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

RWTH Aachen, SS 2022
Release date: June 20th, 2022
Solution: Week starting from June 27th, 2022

Applied Data Analysis

R-Laboratory 9

Logistic Regression – Baseline-Category Logit Model – Cumulative Link Model

Do not use functions from additional R packages (except when it is explicitly stated in the Task, Hint or list of useful packages and functions).

Useful packages and functions:

- `pROC`
- `pROC::roc()`
- `pROC::plot.roc()`
- `pROC::auc()`
- `pchisq()`
- `confint()`
- `ftable()`
- `VGAM`
- `VGAM::vglm()`
- `VGAM::step4()`
- `xtabs()`
- `deviance()`
- `df.residuals()`
- `matplot()`

Task 29

(a) Download the file *FieldGoal.csv* from RWTHmoodle and load it as a data frame into your workspace.

(b) Fit a logistic regression model that predicts `Good.` using `Dist` as explanatory variable and plot the predicted probabilities for a good kick as a function of `Dist`.

(c) Calculate the percentage of correct classified observations according to the model in (b).

(d) Test at significance level $\alpha = 0.05$ the hypotheses

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0$$

in the model from (b), where β_1 is the parameter of the attribute `Dist`. Base your test decision on the Wald's test statistic.

(e) Compute 90% profile likelihood confidence intervals for the parameters of intercept and `Dist`, as well as for the probability of a successful kick when the distance to the goal is 19, 39 and 64 yards.

(f) Fit a logistic regression model, that predicts `Good.` using the attributes `Dist`, `Blk.`, `Pressure`, `Roof.type`, `Altitude` and `Field`. Select the model with smallest AIC using a backwards stepwise selection algorithm. Calculate the percentage of correct classified observations according to this model and compare it with the percentage from (e).

(g) Plot (in a single figure) the ROC curves for the model in (b) and the selected model in (f). Evaluate for both models the area under the curve (AUC).

Task 30

(a) Load the dataset `Hoyt` into your workspace and transform it to a 'flat' contingency table, with 4 columns, each corresponding to a level of the nominal attribute `Status`.

Hint: You may use the function `ftable` with the argument `col.vars="Status"`.

Explanation of `ftable`: The function `ftable` creates a flat contingency table which means that the usual information contained in a contingency table are re-arranged as a matrix. The rows and columns of this matrix correspond to the combinations of the levels of the involved factors. `ftable` can be applied to a data frame or a contingency table generated with `xtabs` for example, where `xtabs` creates a contingency table from cross-classifying all involved factors which is returned as a table and not re-arranged as a matrix

(b) Fit a baseline-category logit model for `Status` with explanatory variables `Rank`, `Occupation`, `Sex` and their pairwise interaction terms.

Hint: You may use the function `vglm` from the package `VGAM` with the key command

`vglm(formula,family,data).`

You can fit grouped data by placing a matrix as response in `formula`.

(c) Given the model you fitted in (b), test at significance level $\alpha = 0.05$, if there is any influence of the explanatory variables on `Status`.

what is the null hypothesis : any of beta is non-zero

(d) Given the model you fitted in (b), test the significance of the effect of `Sex:Rank` on `Status` at significance level $\alpha = 0.05$.

(e) Select the model with smallest AIC and with smallest BIC using a backwards stepwise selection algorithm starting with the model from (b).

Task 31

(a) Load the dataset `Vietnam` into your workspace and transform it to a 'flat' contingency table, with 4 columns, each corresponding to a level of the ordinal attribute `response`.

(b) Compute the baseline-category sample log-odds of `response` at all levels of `sex` and `year` and the cumulative sample log-odds and plot them as a function of `year` (of study) for each gender, with genders indicated by different colors.

Hint: You may use the function `matplot`.

(c) Fit a cumulative logit model for `response` with explanatory variables `sex` and `year`.

Hint: Use the arguments `parallel=FALSE` and `link="logitlink"` within the `family` argument.

(d) Fit a cumulative probit model for `response` with explanatory variables as in (c).

Hint: Use the arguments `parallel=TRUE` and `link="probitlink"` within the `family` argument.

(e) For both models in (b) and (c) compute the Pearson's X^2 and the deviance in order to perform a goodness of fit test at significance level $\alpha = 0.01$.

Hint: You may use the function `predict` with argument `type="response"` to compute the estimated response probabilities for each factor level of `Status`, conditional on the values of the explanatory variables.

```
#########TASK29#########
FieldGoal = read.csv2("R-Lab-Datasets/FieldGoal.csv", header = TRUE, sep = ";")
# b) fit the logistic regression model
FG.fit = glm(Good. ~ Dist, data = FieldGoal, family = "binomial")
# c) plot

pred.binom = function(x, coef){

  bx = coef[1] + x*coef[2]

  return(exp(bx)/(1 + exp(bx)))

}
Good.pred = pred.binom(FieldGoal$Dist, FG.fit$coefficients)
Good.pred = ifelse(Good.pred > 0.5, 1, 0)
field.range = min(FieldGoal$Dist):max(FieldGoal$Dist)
plot(field.range, pred.binom(field.range, FG.fit$coefficients),
     type = 'l', col = "blue")
# generate the confusion amtrix
confusion.mat = table(FieldGoal$Good. , Good.pred)
# accuracy
sum(diag(confusion.mat))/sum(confusion.mat)

#d) significance test with wald statistics
summary(FG.fit)
# Wald's test statistic
z.squared=(FG.fit$coefficients[2]/summary(FG.fit)$coefficients[2,2])^2
p.val=1-pchisq(z.squared, df=1)
p.val

# (e)
CI=confint(FG.fit, level=0.9) #profile likelihoood CI
# probability of good kick with 19,39 and 64 yards to the goal
CI.prob=function(x){
  exp(CI[1, ]+CI[2, ]*x)/(1 + exp(CI[1, ]+CI[2,]*x))
}
CI.prob(19)
CI.prob(39)
CI.prob(64)

# f)

model.select =

# (g) ROC Curves
roc.curve1=roc(Good.~fitted(model.1), data=FG)
roc.curve2=roc(Good.~fitted(model.select), data=FG)
plot.roc(roc.curve1, legacy.axes=TRUE)
plot.roc(roc.curve2, legacy.axes=TRUE, add=TRUE, col="blue")
auc(roc.curve1)
auc(roc.curve2)
```

```
#########TASK30########
library(VGAM)
library(vcdExtra)

# (a) generate the contingency table
Hoyt.tab=ftable(Hoyt, col.vars="Status")
Hoyt.tab
Rank=factor(c(rep("Low", 14), rep("Middle", 14), rep("High", 14)))
Occ=factor(rep(c(rep(1, 2), rep(2, 2), rep(3, 2), rep(4, 2), rep(5,2), rep(6,2), rep(7,2)), 3))
Sex=factor(rep(c("Female", "Male"), 21))

#b) fit the baseline category
formula.blc = Hoyt.tab~ Rank*Occ + Rank*Sex + Occ*Sex
vglm.hoyt = vglm(formula.blc, family=multinomial)

# (d)
fit2=vglm(Hoyt.tab~Rank*Occ+Occ*Sex,family=multinomial)
# p-value of influence of Sex:Rank
1-pchisq(deviance(fit2)-deviance(fit1), df=df.residual(fit2)-df.residual(fit1))

# (e)
model.selectA=step4(fit1, directions="backward")
model.selectB=step4(fit1, directions="backward", k=log(sum(Hoyt.tab)))
```

```
# Task 31
library(VGAM)
library(vcdExtra)

# (a)
Vietnam.tab=ftable(xtabs(Freq~sex+year+response,data=Vietnam), col.vars="response")
Vietnam.tab #flat contingency table with 4 columns
sex=factor(c(rep("Female",5), rep("Male", 5)))
year=factor(rep(1:5,2))



# (b)
# baseline category log odds
baseline.cat.log.odds<-matrix(0,10,3) #we have 10 rows in Vietnam.tab and 4 columns
for (i in 1:3){
  baseline.cat.log.odds[,i] = log(Vietnam.tab[,i]/Vietnam.tab[,4])
}
baseline.cat.log.odds=xtabs(baseline.cat.log.odds~sex+year)



# cumulative log odds
cum.log.odds<-matrix(0,10,3)
cum.log.odds[,1]=log(Vietnam.tab[,1]/rowSums(Vietnam.tab[,2:4]))
cum.log.odds[,2] = log(rowSums(Vietnam.tab[,1:2])/rowSums(Vietnam.tab[,3:4]))
cum.log.odds[,3]=log(rowSums(Vietnam.tab[,1:3])/Vietnam.tab[,4])
cum.log.odds=xtabs(cum.log.odds~sex+year)



# plot of the log-odds
matplot(baseline.cat.log.odds[1, , ],
        type="p",pch=1,xlab="",
        ylab="Estimated Baseline-Category Log-Odds",
        xaxt="n",main="Sample Baseline-Category LO for factor level female")
matplot(baseline.cat.log.odds[2, , ],
        type="p",pch=1,xlab="",
        ylab="Estimated Baseline-Category Log-Odds",
        xaxt="n",main="Sample Baseline-Category LO for factor level male")
matplot(cum.log.odds[1, , ],
        type="p",pch=1,xlab="",
        ylab="Estimated Cumulative Log-Odds",
        xaxt="n",main="Sample Cumulative LO for factor level female")
matplot(cum.log.odds[2, , ],
        type="p",pch=1,xlab="",
        ylab="Estimated Cumulative Log-Odds",
        xaxt="n",main="Sample Cumulative LO for factor level male")
```

```
#(c)
# cumulative logit model
fit.viet.logit=vglm(Vietnam.tab~sex+year,
                    family = cumulative(parallel=FALSE,
                                        link="logitlink"))

# (d)
# cumulative probit model
fit.viet.probit=vglm(Vietnam.tab~sex+year,
                     family = cumulative(parallel=TRUE,
                                         link="probitlink"))

# (e)
# cumulative probit model
# Chi-squared
#need for pearsons X^2
mu=c(predict(fit.viet.probit,type="response")*rowSums(Vietnam.tab))
X.square.probit<-sum((c(Vietnam.tab)-mu)^2/mu) #pearsons X^2
# p-values
1-pchisq(deviance(fit.viet.probit),df.residual(fit.viet.probit))
1-pchisq(X.square.probit,df.residual(fit.viet.probit))

# cumulative logit model
# Chi-squared
mu=c(predict(fit.viet.logit,type="response")*rowSums(Vietnam.tab))
X.square.logit<-sum((c(Vietnam.tab)-mu)^2/mu)
# p-values
1-pchisq(deviance(fit.viet.logit),df.residual(fit.viet.logit))
1-pchisq(X.square.logit,df.residual(fit.viet.logit))
```