RStudio Lab Week 7: Logistic Regression and Model Building

Data

Part 1: Logistic Regression

The logreg.csv is a data set from a study of depression. The objective of this analysis is to use the depression diagnosis of 150 individuals (cases) and assess its association with the sex of the respondent (sex) and their income (in 100 000s Rands) per year (income).

Part 2: Model Building

The step.csv data shows a medical related data set. The objective of the analysis is to model the average length of stay of all patients in a particular hospital. In order to compare the diﬀerences between patients coming from diﬀerent regional areas an indicator variable named region is constructed as well as an indicator variable for medical school aﬃliation of providers. The data set contains ten explanatory variables:

|  |  |
| --- | --- |
| Variable Name | Description |
| length (*y*) | Average length of stay of all patients in the hospital (days) |
| age | Average age of patients (in years) |
| infect | Infection Risk: Average estimated probability of acquiring infection in hos-  pital (%) |
| culture | Ratio of number of cultures performed to number of patients without signs  or symptoms of hospital-acquired infection, multiplied by 100 |
| xray | Ratio of number of X-rays performed to number of patients without signs or  symptoms of pneumonia, multiplied by 100 |
| beds | Average number of Beds in hospital during the study period |
| medschl | Medical School aﬃliation: 1 = Yes , 2 = No |
| region | Geographic region, where 1 =North East, 2=North West, 3=South, 4=West |
| census | Average number of patients in hospital per day during the study period |
| nurses | Average number of full-time equivalent registered and licensed practical  nurses during the study period (number full time plus one half the number part time) |
| facs | Available facilities and Services: Percent of 35 potential facilities and services  that are provided by the hospital |

Table 1: Hospital data

Instructions

Part 1: Logistic Regression

1. Make sure that you have downloaded the logreg.csv data.
2. Copy the dataset into your student drive and make sure you have opened RStudio.
3. Set your working directory to your student drive.
4. Create a new R Script
5. Import the logreg.csv data
6. Perform a logistic regression of cases against gender and income with the **glm()** function and save your result to an object called fit. Hint: *glm(formula = dependent variable ~ all explanatory variables, family = ‘binomial’)*
7. Use the **summary()** function on the fit object to get your model results

Part 2: Model Building

1. Make sure that you have downloaded the step.csv data.
2. Copy the dataset into your student drive and make sure you have opened RStudio.
3. Set your working directory to your student drive.
4. Create a new R Script
5. Import the step.csv data
6. The variables medschl and region are numeric. Convert them to factor variables so that R-studio knows to recognise them as categorical. Use the **as.factor()** function.
7. Perform a multiple linear regression with length as the dependent variable and all other variables as explanatory variables; save your result in an object called fit.full Hint: *formula = y ~ .*
8. Use the summary function to view your model results
9. Create a linear regression model with only an intercept term called fit.empty. Hint: *formula=y~1*
10. Perform a forward selection process using the **step()** function and save your results in an object called step.model. Hint: *step(empty\_model, scope=formula(full\_model), direction = ‘forward’)*
11. Examine the final fitted model using the **summary()** function on your step.model object.