TMA4315: Compulsory exercise 1 (title)

Group 0: Name1, Name2 (subtitle)

09.10.2023

Part 1

Bold

italic

To get a pdf file, make comments of the lines with the "html_document" information, and make the lines with the "pdf_document" information regular, and vice versa.

a)

The log-likelihood function for a binary regression model:

$$\ell(\beta) = \sum_{i=1}^{n} y_i ln(\pi_i) + (1 - y_i) ln(1 - \pi_i)$$

How we arrived at this expression: 1. Start with the link function, aka logarithm of the odds, $ln(\frac{\pi}{1-\pi}) = \sum_{i=0}^{n} \beta_i x_i$, where $x_0 = 1$. This can be written as $\pi = \frac{1}{1+e^{-\sum_{i=0}^{n} \beta_i x_i}}$ 2. Then we formulate the likelihood: $L(\beta) = \prod_{i=1}^{n} P(x_i|\beta) = \prod_{i=1}^{n} \left(\frac{1}{1+\exp(-(\beta x_i))}\right)^{y_i} \left(1 - \frac{1}{1+\exp(-(\beta x_i))}\right)^{1-y_i}$ (here β and x_i are vectors, and y_i is the response-variable). 3. We get the log-likelihood by taking the logarithm: $\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+\exp(-(\beta x_i))}\right) + (1 - y_i) \log\left(1 - \frac{1}{1+\exp(-(\beta x_i))}\right)\right]$ 4. We put $\pi = \frac{1}{1+\exp(-(\beta x_i))}$ and get $\ell(\beta) = \sum_{i=1}^{n} [y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)]$.

b)

```r
# importing data
filepath <- "https://www.math.ntnu.no/emner/TMA4315/2018h/mountains"
mount <- read.table(file = filepath, header = TRUE, col.names = c("height",
    "prominence", "fail", "success"))
```

```r
# fitting model
logreg.mod = glm(cbind(success, fail) ~ height + prominence, data = mount,
    family = "binomial")
summary(logreg.mod)
```

```
##
## Call:
```

```
## glm(formula = cbind(success, fail) ~ height + prominence, family = "binomial",
##     data = mount)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.369e+01  1.064e+00  12.861  < 2e-16 ***
## height      -1.635e-03  1.420e-04 -11.521  < 2e-16 ***
## prominence  -1.740e-04  4.554e-05  -3.821 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 715.29  on 112  degrees of freedom
## Residual deviance: 414.68  on 110  degrees of freedom
## AIC: 686.03
##
## Number of Fisher Scoring iterations: 4
```

```r
# perform likelihood ratio test
logreg.null_mod = glm(cbind(success, fail) ~ 1, data = mount, family = "binomial")
anova(logreg.null_mod, logreg.mod, "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(success, fail) ~ 1
## Model 2: cbind(success, fail) ~ height + prominence
##   Resid. Df Resid. Dev Df Deviance
## 1       112     715.29
## 2       110     414.68  2   300.61
```

```r
# creating 95% confint for beta
cat(" \n Confidence interval: \n")
```

```
##
##  Confidence interval:
```

```r
confint(logreg.mod)
```

```
##                     2.5 %        97.5 %
## (Intercept) 11.6156540313  1.578900e+01
## height      -0.0019157664 -1.359060e-03
## prominence  -0.0002633821 -8.480319e-05
```

```r
cat(" \n exp(CI): \n")
```

```
##
##  exp(CI):
```

```r
exp(confint(logreg.mod))
```

```
##                    2.5 %       97.5 %
## (Intercept) 1.108191e+05 7.195726e+06
## height      9.980861e-01 9.986419e-01
## prominence  9.997367e-01 9.999152e-01
```

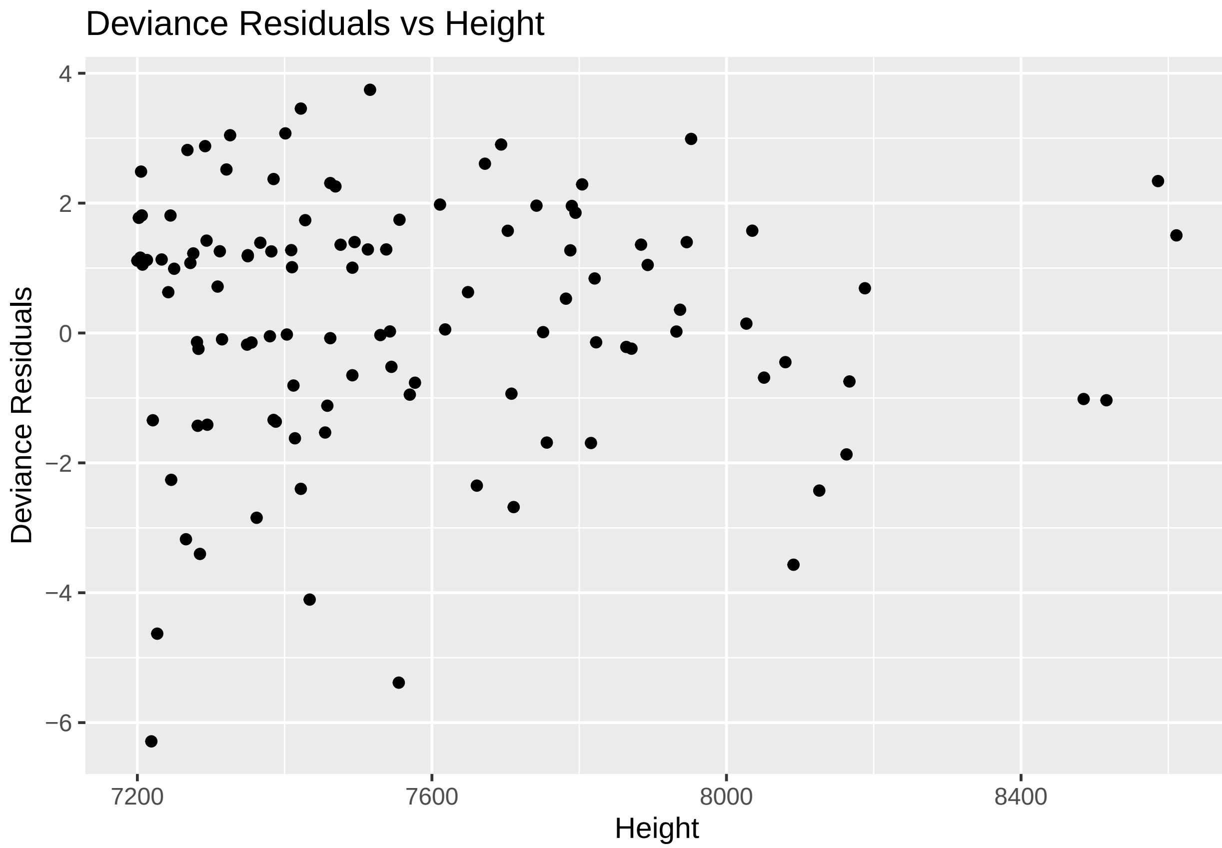
- We can see from the estimates that the log(odds) decrease when height or prominence increase, meaning the chance of success decreases.
- The p-values obtained by performed wald-test suggest that the covariates are very significant. We also performed LRT, where the decrease in residual deviance from 715.29 in Model 1 to 414.68 in Model 2 indicates that Model 2 fits the data significantly better.
- CI for height: (-0.0019157664 -1.359060e-03), which also clearly indicates a negative log-odds relation between height and success.
- In we take $(exp(\beta_L), exp(\beta_H))$ we get the odds instead of the log-odds, which is given above under exp(CI). Here, since odds-confidence interval for height and prominence displays values <1, it indicates that for unit increase in height and prominence the odds for success diminishes.

c)

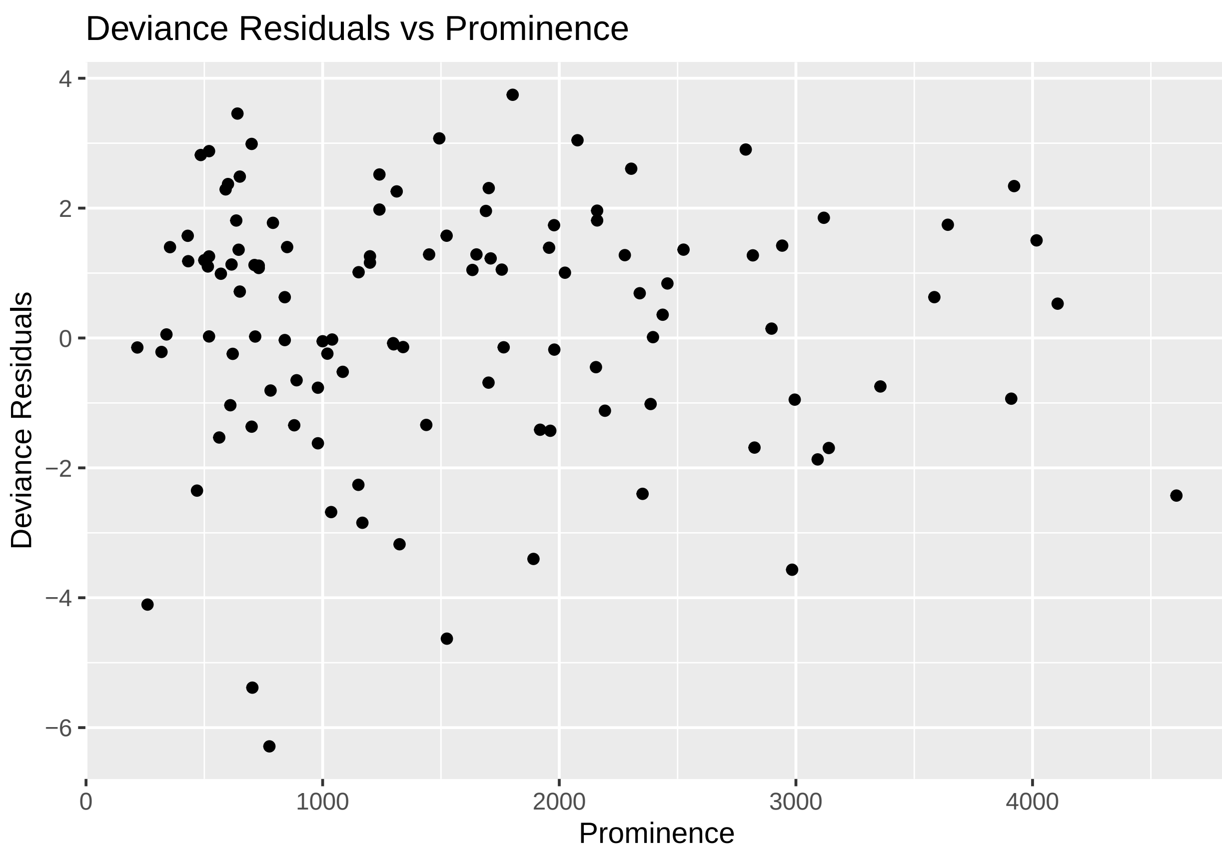
```r
dev_res <- residuals(logreg.mod, type = "deviance")

# Create a data frame to use with ggplot2
plot_data <- data.frame(Height = logreg.mod$data$height, Prominence = logreg.mod$data$prominence,
    DevRes = dev_res)

# Plotting deviance residuals against height
ggplot(plot_data, aes(x = Height, y = DevRes)) + geom_point() + labs(x = "Height",
    y = "Deviance Residuals", title = "Deviance Residuals vs Height")
```




```r
# Plotting deviance residuals against prominence
ggplot(plot_data, aes(x = Prominence, y = DevRes)) + geom_point() + labs(x = "Prominence",
    y = "Deviance Residuals", title = "Deviance Residuals vs Prominence")
```



Part 2

```
filepath <- "https://www.math.ntnu.no/emner/TMA4315/2023h/eliteserien2023.csv"
eliteserie <- read.csv(file = filepath)

NGames <- table(c(eliteserie$home[!is.na(eliteserie$yh)], eliteserie$away[!is.na(eliteserie$yh)]))
RangeofGames <- range(NGames)
```

a)

In this case we have two team categories which can score M goals, where $M = \max_{ij} O_{ij}$, $\mathbf{O} \in \mathbb{R}^{2\times M}$ is the matrix representing the contingency table. The matrix of expected frequencies, whose entries are given by

$$E_{ij} = \frac{\sum_p \mathbf{O}_{ip} + \sum_k \mathbf{O}_{kj}}{\sum_{p,k} \mathbf{O}_{pk}}$$

can be written in matrix form as

$$\mathbf{E} = \frac{\mathbf{O}\mathbf{1}_M \mathbf{1}_2^T \mathbf{O}}{\mathbf{1}_M^T \mathbf{O} \mathbf{1}_2} \tag{1}$$

where $\mathbf{1}_k \in \mathbb{R}^k$ is the k-dimensional vector of ones. Using the definition of $\mathbf{O}$ and Equation @ref{eq:expfreq}, the χ^2-test can be written

$$\chi^2 = \sum_{i,j} \frac{\mathbf{O}_{ij} - \mathbf{E}_{ij}}{\mathbf{E}_{ij}} \sim \chi^2_{1,M-1} \tag{2}$$

```
# O <- rbind(table(eliteserie$yh), table(eliteserie$ya)) O[2,
# (max(eliteserie$ya)+1):max(eliteserie$yh)] <- 0
# max(eliteserie$ya)
```