## Analyzing your data with logistic regression in SPSS

Linear regression is a great tool when your outcome variable is test scores or loan amounts or another continuous variable. But sometimes, your output is a Yes or a No. That type of outcome is known as dichotomous.

You can still do something similar to linear regression because some super smart stats dude awhile back came up with a way to mimic linear regression with a dichotomous outcome variable.

To do logistic regression SPSS, you need to have the "regression models" add-on program. You also have to understand your data and do a little prep work on it.

## Some assumptions for logistic regression

- 1. You don't have a lot of missing values. If you do, you need to create dummy variables so you can test the "missing" category.
- 2. You have enough records (some suggest a minimum of 50)
- 3. You don't have strong collinearity between your independent variables. If you do, you might have to create some composite variables. You can check this by running a correlation matrix on all your independent variables.

## Getting your data ready

- 1. All of your categorical variables, must be made into dichotomous variables. For example, if you have four races, you'll need a category for white (1,0), black (1,0), etc... But you WON'T include all categories. You leave one our as your "reference" category against others will be compared.
- 2. Don't leave anything out. Include every variable you have at the beginning. You may not use it all in the end, but you must make assure nothing else affects your model.
- 3. Beware of data within data. If, for example, you're looking at traffic stops for multiple police departments, you model may be better explaining department differences than differences in individual stops. To account for this, create a dummy variable for every department to eliminate any affect by department.
- 4. Logistic regression is a better testing tool than reporting tool. Once you figure out your key variables in your model, go back and run crosstabs. Use those figures for graphics and reporting, they will make more sense to your readers.

Once your data is ready, you can do the logistic regression. Here's how:

Go to A	Go to ANALYZE   REGRESSION   BINARY LOGISTIC											
🧰 jail_bond.sav - SPSS Data Editor												
File Edit	View Data	Transform	Analyze	Graphs	Utilities	Ac	ld-ons	Windov	v Help			
<b>⊯</b> ∎ 1:	<b>a</b> 🖳 🗠	Reports Descriptive Statistics Tables			* * *							
	CASE		Gener	are Means al Linear N		* *			RACE	SEX		
1	F-0037436	PRICE T		Models		۲.			В	M		
2	F-0053714	RICHMO				×			D	6.4		
3	F-0137341	WHITEH	Regre:			2	Line					
4	F-0148263	RICHMO	Logline			2	Curv	/e Estima	ation			
5	F-0158762	RICHMO	Classify Data Reduction			[]	Binary Logistic MultiNomial Logistic					
6	F-0158767	RICHMO										
7	F-0175837	WOFFO		rametric 1	fests	÷		nal				
8	F-0200041	PRICE T				¥	Prob	oit				
	F-0200701	WASHIN		al		۲	Non	linear				
	F-0200714	JOHNSC	saulus-1	e Respon:	se	۲	Wei	ght Estin	nation			
	F-0200714	BARRET	Missipe	Missing Value Analysis Complex Samples			2-Stage Least Squares					
	F-0225265	JOHNSC	Compl				Opti	mal Scali	ing			
13	E 0030134	FAMRIE			ALLE	-			IR	IM		

SPSS will prompt you for the DEPENDENT and INDENDENT (OR COVARIATE) variables:

E Logistic Regression		×
A RACE   A SEX   Image: Se	Dependent: DISMISSED Block 1 of 1 Previous Covariates: BDND Partbol Method: Enter Selection Variable: Prufe Categorical Save Options	OK Paste Reset Cancel Help

**SAVE**: If you check PROBABILITIES under SAVE. SPSS will save the probability that each variable will have the outcome.

**OPTIONS:** Check the Hosmer and Lemeshow Test for goodness of fit.

Diagnostic tests to help you interpret your results:

- 1. Pseudo R Squared. Because data in a logistic regression does not form a line, there is not real measurement like R squared in linear regression. However, several researchers have come up with what are called "pseudo" R squared calculations that approximate how much of the variation in your dependent variable is explained by your model. Keep in mind that these are rarely as high as they are in a typical linear regression.
- 2. Hosmer and Lemeshow Test for goodness of fit. This is a complicated statistical measure that tells you how good your model is. What you need to look at is the significance measure. If this measure is less than .05, your model does NOT fit your data very well. You want this to be larger than .05.