Logistic Regression

- Logistic regression is used to analyze relationships between a dichotomous dependent variable and metric or dichotomous independent variables.
- Logistic regression combines the independent variables to estimate the probability that a particular event will occur, i.e. a subject will be a member of one of the groups defined by the dichotomous dependent variable. In SPSS, the model is always constructed to predict the group with higher numeric code. If responses are coded 1 for Yes and 2 for No, SPSS will predict membership in the No category. If responses are coded 1 for No and 2 for Yes, SPSS will predict membership in the Yes category. We will refer to the predicted event for a particular analysis as the modeled event.

What logistic regression predicts

- The variate or value produced by logistic regression is a probability value between 0.0 and 1.0.
- If the probability for group membership in the modeled category is above some cut point (the default is 0.50), the subject is predicted to be a member of the modeled group. If the probability is below the cut point, the subject is predicted to be a member of the other group.
- For any given case, logistic regression computes the probability that a case with a particular set of values for the independent variable is a member of the modeled category.

Level of measurement requirements

- Logistic regression analysis requires that the dependent variable be dichotomous.
- Logistic regression analysis requires that the independent variables be metric or dichotomous.
- If an independent variable is nominal level and not dichotomous, the logistic regression procedure in SPSS has a option to dummy code the variable for you.

Assumptions and Sample Size Requirements

- Logistic regression does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variables.

Sample size requirements

- The minimum number of cases per independent variable is 10.
- For preferred case-to-variable ratios, we will use 20 to 1 for simultaneous and hierarchical logistic regression and 50 to 1 for stepwise logistic regression.

Methods for including variables

- There are many methods available for including variables in the regression equation:
  - the simultaneous method in which all independents are included at the same time
  - The stepwise method (forward conditional in SPSS) in which variables are selected in the order in which they maximize the statistically significant contribution to the model.
- For all methods, the contribution to the model is measures by model chi-square is a statistical measure of the fit between the dependent and independent variables, like R^2.

Logistic Regression with 1 Predictor

- Response - Presence/Absence of characteristic
- Predictor - Numeric variable observed for each case
- Model - $\pi(x) \equiv$ Probability of presence at predictor level x

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

- $\beta = 0 \Rightarrow$ P(Presence) is the same at each level of x
- $\beta > 0 \Rightarrow$ P(Presence) increases as x increases
- $\beta < 0 \Rightarrow$ P(Presence) decreases as x increases

Logistic Regression with 1 Predictor

- α, β are unknown parameters and must be estimated using statistical software such as SPSS, SAS, or STATA
- Primary interest in estimating and testing hypotheses regarding β
  - Large-Sample test (Wald Test):
  - H_0: $\beta = 0$ H_A: $\beta \neq 0$

$$T.S.: X^2_{obs} = \left(\frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \right)^2$$

$$R.R.: X^2_{obs} \geq \chi^2_{\alpha,1}$$

$$P-val: P(\chi^2 \geq X^2_{obs})$$

Example - Rizatriptan for Migraine

- Response - Complete Pain Relief at 2 hours (Yes/No)
- Predictor - Dose (*mg*): Placebo (0),2.5,5,10

Dose	# Patients	# Relieved	% Relieved
0	67	2	3.0
2.5	75	7	9.3
5	130	29	22.3
10	145	40	27.6

Source: Gijsmant, et al (1997)

Example - Pain Relief

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	DOSE	.165	.037	19.819	1	.000	1.180
	Constant	-2.490	.285	76.456	1	.000	.083

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

Dependent variable:
Complete Pain Relief at 2 hours (Yes/No)

Independent variable
Dose (*mg*): Placebo (0),2.5,5,10

$$\hat{\pi}(x) = \frac{e^{-2.490+0.165x}}{1+e^{-2.490+0.165x}}$$

$$H_0: \beta = 0 \quad H_A: \beta \neq 0$$

$$T.S.: X^2_{obs} = \left(\frac{0.165}{0.037}\right)^2 = 19.819$$

$$RR: X^2_{obs} \geq \chi^2_{.05,1} = 3.84$$

$$P-val: .000$$

Odds Ratio

- Interpretation of Regression Coefficient (b):
  - In linear regression, the slope coefficient is the change in the mean response as x increases by 1 unit
  - In logistic regression, we can show that:

$$\text{odds}(x) = e^{\beta} = \frac{\pi(x)}{1-\pi(x)}$$

• Thus e^{β} represents the change in the odds of the outcome (multiplicatively) by increasing x by 1 unit

• If $\beta = 0$, the odds and probability are the same at all x levels ($e^{\beta}=1$)

• If $\beta > 0$, the odds and probability increase as x increases ($e^{\beta}>1$)

• If $\beta < 0$, the odds and probability decrease as x increases ($e^{\beta}<1$)

95% Confidence Interval for Odds Ratio

- Step 1: Construct a 95% CI for β:

$$\hat{\beta} \pm 1.96\hat{\sigma}_{\hat{\beta}} \equiv \left(\hat{\beta} - 1.96\hat{\sigma}_{\hat{\beta}}, \hat{\beta} + 1.96\hat{\sigma}_{\hat{\beta}}\right)$$

• Step 2: Raise $e = 2.718$ to the lower and upper bounds of the CI:

$$\left(e^{\hat{\beta}-1.96\hat{\sigma}_{\hat{\beta}}}, e^{\hat{\beta}+1.96\hat{\sigma}_{\hat{\beta}}}\right)$$

• If entire interval is above 1, conclude positive association

• If entire interval is below 1, conclude negative association

• If interval contains 1, cannot conclude there is an association

Example - Pain Relief

• 95% CI for β:

$$\hat{\beta} = 0.165 \qquad \hat{\sigma}_{\hat{\beta}} = 0.037$$

$$95\%\ CI: \quad 0.165 \pm 1.96(0.037) \quad \equiv \quad (0.0925, 0.2375)$$

• 95% CI for population odds ratio:

$$\left(e^{0.0925}, e^{0.2375}\right) \equiv (1.10, 1.27)$$

• Conclude positive association between dose and probability of complete relief

Example - Death Penalty for Crime

- We can compute the odds of receiving a death penalty for each of the groups:
- The odds of receiving a death sentence if the defendant was Black = 28/45 = 0.6222
- The odds of receiving a death sentence if the defendant was not Black = 22/52 = 0.4231

	Blacks	Nonblacks	Total
Death sentence	28	22	50
Life imprisonment	45	52	97
Total	73	74	147

Example - Death Penalty for Crime

Which we interpret as:

- Blacks are 1.47 times more likely to receive a death sentence as non blacks
- The risk of receiving a death sentence are 1.47 times greater for blacks than non blacks
- The odds of a death sentence for blacks are 47% higher than the odds of a death sentence for non blacks. (1.47 - 1.00)
- The predicted odds for black defendants are 1.47 times the odds for non black defendants.
- A one unit change in the independent variable race (nonblack to black) increases the odds of receiving a death penalty by a factor of 1.47.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	BLACKD	.386	.350	1.213	1	.271	1.471
	Constant	-.860	.254	11.439	1	.001	.423

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

Multiple Logistic Regression

- Extension to more than one predictor variable (either numeric or dummy variables).
- With *p* predictors, the model is written:

$$\pi = \frac{e^{\alpha+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\alpha+\beta_1 x_1+\cdots+\beta_p x_p}}$$

• Adjusted Odds ratio for raising x_i by 1 unit, holding all other predictors constant:

$$OR_i = e^{\beta_i}$$

• Inferences on β_i and OR_i are conducted as was described above for the case with a single predictor