```
rm(list=ls())
library(GGally)
library(qqplotr)
library(corrplot)
library(car)
library(reshape)
library(tidyverse)
library(magrittr)
library(dplyr)
library(betareg)
library(statmod)
library(jtools)
library(numDeriv)
library(latex2exp)
library(broom.mixed)
library(car)
library(gridExtra)
library(MASS)
library(data.table)

if (Sys.getenv('USER') == "mortenjohnsen"){
  setwd("/Users/mortenjohnsen/OneDrive - Danmarks Tekniske
Universitet/DTU/10. Semester/02424 - Advanced Dataanalysis and
Statistical Modellling/02424---Assignments/")
}else if (Sys.getenv('USER') == "freja"){
  setwd("~/Documents/Uni/TiendeSemester/Adv. data analysis and stat.
modelling/02424---Assignments")
}else{
  setwd("C:/Users/catdu/OneDrive/DTU/10. semester/Advanced Dataanalysis
and Statistical Modelling/Assignment 1/02424---Assignments/")
}


# Look at data
data <- as.data.table(read.table("Assignment 2/earinfect.txt", header =
T))
summary(data)
data[,sum(persons), by = swimmer]
data[, sum(persons), by =location]
data[, sum(persons), by=age]
data[, sum(persons), by = sex]

data[persons == max(persons)]
data[,freq := infections/persons]
data[freq==max(freq)]


# Load of data
earinfect <- read.table("Assignment 2/earinfect.txt", header = T)
head(earinfect)
earinfect$swimmer <- factor(earinfect$swimmer)
earinfect$location <- factor(earinfect$location)
earinfect$age <- factor(earinfect$age)
earinfect$sex <- factor(earinfect$sex)
earinfect$healthy <- earinfect$persons - earinfect$infections
```

```
# Plots -----------------------------------------------------------------------
--
hist(earinfect$infections)
ggpairs(earinfect)

ggplot(earinfect) +
  geom_histogram(aes(x = infections, y = after_stat(density)), bins = 15,
colour = "white", position = "identity", alpha = 0.4) +
  theme_bw() +
  labs(y = "", colour = "Distribution", title = "Earinfections")
ggsave(file.path(getwd(), "Assignment 2/figs/earinfect.png"), width = 20,
height = 10, units = "cm")

# Is there any difference in mean and variance of infections variable?
mean(earinfect$infections)
var(earinfect$infections)


# Poisson model ---------------------------------------------------------------
--

fit.pois <- glm(infections ~ offset(persons) * age * sex * location *
swimmer,
                data = earinfect, family = poisson(link = 'log'))
fit.pois_full <- fit.pois
1 - pchisq(fit.pois$deviance, df = fit.pois$df.residual)
coefficients(fit.pois)
anova(fit.pois, test = "Chisq")
drop1(fit.pois, test = "Chisq")

# Drop the largest interaction
fit.pois <- update(fit.pois, .~.-age:sex:location:swimmer)
1 - pchisq(fit.pois$deviance, df = fit.pois$df.residual)
length(coefficients(fit.pois))
anova(fit.pois, test = "Chisq")
drop1(fit.pois, test = "Chisq")

fit.pois <- update(fit.pois, .~.-age:sex:swimmer)
1 - pchisq(fit.pois$deviance, df = fit.pois$df.residual)
length(coefficients(fit.pois))
anova(fit.pois, test = "Chisq")
drop1(fit.pois, test = "Chisq")

fit.pois <- update(fit.pois, .~.-sex:location:swimmer)
1 - pchisq(fit.pois$deviance, df = fit.pois$df.residual)
length(coefficients(fit.pois))
anova(fit.pois, test = "Chisq")
drop1(fit.pois, test = "Chisq")

fit.pois <- update(fit.pois, .~.-age:location:swimmer)
1 - pchisq(fit.pois$deviance, df = fit.pois$df.residual)
length(coefficients(fit.pois))
anova(fit.pois, test = "Chisq")
drop1(fit.pois, test = "Chisq")

fit.pois <- update(fit.pois, .~.-age:swimmer)
```

```
1 - pchisq(fit.pois$deviance, df = fit.pois$df.residual)
length(coefficients(fit.pois))
anova(fit.pois, test = "Chisq")
drop1(fit.pois, test = "Chisq")

# Looks good now
summary(fit.pois)
summary(fit.pois_full)
confint(fit.pois)

# Compare the reduced with full model
anova(fit.pois, fit.pois_full, test = "Chisq")

# Residual plot analysis ---------------------------------------------
-------
e.data <- earinfect

e.data$residuals <- fit.pois$residuals
e.data$leverage <- hatvalues(fit.pois)

# Residual analysis
e.data$pred <- predict(fit.pois)
e.data$pearson <- residuals(fit.pois, type = "pearson")

sigma_sq <- fit.pois$deviance / (dim(e.data)[1] -
length(coefficients(fit.pois)))
e.data$stdpearson <- e.data$pearson/sqrt(sigma_sq*(1-e.data$leverage))

first <- ggplot(e.data)+
  geom_point(aes(x = pred, y = residuals))+
  geom_hline(aes(yintercept = 0), colour = "blue", linetype = "dashed")+
  geom_smooth(aes(x = pred, y = residuals), colour = "blue", se = F)+
  theme_bw()+
  labs(x = "Predicted", y = "Residuals")+
  ggtitle("Residuals vs Fitted")

second <- ggplot(e.data)+
  geom_point(aes(x = pred, y = sqrt(stdpearson)))+
  geom_smooth(aes(x=pred,y = sqrt(stdpearson)), colour = "blue", se = F)+
  theme_bw()+
  labs(x = "Predicted", y = TeX("$\\sqrt{Std. Pearson Residuals}$"))+
  ggtitle("Scale-Location")

third <- ggplot(e.data, aes(sample = stdpearson))+
  stat_qq_band(fill = "blue", alpha = 0.2)+
  stat_qq_line(colour = "blue")+
  stat_qq_point()+
  theme_bw()+
  labs(x = "Theoretical quantiles", y = "Std. Pearson Residuals")+
  ggtitle("Normal QQ")

fourth <- ggplot(e.data, aes(x = leverage, y = stdpearson))+
  geom_point()+
  geom_hline(aes(yintercept = 0), colour = "blue", linetype = "dashed")+
  geom_smooth(colour = "blue", se = F)+
  theme_bw()+
  labs(x = "Leverage", y = "Std. Pearson Residuals")+
```

```
  ggtitle("Residuals vs Leverage")

png(filename = file.path(getwd(),"Assignment
2/figs/residualplot_poisson.png"), width = 20, height = 10, units = "cm",
res = 1000)
grid.arrange(first, third, second, fourth, nrow = 2)
dev.off()
#ggsave(file.path(getwd(), "Assignment 2/figs/residualplot_poisson.png"),
width = 20, height = 10, units = "cm")



# Check whether residuals are i.i.d.
ggplot(e.data)+
  geom_boxplot(aes(x = sex, y = pearson, fill = sex))+
  theme_bw()+
  ggtitle("Residual variation difference between genders")
ggsave(file.path(getwd(), "gender_variance.png"), width = 20, height =
10, units = "cm")
# => Residuals are not identically distributed.
# The residual variance being constant is violated (is larger for sex =
"Female")


par(mfrow = c(2, 2))
plot(fit.pois)
# The tails are signs of over-dispersion


## Forestplot
plot_summs(fit.pois)

# Negative binomial model ----------------------------------------------
--
library(MASS)
fit2 <- glm.nb(infections ~ offset(log(persons)) + swimmer + location +
sex + age + age:sex + age:location + sex:location + location:swimmer +
age:sex:location, data = earinfect)

fit2$coefficients
anova(fit2, test = "Chisq")
fit2 <- update(fit2, .~.-location:sex:age)
anova(fit2, test = "Chisq")
fit2 <- update(fit2, .~.-location:age-swimmer:location)
anova(fit2, test = "Chisq")
fit2 <- update(fit2, .~.-sex:age)
anova(fit2, test = "Chisq")
fit2 <- update(fit2, .~.-age)
anova(fit2, test = "Chisq")
fit2 <- update(fit2, .~.-location:sex)
anova(fit2, test = "Chisq")
fit2 <- update(fit2, .~.-sex)
anova(fit2, test = "Chisq")
fit2 <- update(fit2, .~.-swimmer)
anova(fit2, test = "Chisq")

AIC(fit2)
```

```
AIC(fit.pois.final)

pchisq(fit2$deviance, df = fit2$df.residual, lower.tail = F)

plot(fit2)

sumstats <- summary(fit2)
tibble("Parameter" = names(sumstats$coefficients[,1]),
       "Estimate" = sumstats$coefficients[,1],
       "95% lower" = confint(fit2)[,1],
       "95% upper" = confint(fit2)[,2],
       "Std. Error" = sumstats$coefficients[,2],
       "z value" = sumstats$coefficients[,3],
       "p-value" = sumstats$coefficients[,4]
) -> NB_model

print(xtable(NB_model , type = "latex" , caption = "Parameters for the
negative binomial model"
             , label = "tab:NBparms", digits = -1), file = "NBmodel.tex",
caption.placement = "top")

earinfect$residualsNB <- fit2$residuals
earinfect$pearsonNB <- residuals(fit2, type = "pearson")
earinfect$leverageNB <- hatvalues(fit2)
#gender-specific residual analysis
earinfect$predNB <- predict(fit2)

sigma_sq <- fit2$deviance / (dim(earinfect)[1] -
length(coefficients(fit2)))
earinfect$stdpearsonNB <- earinf$pearsonNB/sqrt(sigma_sq*(1-
earinfect$leverageNB))

first <- ggplot(earinfect)+
  geom_point(aes(x = predNB, y = residualsNB))+
  geom_hline(aes(yintercept = 0), colour = "blue", linetype = "dashed")+
  geom_smooth(aes(x = predNB, y = residualsNB), colour = "blue", se = F)+
  theme_bw()+
  labs(x = "Predicted", y = "Residuals")+
  ggtitle("Residuals vs Fitted")

second <- ggplot(earinfect)+
  geom_point(aes(x = predNB, y = sqrt(stdpearsonNB)))+
  geom_smooth(aes(x=predNB,y = sqrt(stdpearsonNB)), colour = "blue", se =
F)+
  theme_bw()+
  labs(x = "Predicted", y = TeX("$\\sqrt{Std. Pearson Residuals}$"))+
  ggtitle("Scale-Location")

third <- ggplot(earinfect, aes(sample = stdpearsonNB))+
  stat_qq_band(fill = "blue", alpha = 0.2)+
  stat_qq_line(colour = "blue")+
  stat_qq_point()+
  theme_bw()+
  labs(x = "Theoretical quantiles", y = "Std. Pearson Residuals")+
  ggtitle("Normal QQ")

fourth <- ggplot(earinfect, aes(x = leverageNB, y = stdpearsonNB))+
```

```
  geom_point()+
  geom_hline(aes(yintercept = 0), colour = "blue", linetype = "dashed")+
  geom_smooth(colour = "blue", se = F)+
  theme_bw()+
  labs(x = "Leverage", y = "Std. Pearson Residuals")+
  ggtitle("Residuals vs Leverage")

png(filename = paste0(figpath,"residual_NB.png"), width = 20, height =
10, units = "cm", res = 1000)
grid.arrange(first, third, second, fourth, nrow = 2)
dev.off()
```