Clase 7 - Módulo 2: Introducción a la analítica

Mauricio Alejandro Mazo Lopera

Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín

UNIVERSIDAD NACIONAL DE COLOMBIA

El modelo de regresión lineal múltiple:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

donde $\epsilon_1, \ldots, \epsilon_n$ son independientes y $\epsilon_i \sim N(0, \sigma^2)$, es:

- Fácil de ajustar.
- Fácil de interpretar.
- Produce un hiperplano.
- No es muy flexible.

Ajustando curvas no lineales: Algunas opciones

Regresión polinomial

Regresión splines

Funciones paso

Splines de suavizamiento

Regresión local

Modelos aditivos generalizados

Regresión polinomial

En muchas situaciones no se puede describir la relación entre dos variables a través de modelo de regresión lineal simple:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

En muchas situaciones no se puede describir la relación entre dos variables a través de modelo de regresión lineal simple:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

por tanto, se plantea la relación polinomial de orden d como

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_d X_i^d + \epsilon_i$$

Regresión polinomial

En muchas situaciones no se puede describir la relación entre dos variables a través de modelo de regresión lineal simple:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

por tanto, se plantea la relación polinomial de orden d como

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_d X_i^d + \epsilon_i$$

La ventaja es que es un modelo más flexible, pero se corre el riesgo de un sobreajuste. Por eso se recomienda usar un d menor o igual a 4.

En general, se plantea el modelo polinomial como:

$$g(\mu_i) = \beta_0 + P_1(X_{i1}) + P_2(X_{i2}) + \cdots + P_p(X_{ip})$$

donde

$$P_j(X_{ij}) = \beta_{1j}X_{ij} + \beta_{2j}X_{ij}^2 + \cdots + \beta_{d_j j}X_{ij}^{d_j}$$

es un polinomio de grado d_j y $j = 1, \ldots, p$. Además, $g(\cdot)$ es conocida como función link y $\mu_i = E(Y_i)$.

Diferentes tipos de formas para g:

Nombre	**Función link** $g()$	**Inversa**
Identidad	$g(\mu) = \mu$	$\mu = g(\mu)$
Inversa Negativa	$g(\mu) = -1/\mu$	$\mu = 1/g(\mu)$
Inversa cuadrática	$g(\mu) = -1/\mu^2$	$\mu = \frac{1}{g(\mu)^{1/2}}$
Log	$g(\mu) = \ln(\mu)$	$\mu = \exp[g(\mu)]$
Logit	$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{1}{1+\exp[-g(\mu)]}$

Ejemplo utilizando la base de datos **wage**:

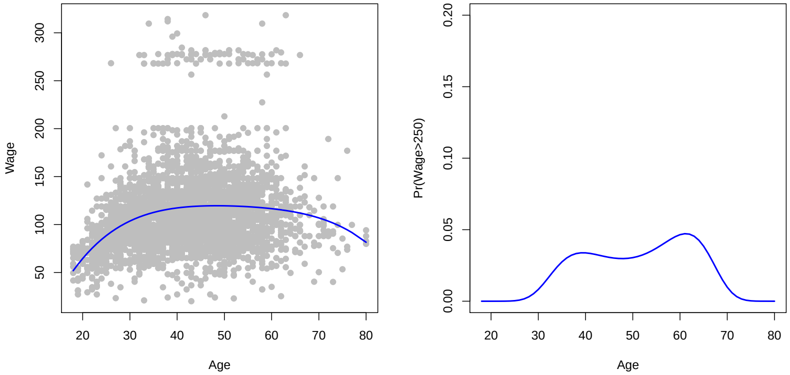
```
require(ISLR)
?Wage # Para obtener información de la base de datos

names(Wage)
```

```
##  [1] "year"       "age"        "maritl"     "race"       "education"
##  [6] "region"     "jobclass"   "health"     "health_ins" "logwage"
## [11] "wage"
```

Ejemplo utilizando la base de datos **wage**:

Asumiendo que existe una relación polinómica entre las variables **wage** y **age**, replique los siguientes gráficos:



Un "problema" que presentan las funciones polinomiales es que imponen una estructura no lineal con respecto al comportamiento no lineal de las covariables $X_1, X_2, \ldots, X_p$.

Un "problema" que presentan las funciones polinomiales es que imponen una estructura no lineal con respecto al comportamiento no lineal de las covariables $X_1, X_2, \ldots, X_p$.

Una alternativa ante dicha situación es usar **funciones paso** con el fin de evitar una estructura global.

En el caso de una sola covariable X, este proceso consiste en particionar el rango de X en varias secciones, lo cual puede ser visto como una categorización de esta covariable considerando una nueva covariable categórica ordenada.

El planteamiento matemático consiste en crear puntos de corte $c_1, c_2, \ldots, c_k$ en el rango de X y luego construir las categorías:

$$
\begin{aligned}
C_0(X) &= I(X < c_1) \\
C_1(X) &= I(c_1 \leq X < c_2) \\
C_2(X) &= I(c_2 \leq X < c_3) \\
&\vdots \\
C_{k-1}(X) &= I(c_{k-1} \leq X < c_k) \\
C_k(X) &= I(c_k \leq X)
\end{aligned}
$$

donde $I(\cdot)$ es la función indicadora que devuelve un 1 en el conjunto indicado y 0 en otro caso.

Las variables $C_i(X)$ son comúnmente conocidas como variables **dummy**.

El modelo que se plantea en este caso sería entonces:

$$Y_i = \beta_0 + \beta_1 C_1(X_i) + \beta_2 C_2(X_i) + \ldots + \beta_k C_k(X_i) + \epsilon_i$$

Funciones paso

Las variables $C_i(X)$ son comúnmente conocidas como variables **dummy**.

El modelo que se plantea en este caso sería entonces:

$$Y_i = \beta_0 + \beta_1 C_1(X_i) + \beta_2 C_2(X_i) + \ldots + \beta_k C_k(X_i) + \epsilon_i$$

¿Cuál sería la interpretación de los β's?

¿Cómo se haría la predicción de la variable aleatoria Y?

Funciones base

La regresión polinomial y la regresión con funciones paso son casos particulares de lo que se conoce como método de las funciones base.

Funciones base

La regresión polinomial y la regresión con funciones paso son casos particulares de lo que se conoce como método de las funciones base.

Este método consiste en aplicar a la v. a. X un conjunto de transformaciones, conocidas y fijas, denotadas por $b_1(X), b_2(X), \ldots, b_k(X)$, de tal manera que el modelo quedaría:

$$Y_i = \beta_0 + \beta_1 b_1(X_i) + \beta_2 b_2(X_i) + \cdots + \beta_k b_k(X_i) + \epsilon_i$$

Funciones base

La regresión polinomial y la regresión con funciones paso son casos particulares de lo que se conoce como método de las funciones base.

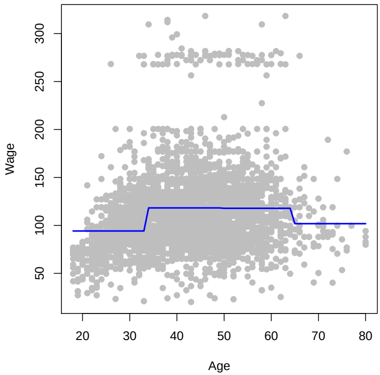
Este método consiste en aplicar a la v. a. X un conjunto de transformaciones, conocidas y fijas, denotadas por $b_1(X), b_2(X), \ldots, b_k(X)$, de tal manera que el modelo quedaría:

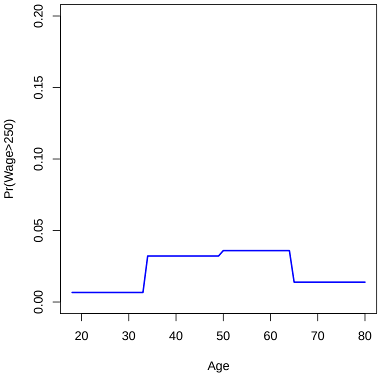
$$Y_i = \beta_0 + \beta_1 b_1(X_i) + \beta_2 b_2(X_i) + \cdots + \beta_k b_k(X_i) + \epsilon_i$$

Cuando $b_j(X_i) = X_i^2$ estamos en el caso de regresión polinomial y cuando $b_j(X_i) = I(c_j \leq X_i < c_{j+1})$ estamos en el caso de funciones paso.

Ejemplo utilizando la base de datos **wage**:

Utilizando funciones paso para **age**, replique los siguientes gráficos:





Considere la base datos **DATOS_C7.txt**.

- Cargue la base de datos en R, guardela como .RData y luego carguela nuevamente. ¿Cuál fue la reducción en tamaño del archivo?
- Realice un análisis para seleccionar las variables más relevantes para explicar Y.

Actividad para realizar en clase:

- Grafique las variables más relevantes versus Y y ajuste un modelo. ¿El comportamiento Y es linear con todas las variables explicativas?
- Ajuste un modelo polinómico y un modelo con funciones paso. Compare ambos modelos. ¿Cuál seleccionaría como el mejor modelo?