Workshop

STAT3015/4030/7030 Generalised Linear Modelling

The Australian National University

Week 5, 2017

Overview

Background information

The question uses the `fruitfly` data in the text file `fruitfly.txt`. The title of the original study conducted by Linda Partridge and Marion Farquhar: "Sexual activity reduces lifespan of male fruitflies" (Partridge and Farquhar , 1981), provides a brief description of their key research question.

Read these data into R and create a new factor variable (`Activity`) to summarise the levels of sexual activity, as follows:

$$\text{Activity} = \begin{cases} \text{A, Partners=0 \& Type=9} \\ \text{B, Partners=1 \& Type=0} \\ \text{C, Partners=1 \& Type=1} \\ \text{D, Partners=8 \& Type=0} \\ \text{E, Partners=8 \& Type=1} \end{cases}$$

Main residual plot

Fit an ordinary (normally distributed) additive linear model with `Longevity` as the response variable, `Activity` as an exploratory factor and `Thorax` as a continuous covariate. Produce a plot of the residuals against the fitted values for this model. Are there any obvious problems with this plot?

Any useful information?

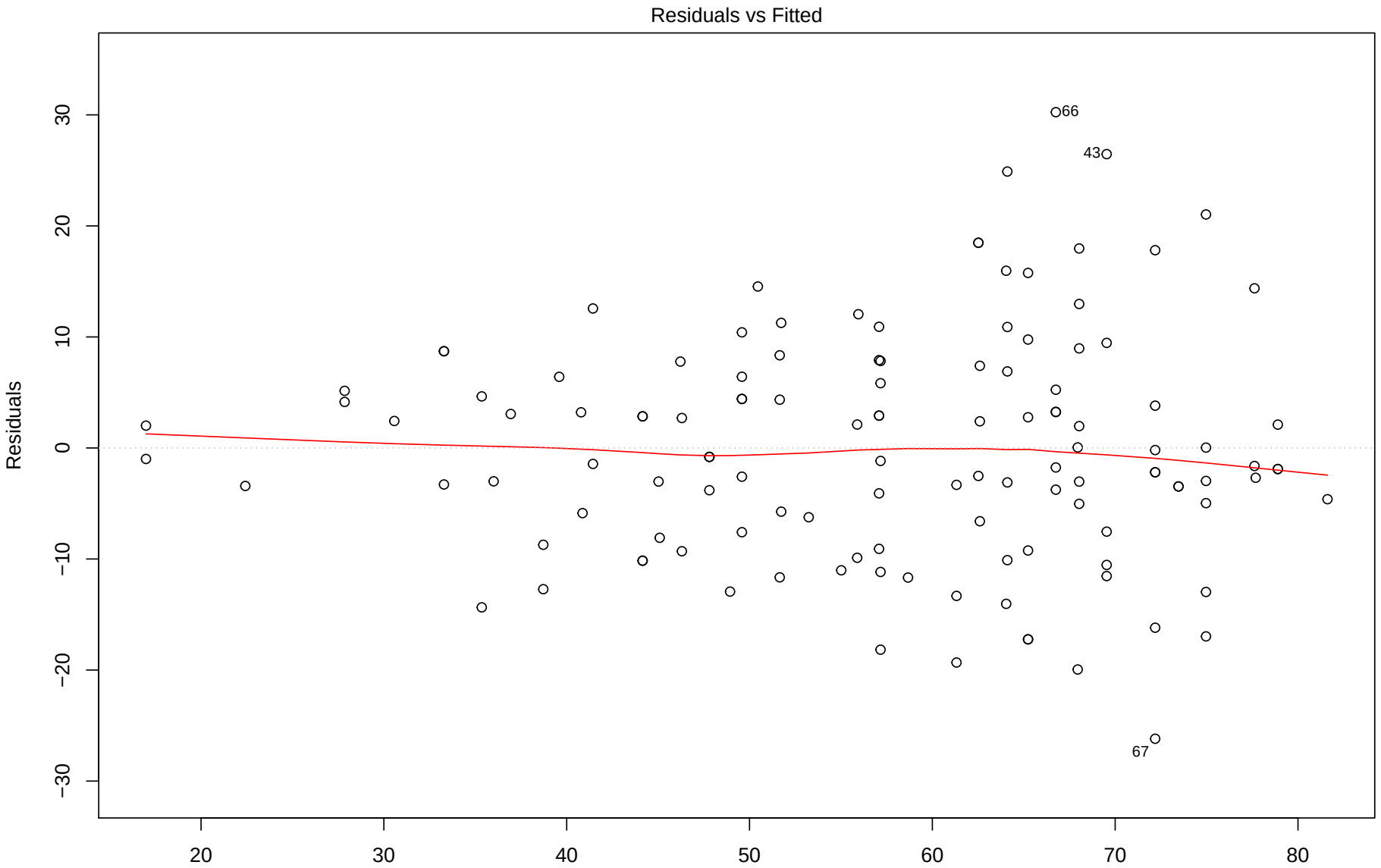
Main residual plot

Fit an ordinary (normally distributed) additive linear model with `Longevity` as the response variable, `Activity` as an exploratory factor and `Thorax` as a continuous covariate. Produce a plot of the residuals against the fitted values for this model. Are there any obvious problems with this plot?

Any useful information?

- Simple Analysis of Covariance (ANCOVA) model without interaction
- Need to create variable `Activity`, possibly by using `ifelse(...)` or other methods
- Main residual plot and assumptions

Main residual plot



Variance stabilising transformation

Refit the model in part (a), applying a `log()` (using the default in R, which are logarithms to base e) transformation to the response variable.

For this modified model, use the `rstandard()` function to produce a plot of the internally Studentised residuals against the fitted values; using different plotting characters for the five different levels of `Activity`.

Variance stabilising transformation

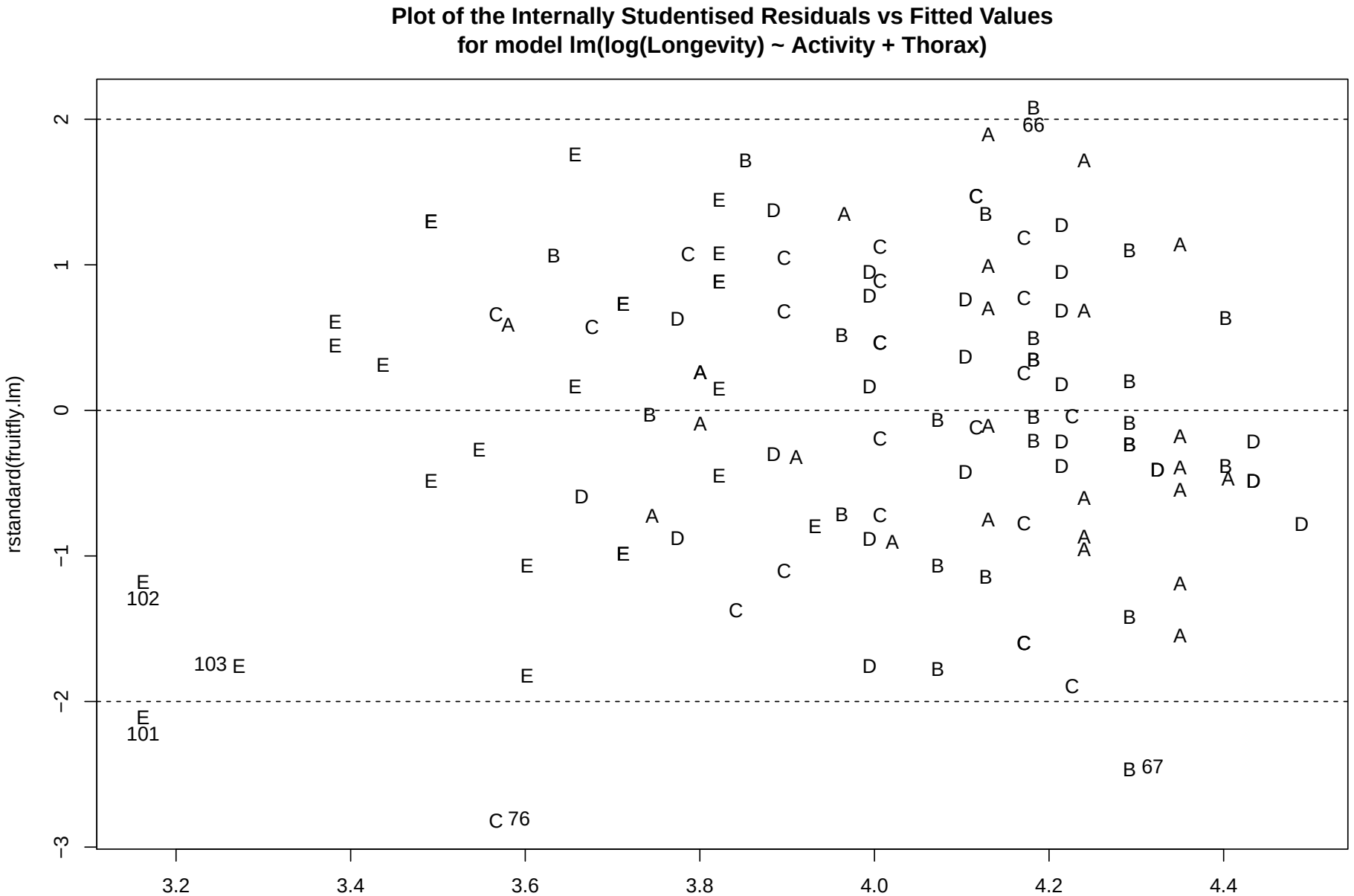
Refit the model in part (a), applying a `log()` (using the default in R, which are logarithms to base e) transformation to the response variable.

For this modified model, use the `rstandard()` function to produce a plot of the internally Studentised residuals against the fitted values; using different plotting characters for the five different levels of `Activity`. Internally

Studentised residual, r_i, is calculated as:

$$\begin{aligned}\frac{e_{i,-i}}{\sqrt{\text{Var}(e_{i,-i})}} &= \frac{e_i}{(1-h_{ii})\text{Var}(e_{i,-i})} \\ &= \frac{e_i}{(1-h_{ii})\sqrt{\sigma^2/(1-h_{ii})}} \\ &= \frac{e_i}{\sigma\sqrt{1-h_{ii}}}.\end{aligned}$$

Internally Studentised residual plot



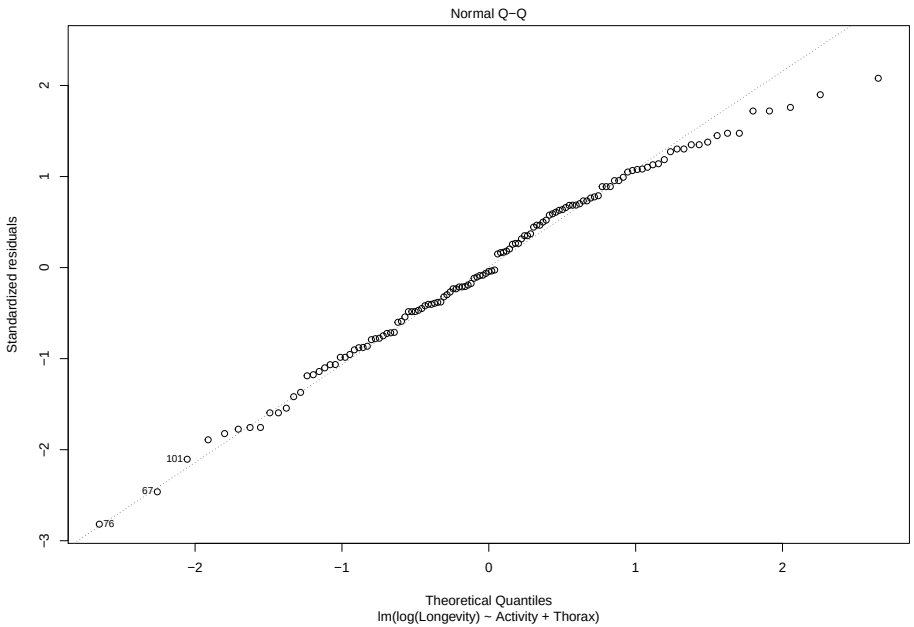
Diagnostic plots

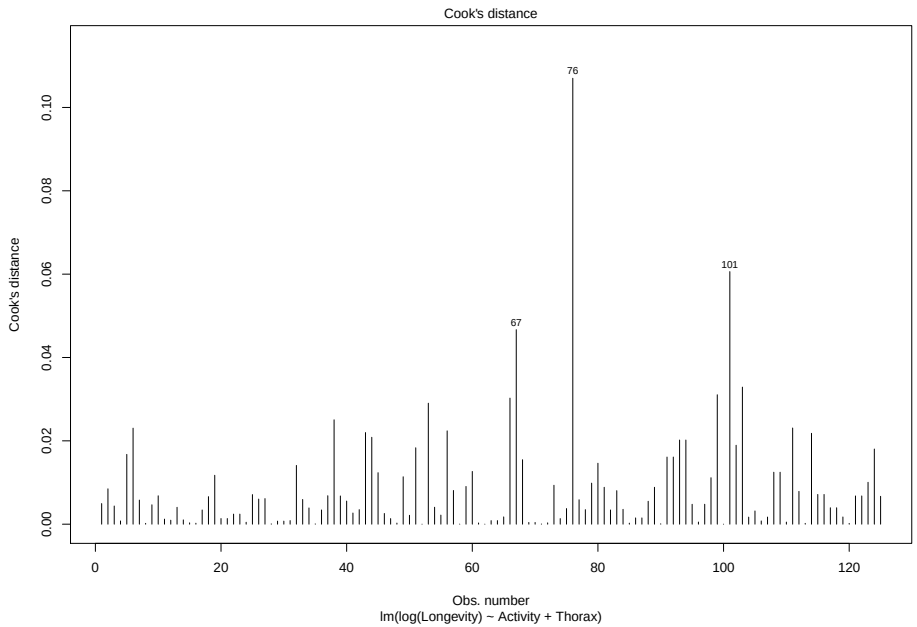
Produce a normal quantile plot of the residuals from the model in part (b). Also produce a bar plot of Cooks Distances for each of the observations.

Use the `rstudent()` and `hatvalues()` functions to calculate the externally Studentised residuals and leverage values for any observations that stand out on these additional residual plots and compare with appropriate cut-offs.

Comment on the plots and statistics you have just produced and discuss whether or not there are any outliers.

Diagnostic plots





Algebraic equation

Give the algebraic equation for the underlying population model fitted in part (b), including any assumptions about the error distribution, full details of the variables included in the model and the constraints applied to any factor variables.

Algebraic equation

The fitted ANCOVA model can be expressed as:

$$\log(\text{Longevity})_{ij} = \beta_0 + \tau_i + \beta_1 \text{Thorax}_{ij} + \varepsilon_{ij} \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

In the above equation, i represents the 5 different levels of `Activity` : $i =$ "A", "B", "C", "D", "E" , and $j = 1, 2, \cdots, 25$ (equivalent to the different values of the `ID` variable) represents the observations within each of the five different `Activity` groups.

Multiplicative ANCOVA model

Compare the model in part (b) with a multiplicative model that includes an interaction term between the factor variable (`Activity`) and the covariate (`Thorax`).

Describe how this additional term modifies the relationship between the response variable and the covariate for the different levels of the factor variable.

Is this additional term a significant improvement to the model? Give full details of an appropriate hypothesis test.

Multiplicative ANCOVA model

The multiplicative model includes an interaction term which allows for different slopes as well as different intercepts for the five different `Activity` groups.

$$\log(\text{Longevity})_{ij} = \beta_0 + \tau_i + \beta_1 \text{Thorax}_{ij} + \gamma_i \text{Thorax}_{ij} + \varepsilon_{ij}$$

$$\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2) \quad \tau_A = \gamma_A = 0$$

```
> anova(lm(log(Longevity) ~ Thorax * Activity))
Analysis of Variance Table

Response: log(Longevity)
                Df Sum Sq Mean Sq  F value Pr(>F)
Thorax           1 6.4256  6.4256 176.4955 <2e-16 ***
Activity         4 4.1499  1.0375  28.4970 <2e-16 ***
Thorax:Activity  4 0.2273  0.0568   1.5611 0.1894
Residuals      115 4.1868  0.0364
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Multiplicative ANCOVA model

The F test associated with the additional Thorax:Activity interaction term tests:

$$\mathrm{H}_0 : \frac{\sigma^2_{Addition}}{\sigma^2_{Error}} = 1 \quad \mathrm{H}_\mathrm{A} : \frac{\sigma^2_{Addition}}{\sigma^2_{Error}} > 1 \quad \equiv \quad \mathrm{H}_0 : \gamma_A = \gamma_B = \gamma_C = \gamma_D = \gamma_E = 0 \quad \mathrm{H}_\mathrm{A} : \text{not all } \gamma_j = 0$$

$$F = \frac{MS_{Addition}}{MS_{Error}} = \frac{0.0568}{0.0364} = 1.5611 \quad \sim \quad F_{4,115}(0.95) = 2.4506$$

So, as $p = 0.1894$ is not less than $\alpha = 0.05$ (or observed F = 1.5611 is not greater than 2.4506), do not reject H_0 in favour of H_A, and conclude that the Thorax:Activity interaction term is not a significant addition to the model and that separate slopes for the different Activity groups are NOT required.

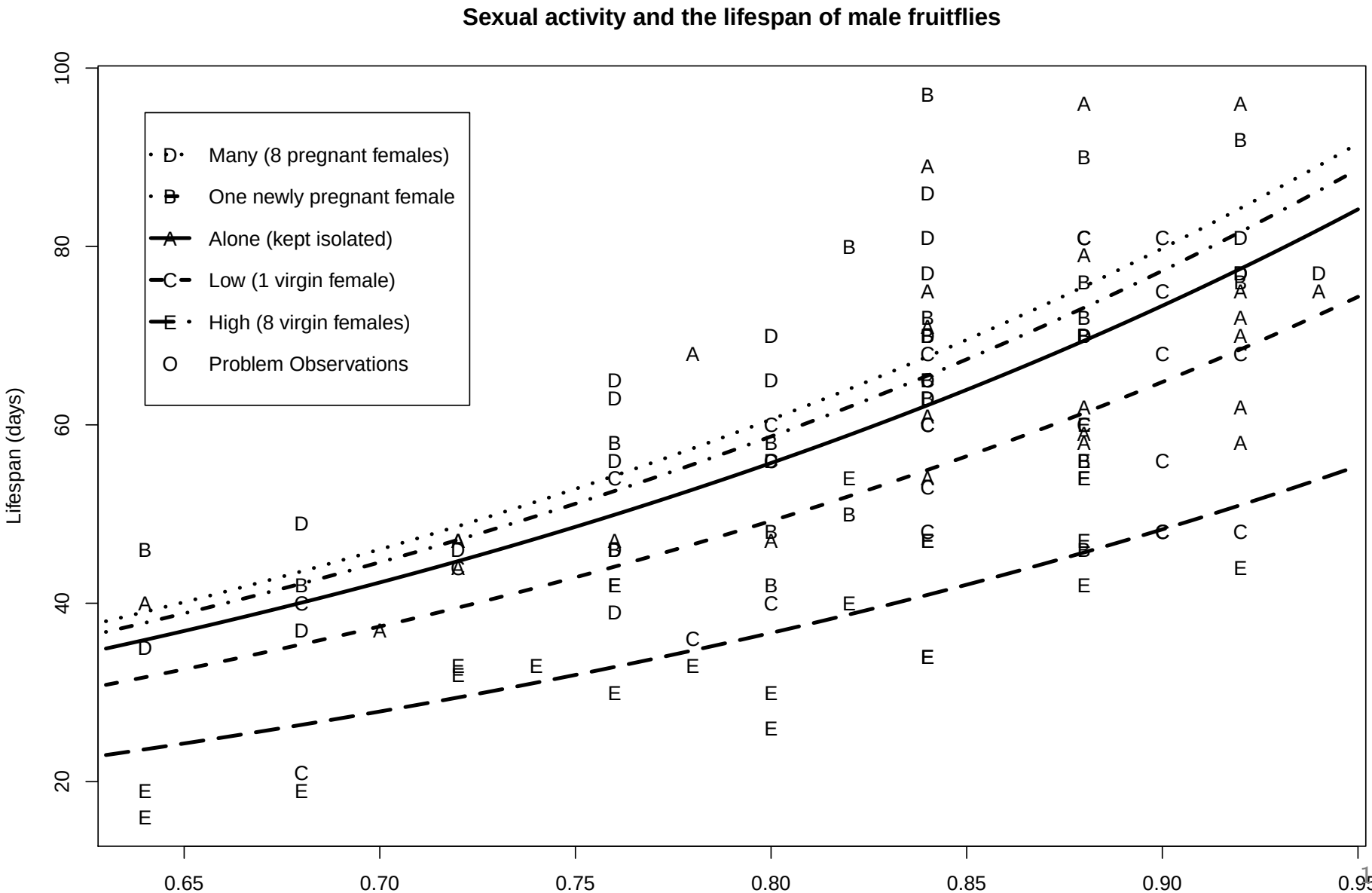
Scatter plot with lines/curves

Produce a plot of the data on the original scale (not the log scale) with different plotting characters for the five levels of `Activity`.

Include five curves on this plot, to represent the fitted model from part (b) for the five levels of `Activity`.

Also highlight on the plot any potential outliers you identified in part (c).

Scatter plot with lines/curves



Discussion of results

Present the ANOVA table and the summary table of the coefficients for the model in part (b). Use these tables and the plot in part (f) to discuss the results of the analysis you have conducted so far.

```
> anova(fruitfly.lm)
Analysis of Variance Table

Response: log(Longevity)
           Df Sum Sq Mean Sq F value    Pr(>F)
Activity    4 5.1809  1.2952  34.918 < 2.2e-16 ***
Thorax      1 5.3946  5.3946 145.435 < 2.2e-16 ***
Residuals 119 4.4141  0.0371
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Discussion of results

```
> summary(fruitfly.lm)

Call:
lm(formula = log(Longevity) ~ Activity + Thorax)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52208 -0.13457 -0.00799  0.13807  0.39234

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.82123    0.19442   9.368 5.89e-16 ***
ActivityB    0.05203    0.05453   0.954   0.3419
ActivityC   -0.12391    0.05448  -2.275   0.0247 *
ActivityD    0.08401    0.05491   1.530   0.1287
ActivityE   -0.41826    0.05509  -7.592 7.79e-12 ***
Thorax       2.74895    0.22795  12.060  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1926 on 119 degrees of freedom
Multiple R-squared:  0.7055, Adjusted R-squared:  0.6932
F-statistic: 57.02 on 5 and 119 DF,  p-value: < 2.2e-16
```

Mixed Effects model

Now modify the model in part (b) to include the `ID` variable as a random effect in an additive mixed effects model.

Describe the changes to the underlying population model described in part (d).

Discuss whether or not this is an appropriate treatment of the ID variable.

Mixed Effects model

The `ID` variable appears to be simply a label for the observations in each group.

`ID` could be interpreted as a blocking factor if there were some sensible connection between fruitflies with the same `ID` value (eg. all fruitflies labelled 1 came from the same genetic stock, which differed from fruitflies labelled 2 etc.)

Blocking is a technique for dealing with nuisance factors. A nuisance factor is a factor that has some effect on the response, but is of no interest to the experimenter; however, the variability it transmits to the response needs to be minimized or explained.

Mixed Effects model

Adding `ID` as a random effect to the model described in part (d):

$$\log(\text{Longevity})_{ij} = \beta_0 + \delta_j + \tau_i + \beta_1 \text{Thorax}_{ij} + \varepsilon_{ij} \quad \varepsilon_{ij}$$

In the above equation, i, j are se before, however, the variance model now has two independent components: $\delta_j \overset{\text{iid}}{\sim} N(0, \sigma^2_\delta)$ and $\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2_\varepsilon)$

Mixed Effects model

```
> fruitfly.lmer <- lmer(log(Longevity) ~ Thorax + Activity + (1|factor(ID)))
> fruitfly.lmer

Linear mixed model fit by REML ['lmerMod']
Formula: log(Longevity) ~ Thorax + Activity + (1 | factor(ID))
REML criterion at convergence: -38.5595
Random effects:
  Groups     Name        Std.Dev.
  factor(ID) (Intercept) 4.911e-09
  Residual               1.926e-01
Number of obs: 125, groups:  factor(ID), 25
Fixed Effects:
(Intercept)       Thorax    ActivityB    ActivityC    ActivityD    ActivityE
    1.82123      2.74895      0.05203     -0.12391      0.08401     -0.41826
```

Intraclass correlation coefficient

Present and examine the analysis of variance table and table of coefficients for the new mixed effects model in part (h).

How has this changed from the summary output presented in part (g)?

Calculate the intraclass correlation coefficient for the mixed effects model and comment on the results.

Intraclass correlation coefficient

```
Random effects:
  Groups      Name        Variance  Std.Dev.
  factor(ID) (Intercept) 2.412e-17 4.911e-09
  Residual               3.709e-02 1.926e-01
Number of obs: 125, groups:  factor(ID), 25
```

Intraclass correlation coefficient is calculated as:

$$\frac{\hat{\sigma}_\delta^2}{\hat{\sigma}_\delta^2 + \hat{\sigma}_\varepsilon^2} = \frac{2.412 \times 10^{-17}}{2.412 \times 10^{-17} + 0.1926} \approx 0$$

So, there does not appear to any real additional information in the `ID` variable (i.e. additional to what is already contained in `Thorax`).

References

Partridge, Linda and Farquhar, Marion (1981), 'Sexual activity reduces lifespan of male fruitflies', *Nature* **294**(58), 580-582.