Prof. Dr. M. Kateri,
Dr. W. Herff,
L. Kaufmann, M.Sc.

RWTH Aachen, SS 2022
Release date: June 13th, 2022
Solution: Week starting from June 20th, 2022

Applied Data Analysis

R-Laboratory 8

Logistic Regression – Link Functions – Model Selection

Task 27

(a) Download the file *transportation.csv* from RWTHmoodle and load it as a data frame into your workspace. Create two boxplots of the attribute `time`, one for each mean of transport.

(b) Fit a logistic regression model, where the attribute `transport` depends only on `time`.

(c) Fit binary regression models for the problem described in (b) but for the link functions probit and cloglog and comment their fit. Plot (on the same plot) the predicted probabilities for using public transport as a function of `time` for all three models considered in (b) and (c).

(d) Fit two logistic regression models as in (b), one for each sex. Plot in `ggplot2` the predicted probabilities for using public transport as a function of time on the same plot, using different colors per gender.

(e) Fit a logistic regression model, where `transport` depends on `time` and `sex`. Calculate the percentage of correct classified observations according to this model. Compare to the corresponding percentage for the model in (b).

(f) Fit a logistic regression model, where `transport` depends on `time`, `sex` and their interaction. Is the interaction statistically significant?

Task 28

(a) Download the file *credits.csv* from RWTHmoodle and load it as a data frame into your workspace. Divide randomly the data of sample size $n = 1000$ into training data (67 % of the data) with 667 rows and test data (the remaining 33% of the data).

(b) Based on the training data fit a logistic regression model, that predicts `repayment` using the attributes:

`account`, `behavior`, `savings`, `employment`, `rate`, `guaran`, `finance`, `furthcred`, `home`, `job`, `time`, `amount`, `age`.
Hint: Some of the explanatory variables are categorical and have to be defined as factors before fitting the model. Use the data set documentation to decide which variables are the factor variables.

(c) Select the best model (nested in the model fitted in (b)) in terms of AIC and BIC using a backwards stepwise selection algorithm. Compute the predicted values for the test data based on the model in (b) and both selected models.

(d) Based on the training data fit a logistic regression model, that predicts `repayment` using the attribute `simple.score` (calculated in R-Laboratory 2, Task 8). Compute the predicted values for the test data. Furthermore, compute a linear predictor, that is equal to $1-\frac{\texttt{simple.score}}{\texttt{max.score}}$, where `max.score` is the largest possible value of `simple.score`.

(e) Compute the *predicted residual sum of squares* (PRESS) based on the test data for all five considered models. Which model do you propose in terms of PRESS?

```
########TASK27########
#a) load data
transport.task27 = read.csv("R-Lab-Datasets/transportation.csv", header = TRUE, sep = ";")
# boxplot
ggplot(transport.task27, aes(as.factor(transport), time, color = transport)) +
geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)
# b)fit the logistic regression
transport.fit = glm(transport ~ time, data = transport.task27, family = "binomial")
#c) fit binary regression using probit
probit.fit = glm(transport ~ time, data = transport.task27, family = binomial(link = "probit"))
# fit binary regression using cloglog
cloglog.fit = glm(transport ~ time, data = transport.task27, family = binomial(link = "cloglog"))

plot(sort(transport.task27$time), sort(fitted(transport.fit)), type="l", col = "green")
lines(sort(transport.task27$time), sort(fitted(probit.fit)), col = "blue")
lines(sort(transport.task27$time), sort(fitted(cloglog.fit)), col = "red")

# predicted probabilities for different link functions
prob.fit = function(model,x,link="logit"){
  pred.lin = model$coefficients[1] + model$coefficients[2] * x
  if(link=="logit"){
    return(exp(pred.lin)/(1 + exp(pred.lin)) )
  }
  if(link=="probit"){
    return(pnorm(pred.lin))
  }
  if(link=="cloglog"){
    return(1-exp(-exp(pred.lin)))
  }
}

# (d)
# divide the data set
transportation.f = transport.task27[transport.task27$sex==1,]
transportation.m = transport.task27[transport.task27$sex==2,]
# fit the model
transp.glm1.f = glm(transport ~ time, family = binomial, data=transportation.f)
Text:
transp.glm1.m = glm(transport ~ time, family = binomial, data=transportation.m)
# predicted probabilities
x = seq(-15,60,0.01)
#use written function above for the predictions
prob.pred.f = prob.fit(transp.glm1.f,x)
prob.pred.m = prob.fit(transp.glm1.m,x)
#write predictions in data frame
prob.pred=data.frame(x=x,prob.pred.f=prob.pred.f, prob.pred.m=prob.pred.m)

col=c("woman"="red", "man"="blue")
ggplot(prob.pred, aes(x=x))+geom_line(aes(y=prob.pred.f, color="woman"))+
  geom_line(aes(y=prob.pred.m, color="man")) +
  labs(x="time", y="prob",color="Legend")+scale_color_manual(values = col)
```

```
# e)
transportation$sex=factor(transportation$sex)
levels(transportation$sex) = c("woman","man")
transp.glm2 = glm(transport ~ time + sex,data=transportation,family = binomial)
#model where response depends on time AND sex
predicted.transport = ifelse(predict(transp.glm2) > 0.5, 1, 0)
#length(predicted.transport == transportation$transport)

# confusion matrix
confusion.mat = table(transportation$transport ,predicted.transport)
# get the accuracy
sum(diag(confusion.mat))/sum(confusion.mat)

# f)
transp.glm3 = glm(transport ~ time * sex,data=transportation,family = binomial)
# different ways to compute the p value
1-pchisq(deviance(transp.glm2) - deviance(transp.glm3),
        df= df.residual(transp.glm2) - df.residual(transp.glm3))
anova(transp.glm3,test="Chisq")
anova(transp.glm2, transp.glm3, test="Chisq")
```