Modelos Lineales y Lineales Generalizados

Manuel Gijón Agudo

Octubre 2018 -

Índice

1. Regresión lineal

1.1. Regresión lineal simple

Objetivo: Nuestro objetivo será siempre explicar el comportamiento de una variable aleatoria Y en función de unos ciertos valores $X_1, ..., X_p$.

Dado un $n \in \mathbb{Z}^+$ denominaremos Y_i a la muestra de Y obtenida cuando $X_j = x_{ij} \in \mathbb{R}\ \forall i, j$.

Definición: denominamos el **Modelo Lineal** como:

$$\boxed{\forall i, \quad Y_i = \beta_0 + x_{i1}\beta_1 + ... + x_{i(p-1)}\beta_{p-1} + e_i = \mu_i + e_i}$$

donde β_0 es denominado **intercepto** (*intercept*) y los términos e_i los **errores**.

Hipótesis:

- $\forall i\ \in \{1, 2, ..., n\},\ e_i \sim N(0, \sigma_i^2)$
- **Homeodasticidad** (*homeodasticity*): $\forall i\ \in \{1, 2, ..., n\},\ \sigma_i^2 = \sigma^2$
- $\forall i, y\ \in \{1, 2, ..., n\},\ i \neq j\ e_i$ es **independiente** de e_j
- Los valores de X son fijos o variables aleatorias **independientes** de los errores.

En **forma matricial** escribiremos el modelo de la siguiente manera:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{n(p-1)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Siendo así, definimos $Y_{n \cdot 1} = (Y_1, Y_2, ..., Y_n)^t,\ X_{n \cdot p} = (x_{ij}),\ \beta_{p \cdot 1} = (\beta_0, \beta_1, ..., \beta_{p-1})^t, e_{n \cdot 1} = (e_1, e_2, ..., e_n)^t$ y escribimos el modelo como:

$$Y = X\beta + e \Longleftrightarrow \mu = E(Y|X) = X\beta$$
$$Y|X \sim N(X\beta, \sigma^2 \cdot Id_n)$$

De la última línea se desprende lo siguiente:

$$E((Y - X\beta)(Y - X\beta)^t) = E\Bigg(\begin{pmatrix} e_1^2 & e_1e_2 & e_1e_3 & \cdots & e_1e_n \\ e_2e_3 & e_2^2 & e_2e_3 & \cdots & e_2e_n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ e_ne_1 & e_ne_2 & e_ne_3 & \cdots & e_n^2 \end{pmatrix}\Bigg) = \sigma^2 \cdot Id_n$$

Observación: Las variables $X_1, X_2, ..., X_n$ pueden ser función de otro conjunto de variables. Por ejemplo, podríamos tener un conjunto $\{Z_1, Z_2, ..., Z_m\}$ tales que, para $m \in \mathbb{N}$:

$$X_i = g_i(Z_1, Z_2, ..., Z_m), \quad i \in \{1, 2, ..., n\}$$

Llegados a este punto es importante hablar de los diferentes tipos de modelos lineales con los que nos encontraremos en función del tipo de dato que sean las variables explicativas.

1.2. Regresión Lineal

1.3. ANOVA

En este caso alguna de las covariables si no todas serán **factores** (también llamadas **variables categóricas**). Estos modelos se denominan **análisis de la varianza (ANOVA)**.

1.3.1. Ejemplos

1.4. ANCOVA

En el caso de que los coeficientes

1.4.1. Ejemplos

Example 1.1 *Tenemos datos sobre coche que utilizan diesel o no.*

2. Regresión lineal generalizada

2.1. Genealidades

2.2. Binomial Response Models

Una variable aleatorio es tal que $Y \sim B(p)$ (Bernulli), $0 \leq p \leq 1$, si y solo sí toma valores 1 ó 0 con las siguientes probabilidades:

$$P(Y = 1) = p \text{ and } P(Y = 0) = 1 - p$$

Una variable aleatoria es tal que $Y \sim Bin(n, p)$ (Binomial), con parámetros.... CONTINUAR AQUÍ

BALA BLA BLA BLA

Definimos los **odds** de una variable aleatoria Binomial como $\text{Odd} = \frac{p}{1-p} \in (0, +\infty)$, tal que verifica:

$$\text{Odd} = \begin{cases} 5 & si \quad x \leq 2 \\ x^2 - 6x + 10 & si \quad 2 < x < 5 \\ 4x - 15 & si \quad x \geq 5 \end{cases}$$

SUSTITUIR APROPIADAMENTE LA MIERDA DE AQUÍ ARRIBA IMPORTANTE!!

Para comparar p_1 con $p_2 \in (-1, 1)$ CONTINUAR A1UÍ BLA BLA BLA BLA

$$\begin{cases} H_0 : p_1 = p_2 & \iff H_0 : p_1 - p_2 = 0 \\ H_1 : p_1 \neq p_2 & \iff H_1 : p_1 - p_2 \neq 0 \end{cases}$$

$$Y_i \sim Bin(m_i, p_i)$$

$$g(\mu) = X\beta \Leftrightarrow g(mp) = X\beta$$

Recordemos el link canónico, el parámetro de ddispersione y la función de varianza son respectivamente:

$$\theta_i = \log\left(\frac{p_i}{1 - p_i}\right)$$

$$\Phi = 1$$

$$V(\mu_i) = \mu_i\left(1 - \frac{\mu_i}{m_i}\right)$$

2.3. Poisson Response Models

La principal característica de estos modelos es que la variable respuesta sigue una distribución de Poisson.

$$Y_i \sim Poiss(\mu_i)$$

$$g(\mu) = X\beta$$

Donde recordemos: $Poiss(Y = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad , E(Y) = \lambda, \quad , V(Y) = \lambda.$

El **índice de dispersión** será:

$$I(Y) = \frac{V(Y)}{E(Y)} = 1$$

—- EXPLICAR MEJOR ESTA MIERDA Ejemplo de las placas de petri overdispersión: tendencia al agrupamiento Poisson: randomness Underdisperssion: uno por casilla —- Estimamos el parámetro $\Phi = 1$

- $\widehat{\Phi} = \frac{X^2}{n-p} >> 1 \Rightarrow$ overdispersión
- $<< 1 \Rightarrow$ underdispersión
- ≈ 1 Poisson

In general $V(Y) = \Phi \cdot \mu$ quasi posisson lo usamos en overdispersion

Estimamos Φ con los datos and like this we will have a more accurate estimation of the true variance.

Una vez tengamos una tabla, si fijamos el parámetro n, lo que nos encontramos es una distribución multinomial. Si no lo fijamos seguimos contando con la distribucuín de poisson

Ejemplo de la tabla $\log(\mu_{ij}) = \beta_0 + \beta_1$Factor $1 + \beta_2$Factor 2, en total $1 + a - 1 + b - 1 = a + b - 1$. Tenemos factores, luego puede haber interacciónes, sea pues el siguiente modelo a considerar:

$$\beta_0 + \beta_1\text{Factor } 1 + \beta_2\text{Factor } 2 + \beta_3\text{Interacción}$$

con un total de $a + b - 1 + (a-1)(b-1) = RELLENAR = ab$ luego se corresponde con el modelo completo (*full model*), lo que no tiene sentido considerar.