QUANTITATIVE METHODS CLASSES

WEEK SEVEN

The regression models studied in previous classes assume that the response variable is quantitative. Often, however, we wish to study social processes that lead to two different outcomes. In these instances, then, the response variable is categorical. For example, when we study unemployment, marriage or voter's choice.

In the case of unemployment, for example, our respondents will either be employed or unemployed. Denote the response on Y by 1 if unemployed and 0 if employed (it is also common to use the term failure and success). The sum of the scores in the sample is then the number of successes (i.e., unemployed respondents). The mean of this response variable (the 0s and 1s scores) equals the proportion of successes (i.e., the proportion of unemployed respondents). Obviously, then, the proportion of employed respondents equals 1-that mean.

Transforming the categorical response variable (0,1) to proportion allows us to think in terns of regression analysis, since the ordinary regression models the mean of the response variable. Let π denote the probability of success, and it is possible to write the following linear equation: $\pi = a + b(X)$

This is the linear probability model, and it implies that the probability of success is a linear function of X.

Unfortunately, this model is often poor. First, it implies probabilities below 0 and above 1, whereas probabilities must fall between 0 and 1. Second, the response variable is not normally distributed, and thus it violates some of the assumptions we make when applying OLS regression. Thus, we need to further 'transform' our response variable.

In other words, we need to describe the relationship between π and X with a *curvilinear* rather than a *linear* function. This can be achieved by the following equation:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X$$

The ratio π/(1- π) equals the ***odds.*** Thus, for example, when the proportion of unemployed individuals (success) in our sample equals 0.20 the odds equals 0.25 (0.2/0.8=0.25), which means that a success is four times less likely as failure.

This equation uses the natural log of the odds, and is called the ***logistic transformation***, or ***logit*** for short. Thus, as π increases from 0 to 1, the odds increases from 0 to +∞, and the logit increases from –∞ to +∞.

Table 1: Ranges of Probability, Odds and Log Odds

		Lowest Level	Mid point	Highest Level
Probability	π	0	.5	1
Odds	π /1- π	0	1	+∞
Log Odds	log (π /1- π) or logit (π)	–∞	0	+∞

Logistic Regression

The model: logit (π) =a+bX is called the logistic regression model.

In logistic regression the parameters of the model are estimated using the **maximum-likelihood method**. That is, the coefficients that make the observed results most likely are selected. For each possible value a parameter might have, SPSS computes the probability that the observed value would have occurred if it were the true value of the parameter. Then, for the estimate, it picks the parameter for which the probability of the actual observation is greatest.

The equation for logistic regression may be given in either the additive or multiplicative forms.

Additive form:

$\log (\pi/1-\pi) = a + ßX$

Multiplicative form:

$\pi/1-\pi = \exp^{(a)}*\exp^{(}ßX^{)}$

π is the proportion with the characteristic (the probability), a is a constant, $ß_1,ß_2$.... are coefficients and X_1,X_2.... are predictor variables. log $\pi/1-\pi$ is known as the log-odds and $\pi/1-\pi$ as the odds. Exponential (ß), are the odds multipliers and interest is in values that differ from 1.

Running Logistic Regression in SPSS

Here we model the probability of being unemployed rather than being employed (EMP86). First, we have to make sure that EMP86 is coded 1 and 0. It is important to code the success as 1. As we are interested in predicting unemployment, the unemployed should be coded 1 and the employed coded 0 (we simply create a new variable UNEMP: 0=employed and 1=unemployed).

We are going to use two predictors: *Class* and *Age*. As *Class* is a categorical variable, we need to recode class to a three-category variable (CLASS1: 10,20=1 (prof); 31,41,42,43,50=2 (inter); and, 32,60,71,72=3 (working)), and then to create dummy variables for *Prof, Inter* and *working*.

```
logistic regression variables=unemp
        /method = enter prof inter.
```

Interpreting the Results of a Logistic Regression Model

1. Assessing the Goodness of Fit of the Model

One way of assessing goodness of fit is to examine how ‘likely’ the sample results are, given the parameter estimates (remember the model attempts to generate the parameter estimates that make the results most likely).

The probability of the observed results given the parameter estimates is known as the **Likelihood.** Since the likelihood is a small number less than 1, it is customary to use -2 times the log likelihood (-2LL) as an estimate of how well the model fits the data. A good model is one that results in a high likelihood of the observed results. This translates into a small value for –2LL (if a model fits perfectly, the likelihood=1 and –2LL=0)

In our model, -2 Log Likelihood = 949.379

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	949.379	.032	.052

This is far from zero, however because there is no upper boundary for –2LL it is difficult to make a statement about the meaning of the score. It is more often used to see whether adding additional variables to the model leads to a significant reduction in the –2LL.

The difference between the –2LL for two models, with the difference in the degrees of freedom (which is equal to the difference between the number of parameters for the two models) has a chi-square distribution. Thus, the significance of this change is derived from the chi-square table.

To assess the change between different models variables must be added in steps or blocks, as was done in the OLS regression.

If we compare the fit statistics for a model with just class and one which adds age we get the following results.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	12.515	1	.000
	Block	12.515	1	.000
	Model	45.119	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	936.864	.045	.071

The chi-square figure for the new block added in the second model shows a change of 12.52 (see omnibus tests table) for 1 additional degree of freedom (*AGE*). This is significant at the .05 level (you can check this on the Chi-sq (X^2) table). This test is comparable to the F-change test in the OLS regression.

2. The Coefficients.

	B	S.E.	Wald	Df	Sig.	Exp(B)
AGE	-.027	.008	12.104	1	.001	.974
PROF	-1.145	.228	25.134	1	.000	.318
INTER	-.722	.211	11.701	1	.001	.486
Constant	-.032	.296	.012	1	.913	.968

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

The **B**s refer to the log-odds of being unemployed. We can insert these into the logistic regression equation as was done in multiple regression.

Additive form:

$$\begin{aligned}\text{logit}(\pi) &= \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots.. + \beta_n X_n \\ &= -.0323 + (-.027)(AGE) + (-1.145)(PROF) + (-.722)(INTER)\end{aligned}$$

This tells us that increasing age decreases the log odds of unemployment (controlling for class). Being in the professional/managerial class or being in the intermediate class also reduces the log-odds of being unemployed relative to those in the working class.

The **Wald statistic** (B/SE)**.** Most software reports its square (B/SE)^2. The significance of the Wald statistic is reported in the column marked **Sig.** This shows that the three predictor variables are significant. If the coefficient is very large the Wald statistic can become unreliable so you should refer to the change in the log likelihood instead (see below).

However, log-odds is not a very straightforward concept. It is probably easier to use the multiplicative form of the equation using **exp(B),** see last column of the SPSS output**.** These are the 'Odds Multipliers'.

Multiplicative form

$$\begin{aligned}\frac{\pi}{(1-\pi)} &= e^{\alpha} \times e^{\beta_1 X_1} \ldots \times e^{\beta_n X_n} \\ &= e^{-0.323} \times e^{-0.027(\text{AGE})} \times e^{-1.145(\text{PROF})} \times e^{-0.722(\text{INTER})} \\ &= .968 \times 0.974(\text{AGE}) \times 0.318(\text{PROF}) \times 0.486(\text{INTER})\end{aligned}$$

Remember **interest is in coefficients that differ from 1**. Values greater than 1 indicate that the variable in question increases the odds of the dependent 'event' occurring and values less than 1 (i.e. between 0 and 1) indicate a decrease in the odds. Effectively the odds for the base category are set to 1.

Using the odds multipliers we can make the more understandable claims that, when other factors in the model are held constant:

- each added year of age leads to about 3% reduction in the odds of being unemployed.

- being in the professional class, *compared to being in the working class,* leads to about 68% reduction in the odds of being unemployed
- being in the intermediate class, *compared to being in the working class,* leads to about 51% reduction in the odds of being unemployed.

2. Estimating Probabilities

The original logistic regression equation can be transformed to show the estimated probability of success by:

$$\hat{\pi} = \frac{e^{(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)}}{1 + e^{(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)}}$$

so for any individual the probability of being unemployed can be calculated.

Example:
What is the estimated probability of being unemployed for a person aged 30 in the professional class?

$$\hat{\pi} = \frac{e^{(-0.32 + (0.27)(30) + (-1.145)(1)}}{1 + e^{(-0.32 + (0.27)(30) + (-1.145)(1)}} = \frac{e^{-1.987}}{1 + e^{-1.987}} = \frac{.137}{1.137} = .12$$

Answer: a 30 year old professional has an estimated probability of being unemployed of .12.