14. Multivariate Models and Confounding

Video Link:
https://www.youtube.com/watch?v=QGz6iHiBM9U&t=0s&index=14&list=PL2fQHGEDK7Yyl1W9tgIo8wpYFTDumgc_j

Section 14.1: Multiple Regression with Quantitative Explanatory Variable
Section 14.2: Multiple Regression with Categorical Explanatory Variable
Section 14.3: Logistic Regression

Section 14.1: Multiple Regression with Correlation

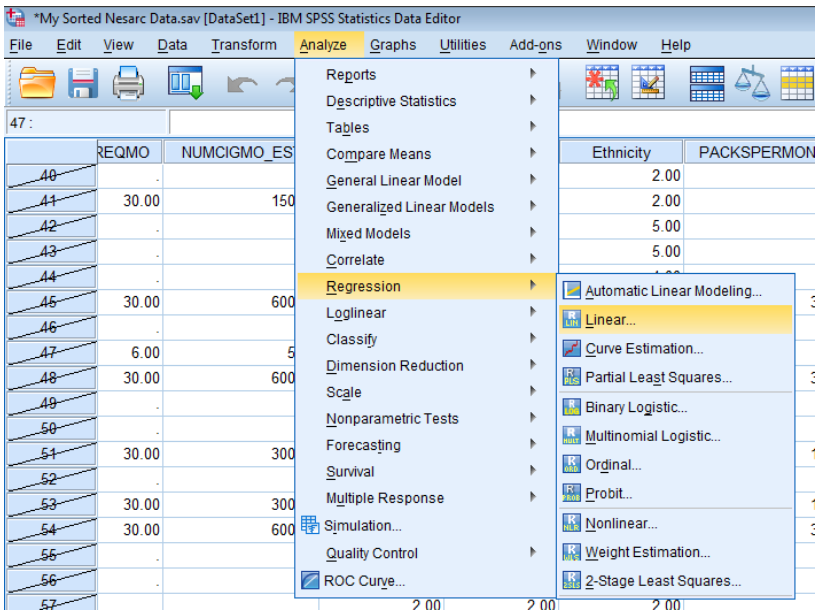
We are going to use the example from correlation in which we asked "Is there an association between Age Onset First General Anxiety Episode and Number of Packs Smoked Per Month in Young Adult Smokers?" With correlation we found a fairly week negative association (r = -.198, p-value = .040). However, with moderation we determined that sex moderated this association (males: r = -.336, p-value = .028; females r = -.126, p-value = .317). Since the relationship is only significant in males we will need to first subset our data to only males.

1. When examining moderation we sorted the data and split the file by our third variables. You will need to sort your data by the Unique Identifier and remove the split file. Review SPSS tutorial 12. Moderation section 12.1 steps #1 through #4.

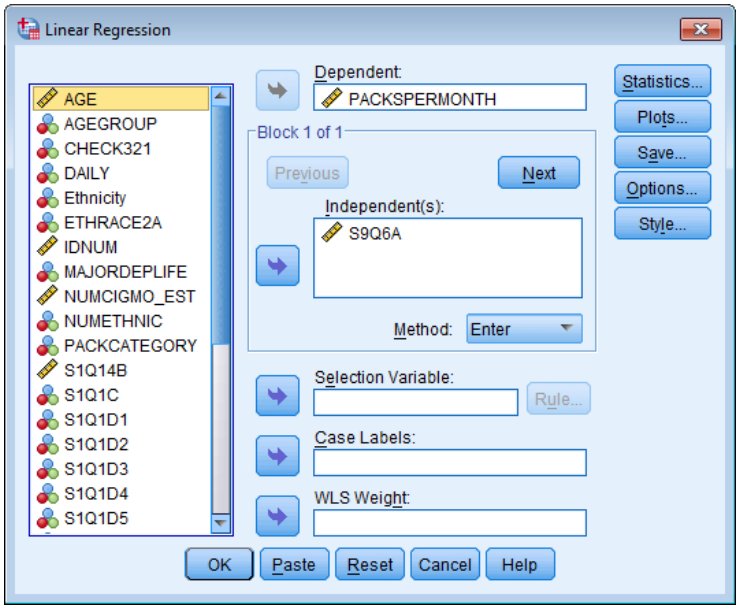
2. To subset, review SPSS tutorial 4. Working with Data steps 4.5 #1 through #4. If you have previously subset your data you will just add to that IF statement (i.e. do not delete your other subset code).

   Now we are ready to find the equation of the best fit line discussed in the video.

3. Go to **Analyze > Regression > Linear…**



4. Using the arrows move your quantitative explanatory variable to the **Independent(s):** window and the quantitative response variable to the **Dependent:** window. Click **OK**.



The **Variables Entered/Removed** table lists your explanatory variable in the **Variables Entered** column and your response variable is listed below the table next to the "a." bullet point.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Age Onset First General Anxiety Episode[b]	.	Enter

a. Dependent Variable: Cigarette packs smoked per month

b. All requested variables entered.

The **Model Summary** table below shows us the **R** or correlation coefficient and **R Square** that we learned about in the Correlation tutorial. Therefore, we do not need to calculate the r^2 value by hand, as it will be directly given to us. We can see this model accounts for 11.3% of the variance in our response variable, PACKSPERMONTH.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.336[a]	.113	.091	13.04827

a. Predictors: (Constant), Age Onset First General Anxiety Episode

The **ANOVA** table shows us the p-value for our explanatory variable's association with the response variable. This p-value will be the same one we get if we run a Pearson Correlation on these two variables. We see that we do indeed have a significant relationship (p = .028) between the explanatory and response variable.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	887.578	1	887.578	5.213	.028[b]
	Residual	6980.552	41	170.257		
	Total	7868.130	42			

a. Dependent Variable: Cigarette packs smoked per month

b. Predictors: (Constant), Age Onset First General Anxiety Episode

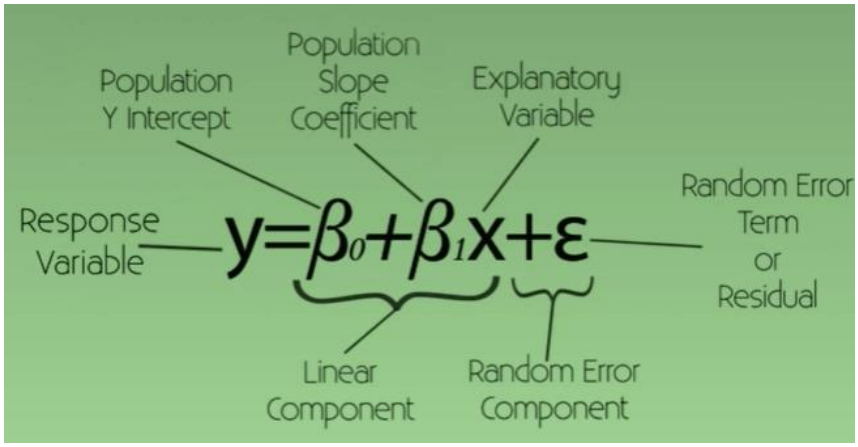
The **Coefficients** table shows us the parameter estimates also known as coefficients or beta weights in the **B** column. The **(Constant)** row lists our Y intercept (B = 42.086) which is our β_0 and our explanatory variable (B = -1.316) which is our β_1. So we now know that our equation for the best fit line of this graph is age of PACKSPERMONTH = 42.086 + -1.316(AGE of ONSET).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	42.086	10.155		4.145	.000
	Age Onset First General Anxiety Episode	-1.316	.576	-.336	-2.283	.028

a. Dependent Variable: Cigarette packs smoked per month

Let's return to the equation for the line that we generated. Look at how our equation is written:

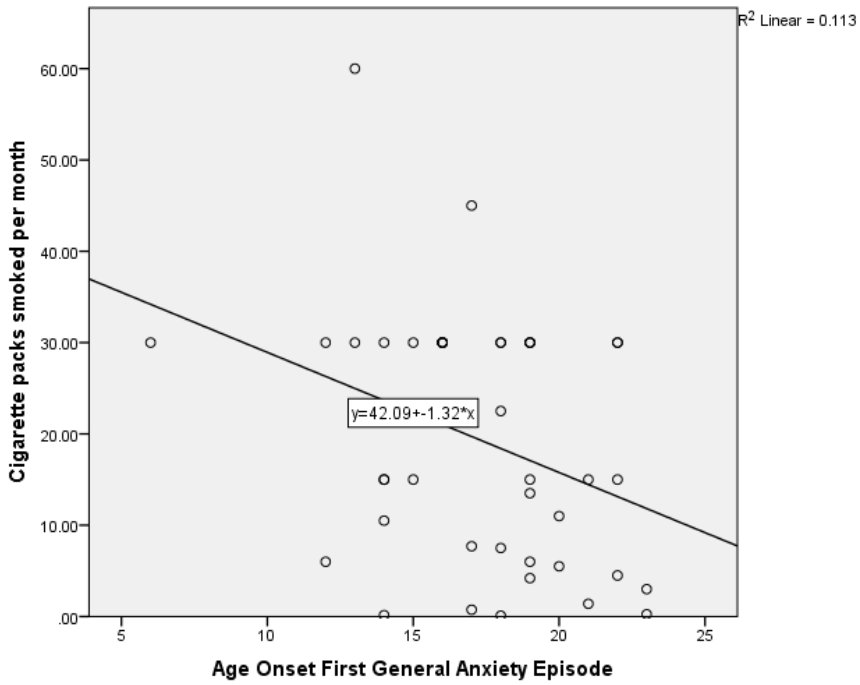


y is a function *of* the variable x and some constant. Thus, as x changes, y will change with it. In building this model we are saying that we believe that x relates to y in some meaningful way. What's exciting about this equation is that we can also use it to generate predicted values for y. The symbol that we use for predicted values of y is "ŷ". For example, let's say we are told that a male's age of onset of first general anxiety episode of 18. Can we predict the amount of cigarette packs smoked per month? Yes! We just plug the value 18 into our equation where we have our x-value.

$\beta_0 = 42.086$
$\beta_1 = -1.316$
Age of Onset = 18

$\hat{y} = 42.086 + -1.316x18 = 18.398$

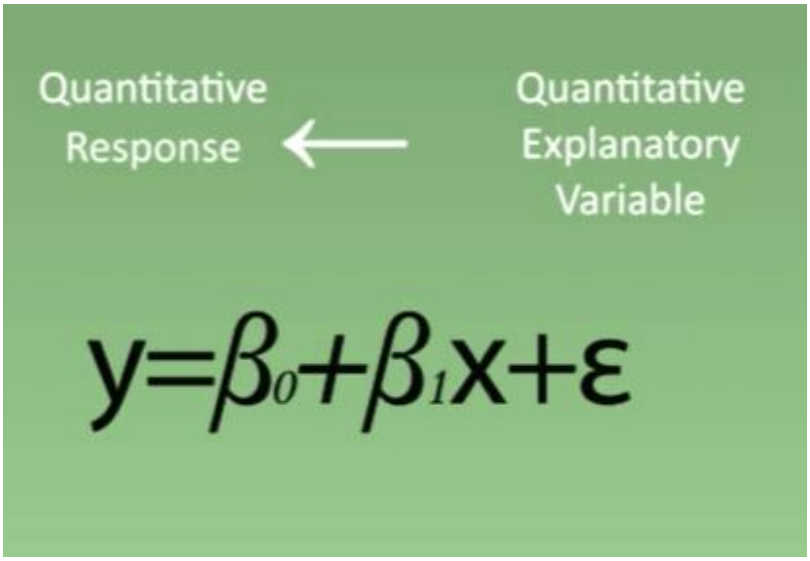
As you can see, if a male's age of onset for first general anxiety episode was 18, we would expect that person to smoke 18.398 packs of cigarettes per month. That would average to about 12 cigarettes a day. Also note from our β_1 that this value is by how much packs per month smoked would increase for every one unit increase in age of onset. For example, if we had a an age of onset of 19, we would know that we would expect their packs smoked per month to be 1.316 less (i.e. 1 less cigarette a day) than a person with age of onset of 18. It is lower because our β_1 is a negative value (i.e., a negative relationship between the explanatory and response variable).



However, note that this is only the *expected* packs smoked per month given what we know about a male's age of onset of first general anxiety episode. It is the value that rests exactly on the best fit line. Unless our data were perfectly correlated, we would anticipate that our expected value and our observed values would differ from one another to some extent.

From our analysis, we now know that there is a statistically significant association between a male's age of first onset of a first general anxiety episode and cigarette packs smoked per month, and we can also tell you what we would expect packs smoked per month in a male to be for a given age of onset of first general anxiety episode. This statistical model has opened the doors to being able to better understand what is really going on between a male's age of first onset of a general anxiety episode and cigarette packs smoked per month. As long as we keep in mind that we are limited by the fact that we imposed the causal model rather than being able to directly test for causation, and that expected data is not the same as observed data, we are still able to explain much about this relationship of interest.

For example, if you look at your scatter plot, you can see a circle on the x-axis representing age of onset at around 23 and packs smoked per month at 0. Our model would predict that males smoke almost 12 packs a month. This is exactly why we include an error term in our model: we are not perfect diviners of the future. What we can do with statistics, however, is identify trends in our data and use those trends to look at what we would expect our data to look like. These trends are incredibly important.



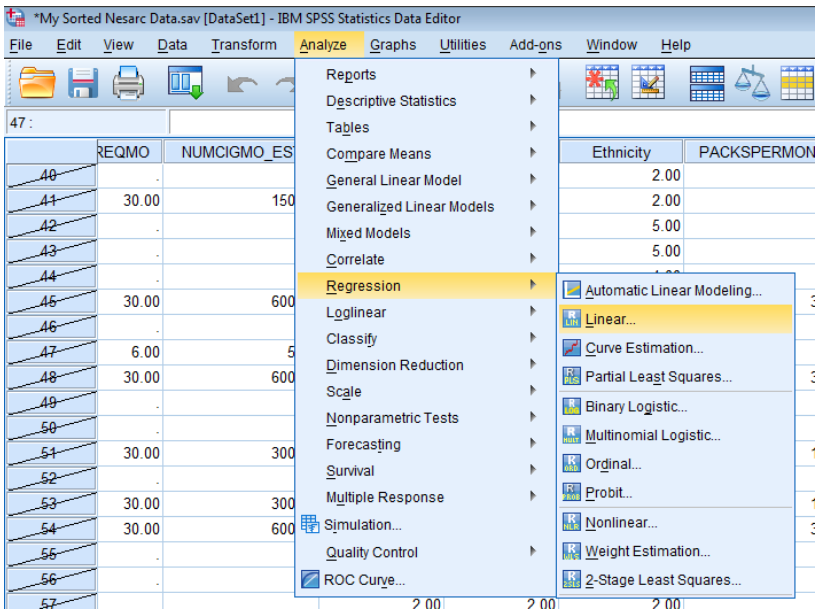
Section 14.2: Multiple Regression with Categorical Explanatory Variable

This equation makes a lot of sense to us when we are working with a quantitative explanatory variable and quantitative response variable, but what about a categorical explanatory variable and quantitative response variable? It obviously wouldn't make very much sense, for example, for us to create a scatterplot and use gender as our predictor variable. However, a regression model will still be informative.

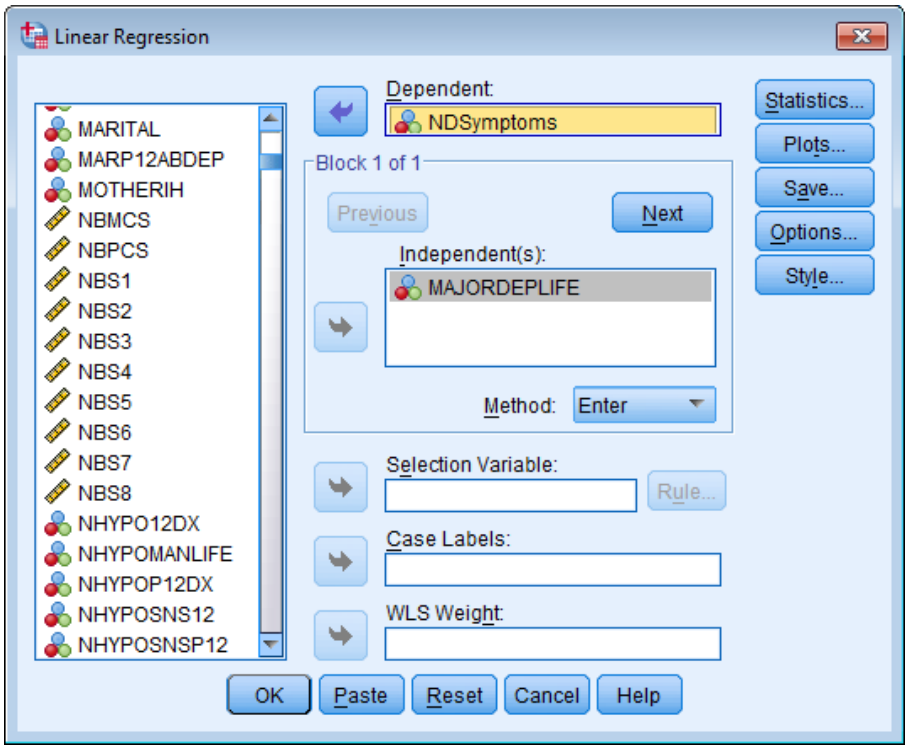
Let's look at the output testing the linear relationship between depression and number of nicotine dependence symptoms, where major depression is a binary categorical explanatory variable and number of nicotine dependence symptoms (ranging from 0 to 7) is a quantitative response variable. Our research question is, "is having major depression associated with an increased number of nicotine dependence symptoms?"

1. In my previous example the data was subset to males. I will need to remove that for this next example. Go to **Data > Select Cases > If…** then remove the "& sex =1" then click **Continue > OK**.

2. Go to **Analyze > Regression > Linear…**



3. Using the arrows, move your categorical explanatory variable to the **Independent(s):** window and the quantitative response variable to the **Dependent:** window. Click **OK**.



We also see the same output format as with the previous regression example.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	MAJORDEPLIFE[b]	.	Enter

a. Dependent Variable: NDSymptoms

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.323[a]	.105	.104	1.772

a. Predictors: (Constant), MAJORDEPLIFE

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	482.270	1	482.270	153.507	.000[b]
	Residual	4131.313	1315	3.142		
	Total	4613.582	1316			

a. Dependent Variable: NDSymptoms

b. Predictors: (Constant), MAJORDEPLIFE

And here are our parameter estimates and the p values.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.186	.057		38.270	.000
	MAJORDEPLIFE	1.365	.110	.323	12.390	.000

a. Dependent Variable: NDSymptoms

Thus, we know that our equation is NDSymptoms = 2.19 + 1.36(MAJORDEPLIFE).

Now let's consider what this equation actually means, since it is not the best fit line of a scatterplot. We know that the variable MAJORDEPLIFE is our depression variable and it takes on the value 0 if the individual does not have major depression and the value 1 if the individual does have major depression. Thus we can plug in the values zero and one into our

MAJORDEPLIFE variable to get the expected number of nicotine dependence symptoms for each group.

As we can see, we would expect daily smokers without depression to have 2.19 nicotine dependence symptoms and daily smokers with depression to have 3.55 nicotine dependence symptoms (remember that we previously subset our data to daily smokers age 18-25).

Notice that this is also the mean number of nicotine dependence symptoms for each group, which we can see by running summary statistics!

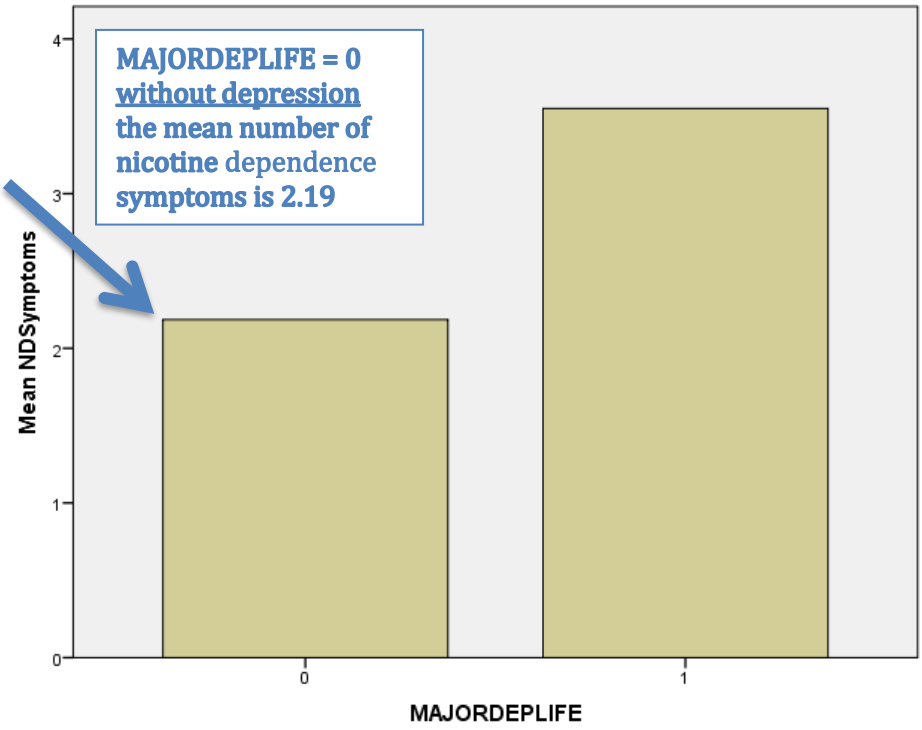
4. To get summary statistics for each group we need to sort the cases by MAJORDEPLIFE (i.e., explanatory variable) and split the file by MAJORDEPLIFE, then run descriptive statistics for NDSymptoms (i.e., response variable). Review SPSS tutorial 12.1 Moderation steps #1 through #4 for how to sort the cases and split the file and SPSS tutorial 6.3 Univariate Graphing steps #1 through #3 for running descriptive statistics.

Descriptive Statistics

MAJORDEPLIFE		N	Minimum	Maximum	Mean	Std. Deviation
0	NDSymptoms	963	0	7	2.19	1.754
	Valid N (listwise)	963				
1	NDSymptoms	354	0	7	3.55	1.822
	Valid N (listwise)	354				

5. Reset, sort the cases and split file removing the MAJORDEPLIFE.

6. Generate a Bivariate graph. Review SPSS tutorial 07. Bivariate Graphing section 7.2 steps #1 through #4.



So although we may not be working with a best fit line, we are still generating important descriptive information out of this equation. Again, this does not mean that everyone in my sample with depression has EXACTLY 3.5 symptoms (obviously, no one can have half of a symptom).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.323[a]	.105	.104	1.772

a. Predictors: (Constant), MAJORDEPLIFE

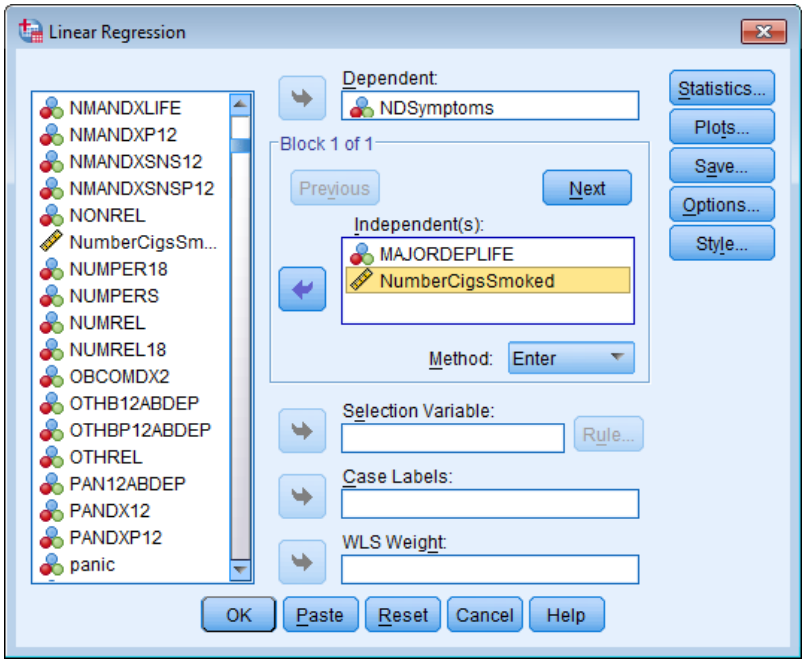
Our low r^2 value (.10) tells us that we're only capturing a small amount of the variability (10%) in the number of nicotine dependence symptoms among daily smokers. But nonetheless this is the value that we would expect given our data. Also note that the categorical variable is a binary categorical variable.

There are a lot of factors that contribute to amount cigarettes smoked and nicotine dependence, the response variables in each of my examples. If we had more information and if we included those other factors in our model it is quite possible that our expected values would be even closer to our observed values.

We could include several explanatory and or predictor variables into our model in order to evaluate both the independent contribution of multiple explanatory variables in predicting our response variable and also in order to evaluate whether specific variables confound the relationship between our explanatory variable of interest and our response variable.
While we now have evidence that depression is significantly associated with the number of nicotine dependence symptoms endorsed by young adult daily smokers (my sample), another likely predictor of nicotine dependence symptoms is of course the number of cigarettes a person smokes each day.

What if number of cigarettes is associated with both our explanatory and response variable (major depression and nicotine dependence symptoms) and that it is really smoking rather than major depression that is associated with number of nicotine dependence symptoms.

7. To evaluate whether this is true, I add number of cigarettes smoked per day to my list of **Independent(s):**



Here is the output

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	NumberCigs Smoked, MAJORDEPLI FE[b]	.	Enter

a. Dependent Variable: NDSymptoms

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.364[a]	.132	.131	1.747

a. Predictors: (Constant), NumberCigsSmoked, MAJORDEPLIFE

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	609.434	2	304.717	99.872	.000[b]
	Residual	3996.922	1310	3.051		
	Total	4606.356	1312			

a. Dependent Variable: NDSymptoms

b. Predictors: (Constant), NumberCigsSmoked, MAJORDEPLIFE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.716	.092		18.554	.000
	MAJORDEPLIFE	1.342	.109	.317	12.327	.000
	NumberCigsSmoked	.036	.006	.166	6.432	.000

a. Dependent Variable: NDSymptoms

We examine the p-values and parameter estimates for each predictor variable (i.e. our explanatory variable – depression and our potential confounder, number of cigarettes smoked). As you can see, both p-values are less than .05 and both of the parameter estimates are positive.

Thus we can conclude that both major depression and number of cigarettes smoked are significantly associated with number of nicotine dependence symptoms after partialing out the portion of the association that can be accounted for by the other. In other words, depression is positively associated with number of nicotine dependence symptoms after controlling for number of cigarettes smoked AND number of cigarettes smoked is positively associated with number of nicotine dependence symptoms after controlling for the presence or absence of depression.

Note that if a parameter estimate is negative and the p-value is significant, it would mean that the there was a negative relationship between that variable and the response variable.

Suppose we started with a different explanatory variable. Dysthymia is pervasive "low level" depression that lasts a long time – often a few years. Suppose we wanted to test the linear relationship between dysthymia, a binary categorical explanatory variable, and number of nicotine dependence symptoms, a quantitative response variable?

Here is the output.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	DYSLIFE[b]	.	Enter

a. Dependent Variable: NDSymptoms

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.150[a]	.023	.022	1.852

a. Predictors: (Constant), DYSLIFE

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	104.066	1	104.066	30.346	.000[b]
	Residual	4509.517	1315	3.429		
	Total	4613.582	1316			

a. Dependent Variable: NDSymptoms

b. Predictors: (Constant), DYSLIFE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.478	.053		46.958	.000
	DYSLIFE	1.138	.207	.150	5.509	.000

a. Dependent Variable: NDSymptoms

You can see from the significant p-value and positive parameter estimates that dysthymia is positively associated with number of nicotine dependence symptoms (i.e. the presence of dysthymia is associated with a larger number of nicotine dependence symptoms and the absence of dysthymia is associated with a smaller number of nicotine dependence symptoms).

While Dysthymia is long-lasting low level depression, major depression is a disorder characterized by a discrete episode of severe depression.

But what happens when we control for major depression (a disorder characterized by a discrete episode of severe depression) in this model?

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	MAJORDEPLIFE, DYSLIFE[b]	.	Enter

a. Dependent Variable: NDSymptoms

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.326[a]	.106	.105	1.771

a. Predictors: (Constant), MAJORDEPLIFE, DYSLIFE

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	490.870	2	245.435	78.226	.000[b]
	Residual	4122.713	1314	3.138		
	Total	4613.582	1316			

a. Dependent Variable: NDSymptoms

b. Predictors: (Constant), MAJORDEPLIFE, DYSLIFE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.181	.057		38.152	.000
	DYSLIFE	.348	.210	.046	1.656	.098
	MAJORDEPLIFE	1.299	.117	.308	11.103	.000

a. Dependent Variable: NDSymptoms

As you can see, dysthymia is no longer significantly associated with number of nicotine dependence symptoms after controlling for major depression. Here we have an example of confounding.

We would say that major depression confounds the relationship between dysthymia and number of nicotine dependence symptoms because the p-value for dysthymia is no longer significant when major depression is included in the model.

As in the previous example, using multiple regression, we can continue to add variables to this model in order to evaluate multiple predictors of our quantitative response variable, number of nicotine dependence symptoms.

Here we can see that when evaluating the independent association among several predictor variables and number of nicotine dependence symptoms, major depression and number of cigarettes smoked are positively and significantly associated with number of nicotine dependence symptoms, while dysthymia, age and gender are not.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.369[a]	.136	.133	1.745

a. Predictors: (Constant), SEX, DYSLIFE, AGE, NumberCigsSmoked, MAJORDEPLIFE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.646	.493		5.363	.000
	DYSLIFE	.275	.209	.036	1.316	.188
	MAJORDEPLIFE	1.297	.116	.307	11.161	.000
	NumberCigsSmoked	.035	.006	.163	6.257	.000
	AGE	-.040	.022	-.047	-1.806	.071
	SEX	-.044	.099	-.012	-.442	.658

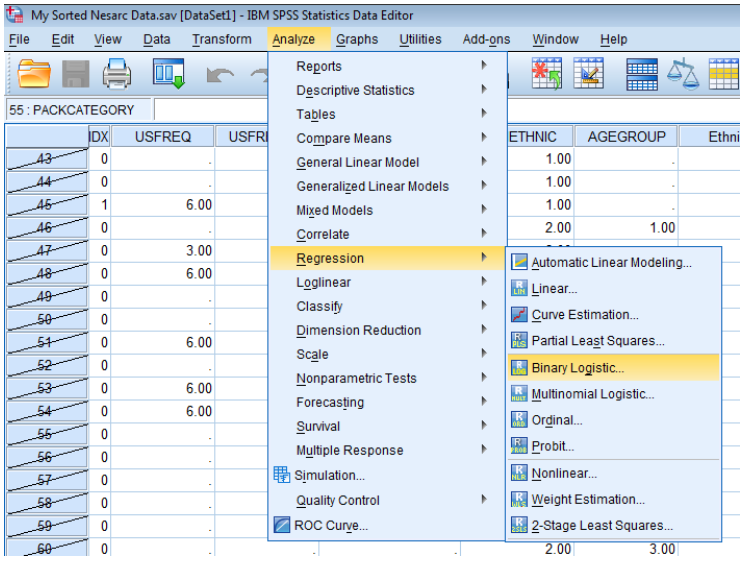
a. Dependent Variable: NDSymptoms

Multiple Regression is the appropriate statistical tool when your response variable is quantitative.

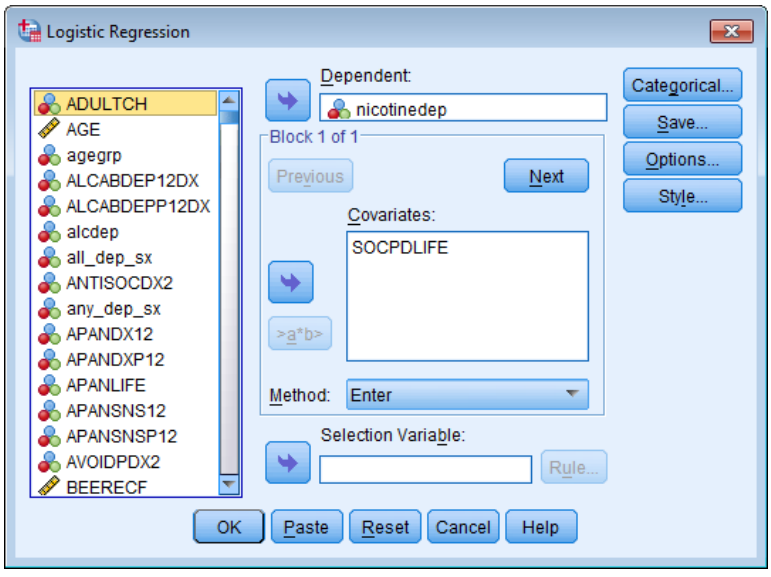
Section 14.3: Logistic Regression

If your response variable is **categorical with two levels** we need to use another multivariate tool, Logistic Regression. My response variable, NICOTINEDEP is binary—yes or no to nicotine dependence—and so I should use a logistic regression. I also have an explanatory variable called "SOCPDLIFE" that indicates the presence or absence of social phobia (an anxiety disorder marked by a strong fear of being judged by others and of being embarrassed).

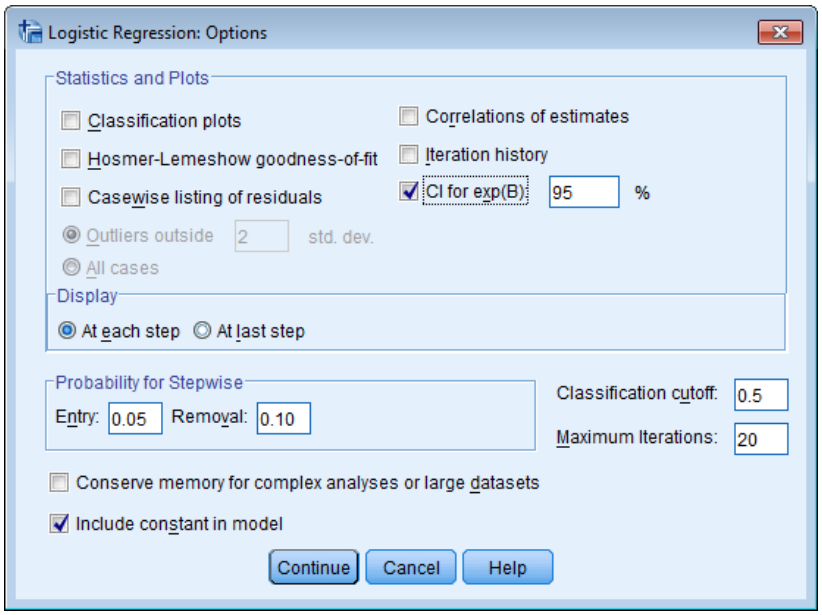
1. Go to **Analyze > Regression > Binary Logistic ...**



2. Using the arrow move your response variable to the **Dependent:** window and your explanatory variable to the **Covariates:** window.



3. Click **Options…** Check **CI for exp(B)**. Click **Continue > OK**.



Let's take a look at the output here:

The **Case Processing Summary** shows the number of participants **Included in Analysis** and **Missing Cases** (i.e., participants excluded).

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1320	100.0
	Missing Cases	0	.0
	Total	1320	100.0
Unselected Cases		0	.0
Total		1320	100.0

a. If weight is in effect, see classification table for the total number of cases.

The **Dependent Variable Encoding** shows the dummy codes used for the response variable.

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Classification Table[a,b]

Observed			Predicted		
			nicotinedep		Percentage Correct
			0	1	
Step 0	nicotinedep	0	0	521	.0
		1	0	799	100.0
	Overall Percentage				60.5

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.428	.056	57.664	1	.000	1.534

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	SOCPDLIFE	15.121	1	.000
	Overall Statistics		15.121	1	.000

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	16.954	1	.000
	Block	16.954	1	.000
	Model	16.954	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1753.965[a]	.013	.017

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table[a]

Observed			Predicted		
			nicotinedep		Percentage Correct
			0	1	
Step 1	nicotinedep	0	0	521	.0
		1	0	799	100.0
	Overall Percentage				60.5

a. The cut value is .500

Similar to the multiple regression output we see a table with the parameter estimates and the p-value. Do note that the explanatory variable is listed in the first row and the y intercept (i.e., Constant) is listed in the second row. With multiple regression the order is switched.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	SOCPDLIFE	1.232	.335	13.500	1	.000	3.427	1.777	6.612
	Constant	.378	.057	43.150	1	.000	1.459		

a. Variable(s) entered on step 1: SOCPDLIFE.

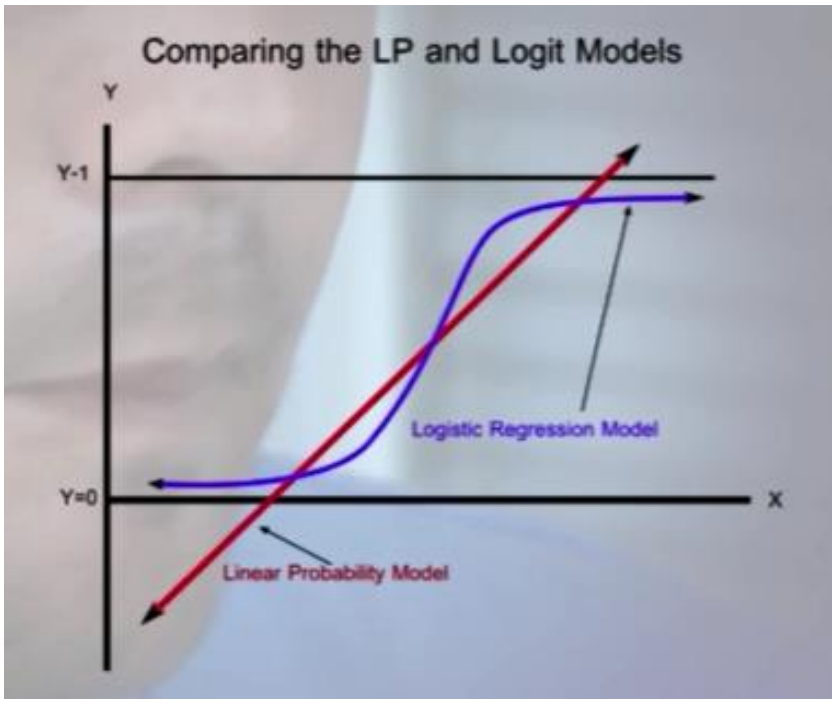
Notice also that our regression is significant at an alpha level of 0.000. Of course, using the parameter estimates we could generate the linear equation:

NICOTINEDEP is a function of .38+1.23(SOCPDLIFE).

But let's really think about this equation some more. In a multiple regression model, our response variable was quantitative, and so it could theoretically take on any value. In a logistic regression, our response variable only takes on the values 0 and 1. Therefore, if I tried to use this equation as a best fit line, I would run in to some problems.

Instead of talking in decimals, it may be more helpful to us to talk about how the PROBABILITY of being nicotine dependent changes based on the presence or absence of social phobia. For example, are those with social phobia more or less likely to be nicotine dependent than those without social phobia? Instead of true expected values, we want probabilities.

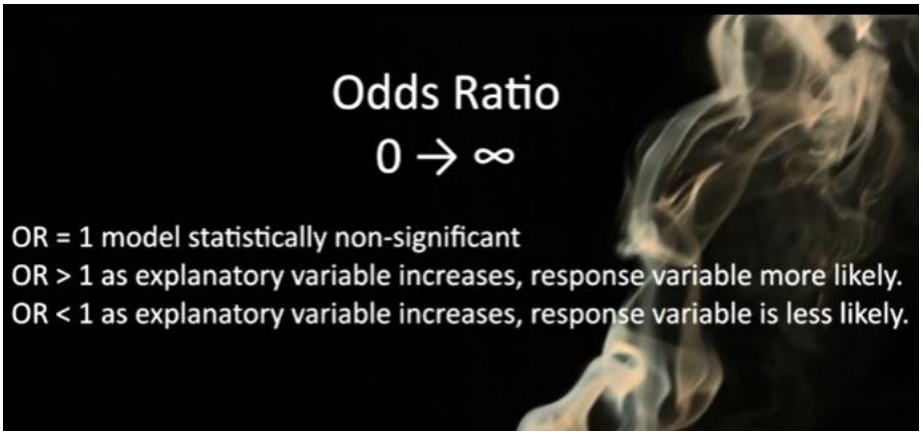
Described visually,



We will no longer find the best fit line, shown in red, very helpful to us, as our outcome variable cannot take on any value. Instead, we are saying that there is somewhere along our x-axis where our outcome variable moves from being more likely to be a zero to being more likely to be a 1. Our goal will be to quantify the probability of getting a one vs. a zero for a given value on our x-axis.

In order to better answer our research question, we will choose to use odds ratios as opposed to coefficients. The "odds ratio" is the probability of an event occurring in one group compared to the probability of an event occurring in another group. Odds ratios are always given in the form of odds, and are not linear. Odds ratios are often a confusing topic for students when they are first introduced to it, so it will be important to go through conceptually and better understand exactly what an odds ratio is and what it means.

An odds ratio can range from zero to positive infinity, and is centered around the value 1. * If we ran our model and got an odds ratio of 1, it would mean that there is an equal probability of nicotine dependence among those with and without social phobia. Those with social phobia are equally as likely to be nicotine dependent as those without. It is also likely, then, that our model would be statistically non-significant. If an odds ratio is greater than 1 it means that the probability of becoming nicotine dependent increases among those with social phobia compared to those without. In contrast, if the odds ratio is below 1 it means that the probability of becoming nicotine dependent is lower among those with social phobia than among those without.



So how do we calculate the odds ratio? It is possible to do this by hand. The odds ratio is the natural exponentiation of our parameter estimate. Thus, all we would need to do is calculate to the power of our parameter estimate. However, we asked SPSS to do this for us when we checked **CI for exp(B)** in step 3.

As you can see, the odds ratio or point estimate and associated confidence interval are part of the SPSS output for logistic regression.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	SOCPDLIFE	1.232	.335	13.500	1	.000	3.427	1.777	6.612
	Constant	.378	.057	43.150	1	.000	1.459		

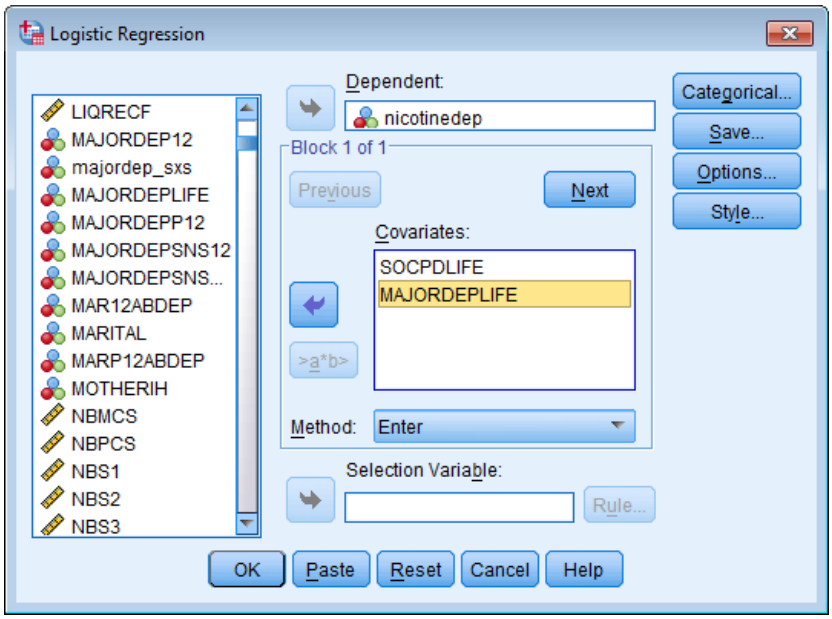
a. Variable(s) entered on step 1: SOCPDLIFE.

Because both my explanatory and response variables in this model are binary coded 0 and 1, I can interpret this odds ratio in the following way. Those young adult daily smokers (my sample) with social phobia are at a 3.4 times greater likelihood of having nicotine dependence than young adult smokers without social phobia.

We also get a confidence interval for our odds ratio. Remember that our data set is just a sample of a population. We do not have every young adult daily smoker in the US. This confidence interval, from 1.85 to 6.97, tells us which values for the odds ratio parameter in the population are plausible. It tells us that we can be 95% confident that, if we select another sample from the population, the odds ratio for that new sample will be somewhere between these two numbers 95 times out of 100. So for example, my odds ratio for social phobia is 3.4. If we were to draw additional samples of young adult daily smokers in the US, 95 times out of 100 the odds ratio would fall somewhere between 1.78 and 6.61.

It is important to keep in mind that the odds ratio is simply a statistic calculated for this sample, and so looking at the confidence interval we can get a better picture of how much this value would change for a different sample drawn from the population. Based on our model, those with social phobia are anywhere from 1.78 to 6.61 times more likely to have nicotine dependence than those without social phobia. Thus, the odds ratio is a sample statistic and the confidence intervals are an estimate of the population parameter.

But what happens when we control for major depression?



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	SOCPDLIFE	.839	.347	5.835	1	.016	2.315	1.172	4.574
	MAJORDEPLIFE	1.307	.152	73.758	1	.000	3.696	2.743	4.981
	Constant	.094	.065	2.086	1	.149	1.098		

a. Variable(s) entered on step 1: SOCPDLIFE, MAJORDEPLIFE.

As you can see, both social phobia and major depression are independently associated with the likelihood of having nicotine dependence

Given that both social phobia and major depression are positively associated with the likelihood of being nicotine dependent, and our predictor or explanatory variables are both binary, we can interpret the odds ratios in the following way.

Young adult daily smokers, the sample population, with social phobia have a 2.3 times greater likelihood of having nicotine dependence than young adult daily smokers without social phobia, after controlling for major depression. Also, daily smokers with major depression are 3.7 times more likely to have nicotine dependence than daily smokers without major depression after controlling for the presence of social phobia. Importantly, because the confidence intervals on our odds ratios overlap, we CANNOT say that major depression is more strongly associated with nicotine dependence than is social phobia. For the population of young adult daily smokers, we can say that those with social phobia are anywhere from 1.2 to 4.6 times more likely to have nicotine dependence than those without social phobia and those with major depression are between 2.7 and 5.0 times more likely to have nicotine dependence than those without major depression. Both of these estimates are calculated after accounting for the alternate disorder.

As with multiple regression, when using logistic regression, we can continue to add variables to our model in order to evaluate multiple predictors of our binary categorical response variable, presence or absence of nicotine dependence.

Another example of confounding occurs when a logistic regression model is run to test the association between panic disorder as the explanatory variable and nicotine dependence -- the response variable. Panic disorder is an anxiety disorder characterized by recurring panic attacks.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	panic	1.033	.269	14.698	1	.000	2.810	1.657	4.765
	Constant	.367	.058	40.012	1	.000	1.443		

a. Variable(s) entered on step 1: panic.

Here we see a significant positive association and note that young adult daily smokers with panic disorder in our sample have a 2.8 times higher likelihood of having nicotine dependence than young adult daily smokers without panic disorder.

However, when we add major depression to the model, panic disorder is no longer significantly associated with nicotine dependence.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	panic	.527	.284	3.451	1	.063	1.695	.971	2.956
	MAJORDEPLIFE	1.298	.154	71.111	1	.000	3.662	2.709	4.952
	Constant	.099	.065	2.310	1	.129	1.104		

a. Variable(s) entered on step 1: panic, MAJORDEPLIFE.

Here we have an example of confounding. We would say that major depression confounds the relationship between panic disorder and nicotine dependence because the p value for panic disorder is no longer significant when major depression is included in the model. Further because panic disorder is no longer associated with nicotine dependence, we would not interpret the corresponding odds ratio but would interpret the significant odds ratio between major depression and nicotine dependence (i.e. that young adult smokers with major depression have a 3.7 times greater likelihood of having nicotine dependence than young adult smokers without major depression, after controlling for panic disorder).

By now, you should be feeling a little more comfortable with the idea of generating a logistic regression model when your outcome variable is binary. Remember to always code your outcome variable such that a 0 means no outcome and a 1 means that an outcome occurred. This is true regardless if your outcome is positive (such as graduation from college) or negative (such as developing nicotine dependence).