

# Medical Statistics – Lab 7

SPSS version

## Part 1: Risk of in-hospital death in patients with acute myocardial infarction

In part 1 of the lab, we are going to analyze the risk of in-hospital death in patients hospitalized because of acute myocardial infarction. The dataset comes from the Worcester Heart Attack Study (WHAS) and includes data from 500 patients admitted in Worcester, Massachusetts in 1997, 1999, and 2001 (file `whas500.sav` from the Datasets menu). The outcome of interest is in-hospital death, measured by the variable “discharge status from hospital” (`dstat`) with values `alive` and `death`.

### Exploratory data analysis

To explore whether gender has an effect on the risk of in-hospital death, we start by creating a contingency table and use the table to calculate the proportion in-hospital death in the two gender subgroups:

Go to **Analyze -> Descriptive Statistics -> Crosstabs**. Select the `gender` variable as the Row(s) and the `stat` variable as the Column(s). Click on **Cells** and under percentages check the box **Row**. Click on **Continue** and **OK** to create the table.

#### Question 1

Based on the group proportions, do you expect gender to have an effect on the risk of in-hospital death?

### Simple logistic regression

To determine whether gender is significantly associated with in-hospital death, we can conduct several statistical tests. As a recap of lab 4, we start by performing a chi-square test of homogeneity.

### Question 2

Perform the chi-square test of homogeneity (see instructions in lab 4 if needed). What conclusion can be drawn from the test?

Another option is to perform logistic regression. To perform this analysis in SPSS, go to **Analyze -> Regression -> Binary Logistic**. Move the **dstat** variable to the **Dependent** box and the **gender** variable to the **Block 1 of 1** box. Press **Categorical**. Specify that the **gender** variable is categorical by moving it from the **Covariates** box to the **Categorical Covariates** box. Press **Continue** to return to the main dialog. Press **Options** and check the box **CI for exp(B)** to include confidence intervals for the estimated hazard ratios. Press **Continue** to return to the main dialog and click on **OK** to run the analysis.

### Question 3

What is the odds ratio for in-hospital death for females compared to males? How should this odds ratio be interpreted in the context of the study?

### Question 4

Based on the estimated regression coefficients (ignoring p-values), what are the predicted proportions of in-hospital deaths for male and female patients? Compare the predicted proportions to the observed proportions from the previously constructed contingency table. Do they match?

### Question 5

What conclusion can be drawn from the logistic regression analysis regarding the association between gender and in-hospital death? Is this in line with the conclusion drawn from the chi-square test?

## Multiple logistic regression

To assess the extent to which the effect of gender is confounded by age, we will fit a multiple regression model with in-hospital death as the dependent variable and gender and age as the independent variables:

Go to **Analyze -> Regression -> Binary Logistic**. Move the **dstat** variable to the **Dependent** box and the **gender** and **age** variables to the **Block 1 of 1** box. Press **Categorical**. Specify that the **gender** variable is categorical by moving it from the **Covariates** box to the **Categorical Covariates** box. Press **Continue** to return to the main dialog. Press **Options** and check the box **CI for exp(B)** to include confidence intervals for

the estimated hazard ratios. Press **Continue** to return to the main dialog and click on **OK** to run the analysis.

#### Question 6

How does adjusting for age affect the estimated odds ratio for in-hospital death for females compared to males?

#### Question 7

Calculate the odds ratio for in-hospital death corresponding to a 10-year increase in age and interpret its meaning.

### Likelihood ratio tests

As explained in the syllabus, the p-values in the table of estimated regression coefficients are derived from Wald tests, which test the null hypothesis that the corresponding regression coefficient is equal to 0

Instead of the Wald tests, we can also obtain p-values using likelihood ratio tests, which compare the goodness of fit of the full model (including the predictor of interest) to a reduced model (excluding the predictor) to test the null hypothesis that the predictor has no effect on the outcome. This approach is particularly useful for testing predictors with non-linear or complex effects, as it evaluates their contribution to the model as a whole. Examples include categorical variables with three or more categories (requiring the creation of multiple dummy variables) and relationships modeled using multiple terms, such as including both a linear and a quadratic term to capture a quadratic relationship.

For example, we can use a likelihood ratio test (LRT) to compare a full model including both sex and age as predictors to a reduced model including only age. To perform this test in SPSS, go to **Analyze -> Regression -> Binary Logistic**. Press **reset** to clear the previous analysis. Move the **dstat** variable to the **Dependent** box and the **age** variables to the **Block 1 of 1** box. Press **Next** to create a second block. Move the **gender** variable to the **Block 2 of 2** box. Press **Categorical**. Specify that the **gender** variable is categorical by moving it from the **Covariates** box to the **Categorical Covariates** box. Press **Continue** to return to the main dialog. Press **Options** and check the box **CI for exp(B)** to include confidence intervals for the estimated hazard ratios. Press **Continue** to return to the main dialog and click on **OK** to run the analysis.

The output will include a table with the results of the likelihood ratio test comparing the full model (variables included in block 1 and 2) to the reduced model (variables included in block 1 only). This table can be found in the section **Block 2: Method = Enter** and is labeled

Omnibus tests of model coefficients. The p-value for the LRT comparing the full model to the reduced model can be found in the row block.

#### Question 8

How does the p-value from the likelihood ratio test compare to the one from the Wald test?

### Evaluating model fit

One way to examine the fit of the logistic regression model is the Hosmer-Lemeshow Goodness of Fit test. In SPSS, this can be achieved by checking the option **Hosmer-Lemeshow goodness-of-fit** in the logistic regression dialog box.

Go to **Analyze -> Regression -> Binary Logistic**. Press **reset** to clear the previous analysis. Move the **dstat** variable to the **Dependent** box and the **gender** and **age** variables to the **Block 1 of 1** box. Press **Categorical**. Specify that the **gender** variable is categorical by moving it from the **Covariates** box to the **Categorical Covariates** box. Press **Continue** to return to the main dialog. Press **Options** and check the box **Hosmer-Lemeshow goodness-of-fit** to include the results of the Hosmer-Lemeshow test in the output produced. Press **Continue** to return to the main dialog and click on **OK** to perform the analysis.

#### Question 9

Based on the results of the Hosmer-Lemeshow goodness-of-fit test, does our model provide a satisfactory fit to the data?

## Part 2: unguided exercises

### Exercise 1

Multiple logistic regression was used to construct a prognostic index to predict coronary artery disease from data on 348 patients with valvular heart disease who had undergone routine coronary arteriography before valve replacement (Ramsdale et al. 1982). The estimated equation was:

$$\text{logit}(p) = \ln(p/(1-p)) = b_0 + 1.167 \times x_1 + 0.0106 \times x_2 + \text{other terms}$$

where  $x_1$  stands for the family history of ischaemic disease (0=no, 1=yes) and  $x_2$  is the estimated total number of cigarettes ever smoked in terms of thousand cigarettes, calculated as the average number smoked annually times the number of years smoking.

- (a) What is the estimated odds ratio for having coronary artery disease for subjects with a positive family history relative to subjects with a negative family history?
- (b) What total number of cigarettes ever smoked carries the same risk as a positive family history? Convert the result into years of smoking 20 cigarettes per day.
- (c) What is the odds ratio for coronary artery disease for someone with a positive family history who had smoked 20 cigarettes a day for 30 years compared to a non smoker with no family history?

## Exercise 2

Data from 37 patients receiving a non-depleted allogenic bone marrow transplant were examined to see which variables were associated with the occurrence of acute graft-versus-host disease (GvHD: 0=no, 1=yes) (Bagot et al., 1988). Possible predictors are TYPE (type of leukemia: 1=AML, acute myeloid leukaemia; 2=ALL, acute lymphocytic leukaemia; 3=CML, chronic myeloid leukemia), PREG (donor pregnancy: 0= no, 1=yes), and LOGIND (the logarithm of an index of mixed epidermal cell-lymphocyte reactions). The data are in the file `GvHD.sav` available from the Downloads menu.

- (a) Perform a likelihood ratio test to determine whether there is a significant association between the type of leukemia and the occurrence of GvHD after adjusting for donor pregnancy and the logarithm of an index of mixed epidermal cell-lymphocyte reactions.
- (b) In the adjusted model, What is the estimated odds ratio for the occurrence of GvHD for patients with ALL compared to those with ALM?
- (c) Use the Hosmer-Lemeshow goodness-of-fit test to evaluate the fit of the model. Based on the results, does the model provide a satisfactory fit to the data?