Project 2: Logistic Regression

MASM22/FMSN30/FMSN40: Linear and Logistic Regression (with Data Gathering), 2022

Peer assessment version: **12.30 on Wednesday 11 May**

Peer assessment comments: **13.00 on Thursday 12 May**

Final version: **17.00 on Friday 13 May**

Cardiovascular diseases in Sweden

Introduction

Diseases of the cardiovascular system constitute a large part of the health care costs in Sweden, and elsewhere. With an increasingly older population these costs can be expected to rise further. The Health economical database HILDA (Health and Individuals. Longitudinal Data and Analysis) is a register database based on Statistics Sweden?s ULF studies (Undersökningar om Levnadsförhållanden) with additional information on, among other things, hospital treatments. We are going to use a manopulated version of this database, that is, the structure and general behaviour of the data is the same as in the true database but the data itself has been manipulated in order to protect personal integrity.

The respondents were interviewed in either 1988 or 1989. Those who were alive on 31 December 1989 were then followed from 1 January 1990 until 31 December 2000 and their number of days in hospital during that period was retrieved from the central registry, *Slutenvårdsregistret*. Those who died during this follow-up period are still part of the data.

Our goal is to model the probability of having at least one day in hospital with a cardiovascular diagnosis during the follow-up period 1990–2000, based on the answer in the interview and the age at the start of the period.

Data file

The semicolon separated datafile `hospital.txt` contains 5921 observations on 10 variables and can be downloaded from the course home page. Save it to your R data directory and then read it into R with

```
mydata <- read.delim("Data/hospital.txt")
```

Name	Description
id	Number
sex	Sex (1 = Male, 2 = Female).
age	Age in 1990 (years)
health	General state of health (1 = good, 2 = bad, 3 = somewhere in between)
civilst	Civil status (1 = unmarried, 2 = married, 3 = divorsed/separated, 4 = widow/widower)
exercise	Exercise habits (0 = practically no exercise at all, 1 = exercises sometimes, 2 = exercises regularly approximately once a week, 3 = exercises regularly approximately twice a week, 4 = exercises regularly rather strenously at least twice a week)
work_norm	Normal hours of work per week (1 = employed 1-19 hours, 2 = employed 20–34 hours, 3 = employed 35–97 hours, 4 = farmer or self-employed, 5 = other, does not work)
inc_hh	Household disposable income (100 SEK)
inc_tot	Individual income from work, capital and pension (100 SEK)
hosp	Number of days in hospital 1/1 1990–31/12 2000 for diagnoses[1] ICD9: 413 Angina pectoris, 410 Acute myocardic infarction, 428 Heart failure, or ICD10: I20 Angina pectoris, I21 Acute myocardic infarction, I50 Heart failure, were the submission and/or release date falls during the period 1990–2000. (0 = no days, 1 = at least one day)

In order to make it easier to colour code the observations according to the value of the response variable, hosp = 0 or 1, it will be convenient to also create a separate factor version of this variable:

```
mydata$hosp_cat <- factor(mydata$hosp,
  levels = c(0, 1),
  labels = c("0 days", "1+ days"))
```

You can then plot with, e.g., `aes(..., y = hosp)` and get the values 0 and 1 on the y-axis, while `aes(..., color = hosp_cat)` will give you different colours for “0 days” and “1+ days”. You can use either version as dependent variable in the glm-function since it will treat the last category (1 or “1+ days”) as “success”.

Part 1. Introduction to logistic regression

(a). Start by examining the relationship between hospital days and general health status by a simple cross-tabulation:

```
table(hospital$health, hospital$hosp)
```

Also calculate the proportions that have at least one day in hospital, for each of the health categories:

```
prop.table(table(hospital$health, hospital$hosp), margin = 1)
```

Does is seem like the general health status has an impact on the probability of having at least one day in hospital?

Then fit a logistic regression modeling the probability of having at least one day in hospital, as a function of health status, using "good" health as reference category.

Present a table with the three β-estimates, the corresponding odds ratios comparing the "bad" and "in between" health categories with the reference category, and 95 % confidence intervals for the odds ratios.

Also present McFadden's adjusted pseudo R^2, AIC and BIC for the model.

Use a suitable test to determine whether this model is significantly better that a model with only an intercept (null model). State the type of test you use, the null hypothesis H_0, the distributon of the test statistic when H_0 is true, the observed value of the test statistic, the P-value and the conclusion.

Use the model to calculate the predicted probabilities of having at least one day in hospital for each of the three health categories, with 95 % confidence intervals. Compare with the proportions calculated from the cross-tabulation.

(b). Plot `hosp` against `age` and add a moving average. Does the probability of having at least one day in hospital increase or decrease, or both, with age? Come up with the probable reason for why the probability starts to decrease again as the age increases. *Hint:* Read the second paragraph of the Introduction again and reflect on the possible consequences.

Ignore the non-monotonous relationship and fit a simple logistic regression model for `hosp` as a function of `age`.

Present the β-estimates, the e^{β}-estimates and their confidence intervals. Also present McFadden's pseudo R^2, AIC and BIC for the model.

Use a suitable test to determine if age has a statistiscally significant impact on the probability of having at least one day in hospital. Report what type of test you use, the null hypothesis H_0, the distribution of the test statistic when H_0 is true, the observed value of the test statistic, the P-value and the conclusion.

Describe how the odds of having at least one day in hospital changes when age increases by 1 year, according to this model. Also describe how the odds changes when age increases by 5 years. Estimate this 5-year change rate and calculate a 95 % confidence interval for it.

Add the predicted probabilities and their 95 % confidence interval to the plot with the moving average. Comment on the result.

(c). We now want to model the non-monotonous relationship. A simple, but efficient, way is to add a quadratic term. This can be done with

```
glm(hosp ~ age + I(age^2), ...)
```

where the term `I(age^2)` is used to produce the square (compare `I(year - 1975)` in the labs). Fit this model and present the β-estimates, the e^{β}-estimates and their confidence intervals. Also present McFadden's pseudo R^2, AIC and BIC for the model.

Use a suitable test to determine whether the square term is statistically significant. Report what type of test you use, the null hypothesis H_0, the distribution of the test statistic when H_0 is true, the observed value of the test statistic, the P-value and the conclusion.

Add the predicted probabilities from this model, and their 95 % confidence interval, to the plot from 1(b). Comment on any interesting differences.

Write down the expression for the relative change in the odds (odds ratio) when the age increases by 1 year, as a function of the starting age. Calculate estimates of these odds ratios for age = 50, 75 and 100 years. Compare with the corresponding odds ratios from the linear model in 1(b).

Comment on the results.

Part 2. Variable selection

(a). Turn the rest of the categorical variables (`sex`, `civilst`, `exercise`, `work_norm`) into factors, present frequency tables and determine which category should be used as reference category, in each case.

(b). Plot the three continuous x-variables (`age`, `inc_hh`, `inc_tot`) against each other and calculate the correlation between each pair. Are there any problematic correlations between any of them?

(c). Fit a full model using age, age-squared, health, sex, civilst, exercise, work_norm, inc_hh, and inc_tot as covariates.

Present McFadden's pseudo R^2, AIC and BIC for the model.

For each of the variables, perform a suitable test of its significance in the model. Also test both age and age-squared at the same time. *Hint*: you can re-estimate a model with some of the x-variables removed by

```
update(model.full, . ~ . -variabletoremove)
```

For each test, state the type of test you use, the null hypothesis H_0, the distribution of the test statistic when H_0 is true, the observed value of the test statistic, the P-value and the conclusion.

(d). Perform a stepwise selection using AIC as criterion, starting with the null model and having the null model as lower scope and the full model from 2(c) as the upper scope.

State the order in which the variables are included and excluded. In particular, note what happens to `work_norm` and try to explain why it is removed again. *Hint*: do a boxplot, `+geom_boxplot()`, with `work_norm` on the x-axis and `age` on the y-axis.

Present the β-estimates, the e^{β}-estimates and their confidence intervals for the resulting model. Also present McFadden's pseudo R^2, AIC and BIC for the model and test whether this reduced model is significantly different from the full model. State the type of test you use, the null hypothesis, the distribution, the observed test statistic, the P-value and the conclusion.

(e). Perform a stepwise selection using BIC as criterion instead of AIC, again starting with the null model and having the null model as lower scope and the full model from 2(c) as the upper scope.

State the order in which the variables are included and excluded.

Present the β-estimates, the e^{β}-estimates and their confidence intervals of the resulting model. Also present McFadden's pseudo R^2, AIC and BIC for the model and test whether this reduced model is significantly different from the full model. State the type of test you use, the null hypothesis, the distribution, the observed test statistic, the P-value and the conclusion.

Is this model nested within the one in 2(d)? If so, also test whether this model is significantly different from that model.

Calculate predicted probabilities, with confidence intervals, for this model and plot them using age on the x-axis with different colour confidence ribbons for the health categories, `aes(..., fill = health)`, in separate plots according to sex, `+facet_wrap(~ sex)`.

Do we really need to separate between all three health categories? Fit the model again but use health status "bad" as reference instead of "good". Conclusion?

Create a new health variable with only two categories and fit a new model replacing `health` with this new variable. Present the β-estimates, the e^{β}-estimates and their confidence intervals. Also present McFadden's pseudo R^2, AIC and BIC for the model. Is this model better that the one with three health categories, according to BIC?

Part 3. Influential observations

(a). Calculate the leverage for the modified BIC-model in 2(e) and plot them against the age, separately for each combintion of sex and god/not good health (use `+facet_grid()`), including a suitable reference line. You may also want to colour code according to `hosp_cat`. Identify the observation with the highest leverage and examine what combination of variable values causes the high leverage.

(b). Calculate the standardized deviance residuals for the modified BIC model and plot them against the linear predictor, with suitable reference lines, and highlight the observation with the largest leverage found in 3(a). Comment on the residuals.

(c). Calculate Cook's distance for the modified BIC-model and plot them against age, separately for each combination of health and sex, including a suitable reference line, and highlight the observation with the largest leverage. Is this also an observation with a high Cook's distance? Identify the observation with the highest Cook's distance.

(d). Calculate the DFBETAS for the modified BIC-model and plot them against age, separately for each combination of health and sex. Which β-parameters were affected by the high leverage observation identified in 3(a) and/or the high Cook's distance observation identified in 3(c).

Part 4. Goodness-of-fit

We will now compare the different models and their ability to predict whether or not a person will have at least one day in hospital. Use the following models: 1(a) *Health*, 1(c) *Age squared*, 2(d) *AIC*, 2(e) *BIC-3* with three health categories, 2(e) *BIC-2* with two health categories.

(a). Use model 2(d) *AIC* and the threshold value 0.5, classifying observations with $\hat{p}_i \leq 0.5$ as "should not have any days in hospital", and observations with $\hat{p}_i > 0.5$ as "should have at least one day in hospital".

Present the resulting confusion matrix for the model and calculate the sensitivity, specificity, accuracy, and precision. Comment on the result.

(b). Plot the ROC-curves for all five models in the same plot and present a table with their AUC-values, including 95 % confidence intervals. Perform pair-wise tests comparing the AUC for model 2(d) *AIC* against each of the other models and discuss the result. Note: these tests are not independent but we perform them here as a crude way of determining whether the performance of the models are significantly different.

(c). For model 2(d) *AIC*, find the optimal threshold for p, where the sensitivity and the specificity are approximately equal, and as large as possible. Use the threshold to calculate a new confusion matrix, sensitivity, specificity, accuracy, and precision, and compare with the result in 4(a).

(d). Perform a Hosmer-Lemeshow goodness-of-fit test for model 2(d) *AIC*. Use a couple of different number of groups, g, starting at $g = p + 2$, in order to se how sensitive the test is to different choices. For each choice of g, report the smallest expected number of observations in any group.

Plot the observed and expected number of successes and failures, using group number on the x-axis. Relate the result of the HL-test to the appearance of the plot and comment on the result.

(e). Taking all the previous results into account, select the model you would prefer as the overall "best" model. Describe the reasons behind your decision.

At the beginning of your project report, include a very short Abstract or Executive summary (no formulas!), describing the overall best model. Describe which variables are left in the model, whether they each are associated with an increase or a decrease in the probability of having at least one day in hospital, and how good the model is at predicting the true outcome (e.g. AUC).

End of Project 2