SB1a Applied Statistics MT15 Sample Practical Report

1. A real practical example

This example is similar in style to the ones that will be given to you as assessed work. Think about how to approach this problem and the exploratory plots you want to make and the models you want to fit. Think about what you would say about these data (perhaps sketch out a report very roughly). Then have a look at the sample answer below. All the R commands are written at the end and can be found here.

`www.stats.ox.ac.uk/~nicholls/sb1a/RnotesForRatsDataAnalysis.R`

It uses just a few statistical ideas you havnt yet seen (but will shortly).

The table below presents the number of mistakes made by 24 rats in a maze. There were three strains of rats: "bright", "mixed" and "dull". Four rats from each strain were reared under "free" conditions and four under "restricted" conditions.

	bright		mixed		dull	
free	24	41	41	26	36	39
	14	16	82	86	87	99
restricted	51	96	39	104	42	92
	35	36	114	92	133	124

Are there differences in the number of mistakes made by rats from the three strains? Is the effect of the conditions under which the experiment is conducted on the number of mistakes made the same for all strains of rats?

The response in this case is actually a count so a poisson distribution is probably more appropriate than a normal distribution. Keeping in mind that in the poisson case the variance equals the mean, show how you can check if the normal approximation works well in this case and if not, find a way to improve it.

Fit appropriate normal linear models to analyse the data. Clearly state the models you have fitted and their assumptions, which you should also formally check. Describe the model selection process and interpret your final model. Include your R-code, preferably as an appendix.

To get you started, here are the data in a `data.frame()`. The `rep()` command is used to produce multiple copies of the different levels of the categorical variables `strain` and `condition` and produce the columns of the data frame.

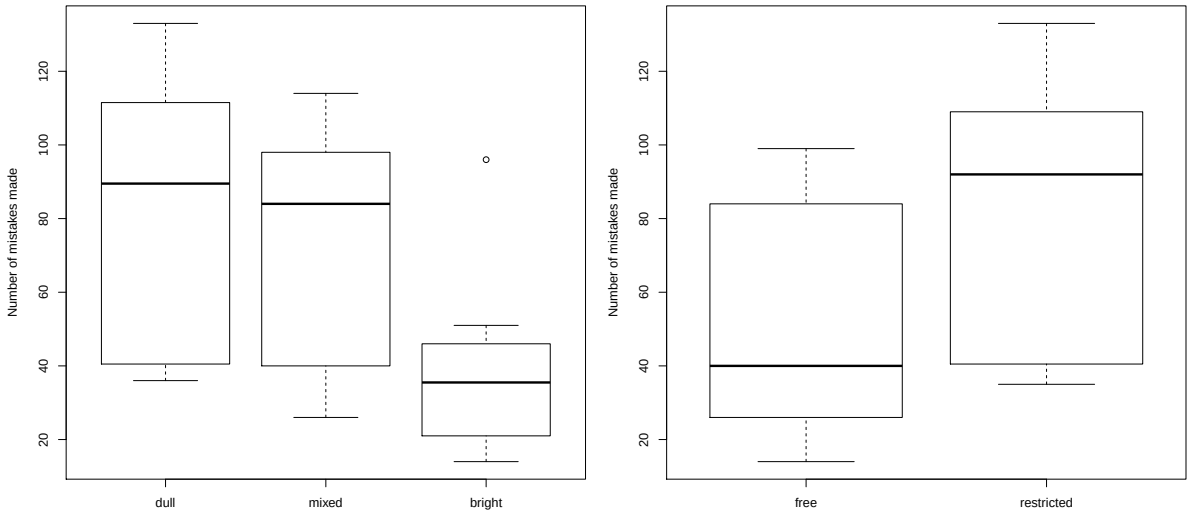
```
rats <- data.frame("mistakes" = c(26,41,41,26,36,39,14,16,82,86,87,99,51,96,39,104,
                    42,92,35,36,114,92,133,124),
                   "strain" = rep(c(rep("bright",2),rep("mixed",2),rep("dull",2)),4),
                   "condition" = c(rep("free",12),rep("restricted",12)))
```

General advice: You are also given hints on report writing, which aim to help you and is advisable to use them. However you will not be marked down if you do not follow these suggestions, any sensible report will be marked accordingly and marks will be awarded based on the conclusions drawn and the justifications given, not based on the structure of the report.

2. Sample report

2.1. **Data.** The data are the result of a 3×2 factorial experiment with 4 replications. The two available explanatory variables are the strain of the rat, bright, mixed or dull, and the conditions under which the experiment was performed, free or restricted.

Box-plots shown in Figure 1 illustrate that rats from the dull and mixed strains tend to make more mistakes than rats from the bright strain and that more mistakes are made under restricted than under free conditions.



(a) (b)

Figure 1. Box-plots of the number of mistakes made by the rats within the levels of the two factors.

An interaction plot shown in Figure 2 suggests that the effect of the two factors on the response is probably additive as the two lines are relatively parallel with all rats, regardless of strain, making more mistakes under restricted conditions than under free. Additionally, rats from the dull strain tend to make the most mistakes, followed by rats from the mixed strain, although the line connecting these two means is almost parallel to the x-axis suggesting that the difference might not be significant, and finally by rats from the bright strain.

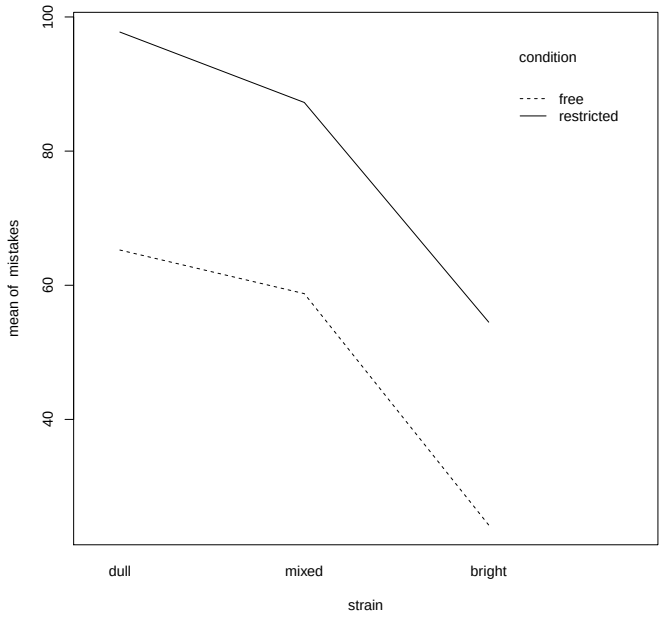


FIGURE 2. Interaction plot: the x-axis indicates the strain in which the rats belong, the two lines indicate the conditions under which the experiment was performed and the y-axis indicates the mean number of mistakes made by the rats in the maze.

2.2. **Modelling.** We start by fitting a normal linear model which includes the main effects and the interactions of the two potential explanatory variables. We assume that the errors are independent and normally distributed with constant variance. The model equation is:

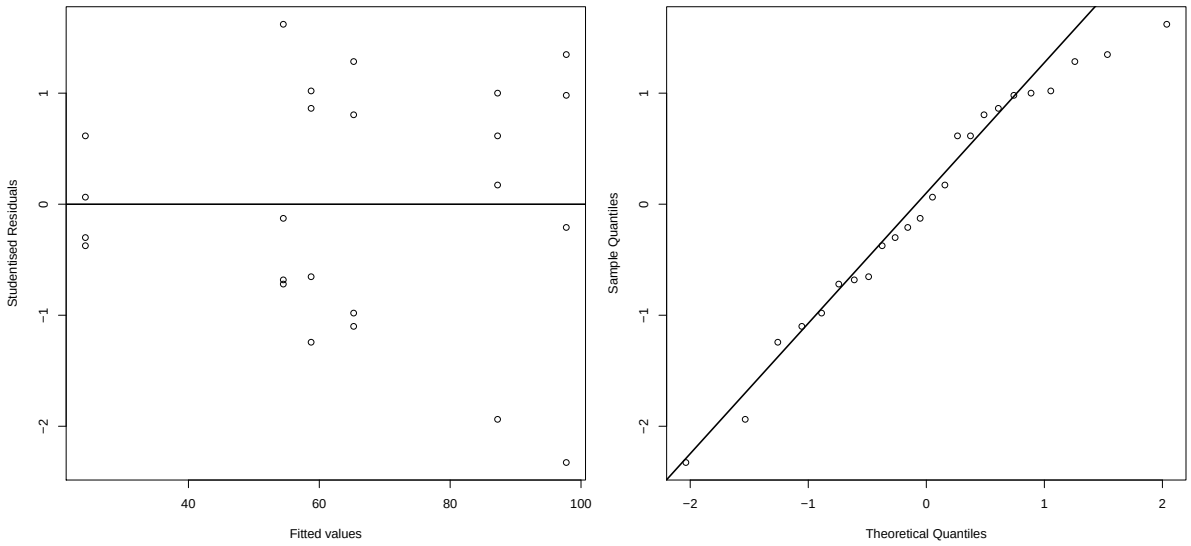
$$E(\text{mistakes}) = \beta_0 + \beta_1 \mathbb{I}_{\text{strain=mixed}} + \beta_2 \mathbb{I}_{\text{strain=bright}} + \beta_3 \mathbb{I}_{\text{condition=restricted}} + \beta_4 \mathbb{I}_{\text{strain=mixed, condition=restricted}} + \beta_5 \mathbb{I}_{\text{strain=bright, condition=restricted}} \quad (1)$$

Levels dull of factor strain and free of factor condition are used as baselines.

Before proceeding with model selection we check the validity of the model assumptions using the studentised residuals of the model. Figure 3 a) suggests that the assumption of constant variance does not hold as the points follow a clear funnel shape. The vertical spread of the points increases as the fitted values increase which is probably due to the fact that the response is a count. A more appropriate distribution instead of the normal would be the poisson which has variance equal to the mean.

A variance-stabilising transformation is required for the response. To identify the most appropriate transformation in this case we can use the Box-Cox family of transformations. The profile log-likelihood plot for parameter λ shown in Figure 4 suggests setting $\lambda = 0$ which corresponds to the log-transformation.

The full model is fitted to the log-transformed response and this time the diagnostic graphs are acceptable; in Figure 5 a) the points are randomly scattered in a relatively even band across the x-axis and in Figure 5 b) the points tend to follow the straight line through the origin. There are no studentised residuals outside the -3,3 range.



(a) (b)

FIGURE 3. Studentised residuals against fitted values, a), and normal QQ-plot of studentised residuals, b), for the model with interaction.

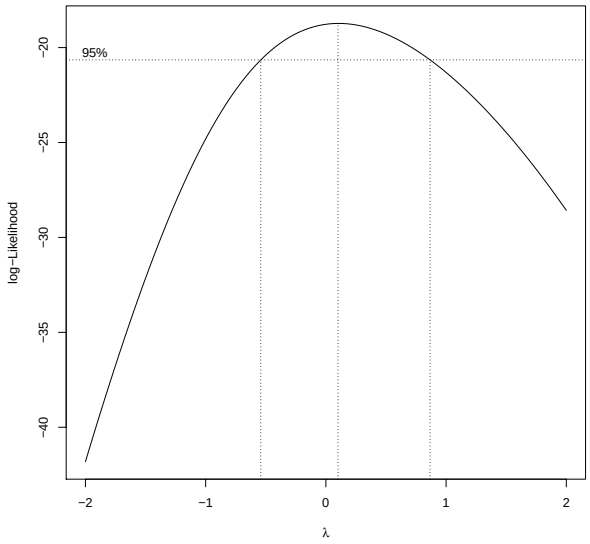
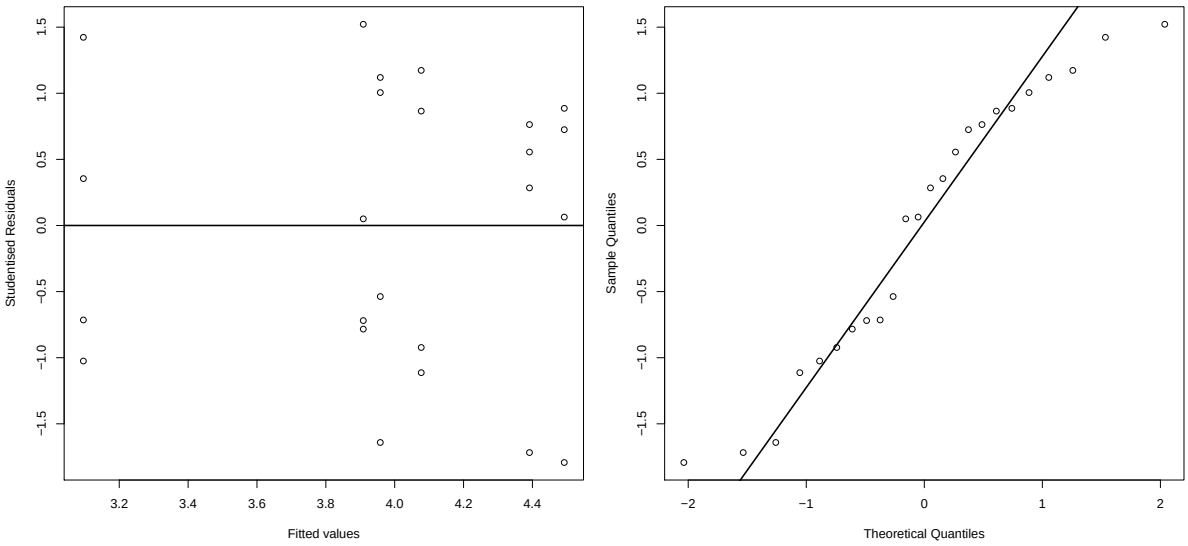


FIGURE 4. Profile log-likelihood plot for parameter λ of the Box-Cox transformation.

Proceeding with model selection, we use the F-test to test H_0 : the interaction between strain and condition is not significant versus the alternative that it is significant. The F-statistic of the test is equal to 0.38 which is not in the 5% critical region of an F distribution with 2 and 18 degrees of freedom (p-value = 0.69). Therefore we have no evidence against the null hypothesis and we can drop the 2 interaction terms from the model. This result agrees with our exploratory analysis of the data.

The additive model is fitted to the log-transformed response and the F-tests for dropping any of the two main effects from the model are significant as the ANOVA table demonstrates. Note



(a) (b)

FIGURE 5. Studentised residuals against fitted values, a), and normal QQ-plot of studentised residuals, b), for the model with interaction with log-transformed response.

that the order in which the terms are added to the model does not affect our conclusion in this case because the two factors are orthogonal since the design is balanced.

```
Analysis of Variance Table

Response: log(mistakes)
          Df Sum Sq Mean Sq F value  Pr(>F)
strain     2 2.8711 1.43556  5.7640 0.01055 *
condition  1 1.8409 1.84089  7.3915 0.01322 *
Residuals 20 4.9811 0.24906
```

Therefore, the model equation of our chosen model is:

$$E(\text{mistakes}) = \beta_0 + \beta_1 \mathbb{I}_{\text{strain=mixed}} + \beta_2 \mathbb{I}_{\text{strain=bright}} + \beta_3 \mathbb{I}_{\text{condition=restricted}} \tag{2}$$

The diagnostic graphs are similar to those obtained for the corresponding model when the interaction terms were included and are therefore acceptable. There are no studentised residuals outside the -3,3 range and no influential observations, which is also demonstrated in Figure 6 since no cook's distance is greater than $8/(n-2p) = 0.5$ in this case.

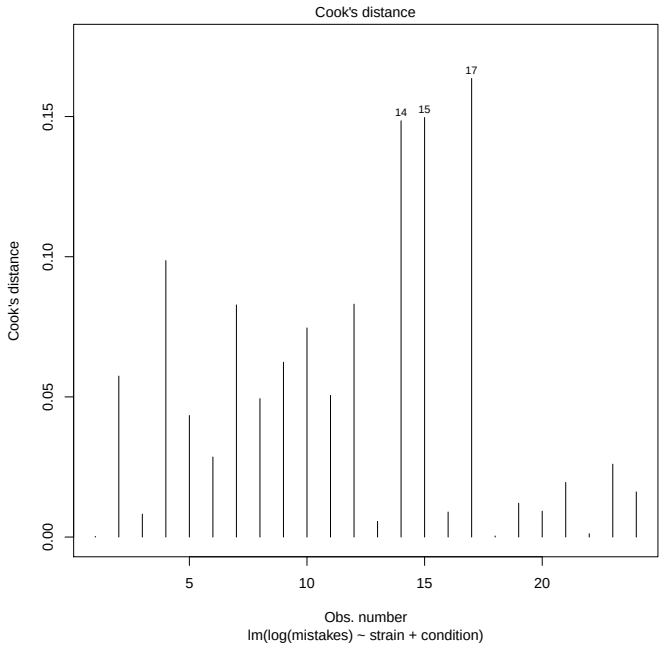


FIGURE 6. Cook's distances for the additive model.

2.3. **Interpretation.** The summary of our selected model is presented below:

```
Call:
lm(formula = log(mistakes) ~ strain + condition)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82406 -0.36427  0.05138  0.36099  0.78510

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.0078     0.2037  19.671 1.48e-14 ***
strainmixed         -0.1099     0.2495  -0.441   0.6642
strainbright        -0.7825     0.2495  -3.136   0.0052 **
conditionrestricted  0.5539     0.2037   2.719   0.0132 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4991 on 20 degrees of freedom
Multiple R-squared: 0.4861,     Adjusted R-squared: 0.409
F-statistic: 6.306 on 3 and 20 DF,  p-value: 0.003466
```

Keeping in mind that the response has been log-transformed we conclude that the number of mistakes made under free conditions by rats from the dull strain is expected to be exp(4), which is around 54 (95% CI: (36, 84)). The corresponding numbers for rats from the mixed and bright strain are exp(-0.11)= 0.9 (95% CI: (0.53, 1.51)) and exp(-0.78)= 0.46 (95% CI: (0.27, 0.76)) times that of the rats from the dull strain, respectively. However, Tukey's Honest Significance test, (Figure 7) suggests that the difference between the baseline i.e. the dull strain and the mixed strain is not significant. Finally, the number of log(mistakes) made under restricted conditions increases by 0.55 (95% CI: (0.13, 0.98)) compared to the log(mistakes) made under free conditions.

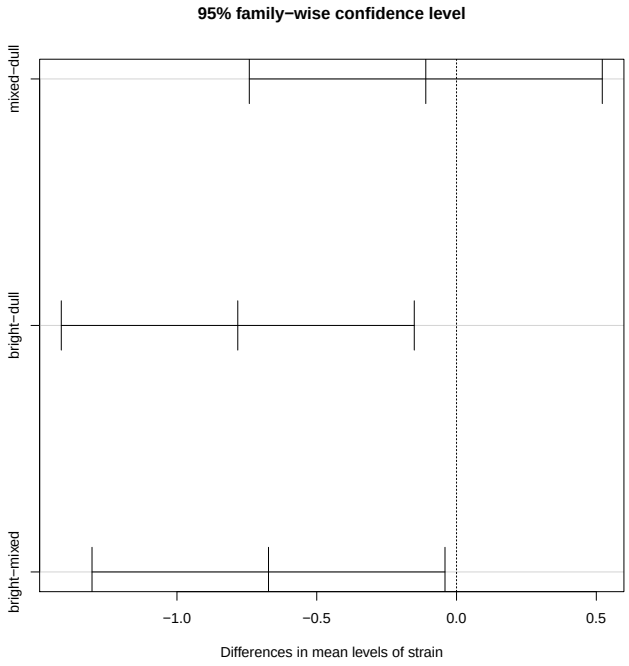


FIGURE 7. Tukey's Honest Significance Test.

2.4. **Conclusions.** After fitting the full model to the data it was noticed that the assumption of constant variance was violated and a logarithmic transformation of the response was used. This is likely to be due to the fact that the response is a count and the poison distribution, which allows for the variance to be equal to the mean instead of restricting it to be constant, is more appropriate in this case. For the model with the log-transformed response the interaction between the two factors was found insignificant and it was removed from the model. The final model included both main effects and suggested that rats from the dull and mixed strain are expected to make more mistakes than rats in the bright strain. It was also found that all rats, regardless of strain, are expected to make more mistakes under restricted conditions.

2.5. **R code.**

```
#Mock practical
#Rats

rats <- data.frame("mistakes" = c(26,41,41,26,36,39,14,16,82,86,87,99,
51,96,39,104,42,92,35,36,114,92,133,124),
       "strain" = rep(c(rep("bright",2),rep("mixed",2),rep("dull",2)),4),
 "condition" = c(rep("free",12),rep("restricted",12)))

attach(rats)

strain <- factor(strain, levels = c("dull", "mixed", "bright"))

pdf("rats1.pdf",height=8,width=8)
interaction.plot(x.factor=strain,trace.factor=condition,
response=mistakes,fixed=T)
dev.off()

pdf("ratsbox1.pdf",height=8,width=8)
boxplot(mistakes~strain, ylab="Number of mistakes made")
dev.off()

pdf("ratsbox2.pdf",height=8,width=8)
boxplot(mistakes~condition, ylab="Number of mistakes made")
dev.off()

rats.lmfit1 <- lm(mistakes ~ strain*condition)

pdf("rats2.pdf",height=8,width=8)
plot(fitted(rats.lmfit1),rstudent(rats.lmfit1),xlab="Fitted values",
ylab="Studentised Residuals")
abline(h=0,lwd=2)
dev.off()

pdf("rats3.pdf",height=8,width=8)
qqnorm(rstudent(rats.lmfit1),main=NULL)
qqline(rstudent(rats.lmfit1),lwd=2)
dev.off()

library(MASS)
```

```
pdf("rats4.pdf",height=8,width=8)
boxcox(rats.lmfit1)
dev.off()


rats.lmfit2 <- lm(log(mistakes) ~ strain*condition)

pdf("rats5.pdf",height=8,width=8)
plot(fitted(rats.lmfit2),rstudent(rats.lmfit2),xlab="Fitted values",
ylab="Studentised Residuals")
abline(h=0,lwd=2)
dev.off()

pdf("rats6.pdf",height=8,width=8)
qqnorm(rstudent(rats.lmfit2),main=NULL)
qqline(rstudent(rats.lmfit2),lwd=2)
dev.off()

anova(rats.lmfit2)

rats.lmfit3 <- lm(log(mistakes) ~  strain + condition)

plot(fitted(rats.lmfit3),rstudent(rats.lmfit3),xlab="Fitted values",
ylab="Studentised Residuals")
abline(h=0,lwd=2)

qqnorm(rstudent(rats.lmfit3),main=NULL)
qqline(rstudent(rats.lmfit3),lwd=2)

pdf("rats7.pdf",height=8,width=8)
plot(rats.lmfit3,4)
dev.off()

anova(rats.lmfit3)

summary(rats.lmfit3)

pdf("rats8.pdf",height=8,width=8)
plot(TukeyHSD(aov(log(mistakes) ~  strain + condition), "strain"))
dev.off()
```