AdvStDaAn, Worksheet, Week 6

Michael Lappert

24 April, 2022

Contents

Exercise 1

```
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 ...
```

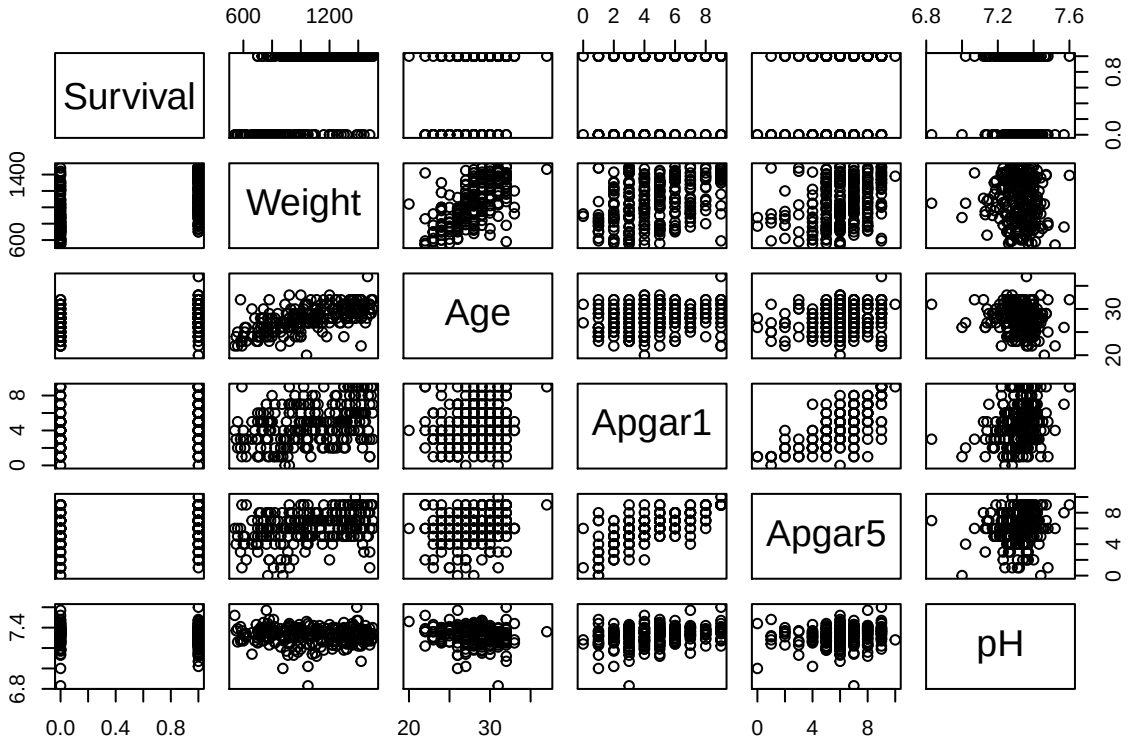
```
head(df)
```

```
##   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)
```

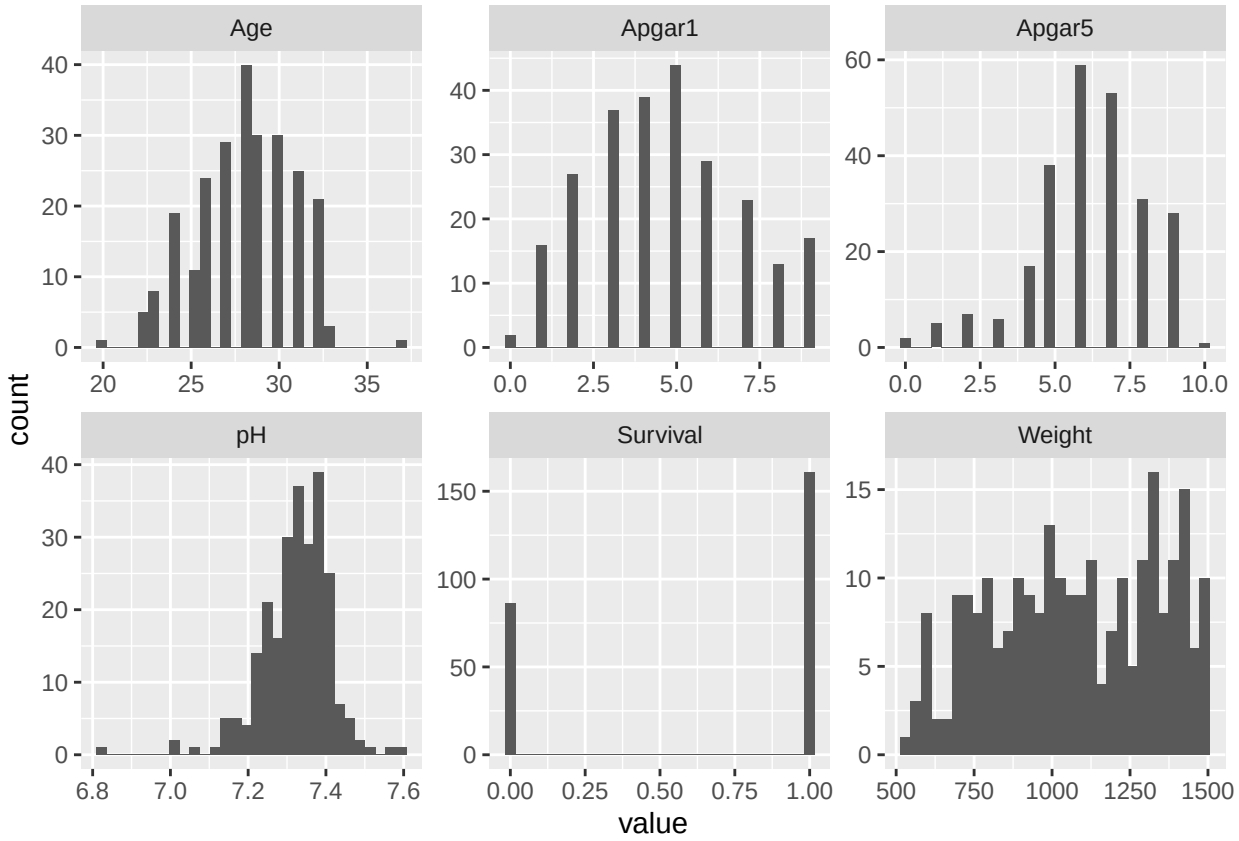
```
##     Survival Weight Age Apgar1 Apgar5   pH
## 242        1   1120  28      7      7 7.33
## 243        1   1020  28      5      7 7.34
## 244        1   1320  28      6      6 7.24
## 245        0    900  27      5      6 7.37
## 246        1   1150  27      4      7 7.37
## 247        0    790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

```
plot(df)
```




```
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.

```
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
```

```
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

```
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
```

```
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
```

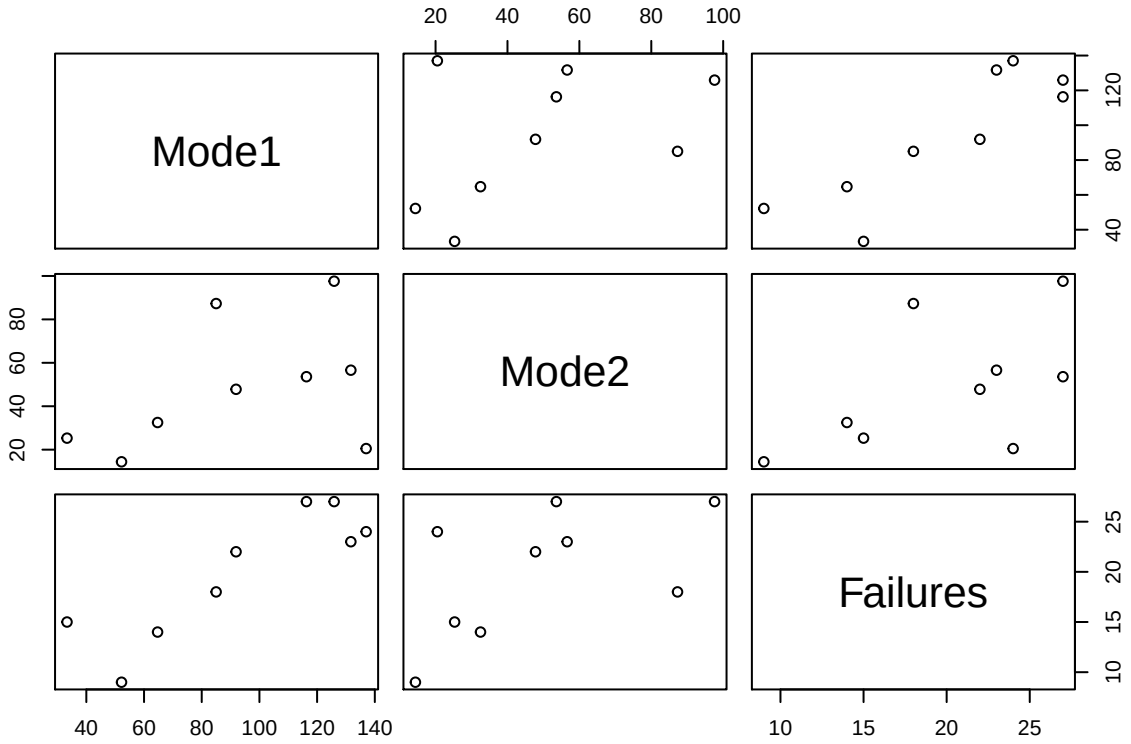
```
head(df)
```

```
##   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
```

```
tail(df)
```

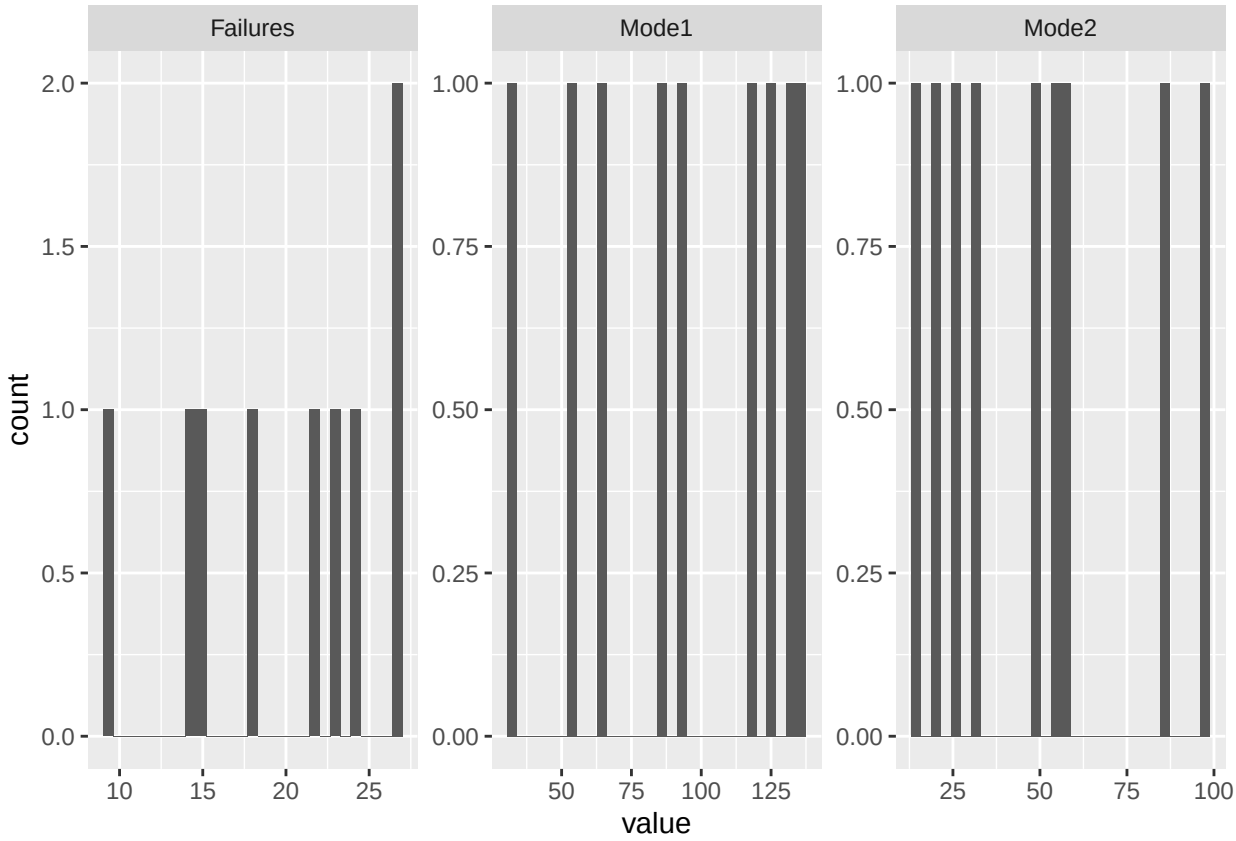
```
##   Mode1 Mode2 Failures
## 4 137.0  20.5       24
## 5 125.9  97.6       27
## 6 116.3  53.6       27
## 7 131.7  56.6       23
## 8  85.0  87.3       18
## 9  91.9  47.8       22
```

```
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 ...
```

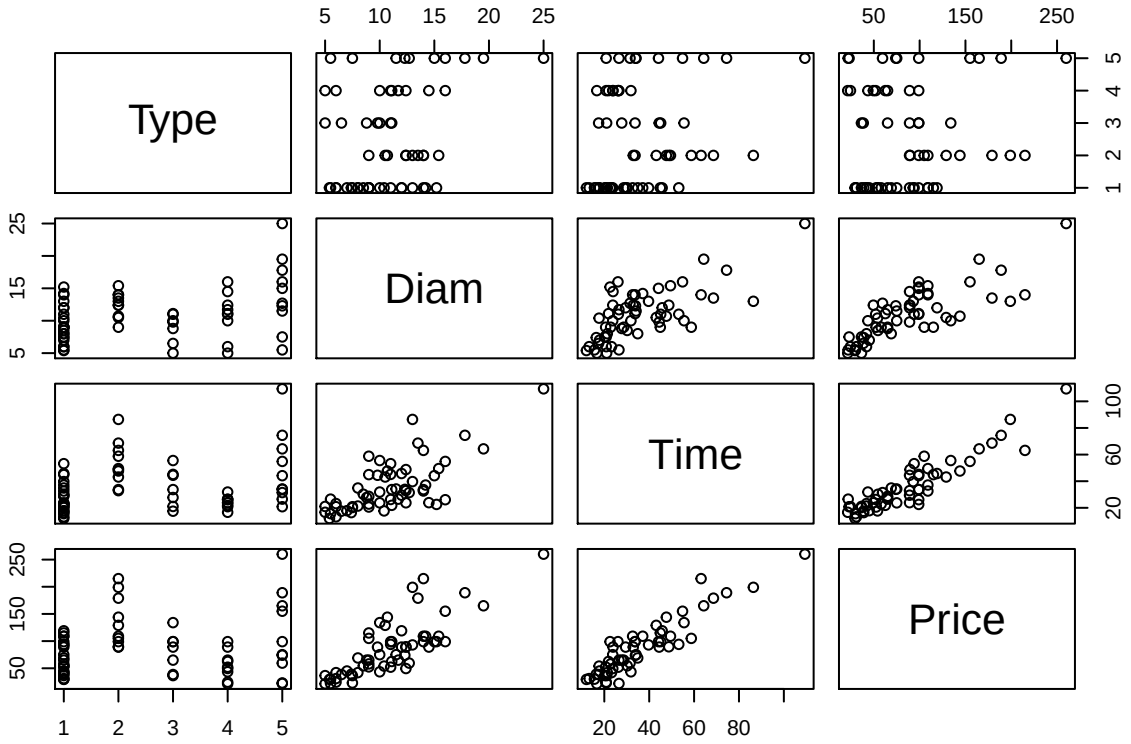
```
head(df)
```

```
##       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)
```

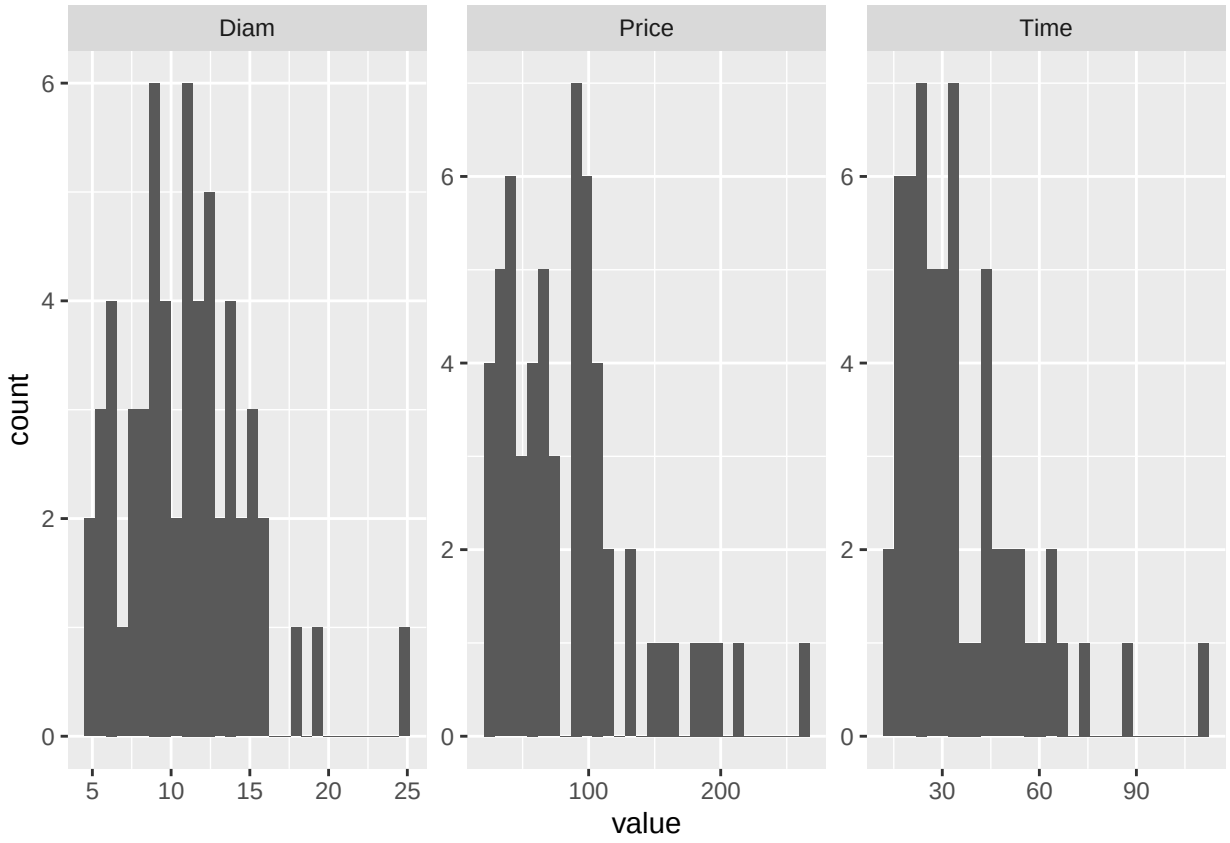
```
##     Type Diam  Time Price
## 54  Bowl  8.5 30.20  54.5
## 55 Plate  6.0 20.85  24.5
## 56 Plate 11.0 26.25  52.0
## 57 Plate 11.1 21.87  62.5
## 58 Plate 14.5 23.88  89.0
## 59 Plate  5.0 16.66  21.5
```

```
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).