar el tema de test de hipótesis visto en cursos de estadística.

Si buscamos realizar un test de hipótesis para un solo parámetro, por ejemplo queremos saber si la variable X_1 es significativa al momento de explicar la variabilidad de nuestra variable Y , podríamos plantear las siguientes hipótesis:

$$H_0 : \beta_1 = 0 \quad Vs. \quad H_1 : \beta_1 \neq 0$$

Entonces, un test de hipótesis con nivel de significación α será

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } |T| > k_\alpha \\ 0 & \text{en otro caso} \end{cases}$$

Con $T = \frac{\hat{\beta}}{S\sqrt{d_{ii}}}$, que bajo H_0 tiene distribución t de student con $n-p$ grados de libertad. (Llamamos d_{ii} al elemento diagonal de la matriz $D = (\mathbf{X}^T \mathbf{X})^{-1}$).

¿Porque elijo esa variable T ? Dada la información que tenemos, podemos decir que $\frac{c^T \hat{\beta} - c^T \beta}{\sqrt{\sigma^2 c^T (\mathbf{X}^T \mathbf{X})^{-1} c}} \sim \mathcal{N}(0, 1)$ y $(n-p)S^2/\sigma^2 \sim \chi_{n-p}^2$, entonces

$$T = \frac{c^T \hat{\beta} - c^T \beta}{\sqrt{S^2 c^T (\mathbf{X}^T \mathbf{X})^{-1} c}} \sim t_{n-p}$$

Bajo H_0 , reemplazando resulta $T = \frac{\hat{\beta}}{S\sqrt{d_{11}}}$

Si llamamos T_{obs} al valor del estadístico T definido calculado en base a las observaciones $(x_i, y_i), i = 1, \dots, n$, se puede calcular el p -valor de la siguiente forma

$$p - \text{valor} = 2\mathbf{P}(T \geq |T_{obs}|), T \sim t_{n-p}$$

ya que se trata de un test a dos colas. Reportar el p-valor cuando uno realiza un test sobre un conjunto de datos siempre permite al lector elegir su punto de corte respecto de rechazar o no una hipótesis.

Caso general

Un test exacto para

$$H_0 : C\beta = \delta \quad Vs. \quad H_1 : C\beta \neq \delta$$

será

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } F > \mathcal{F}_{q,n-p,1-\alpha} \\ 0 & \text{en otro caso} \end{cases}$$

ya que bajo H_0

$$F = \frac{(C\hat{\beta} - \delta)^T (C(\mathbf{X}^T \mathbf{X})^{-1} C^T)^{-1} (C\hat{\beta} - \delta)}{qS^2} \sim \mathcal{F}_{q,n-p}$$

En este caso, el p-valor se calcula como

$$p - \text{valor} = \mathbf{P}(F \geq F_{obs}), \quad F \sim \mathcal{F}_{q,n-p}$$

Para comparar con el caso particular, recordar que $\mathcal{F}_{1,n} = (t_n)^2$

El estadístico F se deduce a partir del test de cociente de verosimilitud mencionado en el *Anexo 1*, donde

$$f(y) = \frac{1}{(2\pi)^{n/2}} \frac{1}{(\sigma^2)^{n/2}} \exp\left\{\frac{-1}{2\sigma^2} \|y - X\beta\|^2\right\}$$

Significación de la regresión

Supongamos un modelo con intercept, vamos a querer saber si la regresión es significativa, esto es si las variables que proponemos como predictoras realmente pueden tener una relación lineal con y bajo el modelo Ω . Entonces, las hipótesis a plantear serán:

$$H_0 : \beta_1 = \dots = \beta_{p-1} = 0 \quad Vs. \quad H_1 : \text{Algun } \beta_i \neq 0, \quad i = 1, \dots, p-1$$

Si lo vemos como un caso particular del caso general presentado anteriormente, podemos ver que

$$\Psi = (\beta_1, \dots, \beta_{p-1})^T, \quad \delta = (0, \dots, 0)^T,$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & \dots & 0 \\ 0 & & & \\ \vdots & & & \\ 0 & \dots & & 1 \end{bmatrix}$$

En este caso, $q = p - 1$, y el test resulta

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } F > \mathcal{F}_{p-1, n-p, 1-\alpha} \\ 0 & \text{en otro caso} \end{cases}$$

Ejemplo: Boston Housing

2.5.2. Intervalos de confianza

Capítulo 3

Anexo 1

Test de hipótesis.

Una hipótesis es una suposición o conjetura sobre la naturaleza, cuyo valor de verdad o falsedad no se conoce. Una hipótesis estadística es una hipótesis sobre una o más poblaciones estadísticas o sobre un fenómeno aleatorio.

Se llama hipótesis estadística a cualquier suposición o enunciado provisorio que se realiza sobre una población o fenómeno aleatorio, cuyo valor de verdad o falsedad no se conoce.

La hipótesis estadística se refiere a la población. Sin embargo el análisis para tratar de comprobar la verdad o falsedad de la hipótesis se realizará sobre una muestra.

Se llama **ensayo de hipótesis** al procedimiento que se sigue para tratar de averiguar si una hipótesis es verdadera o falsa.

En un ensayo de hipótesis se plantean dos hipótesis:

- H_1 : El investigador suele tener una hipótesis de investigación que es la predicción que se deriva de la teoría que está analizando o que sustenta su trabajo. Esta hipótesis de investigación, cuando se formula con una proposición operacional, se llama hipótesis alternativa
- H_0 : Se llama hipótesis nula a la hipótesis objeto del ensayo. Esta se formula con el propósito expreso de ser rechazada. Un ensayo puede rechazar o no la hipótesis nula. Cuando el ensayo no rechaza, diremos que no hay pruebas concluyentes en contra de la hipótesis, pero no quedará probado que la hipótesis sea verdadera. En cambio si se rechaza la hipótesis nula quedará probado que es falsa, y se concluirá que la hipótesis alternativa es correcta. (La conclusión fuerte en un ensayo de hipótesis únicamente se tiene en el rechazo de la hipótesis nula)

Ejemplo: Supongamos que nos encontramos en un juicio, donde una persona será sometido a juicio para decidir si enviarla o no a la cárcel por un crimen. Al comienzo se cuenta con dos hipótesis: la persona es culpable o es inocente. Para tomar una decisión se debe encontrar evidencia suficiente que respalde dicha decisión. La presunción de inocencia que goza el acusado es la hipótesis nula que ensaya el juez. Si el juez encuentra evidencia suficiente de que la hipótesis nula es falsa, entonces rechaza dicha hipótesis y decide enviarlo a la cárcel. Ahora, al momento de tomar una decisión podría llegar a cometer distintos errores: enviar a la cárcel a una persona inocente o dejar libre a una culpable.

Existen dos tipos de errores que pueden cometerse al tomar una decisión:

- Error de tipo I: Es el error que se comete cuando se rechaza una hipótesis nula que en realidad era verdadera. Se trata de que este suceso tenga muy baja probabilidad.
- Error de tipo II: Es el error que se comete cuando no se rechaza una hipótesis nula que en realidad es falsa.

Ahora bien, el Juez también podría tomar una decisión correcta.

La **potencia de un test** es la probabilidad de rechazar una hipótesis nula en función del parámetro desconocido sobre el cual se plantea la hipótesis.

Un **test de hipótesis** es una regla de decisión entre H_0 y H_1 y la expresamos como una función de la muestra aleatoria $\delta(\underline{X})$ que puede tomar los valores 0 o 1. Si $\delta(\underline{X}) = 1$ se rechaza H_0 , y en caso contrario no se la rechaza.

Definición: Sea $\underline{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una población con distribución $F_\theta(x)$, $\theta \in \Theta$. Sean Θ_1 y Θ_2 tal que $\Theta_1 \cup \Theta_2 = \Theta$ y $\Theta_1 \cap \Theta_2 = \emptyset$. Un test para este problema es una regla de decisión basada en \underline{X} para decidir entre 2 hipótesis:

$$H_0 : \theta \in \Theta_1 \quad Vs. \quad H_1 : \theta \in \Theta_2$$

Entonces, de acuerdo a lo definido antes,

$$\mathbf{P}(\text{error tipo I}) = \mathbf{P}_\theta(\text{Rechazar } H_0), \text{ con } \theta \in \Theta_1,$$

$$\mathbf{P}(\text{error tipo II}) = \mathbf{P}_\theta(\text{No rechazar } H_0), \text{ con } \theta \in \Theta_2.$$

La potencia del test es una función de θ :

$$\pi_\delta(\theta) = \mathbf{P}_\theta(\text{Rechazar } H_0) = \mathbf{P}_\theta(\delta(\underline{X}) = 1) = E_\theta(\delta(\underline{X}))$$

Se puede observar que:

$$\mathbf{P}(\text{error tipo I}) = \pi_{\delta}(\theta), \text{ con } \theta \in \Theta_1,$$

$$\mathbf{P}(\text{error tipo II}) = 1 - \pi_{\delta}(\theta), \text{ con } \theta \in \Theta_2.$$

Un buen test deberá tener errores de tipo 1 y 2 pequeños, y por lo tanto deberá tener una función de potencia $\pi_{\delta}(\theta)$ que tome valores cercanos a 0 para $\theta \in \Theta_1$ y valores cercanos a 1 $\theta \in \Theta_2$.

Se llama **Nivel de significación del test** a la máxima probabilidad de cometer un error de tipo I, es decir:

$$\alpha = \sup_{\theta \in \Theta_1} \pi_{\delta}(\theta)$$

Se llama **p-valor** de un test al menor nivel de significación para el cual se rechaza H_0 , para una observación dada. Es la probabilidad de encontrar un valor tan o mas extremo que el que se encontró con la muestra observada.

$$\text{Llamaremos } \beta(\theta) = \mathbf{P}_{\theta}(\text{Error II}) = 1 - \pi_{\delta}(\theta), \text{ con } \theta \in \Theta_2.$$

Ejemplo: La producción diaria de cierta empresa es aleatoria con distribución $\mathcal{N}(10, 1,5^2)$. Se propone comprar una maquina que promete una distribución $\mathcal{N}(11, 1,5^2)$. Plantear las hipótesis correspondientes al problema y un test que permita decidir.

¿Cómo encuentro mi regla de decisión (test)?

Por ahora, no tenemos ningún criterio para elegir entre dos tests, ni entre los muchos otros que podrían definirse. Entonces atacaremos el problema de definir criterios para comparar diferentes tests, y el de elegir un test óptimo.

Idea: Necesito decidir si esta nueva máquina es igual a la que tengo o realmente produce con una media mejor (igual a 11). Un criterio de decisión puede ser: Tomo una muestra de tamaño 16 producidos con la nueva máquina. Si lo que observo tiene mayor probabilidad de ocurrir cuando la media es 10 entonces no compraría la maquina (no rechazo $H_0 : \mu = 10$) pero si tiene mayor probabilidad de ocurrir cuando la media es 11 entonces debe ser que la maquina nueva es mejor, y eso debería llevarme a decidir comprarla (rechazar H_0).Entonces rechazaría H_0 cuando

$$L = \frac{f_{\mu=11}(X)}{f_{\mu=10}(X)} > k_{\alpha}$$

donde k_α es una constante que depende del nivel de significación. Fijado el valor de α , el test será:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } \frac{f_{\mu=11}(X)}{f_{\mu=10}(X)} > k_\alpha \\ 0 & \text{en otro caso} \end{cases}$$

De manera que cumpla:

$$\mathbf{E}_{\mu=10}[\delta(\underline{X})] = \alpha \quad (3.1)$$

3.1. Test para hipótesis simple vs. hipótesis simple

El caso más simple de problema de test de hipótesis es la situación donde Θ_1 y Θ_2 contienen cada uno un elemento. En este caso, se dice que las hipótesis son simples,

$$H_0 : \theta = \theta_1 \quad Vs. \quad H_1 : \theta = \theta_2$$

Por lo tanto, parece razonable plantear la regla de decisión:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } \frac{f_{\theta_2}(X)}{f_{\theta_1}(X)} > k_\alpha \\ 0 & \text{en otro caso} \end{cases} \quad (3.2)$$

Que para un nivel α dado, debemos hallar el valor de k_α que cumpla:

$$\mathbf{E}_{\theta_1}[\delta(\underline{X})] = \alpha = \mathbf{P}_{\theta_1}(\delta(\underline{X}) = 1) \quad (3.3)$$

Un test de la forma (2.2) se llama *test del cociente de verosimilitud*.

(Para el lector más curioso, el teorema de Neyman-Pearson dice que el test (2.2) que cumple con (2.3) es el uniformemente más potente para hipótesis simples. Ver bibliografía recomendada.)

Volviendo al ejemplo. Supongamos ahora que la muestra aleatoria proviene de una población con distribución normal, es decir que $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ independientes,

$$H_0 : \mu = \mu_1 \quad Vs. \quad H_1 : \mu = \mu_2$$

Hacemos el cociente de verosimilitud y despejamos para conseguir la regla de decisión.

Sabemos que

$$f_{\mu, \sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad \mu \in \mathbb{R}, \sigma > 0$$

resultando para la muestra aleatoria

$$f_{\mu, \sigma^2}(\underline{x}) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$$

Según el ejemplo, las dos máquinas provienen de poblaciones normales con diferente media pero igual varianza, entonces $\delta(\underline{X}) = 1$ si

$$L = \frac{f_{\mu=\mu_2}(\underline{X})}{f_{\mu=\mu_1}(\underline{X})} > k_\alpha$$

$$L = \frac{e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_2)^2}}{e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_1)^2}} > k_\alpha$$

Si $\mu_2 > \mu_1$, despejando resulta

$$\sum_{i=1}^n X_i \geq \frac{2\sigma^2 \log(k_\alpha) + n\mu_2^2 - n\mu_1^2}{2(\mu_2 - \mu_1)} = k'_\alpha$$

(Lo llamo k'_α ya que también es una constante que depende de α)

La regla de decisión entonces es

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } \sum_{i=1}^n X_i > k'_\alpha \\ 0 & \text{si no} \end{cases}$$

Para hallar el valor de k' uso la segunda condición:

$$\alpha = \mathbf{P}_{\mu_1}(\delta(\underline{X}) = 1) = \mathbf{P}_{\mu_1} \left(\sum_{i=1}^n X_i > k' \right) = \mathbf{P}_{\mu_1} \left(\frac{\sum_{i=1}^n X_i - n\mu_1}{\sqrt{n\sigma^2}} > z_{1-\alpha} \right)$$

Resultando el test que cumple con las condiciones pedidas:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } \frac{\sum_{i=1}^n X_i - n\mu_1}{\sqrt{n\sigma^2}} > z_{1-\alpha} \\ 0 & \text{si no} \end{cases}$$

De esta manera observamos que la regla de decisión depende de α , n , y el valor que toma el parámetro bajo H_0 (es decir, cuando H_0 es verdadera). Se observa también que la variable aleatoria de la cual depende el test es el estadístico suficiente para μ . La regla de decisión no depende del valor que pusimos en H_1 , lo único que influye en su definición es que el valor que elegimos en H_1 es mayor que el que pusimos en H_0 (si hubiese sido menor, al despejar la desigualdad se hubiese invertido). Por lo tanto la regla de decisión

hubiese sido la misma para una alternativa $H_1 : \mu > \mu_1$.

Estudiemos ahora la función potencia para este caso ($\mu_1 < \mu_2$).

$$\begin{aligned}\pi_\delta(\mu) &= \mathbf{P}_\mu(\delta(\underline{X}) = 1) = \mathbf{P}\left(\frac{\sum_{i=1}^n X_i - n\mu_1}{\sqrt{n\sigma^2}} > z_{1-\alpha}\right) \\ &= \mathbf{P}_\mu\left(\sum_{i=1}^n X_i > z_{1-\alpha}\sqrt{n\sigma^2} + n\mu_1\right) \\ &= 1 - \Phi\left(z_{1-\alpha} - \frac{\sqrt{n}}{\sigma}(\mu - \mu_1)\right)\end{aligned}$$

Propiedades de la función potencia:

- Si μ aumenta entonces $\pi_\delta(\mu)$ aumenta
- $\pi_\delta(\mu)$ es creciente
- $\pi_\delta(\mu_1) = \alpha$
- $\lim_{\mu \rightarrow \infty} \pi_\delta(\mu) = 1$, $\lim_{\mu \rightarrow -\infty} \pi_\delta(\mu) = 0$

A partir de este análisis se puede observar que el mismo test vale para

$$H_0 : \mu \leq \mu_1 \quad Vs. \quad H_1 : \mu > \mu_2$$

Se puede demostrar que el test de cociente de verosimilitud presentado es el test uniformemente mas potente para hipótesis unilaterales (no solo para hipótesis simples). El desarrollo para cada caso puede resultar un poco largo, por lo tanto existen métodos para conseguir el test de una forma más práctica, dadas ciertas condiciones.

3.2. Tests uniformemente mas potentes para hipótesis unilaterales

Sea \underline{X} una muestra aleatoria con distribución perteneciente a una familia exponencial, luego:

1. Si $c(\theta)$ es creciente, el test para

$$H_0 : \theta \leq \theta_1 \quad Vs. \quad H_1 : \theta > \theta_1$$

Será:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } T > k_\alpha \\ 0 & \text{si } T \leq k_\alpha \end{cases}$$

3.2. TESTS UNIFORMEMENTE MAS POTENTES PARA HIPÓTESIS UNILATERALES 35

Que para, un nivel α dado, se tendrá:

$$\alpha = \mathbf{P}_{\theta_1}(\delta(\underline{X}) = 1) \quad (3.4)$$

(también vale la conclusión para distribución uniforme)

2. Si $c(\theta)$ es decreciente, el test para

$$H_0 : \theta \leq \theta_1 \quad V s. \quad H_1 : \theta > \theta_1$$

Será:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } -T > k_\alpha \\ 0 & \text{si } -T \leq k_\alpha \end{cases}$$

Que para, un nivel α dado, se tendrá:

$$\alpha = \mathbf{P}_{\theta_1}(\delta(\underline{X}) = 1) \quad (3.5)$$

Ejemplo: Test de hipótesis para la varianza de una población $\mathcal{N}(\mu, \sigma^2)$ donde μ es un valor conocido. Entonces

$$H_0 : \sigma^2 \leq s \quad V s. \quad H_1 : \sigma^2 > s$$

Como estamos frente a una familia exponencial, sabemos que $T = \sum_{i=1}^n (X_i - \mu)^2$ y que $\eta(\sigma) = -\frac{1}{2\sigma^2}$, la cual es una función creciente, entonces el test correspondiente a las hipótesis planteadas será:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } \sum_{i=1}^n (X_i - \mu)^2 > k_\alpha \\ 0 & \text{si no} \end{cases}$$

A demás $\sum_{i=1}^n (X_i - \mu)^2 / \sigma^2 \sim \chi_n^2$, entonces el test de nivel alfa resulta:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } \frac{\sum_{i=1}^n (X_i - \mu)^2}{s} > \chi_{n, 1-\alpha}^2 \\ 0 & \text{si no} \end{cases}$$

3.3. Test del cociente de máxima verosimilitud

Dada una muestra aleatoria \underline{X} , perteneciente a una población con distribución $f_\theta(x)$, se quiere testear:

$$H_0 : \theta \in \Theta_1 \quad Vs. \quad H_1 : \theta \in \Theta_2$$

Podemos pensar que rechazaríamos H_0 si el valor mas probable de Θ_2 tiene probabilidad considerablemente más grande que el valor más probable de Θ_1

El test de CMV para estas hipótesis será:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } L < k_\alpha \\ 0 & \text{si no} \end{cases}$$

Donde:

$$L = \frac{\sup_{\theta \in \Theta_1} f_\theta(\underline{X})}{\sup_{\theta \in \Theta_2} f_\theta(\underline{X})} \quad (3.6)$$

Si la dimensión de Θ_1 es menor que la dimensión de Θ , entonces es equivalente plantear un test en función de L^* , con

$$L^* = \frac{\sup_{\theta \in \Theta_1} f_\theta(\underline{X})}{\sup_{\theta \in \Theta} f_\theta(\underline{X})}$$

Que en general resulta mucho más fácil.

Podemos verlo de esta forma:

Sea X_1, \dots, X_n una muestra aleatoria de una población con distribución $\mathcal{N}(\mu, \sigma^2)$, hallar el test para

$$H_0 : \mu = \mu_1 \quad Vs. \quad H_1 : \mu \neq \mu_1$$

Proponemos

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } T < k_1 \quad o \quad T > k_2 \\ 0 & \text{si no} \end{cases}$$

Con estadístico del test $T = \sqrt{n} \frac{\bar{X} - \mu_1}{S}$, ya que bajo H_0 , $T \sim t_{n-1}$.

Entonces,

$$\alpha = \mathbf{P}_{\mu_1}(\delta(\underline{X}) = 1) = \mathbf{P}_{\mu_1}(T < k_1 \quad o \quad T > k_2) \quad (3.7)$$

Como la distribución T es simétrica con respecto al cero, entonces resulta

$$\alpha = \mathbf{P}_{\mu_1}(|\sqrt{n} \frac{\bar{X} - \mu_1}{S}| > k) \quad (3.8)$$

con $k = t_{n-1, 1-\alpha}$

A partir de este resultado pueden deducirse los test unilaterales para el caso de poblaciones normales con varianza desconocida.

3.4. Test con nivel de significación asintótico

Para la mayoría de los test de hipótesis se requiere, para encontrar k_α , la distribución del estadístico T para $\theta \in \Theta_1$. Como en muchos casos esta distribución es muy compleja, se puede reemplazar por su distribución asintótica. En este caso el test tendrá un nivel de significación aproximado al deseado para muestras grandes.

Definición: Sea $\underline{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una población con distribución $F_\theta(x)$, $\theta \in \Theta$. Se quiere testear:

$$H_0 : \theta \in \Theta_1 \quad Vs. \quad H_1 : \theta \in \Theta_2$$

Se dirá que una sucesión de test tiene nivel de significación asintótico alfa si:

$$\lim_{n \rightarrow \infty} \sup_{\theta \in \Theta_1} \pi_\delta(\theta) = \alpha$$

Ejemplo:

Encontrar un test de hipótesis de nivel asintótico α para

$$H_0 : p = p_0 \quad Vs. \quad H_1 : p > p_0$$

cuando la población tiene distribución Bernoulli de parámetro p .

Sabemos que la familia $Ber(p)$ es una familia exponencial con $T(\underline{X}) = \sum_{i=1}^n X_i$ y $c(p) = \log(p/(1-p))$ con $c'(p) > 0 \quad \forall 0 < p < 1$, por lo tanto es creciente, entonces usando el teorema sabemos que el test será

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } T > K_\alpha \\ 0 & \text{si no} \end{cases}$$

Pero con la distribución de T no podemos hallar un test de nivel exacto α , ya que corresponde a una variable discreta. Hallamos un test de nivel asintótico, de modo que

$$\lim_{n \rightarrow \infty} \mathbf{P}_{p=p_0}(\delta(\underline{X}) = 1) = \alpha$$

$$\mathbf{P}_{p=p_0}(\delta(\underline{X}) = 1) = \mathbf{P}_{p=p_0} \left(\sum_{i=1}^n X_i > K_\alpha \right) = \mathbf{P}_{p=p_0} \left(\frac{\sum_{i=1}^n X_i - np_0}{\sqrt{np_0(1-p_0)}} > K'_\alpha \right)$$

Aplicando el TCL, sabemos que esa probabilidad es aproximadamente $1 - \phi(K'_\alpha)$. Por lo tanto el test resulta

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } \frac{\sum_{i=1}^n X_i - np_0}{\sqrt{np_0(1-p_0)}} > z_{1-\alpha} \\ 0 & \text{si no} \end{cases}$$

Para encontrar estadísticos para tests asintóticos de nivel α , es útil usar la propiedad de la distribución asintótica de los estimadores de máxima verosimilitud y el método Delta.

3.4.1. Distribución asintótica del test del cociente de máxima verosimilitud

Definición: Sea $\underline{X} = (X_1, \dots, X_n)$ una muestra aleatoria de una población con distribución $F_\theta(x)$, $\theta \in \Theta$. Se quiere testear:

$$H_0 : \theta \in \Theta_1 \quad Vs. \quad H_1 : \theta \in \Theta_2$$

Supongamos que Θ es un abierto en \mathbb{R}^p y supongamos que Θ_1 tiene dimensión $p-j$, entonces bajo H_0 , $-2\log L^* \sim \chi_{j,1-\alpha}^2$, y el test de nivel asintótico α resulta:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si } -2\log L^* \geq \chi_{j,1-\alpha}^2 \\ 0 & \text{si no} \end{cases}$$

Capítulo 4

Bibliografía

1. Hastie T., Tibshirani R., An Introduction to Statistical, Learning with Applications in R
2. Hastie T., Tibshirani R., Elements of Statistical Learning
3. Wackerly, Estadística matemática y aplicaciones
4. Rice J., Mathematical Statistics and Data Analysis
5. Seber G., Lee A., Linear regresión Analysis
6. Kutner M., Applied Linear Statistical Models
7. Flury B., A First Course in Multivariate Statistics
8. Heber G., Multivariate Observations
9. Szretter Noste M., Apunte de regresión lineal, FCEyN