Experimental Design and Data Analysis: Assignment 6

Andrew Bedard(2566978) & Simone van Gompel(2567525)
Group 19

May 23, 2015

Exercise 1

1

This can be added to the data by using the following R-code:

```
data$loglongevity = log(longevity)
```

Now there is a column in the data called loglongevity.

2

In Fig1 a pairplot of the fruitflies data is seen. Here you can see that activity and thorax have a correlation with the longevity and thus also loglongevity. Activity and thorax themselfs have no correlation, which is good, because that is part of the experiment set up.

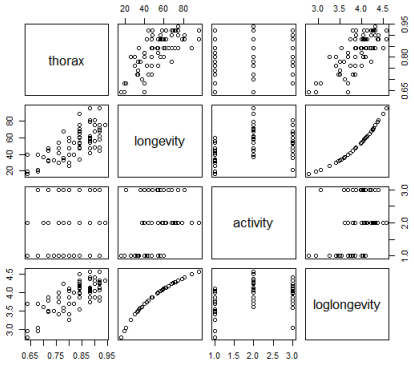


Figure 1: Pairsplot Fruitflies data

3

Using anova without taking thorax length into account has $p = 1.798e - 07$, which means we reject H_0. This means that action has an influence on the loglongevity.

4

In Table1 some information can be found on the effect of action on the loglongevity. It is clear that the more sexual activity the fruit flies have, the shorter their lifespan was. The only data that does not support this is the fly that lived for the longest in low, this is higher than the rest. But this can be because of luck, because the mean and median are still lower than isolated. In Fig3 this influence is graphically shown.

Table 1: Summary data of the different actions and the loglongevity

	isolated	low	high
min	37	21	16
median	62	56	40
mean	63.56	56.76	38.72
max	75	81	60

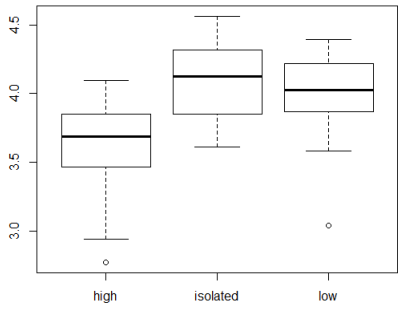


Figure 2: Boxplot of the influence of activity on loglongevity

5

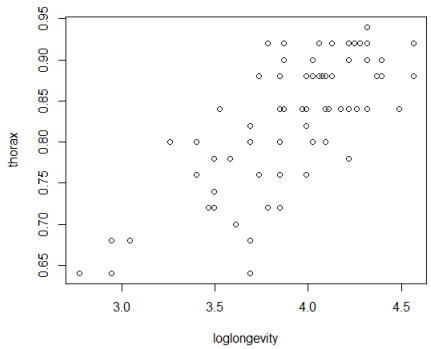
Using anova, this time with taking thorax length into account has p lower than 0.05, which means we reject H_0. This means that both action and thorax have an influence on the loglongevity.

6

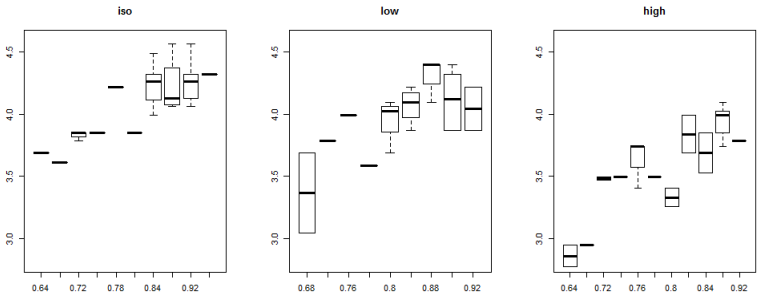
Seen in Fig3, it is clear that sexual activity decreases the loglongevity and thus also the longevity.

7

Thorax length increases the longevity and this stays the same over the different activities. The difference of the longevity between the flies with the smallest thorax and the biggest thorax are lager when there is more sexual activity.



(a) Thorax



(b) Thorax and Activity

Figure 3: Boxplot of the influence of activity on loglongevity

8

My preference goes out to the analysis without thorax length, this is because the analysis of the data is harder the more variables are included. You could say the analysis with thorax length is wrong, this is not important for the research question of the data.

9

In Fig4 the analysis that includes thorax length is evaluated. The qqplot shows that the residuals are normally distributed and the Fitted vs Residuals plot shows that the data has no heteroscedasticity.

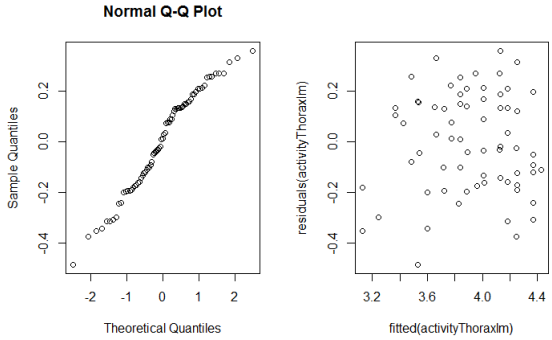


Figure 4: Evaluation of the analysis that includes thorax length

10

The ancova analysis with the longevity instead of the longevity has the following result:

```
Response: longevity
           Df Sum Sq Mean Sq F value    Pr(>F)
activity    2 8239.2  4119.6  38.120 5.686e-12 ***
thorax      1 7686.8  7686.8  71.127 2.624e-12 ***
Residuals  71 7673.0   108.1
```

While the ancova analysis with the loglongevity had the following result:

```
Response: loglongevity
           Df Sum Sq Mean Sq F value    Pr(>F)
activity    2 3.6665  1.8332  44.606 2.838e-13 ***
thorax      1 3.8786  3.8786  94.374 1.139e-14 ***
Residuals  71 2.9180  0.0411
```

This shows that with the longevity the influence of activity and thorax are bigger than with the loglongevity. In Fig4 the analysis of the longevity is evaluated. The qqplot shows that the residuals are normally distributed and the Fitted vs Residuals plot shows that the data has heteroscedasticity.

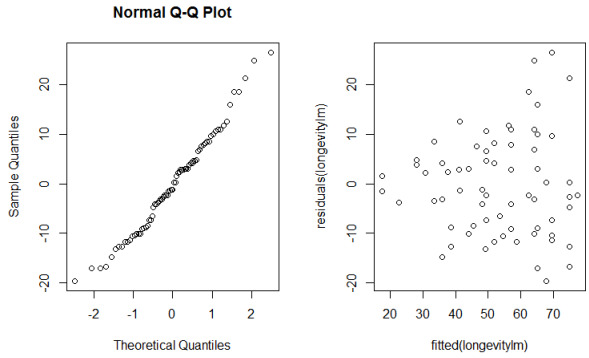


Figure 5: Evaluation of the analysis that includes thorax length

Exercise 2

1

The data contained in *psi.txt* was read in, the following figures were obtained.

2

Fitting a logistic regression model with psi and gpa as explanatory variables for the outcome being that the student passed their assessment or not, we obtain the following table:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -11.602      4.213  -2.754  0.00589 **
psi            2.338      1.041   2.246  0.02470 *
gpa            3.063      1.223   2.505  0.01224 *
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1    1
```

Figure 6: Parameter estimation for logistic regression model

Thus we determine that our logistic regression model should be:

$$Pr(pass = 1) = \frac{\exp(-11.602 + 2.338 * psi + 3.063 * gpa)}{1 + \exp(-11.602 + 2.338 * psi + 3.063 * gpa)} \quad (1)$$

3

Based on the p-value obtained in Figure:6, we reject the null hypothesis that there is no effect of psi on the outcomes of the students final assessment. Further based on our parameters for the logistic regression model, we see that a positive value, ie. 1, for *psi* causes an increase in probability of passing, so we conclude that *psi* does in fact work.

4

To estimate the probability that a student with a *gpa* equal to 3 who receives *psi* passes the assignment, we simply enter our values into equation 1, our logistic regression model.

$$Pr(pass = 1) = \frac{\exp(-11.602 + 2.338 * (1) + 3.063 * (3))}{1 + \exp(-11.602 + 2.338(1) + 3.063 * (3))} = 0.4813$$

So there is a 48.13 % chance of a student with *gpa* of 3 who receives *psi* of passing the final assignment.

Similarly if we wish to estimate the probability that a student with a *gpa* of 3 who does not receive *psi* passing the final assignment, we simply replace the value of *psi* with 0, and substitute this into equation 1

$$Pr(pass = 1) = \frac{\exp(-11.602 + 2.338 * (0) + 3.063 * (3))}{1 + \exp(-11.602 + 2.338(0) + 3.063 * (3))} = 0.0822$$

Thus there is a 8.22 % chance of a student with a *gpa* of 3 who did not take *psi* passing the final assignment.

5

We may investigate the relative change in odds of passing the final assignment for students with *psi* rather than those without *psi* by simply taking the exponential of the change in our linear predictor, in this case taking $psi = 1$ as opposed to 0 increases our linear predictor by 2.338 as we can see in Figure:6, thus our difference in odds is $exp(2.338) = 10.3605$.

6

Using the alternative method of analysis, and obtaining the matrix as is outlined in the R-code section 1.2

```
     [,1] [,2]
[1,]    3    8
[2,]   15    6
```

Column [,1] represents students that did not receive *psi*, where column [,2] are those students that did receive *psi*. Row [1,] represents students that passed the final assignment, and row [2,] are those who did not pass. Thus 15 is the number of students that did not receive *psi* and did not pass the final assignment, where 6 is the number of students that did receive *psi* and also did not pass the final assignment.

Performing the Fisher Test as outline in R-code section 1.2, we obtain the following:

```
        Fisher's Exact Test for Count Data

data: x
p-value = 0.0265
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval: 0.02016297 0.95505763
sample estimates:
odds ratio
0.1605805
```

The conclusion is that with a p-value of 0.0265, we should reject the null hypothesis that the probability of success for students who did and did not receive *psi* is equal, thus we conclude *psi* does increase a students chances of passing.

7

8

Exercise 3

1

2

3

4

5

1 R-Code

1.1 Exercise 1

1.2 Exercise 2

1.3 Exercise 3