Kenan-Flagler Business School
The University of North Carolina

Applied Logistic Regression

Overview

Logistic regression is similar to ordinary multiple regression – except that logistic regression is used when the dependent variable is binary and assumes only two discrete values. Examples include 'yes-no' dependent variables such as whether a customer responded to a marketing campaign or not, whether a person is a homeowner or not, whether a business goes bankrupt or not, or whether a person votes guilty or not guilty. Like ordinary multiple regression, the predictor variables can be metric variables (e.g., age, income, or sales units) or categorical (e.g., gender, religion, or region). Indicator or 'dummy' variables are used to include categorical variables as predictors.

While the basic concepts are similar for multiple linear regression and logistic regression, the interpretation of the regression equation and the coefficients are somewhat different. In multiple regression, the dependent variable is a continuous or metric variable – sales or profits, for example – and can assume many values. The multiple regression coefficients are multiplied by the values of the predictor variables to yield the predicted value for the dependent variable.

In logistic regression, the observed values for the dependent variable take on only two values and are usually represented using a 0-1 dummy variable. The mean of a 0-1 dummy variable is equal to the proportion of observations with a value of 1 – and can be interpreted as a probability. The predicted values in a logistic regression will always range between 0 and 1 and are also interpreted as probabilities.

Suppose we are modeling home ownership (where 1 indicates a homeowner and 0 a non-owner) as a function of income. Each individual in the dataset is either a homeowner or not so the observed values for the dependent variable will be 0 or 1. The predicted value based on the model is interpreted as the probability that the individual is a homeowner. For example, for

Professor Charlotte Mason prepared this note to provide material for class discussion rather than to illustrate either effective or ineffective handling of a business situation.

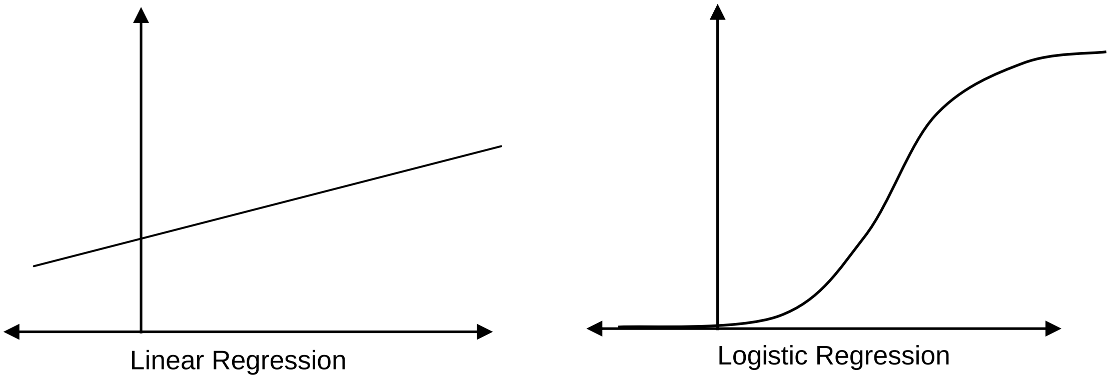
(revised August 3, 2007)

a person with an income of $35,000, the predicted probability may be .22 compared with a predicted probability of .95 for a person with a $250,000 income.

Like linear regression, once a logistic regression model has been estimated it can be used to make predictions for new observations. Assume a bank has data from a past marketing campaign promoting a ‘gold’ credit card – including whether the customer signed up for the offer or not (the dependent variable) as well as information on other bank services the customer used plus financial and demographic customer information (the predictor variables). These data can be used to estimate a logistic regression model. Then the bank could use this model to identify which additional customers to target with this or a similar offer. By inputting values for the predictor variables for each new customer – the logistic model will yield a predicted probability. Customers with high predicted probabilities may be chosen to receive the offer since they seem more likely to respond positively.

A difference between linear and logistic regression is the shape of the model as shown in Exhibit 1. The simple linear regression is represented by a straight line. For the linear regression, an increase of one unit in the predictor variable has a constant effect – equal to the slope of the line. In logistic regression, the relationship between the dependent variable and the predictor variables is assumed to be nonlinear. A logistic regression model with a single predictor is represented by an s-shaped curve.   Moreover, the curve never falls below 0 or exceeds 1 – regardless of the values of the predictor variables. Thus, the predicted values can always be interpreted as probabilities.

Exhibit 1 Simple Linear Regression versus Logistic Regression



Linear Regression

Logistic Regression

In logistic regression, the effect on the predicted probability of a one-unit increase in the predictor variable varies. At the extremes, a one-unit change has very little effect, but has a larger effect in the middle. In many situations, this is intuitive. For example, consider the effect that a $20,000 increase in income might have on the probability of home ownership. The difference in the likelihood that an individual owns a home may not change much as their income increases from $10,000 to $30,000 or from $1,000,000 to $1,020,000 – but may increase quite a bit if income increases from $50,000 to $70,000. Unfortunately, this non-linearity complicates the interpretation of the regression coefficients. In a linear regression, the interpretation of the coefficient for X is straightforward: an increase of 1 unit in X results in a change in the expected value of Y equal to B (the coefficient for X). However, in a logistic regression, we cannot say that a 1 unit increase in X will result in an expected change in Y equal to B. Rather, it depends on where on the curve the value of X is located.

The Simple Logistic Regression Curve

Consider the simple case with a single predictor variable. For now, we assume that the predictor variable is a continuous variable. In a simple linear regression, the model would be:

$$Y = B_0 + B_1X$$

where B_0 is the intercept or constant term (equal to the predicted value of Y when X=0), and B_1 is the slope of the regression line.

In a simple logistic regression, the model is:

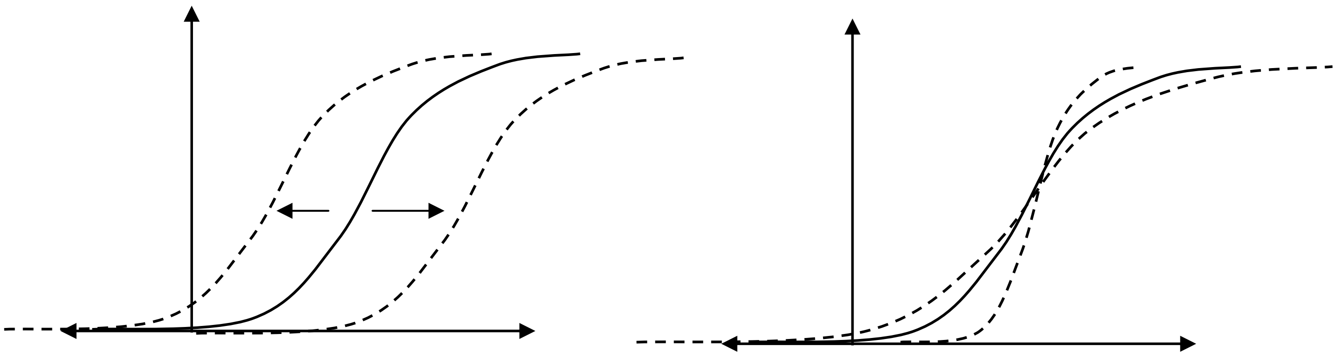
$$\mathrm{Prob}(Y=1) = \frac{e^{B_0+B_1X}}{1+e^{B_0+B_1X}}$$

which can be also be written as:

$$\mathrm{Prob}(Y=1) = \frac{1}{1+e^{-(B_0+B_1X)}}\,.$$

Thus, as in linear regression, there are two coefficients, B_0 and B_1, in a simple logistic regression. These coefficients determine the specific shape of the curve. The B_0 coefficient (also referred to as the constant) determines the location of the logistic curve along the X axis. As the constant increases, the logistic curve shifts left on the X axis. The B_1 coefficient determines the steepness and direction of the curve. A positive B_1 means the curve will increase as X increases. If B_1 is negative, the curve decreases as X increases. Larger values for B_1 indicate a steeper curve.

Exhibit 2 Logistic Regression Coefficients Control the Shape of the Curve



B_0 coefficient ‘shifts’ the curve

B_1 coefficient controls the steepness

Interpreting Logistic Regression Coefficients

In simple linear regression, interpretation of the coefficients is straightforward. The constant term estimates the value of Y when X=0. The B_1 coefficient estimates the change in Y for a one unit increase in X. Because of the nonlinear nature of the logistic regression model, interpretation of the coefficients is more complex. To interpret logistic regression, we start with a discussion of probabilities, odds and odds ratios.

A *probability* is the likelihood of an event and is bounded between 0 and 1. If the weather forecast says the probability of rain is 0.25, then there is a 25% chance of rain. *Odds* are the ratios of two probabilities: the probability that the event will occur divided by the probability that the event will not occur. If the probability of rain is 0.25, then the odds are:

$$\text{Odds} = \frac{\text{Prob (event)}}{\text{Prob (no event)}} = \frac{0.25}{0.75} = \frac{1}{3} = .333$$

Since odds are the ratio of two probabilities, odds are always positive, but may be greater than one. In fact, odds can range from 0 to infinity. When the odds are less than 1, the probability of the event (say, rain) is lower than the probability of no event (no rain). Conversely, odds greater than 1 indicate the probability of the event is greater than the probability of no event. Odds of 1 indicate equal (that is, .50) probabilities of event and no event – meaning that both outcomes are equally likely.

Finally, an *odds ratio* is the ratio of two odds. In logistic regression, the odds ratio for the predictor variable X indicates the expected change in the odds that Prob(Y=1) for a one unit increase in X. The odds ratio is particularly important in logistic regression because, unlike linear regression, the 'slope' of the curve is not constant. However, the odds ratio for a predictor variable is constant. The odds ratio for the predictor variable is computed by raising e to the power B_1, or e^{B_1}. For example, consider a logistic regression to predict the probability of purchase of a newly released movie DVD (the dependent variable) using the value (in dollars) of an 'instant coupon'. If the B_1 coefficient is 0.7, we know – since the coefficient is positive - that increasing the value of the instant coupon increases the probability of purchase. However, because of the s-shaped curve, the magnitude of the increase will depend on whether the increase is, say, from $1 to $2 or from $4 to $5. Since $e^{0.7} = 2.01$, we can say that for every dollar increase in the instant coupon value, the odds of purchase increase by a factor of 2.01. That is, the odds of purchase are twice a large for a $5 coupon compared with a $4 coupon.

Simple Logistic Regression Example Using SPSS

In December 1998 the U.S. Senate voted on two articles of impeachment against President Clinton. The vote of each senator is a binary variable – taking on only two values, guilty or not guilty. Exhibit 3 shows the number of guilty and not guilty votes on the first article of impeachment.

Exhibit 3 Number of Guilty and Not Guilty Votes on Impeachment Article 1

Vote on Article I

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	not guilty	55	55.0	55.0	55.0
	guilty	45	45.0	45.0	100.0
	Total	100	100.0	100.0	

In addition to the each senator's vote (guilty or not guilty) on each of the two articles of impeachment, we have data on variables that might be predictive of how a senator voted. These include:

- Political Party (Republican or Democrat).
- Percent of the vote Clinton received in the 1996 presidential election in the senator's state.
- Degree of ideological conservatism (0 – 100 where 100 is most conservative). These ratings are issued by the American Conservative Union and are based on a senator's voting records.

The simple cross-tabs in Exhibit 4 shows a strong association between political party and the first vote. All of the democratic senators voted not guilty whereas the great majority of republican senators voted guilty.

Exhibit 4 Cross-tabs of Political Party and Vote on Article 1

Count

		Vote on Article I		Total
		not guilty	guilty	
political party	democrat	45		45
	republican	10	45	55
Total		55	45	100

Exhibit 5 reports the correlations between the vote on the first article, degree of ideological conservatism and percent of the vote Clinton received in the senator's state. Degree of conservatism has a high positive correlation (.866) with vote – guilty votes are associated with more conservative senators. In contrast, percent of state vote shows a moderate negative correlation (-.429) with the vote – as a larger percent of the 1996 state vote was for Clinton, the more not guilty votes.

Exhibit 5 Correlations between vote on article 1, conservatism and % of state vote for Clinton

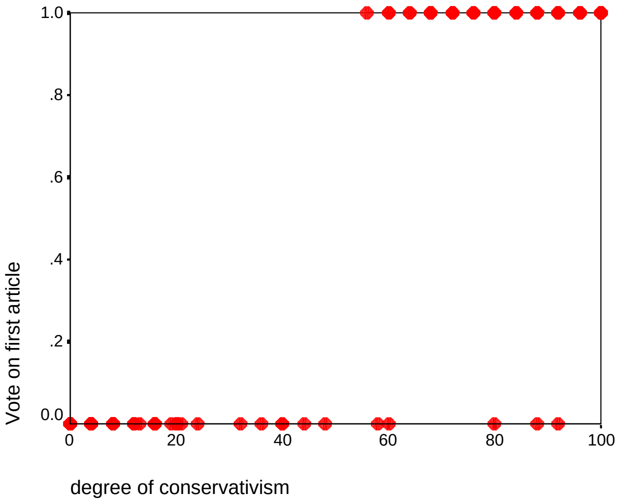
		Vote on Article I	degree of conservativism	% of vote for Clinton in state
Vote on Article I	Pearson Correlation	1.000	.866**	-.429**
	Sig. (2-tailed)	.	.000	.000
	N	100	100	100
degree of conservativism	Pearson Correlation	.866**	1.000	-.447**
	Sig. (2-tailed)	.000	.	.000
	N	100	100	100
% of vote for Clinton in state	Pearson Correlation	-.429**	-.447**	1.000
	Sig. (2-tailed)	.000	.000	.
	N	100	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

Because the variable we want to predict is binary (vote of guilty or not guilty), logistic regression is appropriate. While the other three variables (political party, degree of conservatism and % of state vote for Clinton in 1996 election) are potential predictors of the vote on the first impeachment article, high correlations or collinearity between these predictors is likely – particularly between political party and degree of conservatism. The correlation between these two variables is .906, suggesting that it would be unwise to include both as predictors. Since the ideological conservatism variable captures a broader range than simple party affiliation, let us start by considering that as a single predictor.

A simple scatter plot of vote (0=not guilty, 1=guilty) versus degree of conservatism is shown in Exhibit 6. Because of the 0-1 nature of the dependent variable, this plot is not as insightful as one might hope – although it does show that senators with lower ratings on conservatism tended to vote not guilty, whereas nearly all senators with high conservatism ratings voted guilty.

Exhibit 6 Scatterplot of senate vote (guilty =1) by degree of conservatism



A logistic regression using the vote on Article 1 as the dependent variable and degree of ideological conservatism as the predictor will allow us to quantify the relationship between these variables and to assign a specific probability of a guilty vote to each senator. Exhibit 7 lists the steps to run a logistic regression in SPSS (see the Appendix for additional details and options for running logistic regressions in SPSS).

Exhibit 7 How to specify a logistic regression in SPSS

- Select *Analyze/Regression/Binary logistic*
- Select your dependent variable
- Select your independent variable(s) (called *covariates* by SPSS)
- (optional)To save the predicted values (this will create a new column of data with the predicted probabilities):
    - Click on *Save…*
    - Click on *Probabilities* under *Predicted Values*
    - Click on *Continue*
- Click on *OK*.

Interpreting the SPSS Output

Exhibit 8 contains the full SPSS output from the logistic regression. Key parts of the output have been labeled and are described below.

A: This indicates how many cases are included in the analysis and how many, if any, have missing variables. Here, we see that all 100 cases have been included.

B: This indicates how the dependent variable is coded:
1 indicates guilty and 0 not guilty.

C: Block 0: Beginning Block:
This reports results from an ‘intercept-only’ model and we will skip directly to the Block 1 results. In essence, this model is a basis of comparison for the model specified.

D: Omnibus Tests of Model Coefficients:
These values test whether or not all of the predictor variables entered in this step, in this block, or in the model have a significant effect. Here these values are all the same – they will differ if you request a stepwise regression or enter the variables in ‘blocks’ (i.e., groups of variables). Higher chi-square values indicate more significance which is reported in the column labeled ‘Sig.’ A 0.05 cutoff for significance level is commonly used – here the 0.000 is less than 0.05 so we conclude that the model is statistically significant.

E: Model Summary:

-2 Log likelihood is used to measure how well the model fits the data. Smaller values are better – although we will not focus on this statistic. The two R-square measures are attempts to come up a measure comparable to the familiar R^2 from ordinary multiple regression. However because the observed values for the dependent variable in a logistic regression can take on only two values, 0 or 1, there isn't a measure that is totally analogous. However, as in linear regression, higher values are 'better' and these measures can be used to compare models with different sets of predictor variables.

F: Classification Table:

The classification table compares the predicted values for the dependent variable, based on the model, with the actual observed values in the data. SPSS uses the predicted probabilities for each case (or senator) to compute the predicted value. By default, the SPSS cutoff is 0.50 – although a different value can be chosen.

If the predicted probability is greater than the cutoff of 0.50, SPSS predicts a value of 1 (which would be a guilty vote in this case). If the predicted probability is less than 0.50, SPSS predicts a value of 0 (or a not guilty vote for this data). The actual and predicted values are summarized in a 2x2 table along with the percentages of correct classifications. Here we see that, overall, the model correctly predicts or classifies 94% (=94/100) of the votes. The percent of not guilty votes correctly classified is 90.9% (=50/55) and the percent of guilty votes correctly classified is 97.8% (=44/45).

G: Variables in the Equation:

The Variables in the Equation table summarizes the coefficients, standard errors, significance tests and odds ratios for each variable in the model as well as the constant term.

B: These are the coefficients in the logistic regression equation. In this case the final model equation is:

$$\begin{aligned} P(\text{Guilty}) &= 1/[1+\exp(-\Sigma\ BX)] \\ &= 1/[1+\exp(6.207 - .108\ \text{conservative})] \end{aligned}$$

S.E.: The standard error of B.

Wald: A measure of the significance of the predictor variables. Higher values in conjunction with the degrees of freedom (df) indicate significance.

Sig.: The significance of the Wald test. If the common 0.05 cutoff for significance level is used, values less than 0.05 indicate statistical significance.

Exp(B): This is the odds ratio and is useful for interpreting the effects of the predictor variables.

In summary, a senator's degree of ideological conservatism is a significant predictor of his or her vote on the first article of impeachment. The coefficient for the conservatism variable is positive – so more conservative senators were more likely to cast a guilty vote. Using this single predictor, the model correctly classified 94% of the senators, with a slightly higher prediction accuracy for guilty votes than for not guilty. Exhibit 9 lists the predicted probabilities for different ratings of ideological conservatism. Exhibit 10 shows a plot (based on the numbers in Exhibit 9) of the logistic regression model.

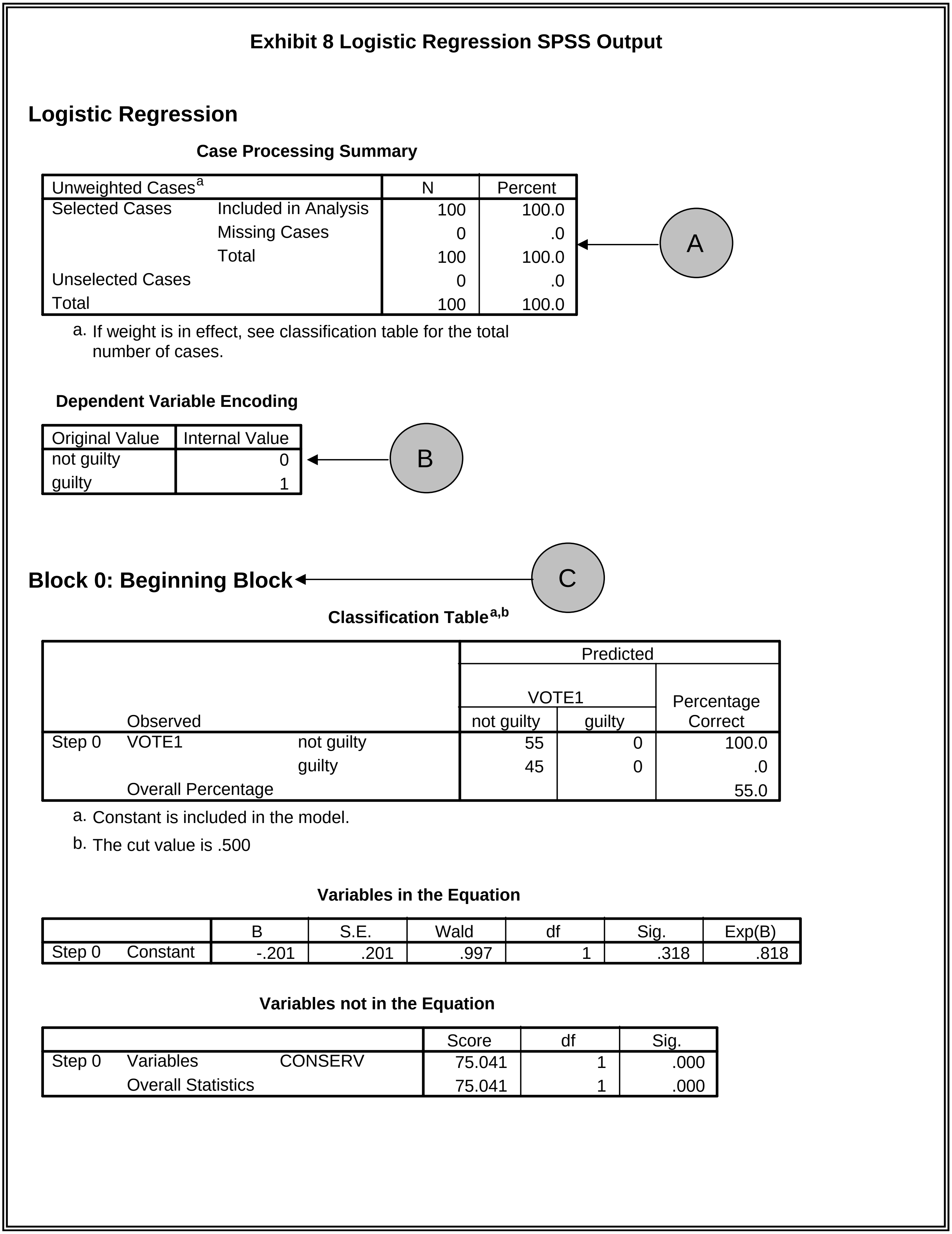


Exhibit 8 Logistic Regression SPSS Output

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	100	100.0
	Missing Cases	0	.0
	Total	100	100.0
Unselected Cases		0	.0
Total		100	100.0

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

Dependent Variable Encoding

Original Value	Internal Value
not guilty	0
guilty	1

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			VOTE1		Percentage Correct
	Observed		not guilty	guilty	
Step 0	VOTE1	not guilty	55	0	100.0
		guilty	45	0	.0
	Overall Percentage				55.0

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	-.201	.201	.997	1	.318	.818

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	CONSERV	75.041	1	.000
	Overall Statistics		75.041	1	.000

Exhibit 8 (continued)

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	100.520	1	.000
	Block	100.520	1	.000
	Model	100.520	1	.000

D

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	37.107	.634	.848

E

Classification Table[a]

			Predicted		
			VOTE1		Percentage Correct
	Observed		not guilty	guilty	
Step 1	VOTE1	not guilty	50	5	90.9
		guilty	1	44	97.8
	Overall Percentage				94.0

[a] The cut value is .500

F

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CONSERV	.108	.024	20.642	1	.000	1.114
	Constant	-6.207	1.566	15.698	1	.000	.002

[a] Variable(s) entered on step 1: CONSERV.

G

We see that a senator with a conservatism rating of 0 (e.g. Boxer of California) has a predicted probability of voting guilty of 0.002. At the other extreme, a senator with a conservatism rating of 100 (e.g. Helms of North Carolina) has a 0.99 predicted probability of voting guilty. The 'dividing line' appears to be between 57 and 58. A conservatism rating of 57 leads to a predicted probability of 0.4873 – just under the 0.50 cutoff. An increase of one to a 58 rating increases the predicted probability to 0.5142. For a simple logistic regression (i.e., one with only one predictor variable), it is easy to compute the smallest value of the predictor needed for a predicted probability of 0.50 or greater. This 50% cutoff value is equal to $-B_0/B_1$. In this example $-B_0/B_1$ = 6.207/.108 = 57.47. Thus, any senator with a conservatism rating of 58 or higher would be expected to vote guilty, those with a rating of 57 or lower are expected to vote not guilty.

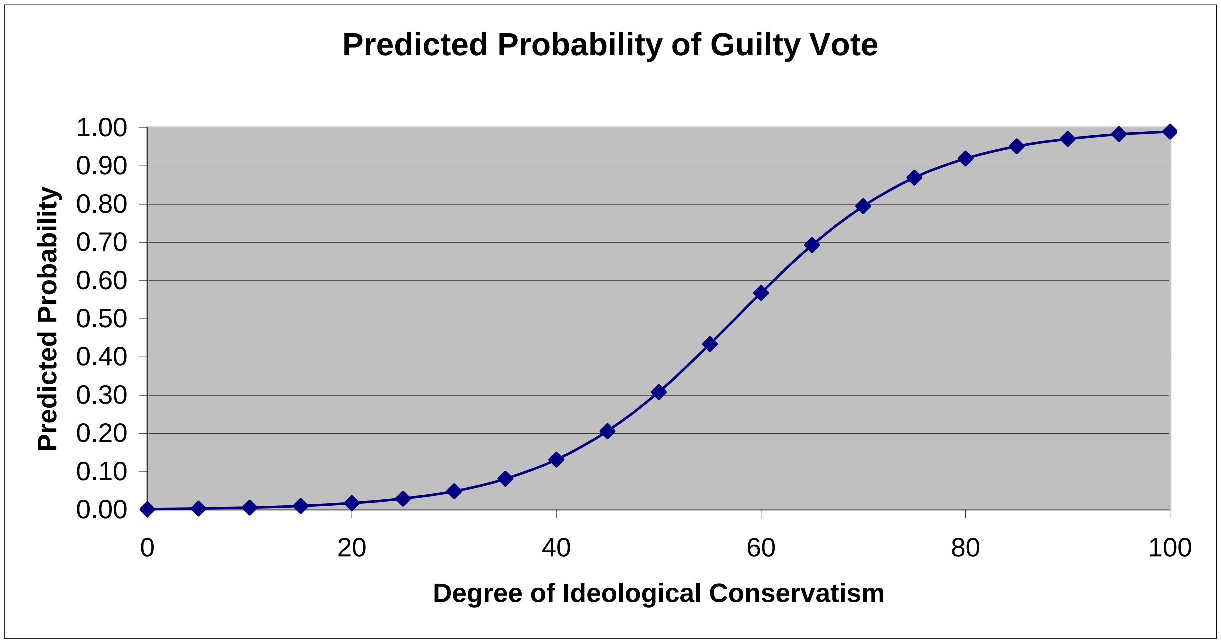
In Exhibit 9 note that although conservatism increases in equal 5-unit increments, the increases in predicted probabilities are not equal. A 5-unit increase from 0 to 5 raises the probability of a guilty vote by only 0.0014 (=0.0034 - 0.0020), whereas an increase from 50 to 55 raises the probability by 0.1251 (=0.4336 - 0.3085).

Exhibit 9 Table of Conservatism Ratings and Predicted Probabilities

Degree of conservatism	**Predicted Probability**[a]
0	0.0020
5	0.0034
10	0.0059
15	0.0101
20	0.0172
25	0.0291
30	0.0489
35	0.0811
40	0.1316
45	0.2064
50	0.3085
55	0.4336
60	0.5678
65	0.6927
70	0.7946
75	0.8691
80	0.9193
85	0.9513
90	0.9711
95	0.9829
100	0.9900

[a] Predicted Probability = $1/[1+\exp(-\Sigma BX)] = 1/[1 + \exp(6.207 - .108X)]$

Exhibit 10 Plot of Logistic Regression Model



Interpreting Logistic Regression Coefficients

In a linear regression model, we use the coefficients and standardized coefficients to describe the magnitude of the effect of that predictor. For example, if the coefficient for a predictor X is 4.5, we would say that for every unit increase in X, the predicted value of Y will increase by 4.5. Unfortunately, the interpretation is more complicated in a logistic regression.

To interpret logistic regression results, you must understand probabilities, odds and odds ratios:

- A *Probability* is the likelihood of an event and is bounded between 0 and 1.

- An *Odds* is the ratio of two probabilities – for example, the probability of voting guilty divided by the probability of voting not guilty. Odds cannot be negative (because probabilities cannot be negative), but odds have no upper bound. In this example, the odds of a senator with a conservatism rating of 100 voting guilty is .99/.01 = 99.

- An *Odds Ratio* is the ratio of two odds. Like odds, the odds ratio cannot be negative, and has no upper bound.

Exhibit 11 shows the probabilities, odds and odds ratios for different values of ideological conservatism. As we noted before, an increase of 1 in ideological conservatism does not have a constant effect on the dependent variable (predicted probability). When conservatism increases from 40 to 41, the probability of a guilty vote increases by 0.1444 - 0.1316 = 0.0128. When conservatism increases from 80 to 81, the probability of a guilty vote increases by 0.9270 - 0.9193 = 0.0077. The effect of an increase in conservatism depends on where you are on the curve in Exhibit 8.

Exhibit 11 Probabilities, Odds and Odds Ratios

Conservatism Rating	Prob(Guilty)	Prob(Not Guilty) =1-Pr(Guilty)	Odds =Pr(Guilty)/Pr(Got Guilty)	Odds Ratio
40	.1316	.8684	.1515	
41	.1444	.8556	.1668	.1668/.1515=1.114
:	:	:	:	
80	.9193	.0807	11.3916	
81	.9270	.0730	12.6986	12.6986/11.3916=1.114

However the odds ratio is constant. In this data, every increase of one in conservatism rating increases the odds of a guilty vote by a factor of 1.114, or 11.4%. So while a one-unit change in a predictor variable does not have a constant effect on the predicted probability, it does have a constant effect on the odds.

The odds ratio for a specific predictor is found using the following formula:

$$\text{Odds Ratio} = \exp(B) = e^{B}$$

For this example, B=0.108 and exp(B) = 1.114. SPSS reports the odds ratio for each predictor in the last column of the 'Variables in the Equation' output. Now consider what the odds ratio will be for positive coefficients, negative coefficients and coefficients of zero:

- Positive coefficients: For any value of $X > 0$, $\exp(X) > 1$. So a positive coefficient will have an odds ratio greater than one, indicating that an increase in that variable will multiply the odds by a factor greater than one – increasing the odds that Y=1.

- Negative coefficients: For any value of $X < 0$, $0 < \exp(X) < 1$. So a negative coefficient will have an odds ratio less than one, indicating that an increase in that variable will multiple the odds by a factor less than one – decreasing the odds that Y=1.

- Coefficients equal to zero: If the coefficient for a predictor variable is 0, then exp(0) = 1, indicating that a one unit change will the odds that Y=1 by a factor of 1. Since multiplying something by one leaves it unchanged – an odds ratio of 1 indicates no effect (corresponding to the zero coefficient).

Note that the magnitudes of positive and negative coefficients can be compared using the absolute values of the coefficients themselves or by taking the inverse of the odds ratio of the negative (or positive) coefficient. For example, an odds ratio of 2.0 has the same 'magnitude' as an odds ratio of 0.5 (1/2 = .5, or 1/.5 = 2). One is a doubling of the odds ratio, the other is a halving.

Adding Another Predictor to the Model

Exhibit 12 reports the key SPSS output from a second logistic regression model that includes the percent of state vote for Clinton in the1996 election in addition to degree of ideological conservatism. The correlations reported earlier showed that senators from states showing strong support for Clinton in the 1996 election were less likely to vote guilty. Will including it as a predictor improve the model?

Exhibit 12 SPSS Output from Second Logistic Regression Model

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	100.856	2	.000
	Block	100.856	2	.000
	Model	100.856	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	36.771	.635	.850

Classification Table[a]

			Predicted		
			VOTE1		Percentage Correct
	Observed		not guilty	guilty	
Step 1	VOTE1	not guilty	50	5	90.9
		guilty	1	44	97.8
	Overall Percentage				94.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CONSERV	.105	.024	18.892	1	.000	1.111
	STATEVOT	-.040	.070	.324	1	.569	.961
	Constant	-4.216	3.774	1.248	1	.264	.015

a. Variable(s) entered on step 1: CONSERV, STATEVOT.

Conclusions drawn from the results in Exhibit 12 are summarized below:

- There is virtually no difference in the chi-square values or overall significance of the model.

- The –2 Log likelihood is nearly the same, as are the R square measures.

- The classification results are identical to the model using only degree of ideological conservatism.

- The ‘variables in the equation’ table shows that the state vote variable is not statistically significant. The coefficient for conservatism is essentially unchanged.

In summary, this additional predictor does not improve the model and need not be included.

Summary

Logistic regression is used when the dependent variable is binary. Like linear regression, the predictor variable can be metric or categorical. In a logistic regression, the predicted values are bounded between 0 and 1 and are interpreted as the probability that the dependent variable equals one. Like linear regression, the coefficients and statistical tests will indicate whether the predictor variables are statistically significant – and whether they have a positive or negative effect on the probability that the dependent variable is one. However, unlike linear regression, the effect of a one-unit change in X on Y is not linear – rather it depends on the value of X. The odds ratios, in combination with the coefficients, are used to interpret the effects of individual predictor variables.

Appendix

Additional Options with Logistic Regression in SPSS

Categorical Predictors

Categorical predictor variables can be included in both linear and logistic regression. With linear regression it is necessary to code categorical predictors using dummy variables. For example, gender (male/female) would be recoded as a single 0-1 binary variable. A categorical variable with more than two categories such as season (fall, winter, spring, summer) must be recoded using n-1 dummy variables where n is the number of categories. Thus, season would need to be recoded into three 0-1 dummy variables.

With logistic regression in SPSS, categorical variables can be included directly without recoding – however it is necessary to tell SPSS which predictors, if any, are categorical. To do this click on the ‘categorical’ box in the main logistic regression dialog box. This will activate a new dialog box listing all the covariates you have selected. Simply highlight any categorical variables in the list and click on the right arrow to transfer them to the ‘categorical covariates’ box. By default SPSS will use standard dummy variable coding (and you can choose whether you want the first or last category as the baseline or reference category).

Obtaining Residuals

As with linear regression, you can request various residuals to be saved as new variables in the data editor. Two residuals that are unique to logistic regression are the *predicted probabilities* and the *predicted group memberships.* The predicted probabilities are the probabilities of *Y* occurring given the values for the predictors for each observation. The predicted group membership predicts which of the two categories of *Y* an observation is most likely to belong to based on the model