Homework 10

Tauqeer Kasam Rumaney

January 16, 2023

Download this R Markdown file, save it on your computer, and perform all the below tasks by inserting your answer in text or by inserting R chunks below. After you are done, upload this file with your solutions on Moodle.

For all exercises, use the KiGGS dataset.

```
dat_link <- url("https://www.dropbox.com/s/pd0z829pv2otzqt/KiGGS03_06.RData?dl=1")
load(dat_link)
dat <- KiGGS03_06
```

Exercise 1: Logistic regression

Choose 1 suitable outcome variable of interest and 3 predictors, and compute a logistic regression model. Interpret the results: which predictor is associated with the outcome and what is the strength of association (odds ratio)? Also, is the model a good fit i.e. can the outcome be predicted well (look at the misclassification table for this)?

```
summary(dat$capi) #to check if it is binary variable or not
```

```
##  Nein    Ja
##   191 17449
```

```
sbp1 <- as.numeric(as.character(dat$sys1))
sbp2 <- as.numeric(as.character(dat$sys2))
pp <- as.numeric(as.character(dat$PPoint))
logit <- glm(dat$capi ~ sbp1+sbp2+pp, data=dat,family="binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = dat$capi ~ sbp1 + sbp2 + pp, family = "binomial",
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2253   0.1342   0.1442   0.1541   0.2381
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.8298963  0.7811766   4.903 9.45e-07 ***
## sbp1         0.0328863  0.0128611   2.557   0.0106 *
## sbp2        -0.0258426  0.0128858  -2.006   0.0449 *
## pp          -0.0008698  0.0016906  -0.514   0.6069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1698.5  on 14564  degrees of freedom
## Residual deviance: 1691.5  on 14561  degrees of freedom
##   (3075 observations deleted due to missingness)
## AIC: 1699.5
##
## Number of Fisher Scoring iterations: 7
```

#Conclusion: Each one-unit change in sbp1 will increase the log odds of getting interviwed by doctor by 0.03, and its p-value(< 0.05) indicates that it is somewhat significant in determining whether someone gets interviwed by doctor or not.

Each one-unit change in sbp2 will decrease the log odds of getting interviwed by doctor by 0.02, and its p-value(< 0.05) indicates that it is somewhat significant in determining whether someone gets interviwed by doctor or not.

Since the p value is greater than 0.05, it is not significant in the regression model. This mean that it is possible for us to remove them from our model since removing it won't significantly affect our results.

Since the difference is only 7 between Null deviance and Residual deviance as well as AIC is very high, the model is not a good fit.

Exercise 2: Poisson regression

Predict the amount of measles vaccinations (Maanzahl) by the number of siblings (e006B1), the sex and age of the children (sex, age2), place of residence (STALA, OW) and the monthly household income (e093), using a Poisson regression model. Interpret the results. Which variables are associated with the outcome? Is the model a good fit to the data?

```
str(dat$Maanzahl)
```

```
##  'labelled' int [1:17640] 3 2 2 2 2 0 0 1 1 NA ...
##  - attr(*, "label")= Named chr "Anzahl der Masernimpfungen"
##   ..- attr(*, "names")= chr "Maanzahl"
```

```
table(dat$Maanzahl)
```

```
##
##     0     1     2     3     4     5     6
##  2023  3163 10407   583   277     4     2
```

```
mean(dat$Maanzahl)
```

```
## [1] NA
```

```
var(dat$Maanzahl)
```

```
## [1] NA
```

```
#I am getting mean and variance as NA, can you explain why?
summary((dat$e006B1))
```

```
##      Kein Geschwisterkind       Ein Geschwisterkind Zwei und mehr Geschwister
##                      4553                      5996                      2058
##                      NA's
##                      5033
```

```
table(dat$sex)
```

```
##
## Männlich Weiblich
##     8985     8655
```

```
table(dat$age2)
```

```
##
##   0 - 1 J.   2 - 3 J.   4 - 5 J.   6 - 7 J.   8 - 9 J. 10 - 11 J. 12 - 13 J.
##       1860       1879       1935       2032       2104       2076       2018
## 14 - 15 J. 16 - 17 J.
##       1972       1764
```

```
summary(dat$STALA)
```

```
##        Ländlich  Kleinstädtisch Mittelstädtisch   Großstädtisch
##            3913            4654            5059            4014
```

```
summary(dat$OW)
```

```
##   Ost  West
##  5899 11741
```

```
summary(dat$e093)
```

```
##           < 500 €     500 - < 750 €   750 - < 1.000 € 1.000 - < 1.250 €
##               194               498               739               919
## 1.250 - < 1.500 € 1.500 - < 1.750 € 1.750 - < 2.000 € 2.000 - < 2.250 €
##              1239              1188              1601              1726
## 2.250 - < 2.500 € 2.500 - < 3.000 € 3.000 - < 4.000 € 4.000 - < 5.000 €
##              1843              2571              2409               982
##        >= 5.000 €              NA's
##               643              1088
```

```
vac <- as.numeric(as.character(dat$Maanzahl))
siblings <- as.numeric(as.character(dat$e006B1))
```

```
## Warning: NAs introduced by coercion
```

```
sex <- as.numeric(as.character(dat$sex))
```

```
## Warning: NAs introduced by coercion
```

```
age <- as.numeric(as.character(dat$age2))
```

```
## Warning: NAs introduced by coercion
```

```
res <- dat$STALA
direction <- dat$OW
income <- dat$e093
#output <- glm(dat$Maanzahl ~ siblings, data = dat, family = poisson(link = "log"))
```

```
#print(summary(output))
#Error in glm.fit(x = numeric(0), y = integer(0), weights = NULL, start = NULL, : object 'fit' not foun
```

```
myvars <- c("Maanzahl", "e006B1", "sex", "age2", "STALA", "OW","e093")
newdata <- dat[myvars]
summary(newdata)
```

```
##     Maanzahl                            e006B1           sex
##  Min.   :0.000   Kein Geschwisterkind     :4553   Männlich:8985
##  1st Qu.:1.000   Ein Geschwisterkind      :5996   Weiblich:8655
##  Median :2.000   Zwei und mehr Geschwister:2058
##  Mean   :1.632   NA's                     :5033
##  3rd Qu.:2.000
##  Max.   :6.000
##  NA's   :1181
##          age2                  STALA         OW                        e093
##  8 - 9 J.  :2104   Ländlich       :3913   Ost : 5899   2.500 - < 3.000 €:2571
##  10 - 11 J.:2076   Kleinstädtisch :4654   West:11741   3.000 - < 4.000 €:2409
##  6 - 7 J.  :2032   Mittelstädtisch:5059                2.250 - < 2.500 €:1843
##  12 - 13 J.:2018   Großstädtisch  :4014                2.000 - < 2.250 €:1726
##  14 - 15 J.:1972                                       1.750 - < 2.000 €:1601
##  4 - 5 J.  :1935                                       (Other)           :6402
##  (Other)   :5503                                       NA's              :1088
```

```
newdata <- newdata[complete.cases(newdata), ]

vac1 <- as.numeric(as.character(newdata$Maanzahl))
siblings1 <- as.numeric(as.character(newdata$e006B1))
```

```
## Warning: NAs introduced by coercion
```

```
sex1 <- as.numeric(as.character(newdata$sex))
```

```
## Warning: NAs introduced by coercion
```

```
age1 <- as.numeric(as.character(newdata$age2))
```

```
## Warning: NAs introduced by coercion
```

```
res1 <- newdata$STALA
direction1 <- newdata$OW
income1 <- newdata$e093
output3 <- glm(newdata$Maanzahl ~ newdata$e006B1 + newdata$sex + newdata$age2 + newdata$STALA + newdata$
print(summary(output3))
```

```
##
## Call:
## glm(formula = newdata$Maanzahl ~ newdata$e006B1 + newdata$sex +
##     newdata$age2 + newdata$STALA + newdata$OW + newdata$e093,
##     family = poisson(link = "log"), data = newdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2020  -0.4455   0.1187   0.2852   1.8808
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                                -0.802697   0.108126  -7.424 1.14e-13
## newdata$e006B1Ein Geschwisterkind          -0.064634   0.015989  -4.042 5.29e-05
## newdata$e006B1Zwei und mehr Geschwister -0.107968   0.023028  -4.688 2.75e-06
## newdata$sexWeiblich                         0.003743   0.014637   0.256   0.7982
## newdata$age22 - 3 J.                        1.245780   0.058150  21.424  < 2e-16
## newdata$age24 - 5 J.                        1.327157   0.057039  23.268  < 2e-16
## newdata$age26 - 7 J.                        1.395824   0.056406  24.746  < 2e-16
## newdata$age28 - 9 J.                        1.398620   0.056312  24.837  < 2e-16
## newdata$age210 - 11 J.                      1.399978   0.056386  24.828  < 2e-16
## newdata$age212 - 13 J.                      1.405903   0.056724  24.785  < 2e-16
## newdata$age214 - 15 J.                      1.495007   0.056632  26.399  < 2e-16
## newdata$age216 - 17 J.                      1.582862   0.057307  27.621  < 2e-16
## newdata$STALAKleinstädtisch                 0.024188   0.020733   1.167   0.2434
## newdata$STALAMittelstädtisch                0.055111   0.020612   2.674   0.0075
## newdata$STALAGroßstädtisch                  0.005487   0.022412   0.245   0.8066
## newdata$OWWest                             -0.154373   0.015873  -9.725  < 2e-16
## newdata$e093500 - < 750 €                   0.083248   0.104523   0.796   0.4258
## newdata$e093750 - < 1.000 €                 0.106071   0.099806   1.063   0.2879
## newdata$e0931.000 - < 1.250 €               0.041858   0.098241   0.426   0.6700
## newdata$e0931.250 - < 1.500 €               0.074999   0.095833   0.783   0.4339
## newdata$e0931.500 - < 1.750 €               0.089603   0.096092   0.932   0.3511
## newdata$e0931.750 - < 2.000 €               0.095711   0.094873   1.009   0.3131
## newdata$e0932.000 - < 2.250 €               0.105829   0.094422   1.121   0.2624
## newdata$e0932.250 - < 2.500 €               0.106139   0.094139   1.127   0.2595
## newdata$e0932.500 - < 3.000 €               0.077446   0.093486   0.828   0.4074
## newdata$e0933.000 - < 4.000 €               0.075624   0.093576   0.808   0.4190
## newdata$e0934.000 - < 5.000 €               0.068920   0.095999   0.718   0.4728
## newdata$e093>= 5.000 €                      0.050665   0.098173   0.516   0.6058
##
## (Intercept)                                ***
## newdata$e006B1Ein Geschwisterkind          ***
## newdata$e006B1Zwei und mehr Geschwister ***
## newdata$sexWeiblich
## newdata$age22 - 3 J.                       ***
## newdata$age24 - 5 J.                       ***
## newdata$age26 - 7 J.                       ***
## newdata$age28 - 9 J.                       ***
## newdata$age210 - 11 J.                     ***
## newdata$age212 - 13 J.                     ***
## newdata$age214 - 15 J.                     ***
## newdata$age216 - 17 J.                     ***
## newdata$STALAKleinstädtisch
## newdata$STALAMittelstädtisch               **
## newdata$STALAGroßstädtisch
## newdata$OWWest                             ***
## newdata$e093500 - < 750 €
## newdata$e093750 - < 1.000 €
## newdata$e0931.000 - < 1.250 €
## newdata$e0931.250 - < 1.500 €
## newdata$e0931.500 - < 1.750 €
## newdata$e0931.750 - < 2.000 €
## newdata$e0932.000 - < 2.250 €
## newdata$e0932.250 - < 2.500 €
## newdata$e0932.500 - < 3.000 €
```

```
## newdata$e0933.000 - < 4.000 €
## newdata$e0934.000 - < 5.000 €
## newdata$e093>= 5.000 €
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5838.0  on 11301  degrees of freedom
## Residual deviance: 4336.3  on 11274  degrees of freedom
## AIC: 29707
##
## Number of Fisher Scoring iterations: 5
```

#Conclusion: The variables which are associated are e006B1, age2 and OW. Other variables are not significant for our model. Since the value of AIC is very high, the model is not a good fit as well as 3/6 variables don't add any significance to our model.

Exercise 3: Negative Binomial regression (optional)

Predict the amount of measles vaccinations (Maanzahl) by the number of siblings (e006B1), the sex and age of the children (sex, age2), place of residence (STALA, OW) and the monthly household income (e093), using a Negative Binomial regression model. Interpret the results. Which variables are associated with the outcome? Is the model a good fit to the data?

```
library(MASS)
output2 <- glm.nb(dat$Maanzahl ~ dat$e006B1 + dat$sex + dat$age2 + dat$STALA + dat$OW + dat$e093, data =
```

```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached
```

```
print(summary(output2))
```

```
##
## Call:
## glm.nb(formula = dat$Maanzahl ~ dat$e006B1 + dat$sex + dat$age2 +
##     dat$STALA + dat$OW + dat$e093, data = dat, init.theta = 60862.88291,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2019  -0.4455   0.1187   0.2852   1.8808
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -0.802697   0.108130  -7.423 1.14e-13 ***
## dat$e006B1Ein Geschwisterkind     -0.064634   0.015989  -4.042 5.29e-05 ***
## dat$e006B1Zwei und mehr Geschwister -0.107968   0.023029  -4.688 2.75e-06 ***
## dat$sexWeiblich                    0.003743   0.014637   0.256   0.7982
## dat$age22 - 3 J.                   1.245780   0.058155  21.422  < 2e-16 ***
## dat$age24 - 5 J.                   1.327158   0.057044  23.266  < 2e-16 ***
## dat$age26 - 7 J.                   1.395824   0.056411  24.744  < 2e-16 ***
## dat$age28 - 9 J.                   1.398620   0.056317  24.835  < 2e-16 ***
```

```
## dat$age210 - 11 J.                   1.399978   0.056392  24.826  < 2e-16 ***
## dat$age212 - 13 J.                   1.405903   0.056729  24.783  < 2e-16 ***
## dat$age214 - 15 J.                   1.495007   0.056637  26.396  < 2e-16 ***
## dat$age216 - 17 J.                   1.582862   0.057312  27.618  < 2e-16 ***
## dat$STALAKleinstädtisch              0.024188   0.020733   1.167   0.2434
## dat$STALAMittelstädtisch             0.055111   0.020613   2.674   0.0075 **
## dat$STALAGroßstädtisch               0.005487   0.022412   0.245   0.8066
## dat$OWWest                          -0.154372   0.015873  -9.725  < 2e-16 ***
## dat$e093500 - < 750 €                0.083248   0.104525   0.796   0.4258
## dat$e093750 - < 1.000 €              0.106070   0.099808   1.063   0.2879
## dat$e0931.000 - < 1.250 €            0.041858   0.098242   0.426   0.6701
## dat$e0931.250 - < 1.500 €            0.074999   0.095835   0.783   0.4339
## dat$e0931.500 - < 1.750 €            0.089603   0.096093   0.932   0.3511
## dat$e0931.750 - < 2.000 €            0.095711   0.094875   1.009   0.3131
## dat$e0932.000 - < 2.250 €            0.105828   0.094424   1.121   0.2624
## dat$e0932.250 - < 2.500 €            0.106139   0.094140   1.127   0.2595
## dat$e0932.500 - < 3.000 €            0.077445   0.093487   0.828   0.4074
## dat$e0933.000 - < 4.000 €            0.075623   0.093577   0.808   0.4190
## dat$e0934.000 - < 5.000 €            0.068920   0.096001   0.718   0.4728
## dat$e093>= 5.000 €                   0.050664   0.098174   0.516   0.6058
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(60862.88) family taken to be 1)
##
##     Null deviance: 5837.9  on 11301  degrees of freedom
## Residual deviance: 4336.3  on 11274  degrees of freedom
##   (6338 observations deleted due to missingness)
## AIC: 29709
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  60863
##           Std. Err.:  68935
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -29651.28
```

#Conclusion: Same as above