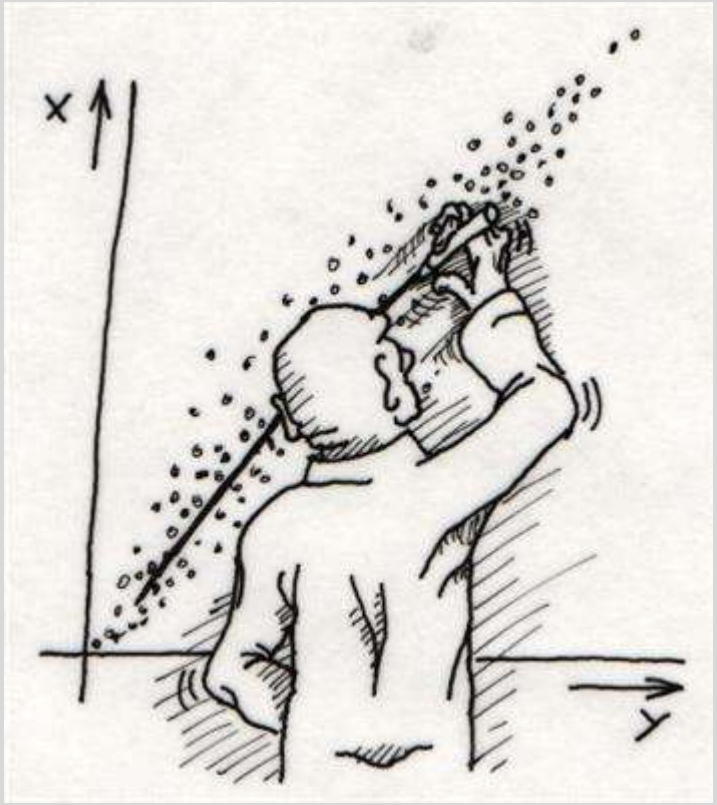


DigitalHouse > Coding School

DATA SCIENCE

MÓDULO 3

Intro a Regresión Lineal -Predictores Cualitativos-

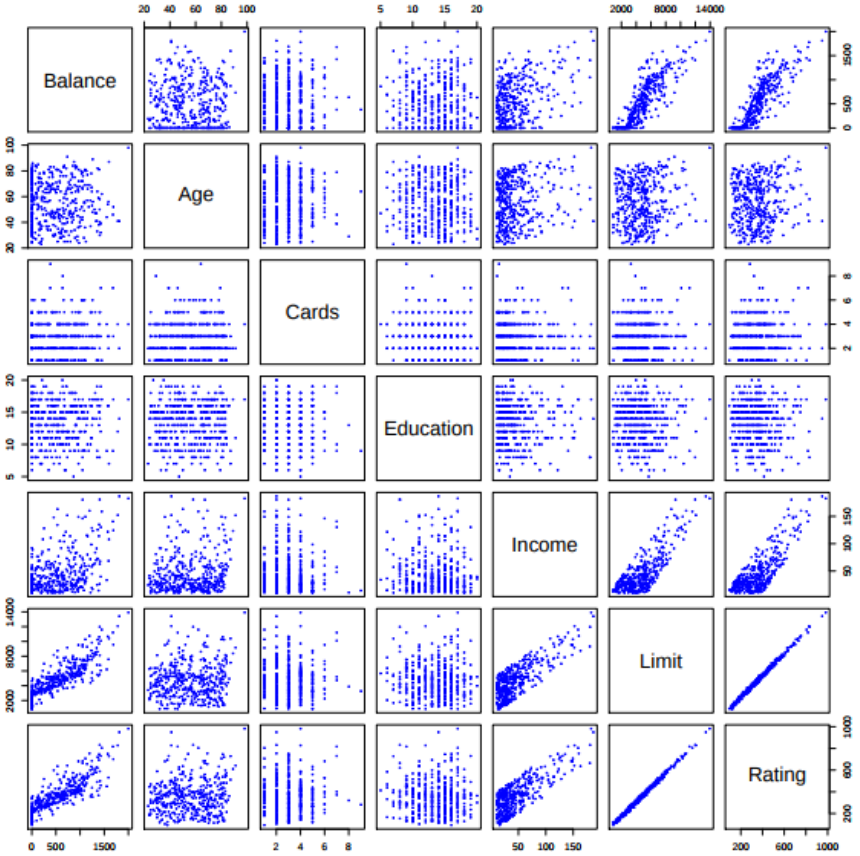
Septiembre de 2017

A veces nos encontraremos con problemas de regresión lineal que presentan predictores cualitativos (variables categóricos 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.)

Features cualitativas / categóricas



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 (xi) 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 diferenca 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 β_1 = -$19,73 (negativo)

Otra forma de codificar el problema:

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

Así, el modelo de regresión toma la siguiente forma

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

β_0 puede ser interpretado como el balance promedio total (de hombres y mujeres)

β_1 es el monto en que las mujeres están **por encima** y los hombres **por debajo** de la media

En este ejemplo, β_0 = \$519,665 (a mitad de camino entre las estimaciones de \$509,80 y \$529,53 del ejemplo anterior.

β_1 = \$9,865 (que es la mitad de \$19,73 en el ejemplo anterior)

Como puede verse, las estimaciones finales de los montos medios son equivalentes, solamente cambia el significado de los coeficientes en ambas codificaciones.

Este enfoque es generalizable de forma directa a variables con más de dos categorías.