Modelo Lineal

Estadística (M)

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).

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 vasta 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 vasta 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:

Aprendizaje Estadístico

Algunos ejemplos de aprendizaje estadístico:

- Predecir si un paciente hospitalizado tendrá un segundo infarto de miocardio o no teniendo en cuenta mediciones clínica, dietas 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 económicos.
- Estimar la cantidad de glucosa en sangre que tendrá un individuo diabético a partir del espectro de adsorción infra-rojo de la sangre.
- Identificar los factores de riesgo de cáncer de próstata, usando mediciones clínicas y variables demográficas.

Vamos a considerar el último ejemplo:

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

- $x_1 =$ **lweight**: log del peso de la próstata
- $x_2 =$ **age**: edad
- $xX_3 =$ **lbph**: log de la cantidad de hiperplasia prostática benigna
- $x_4 =$ **svi**: invasión seminal (si o no)
- $x_5 =$ **lcp**: logaritmo de la penetración calsular
- $x_6 =$ **gleason**: score de Gleason
- $x_7 =$ **pgg46**: porcentaje de scores de Gleason 4 or 5.
- $x_8 =$ **lpsa**: log de PSA

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

Diagramas de Dispersión

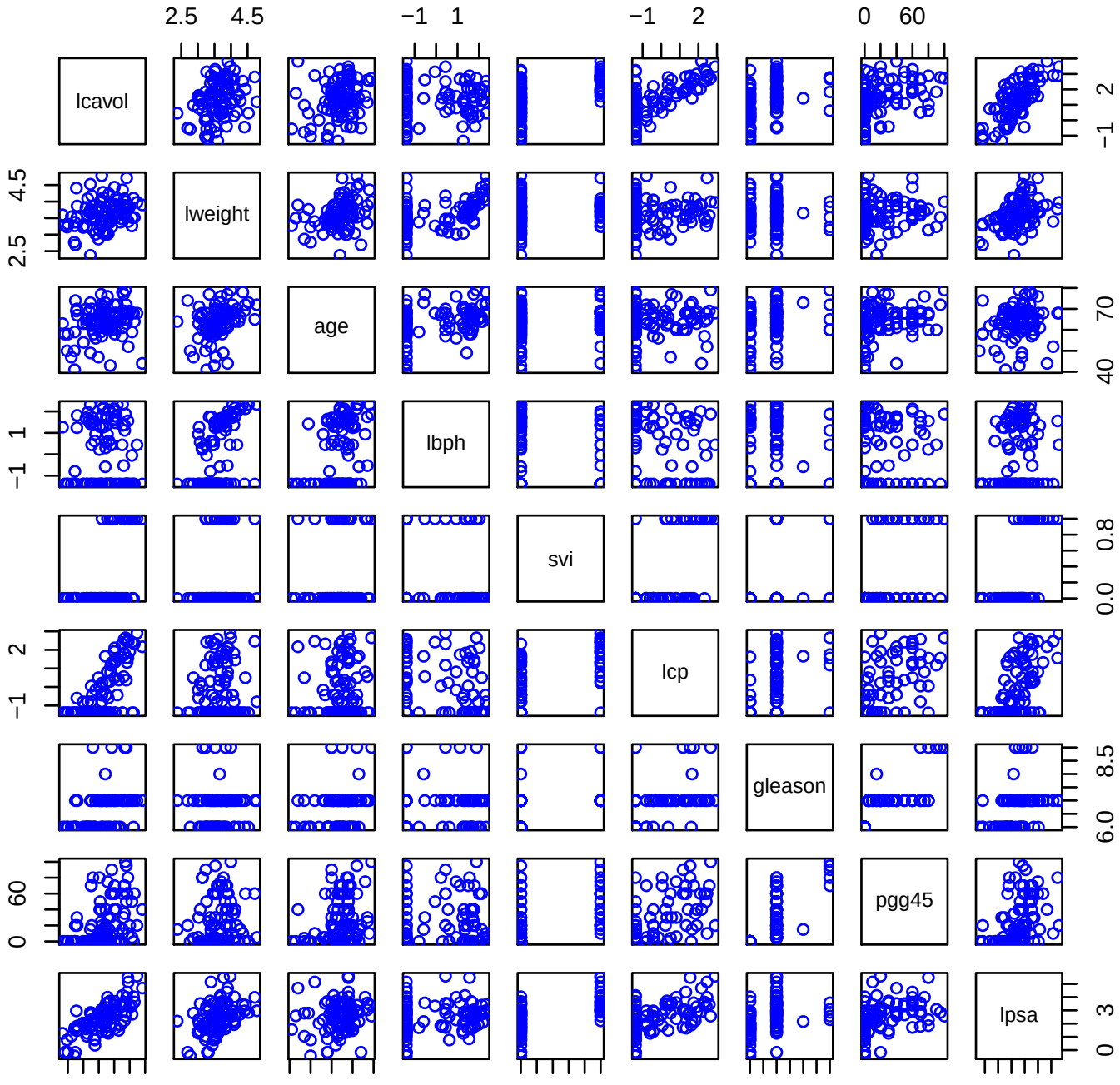
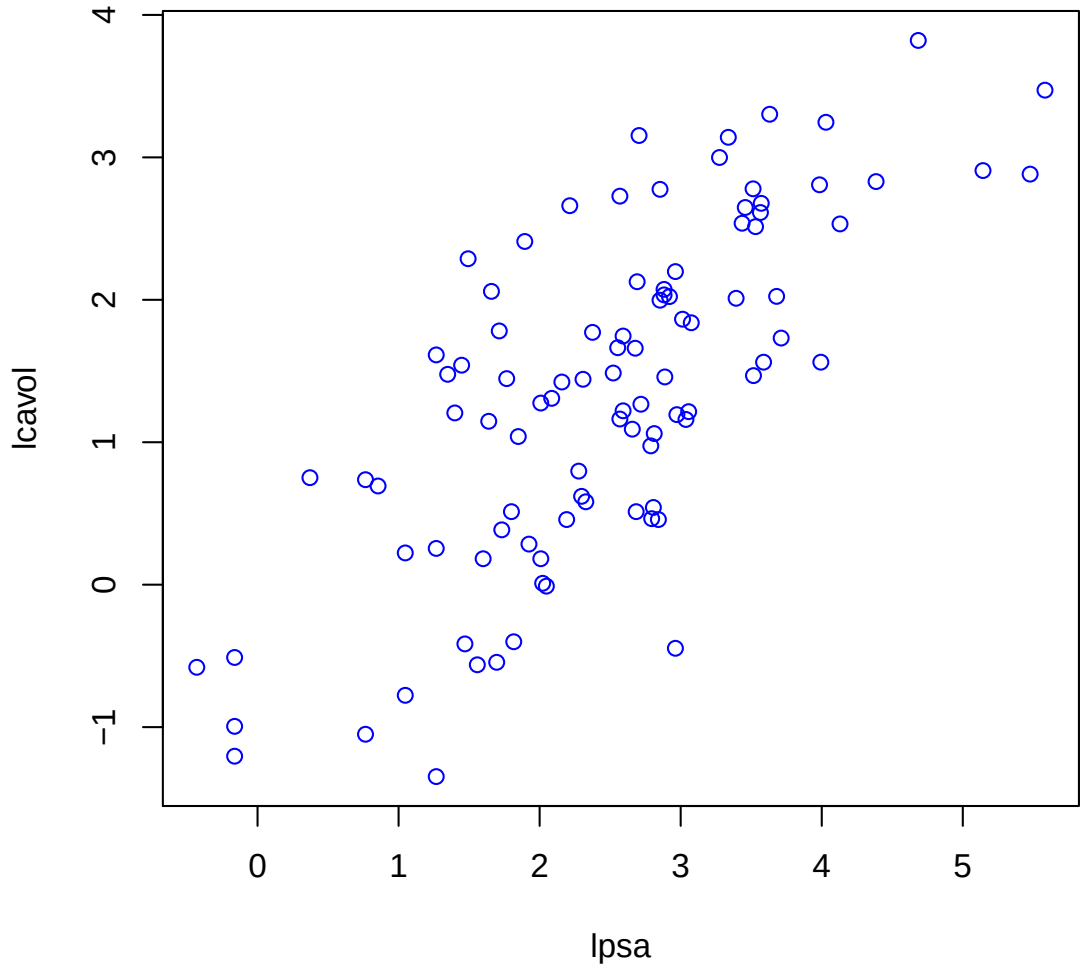


Diagrama de Dispersión



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 y que describa en algún sentido el comportamiento de la variable de interés en función de las demás variables, es decir, las covariables.

En general, identificamos con la letra y a la variable dependiente. El modelo pretende describir cómo el comportamiento de $E(y)$ varía bajo condiciones cambiantes de las otras variables.

En un modelo de regresión se postularía en nuestro caso:

$$y = f(x_1, x_2, x_3, \ldots, x_7) + \epsilon$$

o en general

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

Modelos de regresión

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

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

Una forma de simplificar el problema suponiendo que la familia $\mathcal{F}$ 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

Algunos ejemplos de modelos paramétricos y no paramétricos cuando hay dos variables independientes x_1 y x_2.

Modelos no paramétricos

(i) $y = f(x_1, x_2) + \varepsilon$ donde $f(x_1, x_2)$ es una función continua.

(ii) $y = f(x_1, x_2) + \varepsilon$ donde $f(x_1, x_2)$ es una función continua y derivable.

(iii) $y = f_1(x_1) + f_2(x_2) + \varepsilon$, f_i funciones continuas.

(iv) $y = f(x_1, x_2) + \varepsilon$ donde $f(x_1, x_2)$ es monótona creciente en x_1 y x_2.

Modelos de regresión

Modelos paramétricos

(i) $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 + \varepsilon$

(ii) $y = \beta_1 e^{\beta_2 x_1} + \beta_3 e^{\beta_4 X_2} + \varepsilon$

(iii) $y = \beta_1 x_1^{\beta_2} x_2^{\beta_3} + \varepsilon$

(iv) $y = \beta_1 \log(x_1) + \beta_2 \log(x_2) + \beta_3 x_1^3 + \beta_4 \ \mathsf{sen}(x_2) + \varepsilon$

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) $y = \beta_1 b_1 + \beta_2 x_2 + \beta_3 + \varepsilon$

(iv) $y = \beta_1 \log(x_1) + \beta_2 \log(x_2) + \beta_3 x_1^3 + \beta_4 \,\mathsf{sen}(x_2) + \varepsilon$

Modelo Lineal

En todos estos ejemplos $f(x)$ es **lineal** en los **parámetros**. No es el caso, por ejemplo, de $f(x) = \beta_o e^{-\beta_1 x}$, conocido como creciemiento exponencial, ya que no es lineal como función de los parámetros β_o o β_1.

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

$$\begin{aligned} f(x) &= \beta_o + \beta_1 x \\ f(x) &= \beta_o + \beta_1 x + \beta_2 x^2 \\ f(x) &= \beta_o + \beta_1 \log(x) \end{aligned}$$

Modelo Lineal

En las situaciones más complejas, como en el ejemplo, y depende de un conjunto de p variables $(x_1, \ldots, x_{p-1})$, por lo tanto tendremos

$$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{ip-1} \, .$$

Eventualmente, las $x_i's$ podrían ser funciones de otras variables, tales como $w_1 = \log(x_1)$, $w_2 = \log(x_2)$, $w_3 = x_1^3$, etc., tal como ocurre en el caso iv) y en nuestro ejemplo.

Modelo Lineal

También podríamos introducir variables explicativas que sean categóricas como las dummies que sólo toman los valores 0 y 1 y que sirven para indicar las distintas categorías de una variable categórica.

Este caso es de especial interés pues permite tratar en el marco del modelo lineal el problema de comparar la media de más de dos poblaciones, que se conoce como **Análisis de la Varianza**.

Enfoque matricial

respuesta $y \longleftrightarrow p-1$ variables explicativas x_j

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

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

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \epsilon_i \quad i = 1, \ldots, n \\ E(\varepsilon_i) &= 0 \\ V(\varepsilon_i) &= \sigma^2 \\ cov(\varepsilon_i, \varepsilon_j) &= 0 \quad i \neq j \end{aligned}$$

donde, $\beta_0, \beta_1, \ldots, \beta_{p-1}$ son p parámetros desconocidos a estimar.

Este modelo tiene intercept u ordenada al origen, eventualmente podríamos saber que es 0, en cuyo caso plantearíamos

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

En el caso general tenemos

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p-1} \\ & \ldots & & \ldots & \\ & \ldots & & \ldots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np-1} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_{p-1} \end{pmatrix} \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

$$\Downarrow$$

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

La matriz $\mathbf{X} \in \Re^{n \times p}$ 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.

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:

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

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

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

y que

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

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

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

Matriz de Covarianza

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

$$\{\mathbf{\Sigma_v}\}_{ij} = Cov(\mathbf{v}_i, \mathbf{v}_j) = E[(v_i - \mu_i)(v_j - \mu_j)]$$

Podemos escribirla como:

$$\mathbf{\Sigma_v} = E[(\mathbf{v} - \boldsymbol{\mu})(\mathbf{v} - \boldsymbol{\mu})']$$

donde $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$.

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

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

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

Lema: Sean $\mathbf{A} \in \Re^{m \times n}$, una matriz constante, $\mathbf{d}$ un vector de constantes y $\mathbf{v}$ un vector aleatorio n–dimensional con matriz de covarianza $\mathbf{\Sigma_v}$. Si $\mathbf{w} = \mathbf{A}\mathbf{v} + \mathbf{d}$, entonces:

$$\mathbf{\Sigma_w} = \mathbf{A}\mathbf{\Sigma_v}\mathbf{A}' .$$

El modelo que presentamos más arriba puede escribirse como:

$$\Omega : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad E(\boldsymbol{\epsilon}) = \mathbf{0} \quad \mathbf{\Sigma}_{\boldsymbol{\epsilon}} = \sigma^2\mathbf{I}$$

o equivalentemente

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