UNIVERSIDAD DE BUENOS AIRES
Facultad de Ingeniería

Clases de Aprendizaje Estadístico

Ing. Jemina M. García

Buenos Aires, 2020.

Índice general

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 $X : \Omega \to \mathbb{R}$ 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 $F_X : \mathbf{R} \to [0, 1]$ dada por

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

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. $lim_{x \to -\infty} F_X(x) = 0$ y $lim_{x \to \infty} F_X(x) = 1$

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 $p_X : \mathbb{R} \to [0, 1]$ 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, a demás 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:

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

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, $\mathbf{P}(X = c) = 0, \qquad \forall c \in \mathbb{R}$

Definición: Se dice que X es una V.A. continua si existe una función $f_X : \mathbb{R} \to \mathbb{R}$, 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 $X \sim \mathcal{E}(1)$ 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 alealeatoria 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 Sea $(\Omega, \mathcal{A}, \mathbf{P})$ un espacio de probabilidad, se dice que $X = (X_1, X_2, \ldots, X_n)$ es un vector aleatorio de dimensión n, si para cada $j = 1, 2, \ldots, n$, $X_j : \Omega \to \mathbb{R}$ es una variable aleatoria.

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

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

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

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

1. $lim_{x,y\to\infty} F_X = 1$, $lim_{x\to-\infty} F_X = 0$, $lim_{y\to-\infty} F_X = 0$
2. $F_X(x)$ es monótona no decreciente en cada variable
3. Es continua a derecha en cada variable
4. $\mathbf{P}((a_1, b_1)x(a_2, b_2)) = F_{X,Y}(b1, b2) - F_{X,Y}(b1, a2) - F_{X,Y}(a1, b2) + F_{X,Y}(a1, a2)$

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:

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

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

Entonces para cualquier conjunto A

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

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 $\Sigma \in \mathbb{R}^{pxp}$ (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 $X \sim \mathcal{N}_p(0, diag(\lambda_1, \ldots, \lambda_p))$, entonces $X_1, \ldots, X_p$ son independientes y tienen distribución $\mathcal{N}(0, \lambda_i)$.

2. Si $X \sim \mathcal{N}_p(\mu, \Sigma)$ y $A \in \mathbf{R}^{pxp}$ 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:

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

Además, $X_1, \ldots, X_p$ 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. $A \in \mathbf{R}^{qxp}$, $rg(a) = q$, $AX \sim \mathcal{N}_q(A\mu, A\Sigma A^T)$
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. $(X - \mu)^t\Sigma^{-1}(X - \mu) \sim \chi^2_p$

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 $X_1, \ldots, X_n$ son independientes sii

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

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

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

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

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

vale para casi todo $x_1, \ldots, x_n$

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: $\sigma = \sqrt{\mathbf{var}(X)}$. 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.

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

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

Propiedades de orden:

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

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. $|\mathbf{E}(X)| \geq \mathbf{E}(|X|)$
6. Sea h(X) una función cóncava, $\mathbf{E}(h(X)) \geq h(\mathbf{E}(X))$
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 $X_1, \ldots, X_n$ 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

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

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 $X_i, i = 1, \ldots, p$, y como elemento $(\Sigma_X)_{i,j} = cov(X_i, X_j)$

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., $X = (X_1, X_2, \ldots, X_n)$ 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 $\varphi : sop_X \to \mathbb{R}$. 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 $t : R_X \to \mathbb{R}$ 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. Y1,Y2, V.A. con esperanza finita, X vector aleatorio, E(aY1+bY2—X)=aE(Y1—X)+ bE(Y2—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

$$\begin{aligned}\mathbf{var}(Y|X=x) &= \mathbf{E}((Y - \mathbf{E}(Y|X=x))^2|X=x)\\ \mathbf{var}(Y|X) &= \mathbf{E}((Y - E(Y|X))2|X)\\ \mathbf{var}(Y|X) &= \mathbf{E}(Y2|X) - \mathbf{E}(Y|X)2\end{aligned}$$

Propiedad (Pitágoras)

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

Demostración:

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

Prediccion: "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., $X = (X_1, X_2, \ldots, X_n)$ 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. Su 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.

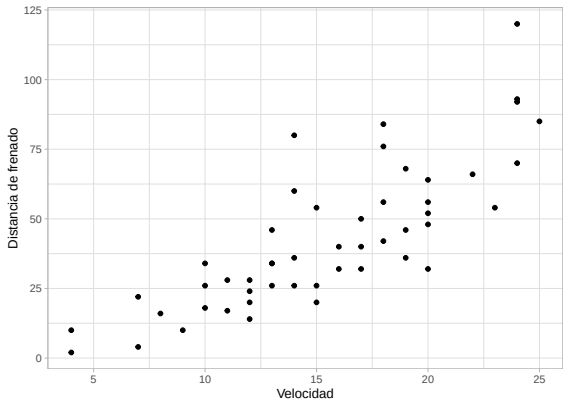

Figura 2.1: Gráfico de dispersión (Scatterplot)

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 $X_1, \ldots, X_P$, 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 $\underline{X} = (X_1, \ldots, X_p)^T$ 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 parámetrica**. 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 monotona 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

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

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

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

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 $i = 1, \ldots, n$.

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

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

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

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

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

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

Podemos escribirlo de forma matricial considerando:
$Y = (y_1, \ldots, y_n)^T$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$, $\beta = (\beta_1, \ldots, \beta_n)^T$,

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

Donde a $\mathbf{X} \in \mathbb{R}^{nxp}$ 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.

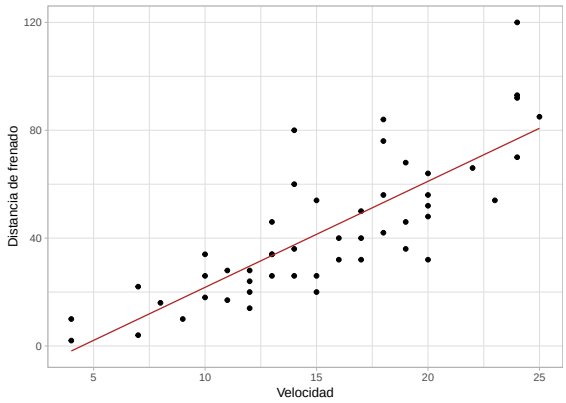

Figura 2.2: Estimación por cuadrados mínimos

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 $i = 1, \ldots, n$, $r_i = y_i - \hat{y}_i$. 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

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

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

el método me 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 la ecuaciones normales, dadas por

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

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 grométrica

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

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

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 mas cerca de Y, que es la proyección ortogonal al subespacio generado por las columnas de $\mathbf{X}$. A dicha proyección la llamamos $\hat{\eta}$.

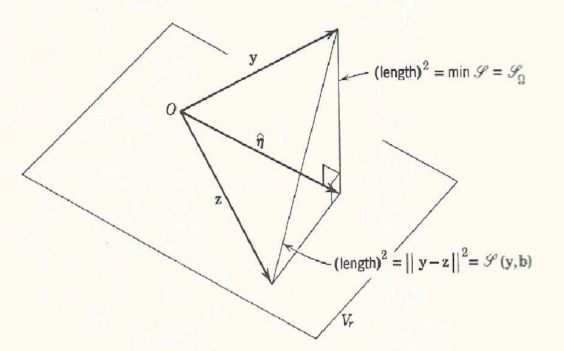

Figura 2.3: Interpretación gromética

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

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

como ya lo habíamos visto. Entonces

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

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

El residuo será

$$r = Y - \hat{Y} = Y - \mathbf{P}Y = (1 - \mathbf{P})Y$$

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

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

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

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

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

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 únivo 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:

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

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 $\Psi_i = c^T\beta$, entonces usaremos como estimador a $\hat{\Psi} = C\hat{\beta}$. Observando su forma matricial, $\Psi = C\beta$ con $C \in \mathbb{R}^{qxp}$, con $c_1, \ldots, c_q$ linealmente independientes (es una propiedad necesaria para que Ψ sea estimable. Para más detalle ver Seber, Linear Analisys). 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

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

Entonces:

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

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

Si tenemo 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

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

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 $(n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$, 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 $(x_i, y_i), i = 1, \ldots, n$, se puede calcular el *p-valor* de la siguiente forma

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

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

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

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

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

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

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

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:

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

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

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

$$\mathbf{C} = \begin{bmatrix} 0 & 1 \ldots & 0 \\ 0 & . & . \\ . & . & . \\ . & . & . \\ 0 & \ldots & 1 \end{bmatrix}$$

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

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

Ejemplo: Boston Housing

2.6. Intervalos de confianza

2.6.1. Intervalos de confianza para la esperanza de una nueva observación independiente

Sabemos que bajo el modelo Ω, $Y \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 I)$, entonces

$$\mathbf{E}(Y_i|X = x_i) = \mathbf{E}(Y_i) = x_i^T\beta.$$

Supongamos ahora que tenemos una nueva observación $(\mathbf{X}_0, Y_0,)$ independiente de todas las demás. Buscaremos un intervalo de confianza para $\mathbf{E}(Y_0) = x_0^T\beta$, nuestro parámetro poblacional a estimar.

Sabemos que $\mathbf{E}(\hat{Y}_0) = \hat{y}_0 = x_0^T\hat{\beta}$ es el estimador de la esperanza que buscamos estimar, y que $\hat{Y_0} \sim \mathcal{N}(x_0^T\beta, \sigma^2 x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0)$.

Un pivote para este caso será

$$T = \frac{\hat{Y_0} - x_0^T\beta}{S\sqrt{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}} \sim t_{n-p}$$

Desarrollando utilizando el método del pivote (ver Anexo 2), resulta que un intervalo de confianza de nivel $1-\alpha$ para $\mathbf{E}(Y_O)$ será

$$IC_{1-\alpha} = \left[\hat{y_0} - t_{n-p,1-\alpha/2}S\sqrt{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}; \hat{y_0} + t_{n-p,1-\alpha/2}S\sqrt{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}\right]$$

Observación: El valor x_0 puede ser o no un valor observado en la muestra, pero siempre debe estar dentro del rango de valores observados para x. Analicemos el caso particular de regresión lineal simple. Si desarrollamos el cálculo de la varianza, podemos ver que

$$\mathbf{var}(\hat{Y_0}) = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right]$$

La variabilidad de nuestra estimación de y_0 se ve afectada por 2 componentes:

1. σ^2, que es la variabilidad de las y cuando conocemos las x
2. Cuan lejos está el valor x_0 de $\bar{X}$ (el promedio observado de la muestra). Entonces la variabilidad de la estimación de la $\mathbf{E}(Y_0)$ será menor cuanto más cercano a la media muestral de $\bar{X}$ este x_0.

2.6.2. Intervalos de predicción

Nuevamente tenemos $(\mathbf{X}_0, Y_0)$ una observación independiente de (x_i, y_i), $i = 1, \ldots, n$. Entonces si $\hat{Y}_0 \sim \mathcal{N}(x_0^T\beta, \sigma^2 x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0)$,

$$\mathbf{E}(Y_0) = \mathbf{E}(\hat{Y}_0) = x_0^T\beta, \qquad Y_0 \sim \mathcal{N}(x_0^T\beta, \sigma^2).$$

También tenemos que

$$\hat{Y}_0 = x_0^T\hat{\beta} = x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$$

es independiente de y_0 ya que es solo función de $y_1, \ldots, y_n$. Entonces resulta

$$\hat{Y} - Y_0 \sim \mathcal{N}(, \sigma^2 + \sigma^2 x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0)$$

una variable variable normal univariada. A partir de ella puedo encontrar un intervalo de confianza de nivel $1-\alpha$ para Y_0 usando el método del pivote, al igual que en el caso anterior.

$$IC_{1-\alpha} = \left[\hat{y_0} - t_{n-p,1-\alpha/2}S\sqrt{1 + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}; \hat{y_0} + t_{n-p,1-\alpha/2}S\sqrt{1 + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}\right]$$

Ahora el error asociado será mayor (busco .ªdivinarçuanto vale Y, no su esperanza)

Ejemplo: Autos

Banda de confianza para la recta estimada

Buscamos obtener una banda de confianza para toda la recta de regresión, es decir una región que, con una confianza de $1-\alpha$ contenga a la recta completa. Esta región tendrá nivel de **al menos** $1-\alpha$ para cada valor de x en particular. La banda de confianza de Sheffé (Working-Hotelling) de nivel $1-\alpha$ para el modelo de regresión lineal está dada por

$$IC_{1-\alpha} = \left[\hat{y_0} - WS\sqrt{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}; \hat{y_0} + WS\sqrt{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}\right]$$

donde $W = \sqrt{q\mathcal{F}_{p,n-p,1-\alpha}}$. Esta idea se desarrollará en detalle para el case de intervalos simultaneos para los parámetros de regresión.

2.6.3. Intervalos de confianza para los parámetros de regresión β_i

Intervalos simples

Bajo el modelo Ω, teníamos que

$$\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{ii})$$

Por lo tanto, podemos construir un intervalo de confianza de nivel $1-\alpha$ para cada β_i usando el método del pivote, con

$$T = \frac{\hat{\beta}_i - \beta_i}{S\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{ii}}} \sim t_{n-p}$$

Resultando

$$IC = [\hat{\beta}_i - t_{n-p,1-\alpha/2}S\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{ii}}, \hat{\beta}_i + t_{n-p,1-\alpha/2}S\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{ii}}]$$

Los valores podemos encontrarlos en la salida de R, donde el *standard error* corresponde a $se = S\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{ii}}$

Intervalos simultáneos

Buscamos intervalos de confianza para q combinaciones lineales $c_i^T\beta$, $i = 1, \ldots, q$ (por ejemplo, cuando buscamos intervalos para comparar parámetros).

Método de Bonferroni

Bajo normalidad, podríamos encontrar un intervalo de confianza para cada combinación lineal (I_i), de manera que

$$\mathbf{P}(c_i^T\beta \in I_i) = 1 - \alpha_i$$

Llamemos $E_i = c_i^T\beta \in I_i$.

Lo que buscamos, es que $1-\alpha$ sea el nivel global, es decir

$$1 - \alpha = \mathbf{P}\left(\bigcap_{i=1}^{q} E_i\right)$$

Desarrollando

$$1 - \alpha = \mathbf{P}\left(\bigcap_{i=1}^{q} E_i\right) = 1 - \mathbf{P}\left(\bigcup_{i=1}^{q} \bar{E}_i\right) \geq 1 - \sum_{i=1}^{q} \alpha_i = 1 - q \cdot \alpha_i$$

Observamos que con un α_i muy pequeño y con un q no tan pequeño, el nivel global da bastante feo (probar con $\alpha_i = 0{,}05$ y $q = 10$). Si totmamos $\alpha_i = \alpha/q$ entonces podemos mantener un nivel global de por los menos $1-\alpha$. Entonces, tomando ese valor, el intervalo para cada combinación lineal será

$$I_i = \left[c_i^T\hat{\beta}_i - t_{n-p,1-\alpha/2q}S\sqrt{c_i^T(\mathbf{X}^T\mathbf{X})^{-1}c_i}, c_i^T\hat{\beta}_i + t_{n-p,1-\alpha/2q}S\sqrt{c_i^T(\mathbf{X}^T\mathbf{X})^{-1}c_i}\right]$$

Ejemplo: Cafeína

Método de Sheffé

Se presenta otra forma, que puede verse como el caso de crear intervalos a partir de invertir la región de aceptación del test de hipótesis. Otra forma de verlo es que estamos usando un pivote ideado a partir del cociente de máxima verosimilitud, realizado para crear el test F.

Supongamos que $c_1, \ldots, c_q$ son L.I., $\Psi = C\beta$, $C \in \mathbb{R}^{qxp}$, $rg(X) = p$, entonces

$$\hat{\Psi} = C\hat{\beta} \sim \mathcal{N}(C\beta, \sigma^2 C(\mathbf{X}^T\mathbf{X})^{-1}C^T)$$

Por lo tanto, mi pivote será

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

Lo que estamos buscando es un intervalo de confianza para todas las combinaciones lineales $\Psi = C\beta$, por lo tanto tenemos que calcular el cuantil para nuestro pivote y luego despejar.

$$\mathbf{P}(F < \mathcal{F}_{q,n-p,1-\alpha}) = 1 - \alpha$$

Despejar de esa expresión no es tarea fácil. Si llamamos $L = C(\mathbf{X}^T\mathbf{X})^{-1}C^T$ y $b = \hat{\Psi} - \Psi$, tenemos que se cumple la siguiente igualdad:

$$b^T L^{-1} b = \max_{h \neq 0} \frac{(h^T b)^2}{h^T L h}$$

Por lo tanto

$$\begin{aligned}
\mathbf{P}(F < \mathcal{F}_{q,n-p,1-\alpha}) &= 1 - \alpha \\
&= \mathbf{P}\left(\forall h \neq 0, \frac{(h^T b)^2}{h^T L h} \leq qS^2 \mathcal{F}_{q,n-p,1-\alpha}\right) \\
&= \mathbf{P}\left(\forall h \neq 0, |h^T\hat{\Psi} - h^T\Psi| \leq \sqrt{q\mathcal{F}_{q,n-p,1-\alpha}} S (h^T L h)^{1/2}\right)
\end{aligned}$$

Entonces, para cualquier función $h^T\Psi$, un intervalo de confianza de nivel $1 - \alpha$ será (siempre que $h \neq 0$)

$$IC = \left[h^T\hat{\Psi} - \sqrt{q\mathcal{F}_{q,n-p,1-\alpha}} S (h^T L h)^{1/2}, h^T\hat{\Psi} + \sqrt{q\mathcal{F}_{q,n-p,1-\alpha}} S (h^T L h)^{1/2}\right]$$

Vemos que si $h = e_i$ el vector canónico con 1 en la posición i, recupero los intervalos de confianza para los $c_i^T\beta$. ¿Qué significa? Dentro del elipsoide de confianza tengo todas

las combinaciones lineales de β con por lo menos ese nivel de confianza. Se puede observar mediante simulaciones que si la cantidad de intervalos simultáneos que buscamos son pocos, Bonferroni da mejores resultados, si no conviene usar Sheffé.

Ejemplo: Café

Otra forma equivalente del test de significación de la regresión

Presento una forma más geométrica de ver el test de la regresión. Bajo el modelo Ω planteamos

$$H_0 : \Psi = \delta \quad Vs. \quad H_1 : \Psi \neq \delta$$

Si definimos $W = H_0 \cap \Omega$,

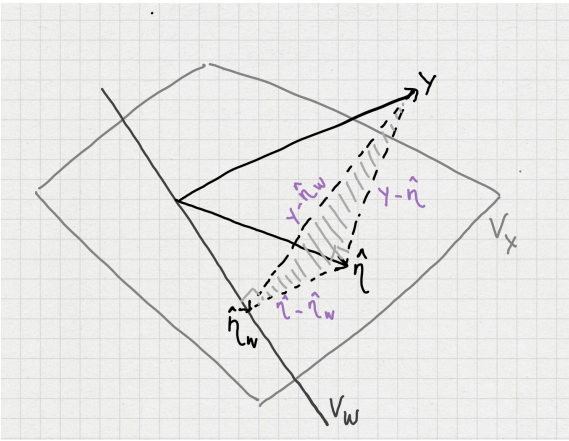

Figura 2.4: Interpretación grométrica

Recordemos que V_X es el subespacio generado por las columnas de la matriz de diseño, llamábamos $\hat{\eta} = \hat{Y}$ a la proyección ortogonal de Y en el subespacio generado por las columnas de X, y resulta en la estimación de la esperanza condicional de Y dado x. Ahora, si suponemos que H_0 es verdadero, entonces el espacio se reduce, a un espacio V_W, que para el caso de significación de la regresión, ese espacio asume que todos los parámetros salvo β_0 son cero, resultando el estimador $\hat{\eta}_W - \bar{Y}$. El gráfico no es muy bonito, pero por ahí se llega a apreciar que $Y - \hat{\eta}$ es ortogonal a $\hat{\eta}$, $Y - \eta\hat{W}$ es ortogonal a $\hat{\eta}_W$, y que $\hat{\eta} - \hat{\eta}_W$ es ortogonal a $\hat{\eta}_W$. Lo que estamos queriendo analizar es si $\hat{\eta}_W$ está muy lejos de $\hat{\eta}$, porque de ser así, entonces el modelo de regresión aporta mucha más información para predecir a la esperanza de Y que usar solo el promedio. Con todo esto en mente, definimos nuevamente el estadístico F pero ahora en función de todas estas distancias.

Ya sabemos que

$$\frac{||Y-\hat{\eta}||^2}{\sigma^2} \sim \chi_{n-r}, \quad \frac{(n-r)S^2}{\sigma^2} \sim \chi_{n-r}$$

Además, bajo H_0 tenemos que

$$\frac{||\hat{\eta}-\hat{\eta}_W||^2}{\sigma^2} \sim \chi_q$$

ya que

$$||\hat{\eta}-\hat{\eta}_W||^2 = (C\hat{\beta}-\delta)^T(C(\mathbf{X}^T\mathbf{X})^{-1}C^T)^{-1}(C\hat{\beta}-\delta)$$

Por lo tanto

$$F = \frac{||\hat{\eta}-\hat{\eta}_W||^2/q\sigma^2}{||Y-\hat{\eta}||^2/(n-r)\sigma^2} \sim \mathcal{F}_{q,n-r}$$

Aplicado al test de significación de la regresión, donde $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$, vale que

$$F = \frac{||\hat{Y}-\bar{Y}||^2/q\sigma^2}{||Y-\hat{Y}||^2/(n-r)\sigma^2} \sim \mathcal{F}_{q,n-r}$$

Y rechazamos H_0 si el valor de F es grande, es decir si la estimación bajo H_0 está muy lejos de la estimación bajo H_1.

Usando pitágoras podemos ver que

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

AL primer término lo llamamos *suma de cuadrados de la regresión* (SCR o RSS en inglés), al segundo *suma de cuadrados totales* (SCT)y al tercero *suma de cuadrados de los residuos*(SCRes).

2.7. Coeficientes de correlación (R) y determinación (R^2)

Se trata de decidir si el hecho de conocer el valor de X mejora nuestro conocimiento de Y. ¿Cuánto de la variabilidad de Y queda explicado por la regresión?. Definimos al coeficiente de determinación como

$$R^2 = \frac{||\hat{Y}-\bar{Y}||^2}{||Y-\bar{Y}||^2} = \frac{SCR}{SCT}$$

Es una medida de la **capacidad de predicción** del modelo. Si R^2 está cerca de 1 significa que el modelo propuesto aporta para explicar la variabilidad de Y. Vemos que

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

Si $||Y-\hat{Y}||^2 = 0$ es porque los puntos observados caen sobre una recta, y si $||\hat{Y}-\bar{Y}||^2 = 0$ significa que el modelo no aporta nada.

Un problema que surge es que a medida que agregamos covariables al modelo, el valor de R^2 aumenta siempre, por lo tanto es una medida que no nos sirve para comparar modelos que tengan diferente cantidad de variables. Necesitamos definir otra variable que no tenga este problema. Definimos al R^2 ajustado como

$$R^2_a = 1 - \frac{n-1}{n-p}\frac{||Y - \hat{Y}||^2}{||Y - \bar{Y}||^2} = 1 - (1 - R^2)\frac{n}{n-p} \tag{2.1}$$

Divide a cada suma de cuadrados por sus grados de libertad, y esta medida aumenta si F aumenta. Esta medida es motivada a partir del test para analizar cuan bueno es un modelo versus el modelo con todas las variables explicativas.

Supongamos que hay K variables disponibles pero el modelo propuesto tiene $p-1 < k$ variables. Si escribimos los coeficientes de determinación para los dos modelos R^2_p y R^2_{k+1} el test F que testea el ajuste del modelo chico vs. el grande será

$$F_p = \frac{R^2_{k+1} - R^2_p}{1 - R^2_{k+1}}\frac{n-k-1}{k-p+1}$$

Entonces, despejando,

$$1 - R^2_p = (1 - R^2_{k+1})\frac{(k-p+1)F_p + n - k - 1}{n-k-1}$$

Usando la ecuación (2.1) tenemos que

$$R^2_{padj} = 1 - (1 - R^2_{k+1})\frac{n}{n-k-1}\frac{(k-p+1)F_p + n - k - 1}{n-p}$$

Ocurre que $\frac{(k-p+1)F_p+n-k-1}{n-p} \geq 1$ si $F_p \geq 1$ entonces $R^2_{padj} \leq R^2_{k+1}$ajustado. Esto muestra que grandes valores de F_p son evidencia a favor del modelo con k variables, lo que también devuelve un valor de R^2_a mayor para el modelo con k variables.

2.8. Análisis de residuos

Si tengo una sola variable predictora, puedo graficar y vs. x y darme cuenta si el supuesto del modelo propuesto es válido, pero con más variables necesito otra forma de visualización. Analicemos los residuos: $r_i = y_i - \hat{y}_i$, que son variables aleatorias don $\mathbf{E}(r_i) = 0$ y $\mathbf{var}(r_i) = \sigma^2(1 - p_{ii})$, donde $P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Si realizo un gráfico de los residuos y el modelo es válido, no debería observarse ninguna estructira, deberían aparecer puntos distribuidos aleatoriamente al rededor del cero.

Ejemplo: Supongamos que se plantea el modelo

$$(1) \qquad \mathbf{E}(Y) = \beta_0 + \beta_1 x$$

pero el verdadero modelo es

$$(2) \qquad \mathbf{E}(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

¿Cuál es el sesgo de la estimación si las observaciones se toman el $x = (-1, 0, 1)$? Sabemos que $sesgo(\hat{\beta}) = \mathbf{E}(\hat{\beta}) - \beta$, si fuera insesgado debería dar cero. Realizando los cálculos de los estimadores y de su esperanza obtenemos que

$$\mathbf{E}[\hat{\beta}_0] = \beta_0 + 2/3\beta_2, \qquad \mathbf{E}[\hat{\beta}_1] = \beta_1$$

El olvidarme de una variable (término cuadrático del modelo verdadero), afectó a la estimación del intercept, no así a la estimación de β_1.

Pero ahora, ¿Cómo afecta ese sesgo en los supuestos? Si X es fijo, tenemos que $r = y - \hat{y} = y - (\beta_0 + \beta_1 x)$, podemos calcular su esperanza (que debería dar cero).

$$\mathbf{E}(r) = \mathbf{E}(y - (\beta_0 + \beta_1 x)) = \beta_0 + \beta_1 x + \beta_2 x^2 - \beta_0 - +2/3\beta_2 - \beta_1 x = \beta_2 x^2 - +2/3\beta_2$$

Lo cual resulta en una forma cuadrática. Si graficamos los residuos en función de los x vamos a observar una estructura cuadrática en el gráfico, lo cual nos está indicando que el modelo propuesto no es el adecuado. Siempre que el gráfico presente "patrones"hay que sospechar que el modelo planteado no es el correcto. Lo mismo pasa si graficamos r_i vs. y.

Si el modelo es el correcto, no debería haber ninguna correlación entre r y $\hat{Y}$. Esto lo vemos desarrollando el cálculo de la correlación entre ambas. Si desarrollamos el cálculo del numerador

$$\sum (r_i - \bar{r})(\hat{y}_i - \bar{\hat{y}}) = \sum r_i(\hat{y}_i - \bar{y}) = \sum r_i \hat{y}_i - \bar{y} \sum r_i = r^T \hat{y} = 0$$

La primera igualdad se cumple solo si el modelo tiene intercept ya que $\sum (y_i - \hat{y}_i) = 0$.

Entonces si el modelo es el correcto y graficamos r_i vs. $\hat{y}_i$ no debe observarse estructura, solo puntos aleatorios al rededor del cero.

Ahora, los r_i no son independientes ni tienen igual varianza, pero si asumimos normalidad, entonces

$$\frac{r_i}{\sqrt{\sigma^2(1 - pii)}} \sim \mathcal{N}(0, 1)$$

Se definen residuos estandarizados a

$$t_i = \frac{r_i}{\sqrt{S^2(1 - pii)}}$$

con $\mathbf{E}(t_i) = 0$, $\mathbf{var}(t_i) = \sigma^2$. Si los r_i son normales entonces los t_i tienen distribución t de Student.

Para analizar homocedasticidad se grafica t_i vs. $\hat{y}_i$. Si quiero analizar graficamente la normalidad, puedo hacer un QQ-norm de los r_i o de los t_i.

¿Para que más sirve el gráfico de los residuos? Para detectar posibles outliers. Pero ahora ya no es confiable el gráfico de los t_i porque de haber outliers la recta estimada podría verse afectada (lo ideal sería utilizar métodos robustos). Lo que se usan son los residuos de cross validation. Si para cada i, $\hat{\beta}_{-i}$ es el estimador de mínimos cuadrados sin la observación i, y $r_{-i} = y_i = x_i^T \hat{\beta}_{-i}$. Parece muy feo tener que hacer toda esta cuenta, pero se puede probar que

$$r_{-i} = \frac{r_i}{1 - p_[ii]}$$

Un gráfico muy útil es el de r_{-i} vs $\hat{y}_i$ para ver posibles outliers, y el QQ-norm de r_{-i} para verificar normalidad. Es conveniente que los residuos estén normalizados, entonces se definen los residuos estudentizados

$$t_{-i} = \frac{r_i}{S_{-i}\sqrt{1 - p_[ii]}}$$

Para un análisis rápido, un $|t_i| > 2{,}5$ se puede considerar sospechoso.

Más adelante veremos el tema de cross validation en detalle.

Para detectar posible correlación entre errores, se usa el test de Durbin Watson (pueden encontrarlo en el Seber)

Si el modelo no es válido o se observa heterocedasticidad, se transforman las variables o se aplica el método de mínimos cuadrados pesados.

Transformaciones de Box y Cox

Proponen una familia de funciones de potencia para la variable respuesta con el objetivo de garantizar el cumplimiento de los supuestos. Estas transformaciones combinan el objetivo de encontrar una relación simple con homogeneidad de varianzas, mejorando la

normalidad.

La transformación de Box y Cox está dada por:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{si} \quad \lambda \neq 0 \\ \log(y) & \text{si} \quad \lambda = 0 \end{cases}$$

Esta familia es continua en λ y monótona creciente para cada λ. No toda distribución puede ser transformada a la normal pero se probó que aunque no se consiga exacta normalidad, los estimadores usuales de lambda conducen a distribuciones cuyos primeros cuatro momentos corresponden a simetría. Supongamos que las observaciones transformadas $y^{(\lambda)} \sim \mathcal{N}(X\beta, \sigma^2 I)$ entonces los parámetros del modelo son λ, β, σ^2. Se pueden estimar por el método de máxima verosimilitud pero la resolución es complicada. Otra forma: se buscan los estimadores de β y σ para cada λ fijo y luego se elije el λ mas adecuado.

A partir de un cambio de variable, se consigue la densidad de y, que depende de los tres parámetros desconocidos. Loego se consiguen los estimadores de máxima verosimilitud. Finalmente se considera la log-verosimilitud evaluada en $\hat{\beta}$ y $\hat{\sigma^2}$ y se maximiza para λ. Usando el cociente de máxima verosimilitud se construye un estadístico, y con la distribución asintótica se arma un intervalo de confianza para λ. En la práctica con R podemos obtener el intervalo de confianza y luego elegir un valor dentro del intervalo, se elige el mas conveniente.

2.8.1. Selección de variables para predicción

La idea es que quiero un vector de parámetros que me de una buena predicción para una Y "nueva", es decir que no pertenece a la muestra. Se usa el criterio del error cuadrático medio. Veamos solo el caso en el que X es fija. Sea y_0 una nueva observación en $\mathbb{R}^n$,

$$ECM(\hat{\beta}) = \mathbf{E}[||y_0 - X\hat{\beta}||^2]/n$$

Ahora, para cada subconjunto de variables calculo el estimador de mínimos cuadrados y me quedo con el que da menor ECM. ¿Cómo puedo estimar su valor? El mejor estimador resulta ser el de validación cruzada:

$$\widehat{ECM} = \frac{1}{n}\sum_{i=1}^{n} r_{-i}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{r^2}{(1-p_{ii})^2}$$

2.8.2. Selección de un subconjunto de subvariables

Método Stepwise tiene 3 versiones: Best subset, Forward y Backward

1. Best subset: estudia las 2^k regresiones y y elijo el que tenga mayor R_a^2 o mejor C_P. Es el único procedimiento que garantiza que se obtenga el modelo final que realmente optimiza la búsqueda del criterio elegido (pero si hay muchas covariables no se puede poner en práctica)

2. Forward: El modelo arranca con el intercept, luego agrega una variable, prueba todos los modelos posibles y se queda con el modelo de mayor F (o menor p-valor). Busca rechazar H_0 cuando esta hipótesis plantea el modelo con p variables vs el modelo con k variables, $k > p$, y el estadístico $F = ||\hat{Y}_p - \hat{Y}_k||^2/S^2$, con $S^2 = ||Y - \hat{Y}_k||^2/(n-k)$

3. Backward: Arranca con todas las variables y va sacando de a una. Elimina la variable que devuelve el valor F más chico. Este método no es bueno si hay colinealidad.

2.8.3. Criterio basado en el error de predicción

Se basa en la idea de elegir al modelo que predice bien.

Cp de Mallow

Sea RSS_p el valor de RSS que resulta de ajustar el modelo con $p < k$ parámetros. Recordemos que $RSS = ||Y - \hat{Y}||^2$ y que el error de predicción de puede medir como $||y_0 - X_0\hat{\beta}||^2$.

Se define Cp de Mallow como $Cp = \frac{RSS_p}{S^2} + 2p - n$. Es una estimación de la esperanza del error del modelo que representa cuan bien el modelo lineal representado por x_0 estima la media μ_0. Entonces

$$Cp = \frac{||Y - \hat{Y}_p||^2}{S^2} + 2p - n, \quad pero \quad S^2 = \frac{||Y - \hat{Y}_k||^2}{n-k}$$

Si el modelo fuera el correcto y $k = p$, entonces $Cp = p$. Un buen modelo debe tener $Cp \sim p$.

2.8.4. Colinealidad

Existe cuando las variables predictoras incluidas en el modelo están correlacionadas entre sí. Problemas que trae:

1. Los coeficientes estimados cambian mucho cuando se agregan o quitan variables

2. La varianza de los estimadores aumenta mucho cuando se agregan covariables correlacionadas (infla la varianza estimada)

3. Los coeficientes pueden ser nosignificativos aún cuando exista asociación entre las variables

Un enfoque apropiado es hacer una regresión de cada variable regresora sobre las demás. Cuando el R^2 da cercano a 1 debemos preocuparnos por el efecto de la multicolinealidad.

Método de detección: uso los factores de inflación de la varianza (VIF). Se calculan para cada covariable como

$$VIF_k = \frac{1}{1 - R_k^2}, \qquad 1 \leq k \leq p - 1$$

R_k^2 es el coeficiente de determinación cuando x_k es la variable regresada en las $p - 2$ restantes covariables. Si $VIF > 10$ decimos que hay señal de colinealidad. Soluciones:

1. Selección de modelos (Stepwise)
2. Regularización o penalización (Lo veremos más adelante)
3. Reducción de la dimensión (componentes principales)

Capítulo 3

Clasificación

El modelo de regresión lineal asume que la variable respuesta Y es cuantitativa. Pero en muchas situaciones la variable puede ser cualitativa. Consideremos una variable categórica que puede indicar la pertenencia a una categoría, o clase, o estado (por ejemplo un paciente puede estar sano o enfermo) y se quiere predecir el estado en función de otras variables $X_1, \ldots, X_p$, por ejemplo peso, edad, colesterol, glucosa en sangre, presión sanguínea, etc. Para predecir este tipo de variables se usa Clasificación, ya que a cada observación se le asigna una categoría (o clase). Hay varios clasificadores:

1. Vecinos más cercanos (K-NN)
2. Regresión logística
3. LDA (linear discriminant analysis) y QDA
4. Redes neuronales
5. etc.

Ejemplo DEFAULT visto en clase

La información que tendremos disponible para poder predecir será:

- Covariables $X = (X_1, \ldots, X_p) \in \mathcal{X}$
- Posibles etiquetas $y \in \mathcal{Y}$
- Clasificador: regla que asigna a un $x \in \mathcal{X}$ un posible valor $y \in \mathcal{Y}$.

3.1. Regla de Bayes - Caso binario

La regla óptima de Bayes dice que el clasificador óptimo estará dado por

$$H^{opt}(x) = \begin{cases} 1 & \text{si} \quad \mathbf{P}(Y=1|X=x) > \mathbf{P}(Y=0|X=x) \\ 0 & \text{si} \quad \mathbf{P}(Y=1|X=x) < \mathbf{P}(Y=0|X=x) \end{cases}$$

Teorema: Para todo H, vale que

$$L(H^{opt}) = \mathbf{P}(H^{opt}(x) \neq Y) \leq \mathbf{P}(H(x) \neq Y) = L(H)$$

Se define el error de clasificación medio como

$$L(H) = P(H(x) \neq Y)$$

El error de clasificación empírico (es decir el estimador del error de clasificación medio) se calcula a partir de los datos $(x_1, y_1), \ldots, (x_n, y_n)$ como

$$\hat{L}_n(H) = \frac{1}{n} \sum_{i=1}^{n} I\{\hat{H}_n(x_i) \neq Y_i\}$$

Como esta estimación puede resultar en overfitting, se recomienda usar el de *leave-one-out*

$$CV = \frac{1}{n} \sum_{i=1}^{n} I\{\hat{H}_n^{-i}(x_i) \neq Y_i\}$$

La pregunta que tenemos que hacernos ahora es, ¿cómo estimamos $P(Y = 1|X = x)$ para poder armar nuestra regla de clasificación?

3.2. Vecinos más cercanos

El método de K- vecinos más cercanos es uno de los métodos existentes para estimar la distribución de Y dado X y después clasificar una observación en la clase con mayor probabilidad estimada. Los pasos a seguir son

1. Elegimos un k entero positivo y un punto x para clasificar
2. Requiere una noción de distancia. Si usamos la euclídea, dados $Z, W \in \mathbb{R}^q$,

$$d(Z, W) = \sqrt{\sum_{i=1}^{q} (z_i - w_i)^2}$$

3. El clasificador $K - NN$ identifica el conjunto de los k puntos más cecanos a x. Sea N_x dicho conjunto
4. Estima $P(Y = 1|X = x)$ por la fracción de puntos en N_x cuya etiqueta es 1

$$\hat{p} = \frac{1}{k} \sum_{i \in N_x} I\{Y_i = 1\}$$

5. El parámetro k se suele elegir por cross validation.

Ventajas y desventajas

1. Es un método muy intuitivo en el que a un nuevo punto se le asigna una categoría por voto de la mayoría entre los k vecinos más cercanos.
2. Se generaliza muy facilmente a un problema con más de dos clases
3. Se pueden usar distintas distancias.
4. Para prevenir que alguno de los atributos tenga más influencia en la medida de distancia que otros se suele escalar, de esta manera una distancia d significa lo mismo para el atributo 1 y para el 2, por ejemplo
5. Atributos irrelevantes podrían incrementar la distancia artificialmente a casos similares.
6. Maldición de la dimensión
7. La clasificación de nuevos registros es más costosa que con otros métodos. Es un clasificador de aprendizaje perezoso (lazy).
8. No construye un modelo explícito.

3.3. Regresión logística

Recordemos que estamos buscando estimar $P(Y = 1|X = x) = p(x)$. ¿Y si modelamos esta probabilidad en función de x?

Si tenemos que $Y = 1$ o $Y = 0$, entonces

$$Y|X = x \sim Ber(p(x)), \qquad \mathbf{E}[Y|X = x] = p(x)$$

Si queremos estimar la esperanza condicional, porqué no usar regresión lineal? Si hacemos esa estimación usando los datos del ejemplo de Default:

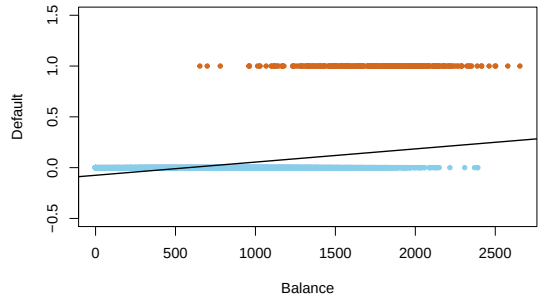

Figura 3.1: Ajuste usando regresión lineal

Observamos que la estimación de la probabilidad no devuelve valores entre 0 y 1, no es una estimación valida para lo que queremos.

Para este caso, se propone el modelo logístico para la estimación de la probabilidad. La función logística es una función acotada entre 0 y 1 dada por:

$$f(t) = \frac{1}{1 + e^{-t}}$$

Se define **ODDS** como la proporción de la probabilidad de que algo ocurra sobre la probabilidad de que no ocurra

$$ODDS = \frac{p(x)}{1 - p(x)}$$

Describe, en un modelo logístico, el riesgo para un individuo x (por ejemplo $odds = 1/3$ significa que las chances de que el evento no ocurra son 3 a 1).

Observemos que si tomamos logaritmo a las *odds*, resulta

$$-\infty < \log\left(\frac{p(x)}{1 - p(x)}\right) < \infty$$

A dicho logaritmo se lo llama función $logit(p(x))$. Como ese valor sí puede ser cualquier valor en los reales, puede ser modelado según un modelo lineal. Si se propone

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

Resulta que despejando tenemos el modelo de regresión logística:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

¿Cómo interpretamos el coeficiente β_1? Podemos interpretar que pasa con *logit* cuando una sola x varía y las otras son fijas. Si $P(Y = 1|X_1 = 1) = p_1(x)$ y $P(Y = 1|X_1 = 0) = p_0(x)$, con todas las otras covariables fijas (si las hay), entonces

$$logit(p_1) - logit(p_0) = \beta_1$$

β_1 indica el cambio en la logit cuando x_1 cambia en una unidad, si todas las otras x_i son fijas.

En general, tendremos un vector de covariables, y el modelo logistico será

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x + \cdots + \beta_{p-1} x_{p-1})}}$$

3.4. Estimación de parámetros

Para estimar los parámetros desconocidos del modelo se utiliza el método de máxima verosimilitud (Ver Anexo 3). A la función de probabilidad o función de densidad conjunta de la muestra aleatoria, vista como una función de θ la llamaremos función de verosimilitud. Lla función de verosimilitud será

$$L(\theta, \underline{x}) = \prod_{i=1}^{n} f_\theta(x_i) \tag{3.1}$$

Donde $f_\theta(x_i)$ será la función de densidad o la función de probabilidad,dependiendo si el vector aleatorio $\underline{X}$ es discreto o continuo. La función de verosimilitud es la probabilidad de haber observado el resultado de la muestra en función del parámetro desconocido. Lo que buscamos por el método de Máxima Verosimilitud es el valor del parámetro que maximiza dicha probabilidad.

Para nuestro caso de regresión logística tenemos que

$$Y_i|X = x_i \sim Ber(p(x_i)), \qquad i = 1, \ldots, n$$

Si llamamos $\theta = p(x_i)$ tenemos que

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}$$

En este caso, maximizar $L(\theta)$ es equivalente a maximizar $l(\theta) = \log(L(\theta))$.

$$l(\theta) = \sum_{i=1}^{n} y_i \log(\theta) + (n - \sum_{i=1}^{n} y_i) \log(1-\theta)$$

Ahora, θ es una función de las x_i y de los parámetros β, por lo tanto

$$l(b) = \sum_{i=1}^{n} [y_i \log(p(x_i, b)) + (1-y_i) \log(1 - p(x_i, b))]$$

Tipicamente, para hallar el EMV se deriva la log-verosimilitud $l(b)$ e igualo a cero. Notar que para el modelo logístico, $p(x, \beta) = h(x^T\beta)$ siendo $h(t) = 1/(1+e^{-t})$, por lo que $h'(t) = h(t)(1-h(t))$. En $l(b)$, cada término es de la forma

$$y \log(p(x, b)) + (1-y) \log(1 - p(x, b))$$

entonces la derivada parcial respecto de cada b_j es de la forma

$$\frac{y}{p(x,b)} p(x,b)(1-p(x,b))x_j - \frac{1-y}{1-p(x,b)} p(x,b)(1-p(x,b))x_j = (y - p(x,b))x_j$$

Resulta entonces

$$\frac{dl(b)}{db_j} = \sum_{i=1}^{n}(y_i - p(x_i, b))x_{i,j}, \qquad j = 0, 1, \ldots, p$$

es de la forma $X^T(Y = P(b))$ don de Y y P representan a los vectores de componentes y_i y $p(x_i, b)$. El EMV resuelve $X^T(Y = P(b)) = 0$. La solución de este sistema no tiene una expresión analítica y debe ser resuelto numéricamente por el método de Newton-Raphson o el método de Fisher-Scoring.

Entendiendo quienes son las funciones de verosimilitud, podemos observar que el modelo logístico asume que la diferencia de los logaritmos de las verosimilitudes es lineal. Se verá más en detalle cuando veamos Análisis Discriminante.

Una vez estimados los parámetros, podemos estimar la probabilidad *a posteriori* dada una muestra. Tendremos que

$$\widehat{p}(x) = \widehat{\mathbf{P}}(Y = 1|X = x) = \frac{1}{1 + e^{-x^T\widehat{\beta}}}$$

Estimada la probabilidad, lo que queremos es clasificar la nueva observación, asignarle una etiqueta. Podemos simplemente utilizar la regla propuesta al inicio, y decir que si dicha probabilidad estimada es mayor a 0,5, clasificamos a la nueva observación como $Y = 1$. Otra forma es utilizar medidas de error o de presición del modelo de clasificación para elegir el parámetro de corte θ. Ya vimos como calcular el error de clasificación, podría elegir el valor de $0 < \theta < 1$ que minimice dicho error. Para calcular el error, es muy útil realizar una *tabla de confusión*, que consiste en

Predicho		0	1
Real	0	TN	FP
	1	FN	TP

Donde TN: true negative, FP: false positive, FN: false negative y TP: true positive.

A partir de la tabla podemos observar que el error de clasificación empírico se calcula como $\hat{L}$=(FN+FP)/n, con n el tamaño de la muestra (n=TP+TN+FP+FN)

Otras medidas son

- Acuracy = (TN+TP)/n
- Precisión = TP/(TP+FP)
- Recall = TP/(TP+FN)
- F_{score}=2TP/(2TP+FP+FN)

Otra herramienta útil son las llamadas **Curvas ROC**. Se definen:

$$TPR(\theta) = \frac{TP(\theta)}{TP(\theta) + FN(\theta)}, \qquad FPR(\theta) = \frac{FP(\theta)}{TN(\theta) + FP(\theta)}$$

Para valores de $0 < \theta < 1$ se calculan los pares $(FPR(\theta), TPR(\theta))$, calculo la distancia al $(0, 1)$ y elijo como corte el valor de θ que minimiza la distancia. La curva generada por los puntos $(FPR(\theta), TPR(\theta))$, $0 < \theta < 1$ se llama Curva ROC. Se puede observar que si la curva cayera en el $(0, 1)$ esto significaría que FN=0 y FP=0, lo cual implica que tengo el mejor clasificador posible. Por lo contrario, si cayera en el $(1, 0)$, TN=0 y TP=0, o sea que siempre clasificaría al revés. Los titas que logran los puntos (1,1) y (0,0) crearían clasificadores completamente inútiles.

3.4.1. Inferencia

Fahrmeir y Kaufmann (1985) estudiaron el comportamiento de los estimadores de máxima verosilmilitud para modelos lineales generalizados bajo condiciones de regularidad. Obtuvieron que para un n suficientemente grande, $\hat{\beta}_n - \beta$ tiene distribución asintótica Normal de parámetros $\mu = 0$ y $\sigma^2 = \mathbf{var}(\hat{\beta}_n) = (X^T W X)^{-1}$, con $W = diag(p(X_1)(1 - p(x_1), \ldots, p(X_n)(1 - p(x_n)))$. Como W (y por lo tanto $\mathbf{var}(\hat{\beta}_n)$) dependen de β, podemos utilizar el estimador *plug-in* $\hat{\mathbf{var}}(\hat{\beta})$ reemplazando por los estimadores de máxima verosimilitud.

Para una función lineal $\psi = a^T\beta$, una aproximación razonable será decir que $(a^T\hat{\beta} - a^T\beta)$ tiene distribución asintótica normal de media 0 y varianza $a^T\mathbf{var}(\hat{\beta})a$. Utilizando esta información, podemos construir intervalos de confianza de nivel $1 - \alpha$ con el método del pivote ya estudiado en el capítulo anterior, resultando un intervalo de confianza para ψ de nivel aproximado $1 - \alpha$

$$a^T\hat{\beta} \pm z_{1-\alpha/2}\sqrt{a^T\mathbf{var}(\hat{\beta})a}$$

Si $a = x_0$ es una nueva observación, puedo obtener una estimación puntual de la probabilidad $p(x_0, \beta)$ y un intervalo de confianza para la misma. Sabemos que

$$\widehat{p}(x_0) = \widehat{\mathbf{P}}(Y = 1|X = x_0) = \frac{1}{1 + e^{-x_0^T\widehat{\beta}}}$$

¿Cómo calculamos el intervalo de confianza? Si n es grande, tenemos que un intervalo de confianza para $x_o\beta$ será de la forma

$$x_0^T\hat{\beta} \pm z_{1-\alpha/2}\sqrt{x_0^T\mathbf{var}(\hat{\beta})x_0}$$

Como $h(t) = 1/(1 + e^{-t})$ es estrictamente creciente, vale que

$$I = \left(\frac{1}{1 + e^{-(x_0^T \hat{\beta} - z_{1-\alpha/2}\sqrt{x_0^T \mathbf{var}(\hat{\beta}) x_0})}}, \frac{1}{1 + e^{-(x_0^T \hat{\beta} + z_{1-\alpha/2}\sqrt{x_0^T \mathbf{var}(\hat{\beta}) x_0})}} \right)$$

es un intervalo de confianza de nivel aproximado $1 - \alpha$ para $p(x_0)$.

Otra opción es usar el método delta mencionado en el anexo 3.

Ejemplo: Pima

3.4.2. Bondad de ajuste

Consideremos la comparación de dos modelos anidados. La diferencia entre los dos modelos es que la componente lineal de un modelo se obtiene anulando alguno de los parámetros del otro.

Definimos nuestra hipótesis nula H_0 como que el modelo correcto es el más simple. Si el modelo más simple ajusta a los datos tan bien como el general, por el principio de parsimonia no rechazaríamos H_0. Para realizar las comparaciones necesitamos medidas de bondad de ajuste.

El modelo más complejo se llama modelo saturado, en el que cada observación se predice por si misma: $\hat{Y}_i = Y_i$, mientras que el modelo mas simple, o modelo nulo, es el que tiene un solo parámetro, y por lo tanto $\hat{Y}_i = \bar{Y}$. La medida de bondad de ajuste más común en modelo lineal generalizado es la **Deviance**.

Para el caso de regresión logística, $D(\beta) = -2\log(L(\beta))$ se llama desviación del modelo i, $d_i = -2(y_i \log(p_i)) + (1 - y_i)\log(1 - p_i))$ es la desviación del dato i y mide el ajuste del modelo al dato (x_i, y_i). Será más grande si la observación está mal explicada por el modelo.

Si $l(\hat{\mu}, \Phi, y)$ es el log-likelihood del modelo de interés y $l(y, \Phi, y)$ es el del modelo saturado, entonces la Deviance será

$$D = 2(l(y, \Phi, y) - l(\hat{\mu}, \Phi, y))$$

Al maximizar la Deviance busco minimizar el error de clasificación . Φ es un parámetro de escala que por ejemplo para las distribuciones Binomial y Poisson vale 1.

Para comparar entre dos modelos anidados, se compara la bondad de ajuste de ambos. Si llamamos M_q al modelo con q componentes, M_p al modelo con p componentes, $p > q$,

entonces

$$H_0 : M_q \qquad vs. \qquad H_1 : M_p$$

tenemos que $D(y, \hat{\mu}_p) \leq D(y, \hat{\mu}_q)$ siempre. Entonces vemos como se comporta la diferencia.

$$\Delta D = D(y, \hat{\mu}_q) - D(y, \hat{\mu}_p)$$

Bajo H_0, $\Delta D \sim \chi^2_{p-q}$, entonces si ΔD es mayor que cierto cuantil de la χ^2 rechazo la hipótesis nula bajo el supuesto de que la alternativa da una mejor descripción de los datos.

Otro criterio muy usado es el **Criterio de Akaike (AIC)**, para comparar modelos que no necesariamente se encuentren anidados. EL modelo con menot AIC es el elegido. Es un modelo basado en la verosimilitud pero que penaliza por el número de parámetros

$$AIC = -2 * \log likelihood + k * p$$

En general, $k = 2$. Si $k = \log(n)$ tendremos el criterio Bayesiano BIC.

Capítulo 4

Anexos

4.1. 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,

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

$\mathbf{P}$(error tipo II) = $\mathbf{P}_\theta$(No rechazar H_0), 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:

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

$\mathbf{P}(\text{error tipo II}) = 1 - \pi_\delta(\theta)$, 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 Ho, para una observación dada. Es la probabilidad de encontrar un valor tan o mas extremo que el que se encontró con la muestra observada.

Llamaremos $\beta(\theta) = \mathbf{P}_\theta(\text{Error II}) = 1 - \pi_\delta(\theta)$, 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 $Ho : \mu = 10$) 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 Ho).Entonces rechazaría H0 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:

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

De manera que cumpla:

$$\mathbf{E}_{\mu=10}[\delta(\underline{X})] = \alpha \tag{4.1}$$

4.1.1. 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:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si} \quad \frac{f_{\theta_2}(X)}{f_{\theta_1}(X)} > k_\alpha \\ 0 & \text{en otro caso} \end{cases} \tag{4.2}$$

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

$$\mathbf{E}_{\theta_1}[\delta(\underline{X})] = \alpha = \mathbf{P}_{\theta_1}(\delta(\underline{X}) = 1) \tag{4.3}$$

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 $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ 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

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

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:

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

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
- $\pi_\delta(\mu_1) = \alpha$
- $\lim_{\mu\to\infty} \pi_\delta(\mu) = 1, \lim_{\mu\to-\infty} \pi_\delta(\mu) = 0$

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.

4.1.2. 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:

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

Que para, un nivel α dado, se tendrá:

$$\alpha = \mathbf{P}_{\theta_1}(\delta(\underline{X}) = 1) \tag{4.4}$$

(también vale la conclusión para distribución uniforme)

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

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

Será:

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

Que para, un nivel α dado, se tendrá:

$$\alpha = \mathbf{P}_{\theta_1}(\delta(\underline{X}) = 1) \tag{4.5}$$

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

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

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:

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

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

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

4.1.3. 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:

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

Donde:

$$L = \frac{\sup_{\theta\in\Theta_1} f_\theta(\underline{X})}{\sup_{\theta\in\Theta_2} f_\theta(\underline{X})} \tag{4.6}$$

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 $X_1, \ldots, X_n$ 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

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

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

Entonces,

$$\alpha = \mathbf{P}_{\mu_1}(\delta(\underline{X}) = 1) = \mathbf{P}_{\mu_1}(T < k_1 \quad o \quad T > k_2) \tag{4.7}$$

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

$$\alpha = \mathbf{P}_{\mu_1}(|\sqrt{n}\frac{\bar{X} - \mu_1}{S}| > k) \tag{4.8}$$

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.

4.1.4. 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:

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

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

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

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

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

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

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

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

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.

4.1.5. 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, $-2logL^* \sim \chi^2_j$, y el test de nivel asintótico α resulta:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si} \quad -2logL^* \geq \chi^2_{j,1-\alpha} \\ 0 & \text{si no} \end{cases}$$

4.2. Anexo 2

En la estimación puntual de parámetros nos ocupamos de dar un único valor (aproximado) para un parámetro desconocido. No podemos conocer el error que se comete al estimar el parámetro desconocido con el estimador elegido, pero podemos buscar un intervalo (aleatorio) que lo contenga con una alta probabilidad.

4.2.1. Intervalos de confianza de nivel exacto $1-\alpha$

Definición: Dada $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población con distribución $F_\theta(x), \theta \in \Theta$, un intervalo de confianza para θ de nivel $1-\alpha$ es un intervalo de extremos aleatorios $[a(\underline{X}), b(\underline{X})]$ tal que

$$\mathbf{P}(\theta \in [a(\underline{X}), b(\underline{X})]) = 1-\alpha$$

A $1-\alpha$ se lo llama nivel de confianza, y α nivel de riesgo.

Definición: Dada $X_1, X_2, \ldots, X_n$ una m.a. de una población con distribución $F_\theta(x), \theta \in \Theta$, una región de confianza para θ de nivel $1-\alpha$ es un conjunto $S(X)$ tal que

$$\mathbf{P}(\theta \in S(X)) = 1-\alpha.$$

4.2.2. Método del pivote

Sean Dada $X_1, X_2, \ldots, X_n$ una m.a. de una población con distribución $F_\theta(x), \theta \in \Theta$ y sea $U = r(X, \theta)$ una variable aleatoria cuya distribución no depente de θ. Sean A y B tales que $\mathbf{P}(A < U < B) = 1-\alpha$.

Luego la región $S(X) = \{\theta : A < r(X, \theta) < B\}$: es una región de confianza de nivel $1-\alpha$ para θ (a U se lo llama Pivote).

Encontrar intervalos de confianza es una tarea relativamente sencilla. El procedimiento más complejo será encontrar pivotes adecuados, y conocer su distribución.

Un mecanismo que suele ser útil para conseguir pivotes es el siguiente: se busca el estimador de máxima verosimilitud del parámetro en estudio, y mediante un cambio de variable se busca la distribución de dicho estimador. Esa distribución probablemente dependa del parámetro desconocido, por lo tanto el paso final será encontrar la transformación que consiga independizar a la distribución del estimador de dicho parámetro. Para las distribuciones vistas en el curso, suele ser fácil encontrar la transformación.

4.2.3. Intervalos de confianza de nivel asintótico $1-\alpha$

Definición: $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población con distribución $F_\theta(x), \theta \in \Theta$. Se dice que $S_n(X)$ es una sucesión de regiones de confianza para θ de nivel asintótico $1-\alpha$ si

$$\lim_{n\to\infty} \mathbf{P}(\theta \in S_n(X)) = 1-\alpha, \qquad \theta \in \Theta$$

Teorema: $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población con distribución $F_\theta(x), \theta \in \Theta$. Supongamos que para cada n se tiene una variable aleatoria $U_n = r_n(X, \theta)$

, donde U es una variable aleatoria cuya distribución no depende de θ. Entonces si A y B son tales que $\mathbf{P}(A < U < B) = 1 - \alpha$ se tiene que:
$S_n(X) = \{\theta : A < r_n(X, \theta) < B\}$ es una región de confianza de nivel asintótico 1–α para θ.

Lo utilizaremos para hallar intervalos de confianza basados en estimadores de máxima verosimilitud

4.2.4. Relación entre regiones de confianza y test

Sea X una variable aleatoria con distribución perteneciente a la familia $F_\theta(x)$ y que para cada θ_0 se tiene un test de nivel α, para $H_0 : \theta = \theta_0 \quad Vs. \quad H_1 : \theta \neq \theta_0$. Se puede definir una región de confianza de nivel $1 - \alpha$ para θ como

$$S(X) = \{\theta : \delta(\underline{X}) = 0\}$$

Es el conjunto de todos los $\theta \in \Theta$ tales que la hipótesis de que el valor verdadero es θ es aceptada cuando se observa $\underline{X}$, S es una región de confianza de nivel $1 - \alpha$. Esto quiere decir que para construir intervalos de confianza siempre podemos conseguirlo si "invertimos" la región de aceptación del test para hipótesis bilaterales.

Por otro lado, si se tiene una región de confianza de nivel $1 - \alpha$, $S(X)$, para θ se puede construir un test de nivel α:

$$\delta(\underline{X}) = \begin{cases} 1 & \text{si} \quad \theta_0 \notin S(X) \\ 0 & \text{si} \quad \theta_0 \in S(X) \end{cases}$$

Esto último es usado para comparaciones de parámetros, uno podrá decir que dos medias son significativamente diferentes, si el intervalo de confianza para la diferencia de medias no contiene al valor cero. No siempre puede construirse un test a partir de un intervalo de confianza, pero siempre puede construirse un intervalo de confianza a partir de un test, invirtiendo la región de aceptación.

4.3. Anexo 3

4.3.1. Método de máxima verosimilitud

Es un método para construir estimadores puntuales. Se basa en que en los experimentos aleatorios los resultados observados deben tener alta probabilidad de ocurrir.

Definición: Diremos que $\hat{\theta}(X)$ es un Estimador de Máxima Verosimilitud de θ si se cumple

$$f(X, \hat{\theta}) = \max_{\theta} f_\theta(x)$$

Es decir, buscamos el valor de θ que hace que la función de verosimilitud $L(\theta)$ sea máxima.
$\hat{\theta}_n = arg\,\text{máx}_\theta\, L(\theta)$

Si Θ es un subconjunto abierto tal que el soporte de $f_\theta(x)$ no depende de θ (familia regular), como la función logaritmo es monótona creciente, maximizar f es lo mismo que maximizar log(f). Luego el EMV debe verificar: $\frac{d}{d\theta} logL(\theta)$.
Si observamos la estructura de las familias exponenciales, basta con hallar la función que maximiza el exponente de la exponencial.

Principio de invariancia: Supongamos $\lambda = q(\theta)$ es una función biunívoca de θ, si $\hat{\theta}$ es el EMV de tita entonces $\hat{\lambda} = q(\hat{\theta})$ es el EMV de λ

Uno de los grandes problemas de la estadística es construir estimadores razonables para el parámetro desconocido θ. Buscaremos entre ellos ciertas propiedades deseables para la estimación.

Bondad de los estimadores

$X_1, X_2, \ldots, X_n \sim F_\theta(x), \theta \in \Theta$ una muestra aleatoria. Estimamos θ por $\hat{\theta}$. Llamamos función de riesgo (error cuadrático medio):

$$R(\theta, \hat{\theta}) = ECM(\hat{\theta}) = \mathbf{E}_\theta[(\theta - \hat{\theta})^2]$$

Un estimador óptimo para θ será $\hat{\theta}^*$, tal que:

$$ECM(\hat{\theta}^*) \leq ECM(\hat{\theta}), \forall \hat{\theta}$$

Definición: Se dice que $\hat{\theta}$ es un estimador **insesgado** para θ si

$$\mathbf{E}_\theta[\hat{\theta}] = \theta,\ \forall \theta \in \Theta$$

En caso contrario, el estimador será sesgado, y definimos su sesgo como

$$B(\hat{\theta}) = \mathbf{E}_\theta[\hat{\theta}] - \theta.$$

Propiedad: dado un estimador de θ , se tiene que

$$ECM(\hat{\theta}) = \mathbf{var}_\theta(\hat{\theta}) + B(\hat{\theta})^2$$

Observadión: Si B=0 entonces $ECM(\hat{\theta}) = \mathbf{var}(\hat{\theta})$.

Estimador asintóticamente insesgado

$$lim_{n\to\infty}\mathbf{E}_\theta[\hat{\theta}] = \theta$$

Consistencia en media cuadrática

$$lim_{n\to\infty}ECM[\hat{\theta}] = 0, \forall\theta \in \Theta$$

Definición: Dada una sucesión de estimadores $\hat{\theta}_n$ de θ, decimos que $T = \hat{\theta}$ es (debilmente) consistente si para todo $\delta > 0$, $\mathbf{P}_\theta(|T - \theta| > \delta) \to 0$

(Lo mínimo que puede pedirse a un estimador es que se aproxime al verdadero valor del parámetro a medida que el tamaño de muestra aumenta)

Teorema: Sea una sucesión de estimadores $\hat{\theta}_n$ de θ. Si $\mathbf{var}(\hat{\theta}) \to 0$ y $\mathbf{E}_\theta(\hat{\theta}) \to \theta$ entonces $\hat{\theta}_n$ es debilmente consistente.

Demostración. Usando Markov:

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}[h(X)]}{h(t)}$$

$$\mathbf{P}_\theta(|\hat{\theta} - \theta| > \epsilon) \leq \frac{\mathbf{E}[(\hat{\theta} - \theta)^2]}{\epsilon^2} = \frac{\mathbf{var}(\hat{\theta}) + (\mathbf{E}[\hat{\theta}] - \theta)^2}{\epsilon^2} \to 0$$

Observación: $h(y) = y^2, \hat{\theta} - \theta = X, \epsilon = t$.

Una vez hallados los estimadores, vamos a querer conocer su distribución. A continuación algunas definiciones que nos servirán para encontrarlas.

4.3.2. Estimadores asintóticamente normales

Se dice que $\hat{\theta}_n$ es una sucesión de estimadores asintóticamente normal si:

$$\sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \sim \mathcal{N}(0, 1)$$

Donde $I(\theta)$ es el número de iniformación de fisher y se calcula como:

$$I(\theta) = \mathbf{E}\left[\left(\frac{d}{d\theta}lnf_\theta(X)\right)^2\right] = -\mathbf{E}\left[\frac{d^2}{d\theta^2}lnf_\theta(X)\right]$$

Propiedad: Sea $\hat{\theta}$ el EMV de θ entonces es asintóticamente normal.

Propiedad: Si $\sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \sim \mathcal{N}(0, 1)$ y $\hat{\theta}$ es un estimador consistente para θ, entonces vale que:

$$\sqrt{nI(\hat{\theta})}(\hat{\theta}_n - \theta) \sim \mathcal{N}(0, 1)$$

Método Delta **Lema:** Sea $X_1, \dots X_n$ una sucesión de variables aleatorias tal que:

$$\sqrt{n}(X_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Sea g una función continua y derivable tal que $g'(\mu) \neq 0$ y $g'(\mu)$ continua en μ, entonces:

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{D} \mathcal{N}(0, g'(\mu)^2\sigma^2)$$

Capítulo 5

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