Session 6

Logistic Regression

Session 6: Logistic Regression

Analysis of Binary Data

Consider the data on age of menarche for a sample of Warsaw girls.

Each girl was asked whether she had had her first period. The data was then grouped by age into fairly narrow age groups. Within each age group the total number of girls (**N**) was recorded and the number who had had their first period (**R**). Also recorded is the mid-point of the age range for that age group (**AGE**).

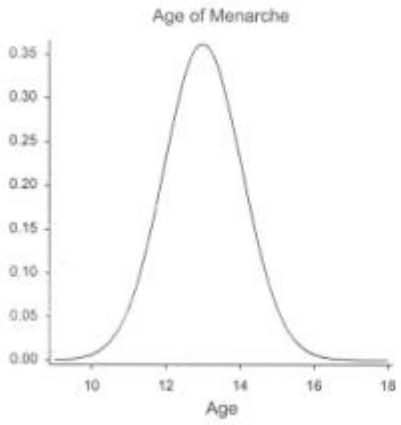
How should we analyse such data?

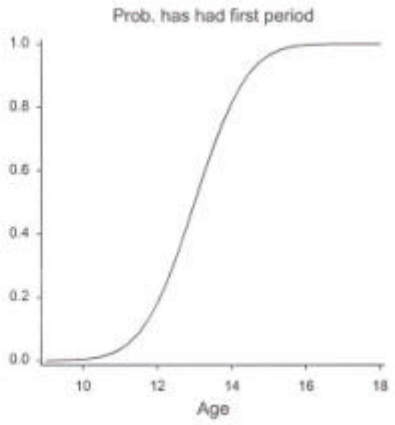
If we had conducted a longitudinal study over a long time we would have been able to establish each girl's age when she had her first period. We could then have studied the distribution of these ages.

However, the study was cross-sectional taken at a particular point in time, so that this is not how the data was recorded since;

i. For those girls who had not had their first period we cannot record what age they would be when this occurred.
ii. For those girls who had had their first period it was felt that their memory of the exact age they were was unreliable. Thus only the fact that they had had their first period was recorded.

R	N	AGE
0	376	9.21
0	200	10.21
0	93	10.58
2	120	10.83
2	90	11.08
5	88	11.33
10	105	11.58
17	111	11.83
16	100	12.08
29	93	12.33
39	100	12.58
51	100	12.83
47	99	13.08
67	106	13.33
81	105	13.58
88	117	13.83
79	98	14.08
90	97	14.33
113	120	14.58
95	102	14.83
117	122	15.08
107	111	15.33
92	94	15.58
112	114	15.83
1049	1049	17.58





It is clear that we can still base our model on the idea of an underlying distribution of the age of menarche.
We now consider the probability that a girl of a certain age will have had her first period at some time previous. This is the probability that her age of menarche is less than her current age. In terms of the distribution of age of menarche it is the cumulative distribution function derived from this distribution

Models for Binary Data

For Normal regression models we have two basic elements:

i. A linear relationship between the mean of the dependent variable and explanatory variables.
ii. A normal distribution which describes the sampling variation of the observations.

Binary data consists of only two possible observations, yes/no, right/wrong, dead/alive etc. We clearly need different modelling assumptions. Consider first the case where we record **R** "positive" responses out of the **N** samples. For example, where we record that **R** girls out of the total of **N** girls in an age group had had their first period. We wish to consider the relationship between the probability of a positive response (i.e. has had her first period) and explanatory variables (age in this case). This must be a non-linear relationship since the probability must lie between 0 and 1 and a linear function would violate this condition at some point. We *linearise* this relationship by applying a suitable transformation. In the case in which we assume the underlying distribution is Normal this transformation is called the **Probit** transformation:

$$\text{Probit}\,(p) = \Phi^{-1}(p) = a + b\text{AGE}$$

Other transformations have been proposed. A popular choice is the **Logit** transformation, which has a relatively simple mathematical form:

$$\text{Logit}\,(p) = \ln\left(\frac{p}{1-p}\right) = a + b\text{AGE}$$

This corresponds to the underlying distribution being a *logistic* distribution which is very similar to the Normal distribution.

How do we estimate the parameters of this relationship? We need some method corresponding to the Least Squares method used for Normal regression models. The general concept is that of **Maximum Likelihood**. The data is of the form of **R** positive responses out of **N** *trials*. For each trial we assume there is a probability **p** of a positive response. The distribution of **R** is the Binomial distribution with parameters **N** and **p**. Thus for a particular choice of parameters **a** and **b** we can compute the corresponding **p** for each age group and hence the probability of obtaining the observed values of **R**. This is called the **Likelihood** of the data. The "best" choice of **a** and **b** is taken to be the values that make the Likelihood a maximum. For the Normal distribution this method results in the Least Squares method.

In general obtaining Maximum Likelihood estimates from a non-linear model requires a complicated numerical procedure which is *iterative*. That is, it goes through a process of refining the current estimates of the parameters until the maximum of the Likelihood is reached. This process is **not** guaranteed to work!

Hypothesis Testing

Having estimated the parameters of the model we would wish to test whether certain of these parameters might be zero. One approach is to attempt to find the sampling distribution of the parameter estimate. In most non-Normal models there are no exact results for the sampling distributions. However, we can use approximations. The approximate (asymptotic) distribution of the parameter estimates is Normal and furthermore we can find the approximate standard error of the estimate. Thus we can apply the usual t-test to the parameter estimate for a test of whether the parameter is zero.

$$\frac{\hat{b}}{s.e.(\hat{b})} \approx t$$

This approximation works best for large samples. For small samples the Normal approximation may not work well since the true sampling distributions are often skew. An alternative approach is to follow the *Analysis of Variance* method. The basis of this is to have a measure of **model fit** which measures the discrepancy between the model and the data. For Normal models this is the *Residual Sum of Squares*. Testing a parameter then is based on how much this measure of discrepancy is reduced when this parameter is introduced into the model. For non-Normal models this measure is called the **Deviance** and with count data is often called **Chi-squared Goodness of Fit**. In general the Deviance based on the value of the (maximised) Likelihood or a log transformation of it. The test is then based on the reduction is this measure of fit. A good approximation to the distribution of this reduction is the *Chi-squared distribution*. Thus we can test the significance of parameters using a Chi-Squared test. This approximation is more reliable than the Normal approximation described earlier. The two tests are not identical as in Normal models.

$$-2\log(Likelihood\ \ ratio) \approx c^{2}$$

If we use the Deviance definition of **Loss** in model fitting we can compare model by *change in Loss*. This is also referred to as the **Likelihood Ratio Test** (**LR**) as it is equivalent to comparing the models by the ratio of their maximised Likelihood values.

Consider binary data with y = 0 or 1, such that:

prob(condition true) = prob(y=1) = p

then the Likelihood for a single observation, y, is p if y=1 and (1-p) if y=0. The Deviance can be expressed conveniently in various ways such as:

$$-2\sum\left(y\log\ p+(1-y)\log(1-p)\right)$$

Interpreting Logistic Regression in SPSS

We have seen that the logit model is given by

$$\text{Logit}\ (p) = \ln\left(\frac{p}{1-p}\right) = a + b\text{AGE}$$

So, using SPSS, we are going to obtain values for coefficients a and b (-21.18 and 1.629). Replacing these into the equation, we obtain

$$\text{Logit}\ (p) = \ln\left(\frac{p}{1-p}\right) = -21.18 + 1.629\text{AGE}$$

In order to interpret this result, let us try to substitute a value for AGE, and let us try with AGE=10. That is, we would like to see how probable it is for a 10 year old to have already had her period.

Substituting, we obtain

$$\text{Logit}\ (p) = \ln\left(\frac{p}{1-p}\right) = -21.18 + 1.629(10) = -4.89$$

But this is NOT probability. To start having some important values, find $\exp^{(-4.89)} = 0.0075$. This is known as the odds value. This means that a change in a unit of AGE multiplies the odds of a girl already having her period by 0.0075. If you want to obtain the probability of a girl aged 10, that has already had her period, use the formula

$$\text{p} = \frac{\exp^{\text{Logit}\,(p)}}{1+\exp^{\text{Logit}\,(p)}} = \frac{\exp^{-4.89}}{1+\exp^{-4..89}} = \frac{0.0075}{1+0.0075} = 0.007$$

This is a very small probability, seeing that it is improbable that a girl less than 10 years of age has her period already.

Similarly, one can show that for a girl aged 18, the probability is 0.9997.

It is important to understand that the probability, the odds, and the logit are three different ways of expressing exactly the same thing.

Logistic Regression in SPSS

There are two ways of fitting Logistic Regression models in SPSS:

1. Regression / Probit

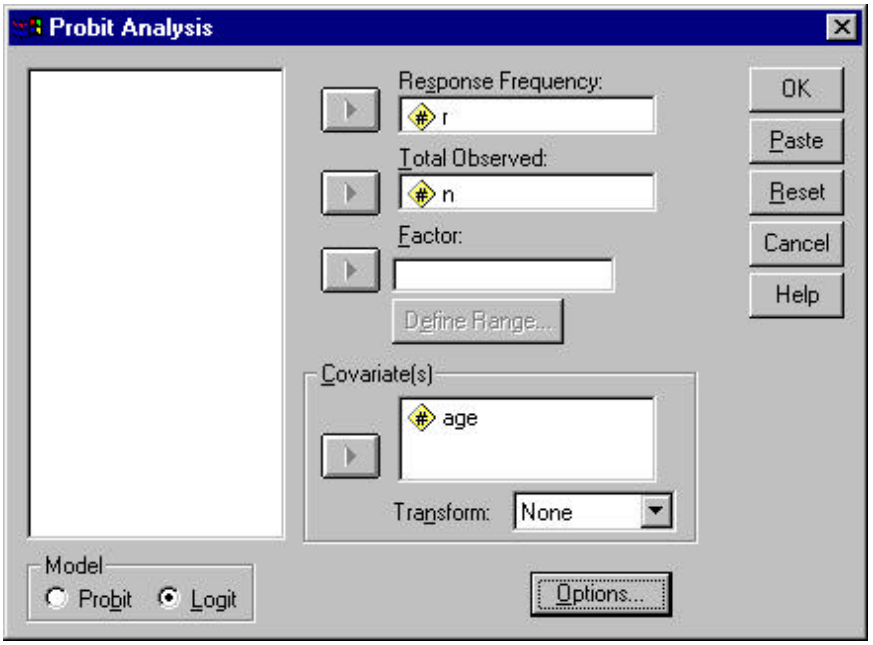
This is designed to fit Probit models but can be switched to Logit models.

The data is expected to be in the R out of N form, that is, each row corresponds to a group of N cases for which R satisfied some condition. The procedure is somewhat limited as it allows only one factor and does not allow the user to specify interactions explicitly (though it will conduct a test for parallelism). Also has a facility for estimating an extra parameter for "Natural Response Rate".

Doesn't allow predicted values to be saved and the output is limited.

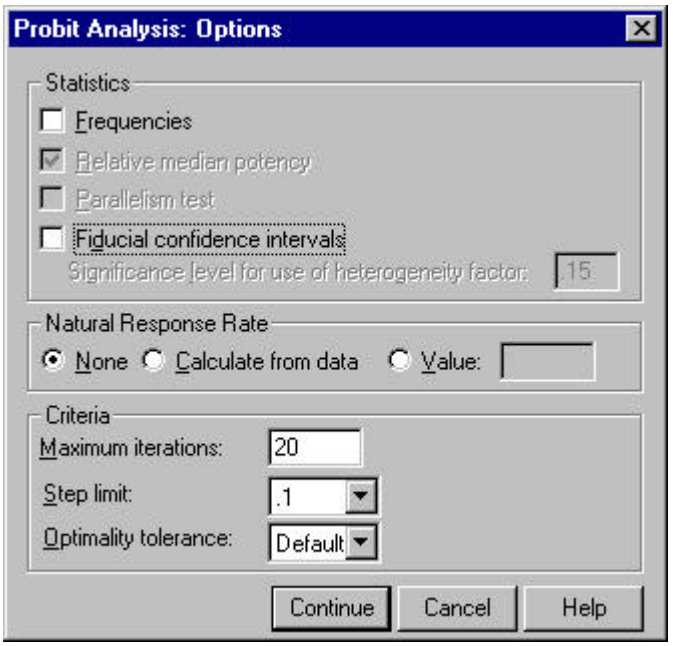
Will work with binary (0/1) data by defining a new variable for N with all values = 1. (Remember to switch OFF frequency output).

So, for the Menarche data, if we want to find the logit model, we press **Analyze**, **Regression**, **Probit**, to obtain the following:



Make sure you move *r* under the *Response Frequency*, and *n* under *Total Observed*. Move *age* under the Covariate box. Choose *Logit* model. Press *Options* to modify the output produced by SPSS.

Under Statistics, switch off *Frequncies* and *Fiducial confidence intervals*. (You can leave them on, however these will produce loads of lines in the output file, and if you need only the model, these will only confuse you more).

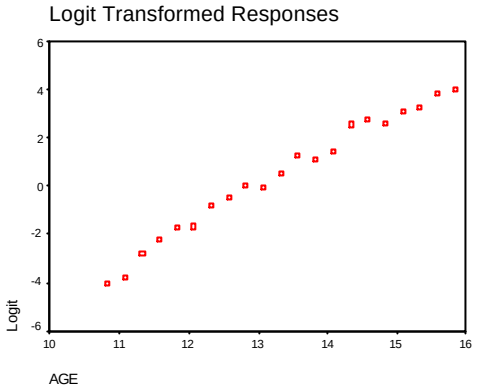


Press Continue, then Ok to obtain the following output.

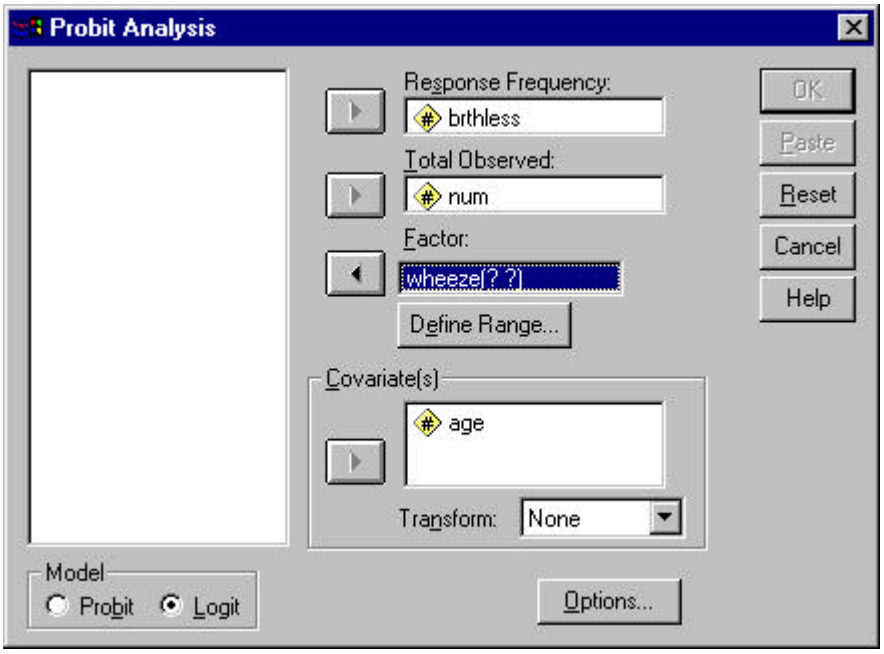
```
* * * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * * * * * * * * * *

 Parameter estimates converged after 17 iterations.
 Optimal solution found.

 Parameter Estimates (LOGIT model:  (LOG(p/(1-p))) = Intercept + BX):

           Regression Coeff.  Standard Error     Coeff./S.E.

   AGE              1.62963          .05885        27.69354


                  Intercept  Standard Error  Intercept/S.E.

                  -21.18218          .76891       -27.54824

  Pearson  Goodness-of-Fit  Chi Square =     23.606    DF = 23   P =  .426

  Since Goodness-of-Fit Chi square is NOT significant, no heterogeneity
  factor is used in the calculation of confidence limits.
```

The Goodness-of-Fit Chi Square, is the log likelihood multiplied by –2. Because the log-likelihood is negative, the Goodness-of-Fit Chi Square is positive, and larger values indicate worse prediction of the dependent variable. Therefore we are after a non-significant p value (as in this case).

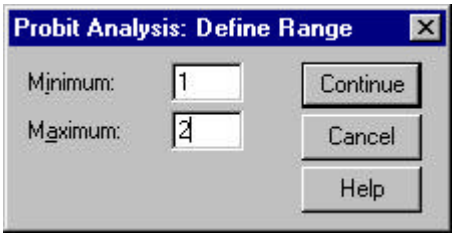
SPSS also outputs a graph of the predicted variable against the covariate age:



Open the file miners.sav. This file contains data about women and whether they have children or not. We are also considering a factor, *wheeze*. To try to find a logit model for *brthless*, using *age* as *covariate* and *wheeze* as *factor*.



Move *brthless* as the *Response Frequency* and *num* as the *Total Observed*. Move *age* as the *Covariate* and *wheeze* as *Factor*. You notice that you have to define the range for *wheeze* as



Click the Options button, and choose the Parallelism test. This will give an indication of the difference between the model fitting due to the factor.

The following output was obtained

```
* * * * * * * * * * * * *  P R O B I T   A N A L Y S I S  * * * * * * * * * * *
Group Information

    WHEEZE    Level  N of Cases    Label
                  1           9        1
                  2           9        2


MODEL Information

      ONLY Logistic Model is requested.


* * * * * * * * * * * * *  P R O B I T   A N A L Y S I S  * * * * * * * * * * *
 Parameter estimates converged after 20 iterations.
 Optimal solution found.

 Parameter Estimates (LOGIT model:  (LOG(p/(1-p))) = Intercept + BX):

           Regression Coeff.  Standard Error     Coeff./S.E.

   AGE                .08667          .00285        30.38923


                   Intercept  Standard Error  Intercept/S.E.  WHEEZE

                    -4.16258          .14283       -29.14388         1
                    -7.00056          .14620       -47.88353         2

  Pearson  Goodness-of-Fit  Chi Square =     35.087    DF = 15   P =  .002
           Parallelism Test Chi Square =     19.458    DF = 1   P =  .000

  Since Goodness-of-Fit Chi square is significant, a heterogeneity
  factor is used in the calculation of confidence limits.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

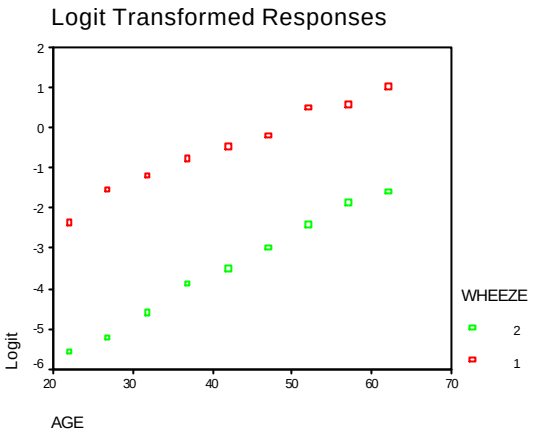
This means that the output gave 2 logits, namely:

$$\text{Logit}\,(p) = \ln\left(\frac{p}{1-p}\right) = -4.16258 + 0.08667(\text{AGE}) \text{ and}$$

$$\text{Logit}\,(p) = \ln\left(\frac{p}{1-p}\right) = -7.00056 + 0.08667(\text{AGE})$$

The first one should be used when the factor *wheeze* is equal to 1, while the second when the factor *wheeze* is equal to 2. Interpretation of the logit is as in the previous example.

The goodness of fit test shows a bad prediction of the dependent variable. The graph is given by



2. Regression / Binary Logistic

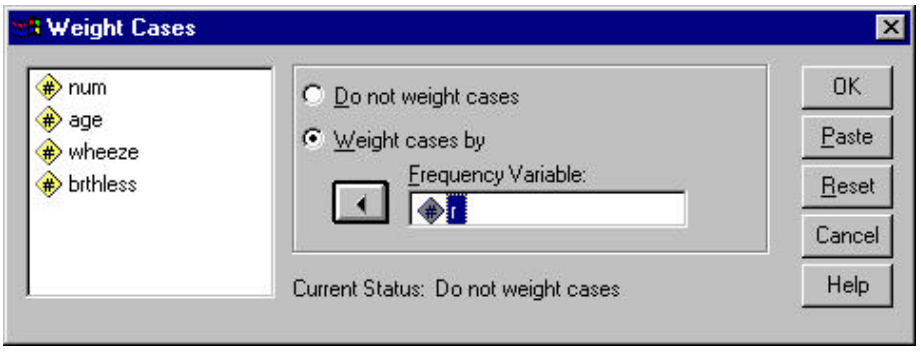
Expects the dependent variable to be binary (any values).

If the data is in the form of R out of N it has to be manipulated into the alternative form. This can be done in the SPSS data editor window (using SPSS Compute and Cut and Paste), but may be easier in some other software such as a spreadsheet. The data will end up with twice as many rows as the values of N-R (not R) must be appended to the R values in a single column. A binary variable has to be set up to distinguish the trues from the falses and all explanatory variables must be duplicated. The variable holding the R and N-R values is then declared as a **weight** variable through **Data / Weight Cases**. Compare the 2 SPSS data files, *miners.sav* and *miners2.sav*.

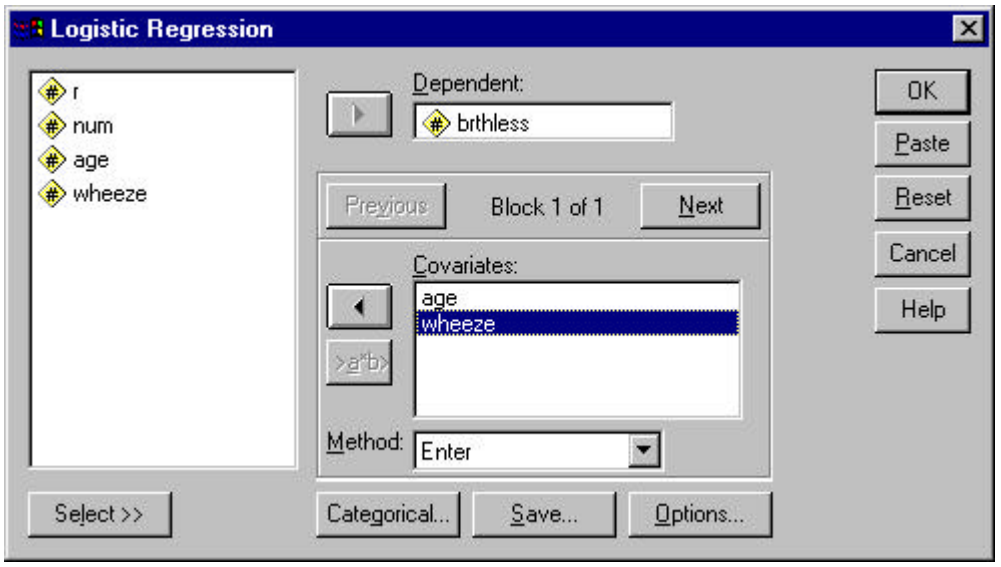
Model specification is more flexible as we can have a mixture of *Factors* and *Covariates* and *Interactions*. However, the method of definition is different from General Linear Models. To include a *Factor* first select the explanatory variable then declare it as *Categorical* by opening the *Categorical box* and selecting it again. Interactions are included by selecting more than one variable simultaneously (Ctrl and click) and entering them via the [>a*b>] button. Stepwise procedures are available (Forward/Backward). **Blocks** of terms are also available, so that model comparison is made easier.

Predicted values (probabilities) can be saved as well as diagnostics. Only the Logit transformation is allowed, cannot switch to Probit.

So in order to test the miners data, remember first to weight by the variable *r*.

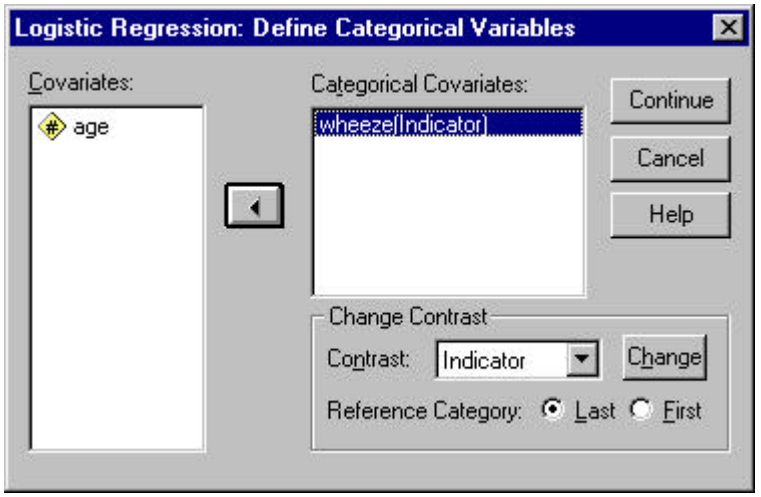


Then, press *Analyze*, *Regression*, *Binary Logistic*.



Move *brthless* as the dependent variable, *age* and *wheeze* as the covariates. You can choose from the different methods of inclusion in the model, by pressing the *Method* drop down list.

In order to change the variable *wheeze* into a categorical variable, press the *Categorical* button. Move *wheeze* under *Categorical Covariates*.



Press Continue, then OK. Some of the output produced by SPSS follows.

Dependent Variable Encoding

Original Value	Internal Value
.00	0
1.00	1

SPSS gives the way it had encoded the binary variable *brthless*. In this case, the original coding was left, but if the variable had been coded as 3 and 4 (e.g.), these would have been recoded to 0 and 1. So in our case, 0 means having a child, and 1 means not having a child.

Categorical Variables Codings

		Frequency	Paramete (1)
WHEEZE	1.00	18	1.000
	2.00	18	.000

SPSS also gives us information of how the factor *wheeze* was recoded. This will be useful when we write down the equation. Note that SPSS works out $\exp^B$, or the odds value.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1.877	.022	7414.155	1	.000	.153

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	AGE	2134.077	1	.000
		WHEEZE(1)	5336.834	1	.000
	Overall Statistics		6039.885	2	.000

Since we chose the Method *Enter*, SPSS starts by insert only a constant in the model. In fact, *age* and *wheeze* are still out of the model.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	5358.839	2	.000
	Block	5358.839	2	.000
	Model	5358.839	2	.000

On Step 1, SPSS enters all the variables in the model. The coefficients here gives us a measure of how well the model fits. We must look mostly at the Model coefficient. It is analogous to the multivariate *F* test for linear regression. The null hypothesis states that information about the independent variables does not allow us to make better prediction of the dependent variable. Therefore we would want that this chi-squared value is significant (as in this example).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.087	.003	923.365	1	.000	1.091
	WHEEZE(1)	2.838	.056	2588.087	1	.000	17.078
	Constant	-7.000	.146	2292.811	1	.000	.001

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

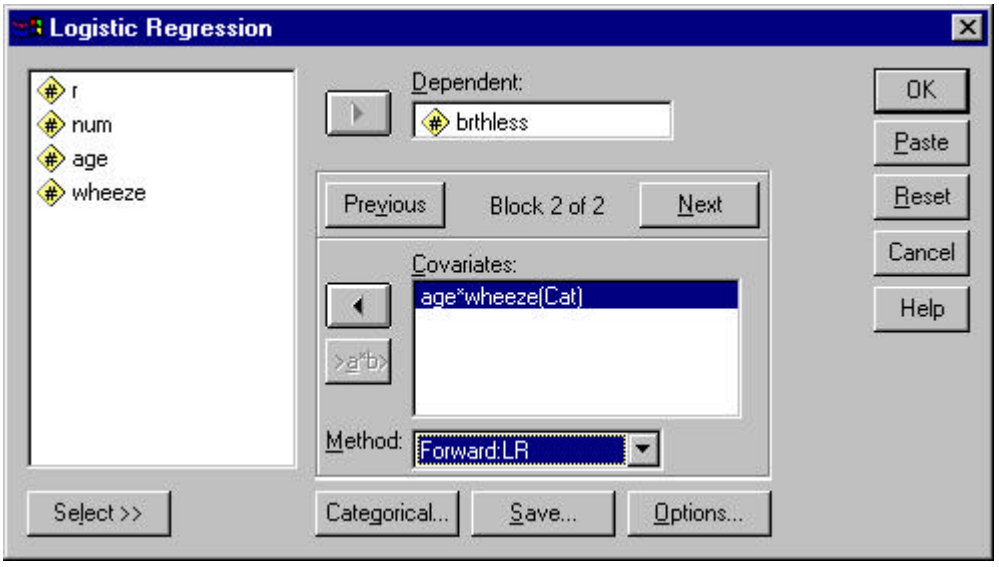
The last table produced by SPSS is the one containing the variable coefficients. The formula should read

$$\text{Logit}\,(p) = \ln\left(\frac{p}{1-p}\right) = -7.000 + 0.08667(\text{AGE}) + 2.838(\text{WHEEZE}(1))$$

Remember that wheeze was recoded to take 0 and 1 as values, and therefore if we substitute 0 for *wheeze(1)*, we obtain the 2nd equation of page 9, and if we substitute 1, we obtain the 1st equation of page 9.

Suppose we wanted to look at the interaction age*wheeze. However we would like to first fit the model without interaction, then add the interaction by using the Forward Stepwise (Likelihood Ratio).

Press *Analyze*, *Regression*, and *Binary Logistic* as before. However, this time press *Next*. Choose *age* and *wheeze* together by using the CTRL button. Press the **>a*b>**, and choose *Forward:LR* as the *method*. Press *OK*.



The first part of the output is identical to the previous example. What differs is when the interaction comes in.

Block 2: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	19.540	1	.000
	Block	19.540	1	.000
	Model	5378.378	3	.000

The model is still highly significant, showing that the independent variables predict the dependent variable well.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.101	.004	515.318	1	.000	1.106
	WHEEZE(1)	4.089	.292	195.870	1	.000	59.697
	AGE by WHEEZE(1)	-.025	.006	19.312	1	.000	.975
	Constant	-7.714	.228	1147.484	1	.000	.000

a. Variable(s) entered on step 1: AGE * WHEEZE .

The equation is given by

$$\text{Logit}\,(p) = -7.714 + 0.101(\text{AGE}) + 4.089(\text{WHEEZE}(1)) - 0.025(\text{AGE} * \text{WHEEZE}(1))$$

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	AGE * WHEEZE	-4479.588	19.540	1	.000

The last piece of output tells us what would happen to the model if the interaction term is removed. As one can see, the –2log likelihood increases significantly, showing a worse fit. Therefore the interaction term improves the overall fit.

To find the probability that a woman aged 25, and having a wheeze condition of 1, has no children, first substitute in equation and find the exp to obtain the odds.

$$\text{Logit}\,(p) = -7.714 + 0.101(25) + 4.089(1) - 0.025(25*1) = -1.725$$

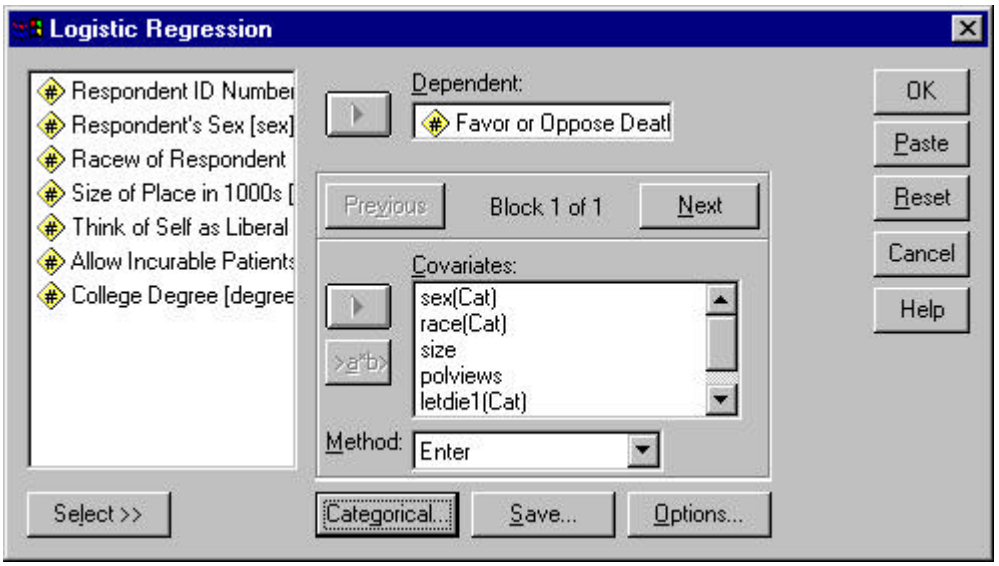
$$\exp^{-1.725} = 0.0188\,.$$

Then transform to obtain the probability $\dfrac{\exp^{-1.725}}{1+\exp^{-1.725}} = 0.0185\,.$

Example

This example uses the *gss93t.sav* data file. This is a subset of the original gss92 data file. The analysis uses as a dependent the attitude variable *cappun*, which is coded '1= favor the death penalty', '2=oppose the death penalty'. The independent variables are *age*, *degree2* (college degree or not), *race*, *sex*, *letdie1* (if would allow the incurable to die), *size* (of place), and *polviews* (liberalism-conservatism).

Press *Analyze*, *Regression* and *Binary Logistic*. Fill in the form as shown.



Remember to choose the variables *degree2*, *race*, *sex* and *letdie1* as categorical variables. Also note that in this case, no weighting is necessary, as each row corresponds to one frequency.

Press Ok.

Dependent Variable Encoding

Original Value	Internal Value
Favor	0
Oppose	1

SPSS lets us know below that it recodes *cappun*, the dependent variable, from 1, 2 coding to 0, 1. So we will find the probability of opposing the death penalty.

Categorical Variables Codings

			Parameter coding	
		Frequency	(1)	(2)
Racew of Respondent	white	355	1.000	.000
	black	42	.000	1.000
	other	26	.000	.000
College Degree	No College degree	321	1.000	
	College degree	102	.000	
Allow Incurable Patients to Die	Yes	278	1.000	
	No	145	.000	
Respondent's Sex	Male	187	1.000	
	Female	236	.000	

Above is SPSS's parameterisation of the categorical independent variables. Note as before that the last category of each variable is ommitted.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	38.718	7	.000
	Block	38.718	7	.000
	Model	38.718	7	.000

The above shows that the overall model predicts the dependent variable.

The list of the coefficients is given by

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SEX(1)	-.525	.263	4.000	1	.046	.591
	RACE			8.281	2	.016	
	RACE(1)	-.533	.471	1.283	1	.257	.587
	RACE(2)	.487	.570	.731	1	.392	1.628
	SIZE	.000	.000	2.974	1	.085	1.000
	POLVIEWS	-.183	.096	3.652	1	.056	.833
	LETDIE1(1)	-.798	.263	9.182	1	.002	.450
	DEGREE2(1)	-.718	.283	6.444	1	.011	.488
	Constant	.948	.673	1.985	1	.159	2.581

a. Variable(s) entered on step 1: SEX, RACE, SIZE, POLVIEWS, LETDIE1, DEGREE2.

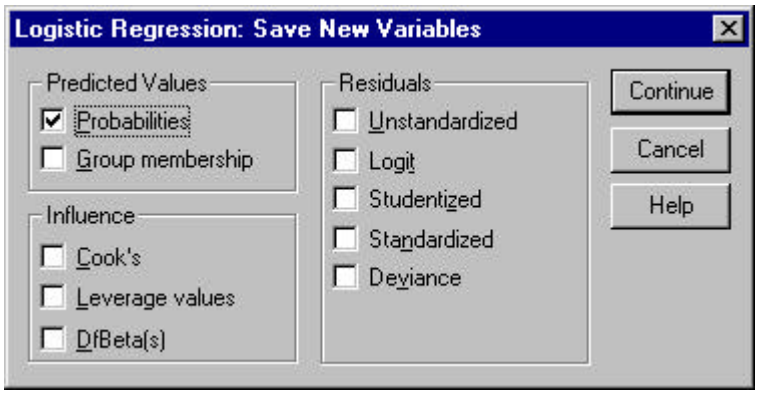
Suppose we wanted to find the probability of opposing the capital sentence, of a white female with a college degree, with a size of place of 400, having a polview of 2 and in favour of letting the incurable die.

$$\text{Logit}\,(p) = 0.948 - 0.533(1) + 0(400) - 0.183(2) - 0.798(1) = -0.749$$

$$\exp^{-0.749} = 0.4728 .$$

Then transform to obtain the probability $\frac{\exp^{-0.749}}{1+\exp^{-0.749}} = 0.32$.

Suppose we just wanted to see how a model with *race*, *age* and their interaction predicts the dependant variable. This time save the predicted values, by pressing *Save* and choosing *Probabilities*.



The model should first enter the variables *age* and *race*, and then their interaction using the method *enter* in both cases.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	36.157	3	.000
	Block	36.157	3	.000
	Model	36.157	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1440.612	.026	.039

This is obtained in the first step, when the 2 variables are entered into the equation. Note the large value for the –2log likelihood.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	-.001	.004	.028	1	.866	.999
	RACE			37.615	2	.000	
	RACE(1)	-.935	.262	12.757	1	.000	.393
	RACE(2)	.049	.301	.027	1	.870	1.051
	Constant	-.454	.292	2.422	1	.120	.635

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

When the interaction was inputted, the following output was obtained:

Block 2: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1.077	2	.583
	Block	1.077	2	.583
	Model	37.235	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1439.534	.027	.040

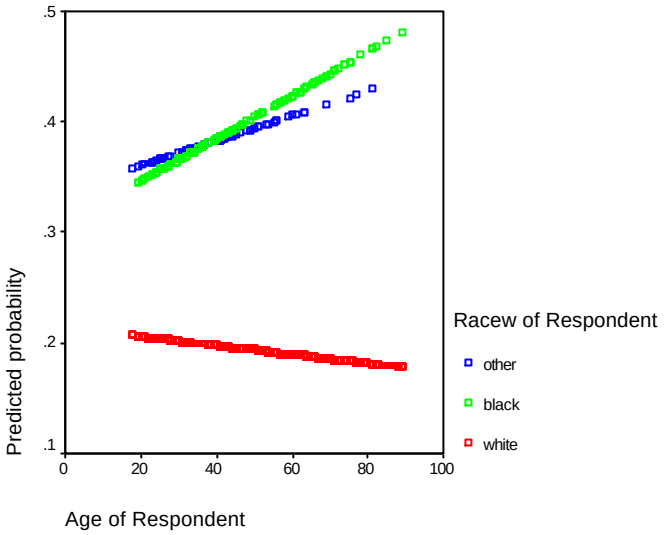
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.005	.017	.074	1	.786	1.005
	RACE			1.469	2	.480	
	RACE(1)	-.632	.763	.686	1	.407	.532
	RACE(2)	-.127	.873	.021	1	.884	.881
	AGE * RACE			1.077	2	.584	
	AGE by RACE(1)	-.007	.018	.167	1	.683	.993
	AGE by RACE(2)	.003	.020	.029	1	.865	1.003
	Constant	-.667	.733	.827	1	.363	.513

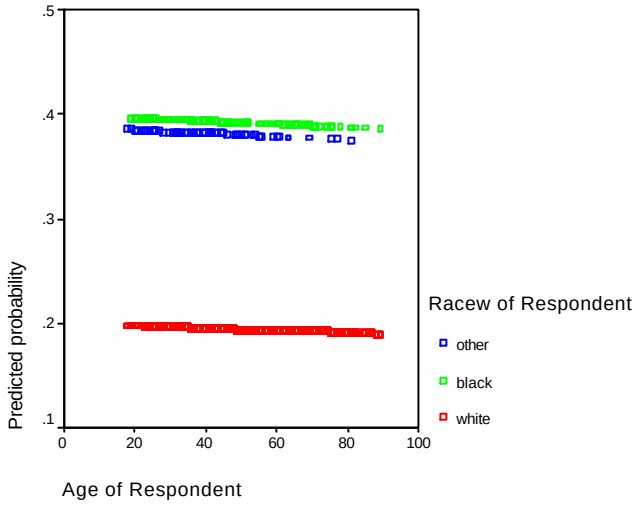
a. Variable(s) entered on step 1: AGE * RACE .

This shows a not significant change in the model, so one should think of removing the interaction between the two variables.

A graph of the predicted probabilities with *age*, using race as *factor* follows.



If the interaction term was removed, the scatter plot is



which has a similar shape for the 3 different races.

Practical Session 6: Logistic Regression Models

1. Repeat the analysis in the lecture.

Using the GSS data (spsswin\data\gss93t.sav)

i. Restrict your analysis to women only.
ii. Fit logistic regression models using **AGE** as covariate and **RACE** as a factor.
iii. Taking **POLVIEWS** as a covariate examine any effects and interactions with other variables.

iv. Insert the other variables (except id), and be careful to clearly mark the factors as categorical variables,

2. Using the data in Q1 include the men into the analysis with **SEX** as a factor. What is the interpretation of this model? What is the probability that a black female with no college degree, with a size of place of 100, having a polview of 3, in favour of letting the incurable die, and aged 25 opposes the capital sentence?

3. Using the voters2 data (voters2.sav) to examine the relationships of voting labour (vote) and other variables **SEX**, **AGE** and any other that you think relevant.