Exam exercise: Logistic regression analysis of Berkely admission data

You may use the combined lecture notes for this module available at https://asta.math.aau.dk to guide you to the relevant methods and R commands for this exam.

The following table shows the total number of admitted and rejected applicants to the six largest departments at University of Berkeley in 1973.

	Admitted	Rejected
Male	1198	1493
Female	557	1278

Use a χ^2-test to check whether the admission statistics for Berkeley show any sign of gender discrimination. To enter the table in R you can do:

```
admit <- matrix(c(1198, 557, 1493, 1278), 2, 2)
rownames(admit) <- c("Male", "Female")
colnames(admit) <- c("Admitted", "Rejected")
admit <- as.table(admit)

# Beregning af forventet procentvis
n <- margin.table(admit)
pctGoals <- round(100 * margin.table(admit, 2)/n, 1)

expected_admit_freq <- (1198+557)/n
expected_reject_freq <- (1493+1278)/n

# Beregning af forventet i tal (rækketotal delt med 100, ganget med forventet procent)
admit_expected_male <- ((1198+1493)/100)*38.8
reject_expected_male <- ((1198+1493)/100)*61.2
admit_expected_fem <- ((557+1278)/100)*38.8
reject_expected_fem <- ((557+1278)/100)*61.2

# Opsætning af forventet tabel for at tjekke overensstemmelse mellem den og funktionens output.
expected_admit <- matrix(c(admit_expected_male, reject_expected_male, admit_expected_fem, reject_expecte
rownames(expected_admit) <- c("Male", "Female")
colnames(expected_admit) <- c("Admitted", "Rejected")
expected_admit <- as.table(expected_admit)

# Chi^2-test
teststat <- chisq.test(admit, correct = FALSE)
round(teststat$expected, 0)
```

```
##        Admitted Rejected
## Male       1043     1648
## Female      712     1123
```

```
teststat
```

```
##
##  Pearson's Chi-squared test
##
## data:  admit
## X-squared = 92.205, df = 1, p-value < 2.2e-16
```

Your analysis should as a minimum contain **arguments** that support:

- Statement of hypotheses

H_0 : De to grupper, mænd og kvinder, er statistisk uafhængige.

H_1 : De to grupper, mænd og kvinder, er statistisk afhængige.

- Calculation of expected frequencies

$\pi_{admit} = 0.388$

$\pi_{reject} = 0.612$

- Calculation of test statistic

$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$

$\chi^2 = 92.205$

- Calculation and interpretation of p-value.

```
p <- 1 - pdist("chisq", 92.205, df = 1, return = "value", plot = FALSE)
p
```

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

$p = 2.2 \cdot 10^{-16} \approx 0$

p-værdien er praktisk talt lig 0, hvorfor nulhypotesen forkastes, og det må konkluderes, at der er en signifikant statistisk afhængighed mellem de to grupper.

A more detailed data set with the admissions for each department is available on the course web page. The variables are:

- `Gender` (male/female)
- `Dept` (department A, B, C, D, E, F)
- `Admit` (frequency of admitted for each combination)
- `Reject` (frequency of rejected for each combination)

Load the data into RStudio:

```
admission <-
    read.table("http://asta.math.aau.dk/dan/static/datasets?file=admission.dat",
               header=TRUE)
admission
```

```
##    Gender Dept Admit Reject
## 1    Male    A   512    313
## 2  Female    A    89     19
## 3    Male    B   353    207
## 4  Female    B    17      8
## 5    Male    C   120    205
## 6  Female    C   202    391
## 7    Male    D   138    279
## 8  Female    D   131    244
## 9    Male    E    53    138
## 10 Female    E    94    299
## 11   Male    F    22    351
## 12 Female    F    24    317
```

In order to do logistic regression for this kind of data, the response is the columns `Admit` and `Reject` (which means that we model the probability of admit) :

```r
m0 <- glm(cbind(Admit, Reject) ~ Gender + Dept, family = binomial, data = admission)
```

The glm-object `m0` is a logistic model with main effects of `Gender` and `Department`.

- Investigate whether there is any effect of these predictors.

```r
mainEffects <- glm(cbind(Admit, Reject)~ ., data=admission, family=binomial)
noEffects <- glm(cbind(Admit, Reject) ~ 1, data=admission, family=binomial)

anova(mainEffects, noEffects, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Admit, Reject) ~ Gender + Dept
## Model 2: cbind(Admit, Reject) ~ 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1         5      20.20
## 2        11     877.06 -6  -856.85 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Med en p-værdi på praktisk talt 0, kan vi kan forkaste $H_0 : \beta_1 = \beta_2 = \ldots = \beta_n = 0$

As a hint you might look at section 9.3 in the combined lecture notes.

```r
summary(m0)
```

```
##
## Call:
## glm(formula = cbind(Admit, Reject) ~ Gender + Dept, family = binomial,
##     data = admission)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
## -1.2487   3.7189  -0.0560   0.2706   1.2533  -0.9243   0.0826  -0.0858
##       9       10       11       12
##  1.2205  -0.8509  -0.2076   0.2052
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.68192    0.09911   6.880 5.97e-12 ***
## GenderMale  -0.09987    0.08085  -1.235    0.217
## DeptB       -0.04340    0.10984  -0.395    0.693
## DeptC       -1.26260    0.10663 -11.841  < 2e-16 ***
## DeptD       -1.29461    0.10582 -12.234  < 2e-16 ***
## DeptE       -1.73931    0.12611 -13.792  < 2e-16 ***
## DeptF       -3.30648    0.16998 -19.452  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 877.056  on 11  degrees of freedom
## Residual deviance:  20.204  on  5  degrees of freedom
## AIC: 103.14
##
## Number of Fisher Scoring iterations: 4
```

Looking at the summary of `m0`:

- Is there a significant gender difference?

Med en p-værdi på 0.217 kan nulhypotesen ikke forkastes, hvorfor køn ikke har en signifikant betydning for at blive optaget på universitetet.

- What is the interpretation of the numbers in the `DeptB`-row?

Med en p-værdi på 0.693 er denne variabel heller ikke signifikant for modellen.

We add the standardized residuals to `admission`:

```
admission$stdRes <- round(rstandard(m0),2)
admission
```

```
##    Gender Dept Admit Reject stdRes
## 1    Male    A   512    313  -4.01
## 2  Female    A    89     19   4.26
## 3    Male    B   353    207  -0.28
## 4  Female    B    17      8   0.28
## 5    Male    C   120    205   1.87
## 6  Female    C   202    391  -1.89
## 7    Male    D   138    279   0.14
## 8  Female    D   131    244  -0.14
## 9    Male    E    53    138   1.61
## 10 Female    E    94    299  -1.65
## 11   Male    F    22    351  -0.30
## 12 Female    F    24    317   0.30
```

- Looking at the standardized residuals, which department deviates heavily from the model?

Department A afviger mest fra de forventede værdier under nulhypotesen.

- What gender is discrimated in this department?

Det standardiserede residual for mænd i Department A er negativ, hvorfor andelen af mænd, der bliver optaget på denne afdeling er væsentlig lavere end forventet under nulhypotesen. Derfor bliver der diskrimineret mod mænd i denne Department.

Next you should fit the model with the interaction `Gender*Dept` and use `anova` to compare this to `m0`.

```
m1 <- glm(cbind(Admit, Reject) ~ Gender * Dept, family = binomial, data = admission)

anova(m0, m1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Admit, Reject) ~ Gender + Dept
## Model 2: cbind(Admit, Reject) ~ Gender * Dept
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         5     20.204
## 2         0      0.000  5   20.204 0.001144 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Explain what interaction means in the current context.

Det betyder, at der er en interaktion mellem afdeling og køn. Eksempelvis er der generelt flere mænd, der søger ind på STEM-uddannelser, og flere kvinder, der søger ind på sygeplejerske-uddannelsen, hvorfor uddannelsen må have en indflydelse på kønnet af ansøgeren.

- Is there a significant interaction?

Ja.

- In the light of your analysis, explain the reason for your answer to the previous question.

p-værdien er lavere end signifikansniveauet, $\alpha = 0.05$. Altså har uddannelsen en indflydelse på kønnet af ansøgeren.