How do I create dummy variables in Excel step-by-step?

Authored by

stats writer

April 25, 2024

Creating dummy variables in Excel allows you to represent categorical data as numerical values, making it easier to analyze and manipulate in your spreadsheet. To create dummy variables in Excel, follow these steps:

1. Identify the categorical variable in your data that you want to convert into dummy variables.

2. Create a new column next to the original variable and label it accordingly.

3. Assign numerical values to each category in the original variable. For example, if your variable has three categories, you can assign 1, 2, and 3 to each one.

4. In the new column, use the IF function to assign a value of 1 if the corresponding cell in the original variable matches the category that you want to represent. For example, if you want to create a dummy variable for the category "Male", your IF function would look like this: =IF(A2="Male",1,0) where A2 is the cell containing the original variable.

5. Copy the formula down the entire column to apply it to all the rows in your data set.

6. Repeat the process for each category in the original variable, creating a new dummy variable column for each one.

7. Once all the dummy variables have been created, you can use them in your analysis and calculations just like any other numerical data.

By following these steps, you can easily create dummy variables in Excel and effectively represent categorical data in your spreadsheet.

Create Dummy Variables in Excel (Step-by-Step)

A is a type of variable that we create in regression analysis so that we can represent a categorical variable as a numerical variable that takes on one of two values: zero or one.

For example, suppose we have the following dataset and we would like to use *age* and *marital status* to

predict *income*:

Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

To use *marital status* as a predictor variable in a regression model, we must convert it into a dummy variable.

Since it is currently a categorical variable that can take on three different values ("Single", "Married", or "Divorced"), we need to create *k*-1 = 3-1 = 2 dummy variables.

To create this dummy variable, we can let "Single" be our baseline value since it occurs most often. Here's how we would convert *marital status* into dummy variables:

Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

Income	Age	Married	Divorced
$45,000	23	0	0
$48,000	25	0	0
$54,000	24	0	0
$57,000	29	0	0
$65,000	38	1	0
$69,000	36	0	0
$78,000	40	1	0
$83,000	59	0	1
$98,000	56	0	1
$104,000	64	1	0
$107,000	53	1	0

This tutorial provides a step-by-step example of how to create dummy variables for this exact dataset in Excel and then perform regression analysis using these dummy variables as predictors.

Step 1: Create the Data

First, let's create the dataset in Excel:

	A	B	C	D	E	F
1	**Income**	**Age**	**Marital Status**			
2	$45,000	23	Single			
3	$48,000	25	Single			
4	$54,000	24	Single			
5	$57,000	29	Single			
6	$65,000	38	Married			
7	$69,000	36	Single			
8	$78,000	40	Married			
9	$83,000	59	Divorced			
10	$98,000	56	Divorced			
11	$104,000	64	Married			
12	$107,000	53	Married			
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						

Step 2: Create the Dummy Variables

Next, we can copy the values in columns A and B to columns E and F, then use the IF() function in Excel to define two new dummy variables: Married and Divorced.

	A	B	C	D	E	F	G	H	I
1	**Income**	**Age**	**Marital Status**		**Income**	**Age**	**Married**	**Divorced**	
2	$45,000	23	Single		$45,000	23	0	0	
3	$48,000	25	Single		$48,000	25	0	0	
4	$54,000	24	Single		$54,000	24	0	0	
5	$57,000	29	Single		$57,000	29	0	0	
6	$65,000	38	Married		$65,000	38	1	0	
7	$69,000	36	Single		$69,000	36	0	0	
8	$78,000	40	Married		$78,000	40	1	0	
9	$83,000	59	Divorced		$83,000	59	0	1	
10	$98,000	56	Divorced		$98,000	56	0	1	
11	$104,000	64	Married		$104,000	64	1	0	
12	$107,000	53	Married		$107,000	53	1	0	
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									

Here is the formula we used in cell G2, which we copied down to the rest of the cells in column G:

=IF(C2 = "Married", 1, 0)

And here is the formula we used in cell H2, which we copied down to the rest of the cells in column H:

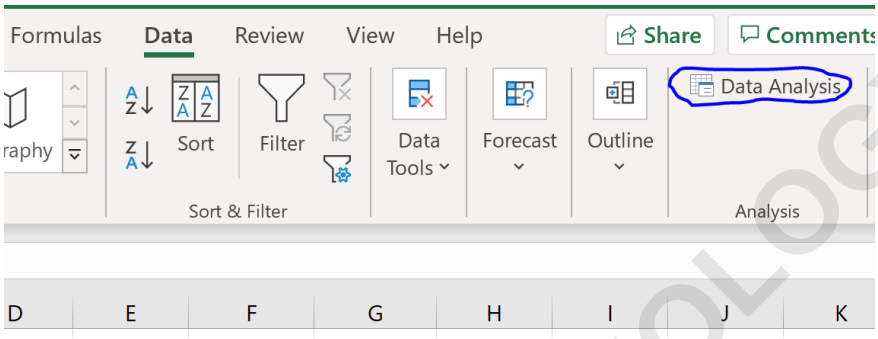
=IF(C2 = "Divorced", 1, 0)

Next, we can use these dummy variables in a

regression model to predict income.

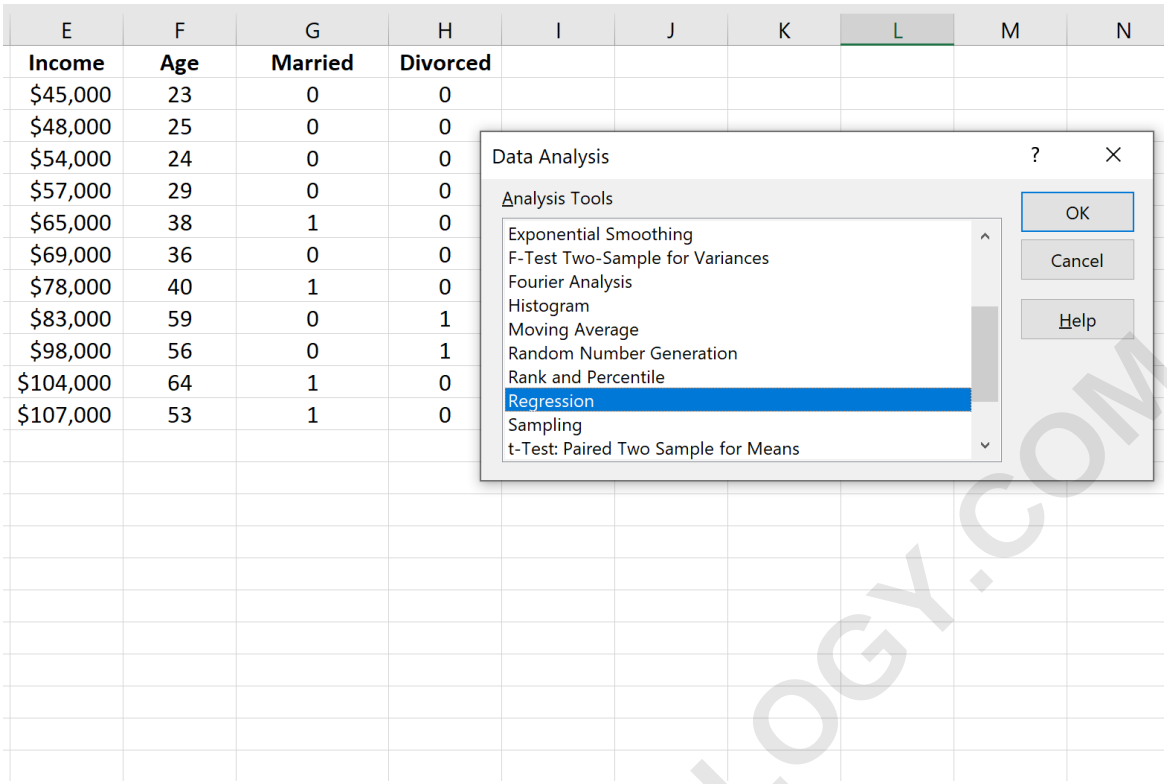
Step 3: Perform Linear Regression

To perform multiple linear regression, we need to click the Data tab along the top ribbon, then Data Analysis within the Analysis section:

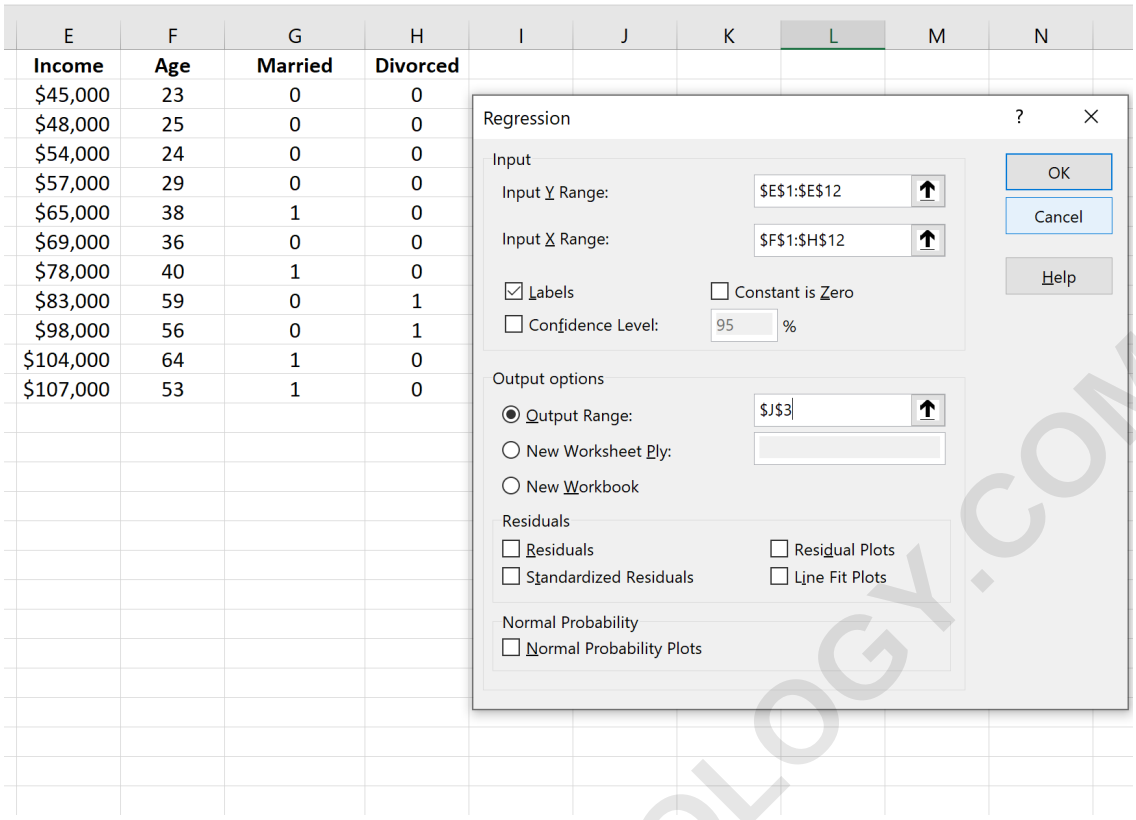


If you don't see this option available, you need to first load the .

In the window that pops up, click Regression and then click OK.



Next, fill in the following information and then click OK.



Income	Age	Married	Divorced
$45,000	23	0	0
$48,000	25	0	0
$54,000	24	0	0
$57,000	29	0	0
$65,000	38	1	0
$69,000	36	0	0
$78,000	40	1	0
$83,000	59	0	1
$98,000	56	0	1
$104,000	64	1	0
$107,000	53	1	0

This produces the following output:

J	K	L	M	N	O	P	Q	R
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.949129568							
R Square	0.900846937							
Adjusted R	0.858352767							
Standard E	8391.005673							
Observatio	11							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	3	4477864439	1.49E+09	21.19931	0.000687			
Residual	7	492862833.4	70408976					
Total	10	4970727273						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	14276.11741	10411.49823	1.371188	0.212656	-10343.2	38895.4	-10343.2	38895.4
Age	1471.674547	354.442326	4.152085	0.004284	633.5516	2309.797	633.5516	2309.797
Married	2479.748416	9431.263403	0.262929	0.800176	-19821.6	24781.14	-19821.6	24781.14
Divorced	-8397.403872	12771.36414	-0.65752	0.531865	-38596.9	21802.07	-38596.9	21802.07

From the output we can see that the fitted regression line is:

Income = 14,276.12 + 1,471.67*(age) + 2,479.75*(married) - 8,397.40*(divorced)

We can use this equation to find the estimated income for an individual based on their age and marital status. For example, an individual who is 35 years old and

married is estimated to have an income of $68,264:

Income = 14,276.12 + 1,471.67*(35) + 2,479.75*(1) - 8,397.40*(0) = $68,264

Here is how to interpret the regression coefficients from the table:

Intercept: The intercept represents the average income for a single individual who is zero years old. Since an individual can't be zero years old, it doesn't make sense to interpret the intercept by itself in this particular regression model.Age: Each one year increase in age is associated with an average increase of $1,471.67 in income. Since the p-value (.004) is less than .05, age is a statistically significant predictor of income.Married: A married individual, on average, earns $2,479.75 more than a single individual. Since the p-value (0.800) is not less than .05, this difference is not statistically significant.Divorced: A divorced individual, on average, earns $8,397.40 less than a single individual. Since the p-value (0.532) is not less than .05, this difference is not statistically significant.

Since both dummy variables were not statistically

significant, we could drop *marital status* as a predictor from the model because it doesn't appear to add any predictive value for income.