AdvStDaAn, Worksheet, Week 7

Michael Lappert

29 April, 2022

Contents

Exercise 1

Exercise 1.a)

Turbine Data (cf. Exercise 1 on Worksheet Week 4) Does the GLM that you have fitted in part 1(b) model the data adequately?

```r
path <- file.path('Datasets', 'turbines.dat')
df <- read.table(path, header=TRUE)

# Fitted model in w4, 1.b)
glm1.1 <- glm(cbind(Fissures, Turbines-Fissures) ~ Hours, family = binomial, data = df)
summary(glm1.1)
```

```
##
## Call:
## glm(formula = cbind(Fissures, Turbines - Fissures) ~ Hours, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5055  -0.7647  -0.3036   0.4901   2.0943
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.9235966  0.3779589 -10.381   <2e-16 ***
## Hours        0.0009992  0.0001142   8.754   <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: 112.670  on 10  degrees of freedom
## Residual deviance:  10.331  on  9  degrees of freedom
## AIC: 49.808
##
## Number of Fisher Scoring iterations: 4
```

Because the response is binomially distributed with m $>$ 1, we can test on overdispersion:

```r
1-pchisq(10.331, 9) # if resulting value > 0.05 -> no overdispersion
```

```
## [1] 0.3243594
```

Because the p-value is $>$ than the significance level of 5% we have no evidence against the null hypothesis that $\phi = 1$ -> no overdispersion.

Or altenatively:

```r
qchisq(0.95, df=9) # if resulting value > Residual deviance -> no overdispersion
```

```
## [1] 16.91898
```

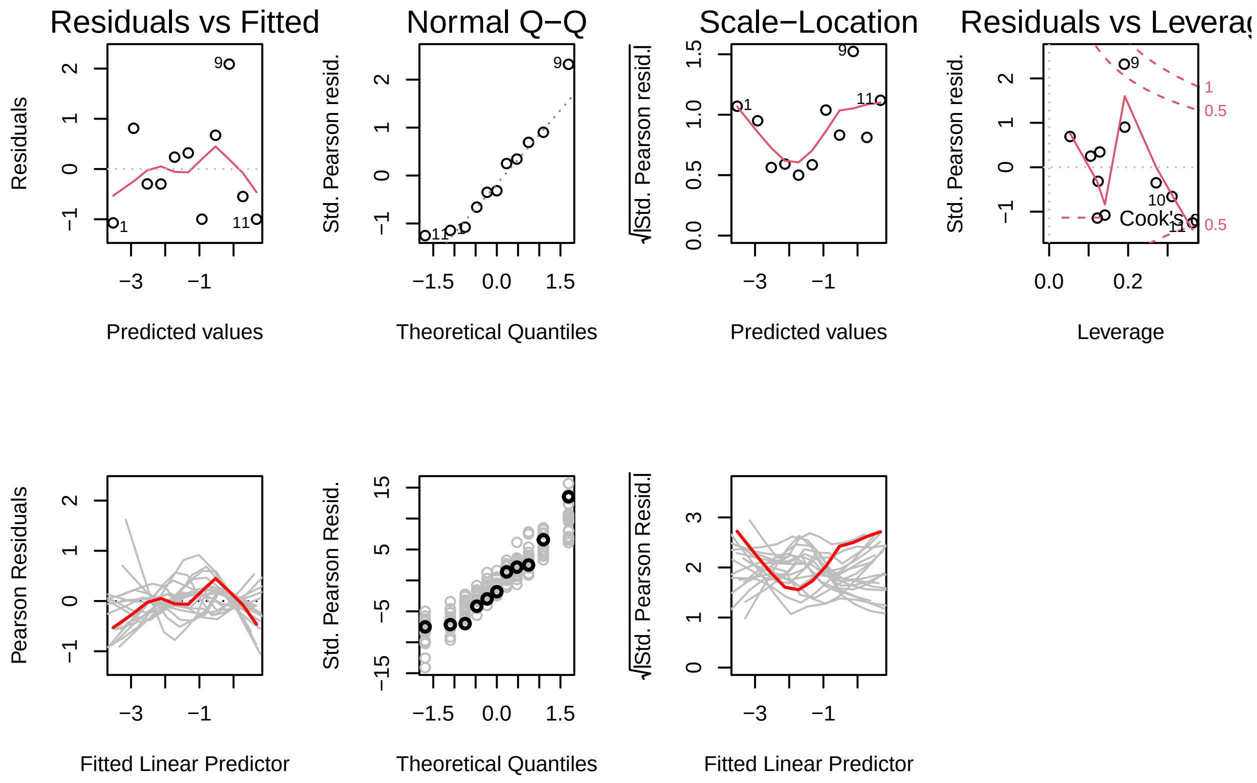
Because the residual deviance is smaller than $q_{0.95}^{\chi_9^2}$ the null hypothesis that $\phi = 1$ cannot be rejected -> no overdispersion.

Question 1.a)

Is the conclusion in the two cells abvoe right? Are these two different methods to come to the same result (looking for overdispersion)?

Performing a residual and sensitivity analysis for the fitted model:

```r
par(mfrow=c(2,4))
plot(glm1.1)
plot.glmSim(glm1.1, SEED = 1)
```



Interpretation:

1. Tukey-Anscombe plot: The smoother shows a banana form, however in the stochastic fluctuation the smoother is not extreme.
   => The assumption of constant expectation is not violated.
2. Q-Q plot: The data scattersnot fully around the straight line but is within the stochastic fluctuation.
   => The assumption of Gaussian distributed errors is violated.
3. Scale-location plot: The smoother has a banana form and stays within the stochastic fluctuation.
   => There is no evidence against the assumption of constant variance of the residuals.
4. Residuals vs. Leverage: All observations have Cook's Distance <1 and therewith no too influential points are present. Leverage points > 2 * 2 [nr. of coefficients] / 11 [nr. of observations] = 0.3636364.
   => No too influential (dangerous) observations

CONCLUSION: The model does fit the data adequately.

Exercise 1.b)

Premature Birth Data (cf. Exercise 2 on Worksheet Week 4) Does the logit model that you have fitted in part 2(c) model the data adequately?

```
path <- file.path('Datasets', 'birth-weight.dat')
df <- read.table(path, header = TRUE)

# Fitted moodel from 2.c), w4
df$lWeight <- log(df$weight)
glm2.1 <- glm(cbind(Y, m-Y) ~ lWeight, family = binomial, data = df)
summary(glm2.1)
```

```
##
## Call:
## glm(formula = cbind(Y, m - Y) ~ lWeight, family = binomial, data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.70057  -0.78559  -0.05153   0.39821   1.41968
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -34.7008     4.7860  -7.250 4.15e-13 ***
## lWeight       5.1193     0.6952   7.363 1.79e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 87.0456  on 9  degrees of freedom
## Residual deviance:  8.7335  on 8  degrees of freedom
## AIC: 41.721
##
## Number of Fisher Scoring iterations: 4
```

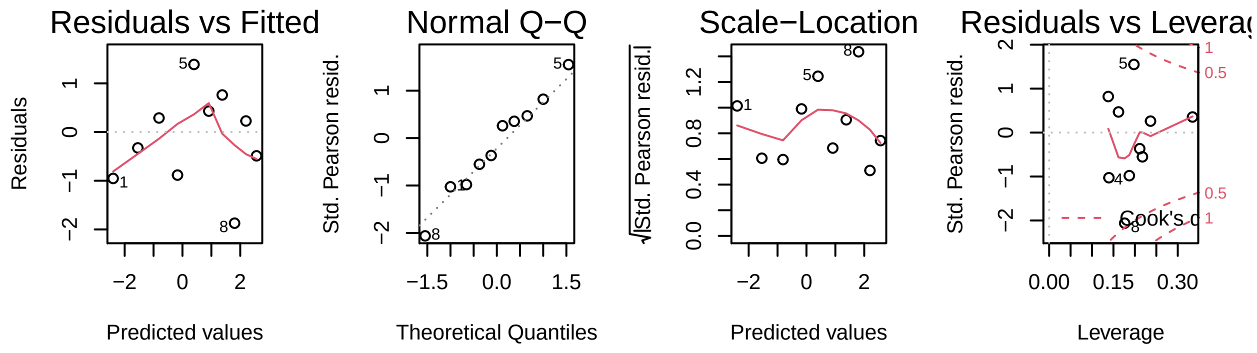
```
# Checking for overdispersion
1 - pchisq(8.7335, 8)
```

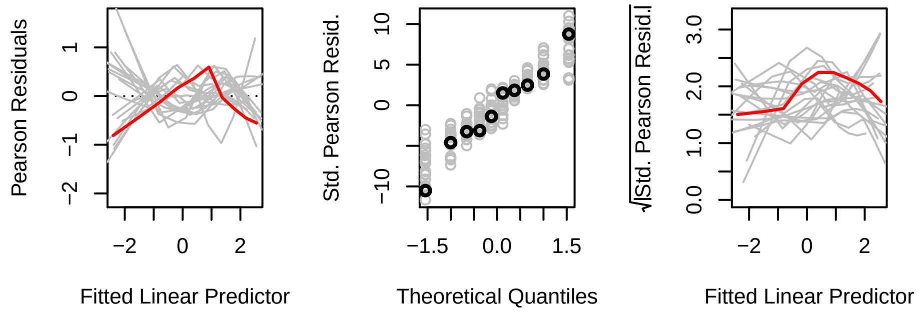
```
## [1] 0.365274
```

Because the p-value is > than the significance level of 5% we have no evidence against the null hypothesis that $\phi = 1$ -> no overdispersion.

Residual and sensitivity analysis:

```
par(mfrow=c(2,4))
plot(glm2.1)
plot.glmSim(glm2.1, SEED = 1)
```





None of the model assumptions is violated, no leverage points (>0.4) and no observations with Cook's Distance > 1.

-> The model fits the data adequately.

Exercise 1.c)

```
path <- file.path('Datasets', 'Dial-a-ride.dat')
df <- read.table(path, header=TRUE)
df <- df[-c(1, 33, 35, 40, 45, 53),]

df$sPOP <- sqrt(df$POP)
df$lAR <- log(df$AR)
df$hHR <- ifelse(df$VH <= 12,0,  df$HR - 12)
df$hVH <- ifelse(df$VH <= 7,0,  df$VH - 7)
df$fF <- as.factor(cut(df$F, breaks=c(0,0.1,0.4,0.7,1)))

glm1.c <- glm(RDR ~ sPOP + lAR + HR + hHR + VH + hVH + fF + IND,
                  family = poisson, data = df)
summary(glm1.c)
```

```
##
## Call:
## glm(formula = RDR ~ sPOP + lAR + HR + hHR + VH + hVH + fF + IND,
##     family = poisson, data = df)
```

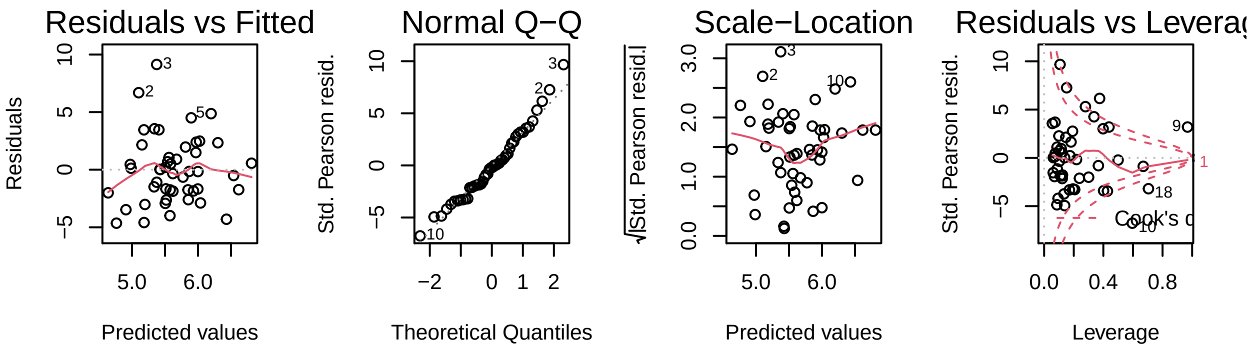
```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0538  -1.8801  -0.1566   1.5864   8.3756
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.4560574  0.0839248  41.180  < 2e-16 ***
## sPOP         0.0027699  0.0002525  10.972  < 2e-16 ***
## lAR         -0.0888673  0.0193671  -4.589 4.46e-06 ***
## HR           0.0399400  0.0032914  12.135  < 2e-16 ***
## hHR         -0.0055008  0.0057019  -0.965    0.335
## VH           0.1491579  0.0096781  15.412  < 2e-16 ***
## hVH         -0.1060012  0.0114986  -9.219  < 2e-16 ***
## fF(0.1,0.4]  0.5319119  0.0463140  11.485  < 2e-16 ***
## fF(0.4,0.7]  0.4902834  0.0436888  11.222  < 2e-16 ***
## fF(0.7,1]    0.7181352  0.0494908  14.510  < 2e-16 ***
## IND          0.4151457  0.0198252  20.940  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3958.4  on 47  degrees of freedom
## Residual deviance:  387.2  on 37  degrees of freedom
## AIC: 766.85
##
## Number of Fisher Scoring iterations: 4
```

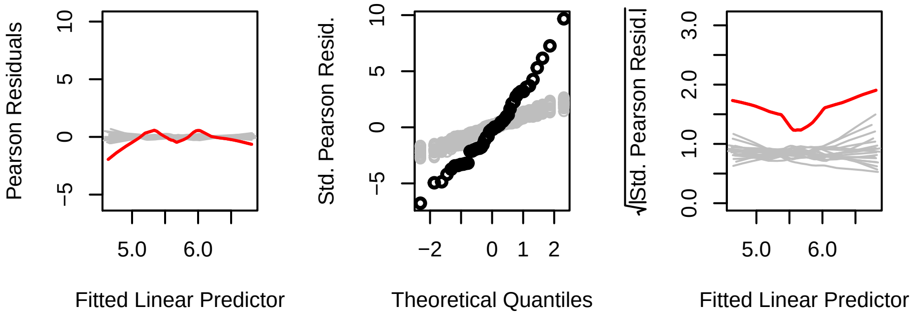
Because the Residual deviance is > 10 times bigger than the corresponding degrees of freedom, it is very clear that there is overdispersion (-> When residual deviance is < than degrees of freedom, there is no overdispersion, otherwise there is.)

```r
par(mfrow=c(2,4))
plot(glm1.c)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs wurden erzeugt

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs wurden erzeugt
```

```r
plot.glmSim(glm1.c, SEED = 1)
```





Interpretation:

1. Tukey-Anscombe plot: The smoother shows an M form which is outside the stochastic fluctuation.
   => The assumption of constant expactation is violated.
2. Q-Q plot: The data scatters around the straight line but is by far not within the stochastic fluctuation.
   => The assumption of Gaussian distributed errors is violated.
3. Scale-location plot: The smoother has a v form and is way above the stochastic fluctuation.
   => The assumption of constant residual variance is violated.
4. Residuals vs. Leverage: There are several observations with Cook's Distance >1 and therewith too influential points are present. Leverage points > 2 * 11 [nr. of coefficients] / 48 [nr. of observations] = 0.4583333 are also apparent.
   => There are too influential (dangerous) observations present

CONCLUSION: The model does not fit the data adequately at all.

-> Better model is presented in the slides of week 8.

Exercise 1.d)

```
df <- read.table("Datasets/transactions.dat", header=T)
str(df)
```

```
## 'data.frame':    261 obs. of  3 variables:
```

```
##  $ Time : int  239627 234827 240326 1351841 1343674 791448 911080 581843 1224988 729993 ...
##  $ Type1: int  0 0 0 51585 62300 39485 40785 24390 53832 1 ...
##  $ Type2: int  116566 165576 89944 331481 396920 308698 292478 148670 409208 279849 ...
```

```
glm1.d  <- glm(Time ~ Type1 + Type2, family=Gamma(link=identity), data=df)
summary(glm1.d)
```

```
## 
## Call:
## glm(formula = Time ~ Type1 + Type2, family = Gamma(link = identity), 
##     data = df)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -0.46888  -0.10719   0.00193   0.08619   0.67961  
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.536e+04  5.183e+03   2.964  0.00332 ** 
## Type1       5.705e+00  4.257e-01  13.401  < 2e-16 ***
## Type2       2.007e+00  5.803e-02  34.582  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for Gamma family taken to be 0.02938966)
## 
##     Null deviance: 92.602  on 260  degrees of freedom
## Residual deviance:  7.478  on 258  degrees of freedom
## AIC: 6725.5
## 
## Number of Fisher Scoring iterations: 4
```

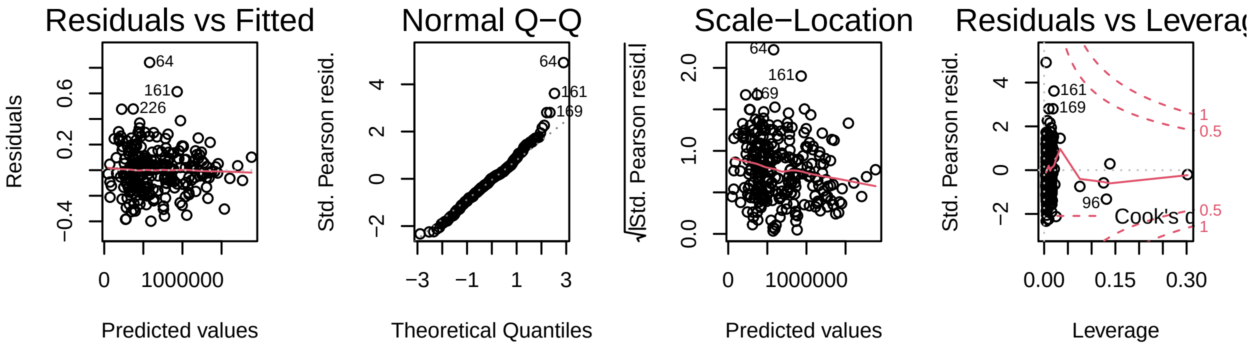
```
1-pchisq(7.478, 258)
```

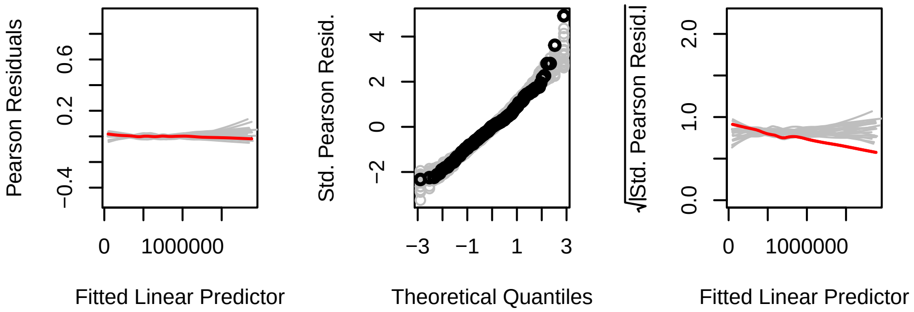
```
## [1] 1
```

There is no overdispersion present (p-value > 0.05).

Residual and sensitivity analysis:

```
par(mfrow=c(2,4))
plot(glm1.d)
plot.glmSim(glm1.d, SEED = 1)
```





Interpretation Tukey Anscombe and Q-Q plot look fine. But the smoother in the scale location plot is outside the stochastic fluctuation. Even none of the observations has Cook's Distance > 1 in the Residuals vs. Leverage plot there are some leverage points apparent with leverage > 0.0842912 but they are not dangerous

CONCLUSION: The model might not fit the data fully adequat.

Exercise 1.e)

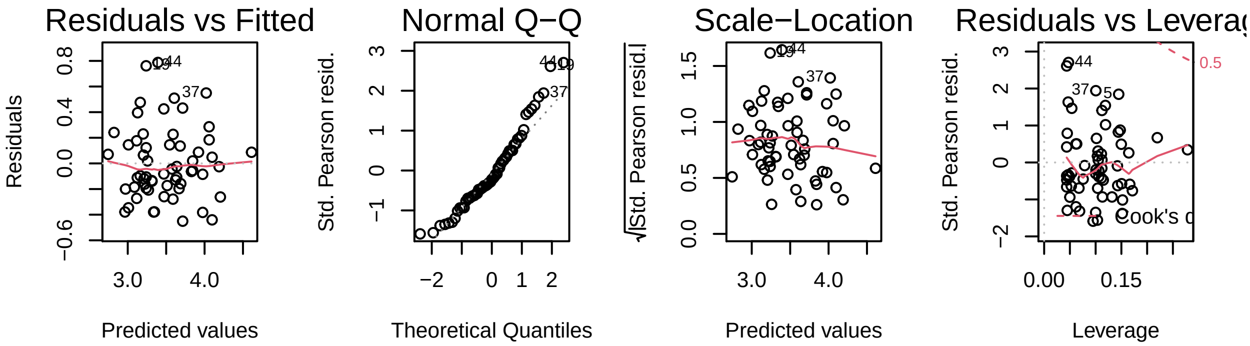
```
df <- read.table("Datasets/nambeware.txt", header=T)
summary(df)
```

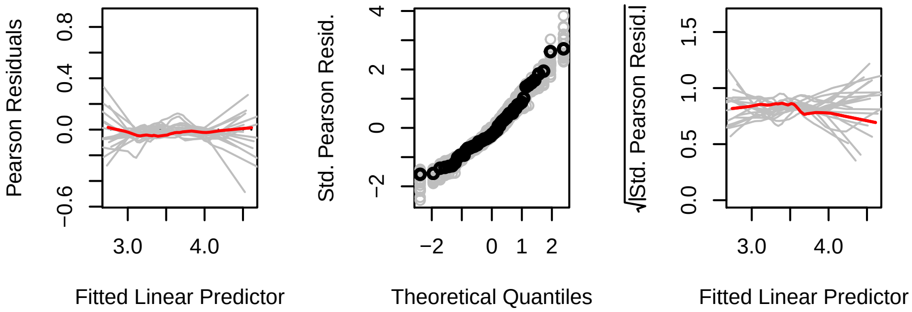
```
##      Type                Diam           Time             Price       
##  Length:59          Min.   : 5.00   Min.   : 12.02   Min.   : 21.50  
##  Class :character   1st Qu.: 8.25   1st Qu.: 22.21   1st Qu.: 47.25  
##  Mode  :character   Median :11.00   Median : 31.46   Median : 75.00  
##                     Mean   :10.93   Mean   : 35.82   Mean   : 86.38  
##                     3rd Qu.:13.00   3rd Qu.: 45.03   3rd Qu.:107.00  
##                     Max.   :25.00   Max.   :109.38   Max.   :260.00
```

```
glm1.d <- glm(Time ~ Diam + Type, family=Gamma(link=log), data=df)
summary(glm1.d)
```

```
##
## Call:
## glm(formula = Time ~ Diam + Type, family = Gamma(link = log),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.54489  -0.20244  -0.06442   0.13852   0.64306
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.54897    0.12721  20.038  < 2e-16 ***
## Diam         0.07671    0.01176   6.525 2.62e-08 ***
## TypeCassDish 0.47516    0.11855   4.008 0.000193 ***
## TypeDish     0.28940    0.12894   2.244 0.029000 *
## TypePlate   -0.18791    0.11847  -1.586 0.118639
## TypeTray     0.14472    0.12652   1.144 0.257816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.08899162)
##
##     Null deviance: 14.0053  on 58  degrees of freedom
## Residual deviance:  4.5039  on 53  degrees of freedom
## AIC: 438.65
##
## Number of Fisher Scoring iterations: 4
```

```r
par(mfrow=c(2,4))
plot(glm1.d,
     panel = function(x,y) panel.smooth(x, y, iter = 1, span = 0.6))
plot.glmSim(glm1.d, SEED = 1,
            smoother = function(x,y) lowess(x, y, iter = 1, f = 0.6))
```





Interpretation None of the model assumptions is violated. There are 2 leverage points with leverage > 0.2033898.

CONCLUSION: The model might be adequate.

Exercise 2

In this exercise, data are presented on the number of fractures (Y) that occur in the upper seams of coal mines in the Appalachian region of western Virginia. Four explanatory variables were reported:

INB inner burden thickness [feet], the shortest distance between seam floor and the lower seam

EXTRP percent extraction of the lower previously mined seam

sHeight lower seam height [feet]

oTime time [year] that the mine has been in operation

```
path <- file.path('Datasets', 'mine.dat')
df <- read.table(path, header = TRUE)
str(df)
```

```
## 'data.frame':    44 obs. of  5 variables:
##  $ Y      : int  2 1 0 4 1 2 0 0 4 4 ...
##  $ INB    : int  50 230 125 75 70 65 65 350 350 160 ...
##  $ EXTRP  : int  70 65 70 65 65 70 60 60 90 80 ...
##  $ sHeight: int  52 42 45 68 53 46 62 54 54 38 ...
##  $ oTime  : num  1 6 1 0.5 0.5 3 1 0.5 0.5 0 ...
```

Exercise 2.a)

Execute the following R command:

```
glm2a <- glm(Y ~ INB + EXTRP + oTime, family = poisson, data = df)
summary(glm2a)
```

```
##
## Call:
## glm(formula = Y ~ INB + EXTRP + oTime, family = poisson, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7727  -0.9073  -0.0107   0.2716   2.1783
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.7206821  0.9788770  -3.801 0.000144 ***
## INB         -0.0014793  0.0008244  -1.794 0.072757 .
## EXTRP        0.0627011  0.0122711   5.110 3.23e-07 ***
## oTime       -0.0316514  0.0163095  -1.941 0.052298 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 74.984  on 43  degrees of freedom
## Residual deviance: 38.031  on 40  degrees of freedom
## AIC: 142.3
##
## Number of Fisher Scoring iterations: 5
```

- The response Y_i is ~ Pois(λ_i), independent with $E(Y_i) = \mu_i$.
- The explanatory variables are IND, EXTRP and oTimes.
- A linear combination of them yield the linear predictor η_i.
- The canonical link, log(), is used: $\log(\mu_i) = \eta_i$
- Estimated coefficients:
  -3.7206821, -0.0014793, 0.0627011, -0.0316514

Exercise 2.b)

Does the residual deviance indicate that the model from part (a) is satisfactory?

The residual deviance is smaller then the degrees of freedom -> so yes this is satisfactory because no overdispersion is present.

Exercise 2.c)

Find approximate 95% Wald confidence intervals on the model parameters and compare them with the 95% profile confidence intervals.

Wald confidence intervals

```
coffs <- summary(glm2a)$coefficients
round(cbind(coffs[,1] - 1.96 * coffs[,2], coffs[,1] + 1.96 * coffs[,2]), 4)
```

```
##                [,1]    [,2]
## (Intercept) -5.6393 -1.8021
## INB         -0.0031  0.0001
## EXTRP        0.0386  0.0868
## oTime       -0.0636  0.0003
```

Profile confidence intervals:

```
round(confint(glm2a), 4)
```

```
## Waiting for profiling to be done...
```

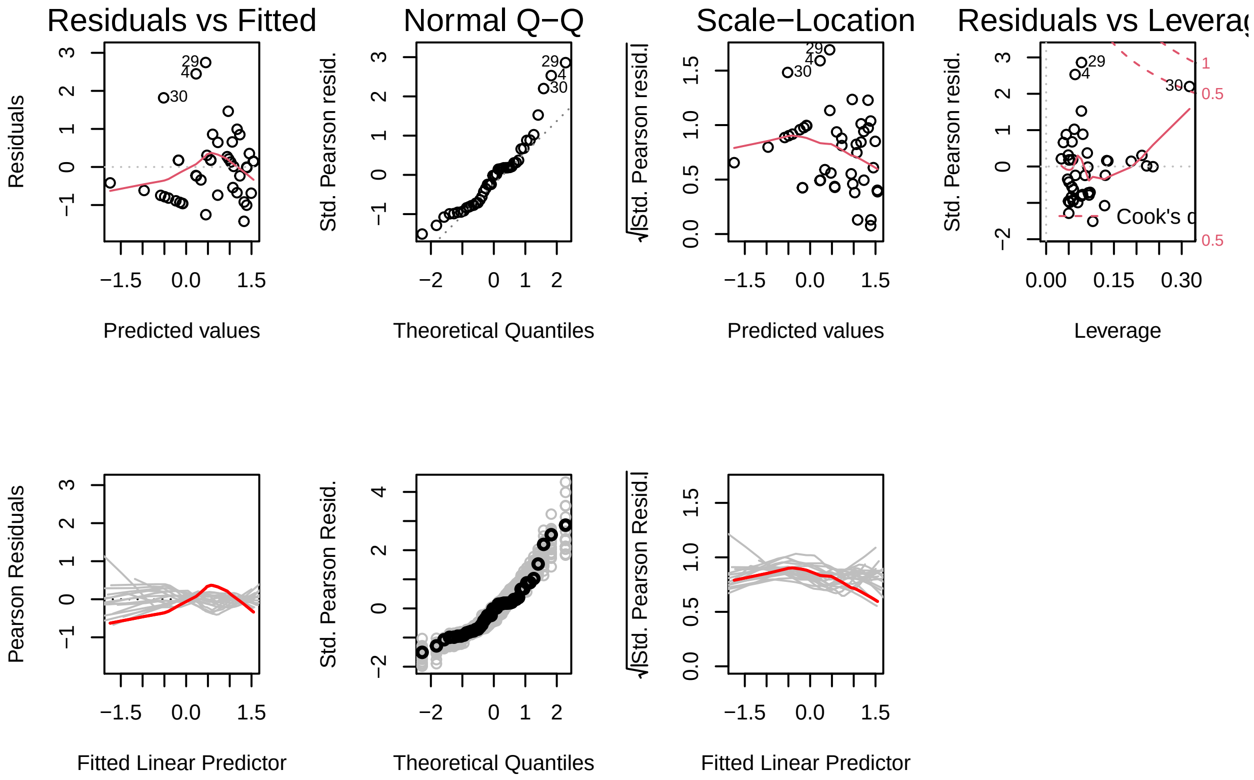
```
##               2.5 %  97.5 %
## (Intercept) -5.7400 -1.8914
## INB         -0.0032  0.0000
## EXTRP        0.0396  0.0879
## oTime       -0.0649 -0.0008
```

There are slight differences in the results. Since we would trust more the profiling data, we would youse this values for further proceedings.

Exercise 2.d)

Perform a thorough residual analysis of the fitted model from part (a).

```
par(mfrow=c(2,4))
plot(glm2a)
plot.glmSim(glm2a, SEED = 1)
```



Interpretation There is evidence in the Tukey Anscombe plot that the assupmtion of constant expactation of the error is violated since the smoother is outside the stochastic fluctuation. The other model assumptions are within the stochastic fluctuation and no observation with Cook's Distance > 1 is visible.

CONCLUSTION: The model might not fit the data fully adequat. -> Trying to improve the model in e)

Exercise 2.e)

Improve the model from a).

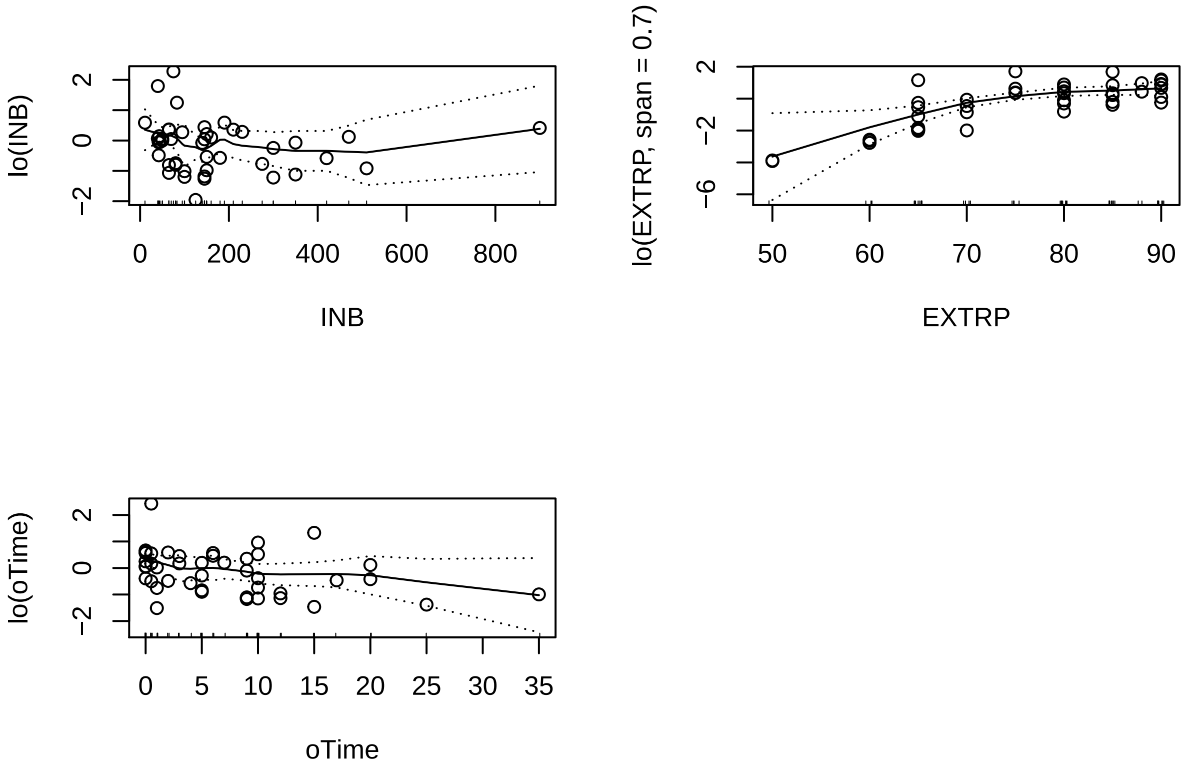
```
library(gam)
```

```
## Loading required package: splines
```

```
## Loading required package: foreach
```

```
## Loaded gam 1.20
```

```
gam2e <- gam(Y ~ lo(INB) + lo(EXTRP, span=0.7) + lo(oTime), family=poisson,
                data=df, bf.maxit=500)
par(mfrow=c(2,2))
plot(gam2e, se=TRUE, residuals=TRUE)
```



The plots do not strongly indicate the needs of transforming INB. But the log-transformation of INB will improve the plots. The variable EXTRP needs a transformation. For lack of a better solution, an upside-down hockey stick transformation is applied. (This needs some justifications by the subject matter experts.)

Question 2.e)

- What does an upside-down hockey stick transformation mean? Why are we allowed to transform the explanatory variables like that? Does it not completely change the data?

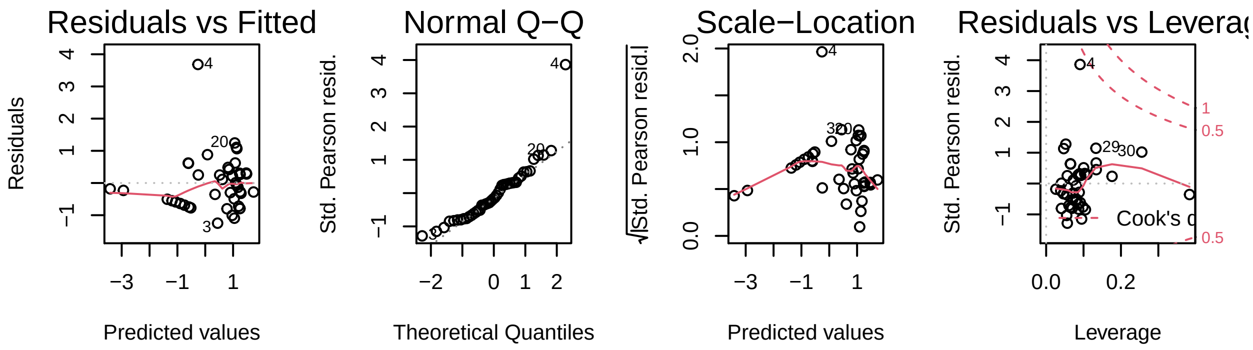
```
df$lINB <- log(df$INB)
df$tEXTRP <- ifelse(df$EXTRP >= 75, 75, df$EXTRP)

# Fitting again the model
glm2e <- glm(Y ~ lINB + tEXTRP + oTime, family=poisson, data=df)
summary(glm2e)
```

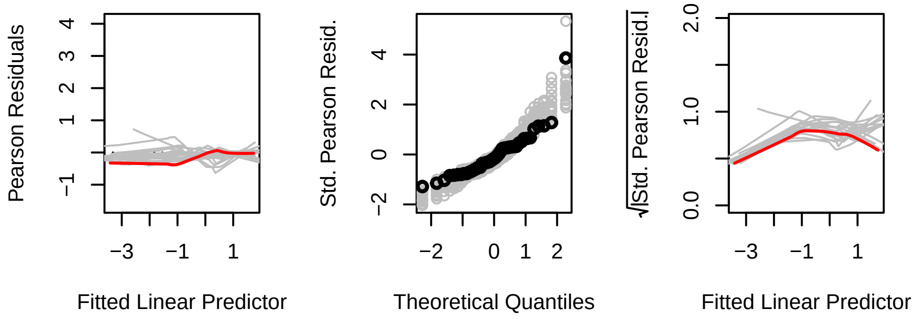
```
##
## Call:
## glm(formula = Y ~ lINB + tEXTRP + oTime, family = poisson, data = df)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7657  -0.7943  -0.1957   0.2970   2.5934
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.01569    2.71324  -3.691 0.000223 ***
## lINB         -0.18668    0.11121  -1.679 0.093227 .
## tEXTRP        0.16264    0.03561   4.567 4.94e-06 ***
## oTime        -0.02377    0.01535  -1.548 0.121661
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 74.984  on 43  degrees of freedom
## Residual deviance: 28.287  on 40  degrees of freedom
## AIC: 132.56
##
## Number of Fisher Scoring iterations: 5
```

Summary looks just fine except, that according to Walds p-value at least one of the explanatory variable would be superfluous.

```r
par(mfrow=c(2,4))
plot(glm2e,
     panel = function(x,y) panel.smooth(x, y, iter = 1, span = 0.6))
plot.glmSim(glm2e, SEED = 1,
            smoother = function(x,y) lowess(x, y, iter = 1, f = 0.8))
```





This's much better. Since the structures are within the stochastic fluctuations, there is no evidence that the model is not yet adaquate. According to the Wald-type inference results there are at least one non-significant explanatory variable:

```
step(glm2e)
```

```
## Start:  AIC=132.56
## Y ~ lINB + tEXTRP + oTime
##
##          Df Deviance    AIC
## <none>        28.287 132.56
## - oTime   1   30.860 133.13
## - lINB    1   31.082 133.35
## - tEXTRP  1   65.803 168.07

##
## Call:  glm(formula = Y ~ lINB + tEXTRP + oTime, family = poisson, data = df)
##
## Coefficients:
## (Intercept)         lINB       tEXTRP        oTime
##   -10.01569     -0.18668      0.16264     -0.02377
##
## Degrees of Freedom: 43 Total (i.e. Null);  40 Residual
## Null Deviance:       74.98
## Residual Deviance: 28.29     AIC: 132.6
```

According to the AIC all explanatory variables are needed.

Exercise 3

An electric utility is interested in developing a model relating peak hour demand (Y) to total energy usage (x, in kilowatt-hours) during the month. This is an important planning problem because while most customers pay directly for energy usage, the generation system must be large enough to meet the maximum demand imposed.

```
path <- file.path('Datasets', 'eUsage.dat')
df <- read.table(path, header = TRUE)
str(df)
```

```
## 'data.frame':    53 obs. of  2 variables:
##  $ Y: num  0.79 0.44 0.56 0.79 2.7 3.64 4.73 9.5 5.34 6.85 ...
##  $ x: int  679 292 1012 493 582 1156 997 2189 1097 2078 ...
```

Exercise 3.a)

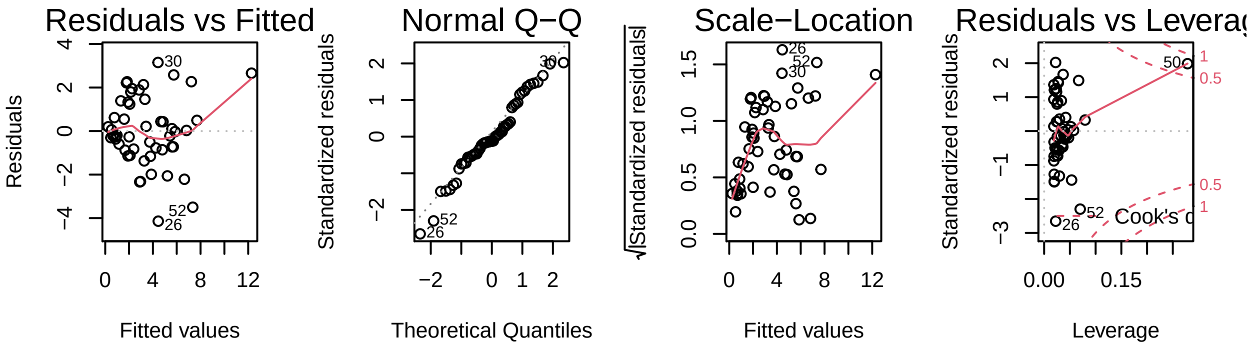
Assume that the response Y is independently Gaussian distributed and fit a simple linear regression model. What are the estimated coefficients?

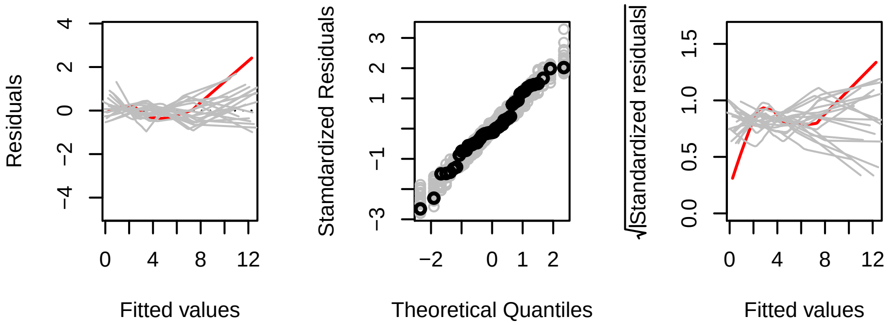
```
lm3a <- lm(Y ~ x, data = df)
coef(lm3a)
```

```
##  (Intercept)            x
## -0.831303660  0.003682843
```

Perform a thorough residual analysis. What are your conclusions?

```
par(mfrow=c(2, 4))
plot(lm3a)
plot.lmSim(lm3a, SEED = 1)
```





Interpretation:

1. Tukey-Anscombe plot: The smoother shows an increasing trend on the r.h.s. and the smoother itself is outside the stochastic fluctuation.
   => The assumption of constant expactation is violated.
2. Q-Q plot: The data scatters around the straight line and is within the stochastic fluctuation.
   => The assumption of Gaussian distributed errors is not violated.
3. Scale-location plot: The smoother is not a straight line and is outside the stochastic fluctuation.
   => The assumption of constant residual variance is violated.
4. Residuals vs. Leverage: There are no observations with Cook's Distance >1 and therewith no too influential points are present. Leverage points > 2 * 2 [nr. of coefficients] / 53 [nr. of observations] = 0.0754717 is one apparent.
   => There are no too influential (dangerous) observations present

CONCLUSION: The model does not fit the data adequately.

Exercise 3.b)

Assume that the response Y is independently gamma distributed and use the identity link for fitting the GLM. What are the estimated coefficients?

```
glm3b <- glm(Y ~ x, family = Gamma(link=identity), data = df)
summary(glm3b)
```

```
##
## Call:
## glm(formula = Y ~ x, family = Gamma(link = identity), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8659  -0.4543  -0.1368   0.3733   0.8969
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.7675127  0.1894508  -4.051 0.000174 ***
## x            0.0036620  0.0003701   9.895 1.85e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.2810353)
##
##     Null deviance: 46.22  on 52  degrees of freedom
## Residual deviance: 18.05  on 51  degrees of freedom
## AIC: 186.23
##
## Number of Fisher Scoring iterations: 7
```

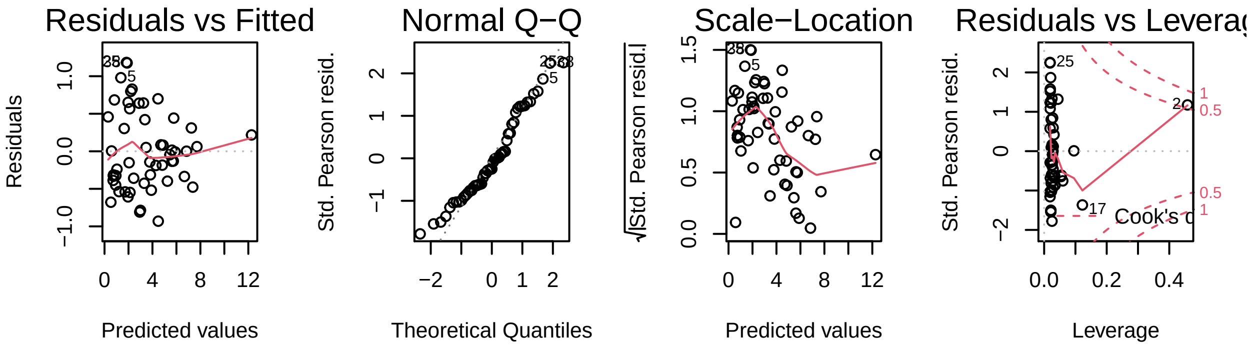
```
coef(glm3b)
```

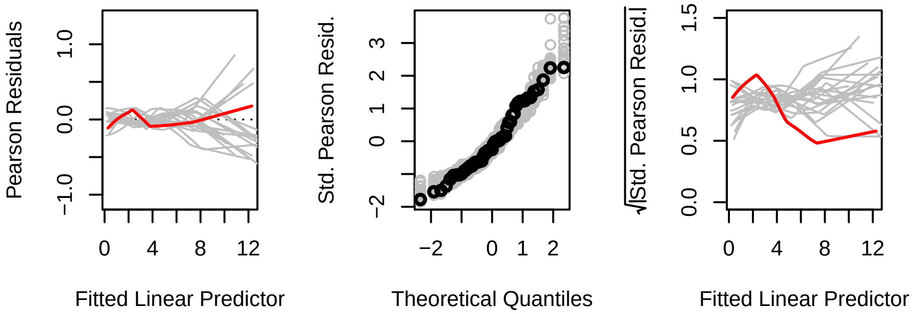
```
##  (Intercept)            x
## -0.767512719  0.003661987
```

The coefficient estimates are very similar to the ones in 3.a)

Is this model more adequate?

```
par(mfrow=c(2,4))
plot(glm3b)
plot.glmSim(glm3b, SEED = 1)
```





Interpretation:

1. Tukey-Anscombe plot: The smoother shows an almost straight line and is within the stochastic fluctuation.
   => The assumption of constant expectation is not violated.
2. Q-Q plot: The data scatters around the straight line and is within the stochastic fluctuation.
   => The assumption of Gaussian distributed errors is not violated.
3. Scale-location plot: The smoother is not a straight line and is outside the stochastic fluctuation.
   => The assumption of constant residual variance is violated.
4. Residuals vs. Leverage: There are no observations with Cook's Distance >1 and therewith no too influential points are present. Leverage points > 2 * 2 [nr. of coefficients] / 53 [nr. of observations] = 0.0754717 are 3 apparent but they do not harm.
   => There are no too influential (dangerous) observations present

CONCLUSION: The model does not fit the data adequately. But maybe a bit better than the one in 3.a).

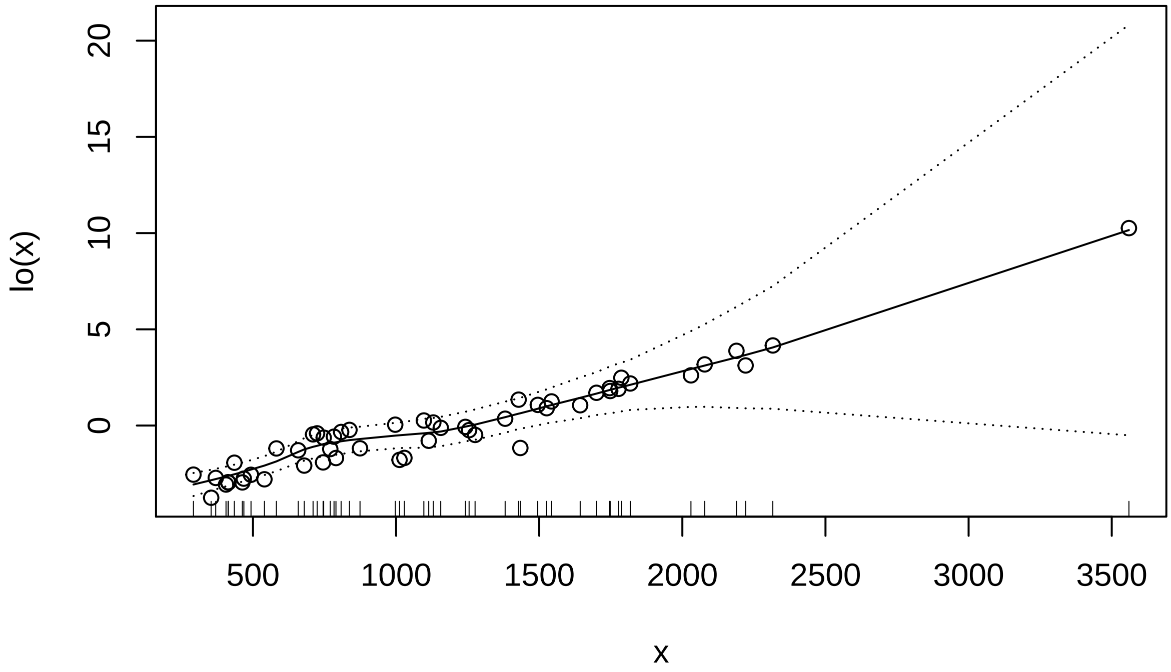
Exercise 3.c)

Clarify with a "GAM-Fit" whether the explanatory variable should be transformed.

```
library(gam)
gam3c <- gam(Y ~ lo(x), family = Gamma(link = identity), data = df)
summary(gam3c)
```

```
##
## Call: gam(formula = Y ~ lo(x), family = Gamma(link = identity), data = df)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80523 -0.41232 -0.01239  0.32903  0.72948
##
## (Dispersion Parameter for Gamma family taken to be 0.2637)
##
##     Null Deviance: 46.2196 on 52 degrees of freedom
## Residual Deviance: 16.96 on 47.6147 degrees of freedom
## AIC: 189.5169
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##               Df Sum Sq Mean Sq F value    Pr(>F)
## lo(x)      1.000 28.461 28.4612  107.92 7.896e-14 ***
## Residuals 47.615 12.557  0.2637
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Anova for Nonparametric Effects
##             Npar Df Npar F  Pr(F)
## (Intercept)
## lo(x)           3.4 1.0314 0.3932
```

```
par(mfrow=c(1,1))
plot(gam3c, se = TRUE, residuals = TRUE)
```



The estimated curve can be approximated well by a straight line and therefore no transformation is needed.

Exercise 3.d)

Find approximate 95% Wald confidence intervals on the slope parameter and compare it with the corresponding 95% profile confidence interval.

Wald Confidence Interval (CI)

```
coefWald <- summary(glm3b)$coefficients
round(c(coefWald[1] - 1.96 * coefWald[2], coefWald[1] + 1.96 * coefWald[2]), 4)
```

```
## [1] -0.7747 -0.7603
```

Profile CI

```
confint(glm3b)[2,]
```

```
## Waiting for profiling to be done...
```

```
## Warning: glm.fit: algorithm did not converge
```

```
##       2.5 %      97.5 %
## 0.002984377 0.004404863
```

Algorithm need more iterations as can be seen in the warning message. So we fit the model again with more iterations:

```
glm3b <- glm(Y ~ x, family = Gamma(link=identity), data = df, maxit = 500)
round(confint(glm3b)[2,], 4)
```

```
## Waiting for profiling to be done...
```

```
##  2.5 % 97.5 %
## 0.0030 0.0044
```

On the lower end a small difference can be seen. On the higher end the estimated CI's are the same. Since we have more confidenct in the profiling approach we would further proceed with the profiling CI.

Exercise 3.e)

Could the response (Y) be exponentially distributed? (Please note all steps of your consideration.)

If the response is exponentially distributed then the dispersion parameter is fixed at 1: So we can test the null hypothesis phi=1 as in the Poisson and the binomial case.

From the summary output we have:
Residual deviance: 18.05 on 51 degrees of freedom

Since the residual deviance is outside of the acceptance region:

```
qchisq(c(0.025,0.975), 51)
```

```
## [1] 33.16179 72.61599
```

the null hypothesis must be rejected on the 5% level. Hence the response cannot be exponentially distributed, it must be a gamma distribution.

Question 3.e)

- Could you explain this exercise in detail? Why would the dispersion parameter be fixed at 1 when the response is exponentially distributed? And how do we come up with the right test (qchisq())?
