Reporting a Simple Analysis: a Primer

R. Labouriau

Spring 2020

```
load("DataFramesStatisticalModelling.RData")
D <- Ex4.2
```

Introduction

Here we will report the exercise 4.2. In this exercise, we will analyse some data on the number of a certain weed found in coffee (open sky) fields when different dosages of nitrogen fertiliser are applied. The number of weed plants per square metre was determined at different (disjoint and relatively far away) places in the space between coffee plants in a standard coffee production field. Five different doses were applied, and for each dose, 20 different places were observed (the 20 repetitions).

The questions of interest are: 1) whether the fertilisation level affects the number of weeds (per square metre) and 2) whether the data is compatible with a linear relationship between the number of weeds and the dose.

The data related to this exercise is stored in the following data-frame

```
str(Ex4.2)
```

```
## 'data.frame':    100 obs. of  2 variables:
##  $ n.weeds: int  7 3 4 5 2 4 2 4 4 4 ...
##  $ Nitr   : int  0 0 0 0 0 0 0 0 0 0 ...
```

Some Preliminairy Exploratory Analyses

There are 20 replicates for each of the 5 fertilisation dose, as displayed in the table below.

```
table(D$Nitr)
```

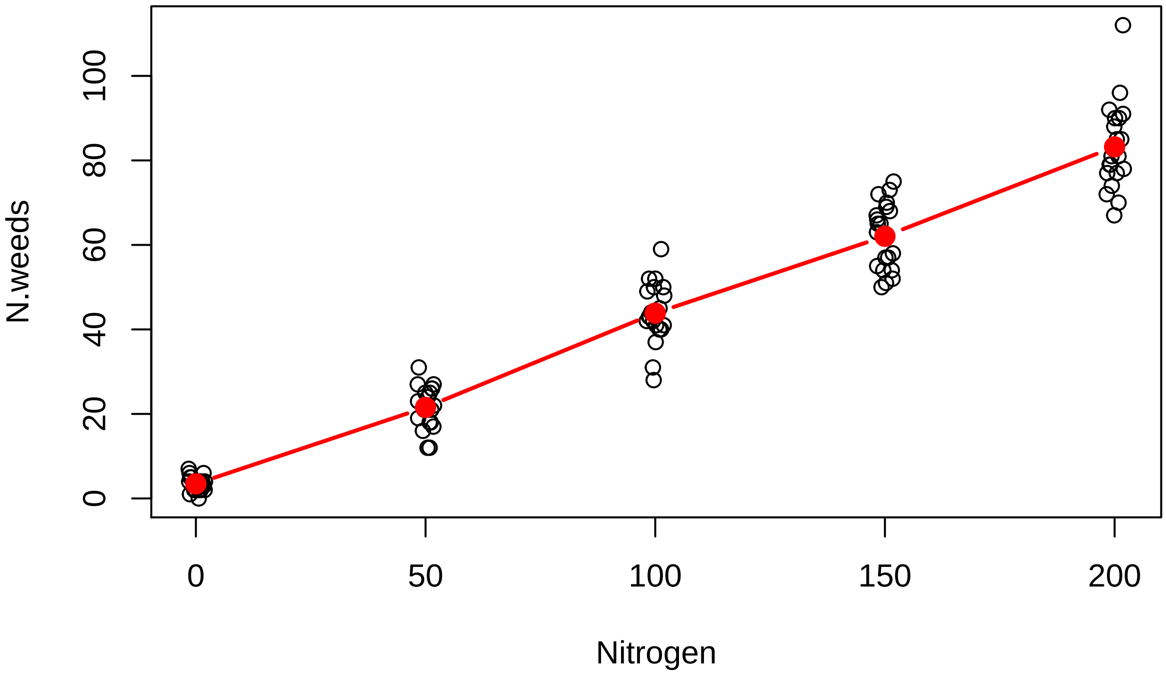
```
##
##   0  50 100 150 200
##  20  20  20  20  20
```

The mean numbers of weed plants for each level of nitrogen fertilisation are:

```
(Means <- tapply(D$n.weeds, D$Nitr, mean))
```

```
##     0    50   100   150   200
##  3.45 21.55 43.85 62.05 83.20
```

The numbers of weed plants are plotted for each observation (vertical axis) against the doses (horizontal axis)



below.
Here we represented the means of the number of plants for each fertilisation level by red dots and connected the means for visualisation. We added some noise to the fertilisation levels for graphically separating superposed points. Note that, apparently, the mean number of weeds increases linearly with the applied dose of N.

As it is apparent from the figure above, the dispersion tends to increase when the means increase. This is also apparent in the calculations and the plot below.

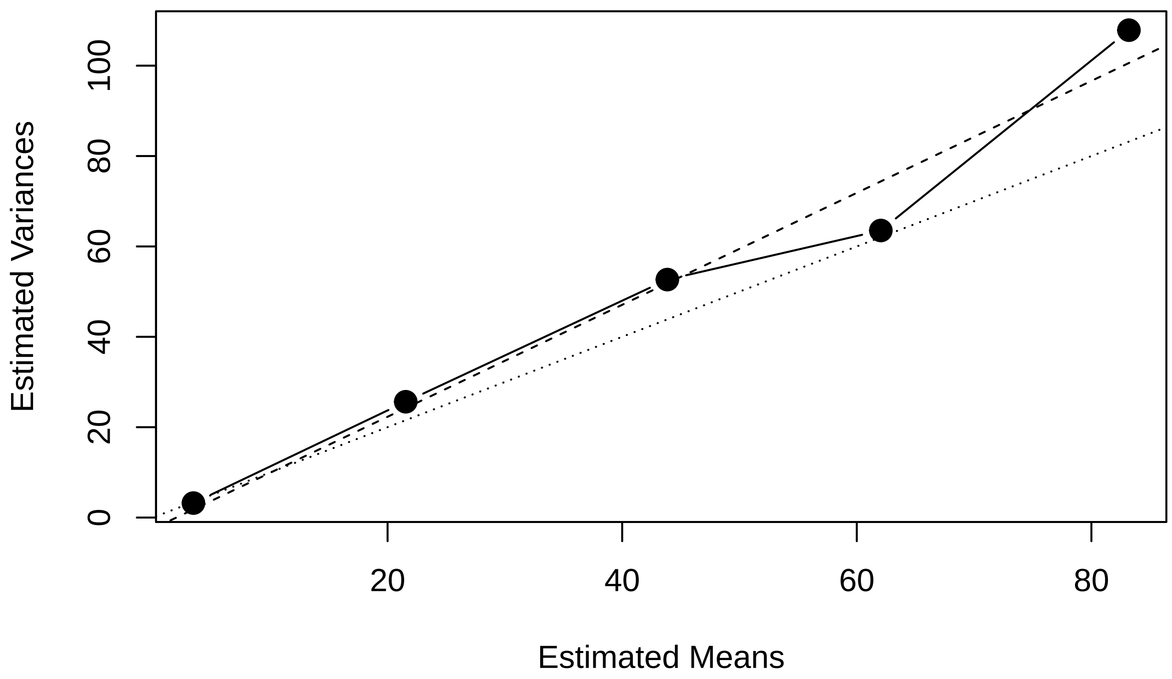
```
(Vars <- tapply(D$n.weeds, D$Nitr, var))
```

```
##          0         50        100        150        200 
##   3.207895  25.628947  52.660526  63.523684 107.852632
```

```
plot(Means, Vars, pch = 19, type = "b", xlab = "Estimated Means",
                                        ylab = "Estimated Variances", ce = 1.5)
```

```
## Warning in plot.window(...): "ce" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ce" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ce" is not a
## graphical parameter

## Warning in box(...): "ce" is not a graphical parameter

## Warning in title(...): "ce" is not a graphical parameter
```

```
abline(lm(Vars ~ Means), lty = 2)
abline(0,1, lty = 3)
```



This is compatible with the assumption that the observed counts are Poisson distributed, since the expectation and the variance of a Poisson distributed random variable are equal.

Fitting a Basic Model

Here we adjust a basic model that assumes that the counts of weeds are independent and Poisson distributed with all the observations that received the same amount of N fertilisation having the same expectation. The formal definition of this model is given below.

Denote by $Y_{n,r}$ the random variable representing the counts of weeds rth repetition of the parcels the received n Kg of N (per ha), where $n = 0, 50, 100, 150, 200$ and $r = 1, ..., 20$. According to the model $Y_{0,1} ..., Y_{100,50}$ are independent, Poisson distributed and have expectations given by

$$E(Y_{n,r}) = \lambda_n, \text{ for } n = 0, 50, 100, 150, 200 \text{ and } r = 1, ..., 20.$$

In short,for \$n = 0, 50, 100, 150, 200, and r = 1, . . . , 20,

$$Y_{n,r} \sim Po(\lambda_n) \, .$$

The model described above will be termed the "free curve model" and can be adjusted in R in the following way.

```
M <- glm(n.weeds ~ factor(Nitr) + 0 ,
         family = poisson(link = "identity"),
         data = D)
```

The parameters of this model (i.e., the expectations related to each N dose) are displayed below (with the respective confidence intervals with coverage of 0.95).

```
cbind(coef(M), confint(M))
```

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

##                            2.5 %    97.5 %
## factor(Nitr)0    3.45  2.698814  4.329355
## factor(Nitr)50  21.55 19.579087 23.649079
## factor(Nitr)100 43.85 41.011624 46.816596
## factor(Nitr)150 62.05 58.661560 65.566674
## factor(Nitr)200 83.20 79.266310 87.261931
```

Some Model Control

We will study two types of residuals or verifying the adequacy of the basic model described above. First we consider the raw residuals, defined by, for \$n = 0, 50, 100, 150, 200, and r = 1, . . . , 20,

$$R_{n,r} = Y_{n,r} - E(Y_{n,r}) = Y_{n,r} - \lambda_n \, ,$$

that is, the difference between the observed counts and the expected counts for each observation. Note that the expectation of those raw residuals are all 0. But the variance of the nth, rth raw residual is $Var(R_{n,r}) = Var(Y_{n,r} - \lambda_n) = Var(Y_{n,r}) = \lambda_n$. Here, the last equality comes from the fact that the expectation and the variance of a Poisson distributed random variable are equal. Note that this equality characterises the Poisson distribution.

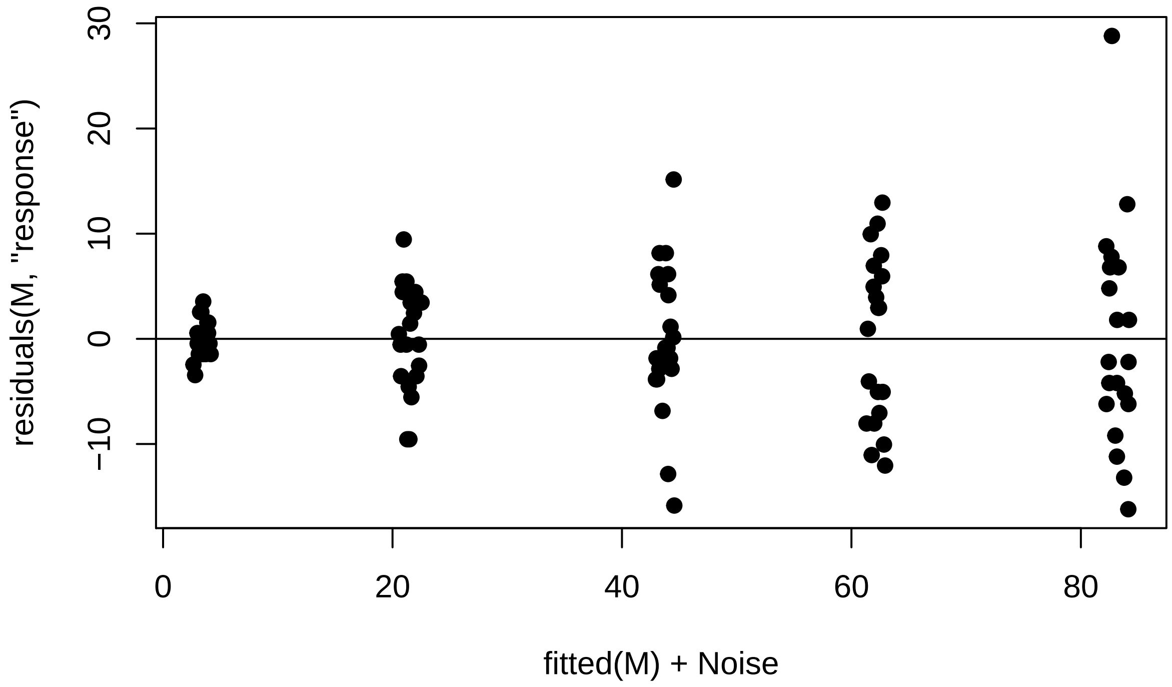
From the discussion above, we expect to see a pattern of increasing dispersion when we plot the raw-residuals against the fitted values. Recall that the fitted values are defined by, for \$n = 0, 50, 100, 150, 200, and r = 1, . . . , 20,

$$F_{n,r} = E(Y_{n,r}) = \lambda_n \, ,$$

i.e., they are the expected values of each observation (under the current model).

The plot of the raw-residuals against the fitted values indeed present the pattern we would expect for the current model, a dispersion increasing with the fitted values.

```r
Noise <- runif(n=100,min=-1, max=1)
plot(fitted(M)+Noise, residuals(M, "response"), pch = 19)
abline(h = 0)
```



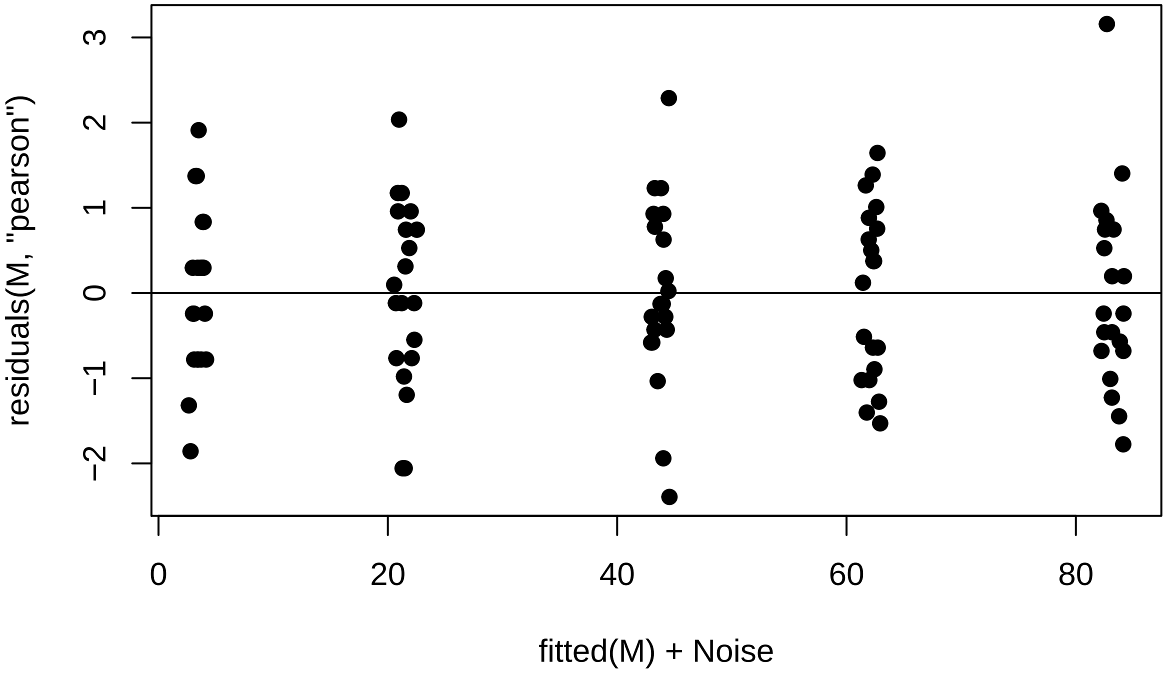
Here, we added sone noise to the fitted values for disintangling superposed points in the graph.

Next, we define the Pearson-residuals, obtained by dividing the raw-residuals by the square root of their theoretical variance. If the Poisson model used is reasonable, the thrombone-form pattern observed in the graph above should disapear. Formally, the Pearson-residuals are defined by, for $n = 0, 50, 100, 150, 200, and r = 1, . . . , 20,

$$P_{n,r} = \frac{Y_{n,r} - E(Y_{n,r})}{\sqrt{Var(Y_{n,r})}} = \frac{Y_{n,r} - \lambda_n}{\sqrt{\lambda_n}} \, .$$

The pattern observed in the plot below indicates that the model used is reasonable (compare with the plot of the raw-residuals)

```
plot(fitted(M)+Noise, residuals(M, "pearson"), pch = 19)
abline(h = 0)
```



Testing homogeneity

An additional model control is obtained by testing homogeneity. To do so, we fit a saturated model and compare the current model with the saturated model. Recal that the saturated model is a model for which there is one paramater for each observation. In the current case the saturated model assumes that, for $n = 0, 50, 100, 150, 200, and r = 1, . . . , 20,

$$Y_{n,r} \sim Po(\lambda_{n,r}) \, .$$

The deviance of the current model is the likelihood ratio statistic for reducing the saturated model to the current model. Here are the calculations related to this test:

```
deviance(M)
```

```
## [1] 111.1696
```

```
pchisq(deviance(M), df = length(D$n.weeds)-length(coef(M)), lower.tail = FALSE)
```

```
## [1] 0.1229912
```

We conclude that there is no evidence against the assumption homogeneity.

On the Effect of Fertilisation

One of the basic questions initially posed in this analysis is “whether the fertilisation level affects the number of weeds”. At this stage of the analysis we have the elements for answering this question. To do so, we adjust a model, termed the “null model”, that assumes the expected number of weed plants to be constant (independent on the amount of fertiliser added). More precisely, according to the null model, for $n = 0, 50, 100, 150, 200, and r = 1, . . . , 20,

$$Y_{n,r} \sim Po(\lambda) \,.$$

We fit the null model and perform the required test below.

```r
M0 <- glm(n.weeds ~ 1 ,
          family = poisson(link = "identity"),
          data = D)
anova(M0, M, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: n.weeds ~ 1
## Model 2: n.weeds ~ factor(Nitr) + 0
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        99    2344.67
## 2        95     111.17  4   2233.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value of the test above is very small, we conclude that there is a strong evidence that the expected values of the number of weed plants were not all the same when we used different fertilisation levels.

Fitting a Linear Regression Model

We define below a linear regression model which assumes that the expected number of weeds depends on the N dose linearly. More precisey, we assume now that, for $n = 0, 50, 100, 150, 200, and r = 1, . . . , 20,

$$Y_{n,r} \sim Po(\alpha + \beta n) \,.$$

That is, we assume that the observations are independent, poisson distributed and have expectations given by, for $n = 0, 50, 100, 150, 200, and r = 1, . . . , 20,

$$E(Y_{n,r}) = \alpha + \beta n \,.$$

The linear regression model above is adjusted in R as follows.

```r
Mlin <- glm(n.weeds ~ Nitr,
          family = poisson(link = "identity"),
          data = D)
```

Note that in this case the explanatory variable “Nitr” is not converted to a factor, as we did in the previous model. The parameters of this model (i.e., the expectations related to each N dose) are displayed below (with the respective confidence intervals with coverage of 0.95).

```r
cbind(coef(Mlin), confint(Mlin))
```

```
## Waiting for profiling to be done...
##                             2.5 %    97.5 %
## (Intercept) 3.3190504 2.6144372 4.1326154
## Nitr        0.3950095 0.3810189 0.4090701
```

Here the estimates of the parameters α and β are, 3.3190504 and 0.3950095, respectively.

Testing linearity

We test the adequacy of the linear regression model above by comparing, via the likelihood ratio test, the reduction of the free curve model to the linear regression model. This test is performed as follows.

```
anova(Mlin, M, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: n.weeds ~ Nitr
## Model 2: n.weeds ~ factor(Nitr) + 0
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        98     114.08
## 2        95     111.17  3   2.9147    0.405
```

Since the p-value for this test is large (0.405), we conclude that thres is no evidence against the adecquacy of the linear regression discussed above.

On the Effect of Fertilisation (second test)

Here we test whether the steepness coeficient in the linear above is equal to zero, which is equivalent to test whether there is an effect of fertilisation (conditional on there is a linear response to fertilisation). This test is done by comparing the linear model and the null model via the likelihood ratio test, as made below.

```
anova(M0, Mlin, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: n.weeds ~ 1
## Model 2: n.weeds ~ Nitr
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        99    2344.67
## 2        98     114.08  1   2230.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value of the test above is very small, we reject the null hypothesis that the steepness coeficient of the linear response of the number of weeds for fertilisation is zero.

Getting some final numbers and confidence intervals

The coeficients of the linear increase of the expected number of weeds are estimated as:

```
cbind(coef(Mlin), confint(Mlin))
```

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

```
##                             2.5 %    97.5 %
## (Intercept) 3.3190504 2.6144372 4.1326154
## Nitr        0.3950095 0.3810189 0.4090701
```

Concluding

We conclude that there is strong evidence that the expected number of weeds is not the same when different levels of N fertilisation are used. Moreover, our results are compatible with the hypothesis that the expected number of weeds increases (in the observed range) linearly with the dose of N used, according to the equation $C = 3.319 + 0.395N$, where N is the dose of nitrogen used and C is the expected counts of weeds.