46114: Estadística Multivariada y Datos Categóricos
Sesión 2: Modelos lineales de regresión

Juan Carlos Martínez-Ovando

juan.martinez.ovando@itam.mx

Enero 21, 2016

Modelos lineales de regresión: Introducción

Los modelos lineales de regresión son motivados por la necesidad de modelar la realización de una variable de interés, Y, con un conjunto de variables auxiliares, $X_1, \ldots, X_p$, a través de la distribución condicional,

$$\mathbb{P}(Y|X_1, \ldots, X_p).$$

Formas de definir la distribución condicional anterior:

- Regresión en media tal que,

$$\mathbb{E}(Y|X_1, \ldots, X_p) = \beta_1 X_1 + \cdots + \beta_p X_p,$$

  y $var(Y|X_1, \ldots, X_p) = \sigma^2$ (constante en las X_js).
- Regresión en varianza tal que,

$$var(Y|X_1, \ldots, X_p) = \beta_1 X_1 + \cdots + \beta_p X_p,$$

  y $\mathbb{E}(Y|X_1, \ldots, X_p) = \mu$ (constante en las X_js).

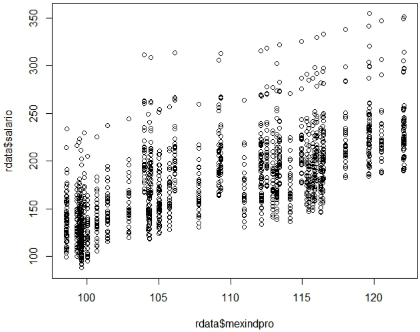
Modelos lineales de regresión: Introducción

¿Cómo elegir el modelo?

- Inspección gráfica.
- Intuición... :)

Ejemplo

- Salario base de cotización en México.
- Nivel de producción industrial (periodo anterior).



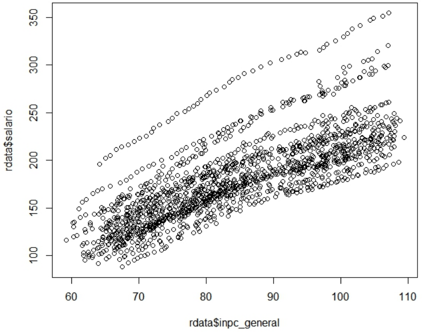
Modelos lineales de regresión: Introducción

¿Cómo elegir el modelo?

- Inspección gráfica.
- Intuición... :)

Ejemplo

- Salario base de cotización en México.
- Nivel del INPC (periodo anterior).



Modelos lineales de regresión: Introducción

La distribución condicional de Y en $X_1, \ldots, X_p$ proviene de,

$$\mathbb{P}(Y, X_1, \ldots, X_p) = \mathbb{P}(Y|X_1, \ldots, X_p) \times \mathbb{P}(X_1, \ldots, X_p).$$

Bajo lo anterior, el enfoque bayesiano de inferencia invoca a la noción de intercambiabilidad (conjunta) en $Y, X_1, \ldots, X_p$, i.e. existen θ y γ tales que

$$\begin{aligned}
\mathbb{P}(\boldsymbol{yx}_1, \ldots, \boldsymbol{yx}_n) &= \int \prod_{i=1}^{n} \mathbb{P}(\boldsymbol{yx}_1, \ldots, \boldsymbol{yx}_n|\theta, \gamma)\pi(\theta, \gamma)d\theta d\gamma \\
&= \int \prod_{i=1}^{n} \mathbb{P}(y_i|x_{i1}, \ldots, x_{ip}, \theta)\mathbb{P}(x_{i1}, \ldots, x_{ip}|\gamma) \\
&\quad \times \pi(\theta|\gamma)d\theta\pi(\gamma)d\gamma,
\end{aligned}$$

donde $\boldsymbol{yx}_j = (y_j, x_{j1}, \ldots, x_{jp})$.

Modelos lineales de regresión: Introducción

- En los modelos lineales de regresión, de lo anterior sólo es de ineterés modelar la relación de dependencia

$$(X_1, \ldots, X_p) \to Y.$$

- Los modelos que involucran la modelación conjunta de Y y las X_js corresponden a *modelos gráficos*.

Considerando los modelos de regresión convencionales, sólo es de interés estudiar $\mathbb{P}(Y|X_1, \ldots, X_p, \theta)$ y $\pi(\theta|\gamma)$, dando origen a la noción de *intercambiabilidad parcial*,

$$\mathbb{P}(y_1, \ldots, y_n|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \int \prod_{i=1}^{n} \mathbb{P}(y_i|X_{i1}, \ldots, X_{ip}, \theta)\pi(\theta|\gamma)d\theta,$$

donde $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jp})$.

Modelos lineales de regresión: Introducción

Concentrándonos en $\mathbb{P}(y_i|X_{i1},\ldots,X_{ip},\theta)$, tenemos los siguientes modelos (entre otros):

- **Modelo lineal en media**

$$\prod_{i=1}^{n} \mathbb{P}(y_i|X_{i1},\ldots,X_{ip},\theta) = \mathrm{N}_n(\boldsymbol{Y}|\boldsymbol{X}'\theta,\sigma^2\boldsymbol{I}).$$

- **Modelo lineal en varianza**

$$\prod_{i=1}^{n} \mathbb{P}(y_i|X_{i1},\ldots,X_{ip},\theta) = \mathrm{N}_n(\boldsymbol{Y}|\boldsymbol{\mu},\mathrm{diag}\{\boldsymbol{x}_1'\theta,\ldots,\boldsymbol{x}_n'\theta\}).$$

- **Modelo lineal generalizado**

$$\prod_{i=1}^{n} \mathbb{P}(y_i|X_{i1},\ldots,X_{ip},\theta) = \mathbb{P}_n(\boldsymbol{Y}|g(\boldsymbol{X}'\theta),\gamma).$$

Modelos lineales de regresión: Inferencia bayesiana

Nos concentraremos ahora en el **modelo de regresión lineal en media**, con la parametrización en términos de la *precisión* en lugar de la *varianza*, i.e.

$$\mathsf{N}_n(\boldsymbol{Y}|\boldsymbol{X}'\theta, \sigma^2\boldsymbol{I}) = \mathsf{N}_n(\boldsymbol{Y}|\boldsymbol{X}'\theta, \lambda\boldsymbol{I}),$$

donde $\lambda = 1/\sigma^2$.
Los parámetros del modelo son $\theta \in \Re^p$ y $\lambda > 0$.
La **distribución inicial** sobre los parámetros se elije como la distribución conjugada, con

$$\pi(\theta, \lambda) = \mathsf{N}_p(\theta|\theta_0, \lambda S_0) \times \mathsf{Ga}(\lambda|\alpha_0, \beta_0).$$

Los hiper-parámetros $\theta_0 \in \Re^p$, S_0 matriz $p \times p$ positivo definida y simétrica, $\alpha_0, \beta_0 > 0$ son *determinados a priori*.

Modelos lineales de regresión: Inferencia bayesiana

Distribución posterior

$$
\begin{aligned}
\pi(\theta, \lambda | \boldsymbol{yx}_1, \ldots, \boldsymbol{yx}_n) &= \pi(\theta | \boldsymbol{yx}_1, \ldots, \boldsymbol{yx}_n, \lambda) \times \pi(\lambda | \boldsymbol{yx}_1, \ldots, \boldsymbol{yx}_n) \\
&= \mathrm{N}_p(\theta | \theta_n, \lambda S_n) \times \mathrm{Ga}(\lambda | \alpha_n, \beta_n)
\end{aligned}
$$

con

$$
\begin{aligned}
S_n &= S_0 + \boldsymbol{X}'\boldsymbol{X}, \\
\theta_n &= S_n^{-1}(S_0\theta_0 + \boldsymbol{X}'\boldsymbol{Y}), \\
\alpha_n &= \alpha_0 + n/2, \\
\beta_n &= \beta_0 + 1/2(\boldsymbol{Y} - \boldsymbol{X}'\theta_n)'\boldsymbol{Y} + 1/2(\theta_0 - \theta_n)'S_0\theta_0.
\end{aligned}
$$

Modelos lineales de regresión: Inferencia bayesiana

Distribución predictiva

$$\begin{aligned}
\mathbb{P}(\boldsymbol{Y}^f|\boldsymbol{X}^f, \boldsymbol{yx}_1, \ldots, \boldsymbol{yx}_n) &= \int\int \mathrm{N}_{n^f}(\boldsymbol{Y}^f|\boldsymbol{X}^{f\prime}\theta, \lambda \boldsymbol{I}_{n^f}) \\
&\quad \times \Pi(d\theta, d\lambda|\boldsymbol{yx}_1, \ldots, \boldsymbol{yx}_n) \\
&= \mathrm{St}_p(\boldsymbol{Y}^f|\boldsymbol{X}^{f\prime}\theta_n, C(\boldsymbol{X}^f)\alpha_n/\beta_n, 2\alpha_n),
\end{aligned}$$

donde

$$C(\boldsymbol{X}^f) = 1 + \boldsymbol{X}^f(\boldsymbol{X}^{f\prime}\boldsymbol{X}^f + S_n)^{-1}\boldsymbol{X}^{f\prime},$$

con θ_n, S_n, α_n y β_n dadas como antes.