

# Una Advertencia

## Restricciones de la Regresión Lineal

1. Linealidad
2. Homocedasticidad
3. Normalidad multivariable
4. Independencia de los errores
5. Ausencia de multicolinealidad

# Variables Dummy

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$y =$

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0$$

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1$$

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2$$



# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3$$

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York		
191,792.06	162,597.70	151,377.59	443,898.53	California		
191,050.39	153,441.51	101,145.55	407,934.54	California		
182,901.99	144,372.41	118,671.85	383,199.62	New York		
166,187.94	142,107.34	91,391.77	366,168.42	California		

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York		
191,792.06	162,597.70	151,377.59	443,898.53	California		
191,050.39	153,441.51	101,145.55	407,934.54	California		
182,901.99	144,372.41	118,671.85	383,199.62	New York		
166,187.94	142,107.34	91,391.77	366,168.42	California		

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	
191,792.06	162,597.70	151,377.59	443,898.53	California	0	
191,050.39	153,441.51	101,145.55	407,934.54	California	0	
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	
166,187.94	142,107.34	91,391.77	366,168.42	California	0	

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

## Variables Dummy

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

## Variables Dummy

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1$$



# Variables Dummy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

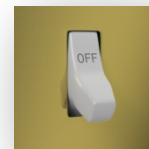
## Variables Dummy

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3$$



$$+ b_4 * D_1$$



# La trampa de las Variables Dummies

# La Trampa de las Variables Dummies

					Variables Dummy	
Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1$$

# La Trampa de las Variables Dummies

Profit	R&D Spend	Admin	Marketing	State	Variables Dummy	
					New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	<b><math>D_2 = 1 - D_1</math></b>		California	0	1
191,050.39	153,441.51			California	0	1
182,901.99	144,372.41			New York	1	0
166,187.94	142,107.34			California	0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3$$



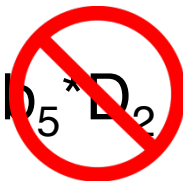
$$+ b_4 * D_1 + \underline{b_5 * D_2}$$



# La Trampa de las Variables Dummies

Profit	R&D Spend	Admin	Marketing	State	Variables Dummy	
					New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

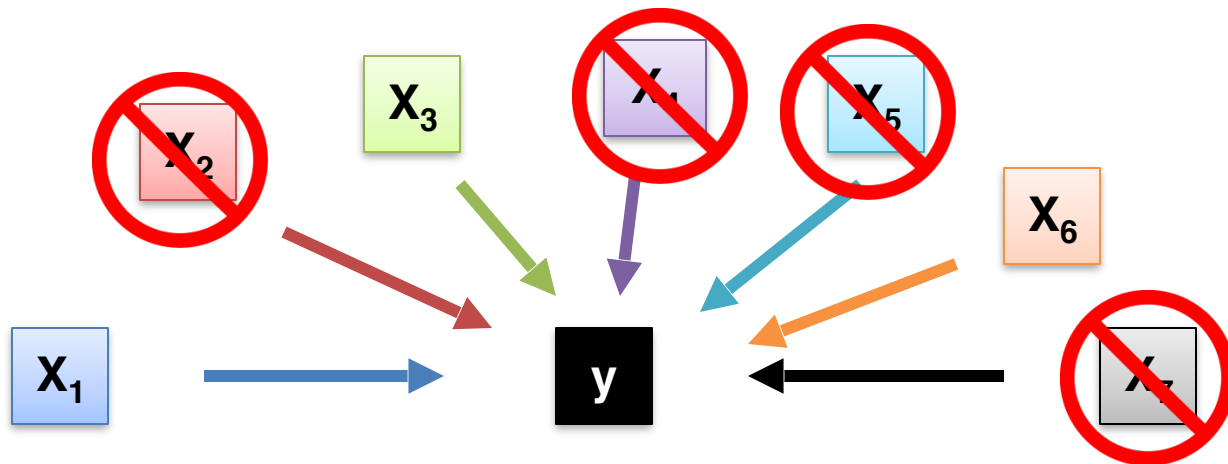
$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3$$

$$+ b_4 * D_1 + b_5 * D_2$$


**Siempre debemos omitir una variable**

# Construir el Modelo (Paso a Paso)

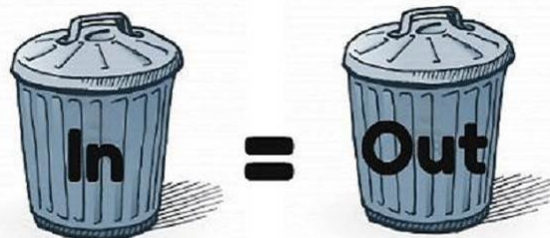
# Construir el Modelo



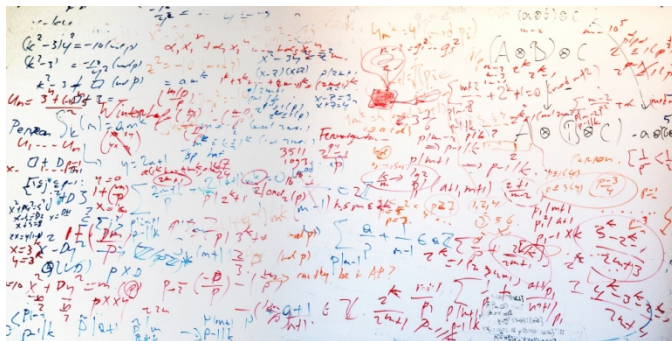
¿Por qué?

# Construir el Modelo

1)



2)





# Construir el Modelo

## 5 métodos para construir modelos:

1. Exhaustivo (All-in)
2. Eliminación hacia atrás
3. Selección hacia adelante
4. Eliminación Bidireccional
5. Comparición de scores

} Regresión paso a paso

# Construir el Modelo

## “All-in” – cases:

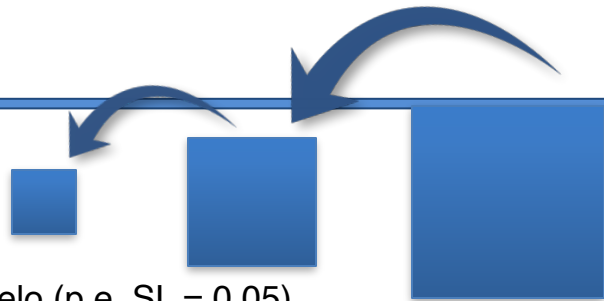
- Conocimiento a priori; OR
- Necesidad; OR
- Preparación previa para Eliminación hacia atrás



# Construir el Modelo

## Eliminación hacia atrás

**PASO 1:** Seleccionar el nivel de significación para permanecer en el modelo (p.e.  $SL = 0.05$ )



**PASO 2:** Se calcula el modelo con todas las posibles variables predictoras

**PASO 3:** Considera la variable predictora con el p-valor más grande. Si  $P > SL$ , entonces vamos al PASO 4, si no vamos a FIN

**PASO 4:** Se elimina la variable predictora

**PASO 5:** Ajustar el modelo sin dicha variable\*

**FIN:** El modelo está listo

# Construir el Modelo

## Selección hacia adelante

**PASO 1:** Elegimos un nivel de significación para entrar en el modelo (p.e.  $SL = 0.05$ )



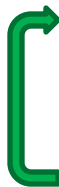
**PASO 2:** Ajustamos todos los modelos de regresión lineal simple  $y \sim x_n$ . Elegimos el que tiene **menor** p-valor.



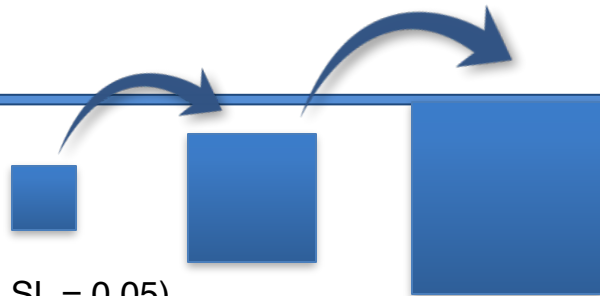
**PASO 3:** Conservamos esta variable, y ajustamos todos los posibles modelos con una variable extra añadida a la(s) que ya tenga(s) el modelo hasta el momento



**PASO 4:** Consideramos la variable predictora con el menor p.valor. Si  $P < SL$ , vamos al PASO 3, si no a FIN



**FIN:** Conservamos el modelo anterior



# Construir el Modelo

## Eliminación Bidireccional

**PASO 1:** Seleccionar un nivel de significación para entrar y salir del modelo  
p.e.:  $SLENTER = 0.05$ ,  $SLSTAY = 0.05$



**PASO 2:** Llevar a cabo el siguiente Paso de Selección hacia adelante (con las nuevas variables con:  
 $P < SLENTER$  para entrar)



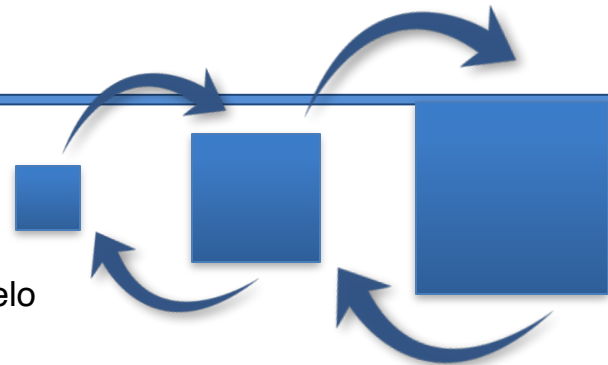
**PASO 3:** Llevar a cabo TODOS los pasos de la Eliminación hacia atrás (las variables antiguas deben tener  
 $P < SLSTAY$  para quedarse)



**PASO 4:** No hay nuevas variables para entrar ni antiguas para salir



**FIN:** El modelo está listo



# Construir el Modelo

## Todos los modelos posibles

**PASO 1:** Seleccionar un criterio de bondad de ajuste (p.e. criterio de Akaike)



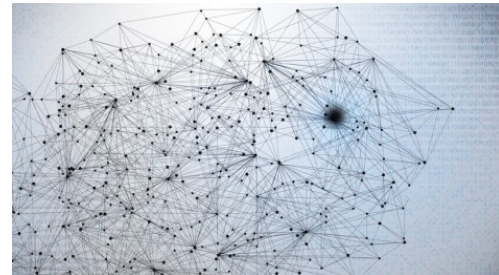
**PASO 2:** Construir todos los posibles modelos de regresión:  $2^N - 1$  combinaciones en total



**PASO 3:** Seleccionar el modelo con el mejor criterio elegido



**FIN:** El modelo está listo



**Por ejemplo:  
10 columnas significan  
1,023 modelos**

# Construir el Modelo

---

## 5 métodos para construir modelos:

1. Exhaustivo (All-in)
2. Eliminación hacia atrás
3. Selección hacia adelante
4. Eliminación Bidireccional
5. Comparición de scores

# Recapitulando



# Recapitulación

---

## En esta sección hemos visto

1. Como crear variables dummies para las categorías de variables independientes
2. Como evitar la trampa de las variables dummies
3. Hacia atrás, hacia adelante, Bidireccional, Todos...
4. Construir un modelo paso a paso!!
5. Como usar el  $R^2$  Ajustado en modelización
6. Como interpretar los coeficientes de una RLM