

DigitalHouse >
Coding School

DATA SCIENCE

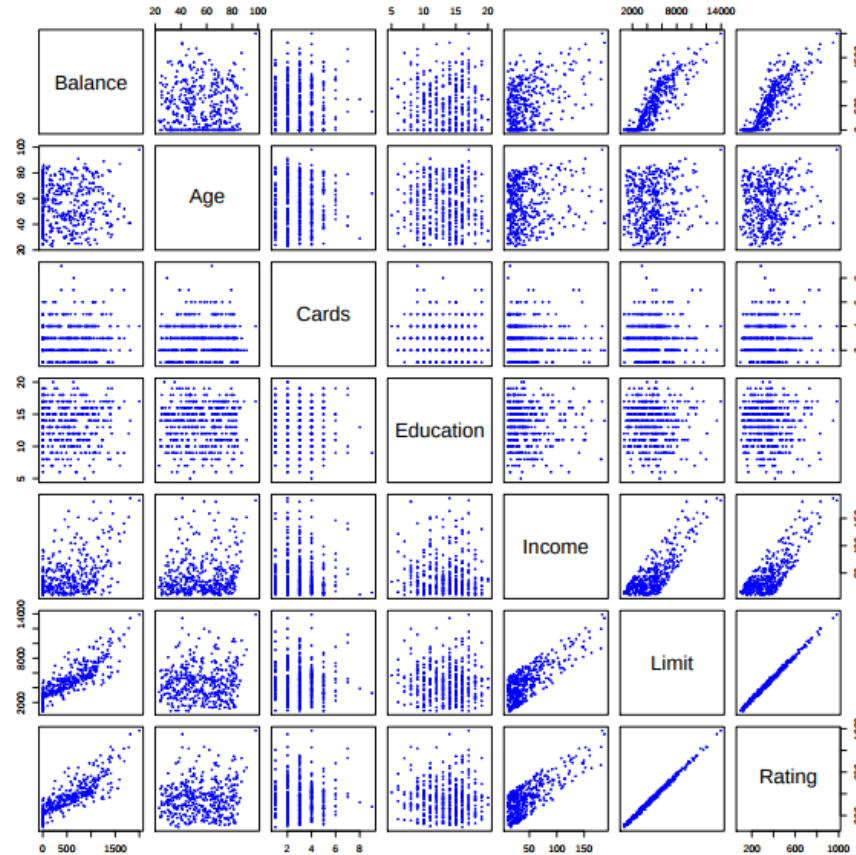
MÓDULO 3

Intro a Regresión Lineal
-Predictores
Cualitativos-
Septiembre de
2017

A veces nos encontraremos con problemas de regresión lineal que presentan predictores cualitativos (variables categóricas nominales u ordinales).

Por ejemplo, en la próxima slide veremos un dataset -**Credit Cards**- que presenta algunos predictores con estas características:

- gender
- student (condición de estudiante)
- status (estado civil)
- ethnicity (caucásico, afroamericano, etc.)



Problema: estimar diferencias entre el balance de la tarjeta de crédito entre hombres y mujeres (ignorando el resto de las variables).

Regresión simple con predictor cualitativo

Variable dummy (x_i) tal que $x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$

Así, el modelo de regresión toma la siguiente forma

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Resultados del modelo anterior...

	Coefficient	Std. Error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

¿Cómo interpretamos esto?

β_0 puede ser interpretado como el balance promedio entre hombres

$\beta_0 + \beta_1$ expresa el balance promedio entre mujeres

β_1 expresa la diferencia media en el balance entre ambos grupos

En la tabla anterior, el crédito medio entre hombres se estimó en \$509,80; el de las mujeres, en cambio, se estimó en \$509,80 + \$19,73 = \$529,53. Es decir, que hay una diferencia de \$19,73.

Sin embargo, notar que el p-value de la variable dummy es demasiado elevado. Esto indica que no parece haber evidencia de una diferencia significativa en el crédito medio entre sexos.

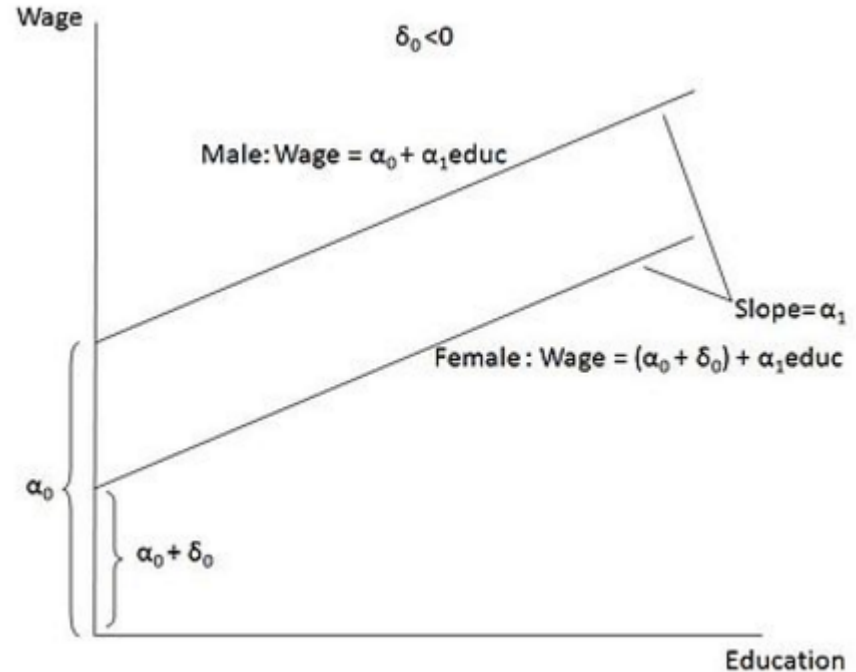
La decisión de codificación es arbitraria. Podríamos haberlo codificado al revés. Solamente hubiesen cambiado los valores de los coeficientes β_0 hubiese sido \$529,53 y $\beta_1 = - \$19,73$ (negativo)

Supongamos un modelo de wage en función de gender y educ donde **gender** es una variable Dummy que **toma los siguientes valores**:

- 0: Male
- 1: Female

Decimos entonces que Male es el **caso base**.

El mc $wage = \alpha_0 + \alpha_1 educ + \delta_0 gender$ ente



Es importante notar lo siguiente:

- La cantidad de parámetros ajustados en total es 3: el intercepto para el caso base, la variación del intercepto y la pendiente.
- Cada grupo (Male y Female) se modela con una recta que tiene la misma pendiente pero distinta ordenada al origen.
- El valor del parámetro δ es negativo, por eso para el caso de las mujeres, la ordenada al origen se encuentra más abajo.

