

Binary Logistic Regression

Krishan K. Pandey (Ph.D.)

Professor & Dean “Office of Doctoral Studies”

Senior Fellow of Jindal Institute of Behavioral Sciences (JIBS)

Jindal Global Business School, O. P. Jindal Global University,

Sonipat-Narela Road, Near Jagdishpur Village, Sonipat, Haryana-131001.

☎: (+91)-130-4091923; 📠: (+91)-130-4091888; Mobile:08396907453.

Email: kkpandey@jgu.edu.in; krishan.pandey@gmail.com;

Webpage: <http://www.krishanpandey.com/>



CONTENTS

| | |
|----------------------------------------|----|
| 1) Introduction..... | 3 |
| Categorical data and 2 x 2 tables..... | 3 |
| Odds and Relative Odds..... | 5 |
| Odds..... | 5 |
| Relative odds..... | 5 |
| 2) Logistic regression theory..... | 6 |
| Introduction..... | 6 |
| The Theory..... | 6 |
| Logistic regression theory..... | 7 |
| Dummy variables..... | 9 |
| Exercise 2..... | 10 |
| 3) Logistic regression in SPSS 13..... | 11 |
| 4) Summary and further comments..... | 42 |
| Summary..... | 42 |
| Further comments..... | 42 |
| Reading list..... | 43 |

1) Introduction

Socio-economic variables are very often categorical, rather than interval scale. In many cases research focuses on models where the dependent variable is categorical. For example, the dependent variable might be ‘unemployed’ / ‘employed’, and we could be interested in how this variable is related to sex, age, ethnic group, etc. In this case we could not carry out a multiple linear regression as many of the assumptions of this technique will not be met, as will be explained theoretically below. Instead we would carry out a **logistic regression** analysis. Hence, logistic regression may be thought of as an approach that is similar to that of multiple linear regression, but takes into account the fact that the dependent variable is categorical.

Categorical data and 2 x 2 tables

We can write categorical data in two forms: list form or table form. The important point to make about this is that whichever way we choose to think about this kind of data, the information is the same. For example, if we were interested in the association between unemployment and sex for a sample of 12 people (this is a smaller sample than we would tend to use in general but it illustrates the point), we could write the data in *list form* as:

| OBS | UNEM | FEMALE |
|-----|------|--------|
| 1 | 0 | 0 |
| 2 | 0 | 1 |
| 3 | 0 | 1 |
| 4 | 0 | 1 |
| 5 | 0 | 0 |
| 6 | 1 | 1 |
| 7 | 0 | 1 |
| 8 | 1 | 0 |
| 9 | 0 | 0 |
| 10 | 0 | 1 |
| 11 | 0 | 0 |
| 12 | 1 | 0 |

Or the same data in table form as:

| | UNEM | NOT UNEM | TOTAL |
|--------|------|----------|-------|
| MALE | 2 | 4 | 6 |
| FEMALE | 1 | 5 | 6 |
| TOTAL | 3 | 9 | 12 |

2 x 2 tables are quite a good way to present the information, because the relationship between the two variables can usually be clearly interpreted. When we are interested in the association between several variables we can, of course, still construct a multi-way table. However, it is less easy to interpret the relationships from the tables when several variables are involved.

Example 1

We will now consider an example from Plewis, I (1997), Chapter 5.

Table 1.

| | Ethnic Group | | |
|--------------------|--------------|-----------|------------|
| Behaviour Problems | White | Black | Total |
| NO | 90 [0.83] | 30 [0.48] | 120 (70%) |
| YES | 19 [0.17] | 33 [0.52] | 52 (30%) |
| Total | 109 (63%) | 63 (37%) | 172 (100%) |

Table 1 is a cross tabulation of two binary variables for a sample of 172 boys in reception classes.

- Whether or not the child is *perceived* by their teacher to have a behaviour problem (which we will later model as the response).
- Ethnic group (which we will later model as the explanatory variable).

We can see that the majority of the sample of boys (70%) are not perceived to have a behaviour problem and that 63% of them are white. The conditional probabilities of having a behaviour problem, given ethnic group are shown in square brackets after each of the cell frequencies. For example the probability of being perceived to have a behaviour problem for white boys is 0.17, and for black boys is 0.52.

Odds and Relative Odds

A useful way of using the information in cross tabulations where one dimension of the table is an outcome of interest (whether 2x2 tables or more complicated ones), is to calculate *odds* and *relative odds* (odds ratios).

Odds

In the above table, the odds of a white boy being seen to have a behaviour problem are $19/90 = 0.21$ or 0.21 to 1. In betting terms that is about 5 : 1 against – much less than even money.

For black boys, the corresponding odds are $33/30 = 1.1$, or 1.1 to 1.

Equivalent to 11 to 10 on, (or a little better than even money.). Note that odds are not the same as probabilities – they are not restricted to the range 0 to 1.

Relative odds

We can also think of the information in the table in terms of relative odds. The relative odds of a black boy compared with a white boy being seen as having a behaviour problem are $1.1 / 0.21$ or 5.2 to 1. In other words a black boy is 5.2 times more likely than a white boy to be seen as having a behaviour problem. Equally, boys perceived to have behaviour problems are 5.2 times more likely to be black rather than white, compared with boys without perceived behaviour problems. Relative odds are symmetrical in that sense; like correlation, we do not think of this measure in terms of a dependent variable and an explanatory variable. We just think in terms of the association between two variables.

Exercise 1

Suppose we are interested in the relationship between unemployment and ethnic group for a sample of 18 year olds and we have the following data

| Unemployed at 18? | Ethnic group | | TOTAL |
|-------------------|--------------|-------|-------|
| | White | Black | |
| No | 1700 | 40 | 1740 |
| Yes | 112 | 8 | 120 |
| Total | 1812 | 48 | 1860 |

Calculate the probabilities, odds and relative odds of being unemployed at 18 for white and black ethnic groups

2) Logistic regression theory

Introduction

When we want to look at a dependence structure, with a dependent variable and a set of explanatory variables (one or more), we can use the logistic regression framework.

Multiple linear regression may be used to investigate the relationship between a continuous (interval scale) dependent variable, such as income, blood pressure or examination score. However, socio-economic variables are very often categorical, rather than interval scale. In many cases research focuses on models where the dependent variable is categorical. For example, the dependent variable might be 'unemployed' or 'not' (as we saw in Exercise 1), and we could be interested in how this variable is related to sex, age, ethnic group, etc. In this case we could not carry out a multiple linear regression as many of the assumptions of this technique will not be met, as will be explained theoretically below. Instead we would carry out a logistic regression.

The Theory

If we wrote the 'perceived behaviour problems' table as data in list format, we would be interested in modelling the variation in the probability of being perceived to have behaviour problems, and for the table data we are interested in modelling the variations in the proportions with perceived behaviour problems amongst black boys compared with white boys. It is important to note that regardless of whether we consider the analysis in terms of data in a list or a table, the results will be exactly the same.

Proportions and probabilities are different from continuous variables in a number of ways. They are bounded by 0 and 1, whereas in theory continuous variables can take any value between plus or minus infinity. This means that we cannot assume normality for a proportion, and we must recognise that proportions have a binomial distribution. Unlike the normal distribution, the mean and variance of the Binomial distribution are not independent. The mean is denoted by P and the variance is denoted by $P*(1-P)/n$, where n is the number of observations, and P is the probability of the event occurring (e.g. the probability of being unemployed, or having 'perceived behavioural problems') in any one 'trial' (for any one individual in this example). If we were considering the data in 'list' rather than table form we would assume that the variable had a mean P and a variance $P*(1-P)$ and this variable would have a Bernoulli distribution.

When we have a proportion as a response, we use a **logistic** or **logit** transformation to link the dependent variable to the set of explanatory variables. The logit link has the form:

$$\text{Logit}(P) = \text{Log} [P / (1-P)]$$

The term within the square brackets is the odds of an event occurring. In the example above this would be the odds of a person being perceived to have behaviour problems.

Using the logit scale changes the scale of a proportion to plus and minus infinity, and also because $\text{Logit}(P) = 0$, when $P=0.5$. When we transform our results back from the logit (log odds) scale to the original probability scale, our predicted values will always be at least 0 and at most 1.

Logistic regression theory

Let:

$$P_i = \text{Pr}(Y = 1 | X = x_i)$$

Then we can write the model:

$$\text{Log} \left(\frac{P_i}{1 - P_i} \right) = \text{logit}(P_i) = \beta_0 + \beta_1 x_i$$

In our example P_i is the probability of being perceived as having behaviour problems, and x_i is the boy's ethnic group. Therefore the parameter β_0 gives the *log odds* of a white boy being perceived to have behaviour problems (when $x_i = 0$) and β_1 shows how these odds differ for black boys (when $x_i = 1$).

We can write the model in terms of **odds** as:

$$P_i / (1 - P_i) = \exp(\beta_0 + \beta_1 x_i)$$

Or in terms of the probability of the outcome (e.g. perceived behaviour problems) occurring as:

$$P_i = \exp(\beta_0 + \beta_1 x_i) / (1 + \exp(\beta_0 + \beta_1 x_i))$$

Conversely the probability of the outcome not occurring is

$$1 - P_i = 1 / (1 + \exp(\beta_0 + \beta_1 x_i))$$

Notice that we have so far not included a residual term in the models, and have instead expressed the model in terms of population probabilities. But we could write it as:

$$p_i = P_i + f_i = \exp(\beta_0 + \beta_1 x_i) / (1 + \exp(\beta_0 + \beta_1 x_i)) + f_i$$

Note that in this case, f_i is not normally distributed, as it was assumed to be for linear regression.

Returning to Example 1: perceived behavioural problems by ethnic group

We will now consider some logistic regression theory and return to Example 1, session 1.

| | Ethnic Group | | |
|--------------------|--------------|-----------|------------|
| Behaviour Problems | White | Black | Total |
| NO | 90 [0.83] | 30 [0.48] | 120 (70%) |
| YES | 19 [0.17] | 33 [0.52] | 52 (30%) |
| Total | 109 (63%) | 63 (37%) | 172 (100%) |

We can fit a logistic regression model to the data in Table 1. We get:

$$\text{Logit } P = -1.56 + 1.65EG$$

Which we can interpret as the log odds of a white boy ($EG=0$) seen as having a behaviour problem being equal to -1.56 , hence the odds of a white boy having a behaviour problem are: $\exp(-1.56) = 0.21$

The log odds of a black boy ($EG=1$) having a perceived behaviour problem are $-1.56 + 1.65 = 0.09$. Hence the odds of a black boy having a perceived behaviour problem are $\exp(0.09) = 1.1$. Alternatively we can say that the odds for black boys are $\exp(1.65)=5.21$ times as high as they are for white boys. That is, the relative odds of a teacher perceiving a black boy to have behavioural problems compared with a white boy are 5.21.

Notice that these results correspond exactly to the results in Table 1. This is because for Table 1 there is one degree of freedom: we can calculate the degrees of freedom in the table as $(r-1)*(c-1) = 1$, where r is the number of rows in the

table and c is the number of columns. So if we fit one parameter (ethnic group, EG) we have used up, or *saturated*, all the degrees of freedom and hence fitted a saturated model. This means we have fitted enough terms in the model to explain everything that is going on in Table 1. Before we fitted the ethnic group term, we could not have explained everything that is going on in the table, and we would hence find a **deviance** (or $-2 \log$ likelihood) of 22.8. The deviance is a measure of how much variation is left having fitted the model (how much is left unexplained by the model). The deviance follows a χ^2 distribution and we would in general compare the difference in deviance in two models to find out if the extra terms we added were significant. In the current example, we cannot really talk in terms of ‘change in deviance’ because once we have fitted EG to the model, we have fitted a saturated model, and the deviance is 0, but in general, assessing the change in deviance is a very useful way of assessing whether we need to add extra terms to our model. In the workshop examples we will see how this works in much more detail.

Dummy variables

When an explanatory variable is categorical we use dummy variables to contrast the different categories. For each variable we choose a baseline category and then contrast all remaining categories with the base line. If an explanatory variable has k categories, we need $k-1$ dummy variables to investigate all the differences in the categories with respect to the dependent variable.

For example suppose the explanatory variable was housing tenure coded like this:

Tenure

- 1: Owner occupier
- 2: renting from a private landlord
- 3: renting from the local authority

We would therefore need to choose a baseline category and create two dummy variables. For example if we chose owner occupier as the baseline category we would code the dummy variables like this

| Tenure: | D1 | D2 |
|------------------------|----|----|
| Owner occupier | 0 | 0 |
| Rented private | 1 | 0 |
| Rented local authority | 0 | 1 |

For logistic regression SPSS can create dummy variables for us from categorical explanatory variables, as we will see later.

Exercise 2

Write down a statistical model to investigate the relationships in the following table

| Unemployed | Ethnic group | | | TOTAL |
|------------|----------------|-----------|--------|-------|
| | Afro Caribbean | Pakistani | Indian | |
| No | 50 | 40 | 45 | 135 |
| Yes | 9 | 5 | 4 | 18 |
| Total | 59 | 45 | 49 | 153 |

3) Logistic regression in SPSS 13

These data are taken from the British Election Study 2005 pre-campaign and post-election panel data. More information:

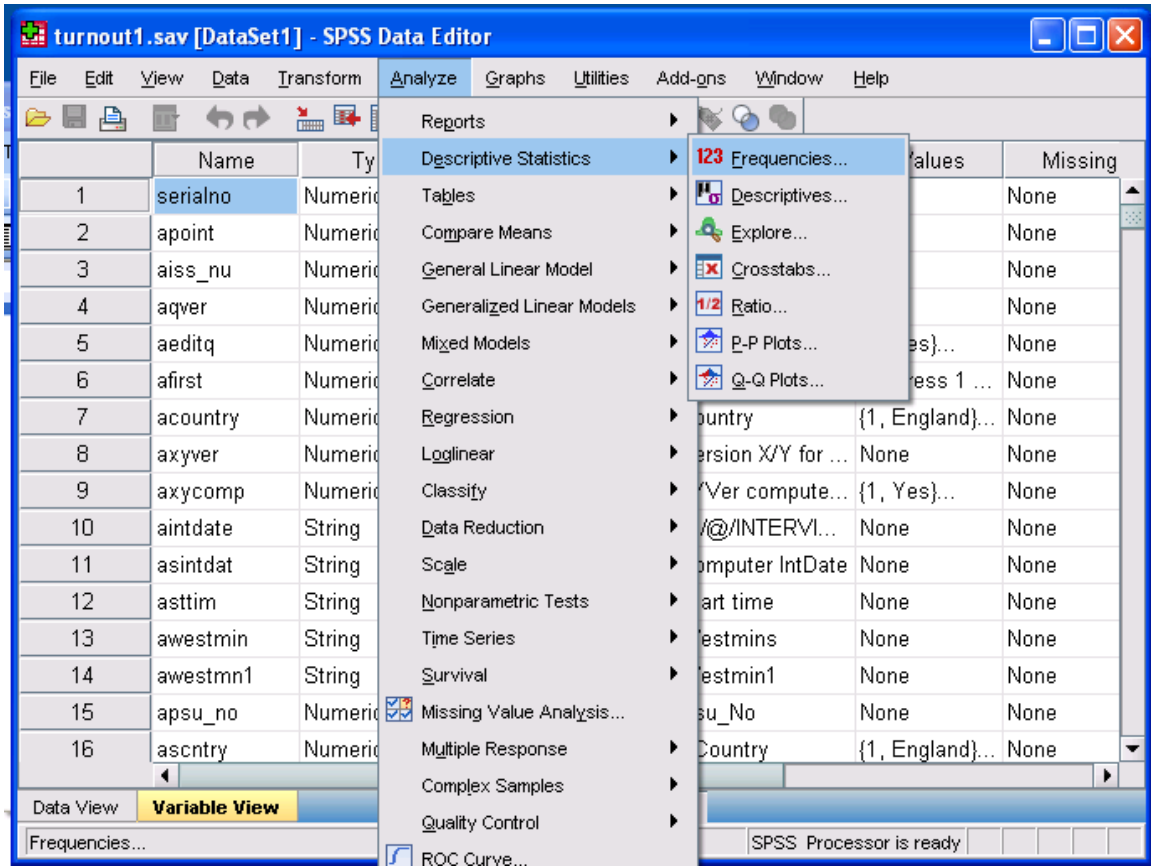
<http://www.essex.ac.uk/bes/>

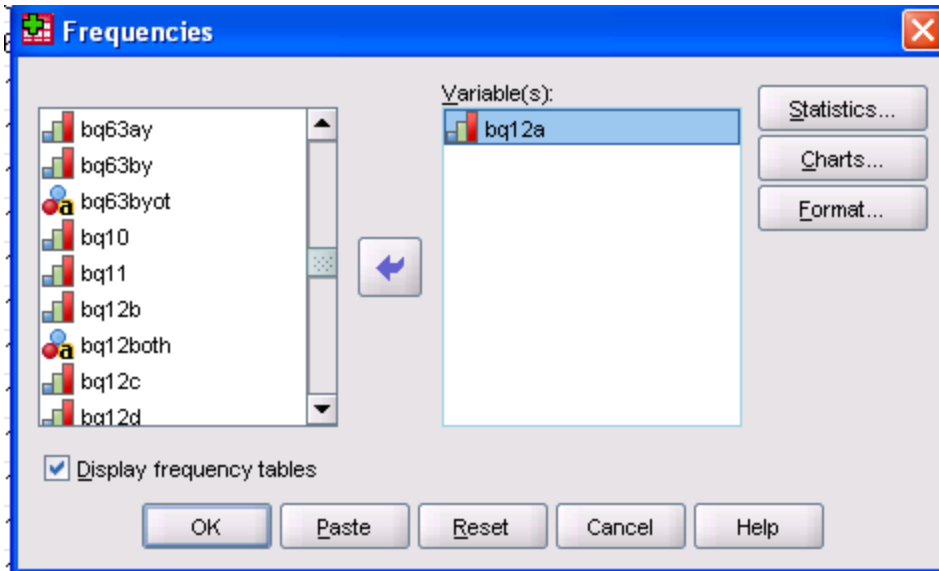
We will consider the propensity to vote (sometimes called ‘turnout’) as the dependent variable, which has 2 categories. 0=did not turn out to vote, 1 turned out to vote.

We will consider turnout in relation to three explanatory variables: gender, age and housing tenure of the respondent. Turnout is known to be lower amongst young people in western democracies, and may also be associated with tenure and gender. We will use logistic regression to investigate the extent of the association between the propensity to turn out to vote, with respect to gender, age and tenure in the 2005 election data.

But first some exploratory data analysis: we will check the distributions of each of the variables and do some filtering of the data and re-coding of the variables.

NB: the dataset which we will use here is called turnout1.sav.



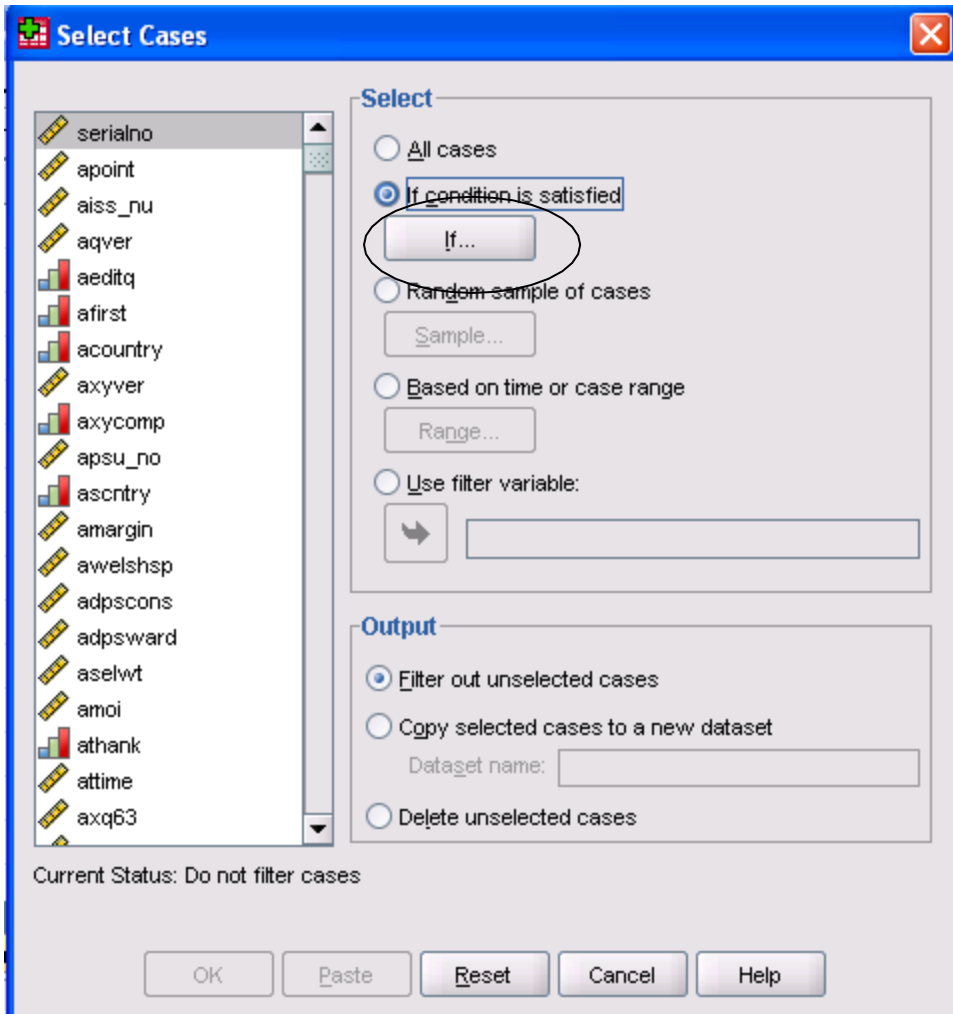


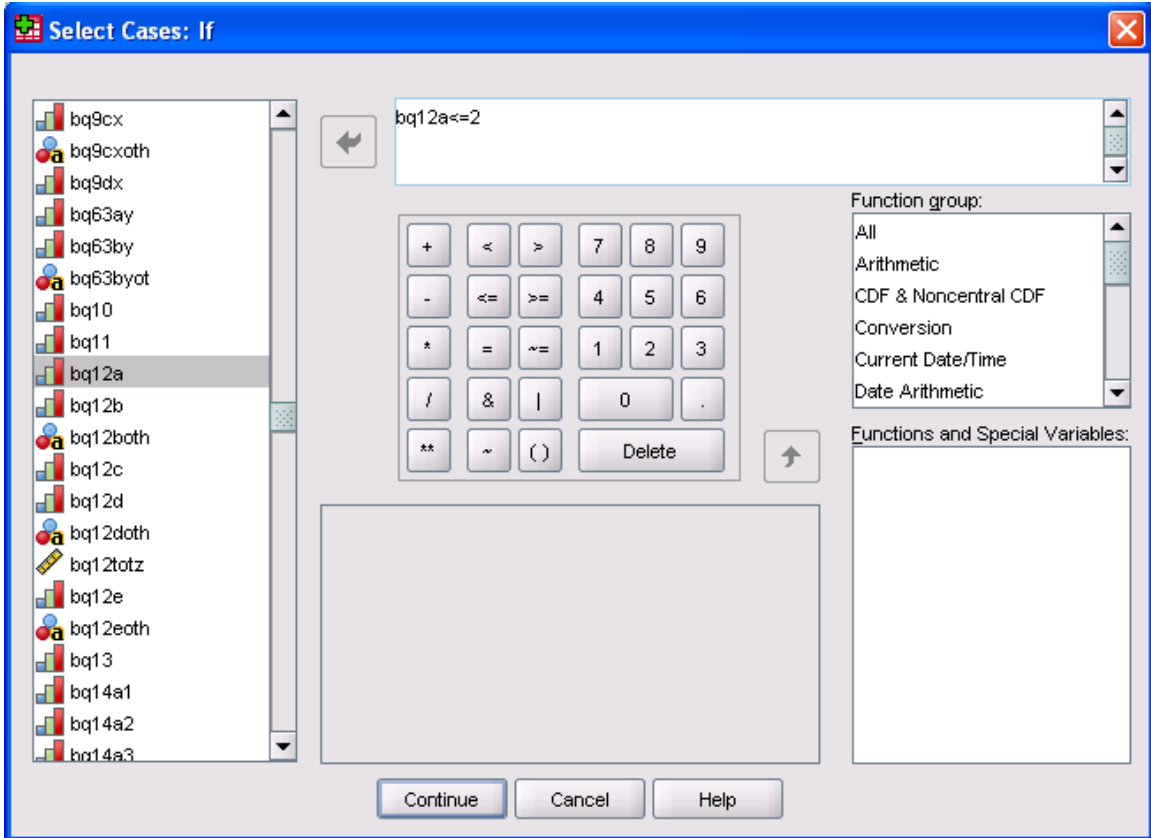
frequency of turnout from unfiltered data shown below.

Vote in General Election?

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---------|------------|-----------|---------|---------------|--------------------|
| Valid | Yes, voted | 3079 | 64.3 | 74.0 | 74.0 |
| | No | 1079 | 22.5 | 25.9 | 99.9 |
| | DK | 3 | .1 | .1 | 100.0 |
| | Total | 4161 | 86.9 | 100.0 | |
| Missing | System | 630 | 13.1 | | |
| Total | | 4791 | 100.0 | | |

We will now filter the dataset so that it only contains those people who either answered yes or no to “did you vote in the general election 2005?”.



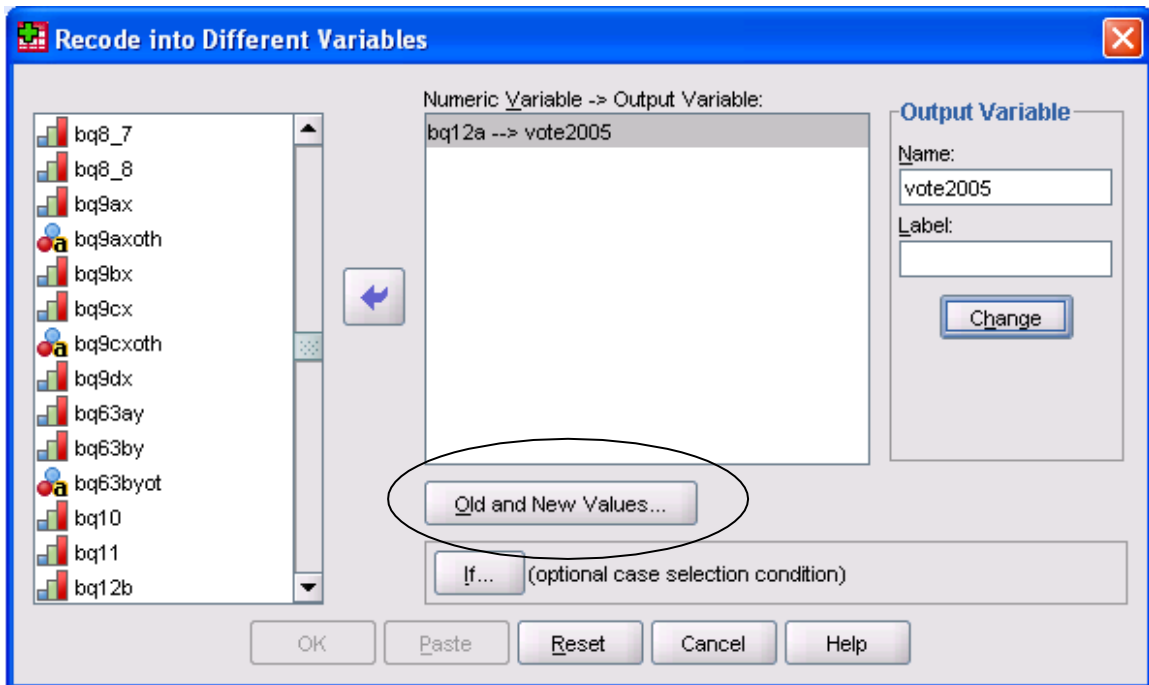
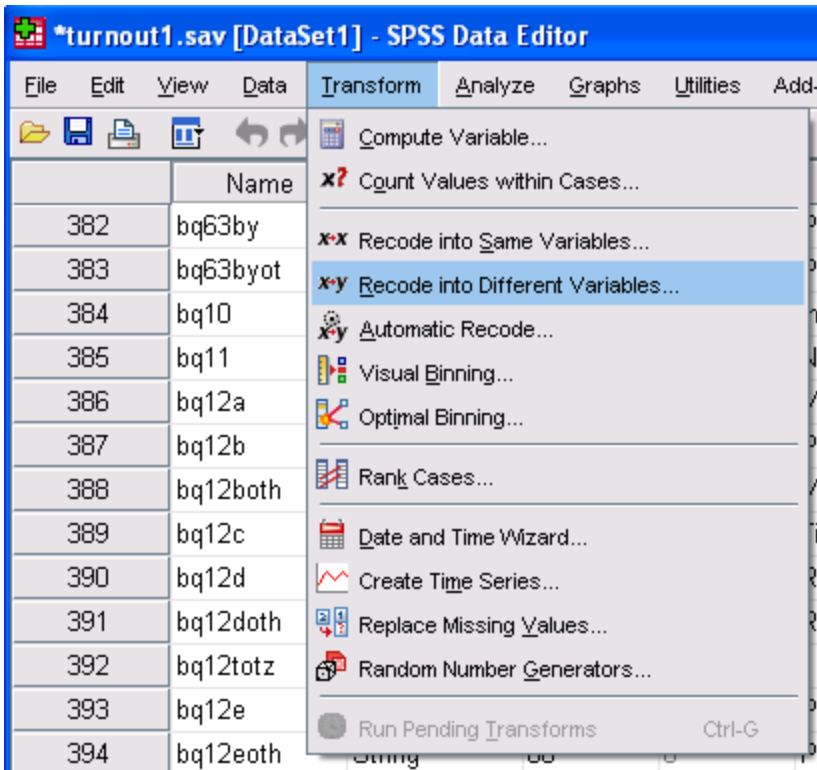


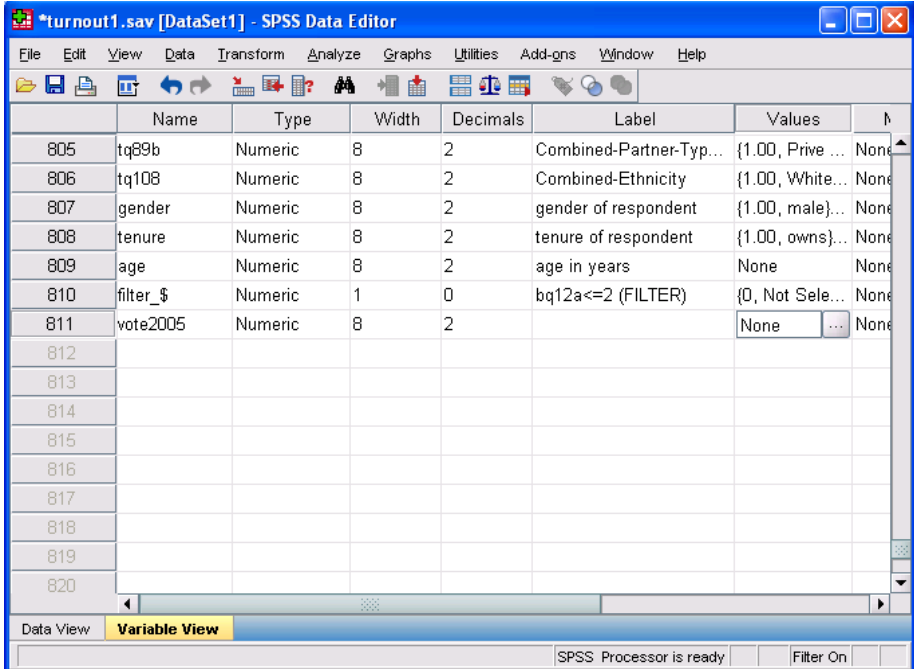
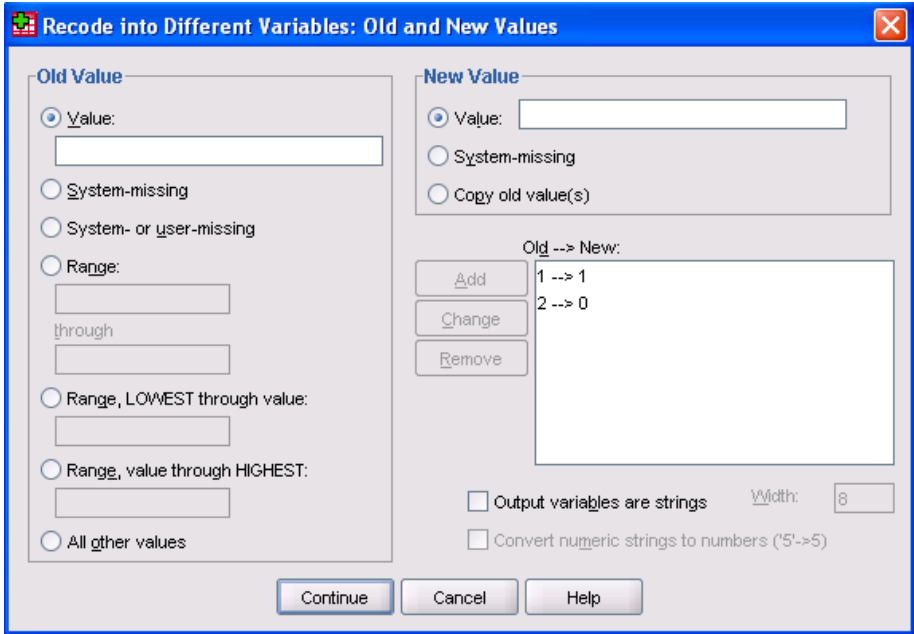
Frequency of turnout from filtered data:

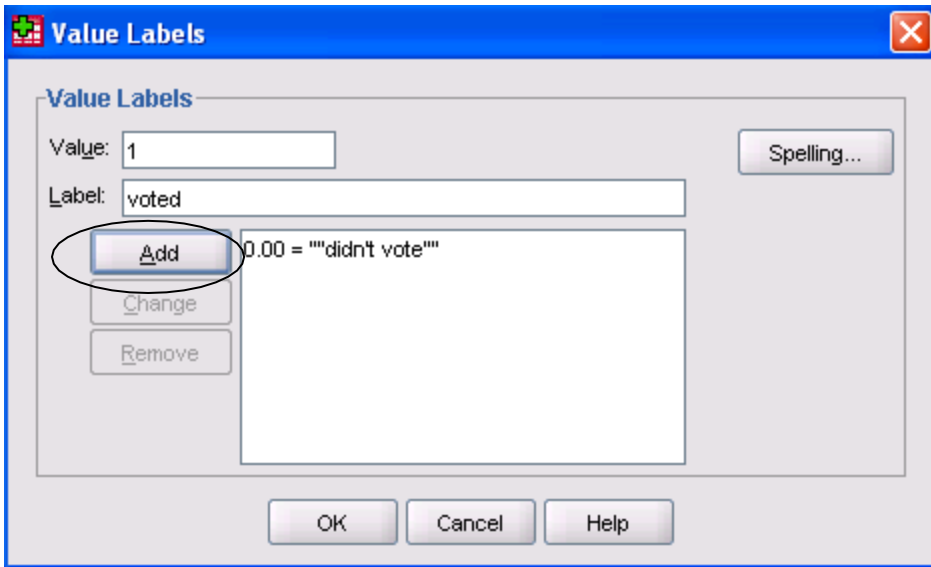
Vote in General Election?

| | Frequency | Percent | Valid Percent | Cumulative Percent |
|------------------|-----------|---------|---------------|--------------------|
| Valid Yes, voted | 3079 | 74.1 | 74.1 | 74.1 |
| No | 1079 | 25.9 | 25.9 | 100.0 |
| Total | 4158 | 100.0 | 100.0 | |

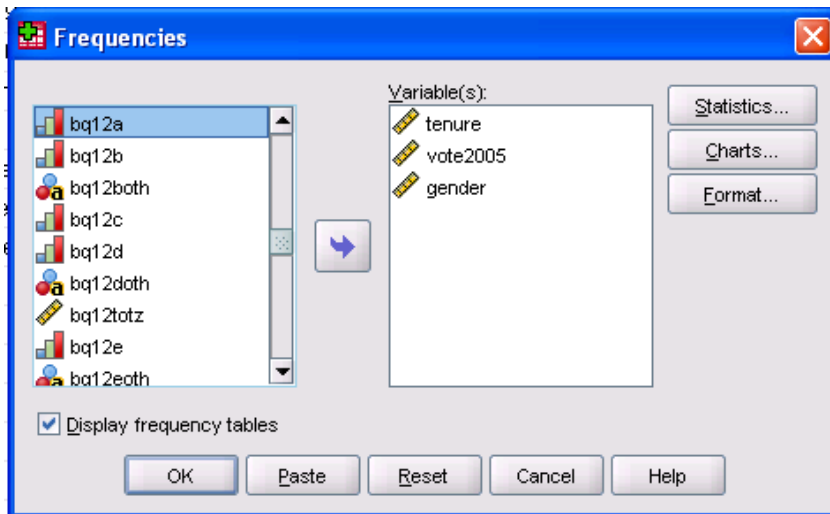
We will now recode the bq12a variable into another variable called 'vote2005' which we will recode as 0=didn't turn out to vote, 1=did turn out to vote. This will enable us to model the probability of turning out to vote, which is the response we require.







Frequencies



vote2005

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---------|-------------|-----------|---------|---------------|--------------------|
| Valid | didn't vote | 1078 | 25.9 | 26.0 | 26.0 |
| | voted | 3075 | 74.0 | 74.0 | 100.0 |
| | Total | 4153 | 99.9 | 100.0 | |
| Missing | System | 3 | .1 | | |
| Total | | 4156 | 100.0 | | |

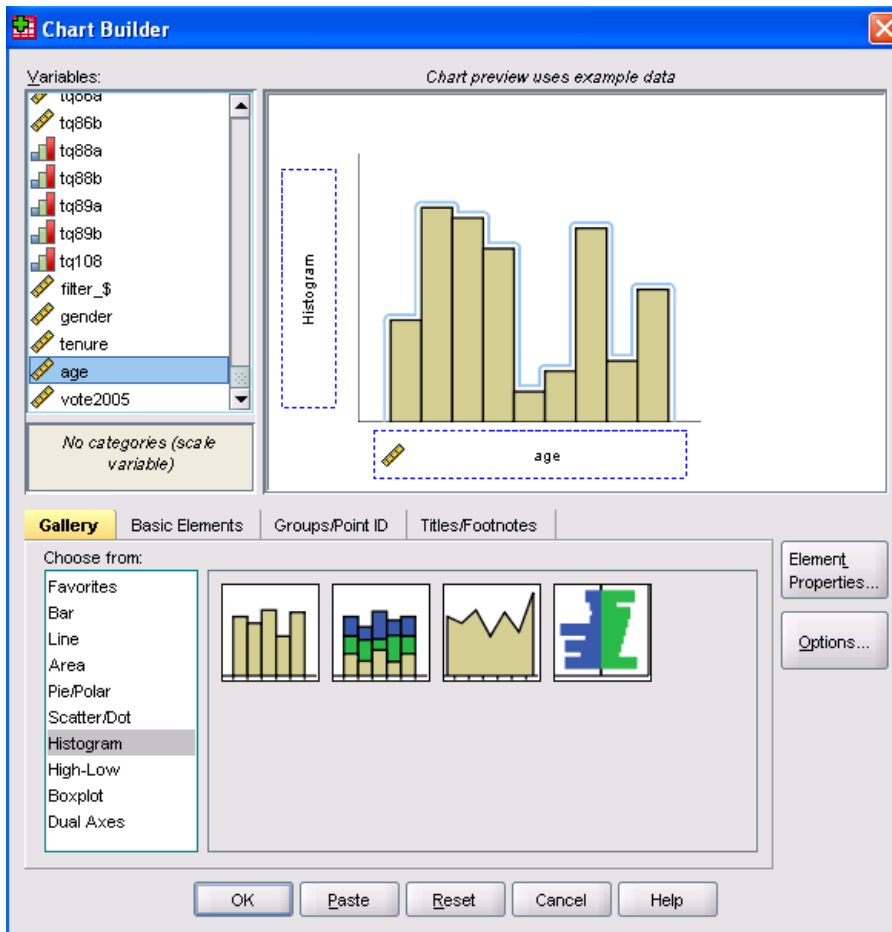
gender of respondent

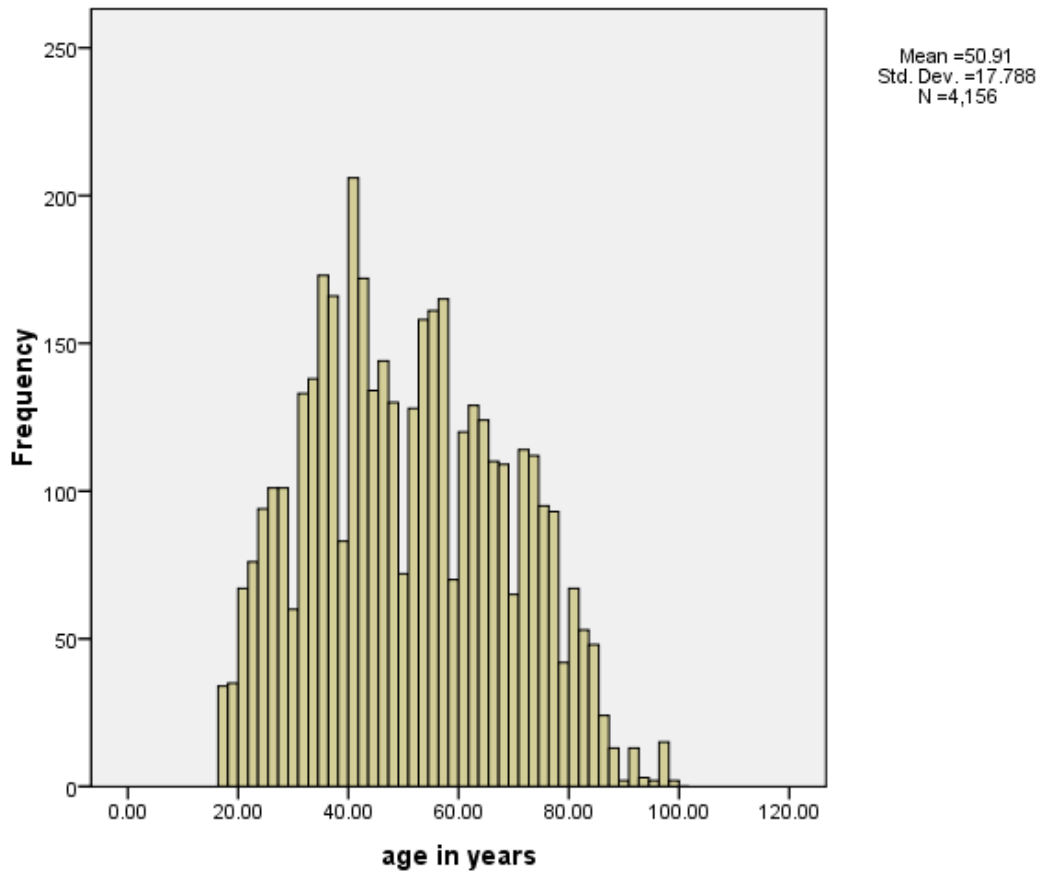
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|-------|--------|-----------|---------|---------------|--------------------|
| Valid | male | 1839 | 44.2 | 44.2 | 44.2 |
| | female | 2317 | 55.8 | 55.8 | 100.0 |
| | Total | 4156 | 100.0 | 100.0 | |

tenure of respondent

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|-------|---------|-----------|---------|---------------|--------------------|
| Valid | owns | 2987 | 71.9 | 71.9 | 71.9 |
| | rents | 1119 | 26.9 | 26.9 | 98.8 |
| | neither | 50 | 1.2 | 1.2 | 100.0 |
| | Total | 4156 | 100.0 | 100.0 | |

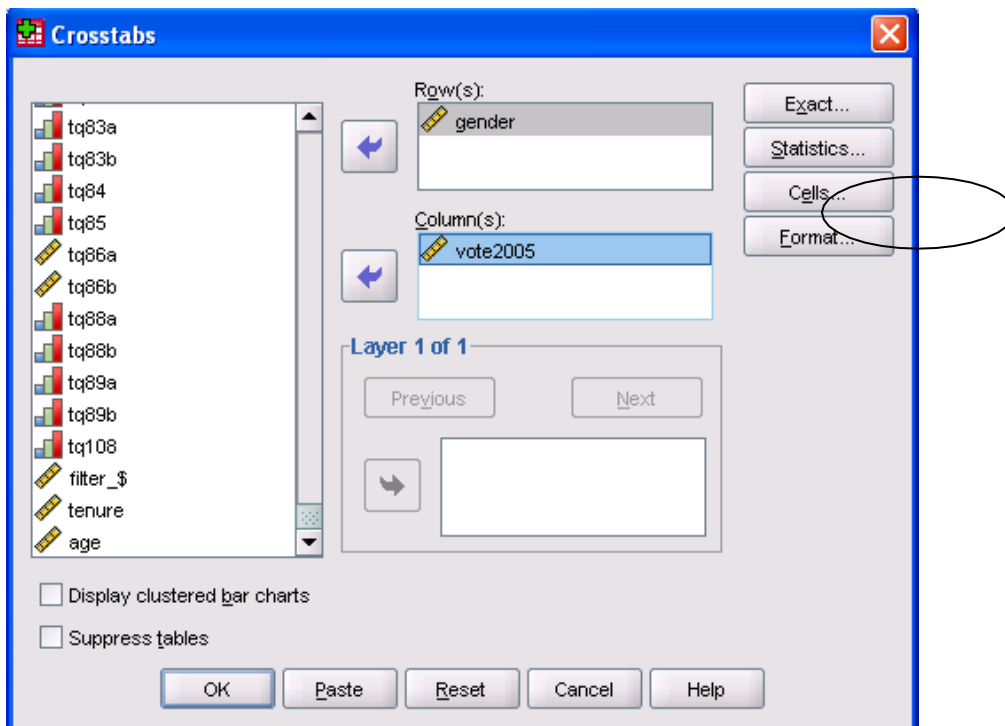
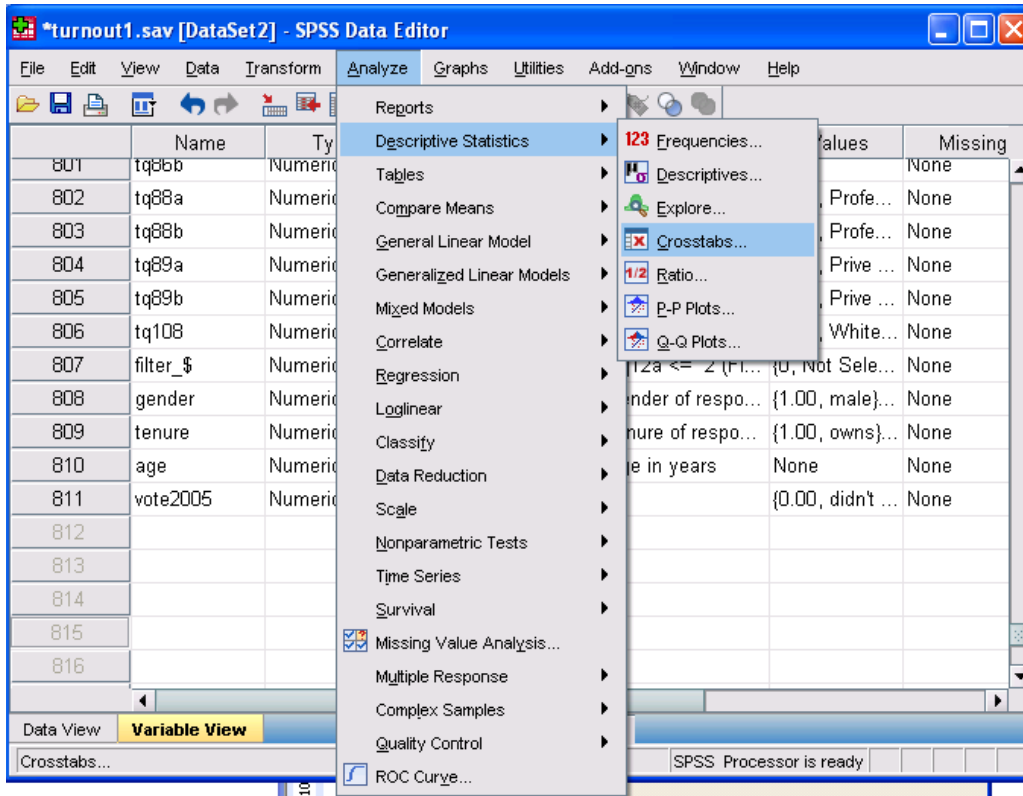
Histogram of age

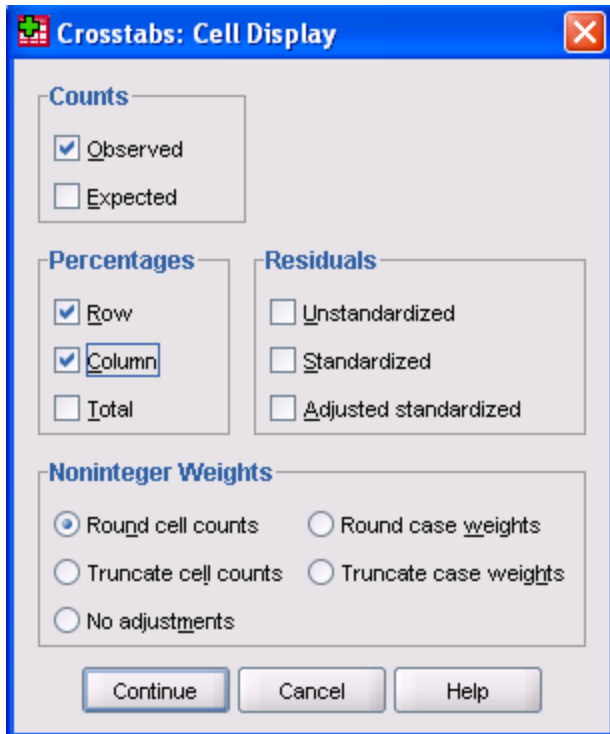




Question why are some bars much lower than their neighbours?

Cross tabulations





gender of respondent * vote2005 Crosstabulation

| | | | vote2005 | | Total |
|----------------------|-------------------------------|-------------------------------|-------------|--------|--------|
| | | | didn't vote | voted | |
| gender of respondent | male | Count | 491 | 1346 | 1837 |
| | | % within gender of respondent | 26.7% | 73.3% | 100.0% |
| | | % within vote2005 | 45.5% | 43.8% | 44.2% |
| | female | Count | 587 | 1729 | 2316 |
| | | % within gender of respondent | 25.3% | 74.7% | 100.0% |
| | | % within vote2005 | 54.5% | 56.2% | 55.8% |
| Total | Count | 1078 | 3075 | 4153 | |
| | % within gender of respondent | 26.0% | 74.0% | 100.0% | |
| | % within vote2005 | 100.0% | 100.0% | 100.0% | |

From these results we can see that:

The conditional probability of male turning out to vote are $1346/1837 = 0.733$

(which we note that when multiplied by 100 is equal to the row % in this table, given the way the cross tab is organised).

The conditional probability of female turning out to vote are $1729/2316 = 0.747$

The odds of a male turning out to vote are:

$$1346/491 = 2.741$$

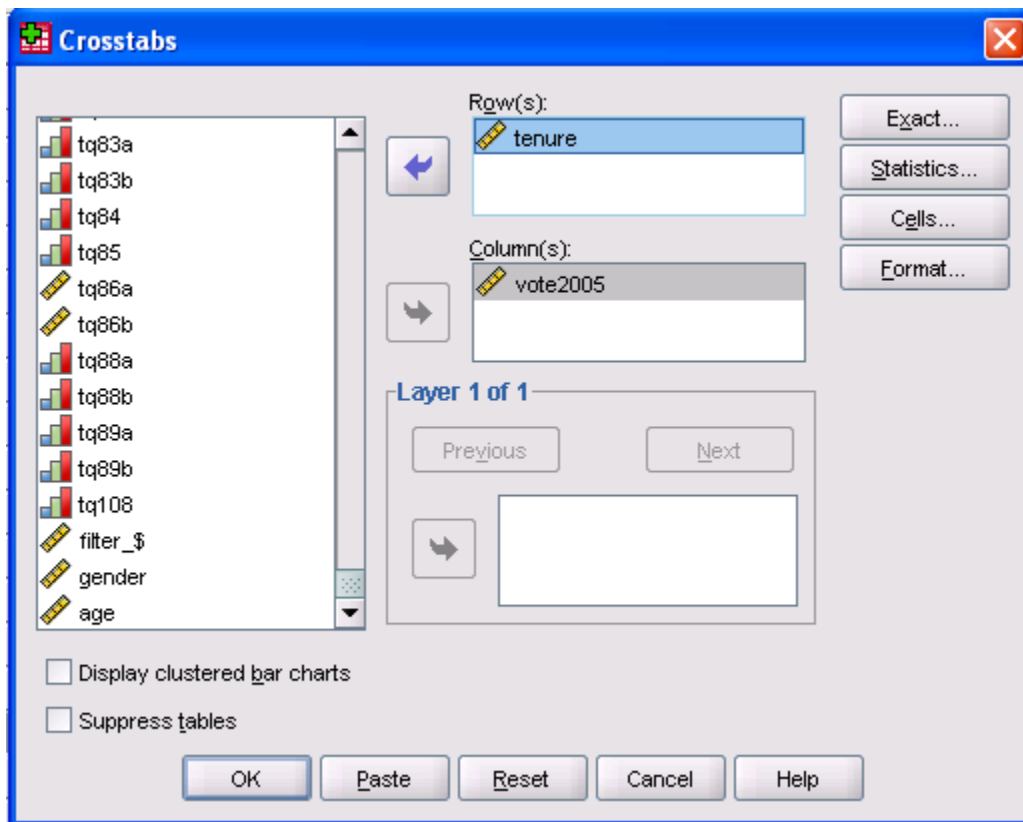
The odds of female turning out to vote are

$$1729/587 = 2.945$$

relative odds (female: male) are

$$(1729/587) / (1346/491) = 1.074$$

We will now cross tabulate vote2005 and housing tenure.



tenure of respondent * vote2005 Crosstabulation

| | | | vote2005 | | Total |
|----------------------|-------------------------------|-------------------------------|-------------|--------|--------|
| | | | didn't vote | voted | |
| tenure of respondent | owns | Count | 583 | 2404 | 2987 |
| | | % within tenure of respondent | 19.5% | 80.5% | 100.0% |
| | | % within vote2005 | 54.1% | 78.2% | 71.9% |
| | rents | Count | 481 | 636 | 1117 |
| | | % within tenure of respondent | 43.1% | 56.9% | 100.0% |
| | | % within vote2005 | 44.6% | 20.7% | 26.9% |
| | neither | Count | 14 | 35 | 49 |
| | | % within tenure of respondent | 28.6% | 71.4% | 100.0% |
| | | % within vote2005 | 1.3% | 1.1% | 1.2% |
| Total | Count | 1078 | 3075 | 4153 | |
| | % within tenure of respondent | 26.0% | 74.0% | 100.0% | |
| | % within vote2005 | 100.0% | 100.0% | 100.0% | |

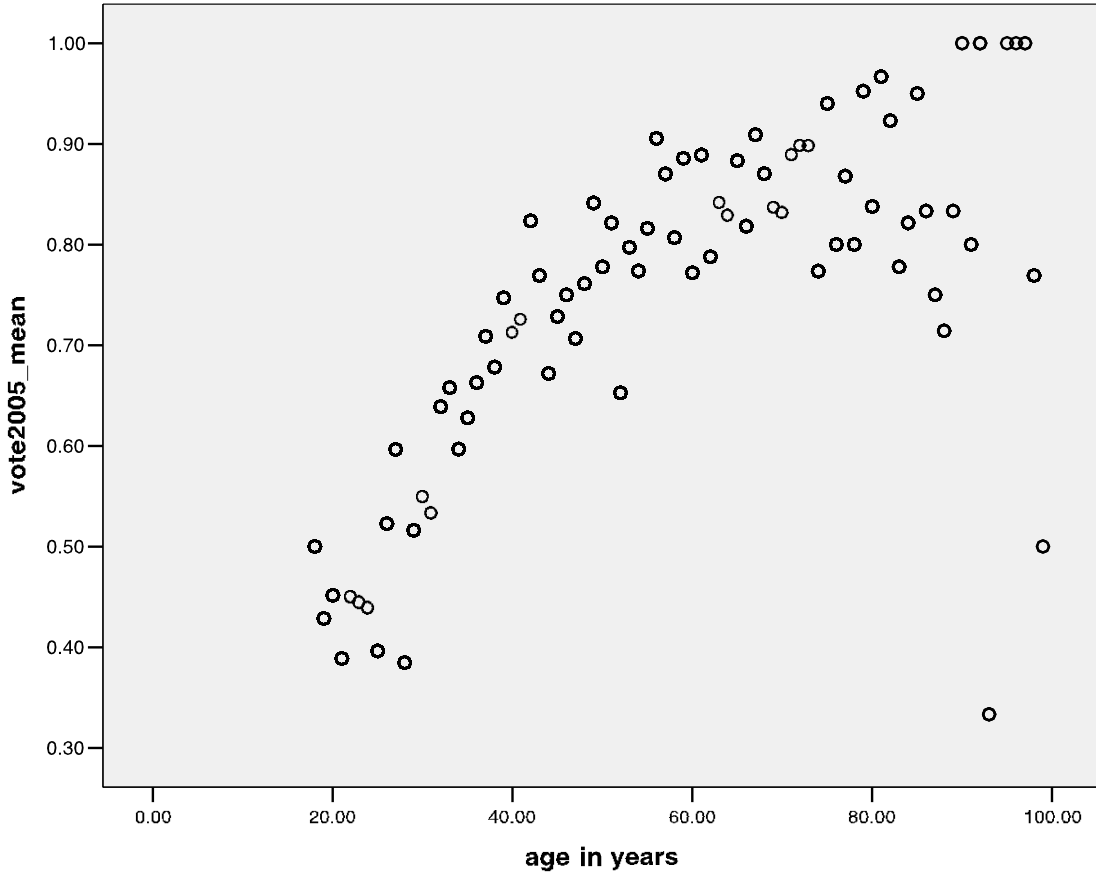
From this table we can see that, according to these data, owner occupier ('owns') are much more likely to turnout to vote than renter ('rents'). The conditional probability of owner occupier turning out to vote is 0.805 whereas for renters it is 0.569.

(If time permits, please work out the odds and relative odds for owns and rents).

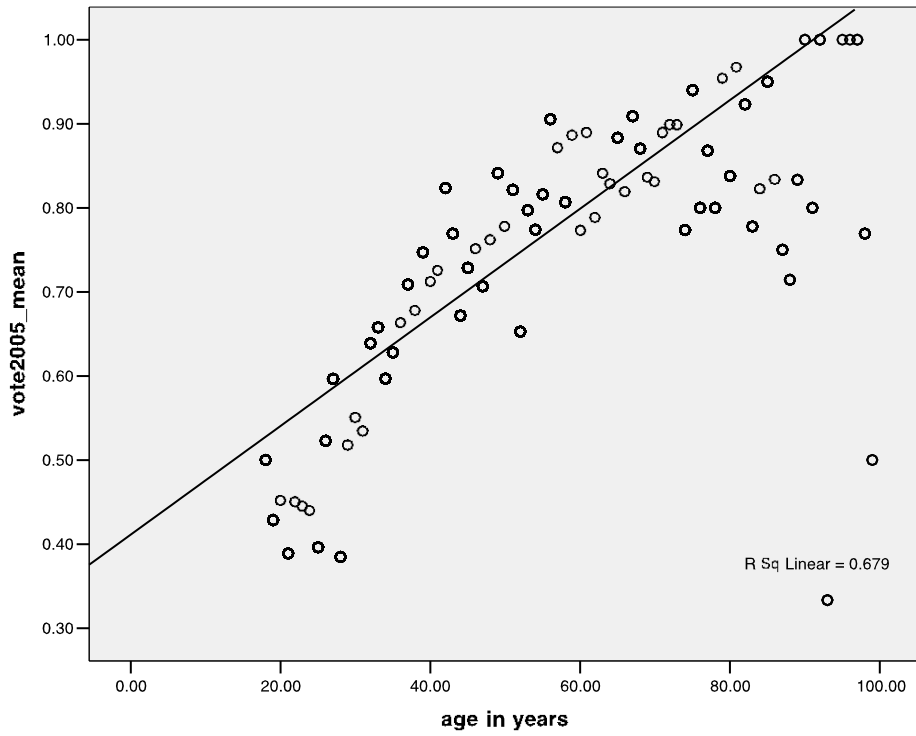
There are a total of 49 people who describe their housing tenure as neither 'owns' or 'rents'.

A scatterplot of the relationship between age and the proportion at each age turning out to vote shows that there is a much higher chance of turning out to vote when you are older.

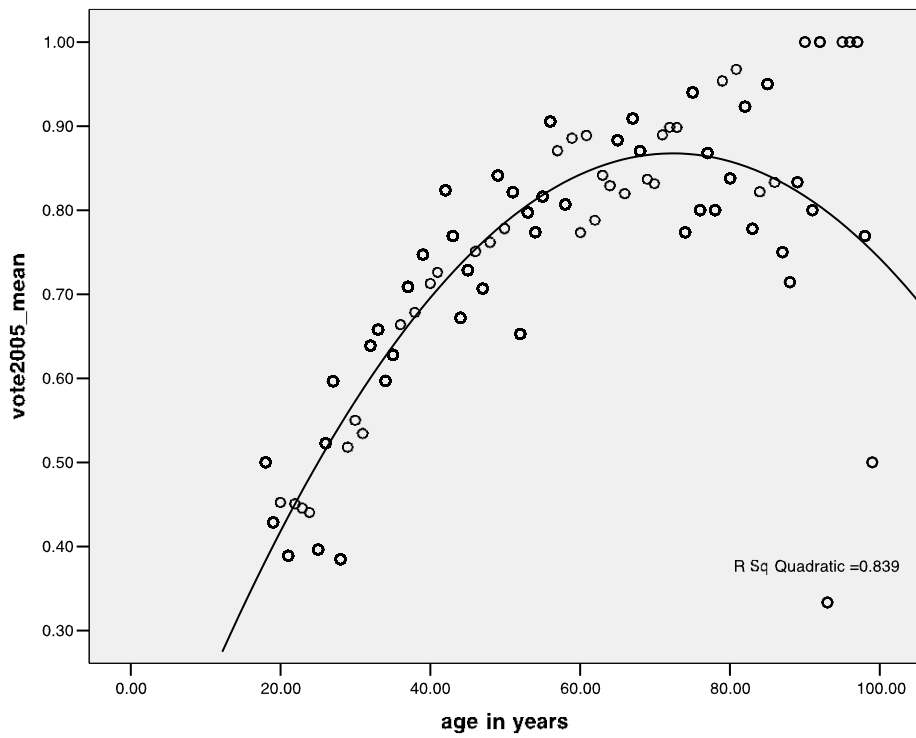
The relationship between age and propensity to turn out to vote (using the variable vote2005_mean).



Line of best fit: linear

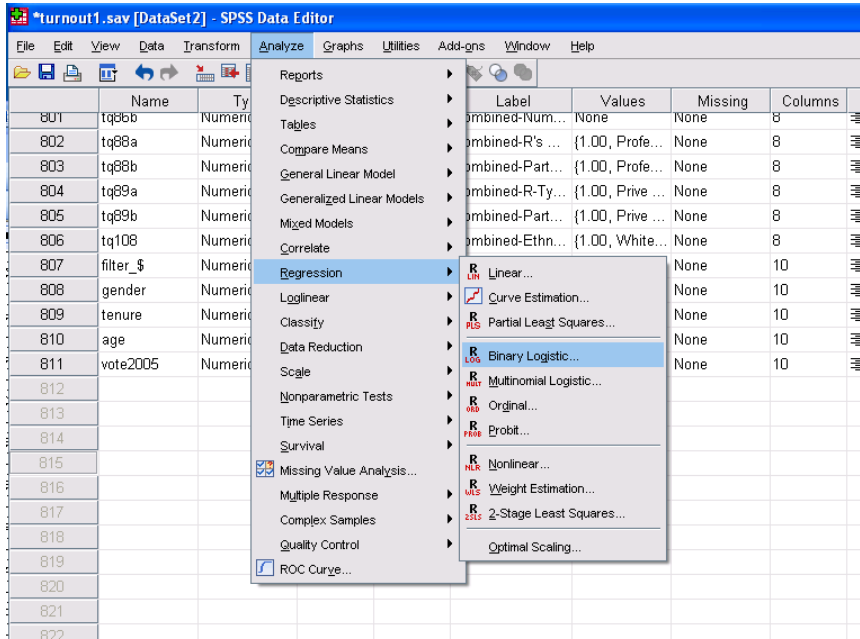


Line of best fit: quadratic

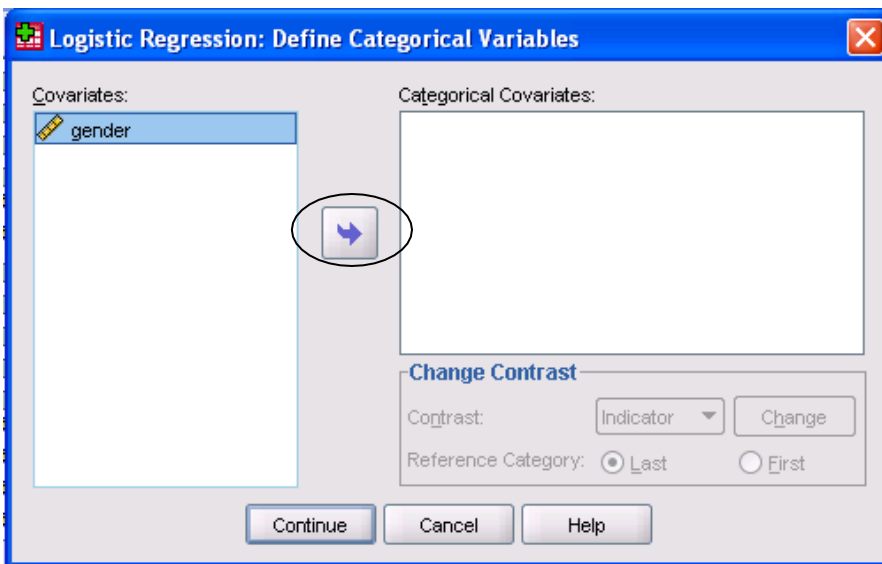
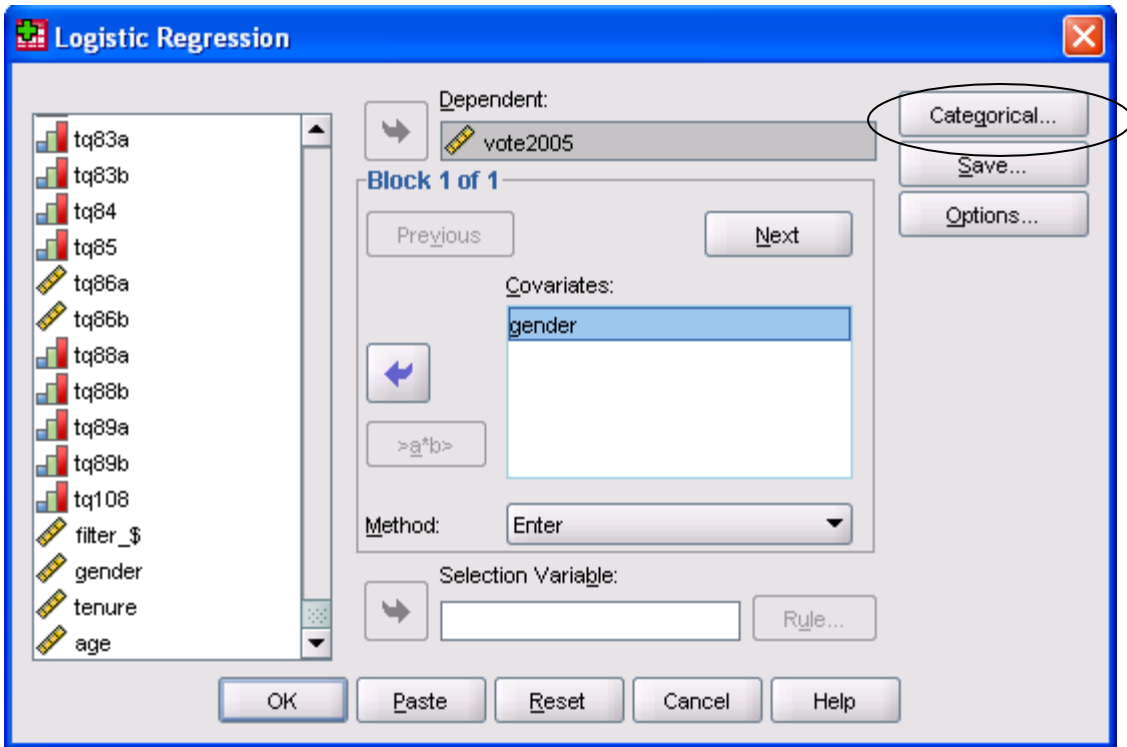


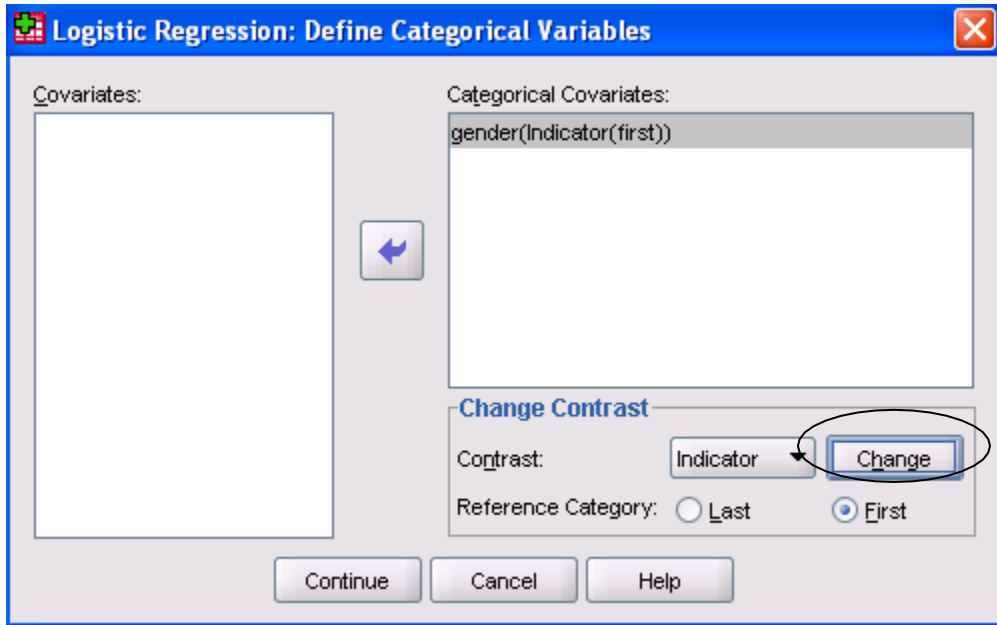
Logistic regression models

We can access the logistic regression procedure in SPSS as follows:



vote2005 is our dependent variable. We begin by adding gender – a categorical variable (which is coded on the dataset as 1=male, 2=female). We must ensure that the fact that gender is a categorical variable is declared in our analysis and we must choose the reference category. We will choose the first category, male, as the reference category as is shown below.





Now Click on Continue and then OK to run the model!

Case Processing Summary

| Unweighted Cases ^a | | N | Percent |
|-------------------------------|----------------------|------|---------|
| Selected Cases | Included in Analysis | 4153 | 99.9 |
| | Missing Cases | 3 | .1 |
| | Total | 4156 | 100.0 |
| Unselected Cases | | 0 | .0 |
| Total | | 4156 | 100.0 |

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

We can see from the table above that we are modeling 4156 cases here (some cases are deleted from the analysis where information is missing. The SPSS default for this is listwise. Only cases where all dependent and explanatory variables are complete are included in the analysis.). The tables below show us firstly that we have coded our dependent variable in the right direction and secondly that the categorical variable for gender has reference category of male. The (1) means that gender (1) in the results refers to female here.

Dependent Variable Encoding

| Original Value | Internal Value |
|----------------|----------------|
| didn't vote | 0 |
| voted | 1 |

Categorical Variables Codings

| | | Frequency | Parameter coding (1) |
|----------------------|--------|-----------|----------------------|
| gender of respondent | male | 1837 | .000 |
| | female | 2316 | 1.000 |

Block 0: Beginning Block

Classification Table^{a,b}

| Observed | | | Predicted | | |
|--------------------|----------|-------------|-------------|-------|--------------------|
| | | | vote2005 | | Percentage Correct |
| | | | didn't vote | voted | |
| Step 0 | vote2005 | didn't vote | 0 | 1078 | .0 |
| | | voted | 0 | 3075 | 100.0 |
| Overall Percentage | | | | | 74.0 |

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 | 1.048 | .035 | 876.977 | 1 | .000 | 2.853 |

Variables not in the Equation

| | | | Score | df | Sig. |
|--------------------|-----------|-----------|-------|----|------|
| Step 0 | Variables | gender(1) | 1.019 | 1 | .313 |
| Overall Statistics | | | 1.019 | 1 | .313 |

Block 1: Method = Enter

Adding one variable to the

Omnibus Tests of Model Coefficients

| | | Chi-square | df | Sig. |
|--------|-------|------------|----|------|
| Step 1 | Step | 1.018 | 1 | .313 |
| | Block | 1.018 | 1 | .313 |
| | Model | 1.018 | 1 | .313 |

We have added one new variable to the model, which has reduced the -2 log likelihood by 1.018 with 1 degree of freedom. The -2 log likelihood is a measure of how well the model explains variations in the outcome of interest, in this example turnout. The -2 log likelihood (sometimes called, deviance) has a chi squared distribution. The p value for the result of adding gender to the model is given in the table above and we can see that this is 0.313 which is greater than the conventional significance level of 0.05. hence we would conclude that the addition of gender to the model is not statistically significant. In other words this variable does not explain variations in turnout.

Model Summary

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|------|-----------------------|----------------------|---------------------|
| 1 | 4755.065 ^a | .000 | .000 |

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table^a

| Observed | | Predicted | | | |
|--------------------|----------|-------------|-------|--------------------|-------|
| | | vote2005 | | Percentage Correct | |
| | | didn't vote | voted | | |
| Step 1 | vote2005 | didn't vote | 0 | 1078 | .0 |
| | | voted | 0 | 3075 | 100.0 |
| Overall Percentage | | | | | 74.0 |

a. The cut value is .500

Variables in the Equation

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|------|-----------|-------|------|---------|----|------|--------|
| Step | gender(1) | .072 | .071 | 1.019 | 1 | .313 | 1.074 |
| 1 | Constant | 1.008 | .053 | 365.868 | 1 | .000 | 2.741 |

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

We see from the table above that the estimated model is

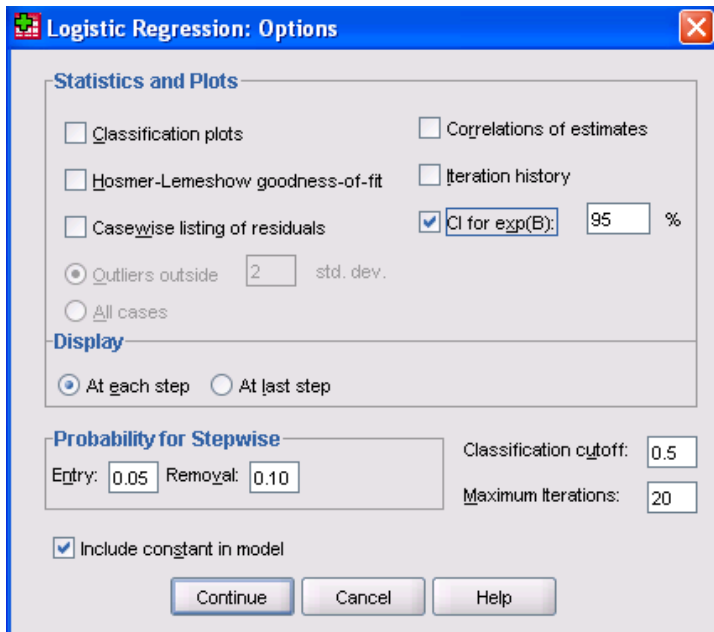
$$\text{Logit}(\text{vote2005}) = 1.008 + 0.072 \text{ gender}(1)$$

As we have recoded gender to 0=male, 1=female, this is equivalent to:

$$\text{Logit}(\text{vote2005}) = 1.008 + 0.072 \text{ female}$$

We can see that the coefficient of gender is non-significant (sig = 0.313 > 0.05).

The Exp(B) column shows the relative odds (odds ratio) and indicates that females are 1.074 times as likely to turnout to vote than males. We can request a confidence interval for this result as shown below.



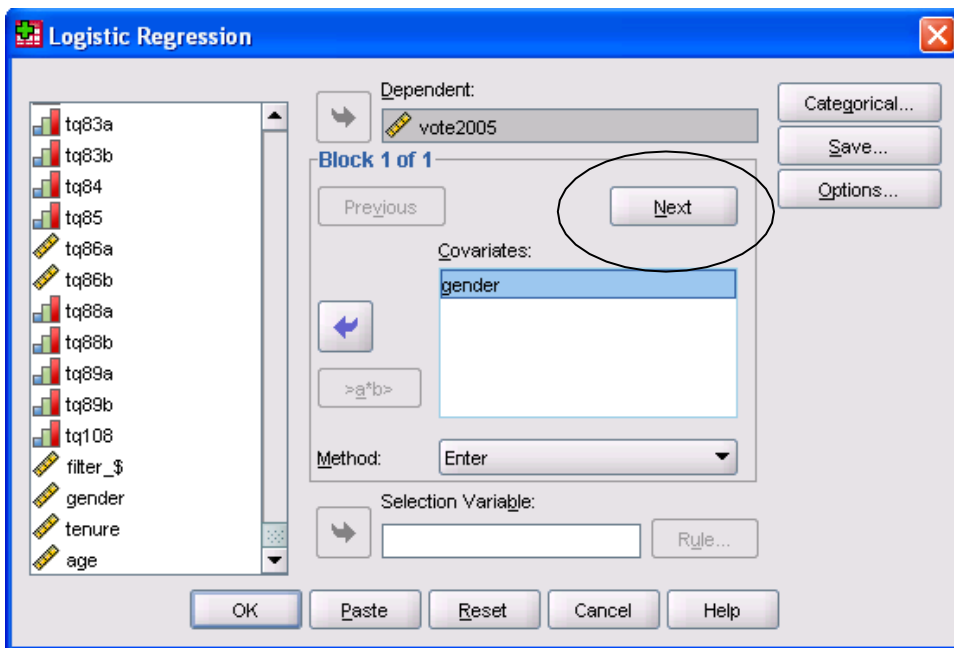
Variables in the Equation

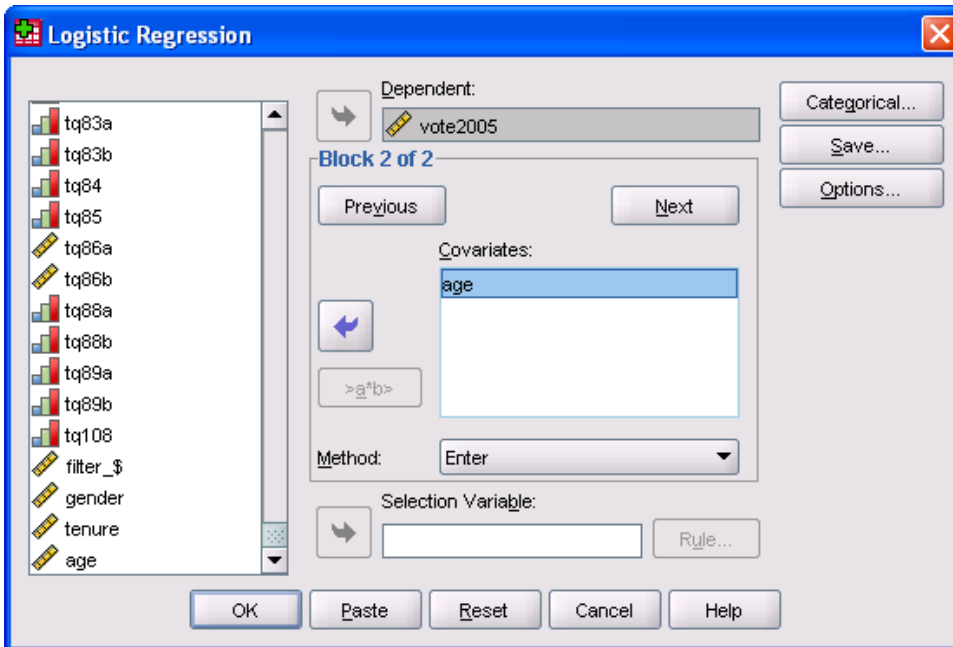
| Step | | B | S.E. | Wald | df | Sig. | Exp(B) | 95.0% C.I. for EXP(B) | |
|------|-----------|-------|------|---------|----|------|--------|-----------------------|-------|
| | | | | | | | | Lower | Upper |
| 1 | gender(1) | .072 | .071 | 1.019 | 1 | .313 | 1.074 | .935 | 1.235 |
| | Constant | 1.008 | .053 | 365.868 | 1 | .000 | 2.741 | | |

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

The confidence interval for exp(B) is 0.935 to 1.235 indicates that females are between 0.935 and 1.235 times as likely to turn out to vote than males. i.e. the range has a lower limit of 'slightly less than males' and upper limit of 'slightly more than males' and therefore includes 'males and females are equally likely to turn out to vote (i.e. exp(B)=1). This is not surprising since we have already concluded that gender has no statistically significant explanatory power in explaining variations in turnout. We will now add an additional variable to the model – age in years (which is a continuous variable, rather than a categorical one).

We will make use of the 'block' procedure to add age to the model, so that we can see both the effect of adding age alone on the -2 log likelihood as well as seeing how a model which includes both age and gender might explain variations in turnout.





Block 2: Method = Enter

Omnibus Tests of Model Coefficients

| | | Chi-square | df | Sig. |
|--------|-------|------------|----|------|
| Step 1 | Step | 300.666 | 1 | .000 |
| | Block | 300.666 | 1 | .000 |
| | Model | 301.684 | 2 | .000 |

The addition of age to the model has, as a single variable, reduced the -2 log likelihood by 300.666 on 1 degree of freedom. The model, which now contains 2 parameters, gender and age has collectively reduced -2 log likelihood by 301.684 but we can see it is age that has the explanatory power, and gender is not adding anything extra.

Model Summary

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|------|-----------------------|----------------------|---------------------|
| 1 | 4454.399 ^a | .070 | .103 |

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

the model which includes gender and age explains between 7 and 10% of the variation in turnout.

Classification Table^a

| Observed | | | Predicted | | Percentage Correct |
|--------------------|----------|-------------|-------------|-------|--------------------|
| | | | vote2005 | | |
| | | | didn't vote | voted | |
| Step 1 | vote2005 | didn't vote | 36 | 1042 | 3.3 |
| | | voted | 42 | 3033 | 98.6 |
| Overall Percentage | | | | | 73.9 |

a. The cut value is .500

Variables in the Equation

| | B | S.E. | Wald | df | Sig. | Exp(B) | 95.0% C.I. for EXP(B) | | |
|--------|-----------|-------|------|---------|------|--------|-----------------------|-------|-------|
| | | | | | | | Lower | Upper | |
| Step 1 | gender(1) | .077 | .074 | 1.087 | 1 | .297 | 1.080 | .935 | 1.248 |
| | age | .037 | .002 | 267.015 | 1 | .000 | 1.038 | 1.033 | 1.042 |
| | Constant | -.779 | .118 | 43.942 | 1 | .000 | .459 | | |

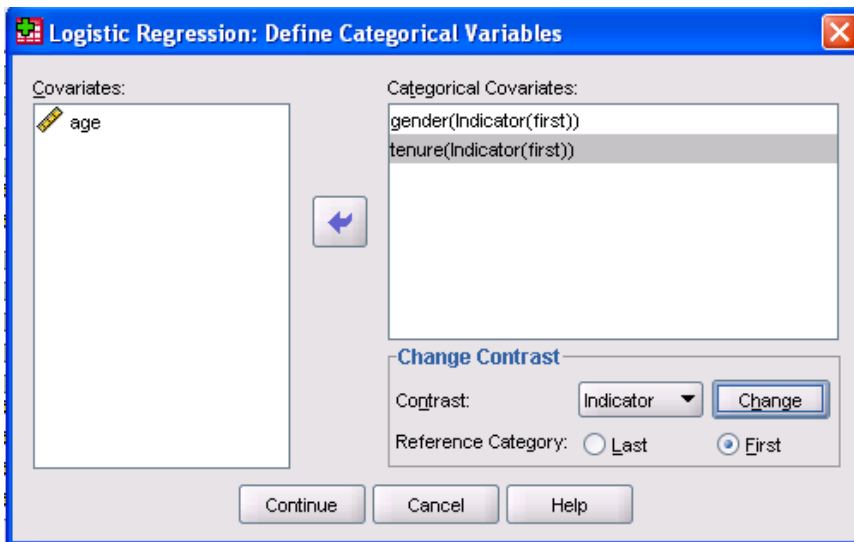
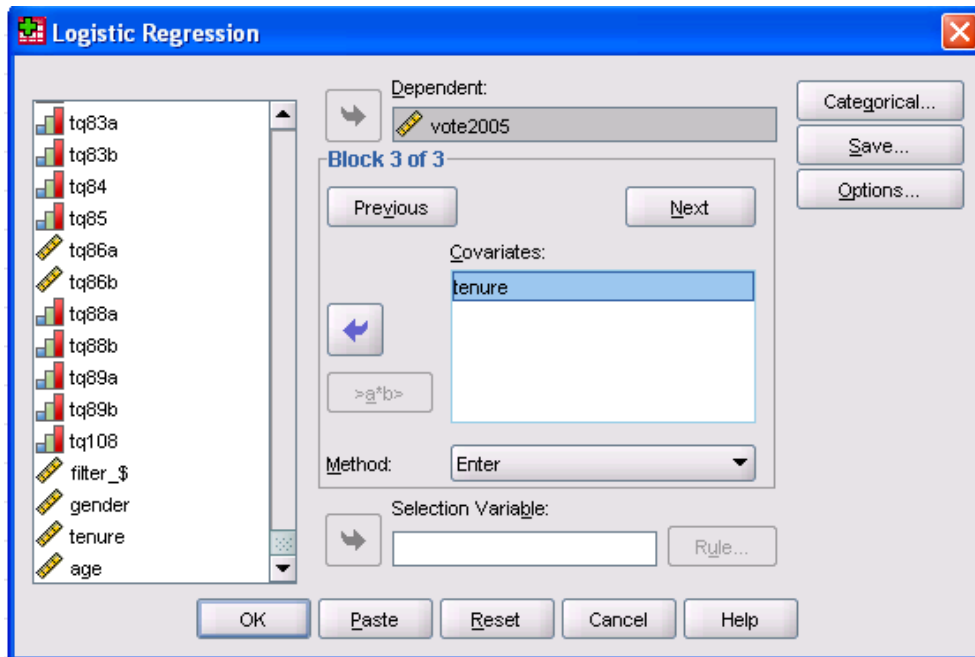
a. Variable(s) entered on step 1: age.

the model is now:

$$\text{logit}(\text{vote2005}) = -.779 + .077\text{gender}(1) + .037\text{age}$$

The age coefficient is statistically significant. Exp(B) for age is 1.038, which means for each year different in age, the person is 1.038 times more likely to turn out to vote, having allowed for gender in the model. Eg. a 21 year old is 1.038 times as likely to turn out to vote than a 20 year old. This might not seem much of a difference but a 20 year difference leads to a person being $1.038^{20} = 2.11$ times more likely to turn out to vote. E.g. a 40 year old is 2.11 times more likely to turn out to vote than a 20 year old, having allowed for gender in the model.

We can add housing tenure to the model, as 'block 3' tenure is a categorical variable, as was gender. By declaring it as categorical we can set up dummy (indicator) variables and make the first category 'owns' the reference category:



Click on Continue and then OK to run the model again.

Case Processing Summary

| Unweighted Cases ^a | | N | Percent |
|-------------------------------|----------------------|------|---------|
| Selected Cases | Included in Analysis | 4153 | 99.9 |
| | Missing Cases | 3 | .1 |
| | Total | 4156 | 100.0 |
| Unselected Cases | | 0 | .0 |
| Total | | 4156 | 100.0 |

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

Dependent Variable Encoding

| Original Value | Internal Value |
|----------------|----------------|
| didn't vote | 0 |
| voted | 1 |

Categorical Variables Codings

| | | Frequency | Parameter coding | |
|----------------------|---------|-----------|------------------|-------|
| | | | (1) | (2) |
| tenure of respondent | owns | 2987 | .000 | .000 |
| | rents | 1117 | 1.000 | .000 |
| | neither | 49 | .000 | 1.000 |
| gender of respondent | male | 1837 | .000 | |
| | female | 2316 | 1.000 | |

Block 0: Beginning Block

Classification Table^{a,b}

| | | | Predicted | | |
|--------------------|----------|-------------|-------------|-------|--------------------|
| | | | vote2005 | | Percentage Correct |
| | | | didn't vote | voted | |
| Observed | vote2005 | didn't vote | 0 | 1078 | .0 |
| | | voted | 0 | 3075 | 100.0 |
| Overall Percentage | | | | | 74.0 |

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 | 1.048 | .035 | 876.977 | 1 | .000 | 2.853 |

Variables not in the Equation

| | Score | df | Sig. |
|----------------------------|-------|----|------|
| Step 0 Variables gender(1) | 1.019 | 1 | .313 |
| Overall Statistics | 1.019 | 1 | .313 |

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

| | Chi-square | df | Sig. |
|-------------|------------|----|------|
| Step 1 Step | 1.018 | 1 | .313 |
| Block | 1.018 | 1 | .313 |
| Model | 1.018 | 1 | .313 |

Model Summary

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|------|-----------------------|----------------------|---------------------|
| 1 | 4755.065 ^a | .000 | .000 |

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table^a

| | | Predicted | | |
|--------------------|-------------|-----------|------|--------------------|
| | | vote2005 | | Percentage Correct |
| Observed | didn't vote | voted | | |
| Step 1 vote2005 | didn't vote | 0 | 1078 | .0 |
| | voted | 0 | 3075 | 100.0 |
| Overall Percentage | | | | 74.0 |

a. The cut value is .500

Variables in the Equation

| | B | S.E. | Wald | df | Sig. | Exp(B) | 95.0% C.I. for EXP(B) | |
|------------------|-------|------|---------|----|------|--------|-----------------------|-------|
| | | | | | | | Lower | Upper |
| Step 1 gender(1) | .072 | .071 | 1.019 | 1 | .313 | 1.074 | .935 | 1.235 |
| Constant | 1.008 | .053 | 365.868 | 1 | .000 | 2.741 | | |

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

Block 2: Method = Enter

Omnibus Tests of Model Coefficients

| | | Chi-square | df | Sig. |
|--------|-------|------------|----|------|
| Step 1 | Step | 300.666 | 1 | .000 |
| | Block | 300.666 | 1 | .000 |
| | Model | 301.684 | 2 | .000 |

Model Summary

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|------|-----------------------|----------------------|---------------------|
| 1 | 4454.399 ^a | .070 | .103 |

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table^a

| Observed | | Predicted | | | |
|--------------------|----------|-------------|-------|--------------------|------|
| | | vote2005 | | Percentage Correct | |
| | | didn't vote | voted | | |
| Step 1 | vote2005 | didn't vote | 36 | 1042 | 3.3 |
| | | voted | 42 | 3033 | 98.6 |
| Overall Percentage | | | | | 73.9 |

a. The cut value is .500

Variables in the Equation

| | B | S.E. | Wald | df | Sig. | Exp(B) | 95.0% C.I. for EXP(B) | | |
|--------|-----------|-------|------|---------|------|--------|-----------------------|-------|-------|
| | | | | | | | Lower | Upper | |
| Step 1 | gender(1) | .077 | .074 | 1.087 | 1 | .297 | 1.080 | .935 | 1.248 |
| | age | .037 | .002 | 267.015 | 1 | .000 | 1.038 | 1.033 | 1.042 |
| | Constant | -.779 | .118 | 43.942 | 1 | .000 | .459 | | |

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

Block 3: Method = Enter

Omnibus Tests of Model Coefficients

| | | Chi-square | df | Sig. |
|--------|-------|------------|----|------|
| Step 1 | Step | 174.499 | 2 | .000 |
| | Block | 174.499 | 2 | .000 |
| | Model | 476.183 | 4 | .000 |

We can immediately see that tenure reduces the -2 log likelihood by 174.499 having added 2 new variables (tenure has 3 categories in all so we need 2 dummy variables). Tenure is statistically significant in this model.

Model Summary

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|------|-----------------------|----------------------|---------------------|
| 1 | 4279.900 ^a | .108 | .159 |

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table^a

| Observed | | | Predicted | | |
|--------------------|----------|-------------|-------------|-------|--------------------|
| | | | vote2005 | | Percentage Correct |
| | | | didn't vote | voted | |
| Step 1 | vote2005 | didn't vote | 273 | 805 | 25.3 |
| | | voted | 180 | 2895 | 94.1 |
| Overall Percentage | | | | | 76.3 |

a. The cut value is .500

Variables in the Equation

| | B | S.E. | Wald | df | Sig. | Exp(B) | 95.0% C.I. for EXP(B) | | |
|--------|-----------|--------|------|---------|------|--------|-----------------------|-------|-------|
| | | | | | | | Lower | Upper | |
| Step 1 | | | | | | | | | |
| | gender(1) | .101 | .076 | 1.769 | 1 | .183 | 1.106 | .953 | 1.283 |
| | age | .034 | .002 | 227.894 | 1 | .000 | 1.035 | 1.030 | 1.039 |
| | tenure | | | 176.580 | 2 | .000 | | | |
| | tenure(1) | -1.053 | .079 | 176.555 | 1 | .000 | .349 | .299 | .407 |
| | tenure(2) | -.324 | .335 | .934 | 1 | .334 | .723 | .375 | 1.396 |
| | Constant | -.326 | .123 | 7.008 | 1 | .008 | .721 | | |

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

The table above shows us that the estimated model is now:

$$\text{logit}(\text{vote2005}) = -.326 + .101\text{gender}(1) + .034\text{age} - 1.053\text{tenure}(1) - .324\text{tenure}(2)$$

in other words

$$\text{logit}(\text{vote2005}) = -.326 + .101\text{female} + .034\text{age} - 1.053\text{rents} - .324\text{neither}$$

tenure(1) which contrasts 'rents' with 'owns' has an $\exp(B)$ of 0.349 which means that a person who rents is only .349 times (i.e. much less) likely to turn out than a person who owns their own property, having allowed for gender and age. If we calculate the inverse of $\exp(B)$ here, i.e. $1/0.349 = 2.87$, we can say that a person who owns their own home is 2.87 times more likely to vote than someone who rents, having allowed for gender and age.

4) Summary and further comments

Summary

We have seen how logistic regression analysis may be used to analyse tabular data where one of the dimensions of the table is an outcome of interest. This morning, we looked at some examples where we calculated the probabilities, odds and relative odds from the table, and we have seen how we can also calculate these (and get the same results) from the model parameter estimates. Some theory was introduced and we saw how the logistic model framework is a good way to investigate associations in multi-way tables where one of the dimensions of the tables is an outcome of interest, with two categories.

We have seen how we can use SPSS to fit logistic regression models to data using an example based on the 2005 UK election. We covered main effects models and models with interactions and we went through the output that SPSS gives us, including the classification table, the deviance, the model coefficients and other useful measures such as $\exp(B)$, which gives the relative odds or odds ratio for a particular explanatory variable, given the other explanatory variables in the model.

Further comments

The term ‘generalised linear model’ is used to describe a procedure for transforming the dependent variable so that the ‘right hand side’ of the model equation can be interpreted as a ‘linear combination’ of the explanatory variables:

$$f(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

In situations where the dependent (y) variable is continuous and can be reasonably assumed to have a normal distribution we do not transform the y variable at all and we can simply run a multiple linear regression analysis.

In situations where the dependent variable is dichotomous or 0/1 as we have seen today the most common procedure is to use logistic regression, using the logit link as we have done today. Other similar types of modelling include probit modelling. (See Dobson, McCullagh and Nelder for further details – details of these references in the reading list).

When the response variable has several categories we can use a model that allows for several categories in the response variable such as multinomial regression. If this response variable is ordinal (as opposed to nominal) we can allow for this in the modelling (see Agresti – reference details in reading list). An alternative is to recode the response variable into just two categories and do a logistic regression analysis (or to fit several logistic regression models to different pairs of categories in the response variable, although this is not as statistically efficient as doing a true multinomial analysis).

Note also that logistic regression models can also be fitted with multilevel components in MLwiN and STATA.

Reading list

Field, A. (2005) *Discovering statistics using SPSS for Windows: advanced techniques for the beginner*, London: Sage. Chapter 6.

Plewis, I (1997) *Statistics in Education*, Edward Arnold. (Especially chapter 5)

Dobson, A (2001) *An introduction to generalized linear models (second edition)*. Chapman and Hall.

McCullagh P and Nelder J.A, (1989) *Generalized linear models (second edition)*. Chapman and Hall

Agresti, A. (1996) *Introduction to categorical data analysis*. John Wiley.