0.1 Exercise Data Set

The exercise data set comes from a survey of home owners conducted by an electricity company about an offer of roof solar panels with a 50% subsidy from the state government as part of the states environmental policy. The variables involve household income measured in units of a thousand dollars, age, monthly mortgage, size of family household, and as the dependent variable, whether the householder would take or decline the offer. The purpose of the exercise is to conduct a logistic regression to determine whether family size and monthly mortgage will predict taking or declining the offer.

For the first demonstration, we will use 'family size and 'mortgage only. For the options, select Classification Plots, Hosmer-Lemeshow Goodness Of Fit, Casewise Listing Of Residuals and select Outliers Outside 2sd. Retain default entries for probability of stepwise, classification cutoff and maximum iterations.

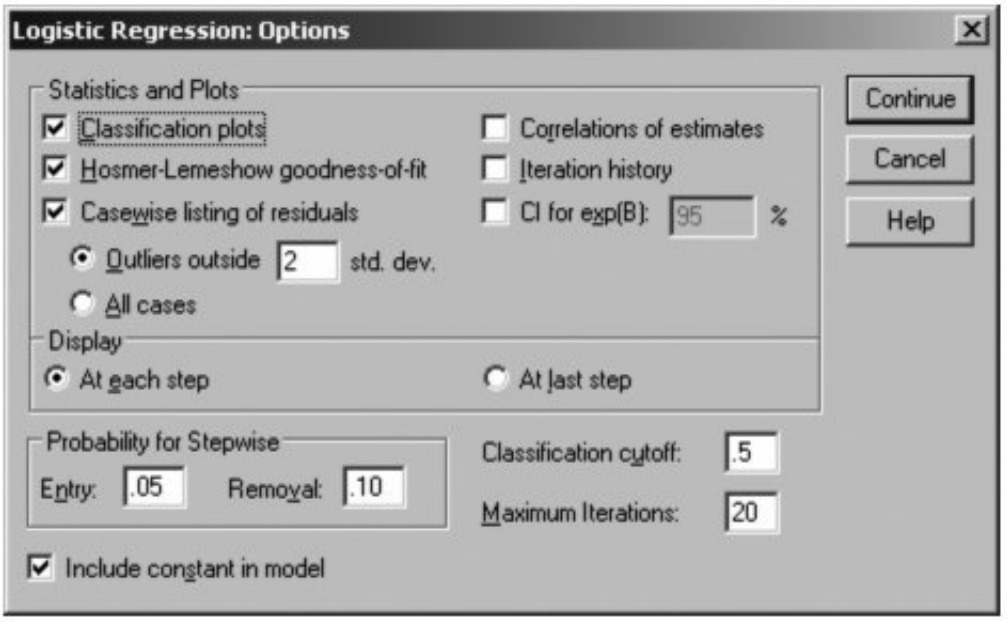

Figure 1: Selected Options for Exercises

We are not using any categorical variables this time. If there are categorical variables, use the ***categorical*** option. For most situations, choose the indicator coding scheme (it is the default).

0.2 Hosmer-Lemeshow Prostate Example

We will now consider a real life example to demonstrate Logistic Regression. This example is taken from a Prostate Cancer Study from Hosmer and Lemeshow (2000). The goal of the analysis is to determine if variables measured at baseline can predict whether a tumour has penetrated the prostatic capsule. The variables are as follows:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.134	.366	.133	1	.715	1.143

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Mortgage	6.520	1	.011
		Famsize	14.632	1	.000
	Overall Statistics		15.085	2	.001

Figure 2: Variables in / not in the equation

Variables from the Dataset Prostate (Hosmer and Lemeshow, 2000):

Variable	Label	Values
ID	Patient ID	1 – 380
Capsule	Tumor Penetration of Prostatic Capsule	0 = No Penetration, 1 = Penetration
Age	Age in Years	Number
Race	Race of Patient	1 = White, 2 = Black
Dpros	Results of the Digital Rectal Exam	1 = No Nodule, 2 = Left Lobe, 3 = Right Lobe, 4 = Both Lobes
Dcaps	Detection of Capsular Involvement	1 = No, 2 = Yes
PSA	Prostatic Specific Antigen Value	mg / ml
Vol	Tumor Volume Obtained from US	cm3
Gleason	Total Gleason Score	2 - 10

Figure 3: Variables

0.3 HSB2 Example

The hsb2 dataset is taken from a national survey of high school seniors. Two hundred observation were randomly sampled from the High School and Beyond survey. Descriptive statistics and exploratory data analysis are shown below. Because we do not have a suitable dichotomous variable to use as our dependent variable, we will create one (which we will call honcomp, for honors composition) based on the continuous variable write. We do not advocate making dichotomous variables out of continuous variables; rather, we do this here only for purposes of this illustration.

Here is the list of variables in the file.

```
obs:           200    highschool and beyond (200 cases)
vars:           12    28 Feb 2005 09:25
-------------------------------------------------------------------------------
variable
variable name   type   about the variable
```

```
-------------------------------------------------------------------------------
id              scale  student id
female        nominal  (0/1)
race          nominal  ethnicity (1=hispanic 2=asian 3=african-amer 4=white)
ses           ordinal  (1=low 2=middle 3=high)
schtyp        nominal  type of school (1=public 2=private)
prog          nominal  type of program (1=general 2=academic 3=vocational)
read            scale  standardized reading score
write           scale  standardized writing score
math            scale  standardized math score
science         scale  standardized science score
socst           scale  standardized social studies score
hon           nominal  honors english (0/1)
```

0.4 Hosmer-Lemeshow Prostate Example

We will now consider a real life example to demonstrate Logistic Regression. This example is taken from a Prostate Cancer Study from Hosmer and Lemeshow (2000). The goal of the analysis is to determine if variables measured at baseline can predict whether a tumour has penetrated the prostatic capsule. The variables are as follows:

Variables from the Dataset Prostate (Hosmer and Lemeshow, 2000):

Variable	Label	Values
ID	Patient ID	1 – 380
Capsule	Tumor Penetration of Prostatic Capsule	0 = No Penetration, 1 = Penetration
Age	Age in Years	Number
Race	Race of Patient	1 = White, 2 = Black
Dpros	Results of the Digital Rectal Exam	1 = No Nodule, 2 = Left Lobe, 3 = Right Lobe, 4 = Both Lobes
Dcaps	Detection of Capsular Involvement	1 = No, 2 = Yes
PSA	Prostatic Specific Antigen Value	mg / ml
Vol	Tumor Volume Obtained from US	cm3
Gleason	Total Gleason Score	2 - 10

Figure 4: Variables

0.5 Kasser and Bruce Infarction Data Example

We use a set of coronary data (Kasser and Bruce, 1969; Kronmal and Tarter, 1974) to see if age, history of angina pectoris (ANGINA: yes, no), history of high blood pressure (HIGHBP: yes, no), and functional class (FUNCTION: none, minimal, moderate, and more than moderate) can be used to predict the probability of past myocardial infarction (INFARCT: yes, no).