Tutorial Business Analytics

Tutorial 4

Exercise 4.1

<u>Note</u>: Use R to solve this exercise(Exercise 4.1_R-template.R).

Load the training data *("admit-train.csv")*into R. Proceed by typing *names(train)* to print the attribute names to the console.

```
library(tidyverse)
train = read_csv("admit-train.csv")
names(train)
```

The attribute *"admit"* indicates whether a student has been admitted to a Master's Course. The attributes *"gre"* and *"gpa"* contain the results of certain exams. The attribute *"rank"* represents the reputation rank of the student's current university. The smaller the rank, the higheris the university's reputation. The functions *summary()*, *table()*, *sd()*, *hist()*, *plot()*etc. provide you with several statistics about the attributes.

a) Briefly describe the data set:

   i. Name the dependent variable and the independent variables
   ii. Which scales of measurement do the variables belong to (e.g. nominal, ordinal, interval or ratio)?

Due to the fact that the dependent attribute *"admit"* is binary, you have decided to use a logistic regression model. Use below-mentioned commands to create a logit-model from the training data and to obtain the results.

```
mylogit = glm(admit~gre+gpa+as.factor(rank), data=train, family=binomial(link="logit"))
summary(mylogit)
```

b) Which attributes are statistically significant regarding a significance level of 5%?

c) Interpret the coefficients.

d) Test the significance of the attribute "*rank*" using a Wald-Test. In order to do that, install the package *aod* (RStudio: Tools -> Install Packages). Then enter the following commands:

```
library(aod)
wald.test(b=coef(mylogit), Sigma=vcov(mylogit), Terms=4:6)
```

e) In order to gain a better understanding of the model, have a look at the predicted probabilities of some observations. Adjust only one parameter and keep the others constant. For example keep "*gre*" and "*gpa*" constant (using their mean/average) and vary "*rank*". This can be done using below-mentioned commands:

```
rank<-c(1,2,3,4)
gre <-c(mean(train$gre))
gpa <-c(mean(train$gpa))
myInstances <- data.frame(gre,gpa,rank)
```

Print the results to console:

```
myInstances
```

In order to find the predicted probability add another variable named *pAdmit* to *myInstances* and fill it with values from the *myLogit* model.

```
myInstances$pAdmit<-predict(mylogit, newdata=myInstances, type="response")
```

Have a look at the results:

```
myInstances
```

Can you draw any conclusions?

f) Find the McFadden ratio and interpret the results:

```
McFadden <- 1 – (mylogit$deviance / mylogit$null.deviance)
```

g) Load the data record "*admit-test.csv*" and predict the probability:

```
test <- read_csv("admit-test.csv")
preds <- predict(mylogit, newdata=test, type='response')
```

Construct the confusion matrix.

```
test = test %>% mutate(pred = round(preds))
test %>% group_by(admit, pred) %>% summarise(count=n())
```

or

```
table(true=test$admit,prediction=round(preds))
```

h) Find the logit model's error rate.

```
incorrectPredictionCount = nrow(test %>% filter(admit!=pred))
totalPredictions = nrow(test)
errorRate = incorrectPredictionCount/totalPredictions
errorRate
```

Solution - Exercise 4.1 - Exercise 4.1_R-Script

Exercise 4.2

You are provided the following numbers from the result of a Poisson Regression model.

Variable	**Estimate**	**Std. Error**
Intercept	1.5499	0.0503
Age	-0.0047	0.0009

a) According to the model above, what qualitative effect does a change in the independent variable *age* (+1) have on the dependent variable *dv*.

b) According to the model above, what quantitative effect (on the *incidence rate* and *log-incidence rate*) does a change in the independent variable *age* (+1) have on the dependent variable *dv*.

Solution - Exercise 4.2

a)

$$ln(dv) \ = \ 1.5499 \ + \ (-0.0047)(age)$$

Given the coefficients in the result, we can write the above equation.
As can be seen - an increase in *age* will decrease the dependent variable, dv.

b)

$$\Delta(log\ incidence\ rate) \ = \ \beta_1 \ = \ -0.0047$$

The log-incidence rate decreases by 0.0047 with an increase of 1 in the variable age.

$$\Delta(incidence\ rate) \ = \ e^{\beta_1} \ = \ 0.9953$$

The incidence rate decreases by a factor of 0.9953 when the variable age increases by 1. This means that there is an approximate 0.5% reduction for each increase in the age by 1.

Exercise 4.3

You are given the following dataset with the dependent binary variable y and the independent variable x.

x	y
1	0
2	0
2.5	1
4	1

Based on these datapoints we want to create a logistic regression model with the logistic function (corresponding to sigmoid function σ($\beta_0 + \beta_1 x$)):

$$\Pr[Y|X] = p(x) = \sigma(\beta_0 + \beta_1 x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

To estimate the logistic regression coefficients, we will use the Maximum Likelihood Estimation.

a) Determine the likelihood function L.

b) Find the gradient for the log of the likelihood function LL(β). The gradient is defined as:

$$\nabla LL(\beta) = \begin{pmatrix} \frac{\partial LL(\beta)}{\partial \beta_0} \\ \frac{\partial LL(\beta)}{\partial \beta_1} \end{pmatrix}$$

Hint: Use the chain rule: $\frac{\partial LL(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial LL(\beta)}{\partial p_i} * \frac{\partial p_i}{\partial z_i} * \frac{\partial z_i}{\partial \beta_j}$ with $z_i = \beta_0 + \beta_1 x_i$

You may use the following derivative of the sigmoid function $\sigma(\cdot)$ without proof: $\sigma'(z_i) = \sigma(z_i) \cdot (1 - \sigma(z_i))$

c) Given the initial values $\beta^{(0)} = \binom{0}{0}$ and $\alpha = 0.2$ calculate the coefficients after the first iteration of gradient ascent.

d) If a linear regression model was fitted to a logistic regression dataset, what could be the problems w.r.t. Gauss Markov properties?

Solution - Exercise 4.3

a) $L = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = p_3 p_4 (1-p_1)(1-p_2)$

b) $LL = \sum_{i=1}^{4} y_i \log p_i + (1-y_i)\log(1-p_i)$
$= \log(p_3) + \log(p_4) + \log(1-p_1) + \log(1-p_2)$

$\nabla LL(\beta) = \begin{pmatrix} \frac{\partial LL}{\partial \beta_0} \\ \frac{\partial LL}{\partial \beta_1} \end{pmatrix}$ where $\frac{\partial LL(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial LL(\beta)}{\partial p_i} * \frac{\partial p_i}{\partial z_i} * \frac{\partial z_i}{\partial \beta_j}$

$$\frac{\partial z_i}{\partial \beta_0} = 1, \frac{\partial z_i}{\partial \beta_1} = x_i$$

$$\frac{\partial p_i}{\partial z_i} = p_i * (1-p_i)$$

$$\frac{\partial LL(\beta)}{\partial p_i} = \frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}$$

Calculate partial derivative for β_0:

$$\frac{\partial LL(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \frac{\partial LL(\beta)}{\partial p_i} * \frac{\partial p_i}{\partial z_i} * \frac{\partial z_i}{\partial \beta_0} = \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}\right) * \left(p_i * (1-p_i)\right) * 1$$

$$= -p_1 - p_2 + (1-p_3) + (1-p_4)$$

Calculate partial derivative for β_1:

$$\frac{\partial LL(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} \frac{\partial LL(\beta)}{\partial p_i} * \frac{\partial p_i}{\partial z_i} * \frac{\partial z_i}{\partial \beta_1} = \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}\right) * (p_i * (1-p_i)) * x_i$$

$$= -p_1 - 2p_2 + 2.5(1-p_3) + 4(1-p_4)$$

c) $\beta^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \xrightarrow{yields} p_i = \frac{1}{2}, \frac{\partial p_i}{\partial z_i} = \frac{1}{4}$

$$\left.\frac{\partial LL(\beta)}{\partial \beta_0}\right|_{\beta=(0,0)} = -\frac{1}{2} - \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 0$$

$$\left.\frac{\partial LL(\beta)}{\partial \beta_1}\right|_{\beta=(0,0)} = -\frac{1}{2} - 2 * \frac{1}{2} + 2.5 * \frac{1}{2} + 4 * \frac{1}{2} = \frac{7}{4}$$

$$\beta_0^{(1)} = \beta_0^{(0)} + \alpha \frac{\partial LL}{\partial \beta_0} = 0.2 * 0 = 0$$

$$\beta_1^{(1)} = \beta_1^{(0)} + \alpha \frac{\partial LL}{\partial \beta_1} = 0.2 * \frac{7}{4} = 0.35$$

$$=> \beta^{(1)} = \begin{pmatrix} 0 \\ 0.35 \end{pmatrix}$$

d) First, if a linear regression model were fitted to this data, the predicted values of $\hat{y}$ could be outside [0,1], which in a binary logistic setting would not be a good prediction since the interpretation as a probability would not make sense.

Second, the properties of Autocorrelation and Homoscedasticity would be violated.

- Autocorrelation: It can be noted that the residuals of a linear model fitted to a setting with a binary dependent variable would result in positive residuals on one side, negative on the other and in the range of $\hat{y}$ = [0,1] they would be ∈ [-1,1]. This would form a pattern that indicates autocorrelation.

- Homoscedasticity: Error terms do not have constant variance because true Y takes on only two values {0,1}, therefore they are heteroscedastic.