Índice general

Capítulo 1

Modelo lineal

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

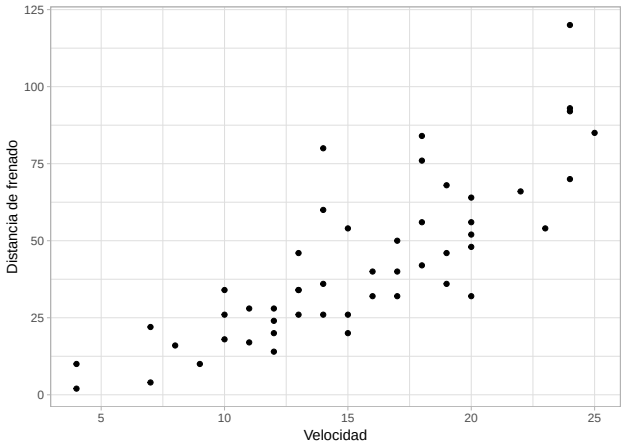

Figura 1.1: Gráfico de dispersión (Scatterplot)

En la Figura 1.1 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.

1.1. 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 1.1 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 $X = x$.

1.2. Un poco de teoría

1.2.1. Momentos para 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. 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:
$Y = (y_1, \ldots, y_n)^T$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$, $\beta = (\beta_1, \ldots, \beta_n)^T$,

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

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.

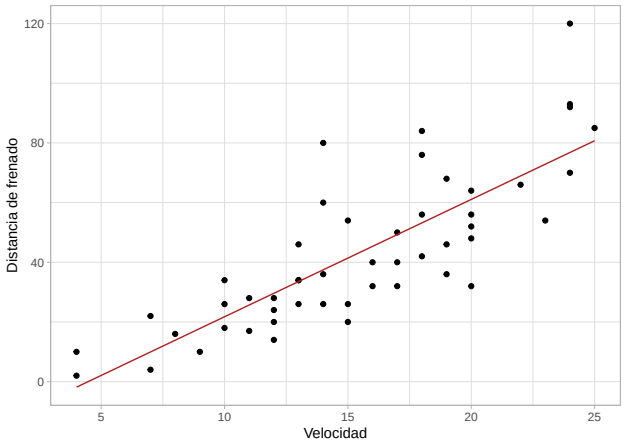

Figura 1.2: Estimación por cuadrados mínimos

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

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

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.

1.3.1. 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 $\mathbf{X}: x^1, \ldots, x^p$. Esto significa que η pertenece al subespacio generado por las columnas de $\mathbf{X}$, 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 $\hat{\eta}$.

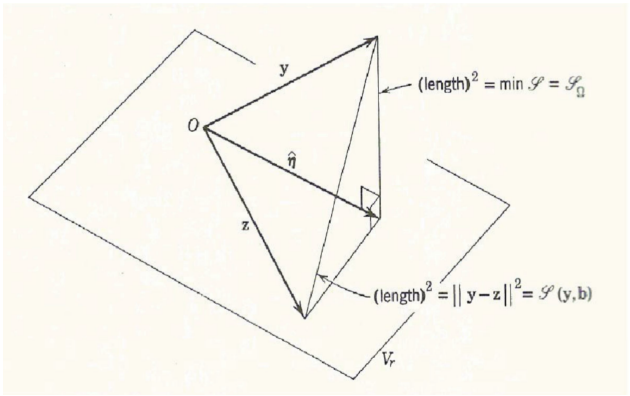

Figura 1.3: Interpretación grométrica

Esta proyección siempre existe y es única (aunque no así los b). Entonces $\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T Y$. 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 1.3, 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

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

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

1.4. 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.