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.

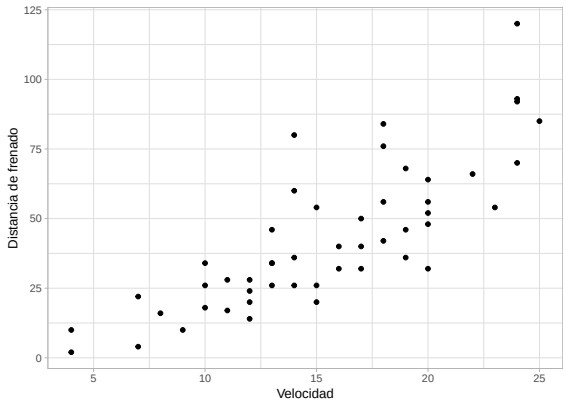

Figura 2.1: Gráfico de dispersión (Scatterplot)

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.

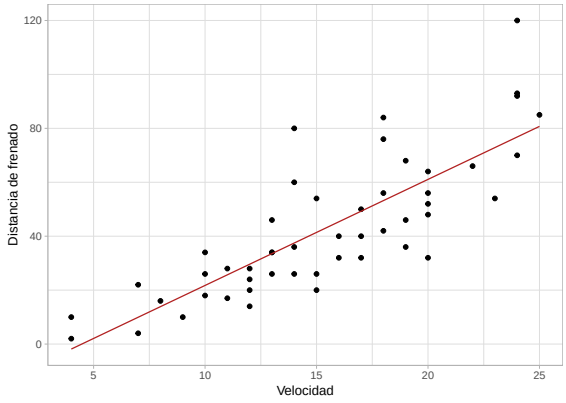

Figura 2.2: Estimación por cuadrados mínimos

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 $\hat{\eta}$.

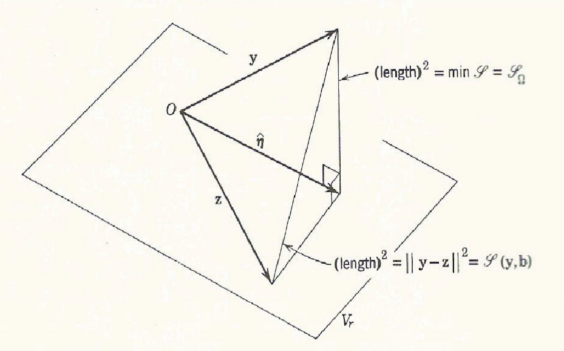

Figura 2.3: Interpretación gromética

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

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

como ya lo habíamos visto. Entonces

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

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

El residuo será

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

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 resifuos 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.

CONTINUARÁ...

Capítulo 3

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