Logistic Regression

Logistic regression is a variation of the regression model. It is used when the dependent response variable is binary in nature. Logistic regression predicts the probability of the dependent response, rather than the value of the response (as in simple linear regression).

In this example, the dependent variable is frequency of sex (less than once per month versus more than once per month). In this case, we are predicting having sex more than once per month.

```
LOGISTIC REGRESSION  freqdum
  /METHOD = ENTER age Married White attend happy Male
  /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1052	38.0
	Missing Cases	1713	62.0
	Total	2765	100.0
Unselected Cases		0	.0
Total		2765	100.0

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

Dependent Variable Encoding

Original Value	Internal Value
Less than of equal to 1/month	0
More than 1/month	1

This table informs you of how the procedure handled the dichotomous dependent variable, which helps you to interpret the values of the parameter coefficients. Here, "less than or equal to once per month" was coded as a 0, while "more than once a month" was coded as a 1.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	311.100	6	.000
	Block	311.100	6	.000
	Model	311.100	6	.000

The omnibus tests are measures of how well the model performs.

The chi-square statistic is the change in the -2 log-likelihood from the previous step, block, or model. If the step was to remove a variable, the exclusion makes sense if the significance of the change is large (i.e., greater than 0.10). If the step was to add a variable, the inclusion makes sense if the significance of the change is small (i.e., less than 0.05). In this example, the change is from Block 0, where no variables are entered.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1125.821[a]	.256	.344

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

The R-Square statistic cannot be exactly computed for logistic regression models, so these approximations are computed instead. Larger pseudo r-square statistics indicate that more of the variation is explained by the model, to a maximum of 1.

Classification Table[a]

			Predicted		
			SEXFREQ (Monthly+ =1)		
	Observed		Less than of equal to 1/month	More than 1/month	Percentage Correct
Step 1	SEXFREQ (Monthly+ =1)	Less than of equal to 1/month	262	189	58.1
		More than 1/month	109	492	81.9
	Overall Percentage				71.7

a. The cut value is .500

The classification table helps you to assess the performance of your model by crosstabulating the observed response categories with the predicted response categories. For each case, the predicted response is the category treated as 1, if that category's predicted probability is greater than the user-specified cutoff. Cells on the diagonal are correct predictions.

Variables in the Equation

		B -b-	S.E. -b-	Wald -c-	df	Sig.	Exp(B) -d-
Step 1[a]	age	-.061	.005	145.748	1	.000	.941
	Married	1.698	.167	103.127	1	.000	5.465
	White	-.149	.178	.699	1	.403	.862
	attend	-.059	.029	4.315	1	.038	.942
	happy	-.318	.123	6.723	1	.010	.727
	Male	.444	.148	8.951	1	.003	1.558
	Constant	3.047	.382	63.773	1	.000	21.054

a. Variable(s) entered on step 1: age, Married, White, attend, happy, Male.

b. B is the estimated coefficient, with standard error, S.E.

c. The ratio of B to S.E., squared, equals the Wald statistic. If the Wald statistic is significant (i.e., less than 0.05) then the parameter is useful to the model.

d. "Exp(B)," or the odds ratio, is the predicted change in odds for a unit increase in the predictor. The "exp" refers to the exponential value of B. When Exp(B) is less than 1, increasing values of the variable correspond to decreasing odds of the event's occurrence. When Exp(B) is greater than 1, increasing values of the variable correspond to increasing odds of the event's occurrence.

If you subtract 1 from the odds ratio and multiply by 100, you get the percent change in odds of the dependent variable having a value of 1. For example, for age:

= 1 - (.941) = .051

= .051 * 100 = 5.1%

The odds ratio for age indicates that every unit increase in age is associated with a 5.1% decrease in the odds of having sex more than once a month.

<u>Regression Equation</u>

$FREQDUM_{PREDICTED}$ = 3.047 - .061*age – 1.698*married - .149*white – .059*attend - .318*happiness + .444*male

If we plug in values in for the independent variables (age = 35 years; married = currently married-1; race = white-1; happiness = very happy-1; church attendance = never attends-0; gender = male-1), we can predict a value for frequency of sex:

$FREQDUM_{predicted}$ = 3.047 - .061*35 – 1.698*1 - .149*1 – .059*0 - .318*1 + .444*1 = 2.587

As this variable is coded, a 35-year old, White, married person with high levels of happiness and who never attends church would be expected to report their frequency of sex between values 4 (weekly) and 5 (2-3 times per week).

If we plug in 70 years, instead, we find that frequency of sex is predicted at 2.455, or approximately 1-2 times per month.

Interpretation

Recall: When Exp(B) is less than 1, increasing values of the variable correspond to decreasing odds of the event's occurrence. When Exp(B) is greater than 1, increasing values of the variable correspond to increasing odds of the event's occurrence.

Constant = Not interpretable in logistic regression.

Age = Increasing values of age correspond with decreasing odds of having sex more than once a month.

Marital = Married persons have a decreased odds of having sex more than once a month.

Race = White persons have a decreased odds of having sex more than once a month. Notice that this variable, however, is not significant.

Church Attendance = Increasing values of church attendance correspond with decreasing odds of having sex more than once a month.

Happiness = Increasing values of general happiness correspond with decreasing odds of having sex more than once a month. Recall that happiness is coded such that higher values indicate less happiness.

Logistic Regression (with non-linear variable)

It is known that some variables are often non-linear, or curvilinear. Such variables may be age or income. In this example, we include the original age variable and an age squared variable.

```
LOGISTIC REGRESSION  freqdum
  /METHOD = ENTER age Married White attend happy Male agesquare
  /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	.036	.027	1.823	1	.177	1.037
	Married	1.625	.172	89.747	1	.000	5.079
	White	-.155	.178	.757	1	.384	.857
	attend	-.057	.029	3.933	1	.047	.944
	happy	-.335	.123	7.427	1	.006	.715
	Male	.492	.150	10.732	1	.001	1.635
	agesquare	-.001	.000	13.055	1	.000	.999
	Constant	1.034	.658	2.465	1	.116	2.812

a. Variable(s) entered on step 1: age, Married, White, attend, happy, Male, agesquare.

The age squared variable is significant, indicating that age is non-linear.

Logistic Regression (with interaction term)

To test for two-way interactions (often thought of as a relationship between an independent variable (IV) and dependent variable (DV), moderated by a third variable), first run a regression analysis, including both independent variables (IV and moderator) and their interaction (product) term. It is highly recommended that the independent variable and moderator are standardized before calculation of the product term, although this is not essential. For this example, two dummy variables were created, for ease of interpretation. Sex was recoded such that 1=Male and 0=Female. Marital status was recoded such that 1=Currently married and 0=Not currently married. The interaction term is a product of these two dummy variables.

Regression Model (*without* interactions)

```
LOGISTIC REGRESSION  freqdum
  /METHOD = ENTER age White attend happy Male Married
  /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-.061	.005	145.748	1	.000	.941
	White	-.149	.178	.699	1	.403	.862
	attend	-.059	.029	4.315	1	.038	.942
	happy	-.318	.123	6.723	1	.010	.727
	Male	.444	.148	8.951	1	.003	1.558
	Married	1.698	.167	103.127	1	.000	5.465
	Constant	3.047	.382	63.773	1	.000	21.054

a. Variable(s) entered on step 1: age, White, attend, happy, Male, Married.

Regression Model (*with* interactions)

```
LOGISTIC REGRESSION freqdum
  /METHOD = ENTER age White attend happy Male Married Male*Married
  /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-.060	.005	139.187	1	.000	.942
	White	-.173	.179	.930	1	.335	.841
	attend	-.056	.029	3.838	1	.050	.945
	happy	-.322	.123	6.838	1	.009	.724
	Male	.649	.193	11.262	1	.001	1.913
	Married	1.936	.222	75.722	1	.000	6.931
	Male by Married	-.504	.302	2.774	1	.096	.604
	Constant	2.929	.387	57.164	1	.000	18.704

a. Variable(s) entered on step 1: age, White, attend, happy, Male, Married, Male * Married .

The product term should be significant in the regression equation in order for the interaction to be interpretable. In this example, the interaction term is significant at the $p < 0.1$ level.

Regression Equation

$FREQDUM_{predicted}$ = 2.93 -.06*age - .17*White -.32*happy - .06*attend + 1.94*married + (.65 - .50*married) * male

Interpretation

Main Effects

The married coefficient represents the main effect for females (the 0 category). The effect for females is then 1.94, or the "marital" coefficient. The effect for males is 1.94 - .50, or 1.44.

The gender coefficient represents the main effect for unmarried persons (the 0 category). The effect for unmarried is then .65, or the "sex" coefficient. The effect for married is .65 - .50, or .15.

Interaction Effects

For a simple interpretation of the interaction term, plug values into the regression equation above.

Married Men = $FREQDUM_{predicted}$ = 2.93 -.06*age - .17*White -.32*happy - .06*attend + 1.94**1* + (.65 - .50**1*) * *1* = 2.43
Married Women = $FREQDUM_{predicted}$ = 2.93 -.06*age - .17*White -.32*happy - .06*attend + 1.94**1* + (.65 - .50**1*) * *0* = 2.28
Unmarried Men = $FREQDUM_{predicted}$ = 2.93 -.06*age - .17*White -.32*happy - .06*attend + 1.94**0* + (.65 - .50**0*) * *1* = 0.49

Unmarried Women = $FREQDUM_{predicted}$ = 2.93 -.06*age - .17*White -.32*happy - .06*attend + 1.94**0* + (.65 - .50**0*) * *0* = 0.34

In this example (age = 35 years; race = white-1; happiness = very happy-1; church attendance = never attends-0), we can see that (1) for both married and unmarried persons, males are reporting higher frequency of sex than females, and (2) married persons report higher frequency of sex than unmarried persons. The interaction tells us that the gender difference is greater for married persons than for unmarried persons.

Odds Ratios

Using "married" as the focus variable, we can say that the effect of being married on having sex more than once per month is greater for females.

Females: $e^{1.936}$ = 6.93

Males: $e^{1.432}$ = 4.20

Using "gender" as the focus variable, we can say that the effect of being male on having sex more than once per month is greater for marrieds.

Marrieds: $e^{0.15}$ = 1.16

Unmarrieds: $e^{0.65}$ = 1.92