MLE01 - Vertiefende statistische Verfahren

2. Übungsblatt SS 2024

Allgemeine Information

Alle Aufgaben sind mit R zu lössen. Die Berechnungen sollen nachvollziehbar und dokumentiert sein. Um die vollständige Punktezahl zu erreichen, müssen alle Ergebnisse und Fragen entsprechend interpretiert bzw. beantwortet werden. Code alleine ist nicht ausreichend! Die Abgabe erfolgt über Moodle entsprechend der Abgaberichtlinien als pdf und Rmd File. Bitte inkludieren Sie namentlich alle beteiligten Gruppenmitglieder sowohl im Bericht als auch im Source Code. Die jeweiligen Datensätze die für diese Übung relevant sind finden Sie ebenfalls in Moodle.

1 Logistische Regressionsanalyse [2P]

Verwenden Sie den Datensatz `birthwt.xlsx`. Dieser Datensatz bezieht sich auf Risikofaktoren im Zusammenhang mit niedrigem Geburtsgewicht von Säuglingen. Eine Beschreibung der einzelnen Variablen entnehmen Sie bitte dem Excel-File.

```
library(readxl)
birthwt <- read_excel("birthwt.xlsx")

# data structure
str(birthwt)
```

```
## tibble [189 x 9] (S3: tbl_df/tbl/data.frame)
##  $ low  : num [1:189] 0 0 0 0 0 0 0 0 0 0 ...
##  $ age  : num [1:189] 19 33 20 21 18 21 22 17 29 26 ...
##  $ lwt  : num [1:189] 182 155 105 108 107 124 118 103 123 113 ...
##  $ race : num [1:189] 2 3 1 1 1 3 1 3 1 1 ...
##  $ smoke: num [1:189] 0 0 1 1 1 0 0 0 1 1 ...
##  $ ptl  : num [1:189] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ht   : num [1:189] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ui   : num [1:189] 1 0 0 1 1 0 0 0 0 0 ...
##  $ ftv  : num [1:189] 0 3 1 2 0 0 1 1 1 0 ...
```

```
summary(birthwt)
```

```
##       low              age             lwt             race      
##  Min.   :0.0000   Min.   :14.00   Min.   : 80.0   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   1st Qu.:1.000  
##  Median :0.0000   Median :23.00   Median :121.0   Median :1.000  
##  Mean   :0.3122   Mean   :23.24   Mean   :129.8   Mean   :1.847  
```

```
##  3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :45.00   Max.   :250.0   Max.   :3.000
##      smoke             ptl               ht                ui
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.3915   Mean   :0.1958   Mean   :0.06349   Mean   :0.1481
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :3.0000   Max.   :1.00000   Max.   :1.0000
##       ftv
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.7937
##  3rd Qu.:1.0000
##  Max.   :6.0000
```

```r
# check for missing or inplausible values
sum(is.na(birthwt)+sum(birthwt$age == 0)+sum(birthwt$lwt == 0))
```

```
## [1] 0
```

Beschreibung der Variablen:

low: low birth weight (<2500g) yes/no age: mother's age in years lwt: mother's weight in pounds at last menstrual period race: Skincolour of the mother (1=white, 2=black, 3=other) smoke: smoking status during pregnancy (1=smoker, 0=non-smoker) ptl: number of episodes of premature labours ht: history of hypertension (1=yes, 0=no) ui: presence of uterine irritability (1=yes, 0=no) ftv: number of physician visits during the first trimester

```r
# transform into factors and set levels
birthwt$low[birthwt$low==0] = "neg"
birthwt$low[birthwt$low==1] = "pos"
birthwt$low <- as.factor(birthwt$low)
birthwt$race <- as.factor(birthwt$race)
birthwt$smoke <- as.factor(birthwt$smoke)
birthwt$ht <- as.factor(birthwt$ht)
birthwt$ui <- as.factor(birthwt$ui)
```

[1P] a: Erstellen Sie ein Modell, welches das Risiko für niedriges Geburtsgewicht (`low`; Gewicht <2500g ja/nein) in Abhängigkeit verschiedener Faktoren beschreibt. Wie lautet die Modellgleichung?

```r
# full logistic regression model
model <- glm(low ~ age + lwt + race + smoke + ptl + ht + ui + ftv,
             data = birthwt, family = binomial)

summary(model)
```

```
##
## Call:
## glm(formula = low ~ age + lwt + race + smoke + ptl + ht + ui +
##     ftv, family = binomial, data = birthwt)
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.480623   1.196888   0.402  0.68801
## age         -0.029549   0.037031  -0.798  0.42489
## lwt         -0.015424   0.006919  -2.229  0.02580 *
## race2        1.272260   0.527357   2.413  0.01584 *
## race3        0.880496   0.440778   1.998  0.04576 *
## smoke1       0.938846   0.402147   2.335  0.01957 *
## ptl          0.543337   0.345403   1.573  0.11571
## ht1          1.863303   0.697533   2.671  0.00756 **
## ui1          0.767648   0.459318   1.671  0.09467 .
## ftv          0.065302   0.172394   0.379  0.70484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 201.28  on 179  degrees of freedom
## AIC: 221.28
##
## Number of Fisher Scoring iterations: 4
```

```
confint(model) # 95% confidence intervals
```

```
## Es wird auf das Profilieren gewartet ...
```

```
##                    2.5 %      97.5 %
## (Intercept) -1.84121370  2.87745202
## age         -0.10373422  0.04207540
## lwt         -0.02978452 -0.00246483
## race2        0.24166064  2.32608774
## race3        0.02661178  1.76511921
## smoke1       0.16158429  1.74790611
## ptl         -0.12346116  1.24603059
## ht1          0.53239257  3.32119843
## ui1         -0.14356295  1.67090307
## ftv         -0.28308378  0.39881567
```

Basierend auf den Koeffizienten des Modells, sieht die Modellgleichung wie folgt aus:

log(Odds) = 0.48 - 0.03 * age - 0.02 * lwt + 1.27 * race2 + 0.88 * race3 + 0.94 * smoke + 0.54 * ptl + 1.86 * ht + 0.77 * ui + 0.07 * ftv

```
library(MASS)
```

```
##
## Attache Paket: 'MASS'
```

```
## Das folgende Objekt ist maskiert durch '.GlobalEnv':
##
##     birthwt
```

```
## Das folgende Objekt ist maskiert 'package:dplyr':
##
##     select
```

```
# find best model with stepwise selection
m2<-stepAIC(model,direction = "backward")
```

```
## Start:  AIC=221.28
## low ~ age + lwt + race + smoke + ptl + ht + ui + ftv
##
##         Df Deviance    AIC
## - ftv    1   201.43 219.43
## - age    1   201.93 219.93
## <none>       201.28 221.28
## - ptl    1   203.83 221.83
## - ui     1   204.03 222.03
## - race   2   208.75 224.75
## - lwt    1   206.80 224.80
## - smoke  1   206.91 224.91
## - ht     1   208.81 226.81
##
## Step:  AIC=219.43
## low ~ age + lwt + race + smoke + ptl + ht + ui
##
##         Df Deviance    AIC
## - age    1   201.99 217.99
## <none>       201.43 219.43
## - ptl    1   203.95 219.95
## - ui     1   204.11 220.11
## - race   2   208.77 222.77
## - lwt    1   206.81 222.81
## - smoke  1   206.92 222.92
## - ht     1   208.81 224.81
##
## Step:  AIC=217.99
## low ~ lwt + race + smoke + ptl + ht + ui
##
##         Df Deviance    AIC
## <none>       201.99 217.99
## - ptl    1   204.22 218.22
## - ui     1   204.90 218.90
## - smoke  1   207.73 221.73
## - lwt    1   208.11 222.11
## - race   2   210.31 222.31
## - ht     1   209.46 223.46
```

```
m3<-stepAIC(model,direction = "forward")
```

```
## Start:  AIC=221.28
## low ~ age + lwt + race + smoke + ptl + ht + ui + ftv
```

```r
m4<-stepAIC(model,direction = "both")
```

```
## Start:  AIC=221.28
## low ~ age + lwt + race + smoke + ptl + ht + ui + ftv
##
##         Df Deviance    AIC
## - ftv    1   201.43 219.43
## - age    1   201.93 219.93
## <none>       201.28 221.28
## - ptl    1   203.83 221.83
## - ui     1   204.03 222.03
## - race   2   208.75 224.75
## - lwt    1   206.80 224.80
## - smoke  1   206.91 224.91
## - ht     1   208.81 226.81
##
## Step:  AIC=219.43
## low ~ age + lwt + race + smoke + ptl + ht + ui
##
##         Df Deviance    AIC
## - age    1   201.99 217.99
## <none>       201.43 219.43
## - ptl    1   203.95 219.95
## - ui     1   204.11 220.11
## + ftv    1   201.28 221.28
## - race   2   208.77 222.77
## - lwt    1   206.81 222.81
## - smoke  1   206.92 222.92
## - ht     1   208.81 224.81
##
## Step:  AIC=217.99
## low ~ lwt + race + smoke + ptl + ht + ui
##
##         Df Deviance    AIC
## <none>       201.99 217.99
## - ptl    1   204.22 218.22
## - ui     1   204.90 218.90
## + age    1   201.43 219.43
## + ftv    1   201.93 219.93
## - smoke  1   207.73 221.73
## - lwt    1   208.11 222.11
## - race   2   210.31 222.31
## - ht     1   209.46 223.46
```

Ein Vergleich der Modelle mit dem Rückwärts-, Vorwärts- und beidseitigen Auswahlverfahren zeigt, dass ein reduziertes Modell der folgenden Form, aufgrund des vergleichbar kleinsten AIC-Werts am besten geeignet wäre um das Risiko für niedriges Geburtsgewicht zu beschreiben: low ~ lwt + race + smoke + ptl + ht + ui. Hierbei wurden die im ursprünglich enthaltenen Modell enthaltenen Variablen age und ftv entfernt.

```r
# reduced model
model_reduced <- glm(low ~ lwt + race + smoke + ptl + ht + ui,
             data = birthwt, family = binomial)
```

```r
summary(model_reduced)
```

```
##
## Call:
## glm(formula = low ~ lwt + race + smoke + ptl + ht + ui, family = binomial,
##     data = birthwt)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.086550   0.951760  -0.091  0.92754
## lwt         -0.015905   0.006855  -2.320  0.02033 *
## race2        1.325719   0.522243   2.539  0.01113 *
## race3        0.897078   0.433881   2.068  0.03868 *
## smoke1       0.938727   0.398717   2.354  0.01855 *
## ptl          0.503215   0.341231   1.475  0.14029
## ht1          1.855042   0.695118   2.669  0.00762 **
## ui1          0.785698   0.456441   1.721  0.08519 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 201.99  on 181  degrees of freedom
## AIC: 217.99
##
## Number of Fisher Scoring iterations: 4
```

```r
# compare moadels
anova(model, model_reduced)
```

```
## Analysis of Deviance Table
##
## Model 1: low ~ age + lwt + race + smoke + ptl + ht + ui + ftv
## Model 2: low ~ lwt + race + smoke + ptl + ht + ui
##   Resid. Df Resid. Dev Df Deviance
## 1       179     201.28
## 2       181     201.99 -2 -0.70079
```

Die Differenz in der Deviance zwischen den beiden Modellen beträgt -0.70079, was darauf hinweist, dass das reduzierte Modell nahezu das gleiche Erklärungspotential wie das volle Modell hat. Das reduzierte Modell besitz jedoch einen geringeren AIC-Wert (217.99) im Vergleich zum vollen Modell (221.28), was euf eine bessere Balance zwischen Modellkomplexität und Passgenauigkeit aufweist.

[1P] b: Überprüfen Sie die Modellvoraussetzungen und bewerten Sie die Güte des Modells. Wie hoch ist die Wahrscheinlichkeit für eine Geburt mit einem Geburtsgewicht <2500g, bei folgenden Daten der Mutter: 38 Jahre alt, 68 kg, weiß, Nichtraucher, 2 Vorgeburten (ptl), keinen Bluthochdruck, keine Reizung der Gebärmutter, ein Arztbesuch im 1. Trimester.

```r
# odds of coefficients
odds <- exp(model$coefficients)
exp(confint(model))
```

```
## Es wird auf das Profilieren gewartet ...
```

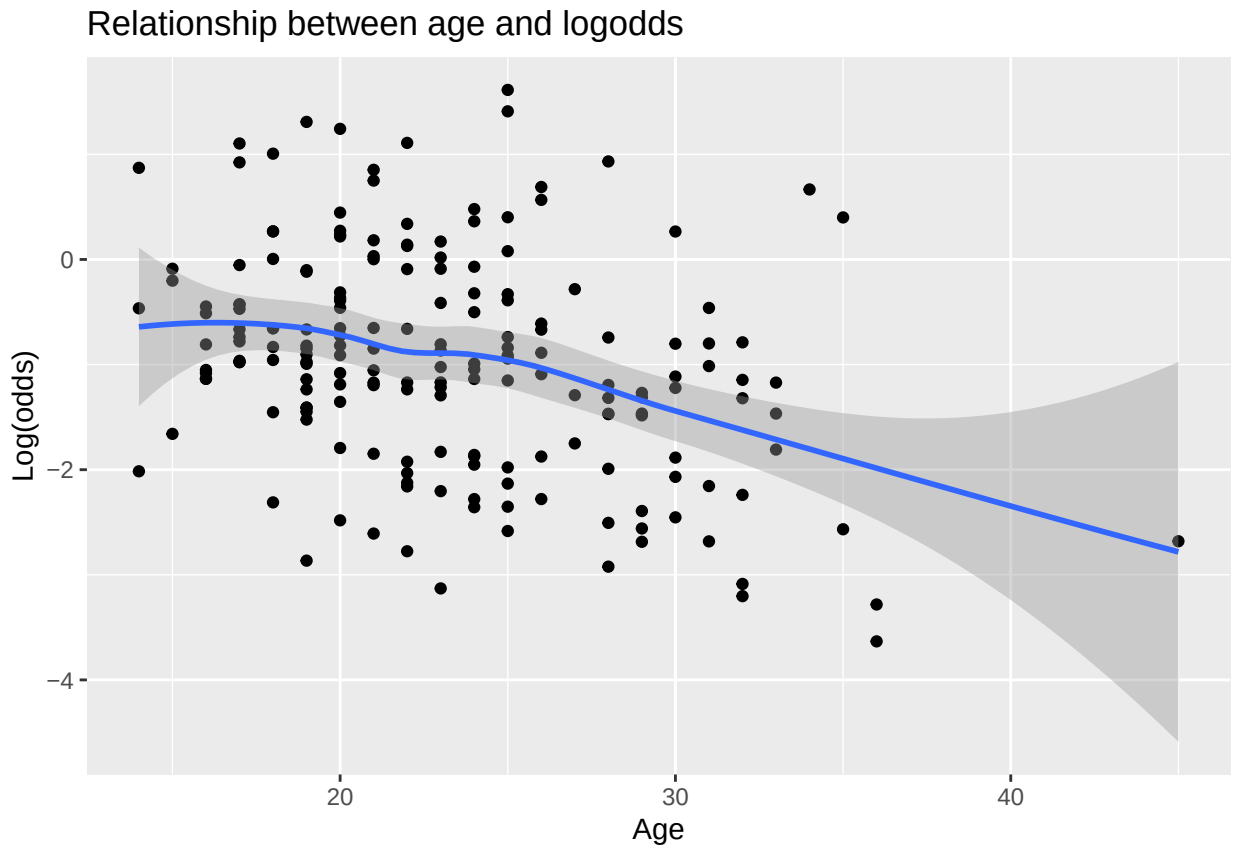
```
##                 2.5 %     97.5 %
## (Intercept) 0.1586248 17.7689406
## age         0.9014649  1.0429731
## lwt         0.9706547  0.9975382
## race2       1.2733620 10.2378101
## race3       1.0269690  5.8422688
## smoke1      1.1753715  5.7425658
## ptl         0.8838560  3.4765158
## ht1         1.7030020 27.6935195
## ui1         0.8662663  5.3169672
## ftv         0.7534567  1.4900589
```

```
# data for evaluation
test.data<- birthwt

# data for prediction
pred.data <- data.frame(age = 38,
                        lwt = 68,
                        race = 1,
                        smoke = 0,
                        ptl = 2,
                        ht = 0,
                        ui = 0,
                        ftv = 1)
# convert kg to pounds
pred.data$lwt <- pred.data$lwt * 2.20462

# transform factors
pred.data$race <- as.factor(pred.data$race)
pred.data$smoke <- as.factor(pred.data$smoke)
pred.data$ht <- as.factor(pred.data$ht)
pred.data$ui <- as.factor(pred.data$ui)

library(tidyverse)

# predict probabilities
probabilities <- model %>% predict(test.data, type = "response")

# calculate logodds
logodds <- model %>% predict(test.data, type = "link")

# predict classes
predicted.classes <- ifelse(probabilities > 0.5, "pos", "neg")

# Accuracy of the model
mean(predicted.classes == test.data$low)
```

```
## [1] 0.7407407
```

```
op<-data.frame(test.data,
               logodds=logodds,
               odds=exp(logodds),
               probabilities=probabilities,
               predicted.classes=predicted.classes)


# check for linearity between logodds and metric predictor values

# age vs logodds
ggplot(op, aes(x = age, y = logodds)) +
  geom_point() +
  geom_smooth(method = "loess", se = T) +
  labs(title = "Relationship between age and logodds",
       x = "Age",
       y = "Log(odds)")
```

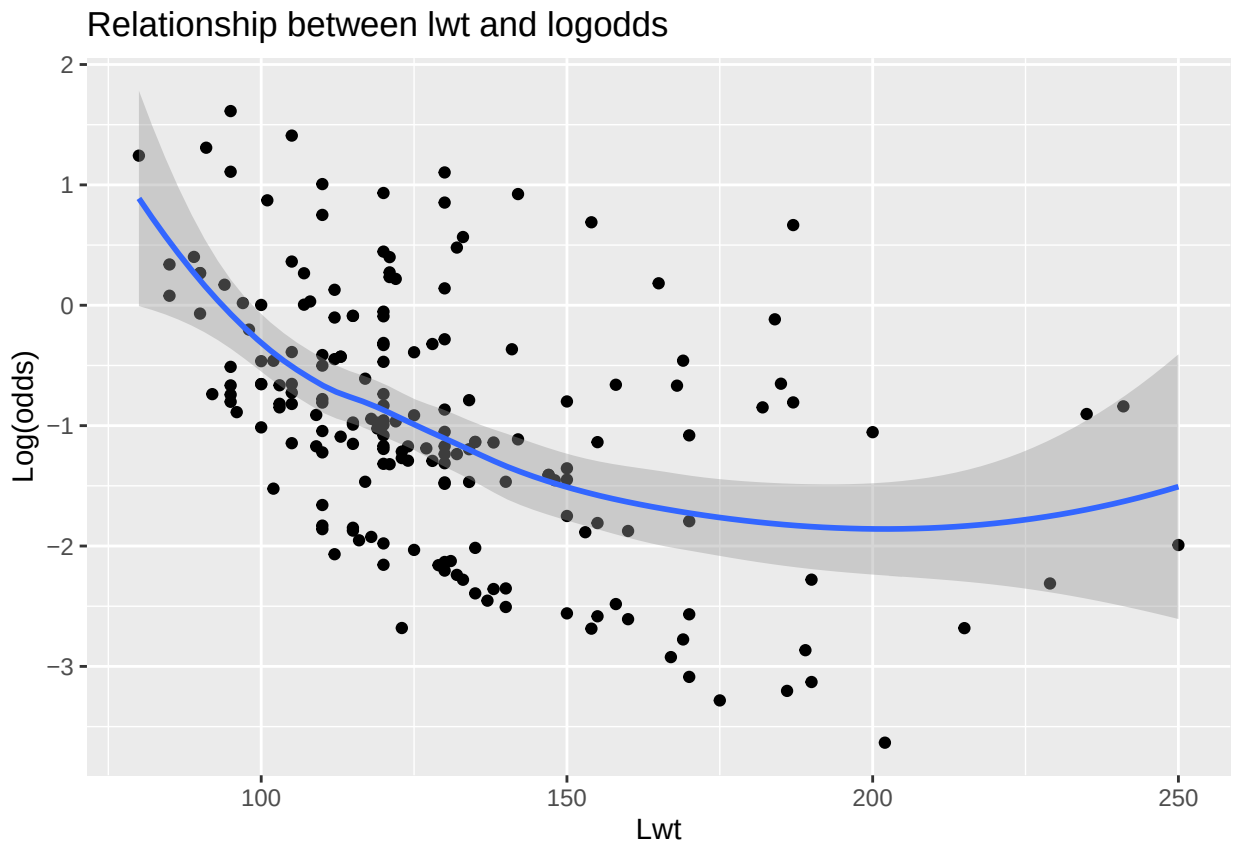
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# lwt vs logodds
ggplot(op, aes(x = lwt, y = logodds)) +
  geom_point() +
  geom_smooth(method = "loess", se = T) +
  labs(title = "Relationship between lwt and logodds",
```

```
       x = "Lwt",
       y = "Log(odds)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# ptl vs logodds
ggplot(op, aes(x = ptl, y = logodds)) +
  geom_point() +
  geom_smooth(method = "loess", se = T) +
  labs(title = "Relationship between ptl and logodds",
       x = "Ptl",
       y = "Log(odds)")
```

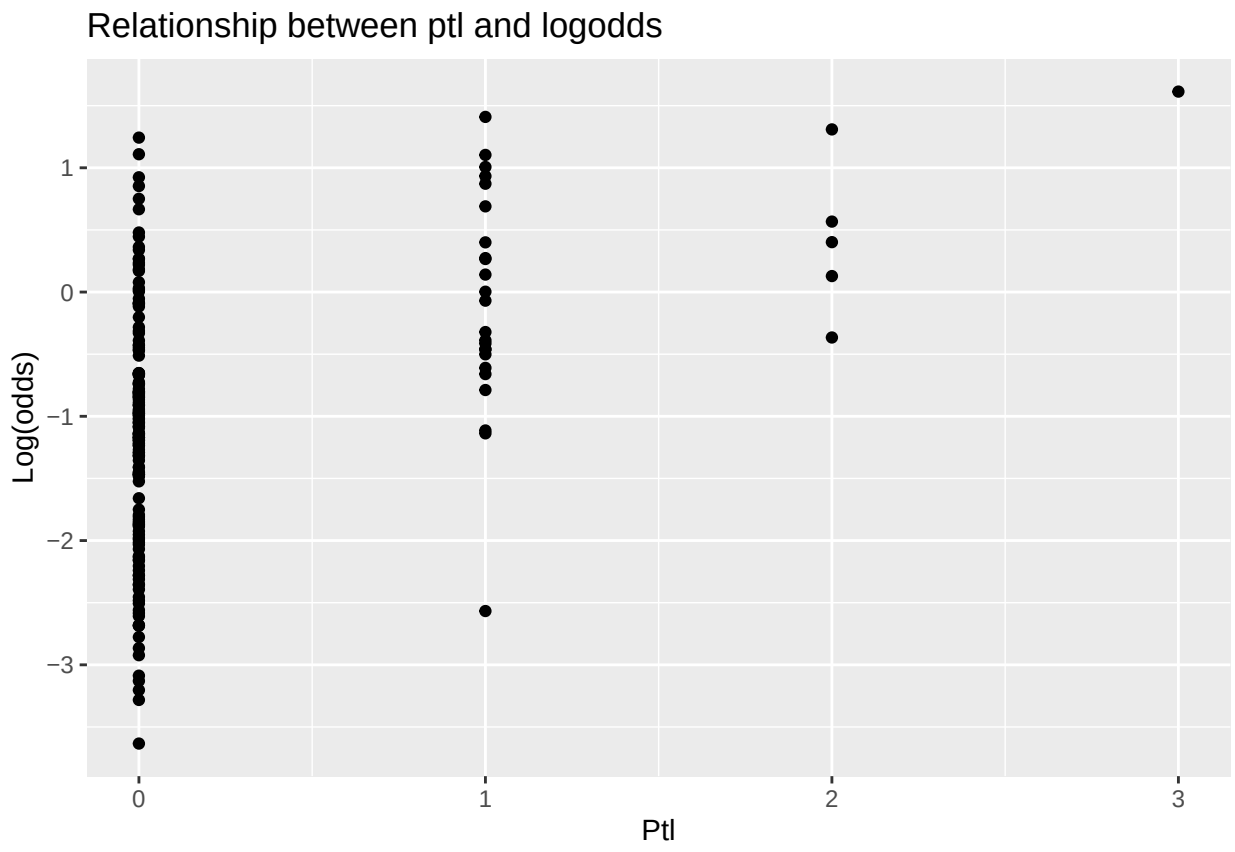
```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : at -0.015
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : radius 0.000225
```

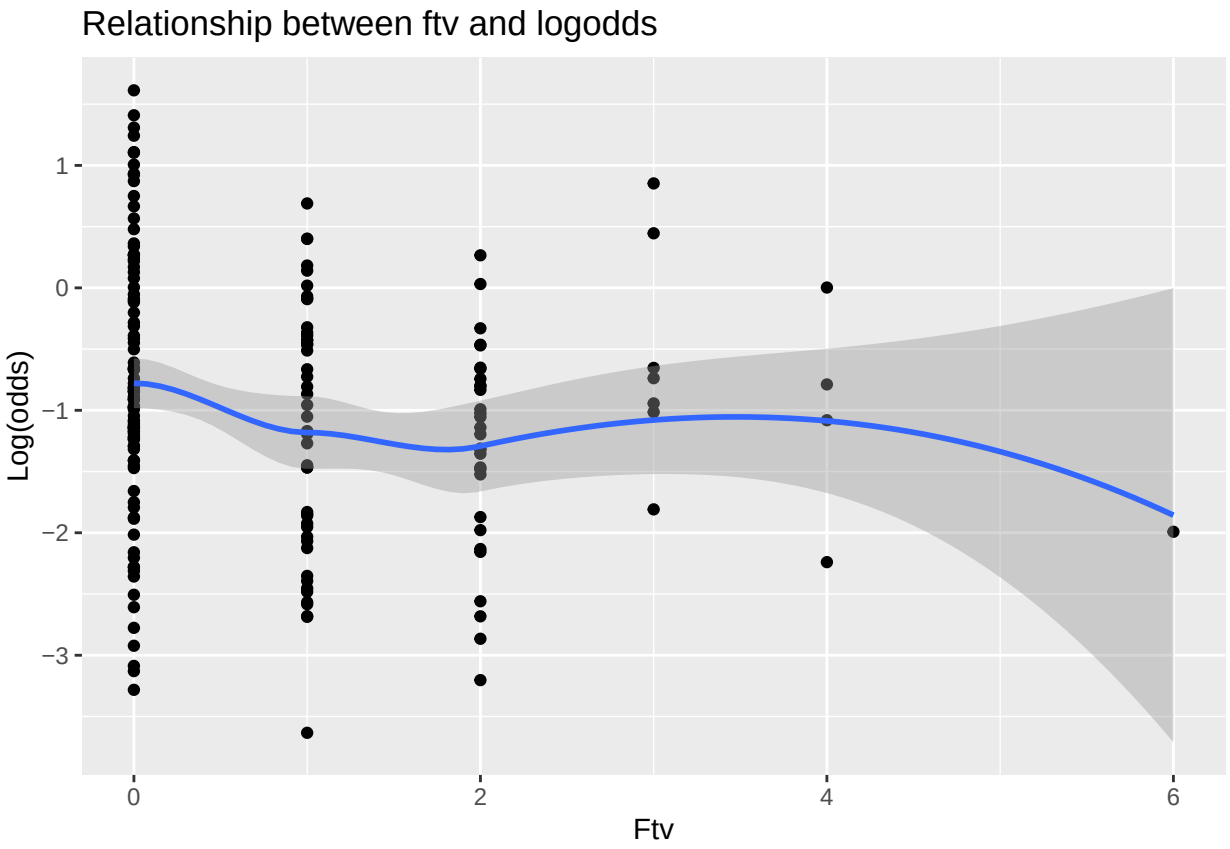
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : all data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at -0.015

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 0.015

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : zero-width neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 1

## Warning: Computation failed in `stat_smooth()`
## Caused by error in `predLoess()`:
## ! NA/NaN/Inf in externem Funktionsaufruf (arg 5)
```

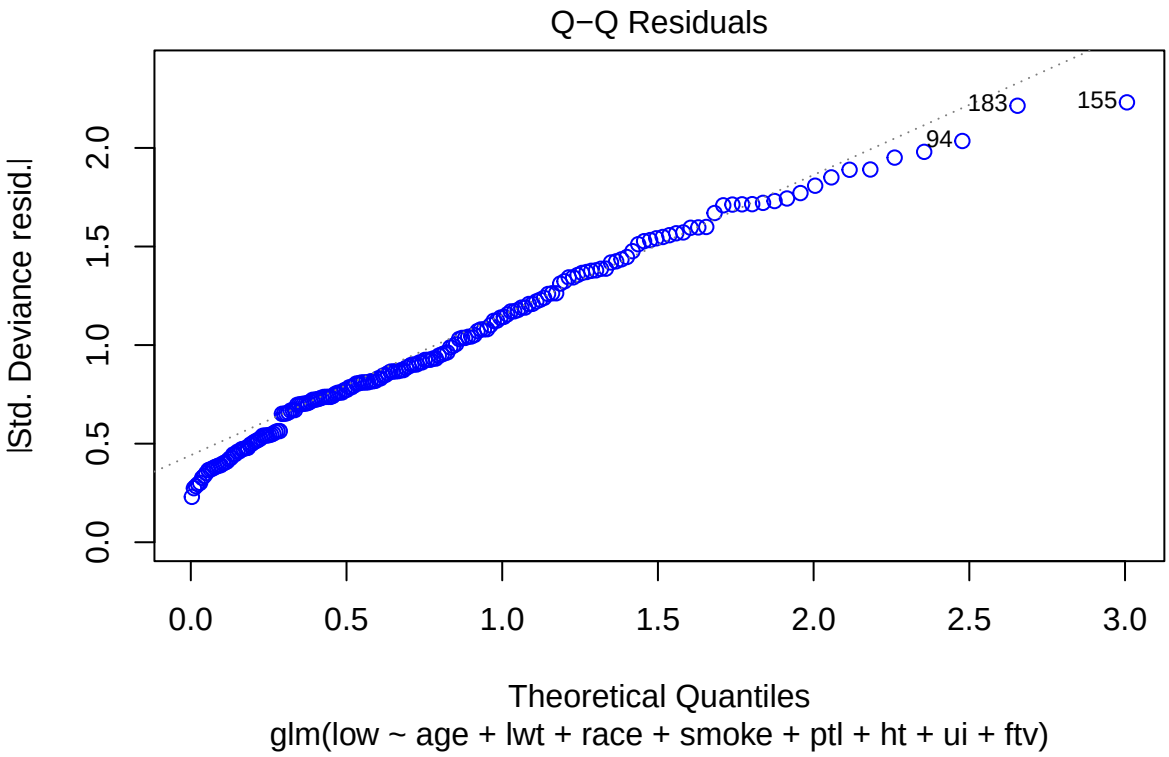



```
# ftv vs logodds
ggplot(op, aes(x = ftv, y = logodds)) +
  geom_point() +
  geom_smooth(method = "loess", se = T) +
  labs(title = "Relationship between ftv and logodds",
       x = "Ftv",
       y = "Log(odds)")
```

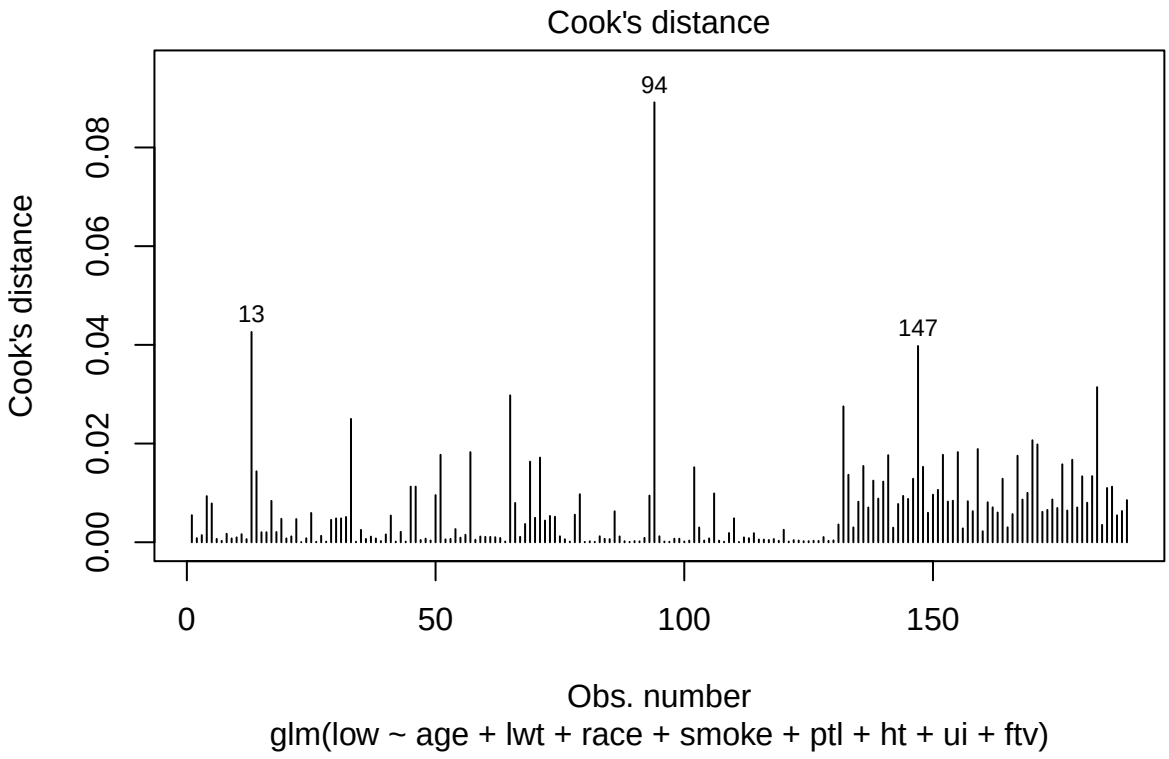
```
## `geom_smooth()` using formula = 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at -0.03

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 1.03

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 1

## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## -0.03

## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 1.03

## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0

## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 1
```




```
# Evaluate model quality

#normalverteilung der residuen
plot(model, which=2, col=c("blue")) # Normal Q-Q Plot
```




```
#influential values: cook's distance
plot(model, which = 4, id.n = 3)
```




```
# check for multicollinearity
library(car)
vif(model)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## age   1.100003  1        1.048810
## lwt   1.303489  1        1.141704
## race  1.510253  2        1.108568
## smoke 1.348243  1        1.161139
## ptl   1.087847  1        1.042999
## ht    1.168419  1        1.080934
## ui    1.063061  1        1.031048
## ftv   1.087144  1        1.042662
```

```
# G?tekriterium
AIC(model)
```

```
## [1] 221.2848
```

```
BIC(model)
```

```
## [1] 253.7023
```

```r
library(pscl)
```

```
## Warning: Paket 'pscl' wurde unter R Version 4.3.3 erstellt
```

```
## Classes and Methods for R originally developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University (2002-2015),
## by and under the direction of Simon Jackman.
## hurdle and zeroinfl functions by Achim Zeileis.
```

```r
#mcfadden R2
pR2(model)
```

```
## fitting null model for pseudo-r2
```

```
##          llh      llhNull           G2     McFadden         r2ML         r2CU
## -100.6423975 -117.3359981   33.3872011    0.1422718    0.1619285    0.2277177
```

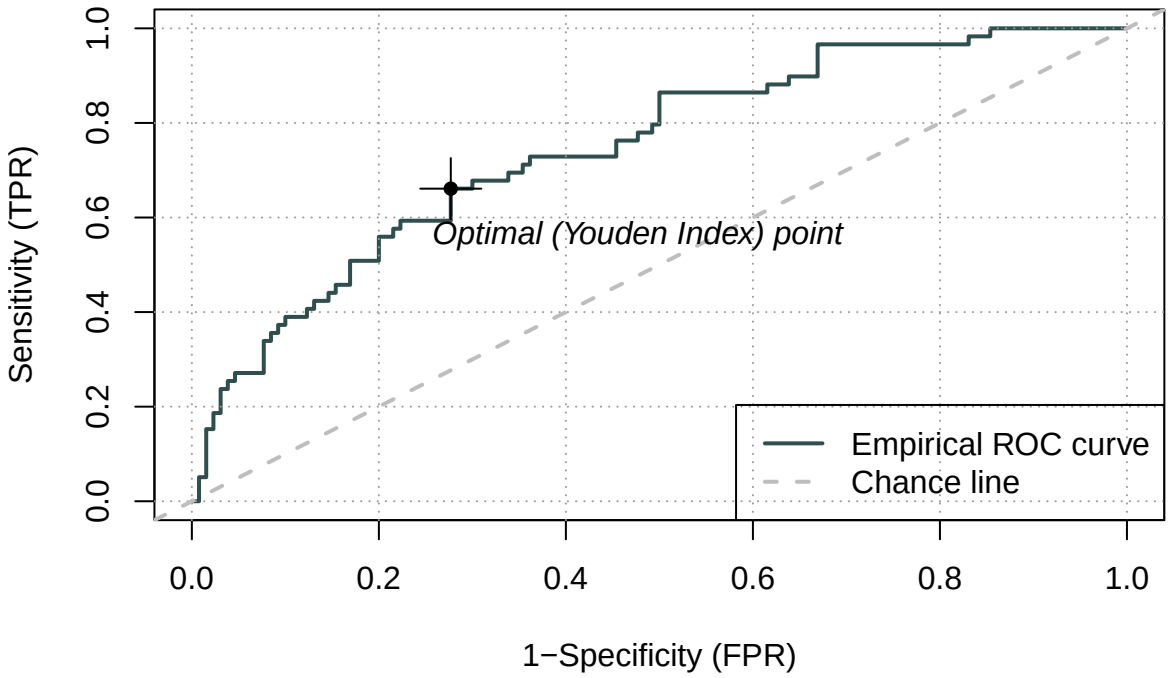
```r
#ROC curve
library(ROCit)
```

```
## Warning: Paket 'ROCit' wurde unter R Version 4.3.3 erstellt
```

```
##
## Attache Paket: 'ROCit'
```

```
## Das folgende Objekt ist maskiert 'package:car':
##
##     logit
```

```r
ROCit_obj <- rocit(score=probabilities,class=birthwt$low)
plot(ROCit_obj)
```



Beurteilung der Modellgüte: Das Modell hat eine Genauigkeit von 0.75, was bedeutet, dass 75% der Vorhersagen korrekt sind. Die Residuen sind normalverteilt und es gibt keine starken Ausreißer. Das Modell hat einen McFadden Pseudo-R2 Wert von 0.14, was darauf hindeutet, dass das Modell nur etwa 14% der Varianz erklärt. Der AIC-Wert beträgt 221 und der BIC-Wert knapp 254. Die VIF-Werte sind alle unter 1.6, was keine erhöhte Multikollinearität nahelegt.

```
# probability prediction for birth weight < 2500g
pred.response <- predict(model, newdata = pred.data, type = "response")

# log odds prediction for birth weight < 2500g
pred.logodds <- predict(model, newdata = pred.data, type = "link")

print(paste("Probability of birth weight < 2500g:", pred.response))
```

```
## [1] "Probability of birth weight < 2500g: 0.141536990497336"
```

Interpretation der Ergebnisse:

2 Logistische Regressionsanalyse [3P]

Verwenden Sie erneut die Framingham-Herz-Studiendaten in `Framingham.sav`. Die Variable `mi_fchd` beschreibt Patienten die einen hospitalisierten Myokardinfarkt oder eine tödliche koronare Herzkrankheit erlitten haben.

[**1.5P**] **a:** Erstellen Sie ein Modell, welches das Risiko für `mi_fchd` in Abhängigkeit von verschiedenen Faktoren beschreibt. Vermeiden Sie nicht relevante bzw. redundante Variablen. Achten Sie auf Ausreißer und fehlende Daten (`NaN, NA's`).

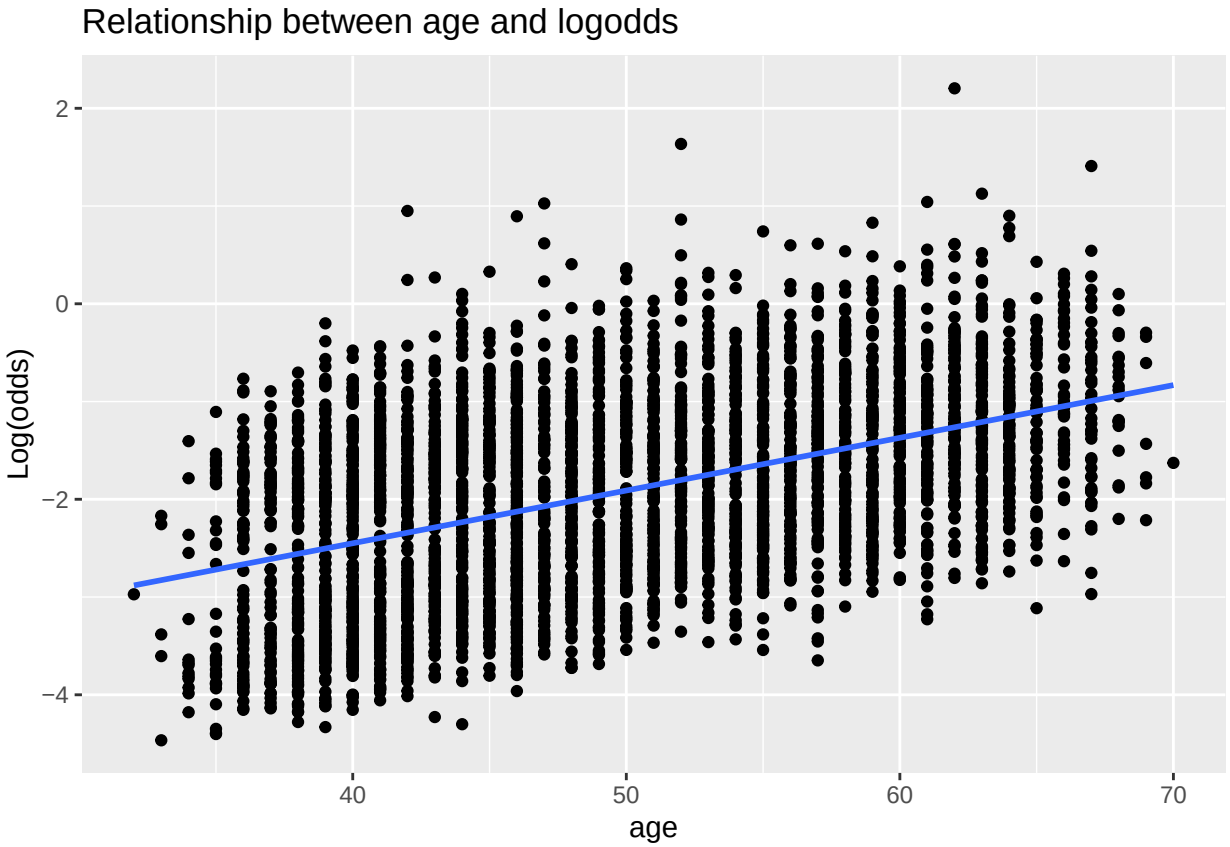
[**1P**] **b:** Überprüfen Sie die Modellvorraussetzungen und evaluieren Sie die Performance des finalen Modells hinsichtlich Genauigkeit und AUROC. Vergleichen Sie die Genauigkeit des Modells mit einem Klassifizierungsmodell, dass immer "no" vorhersagt.

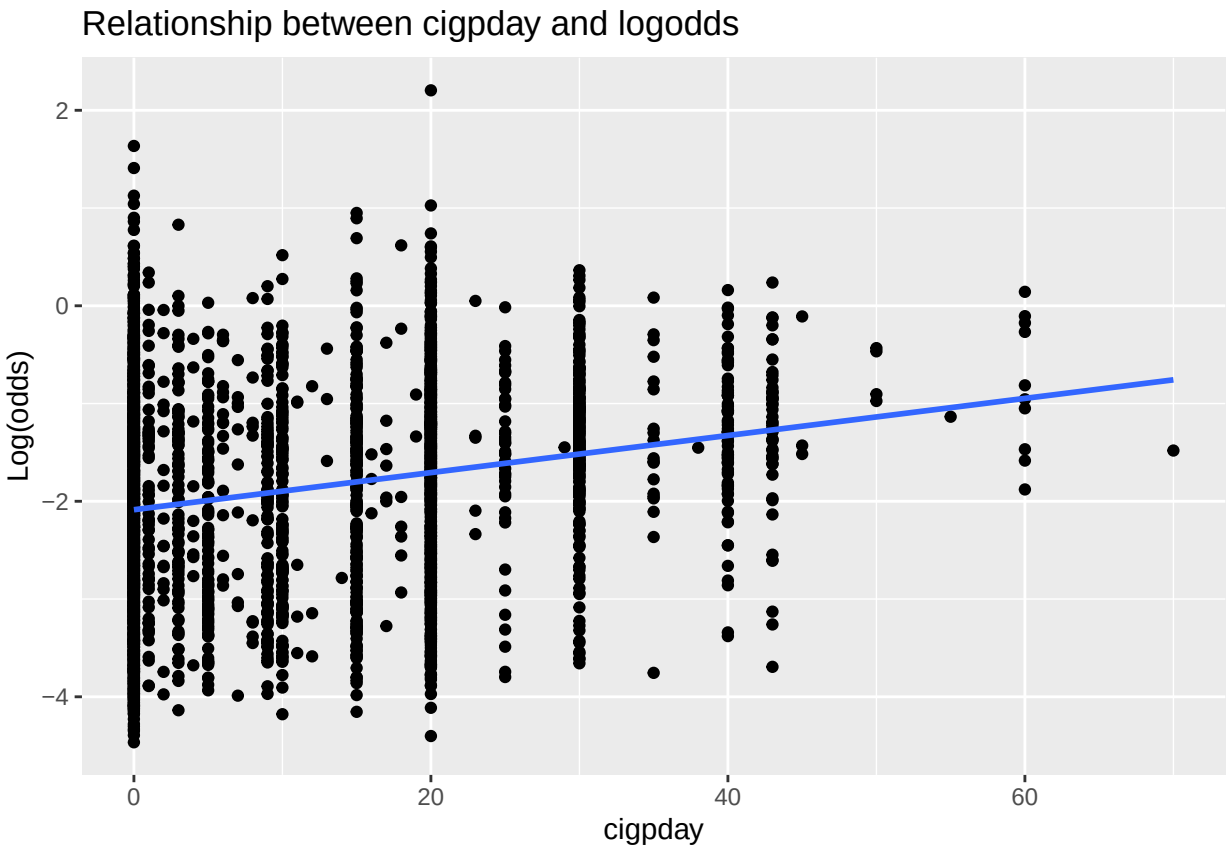
[**0.5P**] **b:** Interpretieren Sie die OR. Welche Aussagen können Sie bezüglich Risikofaktoren für eine Hospitalisierung durch Myokardinfarkt bzw. tödliche koronare Herzkrankheit treffen? Was sind die Top3 Variablen mit dem stärksten Einfluss auf die Zielvariable?

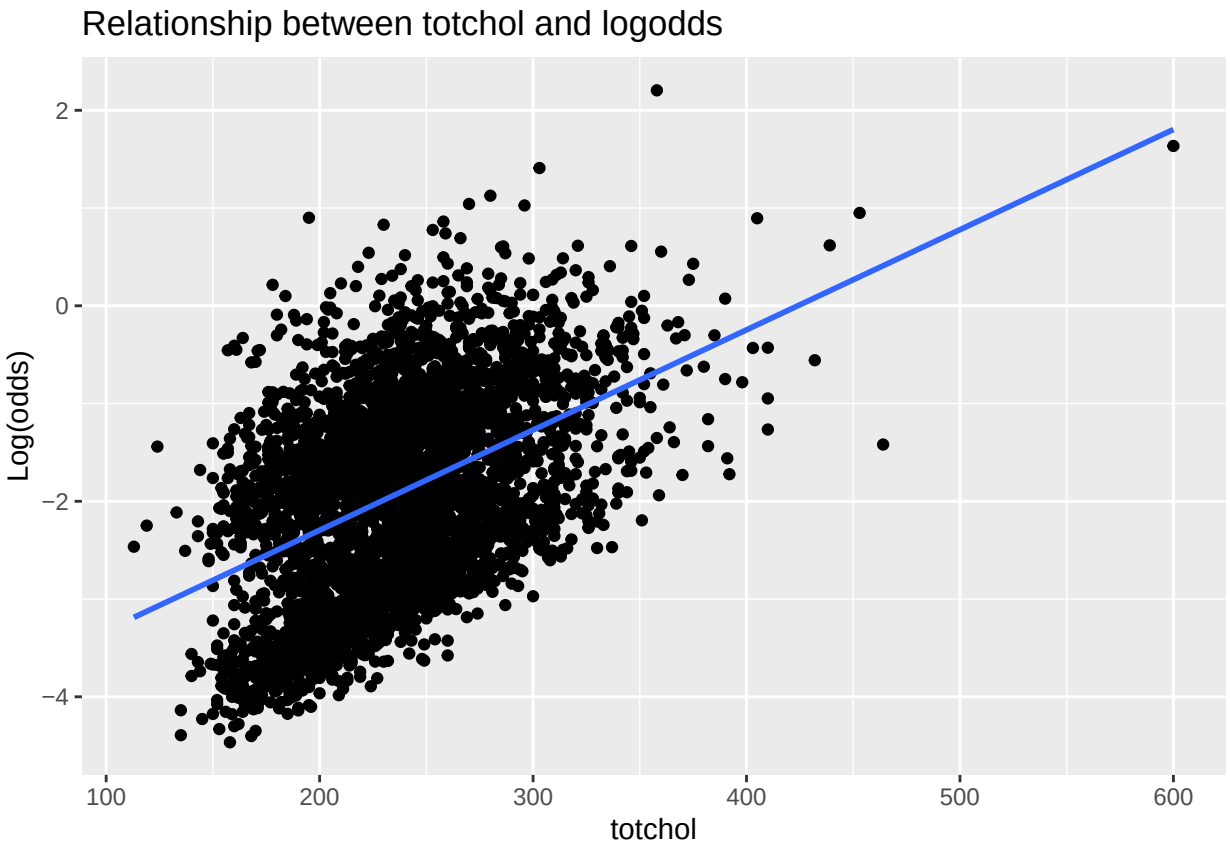
Das volle Modell wird erstellt und begutachtet. Keine Beobachtung mit Cook Distance übe 0.5. Keine Residuen > 3.

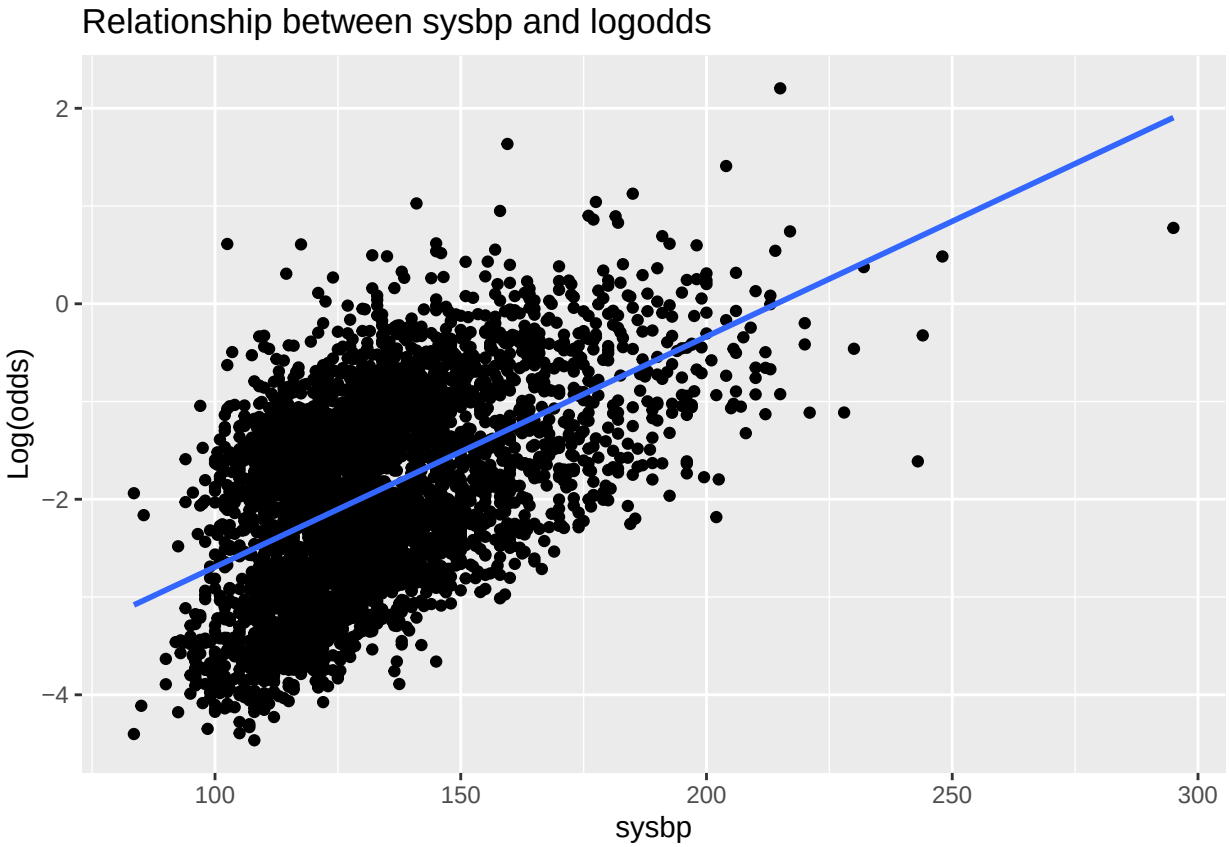
```
load_source()

# Read in framingham data
data.full <- load_and_prepare_framingham("Framingham.sav")

# Analysis of Full Model

glm.full <- run_analysis_of_full_model(data.full)
```

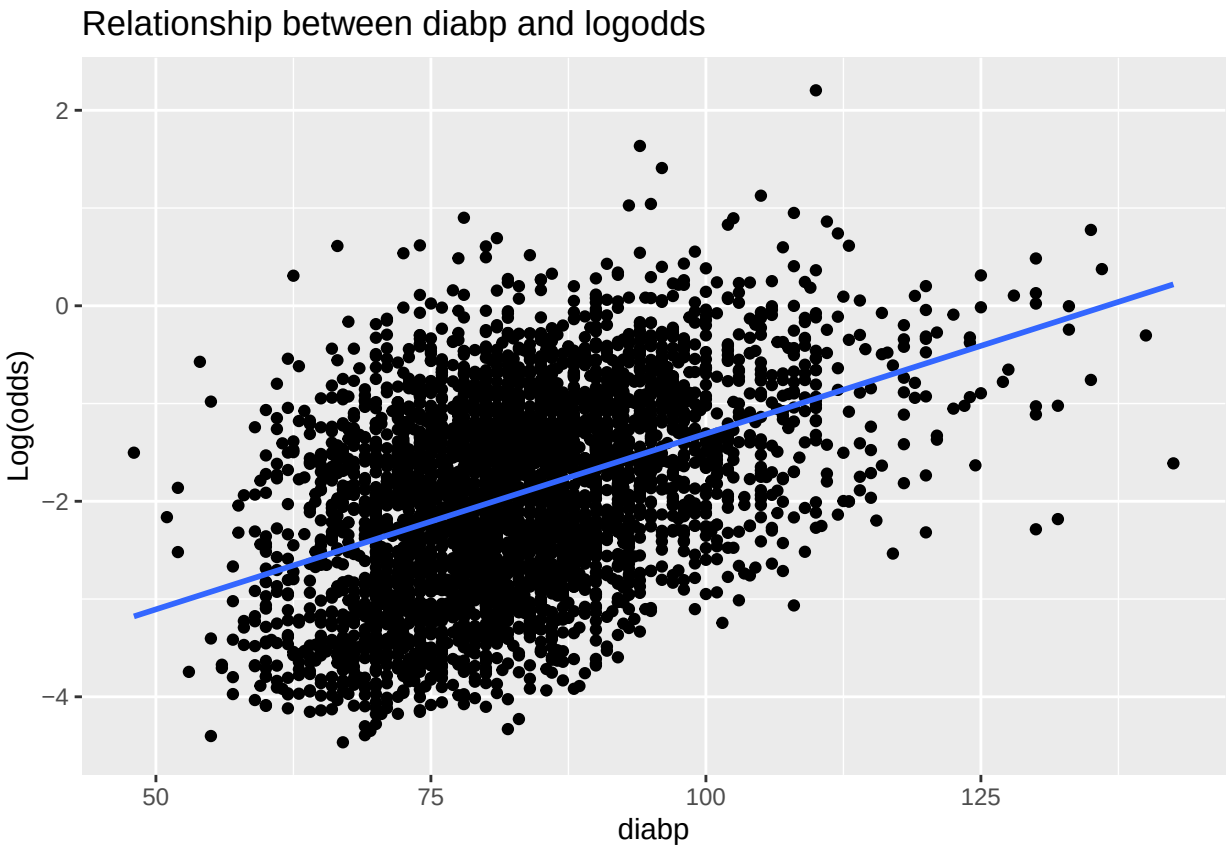
```
##  (Intercept)         sex2          age        educ2        educ3        educ4
## 0.0005581131 0.2608244332 1.0278607087 1.0992992521 0.9922511812 1.1288694823
##    cursmoke1      cigpday      bpmeds1      totchol        sysbp        diabp
## 1.3743945666 1.0000175318 1.2047538542 1.0078738059 1.0181965793 0.9947239088
##          bmi    diabetes1     heartrte      glucose
## 1.0346335955 2.0568579276 0.9972054194 1.0028331702
## [1] "R^2:  0.126198117763004"
## [1] "Goodness of fit:  0.04101093841503"
## [1] "AUROC:  0.752820628296551"
## [1] "Sensitivity: 0.986511919698871"
## [1] "Specificity: 0.0705329153605016"
## [1] "Correctly classified: 83.3768949294302 %"
## [1] "Falsely classified: 16.6231050705698 %"
```

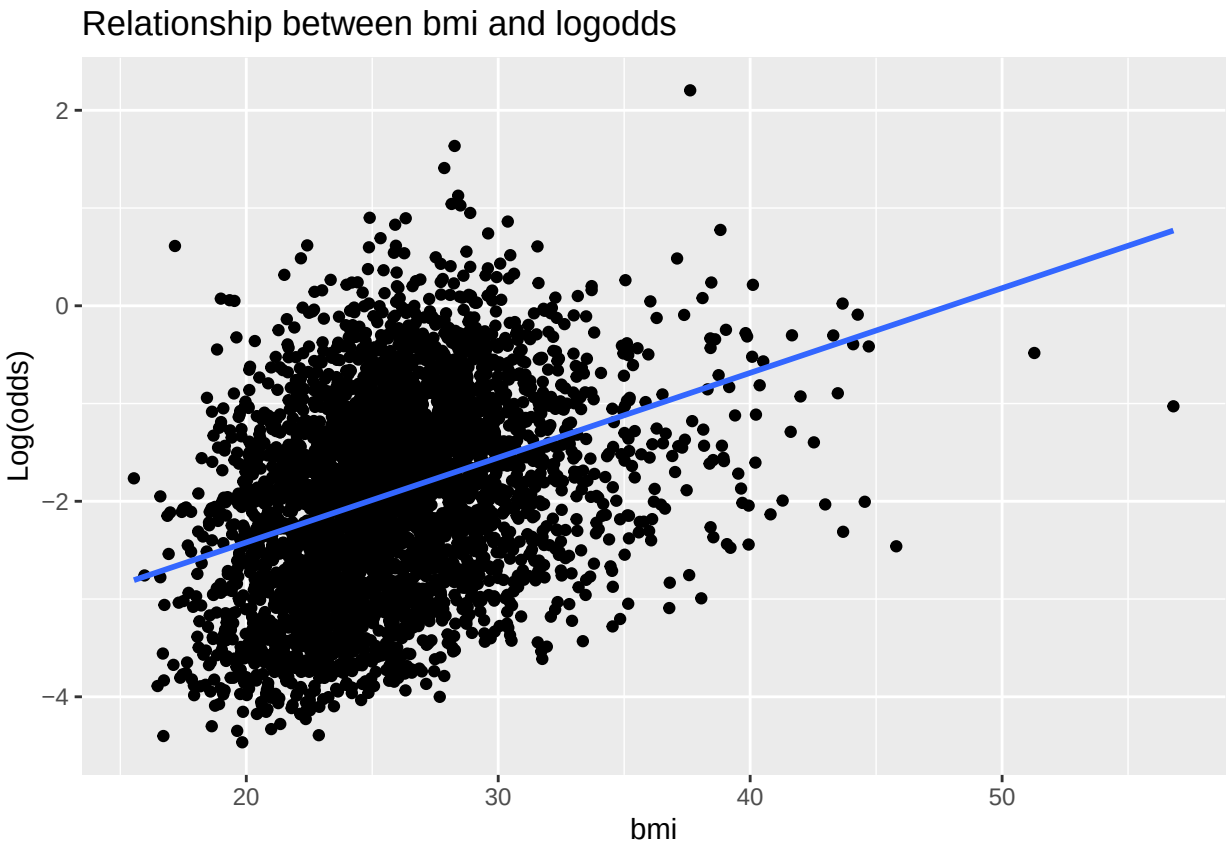

Relationship between age and logodds
Log(odds)
2
0
−2
−4
40
50
60
70
age

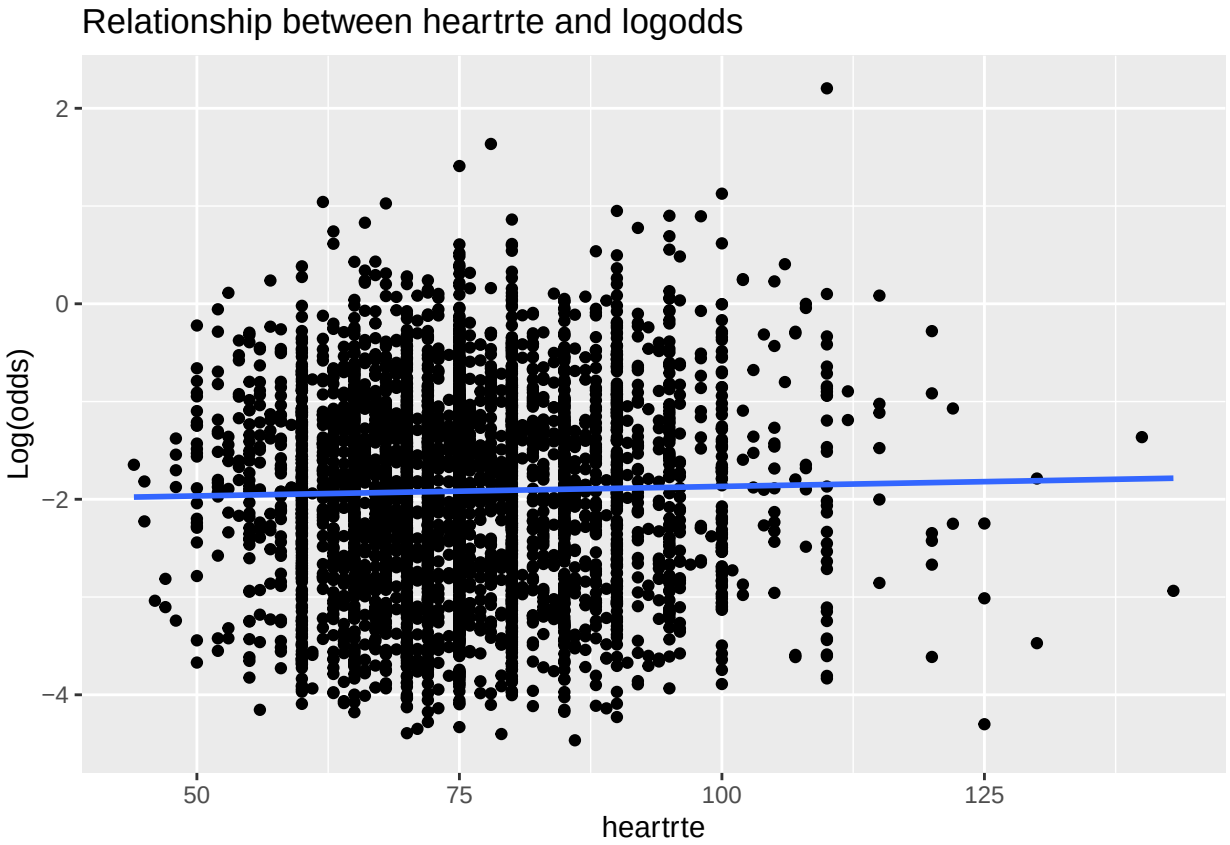

Relationship between cigpday and logodds
Log(odds)
2
0
−2
−4
0
20
40
60
cigpday

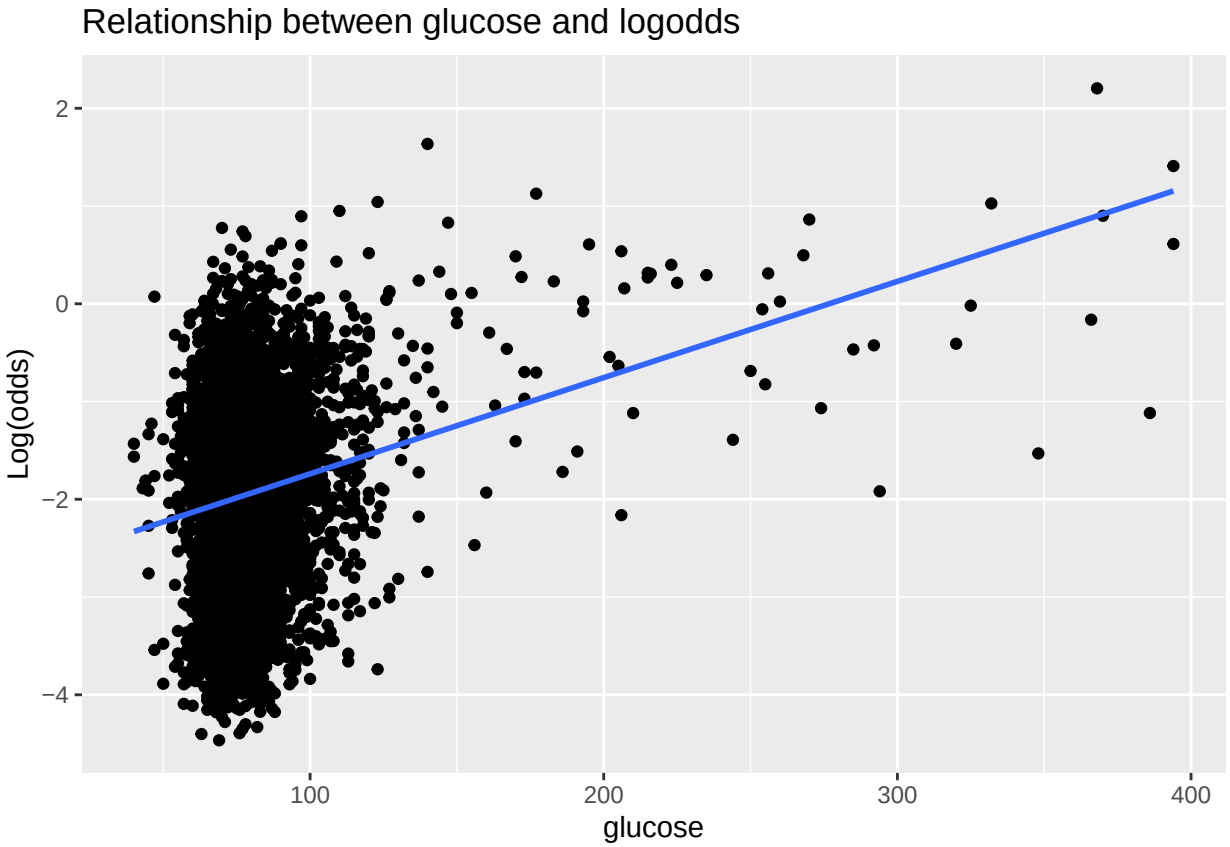

Relationship between totchol and logodds
Log(odds)
2
0
−2
−4
100
200
300
400
500
600
totchol


Relationship between sysbp and logodds
Log(odds)
2
0
−2
−4
100
150
200
250
300
sysbp

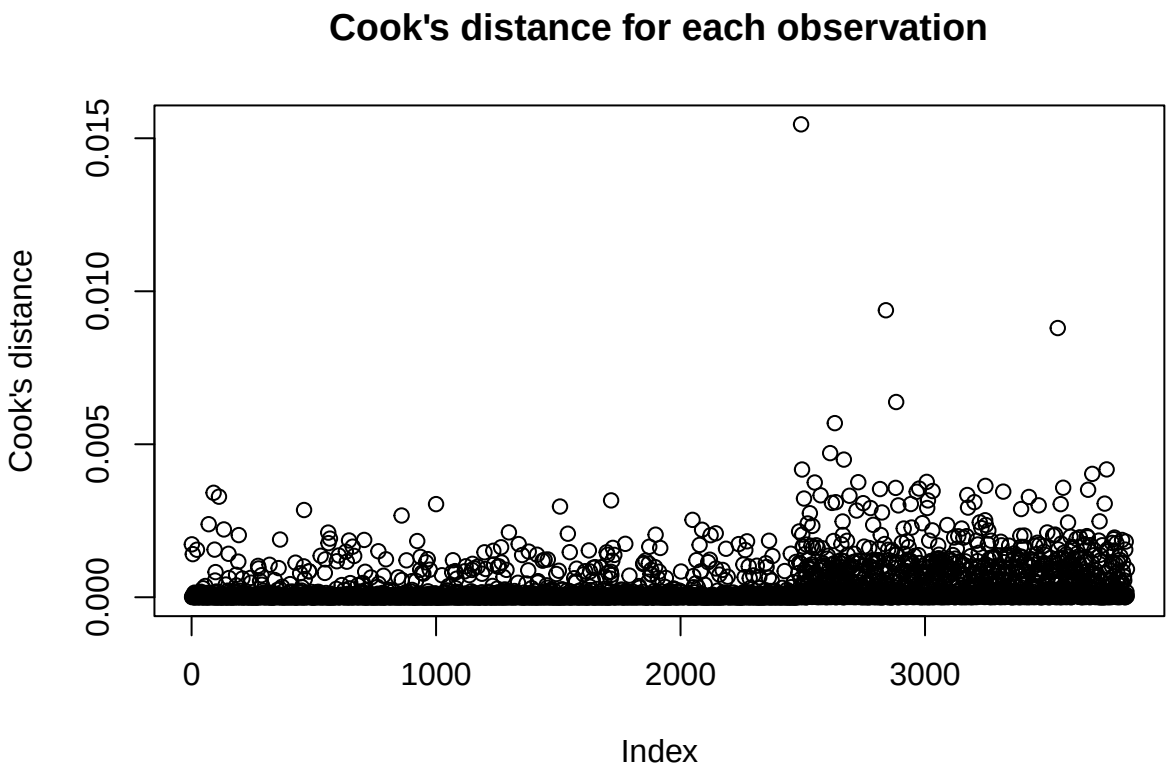

Relationship between diabp and logodds
Log(odds)
2
0
−2
−4
50
75
100
125
diabp


Relationship between bmi and logodds
Log(odds)
2
0
−2
−4
20
30
40
50
bmi

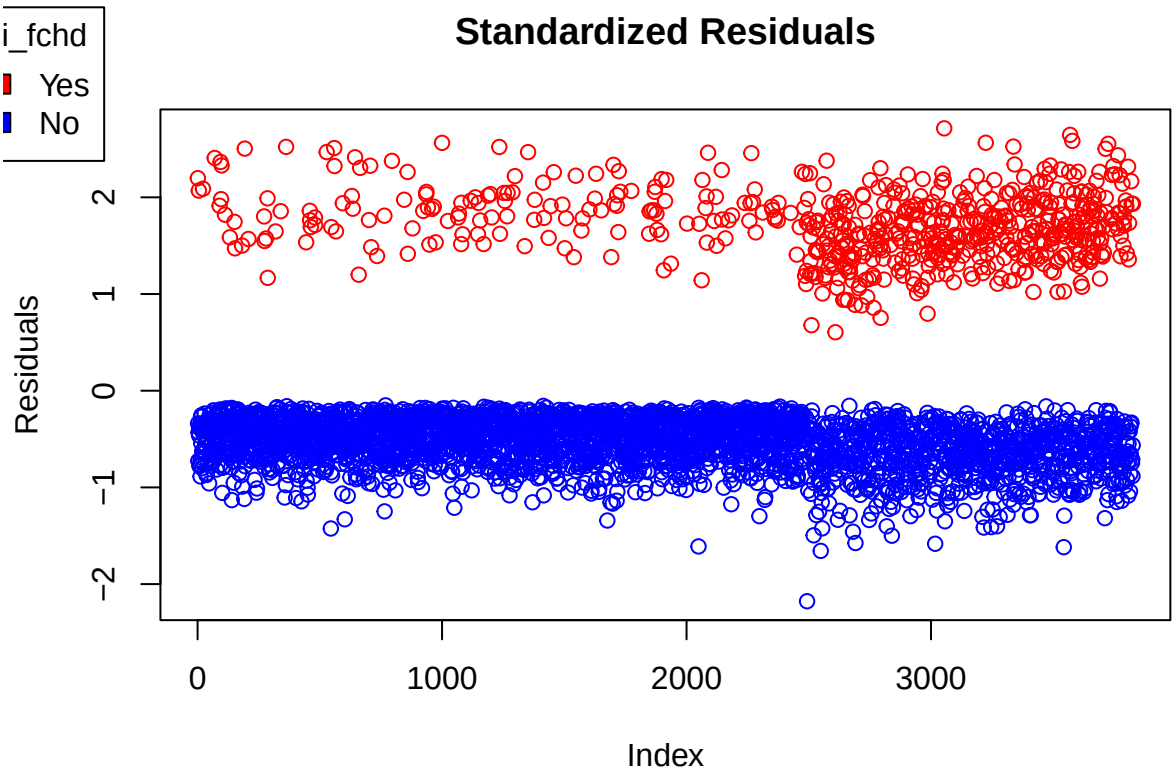

Relationship between heartrte and logodds
Log(odds)
2
0
−2
−4
50
75
100
125
heartrte




```
plot_cooks_distance(glm.full)
```




```
plot_residuals(glm.full,data.full)
```



Außreißer wurden identifiziert. Die Entfernung wirkt sich auf das reduzierte Modell gar nicht oder sogar negativ aus.

```
load_source()

# Remove Outliers

data <- data.full
glm <- glm.full



# Removal of outliers have no effect on goodness of fit regarding cook distance

cooks_d <- cooks.distance(glm)
cooks.data <- data
cooks.data$cooks_d <- cooks_d

# Define the threshold for cook distance. 0.5 would be critical.
# We are far below that.
cooks_threshold <- 0.05

# Remove rows where logit surpasses the threshold
cooks.data <- subset(cooks.data, cooks_d <= cooks_threshold)
```

```
# Removal of outliers decreases goodness of fit regarding logit
# and the variables below. They were initially removed
# due to the visual analysis of the logit plots.

logit.data <- get_logit_data(glm,data)

# Define the threshold for logit
logit_threshold <- 2

# Remove rows where logit surpasses the threshold
logit.data <- subset(logit.data, logodds <= logit_threshold)

# Define the threshold
threshold <- 500

# Remove rows where totchol surpasses the threshold
logit.data <- subset(logit.data, totchol <= threshold)

# Define the threshold for sysbp
sysbp_threshold <- 250

# Remove rows where sysbp surpasses the threshold
logit.data <- subset(logit.data, sysbp <= sysbp_threshold)
```

Das Modell wird durch Rückwertsverfahren reduziert. Der p Wert für die Modellgüte konnte über 0.05 angehoben werden. D.h. die Güte des Modells ist akzeptabel.

```
load_source()

# Repeat Analysis of Performance And Assumptions for the reduced model

# Due to a strange behaviour of step() function, we have to run the analysis inside this file.
# Possible reasons: reduction of dataframe size and calling the step() function from the helper file.

glm <- glm(mi_fchd ~ sex + age + educ + cursmoke + cigpday +
             bpmeds + totchol + sysbp + diabp + bmi + diabetes + heartrte +
             glucose, data = data, family = binomial)

glm <- step(glm, direction = "backward")
```

```
## Start:  AIC=3045.54
## mi_fchd ~ sex + age + educ + cursmoke + cigpday + bpmeds + totchol +
##     sysbp + diabp + bmi + diabetes + heartrte + glucose
##
##            Df Deviance    AIC
## - educ      3   3014.7 3040.7
## - cigpday   1   3013.5 3043.5
## - heartrte  1   3014.0 3044.0
## - bpmeds    1   3014.2 3044.2
## - diabp     1   3014.3 3044.3
## - glucose   1   3015.5 3045.5
```

```
## <none>          3013.5 3045.5
## - cursmoke  1   3018.1 3048.1
## - diabetes  1   3019.8 3049.8
## - bmi       1   3021.1 3051.1
## - age       1   3033.3 3063.3
## - sysbp     1   3041.0 3071.0
## - totchol   1   3068.6 3098.6
## - sex       1   3181.1 3211.1
##
## Step:  AIC=3040.74
## mi_fchd ~ sex + age + cursmoke + cigpday + bpmeds + totchol +
##     sysbp + diabp + bmi + diabetes + heartrte + glucose
##
##            Df Deviance    AIC
## - cigpday   1   3014.7 3038.7
## - heartrte  1   3015.3 3039.3
## - diabp     1   3015.4 3039.4
## - bpmeds    1   3015.5 3039.5
## - glucose   1   3016.6 3040.6
## <none>          3014.7 3040.7
## - cursmoke  1   3019.3 3043.3
## - diabetes  1   3021.0 3045.0
## - bmi       1   3021.9 3045.9
## - age       1   3033.8 3057.8
## - sysbp     1   3041.9 3065.9
## - totchol   1   3070.6 3094.6
## - sex       1   3185.6 3209.6
##
## Step:  AIC=3038.74
## mi_fchd ~ sex + age + cursmoke + bpmeds + totchol + sysbp + diabp +
##     bmi + diabetes + heartrte + glucose
##
##            Df Deviance    AIC
## - heartrte  1   3015.3 3037.3
## - diabp     1   3015.4 3037.4
## - bpmeds    1   3015.5 3037.5
## - glucose   1   3016.6 3038.6
## <none>          3014.7 3038.7
## - diabetes  1   3021.0 3043.0
## - bmi       1   3022.0 3044.0
## - cursmoke  1   3024.9 3046.9
## - age       1   3033.9 3055.9
## - sysbp     1   3041.9 3063.9
## - totchol   1   3070.7 3092.7
## - sex       1   3197.3 3219.3
##
## Step:  AIC=3037.27
## mi_fchd ~ sex + age + cursmoke + bpmeds + totchol + sysbp + diabp +
##     bmi + diabetes + glucose
##
##            Df Deviance    AIC
## - diabp     1   3016.0 3036.0
## - bpmeds    1   3016.0 3036.0
## - glucose   1   3017.0 3037.0
```

```
## <none>          3015.3 3037.3
## - diabetes  1   3021.5 3041.5
## - bmi       1   3022.4 3042.4
## - cursmoke  1   3025.1 3045.1
## - age       1   3034.9 3054.9
## - sysbp     1   3042.0 3062.0
## - totchol   1   3070.7 3090.7
## - sex       1   3201.5 3221.5
##
## Step:  AIC=3036.02
## mi_fchd ~ sex + age + cursmoke + bpmeds + totchol + sysbp + bmi +
##     diabetes + glucose
##
##            Df Deviance    AIC
## - bpmeds    1   3016.8 3034.8
## - glucose   1   3017.9 3035.9
## <none>          3016.0 3036.0
## - diabetes  1   3022.3 3040.3
## - bmi       1   3022.5 3040.5
## - cursmoke  1   3026.0 3044.0
## - age       1   3038.3 3056.3
## - sysbp     1   3063.5 3081.5
## - totchol   1   3071.1 3089.1
## - sex       1   3202.6 3220.6
##
## Step:  AIC=3034.78
## mi_fchd ~ sex + age + cursmoke + totchol + sysbp + bmi + diabetes +
##     glucose
##
##            Df Deviance    AIC
## - glucose   1   3018.7 3034.7
## <none>          3016.8 3034.8
## - diabetes  1   3023.1 3039.1
## - bmi       1   3023.3 3039.3
## - cursmoke  1   3026.7 3042.7
## - age       1   3039.3 3055.3
## - sysbp     1   3070.3 3086.3
## - totchol   1   3072.4 3088.4
## - sex       1   3202.7 3218.7
##
## Step:  AIC=3034.68
## mi_fchd ~ sex + age + cursmoke + totchol + sysbp + bmi + diabetes
##
##            Df Deviance    AIC
## <none>          3018.7 3034.7
## - bmi       1   3025.3 3039.3
## - cursmoke  1   3028.5 3042.5
## - diabetes  1   3037.6 3051.6
## - age       1   3041.5 3055.5
## - sysbp     1   3073.1 3087.1
## - totchol   1   3074.5 3088.5
## - sex       1   3204.6 3218.6
```

```r
summary(glm)
```

```
##
## Call:
## glm(formula = mi_fchd ~ sex + age + cursmoke + totchol + sysbp +
##     bmi + diabetes, family = binomial, data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.548628   0.496823 -15.194  < 2e-16 ***
## sex2        -1.335804   0.102835 -12.990  < 2e-16 ***
## age          0.028057   0.005875   4.775 1.79e-06 ***
## cursmoke1    0.308197   0.098805   3.119  0.00181 **
## totchol      0.007844   0.001055   7.434 1.05e-13 ***
## sysbp        0.016050   0.002166   7.411 1.25e-13 ***
## bmi          0.030823   0.011926   2.584  0.00975 **
## diabetes1    0.968274   0.216035   4.482 7.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3448.8  on 3825  degrees of freedom
## Residual deviance: 3018.7  on 3818  degrees of freedom
## AIC: 3034.7
##
## Number of Fisher Scoring iterations: 5
```

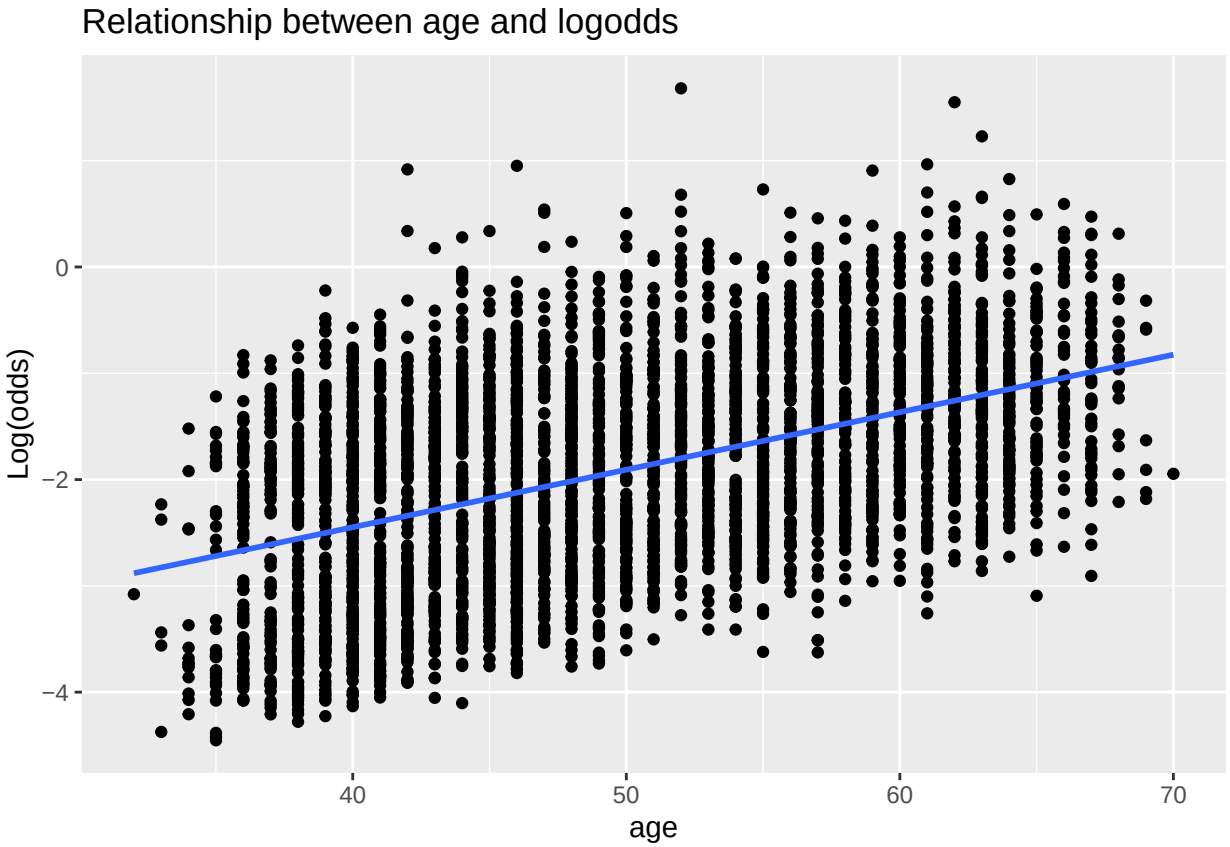
```r
print_odds_ratios(glm)
```

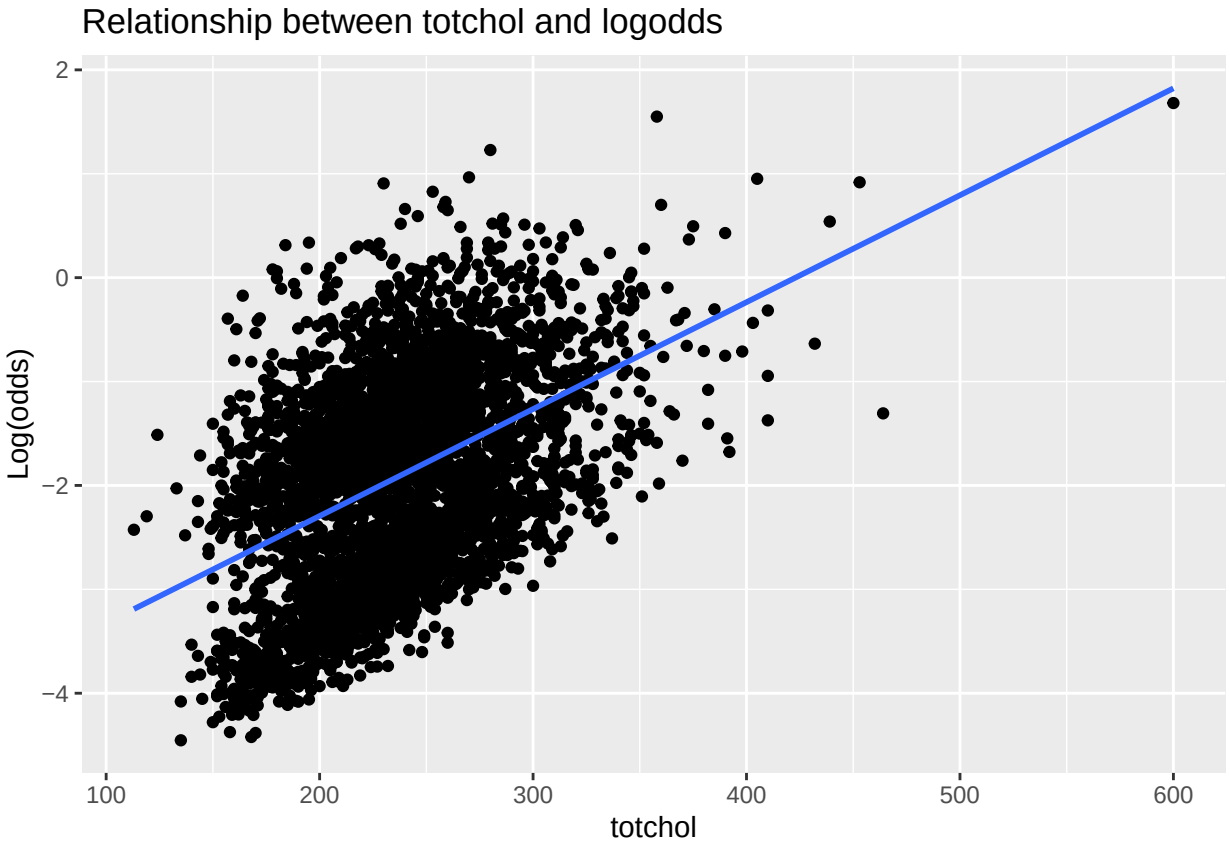
```
##  (Intercept)         sex2          age    cursmoke1      totchol        sysbp
## 0.0005268323 0.2629467880 1.0284543781 1.3609689081 1.0078749109 1.0161790052
##          bmi    diabetes1
## 1.0313034539 2.6333957189
```

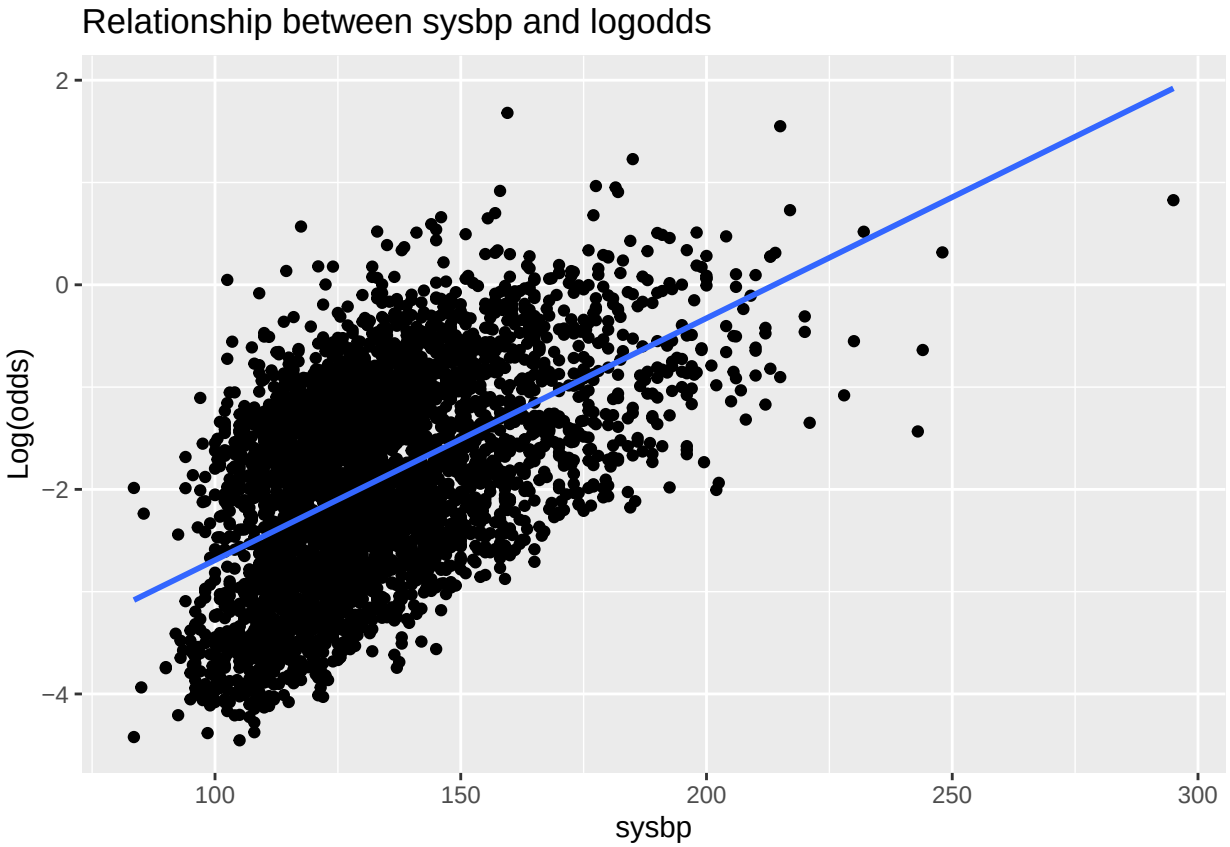
```r
print_glm_parameters(glm,data)
```

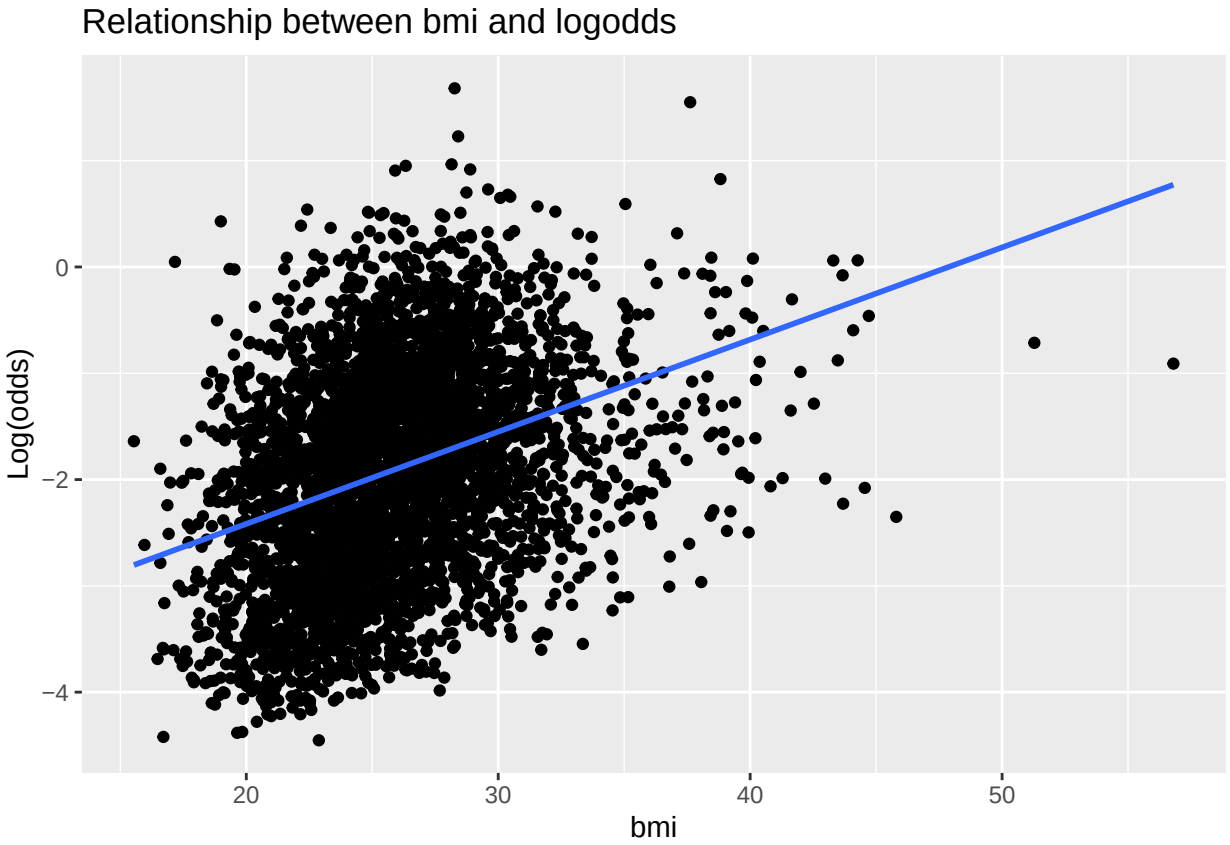
```
## [1] "R^2:  0.12470778901786"
## [1] "Goodness of fit:  0.0574105135437186"
## [1] "AUROC:  0.750870230448823"
## [1] "Sensitivity: 0.98525721455458"
## [1] "Specificity: 0.0689655172413793"
## [1] "Correctly classified: 83.2462101411396 %"
## [1] "Falsely classified: 16.7537898588604 %"
```

```r
# Numerical predictors of reduced model
variables <- c("age", "totchol", "sysbp", "bmi")
logit_data <- get_logit_data(glm,data)
plot_logit(logit_data,variables)
```


Relationship between age and logodds
Log(odds)
0
−2
−4
40
50
60
70
age


Relationship between totchol and logodds
Log(odds)
totchol
2
0
−2
−4
100
200
300
400
500
600

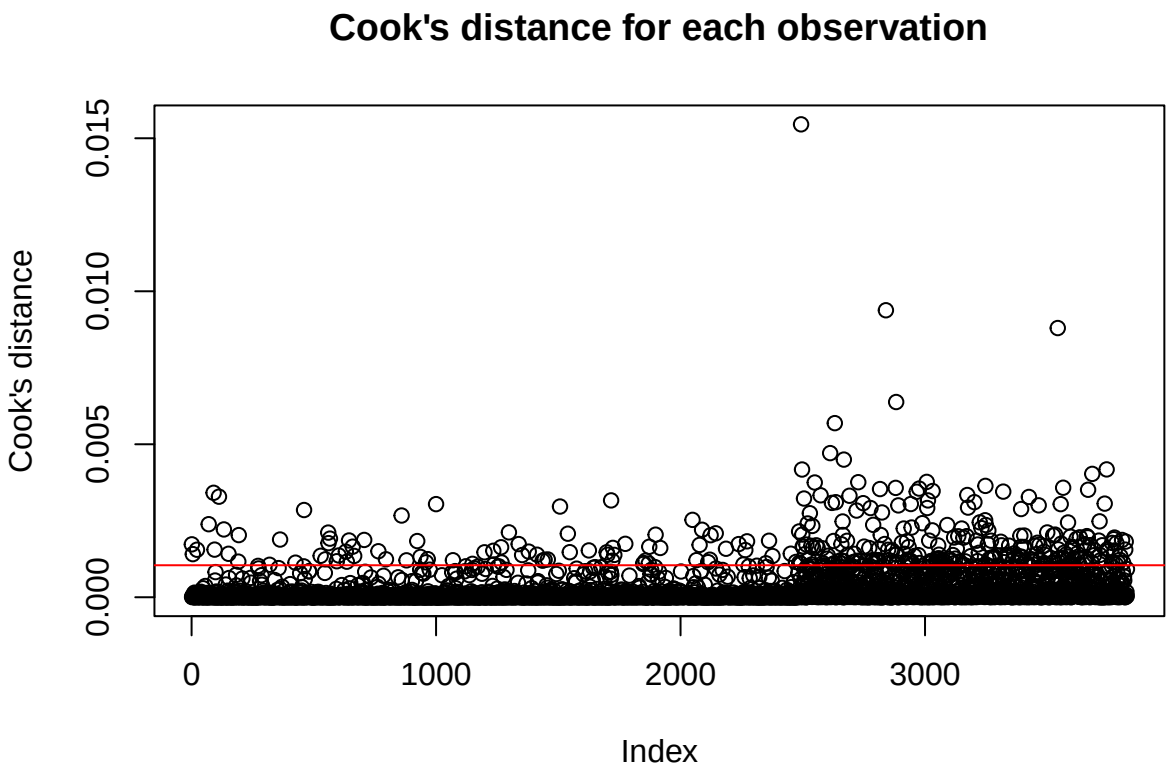

Relationship between sysbp and logodds
Log(odds)
2
0
−2
−4
100
150
200
250
300
sysbp



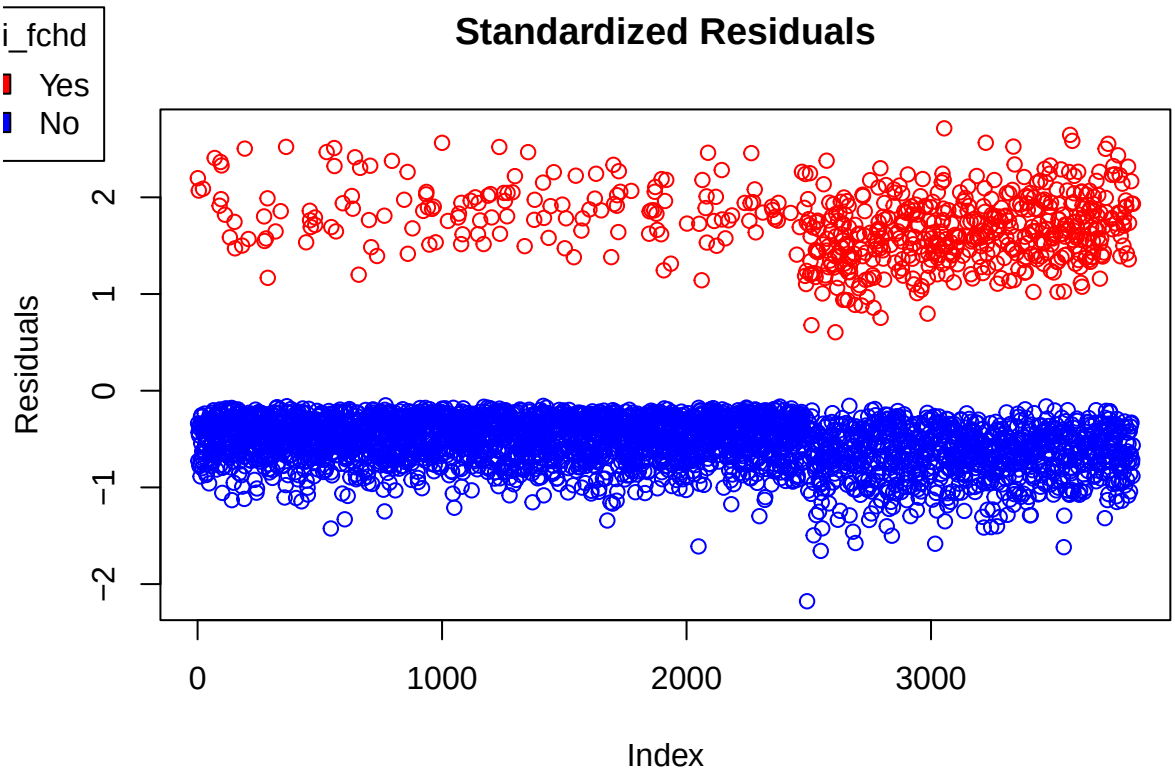

```
# Check for multicollinearity
vif(glm)
```

```
##      sex      age cursmoke  totchol    sysbp      bmi diabetes
## 1.139851 1.193165 1.139720 1.054835 1.281242 1.125249 1.023974
```

```
plot_cooks_distance(glm.full)
```




```
plot_residuals(glm.full,data.full)
```




```
# The always no model

# Count all rows
total_rows <- nrow(data)
total_rows
```

```
## [1] 3826
```

```
# Count all "0" values in mi_fchd
no_values <- sum(data$mi_fchd == "no")
no_values
```

```
## [1] 3188
```

```
# Correctly classified cases
print(no_values/total_rows)
```

```
## [1] 0.8332462
```

Die visuelle Analyse der logit Plots zeigt keine Auffälligkeiten. Residuen und Cook Distanz unauffällig. Keine der Prädiktoren hat einen Varianzinflationsfaktor größer als 1.5 D.h. Multikollinearität ist nicht gegeben.

Der AUROC beträgt 0.75 und die Genauigkeit 83%. Das Modell, das immer “Nein” prognostiziert erreicht eine Genauigkeit von 83%. Damit liegt es gleich auf mit dem reduzierten Modell. Dies lässt an der Sinnhaftigkeit des Modells zweifeln.

Laut OR der Prädiktoren des reduzierten Modells haben Diabetes, Rauchen und weibliches Geschlecht den stärksten Einfluss für das Risiko einer Hospitalisierung durch Myokardinfarkt oder tödliche koronare Herzkrankheit, in der genannten Reihenfolge.

3 Logistische Regressionsanalyse [2P]

Analysieren Sie Tabelle 4 in der Publikation "**Analysis of factors associated with disease outcomes in hospitalized patients with 2019 novel coronavirus disease**". Den Artikel finden Sie als pdf-Version in Moodle, oder direkt unter https://doi.org/10.1097/CM9.0000000000000775.

Beantworten Sie bitte in eigenen Worten folgende Fragen:

- Was ist die Outcome Variable?

Die in der Studie untersuchte Outcome Variable war das Fortschreiten der Kankheit zu einem schwerwiegenderend Zustand. Die Forscher haben die Patienten in zwei Gruppen eingeteilt: Eine Gruppe mit Fällen die sich erholten oder stabil blieben, und eine Gruppe mit Fällen die sich nicht im Zustand verschlechterten. Bezeichnet wurden diese Gruppen mit "improvement/stabilization" und "progression".

- Welche Variablen wurden in die multivariate Analyse aufgenommen und warum?

In die multivariate Analyse wurden jene Variablen aufgenommen, die in der univariaten Analyse einen signifikanten Einfluss auf die Zielvariable hatten. Die Forscher haben die Variablen mit einem p-Wert von < 0.05 in die multivariate Analyse aufgenommen. Konkret waren das die folgende Variablen: Alter, Rauchverhalten, Maxtemperatur bei Aufnahme, Atemstillstand, Schwere Erkrankung anderer Art sowie die Laborparameter Albumin, Creatinin, Procalcitonin, C-reactive Proteine

- Was sind die Top 3 Variablen in der univariaten Analyse mit dem stärksten Einfluss auf die Zielvariable?

Die Top 3 Variablen in der univariaten Analyse mit dem stärksten Einfluss auf die Zielvariable waren: Albunin mit einem OR von 12.5 [2.4;65.2], Rauchverhalten mit einem OR von 12.2 [0.5;7.2] und Alter mit einem OR von 10.6 [2.1;53;4].

- Was sind die Top3 Variablen in der multivariaten Analyse mit dem stärksten Einfluss auf die Zielvariable?

Die Top 3 Variablen in der multivariaten Analyse mit dem stärksten Einfluss auf die Zielvariable waren: Rauchverhalten mit einem OR von 14.28 [1.5;25], der Biomarker C-reactive Proteine mit einem OR von 10.53 [1.3;34.7] und die Maximaltemperatur bei Aufnahme mit einem OR von 9 [1;78.2].

- Welche (medizinische) Bedeutung haben die Variablen im multivariaten Modell?

Die Variablen im multivariaten Modell haben eine ernstzunehmende medizinische Bedeutung, da sie einen signifikanten Einfluss auf das Fortschreiten der Krankheit haben. Die Ergebnisse der Studie zeigen, dass Rauchverhalten, das Vorkommen von C-reactive Proteinen und die Maximaltemperatur bei Aufnahme die aussagekräftigsten Prädiktoren für das Fortschreiten der Krankheit sind.

Sie können dazu dienen, jene Patient:innen zu identifizieren, die ein höheres Risiko für schwerwiegende COVID-19-Verläufe haben: Ältere Patient:innen, Raucher:innen, Patient:innen mit erhöhten C-reactive Proteinen und erhöhter Maximaltemperatur bei Aufnahme.

- Welche Schlussfolgerung ziehen sie für die Praxis?

Höhere Körpertemperatur: Patienten mit hohem Fieber, langen Fieberperioden und schnellem Fieberverlauf sollten engmaschig überwacht werden, da ein hohes Fieber mit einer Verschlechterung der Erkrankung verbunden ist. *Respiratorische Anzeichen:* Atemfrequenz und das Auftreten von Atemversagen in der Vergangenheit sind wichtige Indikatoren für das Risiko einer schweren Erkrankung. *C-reaktives Protein und Albumin:* Ein erhöhter Spiegel von c-reaktiven Proteinen und ein niedriger Albuminspiegel sind mit einem schlechteren Verlauf von COVID-19 assoziiert. Ersterer ist ein wichtiger Indikator für Entzündungen und Albumin zeigt den Ernährungszustand des Körpers an. Diese beiden Parameter sollten während der Behandlung genau überwacht werden, da Veränderungen in diesen Werten die Krankheitsprogression widerspiegeln können. *Frühe Diagnose und Monitoring:* Eine frühzeitige Diagnose und eine dynamische Überwachung von Risikofaktoren sind für eine effektive Behandlung entscheidend. Kliniker sollten die genannten Indikatoren einer Krankheitverschlechterung kennen und bei Hochrisikopatienten frühzeitig intensivierte Behandlungsmaßnahmen in Erwägung ziehen.