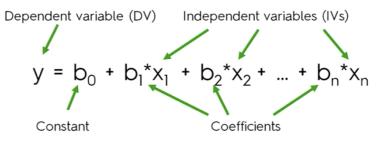
# Regressions

Simple Linear Regression

$$y = b_0 + b_1 x_1$$

Multiple Linear Regression



Data Science Training

© Kirill Eremenko



### **Assumptions of a Linear Regression:**

- 1. Linearity
- 2. Homoscedasticity
- 3. Multivariate normality
- 4. Independence of errors
- 5. Lack of multicollinearity

Data Science Training

© Kirill Eremenko

# **Dummy Variables**

### **Dummy Variables**

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	Wew 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

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

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







Data Science Training

© Kirill Eremenko

## **Dummy Variable Trap**

### **Dummy Variables**

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			nia
191,050.39	153,441.51	Da	= 1 - 1	nia
182,901.99	144,372.41	<b>2</b>		1 ork
166,187.94	142,107.34	,		hia

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

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

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

# **Dummy Variable Trap**

### **Dummy Variables**

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

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

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

Always omit one dummy variable

Data Science Training

© Kirill Eremenko