AdvStDaAn, Worksheet, Week 6

Michael Lappert

24 April, 2022

Contents

Exercise 1

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

summary(df)
```

Dataset loading and sanity check:

```
##     Survival          Weight          Age            Apgar1     
##  Min.   :0.0000   Min.   : 540   Min.   :20.00   Min.   :0.000  
##  1st Qu.:0.0000   1st Qu.: 860   1st Qu.:26.00   1st Qu.:3.000  
##  Median :1.0000   Median :1070   Median :28.00   Median :5.000  
##  Mean   :0.6518   Mean   :1075   Mean   :28.04   Mean   :4.652  
##  3rd Qu.:1.0000   3rd Qu.:1320   3rd Qu.:30.00   3rd Qu.:6.000  
##  Max.   :1.0000   Max.   :1500   Max.   :37.00   Max.   :9.000  
##      Apgar5             pH       
##  Min.   : 0.000   Min.   :6.830  
##  1st Qu.: 5.000   1st Qu.:7.270  
##  Median : 6.000   Median :7.340  
##  Mean   : 6.194   Mean   :7.323  
##  3rd Qu.: 7.000   3rd Qu.:7.380  
##  Max.   :10.000   Max.   :7.600
```

```
str(df)
```

```
## 'data.frame':    247 obs. of  6 variables:
##  $ Survival: int  1 0 0 0 0 0 1 1 0 1 ...
##  $ Weight  : int  1350 725 1090 1300 1200 590 1500 1360 600 1410 ...
##  $ Age     : int  32 27 27 24 31 22 32 29 24 30 ...
##  $ Apgar1  : int  4 5 5 9 5 9 9 9 4 4 ...
##  $ Apgar5  : int  7 6 7 9 5 9 9 9 4 5 ...
##  $ pH      : num  7.25 7.36 7.42 7.37 7.35 7.37 7.29 7.44 7.27 7.35 ...
```

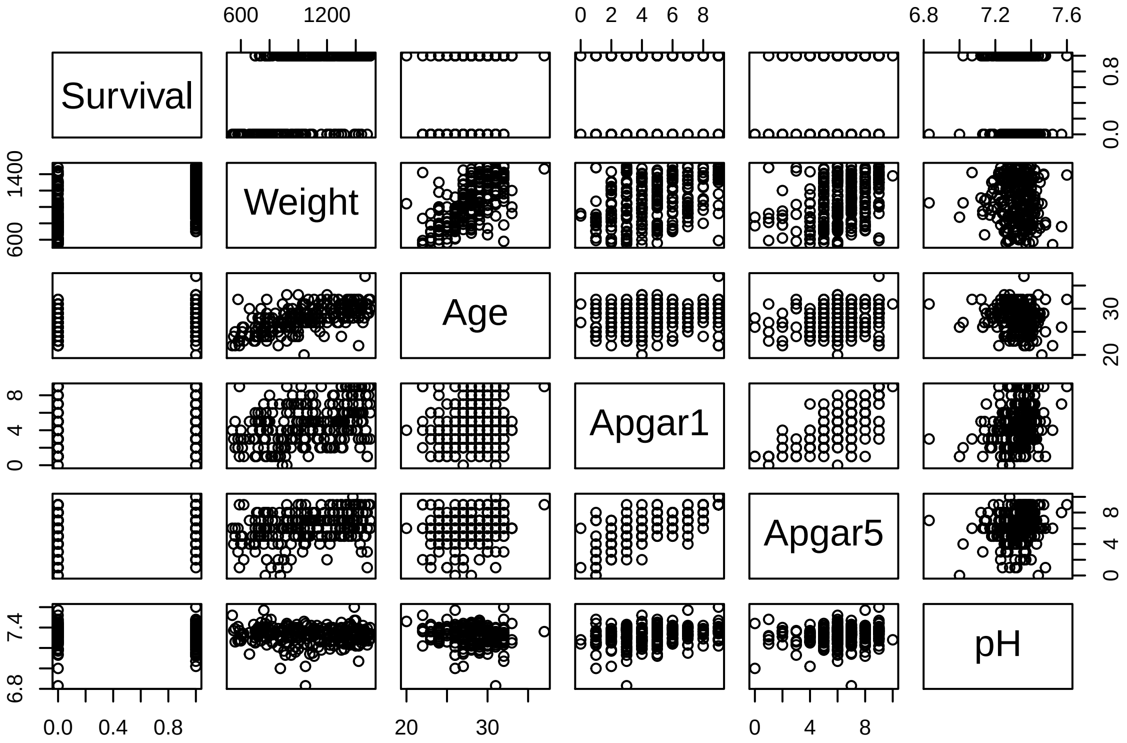
```
head(df)
```

```
##   Survival Weight Age Apgar1 Apgar5   pH
## 1        1   1350  32      4      7 7.25
## 2        0    725  27      5      6 7.36
## 3        0   1090  27      5      7 7.42
## 4        0   1300  24      9      9 7.37
## 5        0   1200  31      5      5 7.35
## 6        0    590  22      9      9 7.37
```

```
tail(df)
```

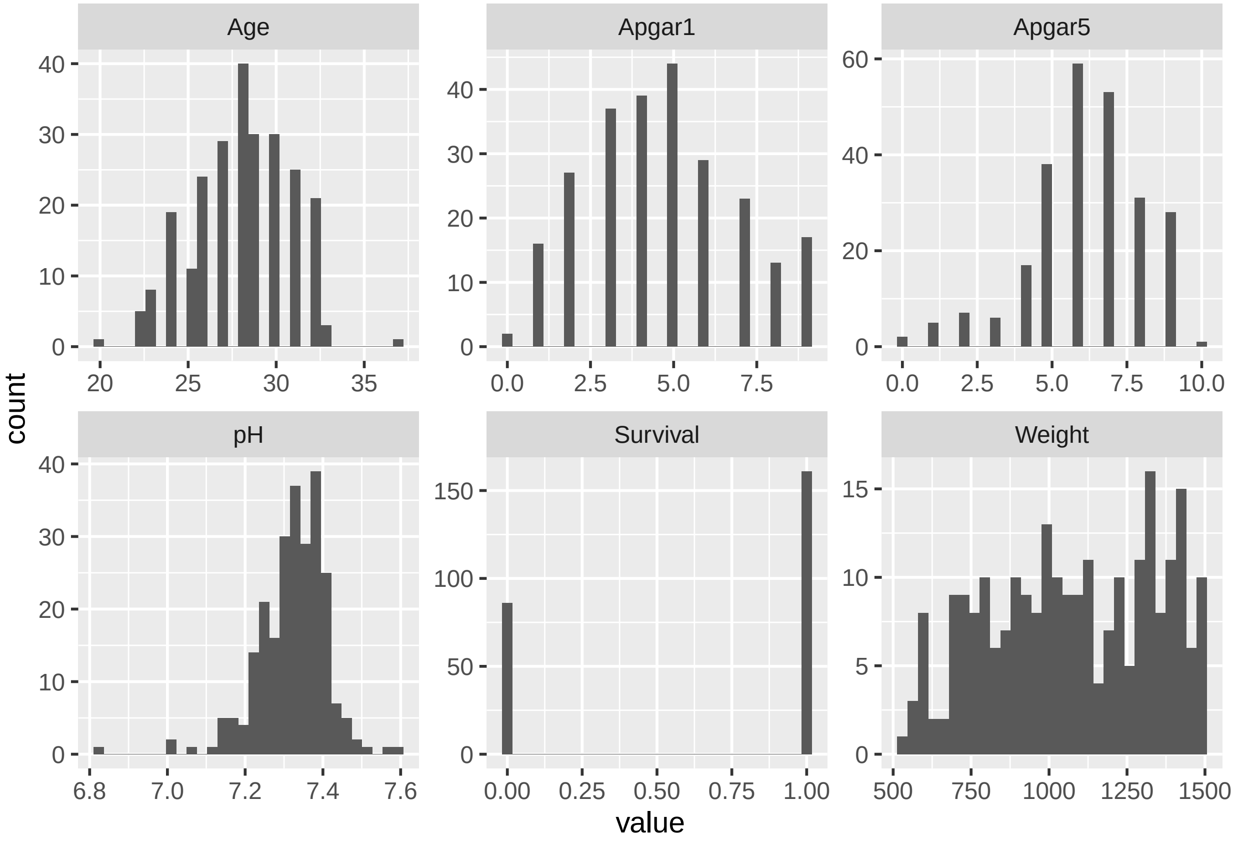
```
##     Survival Weight Age Apgar1 Apgar5   pH
## 242        1   1120  28      7      7 7.33
## 243        1   1020  28      5      7 7.34
## 244        1   1320  28      6      6 7.24
## 245        0    900  27      5      6 7.37
## 246        1   1150  27      4      7 7.37
## 247        0    790  27      4      8 7.35
```

```
plot(df)
```



```r
df %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



Exercise 1.a)

```
glm1.1 <- glm(Survival ~ ., family = binomial, data = df)
summary(glm1.1)
```

```
##
## Call:
## glm(formula = Survival ~ ., family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3994  -0.7393   0.4220   0.7833   1.9445
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.0933685 14.3053767  -0.216   0.8288
## Weight       0.0037341  0.0008468   4.410 1.03e-05 ***
## Age          0.1588001  0.0761061   2.087   0.0369 *
## Apgar1       0.1159864  0.1108339   1.046   0.2953
## Apgar5       0.0611499  0.1202222   0.509   0.6110
## pH          -0.7380214  1.8964578  -0.389   0.6972
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 319.28  on 246  degrees of freedom
## Residual deviance: 236.14  on 241  degrees of freedom
## AIC: 248.14
##
## Number of Fisher Scoring iterations: 4
```

On a first sight, just Weight and Age seem to be sifnificant on the 5% significance level. To test this hypothesis, one must perform a statistical test:

Since (from the summary output)

```
1-pchisq(319.28-236.14, df=246-241) # Compare slide 12&13 from w6
```

```
## [1] 2.220446e-16
```

is smaller than the significant level of 5%, we cannot drop all explanatory variables. At least one of them is significant.

Or without plugging in the numbers explicitly (same as above in other synthax):

```
(h <- summary(glm1.1)$null.deviance - summary(glm1.1)$deviance)
```

```
## [1] 83.1366
```

```
1 - pchisq(h, 246-241)
```

```
## [1] 2.220446e-16
```

This test is identical to

```
glm1.2 <- glm(Survival ~ 1, family=binomial, data = df)
anova(glm1.1, glm1.2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Survival ~ Weight + Age + Apgar1 + Apgar5 + pH
## Model 2: Survival ~ 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       241     236.14
## 2       246     319.28 -5  -83.137 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Where we also conclude that since the p-value of 2.2e-16 is « than the significance level of 0.05 to reject the null hypothesis and assume that the first (full) model describes the data more adequately than the second (empty) one and therefore at least one variable is of significance.

Questions1.a)

- How do we already now, that the response is Bernoulli distributed?

Exercise 1.b)

Performing a stepwise variable selection.

```
glm.step1.1 <- step(glm1.1, scope = list(upper =~ .,
                                         lower =~ 1),
                    direction = 'both')
```

```
## Start:  AIC=248.14
## Survival ~ Weight + Age + Apgar1 + Apgar5 + pH
## 
##          Df Deviance    AIC
## - pH      1   236.29 246.29
## - Apgar5  1   236.40 246.40
## - Apgar1  1   237.25 247.25
## <none>        236.14 248.14
## - Age     1   240.55 250.55
## - Weight  1   257.93 267.93
## 
## Step:  AIC=246.29
## Survival ~ Weight + Age + Apgar1 + Apgar5
## 
##          Df Deviance    AIC
## - Apgar5  1   236.56 244.56
## - Apgar1  1   237.26 245.26
## <none>        236.29 246.29
## + pH      1   236.14 248.14
## - Age     1   241.17 249.17
## - Weight  1   258.35 266.35
## 
## Step:  AIC=244.56
## Survival ~ Weight + Age + Apgar1
## 
##          Df Deviance    AIC
## <none>        236.56 244.56
## - Apgar1  1   239.85 245.85
## + Apgar5  1   236.29 246.29
## + pH      1   236.40 246.40
## - Age     1   241.56 247.56
## - Weight  1   259.10 265.10
```

```
summary(glm.step1.1)
```

```
## 
## Call:
## glm(formula = Survival ~ Weight + Age + Apgar1, family = binomial, 
##     data = df)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.4320  -0.7431   0.4180   0.7694   1.9416  
## 
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -8.4841905  1.8177415  -4.667 3.05e-06 ***
## Weight       0.0037911  0.0008449   4.487 7.22e-06 ***
## Age          0.1652973  0.0745653   2.217   0.0266 *  
## Apgar1       0.1429887  0.0795671   1.797   0.0723 .  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 319.28  on 246  degrees of freedom
## Residual deviance: 236.56  on 243  degrees of freedom
## AIC: 244.56
##
## Number of Fisher Scoring iterations: 5
```

The variables Agpar5 and pH got dropped.

Exercise 1.c)

Fitting a model with the explanatory variables Weight and Age and comparing it with anova at the 5% significance level.

```r
glm1.3 <- glm(Survival ~ Weight + Age, family = binomial, data = df)
summary(glm1.3)
```

```
##
## Call:
## glm(formula = Survival ~ Weight + Age, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max  
## -2.3626  -0.7749   0.4141   0.7842   1.7730  
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -8.0983782  1.7808798  -4.547 5.43e-06 ***
## Weight       0.0041919  0.0008156   5.140 2.75e-07 ***
## Age          0.1593810  0.0734420   2.170     0.03 *  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 319.28  on 246  degrees of freedom
## Residual deviance: 239.85  on 244  degrees of freedom
## AIC: 245.85
##
## Number of Fisher Scoring iterations: 4
```

```r
anova(glm1.1, glm1.3, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: Survival ~ Weight + Age + Apgar1 + Apgar5 + pH
## Model 2: Survival ~ Weight + Age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       241     236.14
## 2       244     239.85 -3  -3.7091   0.2946
```

Since the p-value is 0.29 and therefore bigger than the significance level of 5% the null Hypothesis can not be rejected concluding that both models describe the data in the same adequacy. Therefore one can conclude that the model 'Survival ~ Weight + Age' describes the data statistically equally well as the full one.

Exersice 2

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

summary(df)
```

Dataset loading and sanity check:

```
##      Mode1            Mode2         Failures
##  Min.   : 33.30   Min.   :14.4   Min.   : 9.00
##  1st Qu.: 64.70   1st Qu.:25.3   1st Qu.:15.00
##  Median : 91.90   Median :47.8   Median :22.00
##  Mean   : 93.11   Mean   :48.4   Mean   :19.89
##  3rd Qu.:125.90   3rd Qu.:56.6   3rd Qu.:24.00
##  Max.   :137.00   Max.   :97.6   Max.   :27.00
```

```r
str(df)
```

```
## 'data.frame':    9 obs. of  3 variables:
##  $ Mode1   : num  33.3 52.2 64.7 137 125.9 ...
##  $ Mode2   : num  25.3 14.4 32.5 20.5 97.6 53.6 56.6 87.3 47.8
##  $ Failures: int  15 9 14 24 27 27 23 18 22
```

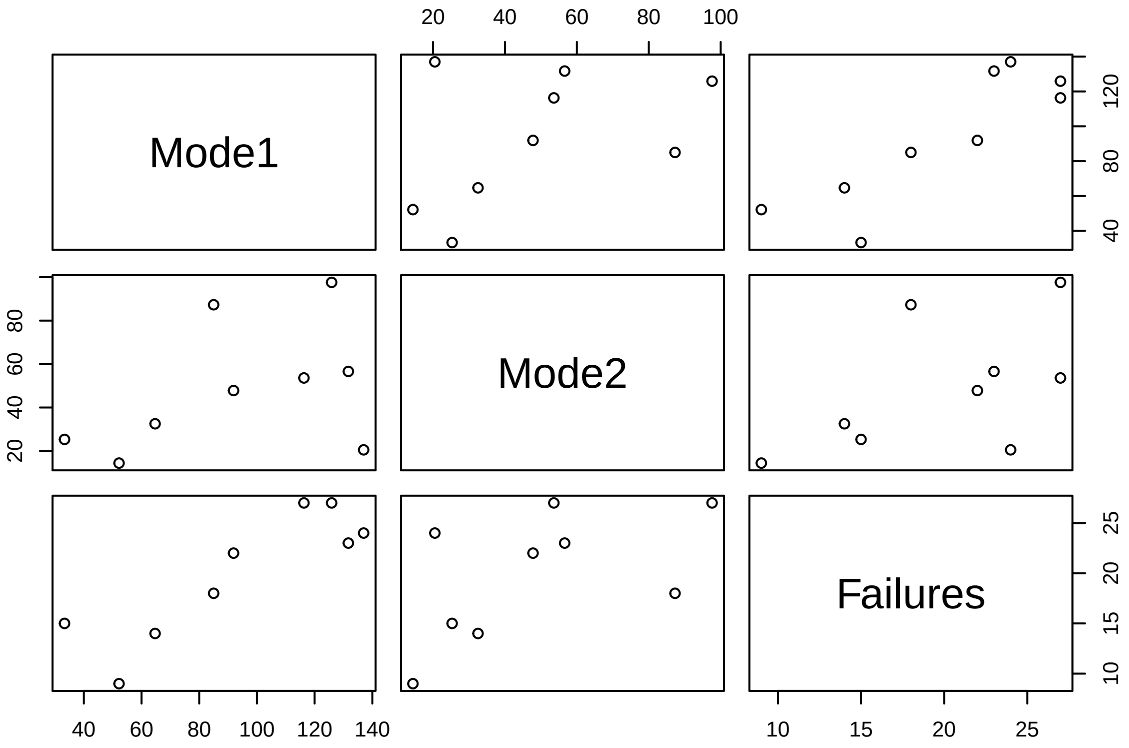
```r
head(df)
```

```
##   Mode1 Mode2 Failures
## 1  33.3  25.3       15
## 2  52.2  14.4        9
## 3  64.7  32.5       14
## 4 137.0  20.5       24
## 5 125.9  97.6       27
## 6 116.3  53.6       27
```

```r
tail(df)
```

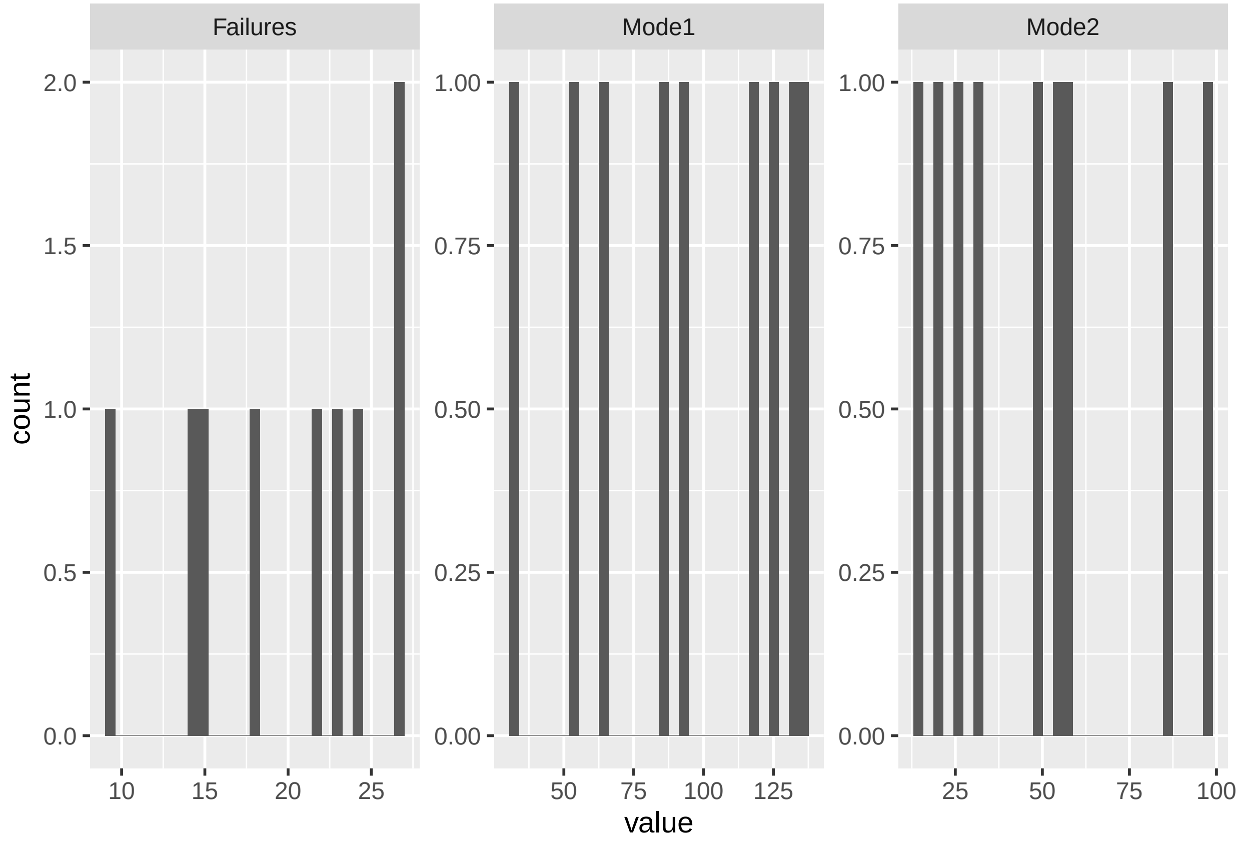
```
##   Mode1 Mode2 Failures
## 4 137.0  20.5       24
## 5 125.9  97.6       27
## 6 116.3  53.6       27
## 7 131.7  56.6       23
## 8  85.0  87.3       18
## 9  91.9  47.8       22
```

```
plot(df)
```



```
df %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



Exercise 2.a)

- Response: Failures
- Distribution: Poisson
- Explanatory variables: mode1 & mode2
- Link function: ~~log~~ -> rather identity, because one rather wants a direct influence of the operating time on the failure rate in each mode. This choice is supported by the fact that both operating times are positive explanatory variables, and thus, with positive parameter values, the linear predictor is also positive. Therefore, the link "identity" guarantees a positive failure rate.– But the log link is not excluded by these arguments!

Question 2.a)

What is a good suggestion of the procedure to finde the right model parameters like distribution and especially link function?

Exercise 2.b)

Fit the suggested model in a):

```
glm2.1 <- glm(Failures ~ ., family = poisson(link = 'identity'), data = df)
summary(glm2.1)
```

```
##
## Call:
## glm(formula = Failures ~ ., family = poisson(link = "identity"),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.19870  -0.40947   0.06809   0.50632   1.01581
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.99773    3.63545   1.650  0.09899 .
## Mode1        0.12081    0.04578   2.639  0.00832 **
## Mode2        0.05459    0.06356   0.859  0.39037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16.9964  on 8  degrees of freedom
## Residual deviance:  4.1971  on 6  degrees of freedom
## AIC: 53.254
##
## Number of Fisher Scoring iterations: 6
```

Since the coefficients have positive signs and therefore are positive linear predictors the signs are correct.

Exercise 2.c)

Another model that can be considered, as stated in the worksheet, uses neither an intercept nor the explanatory variable *mode2*; that is, *Failures ~ -1 + mode1*

What are the pros and cons of this reduced model?

- Pros
  - in practical application, it has been repeatedly shown that the intercept collects systematic errors in both the response and the explanatory variables, which would be avoided this way
- Cons
  - The intercept must be interpreted somehow, but is not included in this model

Fitting the suggested model and comparing it to the original one fitted in b):

```
glm2.2 <- glm(Failures ~ -1 + Mode1, family = poisson(link = 'identity'), data = df)
summary(glm2.2)
```

```
##
## Call:
## glm(formula = Failures ~ -1 + Mode1, family = poisson(link = "identity"),
##     data = df)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -1.00464  -0.66647   0.02067   0.42689   2.57095
##
## Coefficients:
##       Estimate Std. Error z value Pr(>|z|)
## Mode1  0.21360    0.01597   13.38   <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:    Inf  on 9  degrees of freedom
## Residual deviance: 9.5237  on 8  degrees of freedom
## AIC: 54.58
##
## Number of Fisher Scoring iterations: 3
```

Question 2.c)

Is this explanation right, why the Null deviance is inf? The null deviance is Inf (infinite) because it describes the residuals with only the intercept and because there is no intercept in this model, the model has no residuals there.

```
anova(glm2.1, glm2.2, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: Failures ~ Mode1 + Mode2
## Model 2: Failures ~ -1 + Mode1
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         6     4.1971
## 2         8     9.5237 -2  -5.3265  0.06972 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value of the newly suggested (reduced) model is 0.06972 and therefore > than the significance level of 5% we can not reject the null Hypothesis and conclude, that both models describe the model statistically equally well.

Exercise 3

```


path <- file.path('Datasets', 'nambeware.txt')
df <- read.table(path, header=TRUE)

summary(df)
```

Dataset loading and sanity check:

```
##      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  
```

```
str(df)
```

```
## ’data.frame’:    59 obs. of  4 variables:
##  $ Type : chr  "CassDish" "CassDish" "CassDish" "Bowl" ...
##  $ Diam : num  10.7 14 9 8 10 10.5 16 15 6.5 5 ...
##  $ Time : num  47.6 63.1 58.8 34.9 55.5 ...
##  $ Price: num  144 215 105 69 134 129 155 99 38.5 36.5 ...
```

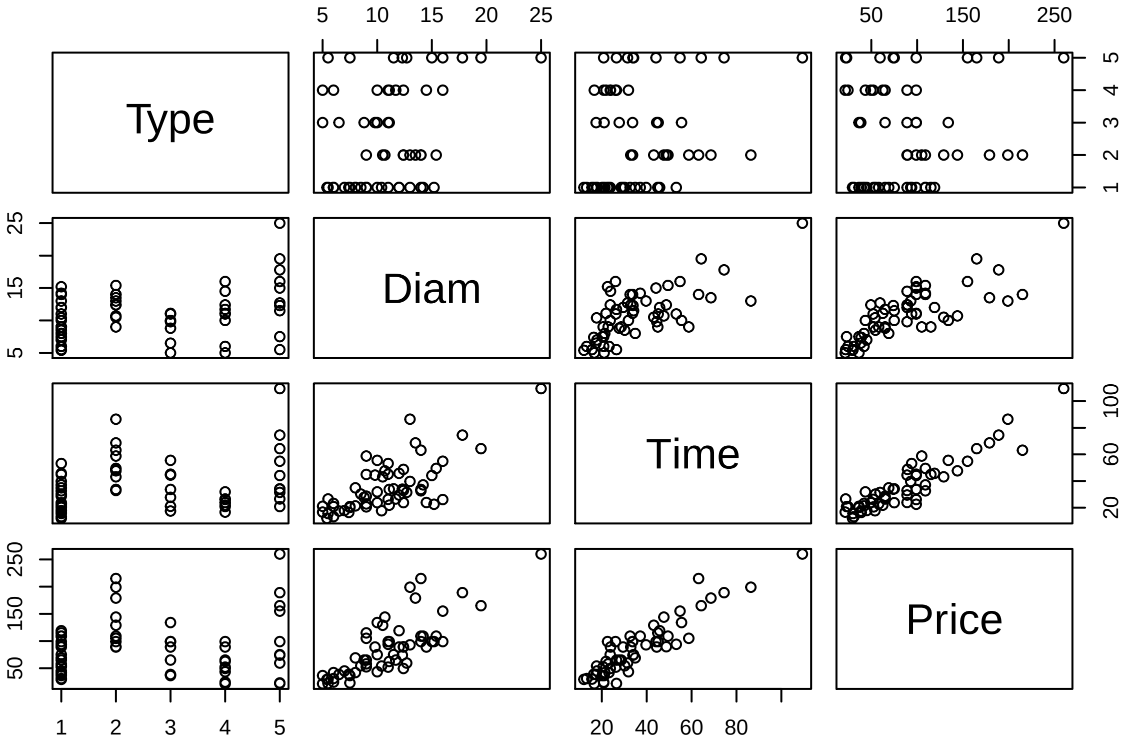
```
head(df)
```

```
##       Type Diam  Time Price
## 1 CassDish 10.7 47.65   144
## 2 CassDish 14.0 63.13   215
## 3 CassDish  9.0 58.76   105
## 4     Bowl  8.0 34.88    69
## 5     Dish 10.0 55.53   134
## 6 CassDish 10.5 43.14   129
```

```
tail(df)
```

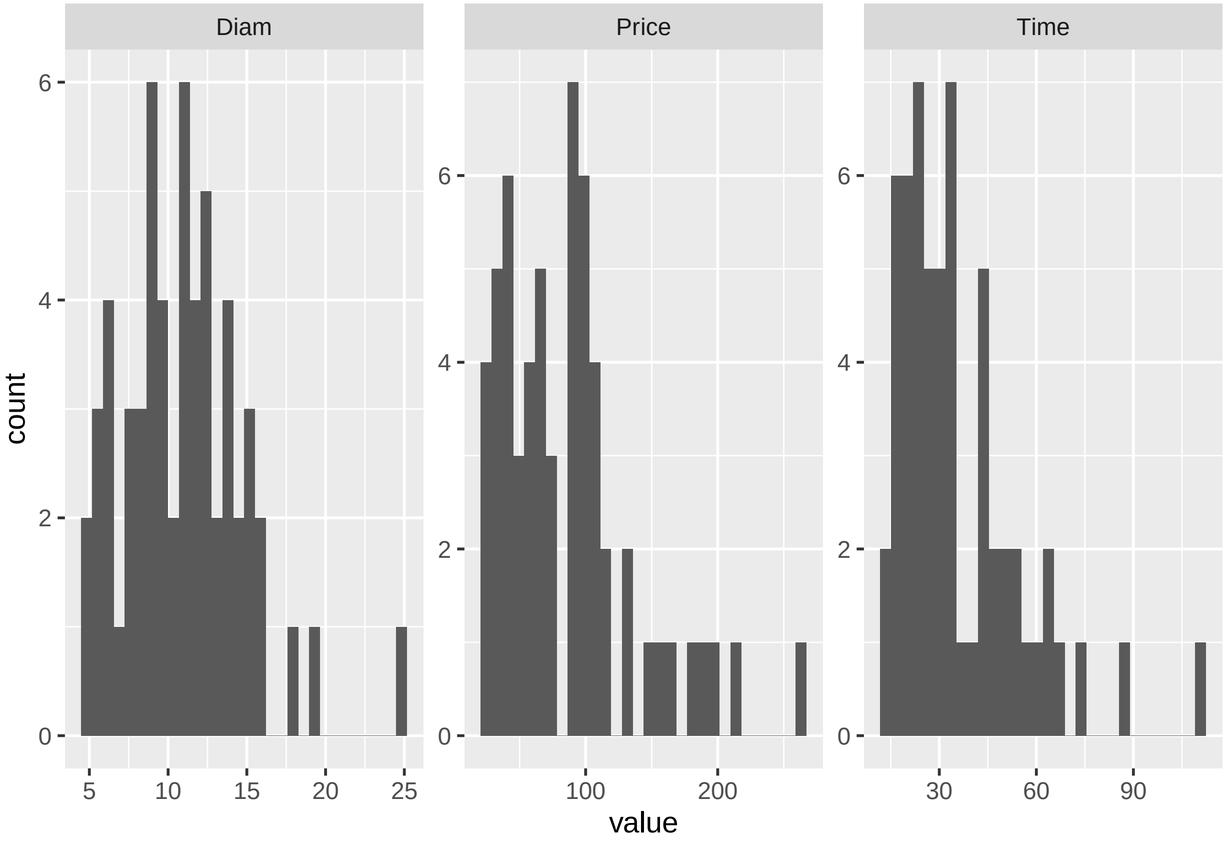
```
##     Type Diam  Time Price
## 54  Bowl  8.5 30.20  54.5
## 55 Plate  6.0 20.85  24.5
## 56 Plate 11.0 26.25  52.0
## 57 Plate 11.1 21.87  62.5
## 58 Plate 14.5 23.88  89.0
## 59 Plate  5.0 16.66  21.5
```

```
plot(df)
```



```
df %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



Exercise 3.a)

Testing the if the model using the linear predictor 'Diam * Type' describe the data of Nambeware better than the model with the linear predictor 'Diam + Type':

```
glm3.1 <- glm(Time ~ Diam * Type, family = Gamma(link = log), data = df)
glm3.2 <- glm(Time ~ Diam + Type, family = Gamma(link = log), data = df)

anova(glm3.1, glm3.2, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: Time ~ Diam * Type
## Model 2: Time ~ Diam + Type
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        49     3.9210
## 2        53     4.5039 -4 -0.58292   0.1442
```

Since the p-value of the second model is > than the significance level of 5% we can not reject the null Hypothesis and conclude that both models describe the data equally well and use therefore the reduced model (glm3.2).

Exercise 4

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

summary(df)
```

Dataset loading and sanity check:

```
##      Fails              m          Pres            Temp
##  Min.   :0.0000   Min.   :6   Min.   : 50.0   Min.   :53.00
##  1st Qu.:0.0000   1st Qu.:6   1st Qu.: 50.0   1st Qu.:67.00
##  Median :0.0000   Median :6   Median :200.0   Median :70.00
##  Mean   :0.3913   Mean   :6   Mean   :145.7   Mean   :69.57
##  3rd Qu.:1.0000   3rd Qu.:6   3rd Qu.:200.0   3rd Qu.:75.00
##  Max.   :2.0000   Max.   :6   Max.   :200.0   Max.   :81.00
```

```
str(df)
```

```
## ’data.frame’:    23 obs. of  4 variables:
##  $ Fails: int  0 1 0 0 0 0 0 0 1 1 ...
##  $ m    : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ Pres : int  50 50 50 50 50 50 50 100 100 200 ...
##  $ Temp : int  66 70 69 68 67 72 73 70 57 63 ...
```

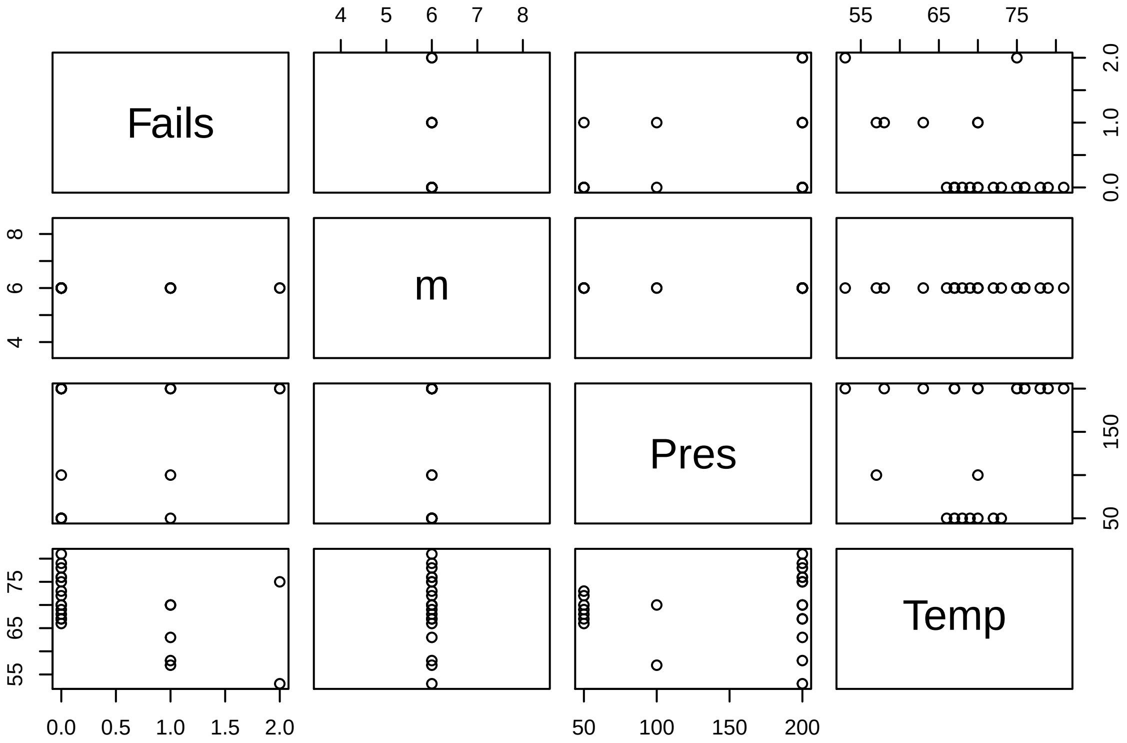
```
head(df)
```

```
##   Fails m Pres Temp
## 1     0 6   50   66
## 2     1 6   50   70
## 3     0 6   50   69
## 4     0 6   50   68
## 5     0 6   50   67
## 6     0 6   50   72
```

```
tail(df)
```

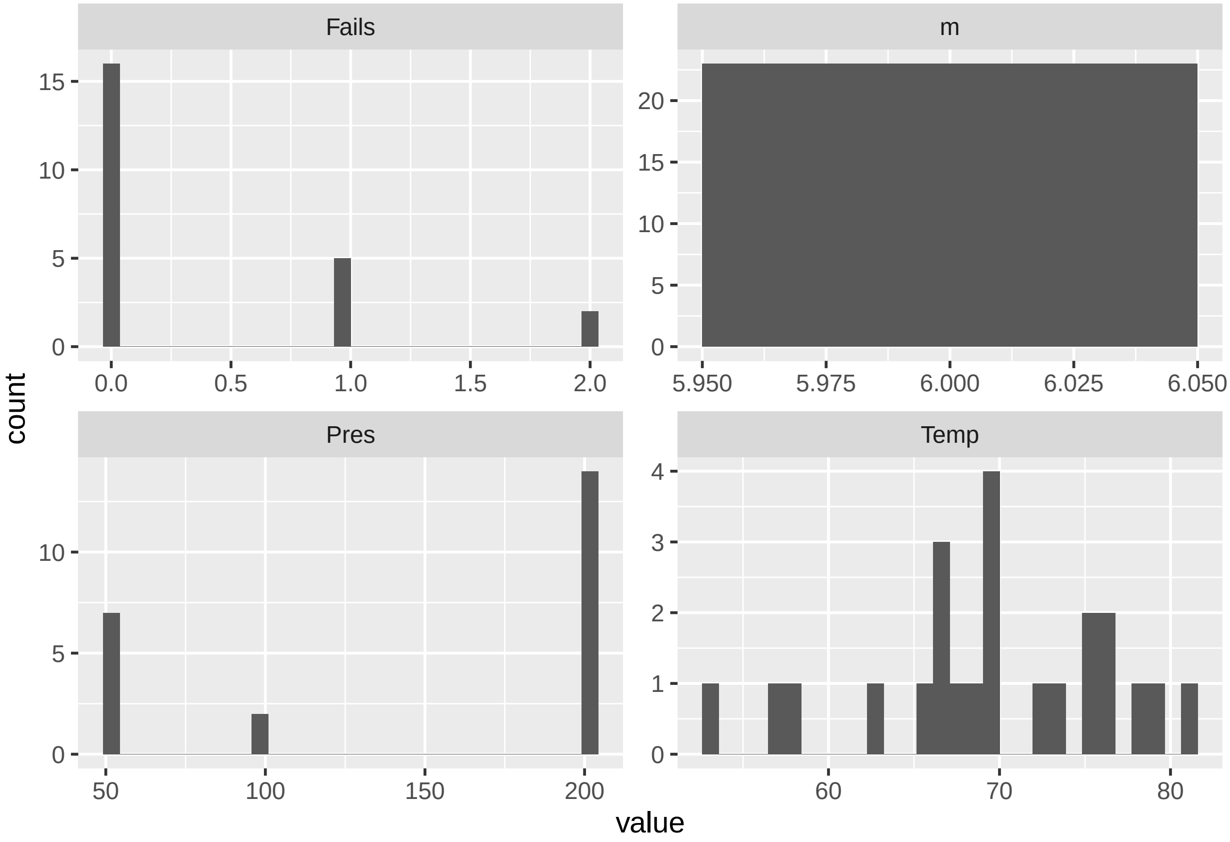
```
##    Fails m Pres Temp
## 18     0 6  200   81
## 19     0 6  200   76
## 20     0 6  200   79
## 21     2 6  200   75
## 22     0 6  200   76
## 23     1 6  200   58
```

```
plot(df)
```

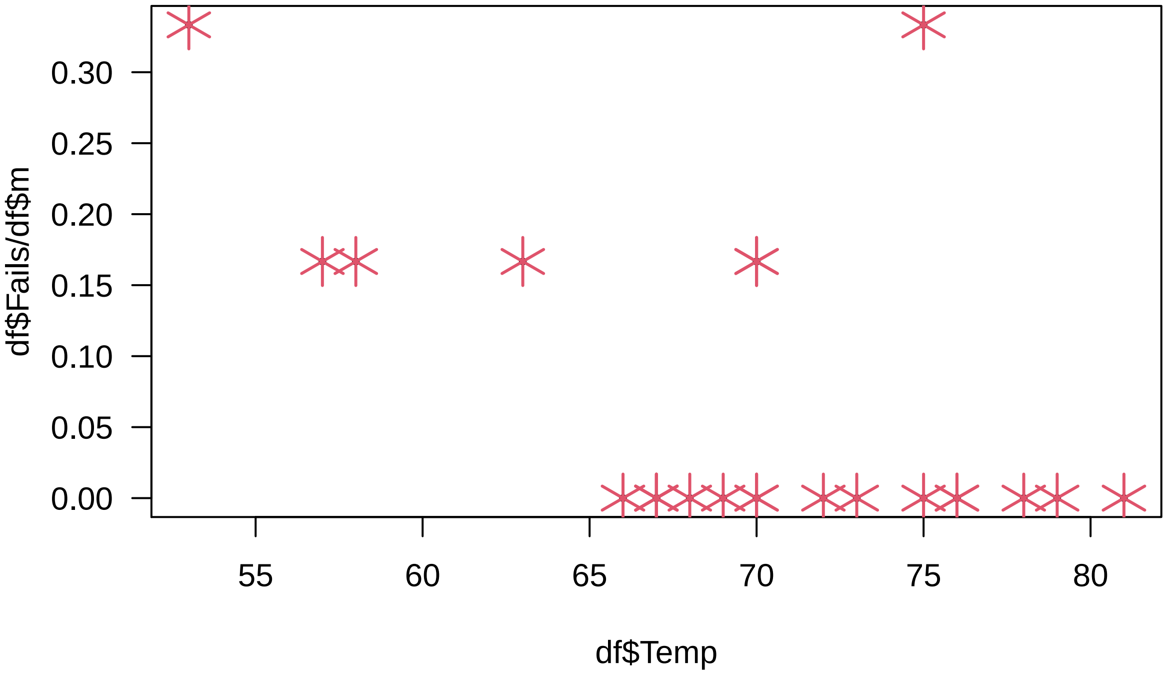


```
df %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
par(mfrow=c(1,1))
sunflowerplot(df$Temp, df$Fails/df$m, number=df$m, las=1)
```



Exercise 4.a)

- Response: Fails
- Distribution: binomial with expectation mu=pi_i and size=m_i
- Explanatory Variables: m, Temp, Pres
- Link function: canonical link because there is none mentioned explicitly. But alternative: + complementary log-log link because the topic is material fatigue
- Model:
  - glm(cbind(Failures, m-Failures) ~ Temp + Pres, family = binomial(link = logit), data = df)

Question 4.a)

How do we knoe that the link function is logit? With just stated to use the canonical link, as in the solutions, to me is not clear why logit is the reasonable choice.

Exercise 4.b)

Fit the model proposed in a)

```
glm4.1 <- glm(cbind(Fails, m-Fails) ~ Temp + Pres, family = binomial(link = logit),
              data = df)
summary(glm4.1)
```

```
##
## Call:
## glm(formula = cbind(Fails, m - Fails) ~ Temp + Pres, family = binomial(link = logit),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.05383  -0.65352  -0.56140  -0.03971   2.37171
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.409728   3.178539   1.073   0.2834
## Temp        -0.107747   0.044648  -2.413   0.0158 *
## Pres         0.007380   0.006447   1.145   0.2523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.565  on 20  degrees of freedom
## AIC: 36.125
##
## Number of Fisher Scoring iterations: 5
```

The relevance of the pressure at which safety testing for field join leaks was performed to the failure process was unclear. The p-value of Pres is > than the significance level of 5% and we conclude that it is not significant to describe the data (Wald statistics). But lets compare another model without the pressure to the first model.

```
glm4.2 <- glm(cbind(Fails, m-Fails) ~ Temp, family = binomial(link = logit), data = df)
anova(glm4.1, glm4.2, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Fails, m - Fails) ~ Temp + Pres
## Model 2: cbind(Fails, m - Fails) ~ Temp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        20     16.565
## 2        21     18.086 -1  -1.5212   0.2174
```

Since the p-value of 0.2174 is > than the significance level of 5% we can not reject the null hypothesis and conclude that both models describe the data statistically equally well and can use in practice the reduced one (based on the preferred deviance statistics here). However, we do not know what risk of using the reduced model is (i.e. the probability of type II error)

Testing using the confidence intervals Wald:

```
(h <- summary(glm4.1)$coefficients)
```

```
##                 Estimate  Std. Error   z value   Pr(>|z|)
## (Intercept)  3.409727546 3.178539015  1.072734 0.28339039
## Temp        -0.107747167 0.044648409 -2.413236 0.01581156
## Pres         0.007379589 0.006446755  1.144698 0.25233419
```

```
h[3,1] + c(-1,1)*qnorm(0.975) * h[3,2]  ##  -0.0053  0.0200
```

```
## [1] -0.005255819  0.020014998
```

The 95% confidence interval covers the null hypothesis ‘beta1 = 0’ (0 is between -0.0053 and 0.02). Hence we have no evidence against the null hypothesis.

Deviance (via profiling):

```
confint(glm4.1)
```

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

##                    2.5 %      97.5 %
## (Intercept) -2.776236540  9.93358512
## Temp        -0.201164111 -0.02229717
## Pres        -0.004030283  0.02272544
```

This 95% confidence interval ([-0.004030283, 0.02272544]) covers the null hypothesis ‘beta1 = 0’ as well. So we obtain the same conclusion.

Exercise 4.c)

Predict the probability that an O-ring will leak at the expected tempreature of 31°F at launch.

```
preds1 <- predict(glm4.2, type = 'response',
                   se.fit = TRUE, newdata = data.frame(Temp = 31))
preds1$fit
```

```
##         1
## 0.8177744
```

```
preds1$fit + c(-1, 1) * qnorm(0.975) * preds1$se.fit
```

```
## [1] 0.346496 1.289053
```

This leads to a very large confidence interval and lies not within the support of [0,1]. Therefore we try another approach:

```
preds2 <- predict(glm4.2, type="link",
                  se.fit = TRUE, newdata=data.frame(Temp=31))
(preds2adj <- preds2$fit + c(-1,1) * qnorm(0.975) * preds2$se.fit)
```

```
## [1] -1.661189  4.663871
```

```
1/(1 + exp(-preds2adj))
```

```
## [1] 0.1596025 0.9906582
```

The 95% confidence interval covers almost the whole support except the area to 0. But the probability that an o-ring will fail may be close to 1! There is not much confidence that the o-ring will sustain.

Exercise 4.d)

Repeating the above analysis with just those observatinos in which at least one failure occured:

```
df2 <- df[(df$Fails != 0), ]
str(df2)
```

```
## 'data.frame':    7 obs. of  4 variables:
##  $ Fails: int  1 1 1 1 2 2 1
##  $ m    : int  6 6 6 6 6 6 6
##  $ Pres : int  50 100 200 200 200 200 200
##  $ Temp : int  70 57 63 70 53 75 58
```

```
summary(df2)
```

```
##      Fails             m          Pres            Temp      
##  Min.   :1.000   Min.   :6   Min.   : 50.0   Min.   :53.00  
##  1st Qu.:1.000   1st Qu.:6   1st Qu.:150.0   1st Qu.:57.50  
##  Median :1.000   Median :6   Median :200.0   Median :63.00  
##  Mean   :1.286   Mean   :6   Mean   :164.3   Mean   :63.71  
##  3rd Qu.:1.500   3rd Qu.:6   3rd Qu.:200.0   3rd Qu.:70.00  
##  Max.   :2.000   Max.   :6   Max.   :200.0   Max.   :75.00
```

```
nrow(df) - nrow(df2)
```

```
## [1] 16
```

```
nrow(df2)
```

```
## [1] 7
```

```
glm4.3 <- glm(cbind(Fails, m-Fails) ~ Pres + Temp,
              family = binomial(link = logit), data = df2)
glm4.4 <- glm(cbind(Fails, m-Fails) ~ Temp,
              family = binomial(link = logit), data = df2)
glm4.5 <- glm(cbind(Fails, m-Fails) ~ 1,
              family = binomial(link = logit), data = df2)

anova(glm4.3, glm4.4, test = 'Chisq')
```

```
## Analysis of Deviance Table
## 
## Model 1: cbind(Fails, m - Fails) ~ Pres + Temp
## Model 2: cbind(Fails, m - Fails) ~ Temp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         4     1.1070                     
## 2         5     1.3339 -1 -0.22684   0.6339
```

The two model describe the data statistically equally well, so we use the reduced one.

```
anova(glm4.4, glm4.5, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Fails, m - Fails) ~ Temp
## Model 2: cbind(Fails, m - Fails) ~ 1
##   Resid. Df Resid. Dev Df    Deviance Pr(>Chi)
## 1         5     1.3339
## 2         6     1.3347 -1 -0.00080913   0.9773
```

Also this models describe the data statistically equally well, so we would again use the reduced one (with just the intercept…). Justify this by a variable selection with step():

```
step(glm4.3)
```

```
## Start:  AIC=20.67
## cbind(Fails, m - Fails) ~ Pres + Temp
##
##        Df Deviance    AIC
## - Temp  1   1.1111 18.671
## - Pres  1   1.3339 18.894
## <none>      1.1070 20.667
##
## Step:  AIC=18.67
## cbind(Fails, m - Fails) ~ Pres
##
##        Df Deviance    AIC
## - Pres  1   1.3347 16.895
## <none>      1.1111 18.671
##
## Step:  AIC=16.89
## cbind(Fails, m - Fails) ~ 1
```

```
##
## Call:  glm(formula = cbind(Fails, m - Fails) ~ 1, family = binomial(link = logit),
##     data = df2)
##
## Coefficients:
## (Intercept)
##      -1.299
##
## Degrees of Freedom: 6 Total (i.e. Null);  6 Residual
## Null Deviance:       1.335
## Residual Deviance: 1.335     AIC: 16.89
```

The best model is again the one with just the intercept. So we could start at any temperature.

95% confidence interval for the probability of a defect o-ring:

```
preds3 <- predict(glm4.5, newdata=data.frame(Temp=31), type="link", se=T)
family(glm4.5)$linkinv(preds3$fit + c(-1,1) * qnorm(0.975) * preds3$se.fit)
```

```
## [1] 0.1154411 0.3630300
```

To compare with CI based on glm4.2

```
preds4 <- predict(glm4.2, newdata=data.frame(Temp=31), type="link", se=T)
family(glm4.2)$linkinv(preds4$fit + c(-1,1) * qnorm(0.975) * preds4$se.fit)
```

```
## [1] 0.1596025 0.9906582
```

this one is much wider than the one before.

(from solutions)
Based on this “reduced” dataset, one could easily be convinced that temperature does not affect O-ring performance. Hence, based on this “reduced” dataset the conclusion which the scientists and engineers drew was correct.

But, when you conduct a statistical analysis on a sample of the available data, you can induce what in statistics is known as a sample selection problem. Running an analysis on less than the entire data set is not always a problem, but it can lead to mistaken conclusions depending on the question you are trying to answer.

Lessons learned: 1. Be very, very careful when predicting “out-of-sample” support. 2. Don’t sample when all data points are available … all launches, not just ones with O-ring distress.

Exercise 4.e) (copied from solutions)

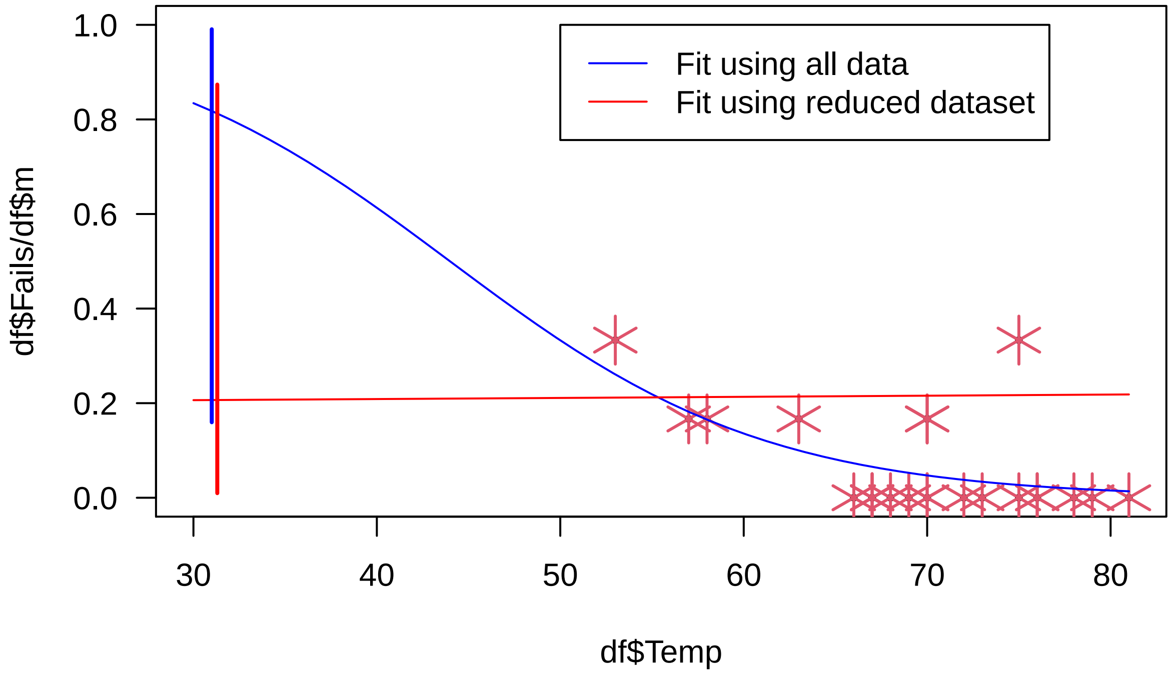
Display the data properly assuming just an effect of temperature on the response and overlay the corresponding fit using all data or the reduced dataset. In addtion, overlay both 95% confidence intervals at a temperature of 31 ∘ F. What do you conclude from this?

```
h.xlim <- c(30, max(df$Temp))
new.df <- data.frame(Temp=seq(h.xlim[1], h.xlim[2], length=50))
h.predGLM2 <- predict(glm4.2, newdata=new.df, type="response")
h.pred1GLM2 <- predict(glm4.4, newdata=new.df, type="response")

sunflowerplot(df$Temp, df$Fails/df$m, number=df$m, las=1,
              xlim=h.xlim, ylim=c(0,1))
## fit
lines(new.df$Temp, h.predGLM2, col="blue")
lines(new.df$Temp, h.pred1GLM2, col="red")
legend(x=50, y=1, legend=c("Fit using all data", "Fit using reduced dataset"),
       col=c("blue", "red"), lty=c(1,1))

## confidence intervals
h2 <- predict(glm4.2, newdata=data.frame(Temp=31), type="link", se=T)
h2.ci <- family(glm4.2)$linkinv(h2$fit + c(-1,1)*qnorm(0.975)*h2$se.fit)
h12 <- predict(glm4.4, newdata=data.frame(Temp=31), type="link", se=T)
h12.ci <- family(glm4.4)$linkinv(h12$fit + c(-1,1)*qnorm(0.975)*h12$se.fit)

lines(c(31,31), h2.ci, col="blue", lwd=2)
lines(c(31.3,31.3), h12.ci, col="red", lwd=2)
```



Both confidence intervals are huge indicating that there is a great uncertainty in the predicted probabilities independent of the applied fit.