Logistic Regression in SPSS

Based on *dale_jan.sav*, data collected by two Macquarie Masters students.

```
freq vars=wktotrec.
```

WKTOTREC

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 None	33	9.2	9.2	9.2
	1 >0 -> 4	114	31.8	31.8	41.1
	2 >4 -> 8	118	33.0	33.0	74.0
	3 >8 -> hi	93	26.0	26.0	100.0
	Total	358	100.0	100.0	

For this example, the variable *wktotrec* is recoded into two categories, one containing those who do 0 to 4 hours exercise per week and the other containing those who do more than four hours exercise a week.

```
recode wktotrec (0,1=0)(2,3=1) into wktot2.
freq vars=wktot2.
```

WKTOT2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 0->4	147	41.1	41.1	41.1
	1 >4+	211	58.9	58.9	100.0
	Total	358	100.0	100.0	

```
print format wktot2 (f1).
value labels wktot2 0 '0->4' 1 '>4+'.
```

The variable *enjoype* is recoded for this example into three categories.

```
recode enjoype (1,2=2)(3=1)(4,5=0) into enjoyrec.
print format enjoyrec (f1).
value labels enjoyrec 0 'Never or rarely' 1 'Sometimes'
  2 'Always or often'.
crosstabs wktot2 by enjoyrec nesb/cells=count column/statistics=chisq.
```

Crosstab

			ENJOYREC			
			0 Never or rarely	1 Sometimes	2 Always or often	Total
WKTOT2	0 0->4	Count	26	73	48	147
		% within ENJOYREC	49.1%	53.3%	28.6%	41.1%
	1 >4+	Count	27	64	120	211
		% within ENJOYREC	50.9%	46.7%	71.4%	58.9%
Total		Count	53	137	168	358
		% within ENJOYREC	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	20.687[a]	2	.000
Likelihood Ratio	20.986	2	.000
Linear-by-Linear Association	14.240	1	.000
N of Valid Cases	358		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 21.76.

There is a highly significant relationship between enjoyment of physical education and amount of exercise. However, it isn't entirely consistent. For students who 'never' or 'rarely' enjoy physical education, the odds of doing more than four hours exercise per week are 27 to 26 (27/26 = 1.039). The corresponding figures for those who 'sometimes' enjoy physical activity and those who 'always' or 'often' enjoy it are 64 to 73 (64/73 = .877) and 120 to 48 (120/48 = 2.5) respectively. In other words, students who 'sometimes' enjoy physical education are a bit less likely to do more than four hours exercise per week than those who 'never' or 'rarely' enjoy physical education. But the difference is not great: the ratio of the two odds (the OR) is .877/1.039 = .844, not all that different from one. If the odds for students who 'often' or 'always' enjoy physical education are compared to those for those who 'never' or 'rarely' enjoy physical education, a much higher OR is obtained: 2.5/1.039 = 2.406.

```
logistic regression wktot2 with enjoyrec/
 categorical=enjoyrec/
 contrast(enjoyrec)=indic(1)/
 method=enter.
```

Point-and-click

- Analysis➔Regression➔Binary Logistic
- Select *wktot2* as Dependent, *enjoyrec* as Covariate
- Click on Categorical, select *enjoyrec.* Select *Indicator* as Contrast. Select *First* as Reference Category, click on Change. Click on *Continue*.
- Select *Enter* as Method
- Click on *OK*

This logistic regression analysis with ***wktot2*** as the outcome and ***enjoyrec*** as the only predictor is equivalent to the crosstabs analysis above. The ***contrast*** sub-command asks that ***enjoyrec*** be dummy- (or indicator-) coded, with the lowest-numbered category, 'never' or 'rarely', as the reference category.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	358	100.0
	Missing Cases	0	.0
	Total	358	100.0
Unselected Cases		0	.0
Total		358	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->4	0
1 >4+	1

Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
ENJOYREC	0 Never or rarely	53	.000	.000
	1 Sometimes	137	1.000	.000
	2 Always or often	168	.000	1.000

Classification Table[a,b]

Observed			Predicted WKTOT2 0->4	Predicted WKTOT2 >4+	Percentage Correct
Step 0	WKTOT2	0->4	0	147	.0
		>4+	0	211	100.0
	Overall Percentage				58.9

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	.361	.107	11.318	1	.001	1.435

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	ENJOYREC	20.687	2	.000
		ENJOYREC(1)	13.701	1	.000
		ENJOYREC(2)	20.405	1	.000
	Overall Statistics		20.687	2	.000

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	20.986	2	.000
	Block	20.986	2	.000
	Model	20.986	2	.000

The Model Chi-Square, 20.986, is the same as the likelihood-ratio chi-square obtained in the crosstabs analysis. In this case, it is calculated as the difference between the -2LL for a model with only a constant (484.790) and the -2LL for the model with *enjoyrec* (463.804). This tells us that, in terms of predicting *wktot2*, the model containing *enjoyrec* is a significant improvement over the model with just a constant.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	463.804	.057	.077

The two R Square values are analogous to the R^2 values given for ordinary least-squares regression. They are based on the log likelihood.

Classification Table[a]

Observed			Predicted		
			WKTOT2		Percentage Correct
			0->4	>4+	
Step 1	WKTOT2	0->4	73	74	49.7
		>4+	64	147	69.7
	Overall Percentage				61.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ENJOYREC			20.179	2	.000	
	ENJOYREC(1)	-.1693	.324	.273	1	.601	.844
	ENJOYREC(2)	.8786	.324	7.374	1	.007	2.407
	Constant	.0377	.275	.019	1	.891	1.038

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

The Equation

The regression equation shown above is

Log odds(*wktot2* = 1) = B_0 + B_1 x *enjoyrec*(1)+ B_2 x *enjoyrec*(2) (Equation 1)

or, with the actual values of the regression coefficients,

Log odds(*wktot2* = 1) = .0377 - .1693 x *enjoyrec*(1) + .8786 x *enjoyrec*(2)

where *enjoyrec*(1) is the dummy variable for the second category of *enjoyrec* ('sometimes') and *enjoyrec*(2) is the dummy variable for the third category of *enjoyrec* ('always' or 'often'). Thus, for example, the log odds of *wktot2* equalling one for subjects in the reference category of *enjoyrec* ('never' or 'rarely') is

.0377 - .1693 x 0 + .8786 x 0 = .0377.

The odds are therefore $e^{.0377}$, or 1.038, which agrees with the figure calculated from the *crosstabs* table above (*e* is the base of the natural logarithms, 2.71828; your calculator should have an e^x key to do this calculation). For the second category the log odds of *wktot2* equalling one are

.0377 - .1693 x 1 + .8786 x 0 = -.1316

which translates to odds of .877, which also agrees with the previous calculation.

Working with log odds and odds is all very well, but sometimes it would be useful to estimate the probability that *wktot2* is equal to one. The formula for this is:

$$P(wktot2 = 1) = e^{[B0 + B1 \times enjoyrec(1) + B2 \times enjoyrec(2)]} / (1 + e^{[B0 + B1 \times enjoyrec(1) + B2 \times enjoyrec(2)]}) \quad \text{(Equation 2)}$$

If we take the result of the calculation of the log odds for the reference group, carried out above (.0377) and put it in Equation 2, we get

$$P(wktot2 = 1) = e^{.0377} / (1 + e^{.0377}) = .5094$$

which agrees with the percentage in the *crosstabs* table of *wktot2* by *enjoyrec* above.

The Wald Statistic

The Wald statistic for a coefficient is the square of the result of dividing the coefficient by its standard error; this quantity is distributed as chi-squared. The statistic is also calculated for a group of variables which define a "factor".

Exp(B) - The estimated odds ratio

The final column of the output shows Exp(B), e^B, which is an estimate of the odds ratio. The figures given in this column are equivalent to those calculated for the crosstabulation of *wktot2* and *enjoyrec* earlier.

Version 10 of SPSS optionally calculates confidence intervals for the odds ratios. (Click on Options➔CIforexp(B)in point-and-click, add */options=CI(95)* in syntax). You can do it yourself by multiplying the standard error of the original coefficient by 1.96 (assuming 95% confidence intervals are required), subtracting the result from the coefficient to get the lower bound and adding it to the coefficient to get the upper bound, then calculating e^x for both values to get the bounds for the odds ratio. The process for the contrast between the hightest category of *enjoyrec* ('always' or 'often') and the reference category ('never' or 'rarely') is as follows:

Multiply the s.e. by 1.96: .3235 x 1.96 = .6341

Subtract the result from the coefficient: .8786 - .6341 = .2445
Add the result to the coefficient: .8786 + .6341 = 1.5127

The lower bound for the CI is $e^{.2445}$ = 1.277
The Upper limit for the CI is $e^{1.5127}$ = 4.539

Note that the interval for the estimate of the OR does not contain one.

```
logistic regression wktot2 with enjoyrec age/
  categorical=enjoyrec/
  contrast(enjoyrec)=indic(1)/
  method=enter.
```

This example is very similar to the last one, except that age, measured in years, has been added.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	21.181	3	.000
	Block	21.181	3	.000
	Model	21.181	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	463.609	.057	.077

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ENJOYREC			19.738	2	.000	
	ENJOYREC(1)	-.177	.324	.299	1	.585	.838
	ENJOYREC(2)	.864	.325	7.062	1	.008	2.373
	AGE	-.032	.072	.195	1	.659	.969
	Constant	.494	1.071	.213	1	.644	1.639

a. Variable(s) entered on step 1: ENJOYREC, AGE.

The inclusion of *age* has very little effect: Its OR is very close to one, and it is not significant. Also, its inclusion has very little effect on the result for *enjoyrec*. If we did want to interpret the OR for *age*, though, what would it mean? As always, the coefficient, and therefore the corresponding OR, shows the effect of a one-unit increase, i.e., of being a year older. In this case the OR is less than one, indicating that as students get older, the odds of their doing more than four hours exercise a week become smaller.