8. Introduction to Logistic Regression

Regression analysis enables you to characterise the relationship between a response variable and one or more regressors. In linear regression, the response variable is continuous. In *logistic regression*, the response variable is categorical. Depending on the definition of your categorical variable, you can choose one of the three types of the logistic regression models.

- If the response variable is dichotomous (two categories), the appropriate logistic regression model is **binary logistic** regression.

If you have more than two categories (levels) within the response variable, then there are two possible logistic regression models:

- If there are three or more categories with no natural ordering to the levels, we fit a **nominal logistic** regression model.
- If the response variable is ordinal, an **ordinal logistic** regression model is used.

Using a Linear Probability Model

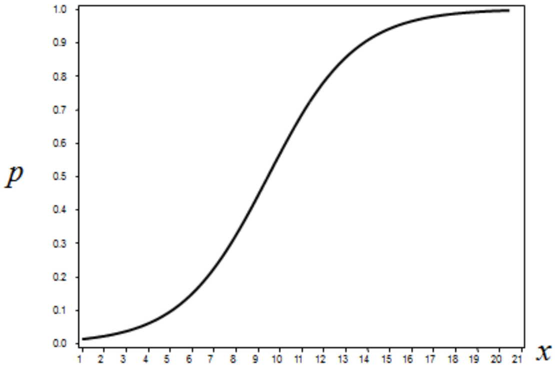
One might be tempted to analyse a regression model with a binary response variable using procedures specified for linear regressions (LS estimation of the model). However, there are problems with that. Besides the arbitrary nature of the coding, there is the problem that the predicted values will take on values that have no intrinsic meaning, with regards to your response variable. There is also the mathematical inconvenience of not being able to assume normality and constant variance when the response variable has only two values.

Therefore, instead of modelling the 0's and 1's directly, another way of thinking about modelling a binary variable is to model the probability of either the 0 or the 1. If you can model the probability of the 1 (call that *p*), then you have also modelled the probability of the 0, which would be (1 – *p*). Probabilities are truly continuous and so this line of thinking might sound compelling at first.

$$E(y=1|x)=p=\beta_1+\beta_2 x+\varepsilon$$

One problem is that the predicted values from a linear model can assume, theoretically, any value. However, probabilities are by definition bounded between 0 and 1.

Another problem is that the relationship between the probability of the outcome and a predictor variable is usually non-linear rather than linear. In fact, the relationship often resembles an S-shaped curve (a “sigmoidal” relationship). Probabilities do not have a random normal error associated with them, but rather a *binomial* error of $p(1-p)$. That error is greatest at probabilities close to 0.5 and lowest near 0 and 1. They do not have constant error associated with them.



This plot shows a model of the relationship between a continuous predictor and the probability of an event or outcome. The linear model clearly will not fit if this is the true relationship between *x* and the probability. In order to model this relationship directly, you must use a nonlinear function. You may recall that an appropriate function can be the **cdf** of a continuous random variable. One such function is the *logistic* cdf which implies:

$$E(y=1|x_i) = p_i = \frac{1}{1+e^{-(\beta_1+\beta_2 x_i)}}$$

(Alternatively, the cdf of a standard normal variable may be used which is called **Probit** model).

The parameter estimate of this curve determines the rate of increase or decrease of the estimated curve. When the parameter estimate is greater than 0, the probability of the outcome increases as the predictor variable values increase. When the parameter estimate is less than 0, the probability decreases as the predictor variable values increase. As the absolute value of the parameter estimate increases, the curve has a steeper rate of change. When the parameter estimate is equal to 0, the curve can be represented by a straight, horizontal line that shows an equal probability of the event for everyone.

Logit Transformation

The previous expression, in a simplified notation, can be re-written as:

$$p = \frac{1}{1+e^{-(\beta_1+\beta_2 x)}} = \frac{1}{1+e^{-z}} = \frac{1}{1+\frac{1}{e^z}} = \frac{e^z}{1+e^z}$$

$$1-p = 1-\frac{e^z}{1+e^z} = \frac{1}{1+e^z}$$

$$\frac{p}{1-p} = \frac{1+e^z}{1+e^{-z}} = e^z$$

The last expression is simply the **odds**. Taking the natural log of this we will get the logit transformation:

$$\ln\left(\frac{p_i}{1-p_i}\right) = z_i = \beta_1 + \beta_2 x_i$$

That is, the log of the odds (the *logit*) is a linear function of *x* with the following features:

1. As $p \to 1, \ln\left(\frac{p_i}{1-p_i}\right) \to +\infty$. As $p \to 0, \ln\left(\frac{p_i}{1-p_i}\right) \to -\infty$. So, the logit has no upper or lower bounds.
2. If you can model the logit, then simple algebra will allow you to model the odds or the probability.
3. The logit transformation ensures that the model generates estimated probabilities between 0 and 1.

Note: The logit transformation of the probabilities results in a linear relationship with the predictor variables. To verify this assumption, it would be useful to plot the logits by the predictor variable.

Multiple Binary Logistic Regression

The results mentioned for a simple logistic model can easily be extended to represent the multiple case:

- $p=\dfrac{e^{\beta_1+\beta_2x_2+\ldots+\beta_kx_k}}{1+e^{\beta_1+\beta_2x_2+\ldots+\beta_kx_k}}=\dfrac{e^{\boldsymbol{X\beta}}}{1+e^{\boldsymbol{X\beta}}}$

- $\dfrac{p}{1-p}=e^{\beta_1+\beta_2x_2+\ldots+\beta_kx_k}=e^{\boldsymbol{X\beta}}$

- $\ln\left(\dfrac{p}{1-p}\right)=\beta_1+\beta_2x_2+\ldots+\beta_kx_k=\boldsymbol{X\beta}$

These are algebraically equivalent, however, the last representation is commonly used.

Estimation Methods

Logistic models are usually estimated by the method of *Maximum Likelihood* (ML) in which the likelihood function is maximised. For a sample of size *n*, the likelihood for a binary logistic regression is given by:

$$L(\beta;\boldsymbol{y},\boldsymbol{X})=\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{1-y_i}$$

However, the maximisation is applied to the log of the likelihood function (lnL) which usually has a simpler form and the same solutions.

Maximizing the likelihood (or log likelihood) has no closed-form solution, so an iterative technique is used to find an estimate of the regression coefficients, ***b***. Once this estimate has been obtained, we may proceed to define some various goodness-of-fit measures and calculated residuals.

Odds Ratio

The odds ratio determines the relationship between a predictor and response . The odds ratio can be any non-negative number. An odds ratio of 1 serves as the baseline for comparison and indicates there is no association between the response and predictor. If the odds ratio is greater than 1, then the odds of success are higher for the reference level of the factor (or for higher levels of a continuous predictor). If the odds ratio is less than 1, then the odds of success are less for the reference level of the factor (or for lower levels of a continuous predictor). Values farther from 1 represent stronger degrees of association. For binary logistic regression, the odds of success are:

$$\frac{p}{1-p} = e^{\boldsymbol{x\beta}}$$

This exponential relationship provides an interpretation for $\boldsymbol{\beta}$. The odds increase multiplicatively by e^{β_j} for every one-unit increase in x_j. More formally, the odds ratio between two sets of predictors (say x_j and x_l) is given by:

$$\frac{\left(p_j / (1-p_j)\right)\left|x_j\right.}{\left(p_l / (1-p_l)\right)\left|x_l\right.}$$

Example 8.1

Suppose that seven out of 10 males are admitted to an engineering school while two out of 10 females are admitted. The probabilities for admitting a male are, **p = .7 ; q = 1 – .7 = .3**.

Here are the same probabilities for females, **p = .2 ; q = .8**.

Now we can use the probabilities to compute the admission odds for both males and females,

odds(male) = .7/.3 = 2.33 **odds(female) = .2/.8 = .25**

The odds ratio for admission, **OR = 2.33/.25 = 9.33**

Thus, the odds of a male being admitted are **9.33** times greater than for a female.

Example 8.2

A car dealer has a mailing list of customers and wants to identify those who are most likely to purchase a new car in the next 12 months. Past records will be used to find what this depends on.

$$y = \begin{cases} 1, & \text{Person will buy a new car in next 12 months} \\ 0, & \text{Person will not buy a new car in next 12 months} \end{cases}$$

p = Probability that an individual will buy a car in the next 12 months

Two independent variables are both categorical and will be represented in the analysis by dummy variables:

$x_2 = \text{Sex}$

$$D_2 = \begin{cases} 1, & \text{Man} \\ 0, & \text{Woman} \end{cases}$$

$x_3 =$ Price range of last car (Low, Medium, High)

$$D_3 = \begin{cases} 1, & \text{Previous car price low} \\ 0, & \text{Otherwise} \end{cases}$$

$$D_4 = \begin{cases} 1, & \text{Previous car price medium} \\ 0, & \text{Otherwise} \end{cases}$$

Logistic regression model for individual i is:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{4i} + \varepsilon_i$$

Or equivalently:

$$\text{Odds}_i = \frac{p_i}{1-p_i} = e^{\beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{4i} + \varepsilon_i}$$

What odds does the model predict for individuals in different categories?

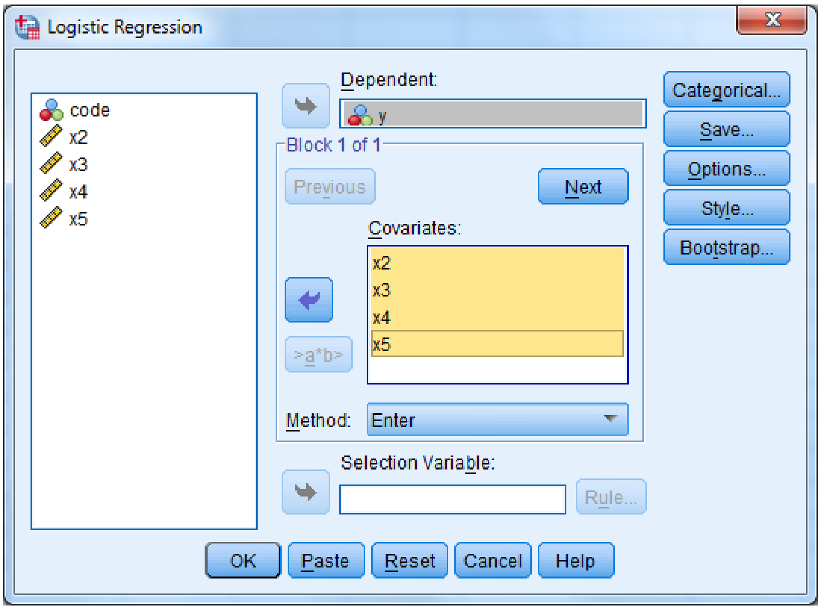
	Last price: Low $D_3 = 1, D_4 = 0$	**Last price: Medium** $D_3 = 0, D_4 = 1$	**Last price: High** $D_3 = 0, D_4 = 0$
Sex: Men $D_2 = 1$	$e^{\beta_1+\beta_2+\beta_3} =$ $e^{\beta_1} \times e^{\beta_2} \times e^{\beta_3}$		
Sex: Women $D_2 = 0$			

Note that estimates of each e^{β_j} is given by SPSS output.

Example 8.3

Suppose you have a model which contains 1 dichotomous categorical outcome variable, y, and 4 regressors, $x_2, \ldots, x_5$.

We Begin by clicking on **Analyze → Regression → Binary Logistic**.


Logistic Regression

code
x2
x3
x4
x5

Dependent: y

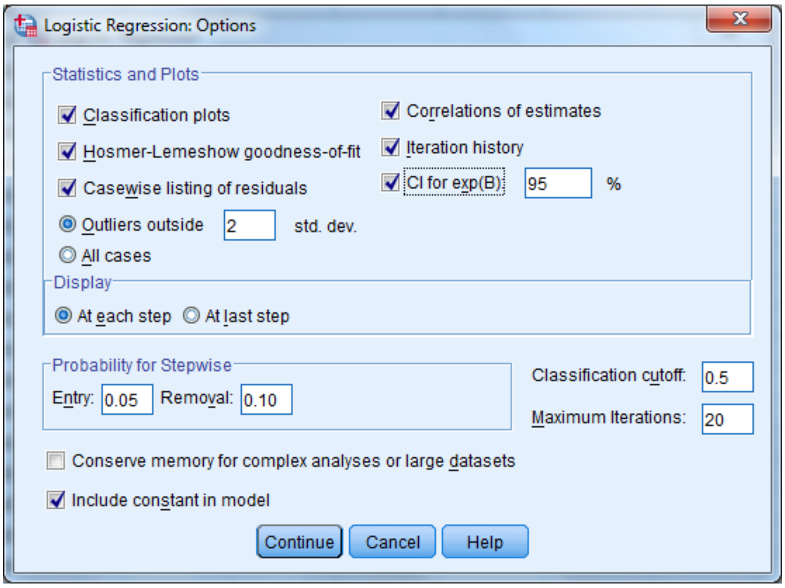
Block 1 of 1

Covariates:
x2
x3
x4
x5

Method: Enter

Selection Variable:

If one or more of the predictors was categorical, we would need to click on the **Categorical** button to specify them as such. Click on the **Options** button and select **Classification plots**, **Hosmer-Lemeshow goodness-of-fit**, **Casewise listing of residuals**, **Correlations of estimates**, **Iteration history** and **CI for exp(B)**. Then, click the **Continue** button, then click the **OK**.


Logistic Regression: Options

Statistics and Plots

☑ Classification plots
☑ Hosmer-Lemeshow goodness-of-fit
☑ Casewise listing of residuals
● Outliers outside 2 std. dev.
○ All cases
☑ Correlations of estimates
☑ Iteration history
☑ CI for exp(B): 95 %

Display

● At each step ○ At last step

Probability for Stepwise

Entry: 0.05 Removal: 0.10

Classification cutoff: 0.5

Maximum Iterations: 20

☐ Conserve memory for complex analyses or large datasets

☑ Include constant in model

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	400	100.0
	Missing Cases	0	.0
	Total	400	100.0
Unselected Cases		0	.0
Total		400	100.0

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

The Case Processing Summary table provides an overview of missing data; here there was no missing data.

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

The Dependent Variable Encoding table shows how the outcome variable was coded, if it was coded. Here, the outcome variable was not coded. By default, the binary logistic regression predicts the odds of membership in the outcome category with the highest value; here predicting membership in the 1 value, as opposed to membership in the 0 value.

Block 0: Beginning Block

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	554.518	.000

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 554.518

c. Estimation terminated at iteration number 1 because parameter estimates changed by less than .001.

The Beginning Block evaluates the model with only the constant in the equation (the null model). The constant is analogous to the *y*-intercept in LS regression. The iteration history was specified in the options, is displayed throughout the output file. The first Iteration History table shows that estimation was terminated at iteration 1 because the parameter estimates did not change by more than 0.001. The (-2 Log likelihood) is a likelihood ratio and represents the unexplained variance in the outcome variable. Therefore, the smaller the value, the better the fit!

Classification Table[a,b]

Observed			Predicted		
			y		Percentage Correct
			0	1	
Step 0	y	0	0	200	.0
		1	0	200	100.0
	Overall Percentage				50.0

a. Constant is included in the model.

b. The cut value is .500

The Classification Table shows how well the null model correctly classifies cases. The rows represent the number of cases in each category in the actual data and the columns represent the number of cases in each category as classified by the null model. The key piece of information is the overall percentage in the lower right corner which shows our null model is only 50% accurate which is equal to the accuracy of random guessing.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.000	.100	.000	1	1.000	1.000

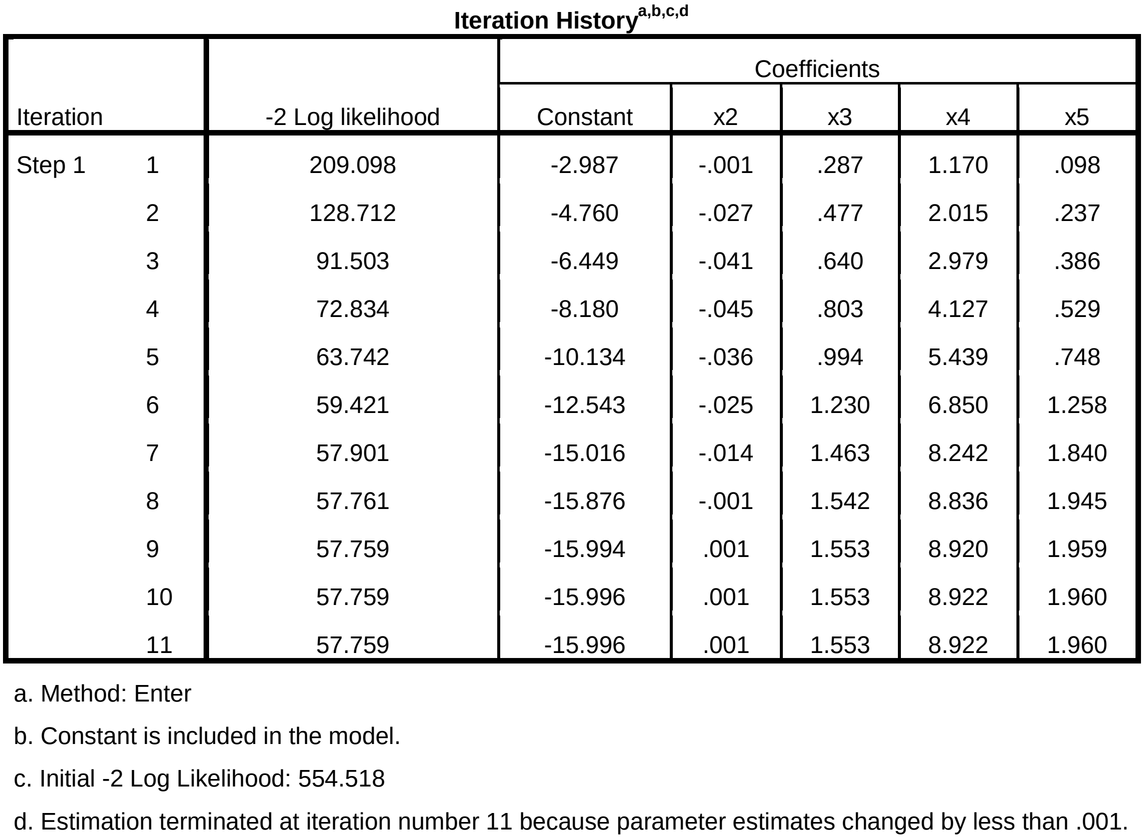
Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	x2	.074	1	.786
		x3	75.217	1	.000
		x4	283.077	1	.000
		x5	19.966	1	.000
	Overall Statistics		297.443	4	.000

The Variables in the Equation table shows the logistic coefficient (B) associated with the intercept as it is included in the model. The Wald statistic is a chi-square 'type' of statistic and is used to test the significance of the variable in the model. The Exp(B) refers to the change in odds ratio attributed to the variable. The Variables not in the Equation table simply lists the Wald test score, *df*, and *p*-value for each of the variables *not* included in the beginning block model. Notice the Overall Statistics is not a total, but rather an estimate of overall Wald statistic associated with the model had all the variables been included.

The number of blocks will increase with and correspond to the number of *blocks of covariates or predictors* entered into the model. Meaning, when specifying the variable for inclusion into the model, you notice above in the second figure (Logistic Regression dialog box) we could have clicked the Next button and entered more variables as a distinct block (as is done in sequential or hierarchical regression). Here, we only have one set of predictors so there is only the intercept model block (Block 0) and the complete model (Block 1).

Block 1: Method = Enter



Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients				
			Constant	x2	x3	x4	x5
Step 1	1	209.098	-2.987	-.001	.287	1.170	.098
	2	128.712	-4.760	-.027	.477	2.015	.237
	3	91.503	-6.449	-.041	.640	2.979	.386
	4	72.834	-8.180	-.045	.803	4.127	.529
	5	63.742	-10.134	-.036	.994	5.439	.748
	6	59.421	-12.543	-.025	1.230	6.850	1.258
	7	57.901	-15.016	-.014	1.463	8.242	1.840
	8	57.761	-15.876	-.001	1.542	8.836	1.945
	9	57.759	-15.994	.001	1.553	8.920	1.959
	10	57.759	-15.996	.001	1.553	8.922	1.960
	11	57.759	-15.996	.001	1.553	8.922	1.960

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 554.518

d. Estimation terminated at iteration number 11 because parameter estimates changed by less than .001.

The Iteration History table shows that estimation was terminated at iteration 11 because the parameter estimates did not change by more than 0.001. The (-2 Log likelihood) is a likelihood ratio and represents the unexplained variance in the outcome variable. Therefore, the smaller the value, the better the fit. Notice here the (-2 Log likelihood) is (57.759) which is substantially lower than that given above for the null model (554.518).

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	496.759	4	.000
	Block	496.759	4	.000
	Model	496.759	4	.000

The Omnibus Tests of Model Coefficients table reports the chi-square associated with each step in a stepwise model. Here, there is only one step from the constant model to the block containing predictors so all three values are the same. The *p*-value indicates the complete model (Block 1) is significantly different from the null model meaning there is a significant effect for the combined predictors on the outcome variable.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	57.759[a]	.711	.948

a. Estimation terminated at iteration number 11 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	14.559	8	.068

The Model Summary table displays the -2ln*L* as was shown and discussed directly above. The two R^2 estimates are not truly R^2 estimates; they are pseudo-R^2, meaning they are analogous to R^2 in standard multiple regression, but do not carry the same interpretation. They are not representative of the amount of variance in the outcome variable accounted for by all the predictor variables. The Nagelkerke estimate is calculated in such a way as to be constrained between 0 and 1. So, it can be evaluated as indicating model fit; with a better model displaying a value closer to 1. The larger Cox & Snell estimate is the better the model; but it can be greater than 1. These metrics should be interpreted with caution. The Hosmer and Lemeshow Test table is the preferred test of goodness-of-fit. As with most chi-square based tests however, it is prone to inflation as sample size increases. Here, we see model fit is acceptable χ^2 (9) = 14.559, *p* = .068, which indicates our model predicts values not significantly different from what we observed. To be clear, you want the *p*-value to be *greater than* your established cut-off (generally 0.05) to indicate good fit.

The Contingency Table for Hosmer and Lemeshow Test simply shows the observed and expected values for each category of the outcome variable as used to calculate the Hosmer and Lemeshow chi-square.

Contingency Table for Hosmer and Lemeshow Test

		y = 0		y = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	39	39.919	1	.081	40
	2	39	39.592	1	.408	40
	3	40	38.684	0	1.316	40
	4	40	38.651	0	1.349	40
	5	38	37.733	2	2.267	40
	6	4	5.421	36	34.579	40
	7	0	.000	40	40.000	40
	8	0	.000	40	40.000	40
	9	0	.000	40	40.000	40
	10	0	.000	40	40.000	40

Classification Table[a]

Observed			Predicted		
			y		Percentage Correct
			0	1	
Step 1	y	0	200	0	100.0
		1	7	193	96.5
	Overall Percentage				98.3

a. The cut value is .500

The Classification Table shows how well our full model correctly classifies cases. A perfect model would show only values in the diagonal and zeros off –diagonal. Adding across the rows represents the number of cases in each category in the actual data and adding down the columns represents the number of cases in each category as classified by the full model. The key piece of information is the overall percentage in the lower right corner which shows the full model is 98.3% accurate which is excellent. One way of assessing the model's fit is to compare the overall percentage in the full model's table to the overall percentage in the null model table.

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	x2	.001	.406	.000	1	.998	1.001	.452	2.216
	x3	1.553	.340	20.902	1	.000	4.727	2.429	9.201
	x4	8.922	1.922	21.548	1	.000	7494.443	173.272	324152.616
	x5	1.960	.745	6.923	1	.009	7.097	1.649	30.553
	Constant	-15.996	3.347	22.841	1	.000	.000		

a. Variable(s) entered on step 1: x2, x3, x4, x5.

The table above shows the logistic coefficient (B) for each predictor variable. The logistic coefficient is the expected amount of change in the logit for each one unit change in the predictor. It is the odds of membership in the category of the outcome variable with the numerically higher value. The closer a logistic coefficient is to zero, the less influence it has in predicting the logit. The table also displays the standard error, Wald statistic, *df*, (*p*-value), as well as the Exp(B) and confidence interval for the Exp(B). The Wald test (and associated *p*-value) is used to evaluate whether or not the logistic coefficient is different than zero. The Exp(B) is the odds ratio associated with each predictor. We expect predictors which increase the logit to display Exp(B) greater than 1.0, those predictors which do not have an effect on the logit will display an Exp(B) of 1.0 and predictors which decease the logit will have Exp(B) values less than 1.0. Note that the Exp(B) is wildly large for the x_4 predictor. This is due to a combination of the strong relationship between that variable and the outcome variable and the fact that x_4 is nearly categorical itself. Generally, when using continuous variables as predictors, you will not see such large Exp(B).

Correlation Matrix

		Constant	x2	x3	x4	x5
Step 1	Constant	1.000	-.425	-.822	-.863	-.771
	x2	-.425	1.000	.069	.048	.128
	x3	-.822	.069	1.000	.783	.628
	x4	-.863	.048	.783	1.000	.591
	x5	-.771	.128	.628	.591	1.000

The Correlation Matrix table simply shows the correlations between each of the regressors.

```
             Step number: 1

             Observed Groups and Predicted Probabilities

     200 +                                                                                                    +
         I                                                                                                   1I
         I                                                                                                   1I
F        I                                                                                                   1I
R    150 +                                                                                                   1+
E        I                                                                                                   1I
Q        I                                                                                                   1I
U        I                                                                                                   1I
E    100 +   0                                                                                               1+
N        I   0                                                                                               1I
C        I   0                                                                                               1I
Y        I0  0                                                                                               1I
      50 +0  0                                                                                               1+
         I0  0                                                                                               1I
         I0  0                                                                                               1I
         I00 0                                                                                               1I
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0        .1        .2        .3        .4        .5        .6        .7        .8        .9        1
  Group:  0000000000000000000000000000000000000000000000000011111111111111111111111111111111111111111111111111

          Predicted Probability is of Membership for 1
          The Cut Value is .50
          Symbols: 0 - 0
                   1 - 1
          Each Symbol Represents 12.5 Cases.
```

The graph above shows how the full model predicts membership. It is unusually clear in the middle because the model was so accurate. When a model is less accurate, more symbols (here 1 and 0) would appear in the middle, displaying their probability (*x*-axis). The better the model, the clearer the middle of the graph!

Casewise List[b]

		Observed			Temporary Variable	
Case	Selected Status[a]	y	Predicted	Predicted Group	Resid	ZResid
202	S	1**	.011	0	.989	9.307
256	S	1**	.000	0	1.000	224.790
317	S	1**	.436	0	.564	1.136
367	S	1**	.081	0	.919	3.374

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.000 are listed.

The Casewise List table displays cases which were incorrectly classified by the model. Here, we only have four cases misclassified.