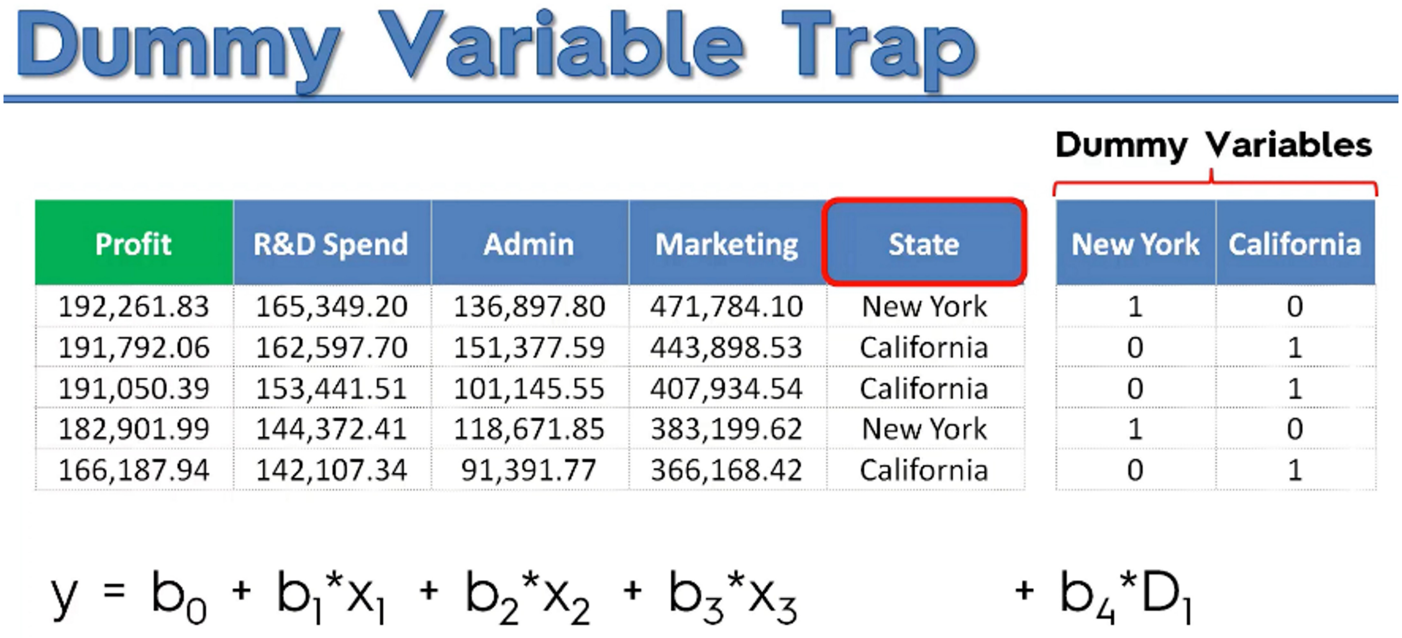
Step 4: Multiple Linear Regression Intuition

Friday, April 8, 2022 4:26 PM

Dummy Variable Trap



Dummy Variable Trap

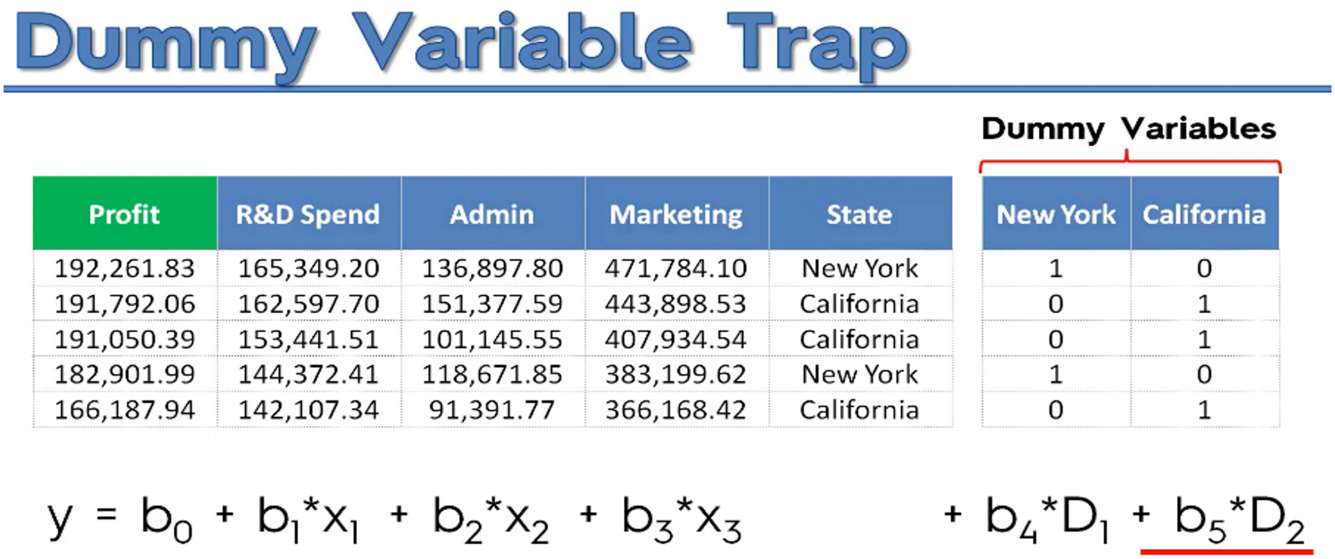
Dummy Variables

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

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + b_3{*}x_3 \qquad + b_4{*}D_1$$

In previous step, we learned how to create dummy variable to replace our Categorical predictors i.e. "State" in the model. And we also learned that we can never include both dummy variables at the same time.

In our example we omitted(excluded) "California" dummy. Now why is that? What will happen if we include second dummy variable in the model as well?



Dummy Variable Trap

Dummy Variables

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

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + b_3{*}x_3 \qquad + b_4{*}D_1 \underline{+ b_5{*}D_2}$$

=>The thing is if we include second dummy variable then we will basically be duplicating a variable. This is because D_2 always equals to $1 - D_1$; the phenomenon where one or several independent variables in a linear regression predict another is called **Multicollinearity**. As a result of this effect the model cannot distinguish between the effects of D_1 from the effects of D_2.

Hence, if we include second dummy variable (Canada) then we will add up with

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

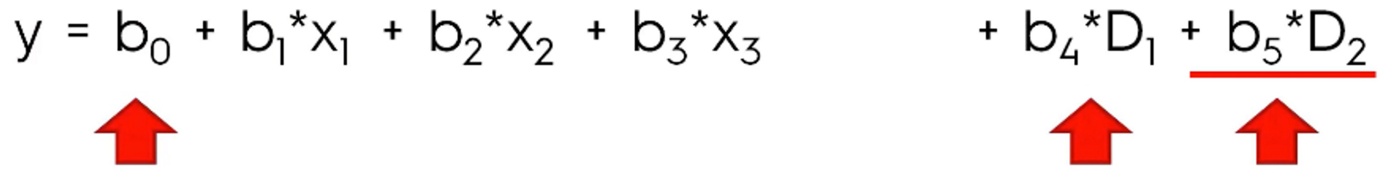
which then will be equals to:

=>$b_4*D_1 + b_5*(1-D_1)$ (because: *D2 = 1 - D1)*

So we cannot include second dummy variable. As a result of this effect the model cannot distinguish between the effects of D_1 from the effects of D_2.

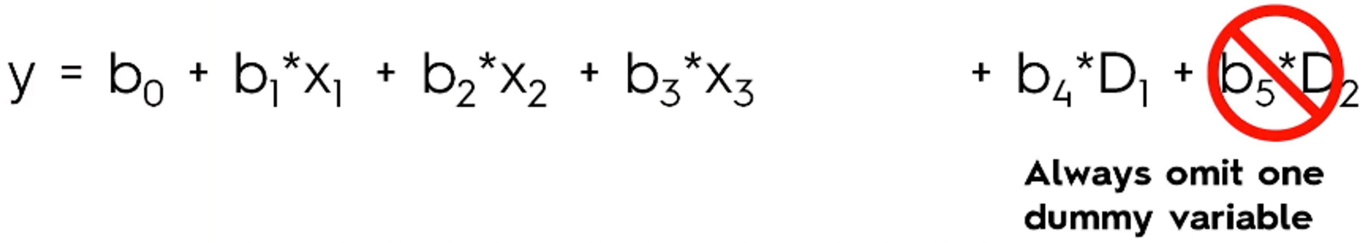
And this is called the Dummy Variable Trap.

The real problem is that we cannot have these 3 (red-arrowed) elements in our model at the same time. The constant(b_0) and both the dummy variable($b_4*D_1 + b_5*D_2$)



$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 \qquad + b_4*D_1 + \underline{b_5*D_2}$$

To sum up, **whenever we are building a model always omit(exclude) one dummy variable**. **And this applies irrespective of the number of dummy variable there are in that specific dummy set. If we have 9 then we should only include 8. If we have a 100 then we should only include 99.**



$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 \qquad + b_4*D_1 + \cancel{b_5*D_2}$$

Always omit one dummy variable