

Modelo Lineal

IECD

Modelización y Predicción

Algunas de los métodos estadísticos más extendidos se ocupan de la modelización de datos y de la predicción.

Muchas de estas técnicas estadísticas se encuadran en lo que hoy se conoce como **aprendizaje estadístico** (AE).

El AE abarca una amplia cantidad de técnicas que ayudan a comprender los datos cuando se analizan varias variables al mismo tiempo, ya sea postulando modelos o encontrando relaciones entre las variables o estructuras que ayudan a su comprensión.

Modelización y Predicción

Algunas de los métodos estadísticos más extendidos se ocupan de la modelización de datos y de la predicción.

Muchas de estas técnicas estadísticas se encuadran en lo que hoy se conoce como **aprendizaje estadístico** (AE).

El AE abarca una amplia cantidad de técnicas que ayudan a comprender los datos cuando se analizan varias variables al mismo tiempo, ya sea postulando modelos o encontrando relaciones entre las variables o estructuras que ayudan a su comprensión.

Los métodos de AE pueden reunirse en dos grandes grupos:

- **Aprendizaje Supervisado**: Aquí una de las variables es identificada como una respuesta.
- **Aprendizaje No Supervisado**: todas las variables cumplen un rol análogo.

Aprendizaje Estadístico

Algunos ejemplos de aprendizaje estadístico:

- Predecir si un paciente hospitalizado tendrá un segundo infarto de miocardio o no teniendo en cuenta mediciones clínicas, dietas y variables demográficas.
- Identificar los factores de riesgo de cáncer de próstata, usando mediciones clínicas y variables demográficas.
- Predecir los precios que tendrán en 6 meses las acciones de ciertas compañías a partir de mediciones del rendimiento de las compañías y datos macroeconómicos.
- Estimar la cantidad de glucosa en sangre que tendrá un individuo diabético a partir del espectro de adsorción infra-rojo de la sangre.

Los dos primeros problemas están en el campo de lo que conocemos como clasificación, los dos últimos son un problema de regresión.

Regresión: Galton

ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.

By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

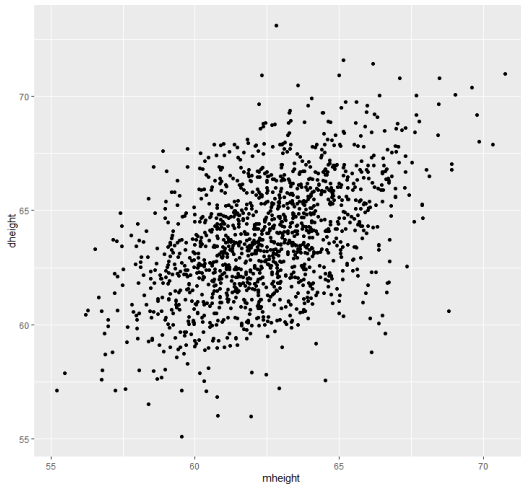
Ejemplo: alturas de madres e hijas

Pearson organizó la recolección de datos de 1100 familias en Inglaterra en el período 1893 a 1898. El paquete `alr4` de R contiene el conjunto de datos `Heights` que corresponde a la altura en pulgadas de madres e hijas (con hasta dos hijas por madre).

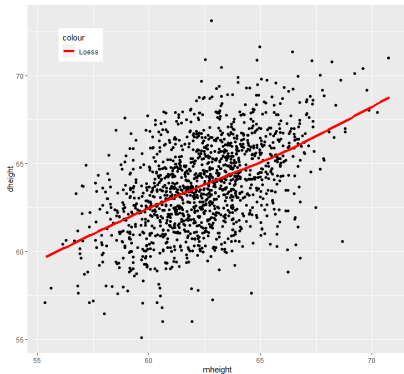
Las hijas tienen por lo menos 18 años y todas las madres son menores de 65 años.

Originalmente los datos aparecen en una tabla redondeados a la pulgada más cercana. Sin embargo, la librería de R les agrega un error de redondeo para que el gráfico no sea discreto.

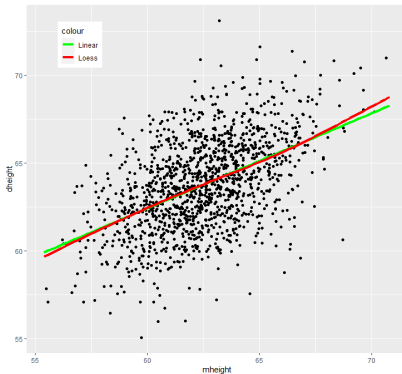
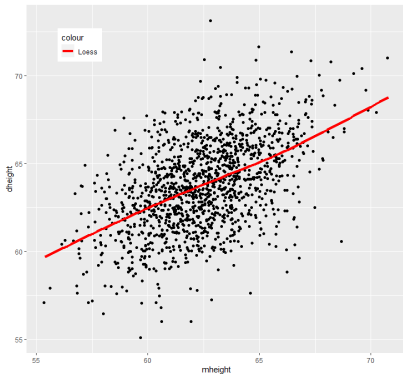
Ejemplo: alturas de madres e hijas



Ejemplo: alturas de madres e hijas



Ejemplo: alturas de madres e hijas



¿Cuánto medirá de grande?



¿Cuánto medirá de grande?



¿Qué información adicional tenemos?

Escenario 1: ninguna

Escenario 2: sexo: será niña

sabemos algo de la familia....

Escenario 3 : será niña y la madre mide 156 cm.

Escenario 4 : será niña, la madre mide 156 cm. y el padre 178 cm.

Escenario 1: Mejor predictor constante

Queremos predecir cuánto medirá le hije de grande.

¿Cuál es la mejor predicción que podemos hacer? Sin información adicional, seguramente nuestra predicción será mediante una constante c . Queremos que la diferencia entre Y y c , digamos $(Y - c)^2$ sea pequeña, pero recordemos que es una v.a.

Podemos resumirla en un número, por ej., mediante el valor esperado:

Error cuadrático medio de predicción de c :

$$ECM(Y, c) = \mathbb{E}[(Y - c)^2]$$

¿Cuál es el mejor predictor constante?

Si Y es una va.a. con $\mathbb{E}(Y^2) < \infty$, entonces $\mu = \mathbb{E}(Y)$ es el mejor predictor constante en términos del error cuadrático medio (de predicción), en tanto para toda $c \in \mathbb{R}$

$$ECM(Y, \mu) \leq ECM(Y, c)$$

¿Cuál es el mejor predictor constante?

Notemos además que

$$\mathbb{E}[(Y - c)^2] = \mathbb{E}[(Y - \mu)^2] + (\mu - c)^2 = \mathbb{V}ar(Y) + (\mathbb{E}[Y] - c)^2$$

Por lo tanto, como $ECM(Y, \mu) = \mathbb{V}ar(Y)$, podemos pensar que éste es el precio que pagamos por la predecir a la v.a. Y por una constante.

Otros escenarios

¿Cuál es el mejor predictor de Y basado en X ?

En los otros escenarios tenemos información adicional dada a través de X , otra(s) variables(s).

Conociendo X ahora queremos predecir a Y , sabemos que será niña o sabemos la altura de la madre o ambas cosas, por ejemplo.

En este caso nuestra predicción será una función de X , digamos $m(X)$ y busquemos ahora que

$$ECM(Y, m(X)) = \mathbb{E} [(Y - m(X))^2]$$

sea tan pequeño como sea posible.

Otros escenarios

¿Cuál es el mejor predictor de Y basado en X ?

En los otros escenarios tenemos información adicional dada a través de X , otra(s) variables(s).

Conociendo X ahora queremos predecir a Y , sabemos que será niña o sabemos la altura de la madre o ambas cosas, por ejemplo.

En este caso nuestra predicción será una función de X , digamos $m(X)$ y busquemos ahora que

$$ECM(Y, m(X)) = \mathbb{E} [(Y - m(X))^2]$$

sea tan pequeño como sea posible.

Debemos elegir, entonces a m como **la esperanza condicional**.

En efecto, si Y es una v.a. con $\mathbb{E}(Y^2) < \infty$, entonces $\tilde{m}(X) = \mathbb{E}(Y | X)$ es el mejor predictor basado en X , es decir es la mejor aproximación en términos del error cuadrático medio en tanto para toda $m(X)$ con esperanza cuadrada finita tenemos que

$$ECM(Y, \tilde{m}(X)) \leq ECM(Y, m(X))$$

¿Cuál es el mejor predictor basado en X ?

Este es un resultado de Probabilidades que recordaremos:

$$\begin{aligned}\mathbb{E}[(Y - m(X))^2] &= \mathbb{E}[(Y - \tilde{m}(X) + \tilde{m}(X) - m(X))^2] \\ &= \mathbb{E}[(Y - \tilde{m}(X))^2 + (\tilde{m}(X) - m(X))^2 \\ &\quad + 2\mathbb{E}[(Y - \tilde{m}(X))(\tilde{m}(X) - m(X))]]\end{aligned}$$

Veamos que el tercer sumando se anula:

$$\begin{aligned}\underbrace{\mathbb{E}[(Y - \tilde{m}(X))]}_{u(X,Y)} \underbrace{(\tilde{m}(X) - m(X))}_{v(X)} &= E[u(X, Y)v(X)] \\ &= \mathbb{E}[\mathbb{E}[u(X, Y)v(X) \mid X]] \\ &= \mathbb{E}[v(X)E[u(X, Y) \mid X]]\end{aligned}$$

Como $\mathbb{E}[u(X, Y) \mid X] = E[Y - \tilde{m}(X) \mid X] = \mathbb{E}[Y \mid X] - \mathbb{E}[\tilde{m}(X) \mid X]$
 $= \tilde{m}(X) - \tilde{m}(X) = 0$

¿Cuál es el mejor predictor basado en X ?

Luego, como el primer sumando no depende de $m(X)$, tenemos que minimizar a $ECM(Y, m(X))$ equivale a minimizar a

$$\mathbb{E} \left[(\tilde{m}(X) - m(X))^2 \right]$$

y eso se logra tomando

$$m(X) = \tilde{m}(X) = \mathbb{E}(Y \mid X)$$

Ahora bien, si quisiéramos calcular a $\tilde{m}(X) = \mathbb{E}(Y \mid X)$ necesitaríamos la distribución conjunta de (Y, X) y en la práctica esa información, en general, no está disponible.

Usualmente, tomamos un atajo haciendo un modelado e intentamos responder, por ejemplo, cuál es la función lineal en X que mejor aproxima a Y .

¿Cuál es el mejor predictor lineal de Y basado en X ?

En este caso buscamos la mejor aproximación dentro de la familia de funciones de la forma:

$$m(X) = \theta_0 + \theta_1 X$$

Buscamos los valores $\tilde{\theta}_0$ y $\tilde{\theta}_1$ que minimizan

$$ECM(Y, \theta_0 + \theta_1 X) = \mathbb{E} \left[(Y - (\theta_0 + \theta_1 X))^2 \right]$$

Es un ejercicio de la práctica comprobar, derivando e igualando a 0, que:

- $\tilde{\theta}_0 = \mathbb{E}(Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \mathbb{E}(X)$
- $\tilde{\theta}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$

En otras palabras, el mejor predictor lineal es de la forma:

$$m_L(X) = \mathbb{E}(Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \mathbb{E}(X) + \frac{\text{Cov}(X, Y)}{\text{Var}(X)} X$$

¿Cuál es el mejor predictor lineal de Y basado en X ?

Notemos que

- $m_L(\mathbb{E}(X)) = \mathbb{E}(Y)$
- En términos absolutos, la pendiente aumenta con la $\mathbb{Cov}(X, Y)$, que mide la asociación lineal entre las variables: $|\tilde{\theta}_1|$ aumenta con $|\mathbb{Cov}(X, Y)|$.
- En términos absolutos, la pendiente decrece a 0 con la $\mathbb{Var}(X)$, que mide la variabilidad de X : cuanto más dispersión en X , más se acerca $|\tilde{\theta}_1|$ a 0.
- La aproximación lineal óptima obtenida no asume ninguna relación particular entre X e Y . Es de esperar que sea buena cuando se vinculan aproximadamente en forma lineal, pero puede ser muy mala si la relación entre ellas está muy lejos de la linealidad.

Otro ejemplo

En 97 pacientes que van a ser tratados con una prostatectomía radical se miden las siguientes variables:

- $x_1 = \text{**lweight**}$: log del peso de la próstata
- $x_2 = \text{**age**}$: edad
- $x_3 = \text{**lbph**}$: log de la cantidad de hiperplasia prostática benigna
- $x_4 = \text{**svi**}$: invasión seminal (si o no)
- $x_5 = \text{**lcp**}$: logaritmo de la penetración capsular
- $x_6 = \text{**gleason**}$: score de Gleason
- $x_7 = \text{**pgg46**}$: porcentaje de scores de Gleason 4 or 5.
- $x_8 = \text{**lpsa**}$: log de PSA

El objetivo es poder predecir el logaritmo del volumen del tumor ($y = \text{**lcavol**}$).

Diagrama de Dispersión

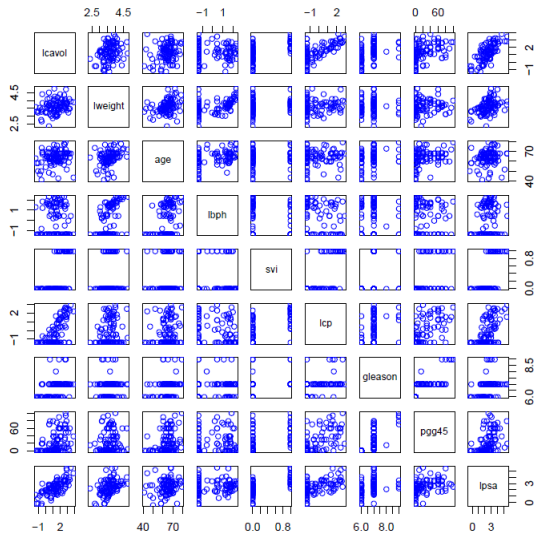
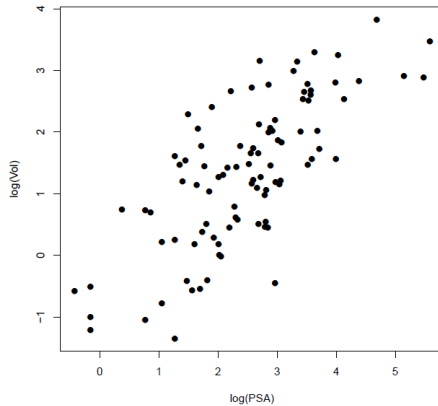


Diagrama de Dispersión



Modelos de regresión

Buscamos un modelo que exprese a la variable de respuesta en términos de las otras variables presentes (covariables).

Cuando hablamos de un modelo nos referimos a una expresión matemática que:

- sea válida aproximadamente
- que describa en algún sentido el comportamiento de la variable de interés en función de las demás variables predictoras.

En un modelo de regresión se postularía:

$$y = m(x_1, x_2, x_3, \dots, x_p) + \epsilon$$

o en general si $\mathbf{x} = (x_1, x_2, x_3, \dots, x_p)$

$$y = m(\mathbf{x}) + \epsilon$$

Modelos de regresión paramétrica

$$y = m(\mathbf{x}) + \epsilon$$

Las posibles funciones de regresión m pertenecen a una clase \mathcal{M} tan grande que es frecuente que se simplifique el problema suponiendo cierta forma o ciertas propiedades de la función de regresión m .

Una forma de simplificar el problema suponiendo que la familia \mathcal{M} puede expresarse en función de un número finito de constantes desconocidas, a estimar, llamadas **parámetros**, que controlan el comportamiento del modelo. En este sentido diremos que el **modelo de regresión es paramétrico**.

Modelos de regresión paramétrica

Para simplificar, pensemos que tenemos dos covariables x_1 y x_2 que son determinísticas.

Modelos paramétricos

$$(i) \quad y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon$$

$$(ii) \quad y = \alpha \log(x_1) + \beta \log(x_2) + \gamma x_1^3 + \delta \sin(x_2) + \varepsilon$$

$$(iii) \quad y = \alpha x_1^\beta x_2^\delta + \varepsilon$$

$$(iv) \quad y = \alpha e^{\beta x_1} + \gamma e^{\delta x_2} + \varepsilon$$

Modelo Lineal

Uno de los modelos más sencillos es el **modelo lineal**, en el que los parámetros intervienen como simples coeficientes de las variables independientes o de funciones de éstas.

Es el caso de:

(i) $y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon$

(ii) $y = \alpha \log(x_1) + \beta \log(x_2) + \gamma x_1^3 + \delta \sin(x_2) + \varepsilon$

Modelo Lineal

En estos dos ejemplos $m(x)$ es **lineal** en los **parámetros**. No es el caso, por ejemplo, de $m(x) = \alpha e^{-\beta x}$, conocido como crecimiento exponencial, ya que no es lineal como función de los parámetros α o β .

Algunos ejemplos sencillos de modelos lineales dependientes de una sola variable son:

$$m(x) = \alpha + \beta x$$

$$m(x) = \alpha + \beta x + \gamma x^2$$

$$m(x) = \alpha + \beta \log(x)$$

Modelo Lineal

En general, en un modelo lineal tendremos que la i -ésima respuesta y_i está asociada a un vector de covariables $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ de la siguiente forma

$$y_i = \theta_1 x_{i1} + \dots + \theta_p x_{ip} + \epsilon_i .$$

es decir

$$m(\mathbf{x}_i) = \theta_1 x_{i1} + \dots + \theta_p x_{ip} = \boldsymbol{\theta}^t \mathbf{x}_i$$

siendo $\boldsymbol{\theta}^t = (\theta_1, \dots, \theta_p)$.

Modelo Lineal

En general, en un modelo lineal tendremos que la i -ésima respuesta y_i está asociada a un vector de covariables $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ de la siguiente forma

$$y_i = \theta_1 x_{i1} + \dots + \theta_p x_{ip} + \epsilon_i.$$

es decir

$$m(\mathbf{x}_i) = \theta_1 x_{i1} + \dots + \theta_p x_{ip} = \boldsymbol{\theta}^t \mathbf{x}_i$$

siendo $\boldsymbol{\theta}^t = (\theta_1, \dots, \theta_p)$.

Eventualmente

- las covariables podrían ser funciones de otras variables, tal como ocurre en el caso ii) y en nuestro ejemplo.
- $x_{i1} = 1$ para todo i , en este caso decimos que el modelo tiene intercept u ordenada al origen.

Modelo Lineal

Supondremos $x_j, 1 \leq j \leq p$ determinísticas.

Muestra $(x_{i1}, \dots, x_{ip}, y_i), 1 \leq i \leq n$ que cumplen el modelo Ω :

$$\begin{aligned}y_i &= \theta_1 x_{i1} + \dots + \theta_p x_{ip} + \epsilon_i \quad i = 1, \dots, n \\ \mathbb{E}(\epsilon_i) &= 0 \\ \mathbb{V}(\epsilon_i) &= \sigma^2 \\ \text{Cov}(\epsilon_i, \epsilon_j) &= 0 \quad i \neq j\end{aligned}$$

donde, $\theta_1, \dots, \theta_p$ son p parámetros desconocidos a estimar.

Enfoque matricial

En forma matricial plantearíamos

$$y_i = \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \epsilon_i \quad i = 1, \dots, n$$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & & \cdots & \\ \cdots & & \cdots & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

\Downarrow

$$\begin{array}{ccccccc} \mathbf{Y} & = & \mathbf{X} & \boldsymbol{\theta} & + & \boldsymbol{\epsilon} \\ n \times 1 & & n \times p & p \times 1 & & n \times 1 \end{array}$$

La matriz $\mathbf{X} \in \mathbb{R}^{n \times p}$ recibe el nombre de **matriz de regresión** o de **diseño**.

En general, se elige de tal forma que tenga rango máximo, es decir $\text{rg}(\mathbf{X}) = p$, sin embargo esto no siempre es posible, como en el caso de algunos diseños tratados en análisis de la varianza (ANOVA).

Trataremos el caso de rango completo.

La teoría que veremos no necesita que la primera columna sea de 1's, es decir que el modelo tenga intercept, por lo tanto estudiaremos el caso general.

Algunos ejemplos: Modelo de regresión simple

En el caso más sencillo de regresión simple tendríamos

$$y_i = \theta_1 + \theta_2 x_i + \epsilon_i \quad 1 \leq i \leq n$$

$$p = 2 \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

Modelo de posición

$$y_i = \mu + \epsilon_i \quad 1 \leq i \leq n$$

$$p = 1 \quad \mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ 1 \end{pmatrix} \quad \theta = \mu$$

Modelo de 2 muestras

$$y_{i1} = \mu_1 + \epsilon_{i1} \quad 1 \leq i \leq n_1$$

$$y_{i2} = \mu_2 + \epsilon_{i2} \quad 1 \leq i \leq n_2$$

$$p = 2 \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ 0 & 1 \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

Modelo de regresión cuadrática

$$y_i = \theta_1 + \theta_2 x_i + \theta_3 x_i^2 + \epsilon_i \quad 1 \leq i \leq n$$

$$p = 3 \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & x_n & x_n^2 \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}$$

Propiedades de vectores y matrices aleatorias

Dada una matriz \mathbf{V} ($r \times s$) de variables aleatorias conjuntamente distribuidas $\{V_{ij}\}$ con esperanza finita, definimos la matriz o vector de esperanzas como:

$$\{\mathbb{E}(\mathbf{V})\}_{ij} = \mathbb{E}(V_{ij})$$

En el caso del modelo Ω , esto nos permite decir que el vector de errores es tal que

$$\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$$

y que

$$\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t) = \mathbb{E} \begin{pmatrix} \epsilon_1\epsilon_1 & \epsilon_1\epsilon_2 & \dots & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2\epsilon_2 & \dots & \epsilon_2\epsilon_n \\ \dots & & \dots & \\ \dots & & \dots & \\ \epsilon_n\epsilon_1 & \epsilon_n\epsilon_2 & \dots & \epsilon_n\epsilon_n \end{pmatrix} = \sigma^2 \mathbf{I}$$

Lema 1: Sean $\mathbf{A} \in \mathbb{R}^{q \times r}$, $\mathbf{B} \in \mathbb{R}^{s \times t}$ y $\mathbf{C} \in \mathbb{R}^{q \times t}$ matrices constantes y \mathbf{V} una matriz aleatoria de dimensión $r \times s$, entonces:

$$\mathbb{E}(\mathbf{A}\mathbf{V}\mathbf{B} + \mathbf{C}) = \mathbf{A}\mathbb{E}(\mathbf{V})\mathbf{B} + \mathbf{C}.$$

Matriz de Covarianza

Sea $\mathbf{v} = (v_1, \dots, v_n)^t$ un vector aleatorio de variables con $E(v_i) = \mu_i$ y varianza finita. Definimos la matriz de covarianza de \mathbf{v} como:

$$\{\Sigma_{\mathbf{v}}\}_{ij} = \text{Cov}(v_i, v_j) = \mathbb{E}[(v_i - \mu_i)(v_j - \mu_j)]$$

Podemos escribirla como:

$$\Sigma_{\mathbf{v}} = \mathbb{E}[(\mathbf{v} - \boldsymbol{\mu})(\mathbf{v} - \boldsymbol{\mu})^t]$$

donde $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^t$.

En este sentido, como $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, entonces hemos visto que

$$\Sigma_{\boldsymbol{\epsilon}} = \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t) = \sigma^2 \mathbf{I}$$

Usaremos el siguiente resultado basado en las propiedades de linealidad de la esperanza:

Lema 2: Sean $\mathbf{A} \in \mathbb{R}^{m \times n}$, una matriz constante, \mathbf{d} un vector de constantes y \mathbf{v} un vector aleatorio n -dimensional con esperanza $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}_{\mathbf{v}}$. Si $\mathbf{w} = \mathbf{A}\mathbf{v} + \mathbf{d}$, entonces:

$$\boldsymbol{\Sigma}_{\mathbf{w}} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{v}}\mathbf{A}^t.$$

El modelo que presentamos más arriba puede escribirse como:

$$\Omega : \mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0} \quad \Sigma_{\boldsymbol{\epsilon}} = \sigma^2 \mathbf{I}$$

o equivalentemente

$$\Omega : \mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta} \quad \Sigma_{\mathbf{Y}} = \sigma^2 \mathbf{I}$$

El modelo lineal: Interpretación Geométrica

Supondremos que $n \geq p$. Podemos pensar a la matriz de diseño de dos maneras:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^t \\ \vdots \\ \mathbf{x}_n^t \end{pmatrix} = \left(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)} \right)$$

Sea \mathcal{V} el subespacio de \mathbb{R}^n generado por las columnas de \mathbf{X} .

$$\dim(\mathcal{V}) = \text{rango}(\mathbf{X})$$

El modelo lineal: Interpretación Geométrica

Supondremos que $n \geq p$. Podemos pensar a la matriz de diseño de dos maneras:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^t \\ \vdots \\ \mathbf{x}_n^t \end{pmatrix} = \left(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)} \right)$$

Sea \mathcal{V} el subespacio de \mathbb{R}^n generado por las columnas de \mathbf{X} .

$$\dim(\mathcal{V}) = \text{rango}(\mathbf{X})$$

El modelo supuesto

$$\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta} = \theta_1 \mathbf{x}^{(1)} + \dots + \theta_p \mathbf{x}^{(p)}$$

El modelo lineal: Interpretación Geométrica

Supondremos que $n \geq p$. Podemos pensar a la matriz de diseño de dos maneras:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^t \\ \vdots \\ \mathbf{x}_n^t \end{pmatrix} = \left(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)} \right)$$

Sea \mathcal{V} el subespacio de \mathbb{R}^n generado por las columnas de \mathbf{X} .

$$\dim(\mathcal{V}) = \text{rango}(\mathbf{X})$$

El modelo supuesto

$$\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta} = \theta_1 \mathbf{x}^{(1)} + \dots + \theta_p \mathbf{x}^{(p)}$$

es decir que $\mathbb{E}(\mathbf{Y})$ es una combinación lineal de las columnas de \mathbf{X} o sea

$$\mathbb{E}(\mathbf{Y}) \in \mathcal{V}$$

El modelo lineal: Interpretación Geométrica

- Si $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)}$ son linealmente independientes o sea, si $\dim(\mathcal{V}) = \text{rango}(\mathbf{X}) = p$, existe un **único** $\boldsymbol{\theta}$ tal que $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}$.

El modelo lineal: Interpretación Geométrica

- Si $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)}$ son linealmente independientes o sea, si $\dim(\mathcal{V}) = \text{rango}(\mathbf{X}) = p$, existe un **único** $\boldsymbol{\theta}$ tal que $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}$.
- Si $\dim(\mathcal{V}) = q < p$, $\boldsymbol{\theta}$ **no está determinado** aunque $\mathbb{E}(\mathbf{Y})$ sí lo está. En este caso, tenemos dos opciones

El modelo lineal: Interpretación Geométrica

- Si $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)}$ son linealmente independientes o sea, si $\dim(\mathcal{V}) = \text{rango}(\mathbf{X}) = p$, existe un **único** $\boldsymbol{\theta}$ tal que $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}$.
- Si $\dim(\mathcal{V}) = q < p$, $\boldsymbol{\theta}$ **no está determinado** aunque $\mathbb{E}(\mathbf{Y})$ sí lo está. En este caso, tenemos dos opciones
 - Si $\dim(\mathcal{V}) = q$ y $\mathbf{x}^{(j_1)}, \dots, \mathbf{x}^{(j_q)}$ generan \mathcal{V} , tenemos que

$$\mathbb{E}(\mathbf{Y}) = \sum_{s=1}^q \tilde{\theta}_s \mathbf{x}^{(j_s)}$$

y el vector $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \dots, \tilde{\theta}_q)^t$ está bien determinado.

El modelo lineal: Interpretación Geométrica

- Si $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)}$ son linealmente independientes o sea, si $\dim(\mathcal{V}) = \text{rango}(\mathbf{X}) = p$, existe un **único** $\boldsymbol{\theta}$ tal que $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}$.
- Si $\dim(\mathcal{V}) = q < p$, $\boldsymbol{\theta}$ **no está determinado** aunque $\mathbb{E}(\mathbf{Y})$ sí lo está. En este caso, tenemos dos opciones
 - Si $\dim(\mathcal{V}) = q$ y $\mathbf{x}^{(j_1)}, \dots, \mathbf{x}^{(j_q)}$ generan \mathcal{V} , tenemos que

$$\mathbb{E}(\mathbf{Y}) = \sum_{s=1}^q \tilde{\theta}_s \mathbf{x}^{(j_s)}$$

y el vector $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \dots, \tilde{\theta}_q)^t$ está bien determinado.

- La otra opción es **imponer restricciones lineales a $\boldsymbol{\theta}$** mediante

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{0}$$

con \mathbf{A} una matriz elegida para que el vector quede unívocamente determinado.