LOGISTIC REGRESSION | SPSS ANNOTATED OUTPUT

This page shows an example of logistic regression with footnotes explaining the output. These data were collected on 200 high schools students and are scores on various tests, including science, math, reading and social studies (**socst**). The variable **female** is a dichotomous variable coded 1 if the student was female and 0 if male.

In the syntax below, the **get file** command is used to load the data into SPSS. In quotes, you need to specify where the data file is located on your computer. Remember that you need to use the use extension and that you need to end the command with a period. By default, SPSS does a listwise deletion of missing values. This means that only cases with non-missing values for the dependent as well as all independent variables will be used in the analysis.

Because we do not have a suitable dichotomous variable to use as our dependent variable, we will create one (which we will call **honcomp** for honors composition) based on the continuous variable **write**. We do not advocate making dichotomous variables out of continuous variables; rather, we do this here only for purposes of this illustration.

Use the keyword **with** after the dependent variable to indicate all of the variables (both continuous and categorical) that you want included in the model. If you have a categorical variable with more than two levels, for example, a three-level ses variable (low, medium and high), you can use the **categorical** subcommand to tell SPSS to create the dummy variables necessary to include the variable in the logistic regression, as shown below. You can use the keyword **by** to create interaction terms. For example, the command **logistic regression honcomp with read female read by female.** will create a model with the main effects of **read** and **female**, as well as the interaction of **read** by **female**.

We will start by showing the SPSS commands to open the data file, creating the dichotomous dependent variable, and then running the logistic regression. We will show the entire output, and then break up the output with explanation.

```
get file 'c:\data\hsb2.sav'.
*create honcomp = (write gt 60).
exe.

logistic regression honcomp with read science ses
/categorical ses.
```

Logistic Regression



Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	200	100.0
	Missing Cases	0	.0
	Total	200	100.0
Unselected Cases		0	.0
Total		200	100.0

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

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Categorical Variables Codings[a]

		Parameter coding	
	Frequency	(1)	(2)
ses	1.00	47	.000
	2.00	95	1.000
	3.00	58	.000

a. This coding results in indicator coefficients.

Block 0: Beginning Block



Classification Table[a,,b]

Observed		Predicted		
		honcomp		Percentage Correct
		.00	1.00	
Step 0 honcomp	.00	151	0	100.0
	1.00	49	0	.0
Overall Percentage				75.5

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	-1.132	.170	44.564	1	.000	.322

Variables not in the Equation

		Score	df	Sig.
Step 0	Variables	read	34.861	.000
	science	34.861	1	.000
	ses	14.743	1	.001
	ses(1)	13.333	1	.000
	ses(2)	3.898	1	.048
	Overall Statistics	49.544	3	.000

Block 1: Method = Enter



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	49.544	3	.000
	Block	49.544	3	.000
	Model	49.544	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	165.463[a]	.217	.290

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.



Classification Table[a]

Observed		Predicted		
		honcomp		Percentage Correct
		.00	1.00	
Step 1 honcomp	.00	138	13	91.4
	1.00	37	12	24.5
Overall Percentage				77.5

a. The cut value is .500

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a] read	.129	.023	31.564	1	.000	1.138
science	.099	.027	13.307	1	.000	1.104
ses			5.867	2	.053	
ses(1)	.658	.512	1.652	1	.199	1.931
ses(2)	1.013	.484	4.379	1	.036	2.754
Constant	-10.809	1.462	54.685	1	.000	.000

a. Variable(s) entered on step 1: read, science, ses.

This part of the output tells you about the cases that were included and excluded from the analysis, the coding of the dependent variable, and coding of any categorical variables listed on the **categorical** subcommand. (Note: You will not get the third table ("Categorical Variable Codings") if you do specify the **categorical** subcommand.)

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	200	100.0
	Missing Cases	0	.0
	Total	200	100.0
Unselected Cases		0	.0
Total		200	100.0

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

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Categorical Variables Codings[a]

		Parameter coding	
	Frequency	(1)	(2)
ses	1.00	47	.000
	2.00	95	1.000
	3.00	58	.000

a. This coding results in indicator coefficients.

Block 0: Beginning Block

Classification Table[a,,b]

Observed		Predicted		
		honcomp		Percentage Correct
		.00	1.00	
Step 0 honcomp	.00	151	0	100.0
	1.00	49	0	.0
Overall Percentage				75.5

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	-1.132	.170	44.564	1	.000	.322

Variables not in the Equation

		Score	df	Sig.
Step 0	Variables	read	34.861	.000
	science	34.861	1	.000
	ses	14.743	1	.001
	ses(1)	13.333	1	.000
	ses(2)	3.898	1	.048
	Overall Statistics	49.544	3	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	49.544	3	.000
	Block	49.544	3	.000
	Model	49.544	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	165.463[a]	.217	.290

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

Observed		Predicted		
		honcomp		Percentage Correct
		.00	1.00	
Step 1 honcomp	.00	138	13	91.4
	1.00	37	12	24.5
Overall Percentage				77.5

a. The cut value is .500

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a] read	.129	.023	31.564	1	.000	1.138
science	.099	.027	13.307	1	.000	1.104
ses			5.867	2	.053	
ses(1)	.658	.512	1.652	1	.199	1.931
ses(2)	1.013	.484	4.379	1	.036	2.754
Constant	-10.809	1.462	54.685	1	.000	.000

a. Variable(s) entered on step 1: read, science, ses.

b. **Step 1** – This is the first step (or model) with predictors in it. In this case, it is the full model that we specified in the **logistic regression** command. You can have more steps if you do stepwise or use blocking of variables.

c. **Chi-square** and **Sig.** – This is the chi-square statistic and its significance level. In this example, the statistics for the Step, Model and Block are the same because we have not used stepwise logistic regression or blocking. The value given in the Sig. column is the probability of obtaining the chi-square statistic given that the null hypothesis is true. In other words, this is the probability of obtaining this chi-square statistic (65.588) if there is in fact no effect of the independent variables, taken together, on the dependent variable. This is the overall test of the model. Of course, the p-value, which is compared to a critical value, perhaps .05 or .01 to determine if the overall model is statistically significant. In this case, the model is statistically significant because the p-value is less than .000.

d. **df** – This is the number of degrees of freedom for the model. There is one degree of freedom for each predictor in the model. In this example, we have four predictors: **read**, **write** and two dummies for **ses** (because there are three levels of **ses**).

e. **-2 Log likelihood** – This is the -2 log likelihood for the final model. By itself, this number is not very informative. However, it can be used to compare nested (reduced) models.

f. **Cox & Snell R Square** and **Nagelkerke R Square** – These are pseudo R-squares. Logistic regression does not have an equivalent to the R-squared that is found in OLS regression; however, many people have tried to come up with one. There are a wide variety of pseudo R-square statistics (these are only two of them). Because this statistic does not mean what R-squared means in OLS regression (the proportion of variance explained by the predictors), we suggest interpreting this statistic with great caution.

g. **Observed** – This indicates the number of 0's and 1's that are observed in the dependent variable.

h. **Predicted** – These are the predicted values of the dependent variable based on the full logistic regression model. This table shows how many cases are correctly predicted (122 cases are observed to be 0 and are correctly predicted to be 0; 27 cases are observed to be 1 and are correctly predicted to be 1), and how many cases are not correctly predicted (29 cases are observed to be 0 but are predicted to be 1; 26 cases are observed to be 1 but are predicted to be 0).

i. **Overall Percentage** – This gives the overall percent of cases that are correctly predicted by the model (in this case, the full model that we specified). As you can see, this percentage has increased from 73.5 for the model with only the intercept to 79.5 for the full model.

j. **B** – These are the values for the logistic regression equation for predicting the dependent variable from the independent variable. They are in log-odds units. Similar to OLS regression, the prediction equation is

$$\text{logit}\,\hat{p} = b0 + b1*r1 + b2*r2 + b3*r3 + b3*r4*w$$

where $\hat{p}$ is the probability of being in honors composition. Expressed in terms of the variables used in this example, the logistic regression equation is

$$\text{logit}\,\hat{p} = -9.561 + 0.098*\text{read} + 0.066*\text{science} + 0.058*\text{ses}(1) + 1.057*\text{ses}(2)$$

These estimates tell you about the relationship between the independent variables and the dependent variable, where the dependent variable is on the logit scale. These estimates tell the amount of increase (or decrease, if the sign of the coefficient is negative) in the predicted log odds of honcomp = 1 that would be expected by a one unit increase in the predictor, holding all other predictors constant. **Note**: For the dummy variables created for **ses**, the interpretation is that the coefficients are not significantly different from 0, which should be taken into account when interpreting the coefficients. (See the columns labeled Wald and Sig. regarding testing whether the coefficients are statistically significant). Because these coefficients are in log-odds units, they are often difficult to interpret, so they are often converted into odds ratios. You can do this by hand by exponentiating the coefficient, or by looking at the right-most column in the Variables in the Equation table labeled "Exp(B)". **read** – For every one-unit increase in reading score, we expect a one-unit increase in the log-odds of **honcomp**, holding all other independent variables constant. **science** – For every one-unit increase in science score, we expect a 0.066 increase in the log-odds of **honcomp**, holding all other independent variables constant. **ses** – This tells you if the overall variable **ses** is statistically significant. Here, it is not significant at the .05 level, but it is significant at the .10 level. Remember that you cannot look at the significance of either of those **ses** coefficients. However, as you can see in this example, the coefficient for one of the dummies is statistically significant while the other one is not. The statistic given on this row tells you if the dummies that represent **ses**, taken together, are statistically significant. Because there are two dummies, this test has two degrees of freedom. This is equivalent to using the **test** statement in SAS or the **test** command in Stata. **ses(1)** – The reference group is level 3 (see the Categorical Variables Codings table above), so this coefficient represents the difference between level 1 of **ses** and level 3. Note: The number in the parentheses indicates the number of the level of the variable, not the value used to code the level. This is important to remember for interpretation. For example, if you changed the reference group from level 3 to level 1 the labeling of the dummy variables in the output would not change. **ses(2)** – The reference group is level 3 (see the Categorical Variables Codings table above), so this coefficient represents the difference between level 2 of **ses** and level 3. Note: The number in the parentheses only indicate the number of the level of the variable. It does not tell you anything about which levels of the categorical variable are being compared. For example, if you changed the reference group from level 3 to level 1 the labeling of the dummy variables in the output would not change. **constant** – This is the expected value of the log-odds of **honcomp** when all of the predictor variables equal zero. In most cases, this is not interesting. Also, oftentimes zero is not a realistic value for a variable to take.

k. **S.E.** – These are the standard errors associated with the coefficients. The standard error is used for testing whether the parameter is significantly different from 0; by dividing the parameter estimate by the standard error you obtain a t-value. The standard errors can also be used to form a confidence interval for the parameter.

l. **Wald** and **Sig.** – These columns provide the Wald chi-square value and 2-tailed p-value used in testing the null hypothesis that the coefficient (parameter) is 0. If you use a 2-tailed test, then you would compare each p-value to your preselected value of alpha. Coefficients having p-values less than alpha are statistically significant. For example, if you chose alpha to be 0.05, coefficients having a p-value of 0.05 or less would be statistically significant (i.e., you can reject the null hypothesis and say that the coefficient is significantly different from 0). If you use a 1-tailed test (i.e., you predict that the parameter will go in a particular direction), then you can divide the p-value by 2 before comparing it to your preselected alpha level. For the variable **read**, the p-value is .000, so the probability of obtaining this chi-square value (31.564), or one larger, is less than .000 if in fact **read** has no effect in the population. If you were using an alpha of 0.05, **read** would be rejected. For the variable **ses**, the p-value is .053, so the null hypothesis that the coefficient equals 0 would be rejected. Because the test of the overall variable is statistically significant, you can look at the one degree of freedom tests for the dummies **ses(1)** and **ses(2)**. The dummy **ses(1)** is not statistically significantly different from the dummy **ses(3)** (which is the omitted, or reference, category), but the dummy **ses(2)** is statistically significantly different from the dummy **ses(3)** with a p-value of 0222.

m. **df** – This column lists the degrees of freedom for each of the tests of the coefficients.

n. **Exp(B)** – These are the odds ratios for the predictors. They are the exponentiation of the coefficients. There is no odds ratio for the variable **ses** because **ses** (as a variable with 2 degrees of freedom) was not entered into the logistic regression equation.

Odds Ratios

In this next example, we will illustrate the interpretation of odds ratios. In this example, we will simplify our model so that we have only one predictor, the binary variable **female**. Before we run the logistic regression, we will use the **crosstabs** command to obtain a crosstab of the two variables.

```
crosstabs female by honcomp.
```

female * honcomp Crosstabulation

Count			
		honcomp	
		.00	1.00
female .00		74	21
	1.00	77	28
Total		151	49

If we divide the number of males who are in honors composition, 18, by the number of males who are not in honors composition, 73, we get the odds of being in honors composition for males, 18/73 = .246. If we do the same thing for females, we get 28/.541 = .472. To get the odds ratio, which is the ratio of the two odds (.246 and .472), we would divide .246 by .472 and get .521. As we will see in the output below, this is exactly the odds ratio we obtain from the logistic regression. The thing to remember here is that we want the group coded as 1 over the group coded as 0, so honcomp=1/honcomp=0 for both males and females, and then the odds for females/odds for males, because the females are coded as 1.

You can get the odds ratio from the **crosstabs** command by using the **/statistics risk** subcommand, as shown below.

```
crosstabs female by honcomp
/statistics risk.
```

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for female (.00 / 1.00)	.521	.267	1.020
For cohort honcomp = .00	.814	.667	.993
For cohort honcomp = 1.00	1.561	.975	2.501
N of Valid Cases	200		

As you can see in the output below, we get the same odds ratio when we run the logistic regression. (NOTE: Although it is equivalent to the odds ratio estimated from the logistic regression, the odds ratio in the "Risk Estimate" table is calculated as the ratio of the odds of honcomp=0 for males over the odds of honcomp=0 for females, which is the opposite of what we did above. Because we want to interpret the odds ratio for females relative to males, we would not want this to include the value of 1. When we were considering the coefficients, we did not want the confidence interval to include 0. If we exponentiate 0, we get 1 (exp(0) = 1. Hence, this is two ways of saying the same thing. As you can see, the 95% confidence interval includes 1; hence, the odds ratio is not statistically significant. Because the lower bound of the 95% confidence interval is so close to 1, the p-value is very close to .05. We can use the **/print = n(95)** subcommand to get the 95% confidence intervals included in our output.

There are a few other things to note about the output below. The first is that although we have only one predictor variable, the test for the odds ratio does not match with the overall test of the model. This is because the test of the coefficient is a Wald chi-square test, while the test of the overall model is a likelihood ratio chi-square test. While these two types of chi-square tests are asymptotically equivalent, in small samples they can differ, as they do here. Also, we have the unfortunate situation in which the results of the two tests give different conclusions. This does not happen very often. In a situation like this, it is difficult to know what to conclude. One might consider the power, or one might decide if an odds ratio of this magnitude is important from a clinical or practical standpoint.

```
logistic regression honcomp with female
/print = ci(95).
```

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	2.633	1	.105
	Block	2.633	1	.105
	Model	2.633	1	.105

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	226.379[a]	.013	.018

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

Classification Table[a]

Observed		Predicted		
		honcomp		Percentage Correct
		.00	1.00	
Step 1 honcomp	.00	151	0	100.0
	1.00	49	0	.0
Overall Percentage				75.5

a. The cut value is .500

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
							Lower	Upper
Step 1[a] female	.521	.321	2.633	1	.105	1.684	.897	3.160
Constant	-1.359	.267	25.884	1	.000	.257		

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