Regresión lineal multiple

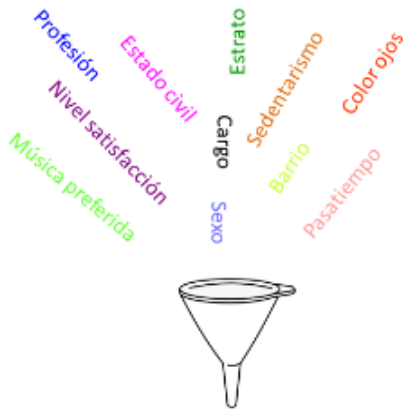
Luis Fernando Delgado Muñoz - Ing. Agroindustrial, M.Sc

lfdelgadom@unal.edu.co

Universidad Nacional de Colombia Facultad de Ingeniería y Administración Departamento de Ingenieria

Modelos con variables explicativas cualitativas (variables indicadoras)

En muchas ocasiones para analizar el efecto de una variable cualitativa, esta se clasifica en 1 ó 0. Para resolver este interrogante se plantea el modelo general teniendo en cuenta la variable cualitativa y su interacción con cada una de las variables regresoras.



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Modelos con variables explicativas cualitativas (variables indicadoras)

Ejemplo: En una planta industrial dedicada a la producción de cierto tipo de material se ha observado el rendimiento de 5 operarios experimentados y no experimentados que han recibido entrenamiento previo en dicha tarea, obteniendo los siguientes resultados. use $\alpha = 0.08$

Horas_Entrenamiento	Experiencia	Tiempo_pn
4	No	26
8	No	19
10	Si	14
13	Si	10
20	Si	6
25	Si	5

Modelos con variables explicativas cualitativas (variables indicadoras)

Ejemplo: En una planta industrial dedicada a la producción de cierto tipo de material se ha observado el rendimiento de 5 operarios experimentados y no experimentados que han recibido entrenamiento previo en dicha tarea, obteniendo los siguientes resultados.

Horas_Entrenamiento	Experiencia	Tiempo_pn
4	No	26
8	No	19
10	Si	14
13	Si	10
20	Si	6
25	Si	5

Modelos con variables explicativas cualitativas (variables indicadoras)

En este caso la variable categorica puede tomar 2 variables (Si y No) y se puede modelar con dos variables Dummy

Hora_Entrenamiento	Tiempo_pn	Experiencia	Si	No
4	26	No	0	1
8	19	No	0	1
10	14	Si	1	0
13	10	Si	1	0
20	6	Si	1	0
25	5	Si	1	0

Modelos con variables explicativas cualitativas (variables indicadoras)

Si la variable explicativa categórica tiene “C” categorías, la regla es ingresar a la ecuación de regresión C-1 variables Dummy; esto para poder estimar el modelo.

Hora_Entrenamiento	Tiempo_pn	Experiencia	Si
4	26	No	0
8	19	No	0
10	14	Si	1
13	10	Si	1
20	6	Si	1
25	5	Si	1

Es indiferente quitar SI ó No, la variable que se quite estará representada por el intercepto y se llegará a la misma conclusión.

Modelos con variables explicativas cualitativas (variables indicadoras)

El modelo de regresión lineal múltiple a estimar es:

$$Y = b_0 + b_1 x_1 + b_2 Si$$

Donde "Si" puede tomar valores entre 0 y 1:

- Cuando "Si" = 1;

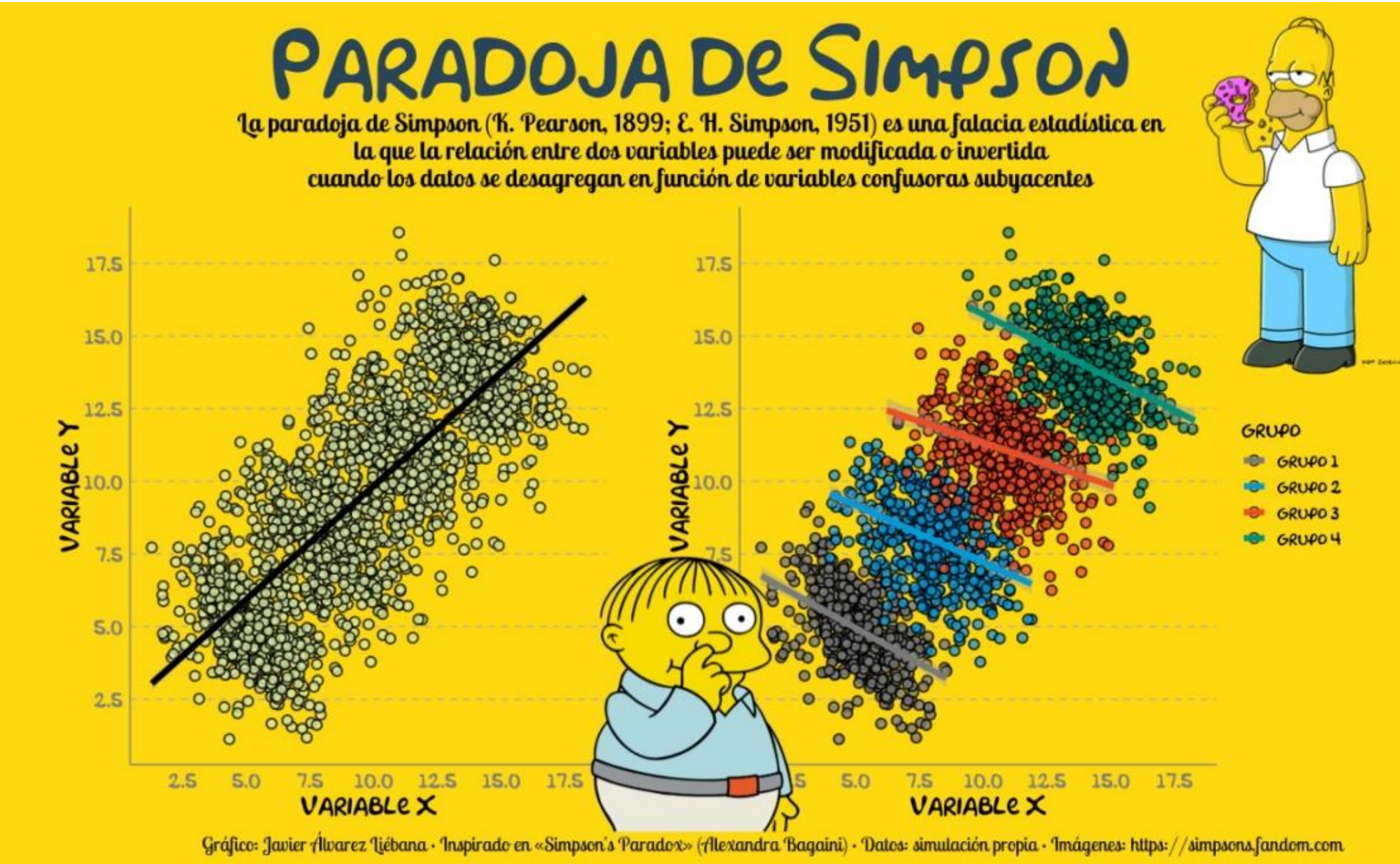
$$Y = b_0 + b_1 x_1 + b_2 (1)$$

$$Y = (b_0 + b_2) + b_1 x_1$$

- Cuando "Si" = 0;

$$Y = b_0 + b_1 x_1 + b_2 (0)$$

$$Y = b_0 + b_1 x_1$$


PARADOJA DE SIMPSON
La paradoja de Simpson (K. Pearson, 1899; E. H. Simpson, 1951) es una falacia estadística en la que la relación entre dos variables puede ser modificada o invertida cuando los datos se desagregan en función de variables confusoras subyacentes
VARIABLE Y
VARIABLE X
17.5
15.0
12.5
10.0
7.5
5.0
2.5
GRUPO
GRUPO 1
GRUPO 2
GRUPO 3
GRUPO 4
Gráfico: Javier Álvarez Liébana · Inspirado en «Simpson's Paradox» (Alexandra Bagaini) · Datos: simulación propia · Imágenes: https://simpsons.fandom.com

Regresión lineal múltiple

Regresión lineal múltiple

En muchas situaciones prácticas existen varias variables independientes que se cree que influyen o están relacionadas con una variable de respuesta Y, y por lo tanto será necesario tomar en cuenta si se quiere predecir o entender mejor el comportamiento de Y.

Por ejemplo, para explicar o predecir el consumo de electricidad en una casa habitación tal vez sea necesario considerar el tipo de residencia, el número de personas que la habitan, la temperatura promedio de la zona, etcétera.

Modelo

$$Y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k + e$$

- Y: es la variable de respuesta o variable dependiente.
- b_j: son los parámetros del modelo que se conocen como coeficientes de
- Regresión
- e: es una componente de error aleatorio.

Si k = 1, estamos en el caso de regresión lineal simple y el modelo es una línea recta; si k = 2, tal ecuación representa un plano.

En general la ecuación la representa un hiperplano en el espacio de k dimensiones generado por las variables $\{X_j\}$

Evaluación de significancia de un modelo

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	?F	Valor-p
Regresión	k	SSR	$MSR = \frac{SSR}{k}$	$F_c = \frac{MSR}{MSE}$	$P(F_c > F_{tabla})$
Residuos	n-(k+1)	SSE	$MSE = \frac{SSE}{n-(k+1)}$		
Total	n-1	Total			

Coeficiente de determinación

Otro criterio para determinar si un modelo es adecuado es el coeficiente de determinación múltiple R^2

El coeficiente de determinación tiene problemas ya que su valor aumenta introduciendo nuevas variables en el modelo aunque su efecto no sea significativo, por lo que siempre se puede aumentar artificialmente y esto llevaría a malas interpretaciones.

Pruebas de hipótesis

El aporte de cada variable al modelo se puede evaluar planteando las hipótesis.

$$H_0: bi = 0$$
$$H_a: bi \neq 0$$

Multicolinealidad

En la regresión múltiple se debe evaluar el supuesto de multicolinealidad, es decir, que no haya relación entre las covariables. El problema es que reduce el poder predictivo de la variable independiente.

Una forma de diagnosticar la multicolinealidad es:

1. Matriz de correlación

2. Valor de tolerancia .

3. Factor de inflación de varianza (VIF): si este valor es >10 hay problemas de multicolinealidad y hay que corregir

Corregir multicolinealidad

- Eliminar la variable causante de la multicolinealidad.
- El sesgo de especiación ocurre cuando se elimina una variable de un modelo pero esta operación va en contravía del modelo teórico.

Construcción del modelo

El objetivo es encontrar el modelo que permita hacer las predicciones mas cercanas a la realidad, para esto se requiere de una evaluación de diversos elementos.

1. Numero apropiado de variables.
2. Transformaciones posibles de las variables predictoras.
3. Potencias de orden superior de los predictores.
4. Interacciones entre los predictores básicos

Métodos de construcción de modelos

Se inicia con una única variable predictora X y luego se incluyen al resto dependiendo del valor p de cada variable.
El investigador debe asumir un valor p para la inclusión de la variable en el modelo.

- Forward
- Backward
- Stepwise

Tarea

Leer sobre los 3 métodos de construcción de modelos.

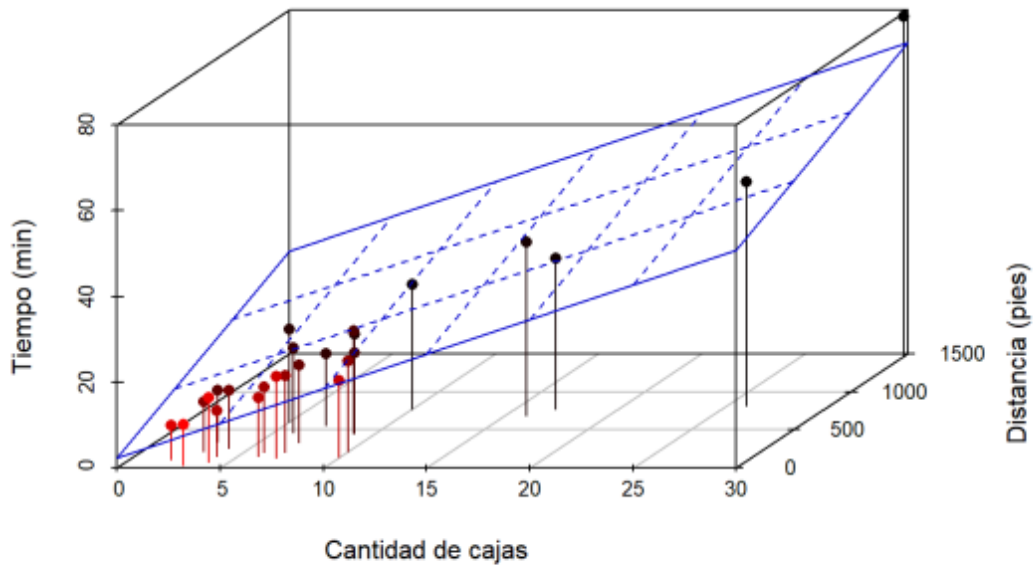
Como seleccionar variables

Para seleccionar las variables del modelo se debe tener en cuenta:

- Coeciente de determinación
- Coeciente de determinación ajustado
- Menor error estandar.
- El estadstico del C_p de Mallows (investigar en que consiste este metodo).

Ejemplo: Tiempo necesario para que un trabajador haga el mantenimiento y surta una máquina dispensadora de refrescos en función de las variables Número de Cajas y Distancia.

$$Tiempo = b_0 + b_1 * cantidad \text{ de cajas} + b_2 * distacia$$



MUCHAS GRACIAS