RESEARCH SCHOOL OF
FINANCE, ACTUARIAL STUDIES AND APPLIED STATISTICS
College of Business & Economics, The Australian National University

GENERALISED LINEAR MODELS

(STAT3015/STAT4030/STAT7030)

Solutions to Assignment 1 for 2016

Data

Many of the projects I have worked on as a statistician have involved data that was considered private (such as health data) or data to which access was restricted (for example, data that was designated "commercial-in-confidence"). For these reasons, it is not always easy to source realistic data for use in teaching statistics and so groups of statisticians maintain repositories of examples of real data that are in the "public domain". In many countries, there are Internet repositories of data available for use in the teaching of introductory statistics.

The data to be used in this year's assignments come from one such repository: the data archive associated with the Journal of Statistics Education (JSE), a publication of the American Statistical Association (www.amstat.org/publications/jse/jse_data_archive.htm).

Datasets in the JSE data archive are typically accompanied by a file which give a description of the variables included in the data (the "meta-data") and are also often accompanied by an associated article in the journal (and occasionally even by references to other sources). The fruitfly data, which we will be using in question 1 of this year's assignments, includes both of the above accompanying documents.

You can download a text file containing the fruitfly data and the associated documents from the JSE website (www.amstat.org/publications/jse/jse_data_archive.htm) or the data is also available on Wattle in the file fruitfly.csv, which includes a header row with the variable names. I have also downloaded a copy of the meta-data text file (fruitfly.txt), and made this file available on Wattle.

In these model solutions, I have included the assignment questions and a few additional comments in italics. These are not essential parts of the solutions to the assignment, but they do make this document into a more readable, self-contained report. Even with these optional "extras" this report is well within the 10 *page limit and does so without using an unreasonably small font size or any unreadable formatting.*

The essential solutions are the parts NOT in italics, but there are a number of "tangents" or different approaches you could have included in your analysis. I discuss a number of these "tangents" in the comments which I have included in the appendix of R *commands. Your solutions may sensibly include some of these "tangents", but if you include too many and go over the page limit as a result, you may well use lose marks. On the other hand, good additional discussion, that is concise and well expressed, may well compensate if you miss a few of the points that we were expecting to find.*

Question 1 **(22 marks)**

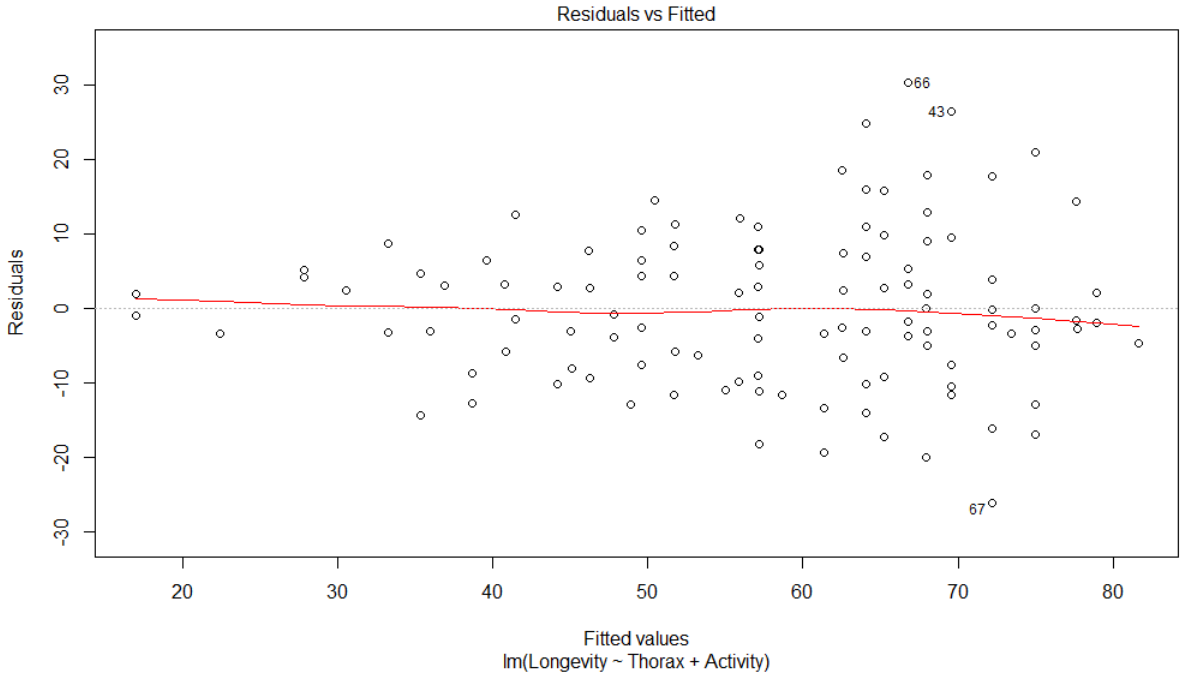
Read the description of the fruitfly data in the text file fruitfly.txt, which is available on Wattle (you may also choose to read the other articles referenced in this file, which are all available on-line or through the ANU library e-resources). The title of the original study conducted by Linda Partridge and Marion Farquhar: "Sexual activity reduces lifespan of male fruitflies" (*Nature*, Vol. 294, 10 December 1981, pp. 580-582), provides a brief description of their key research question (locate and read this article for further details).

The original data for this study is available in the file fruitfly.csv, which is also available on Wattle. Read these data into R and create a new factor variable (Activity) to summarise the levels of sexual activity, as follows:

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

Note that there is a copy of the data in the faraway library, where the above levels are listed as "isolated", "one", "low", "many" and "high", respectively. However, the data in the faraway library is missing the first observation, so do NOT use that data for this assignment.

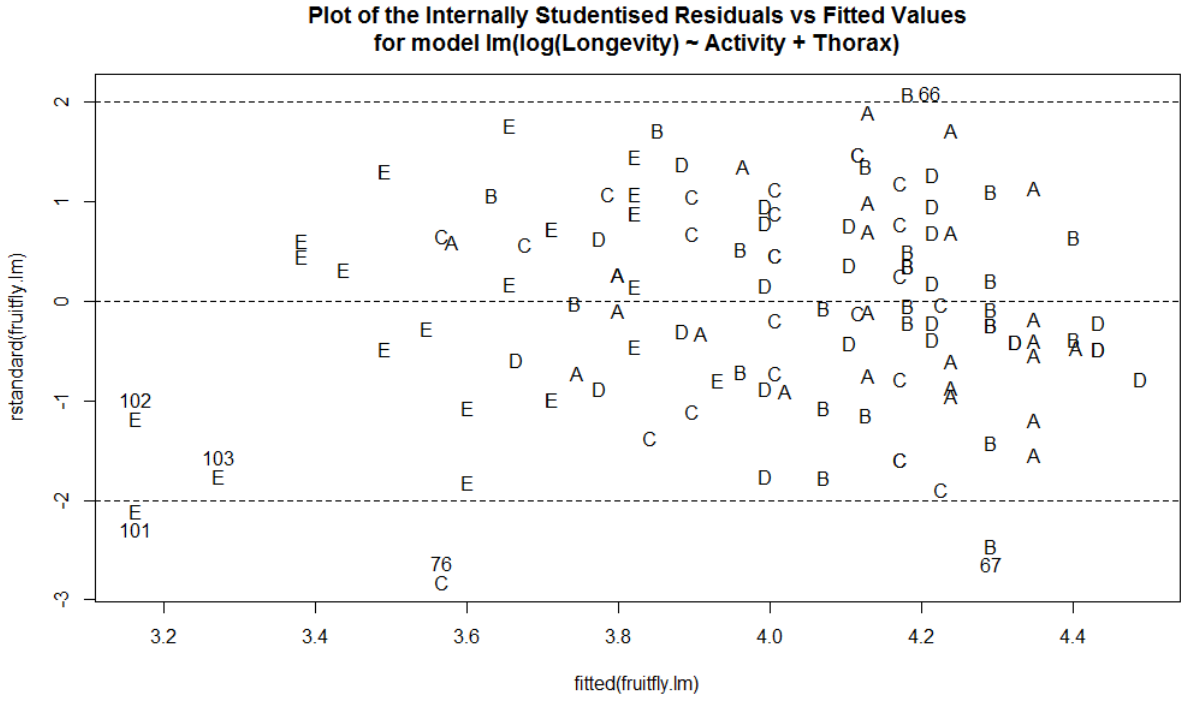
(a) Fit an ordinary (normally distributed) additive linear model with Longevity as the response variable, Activity as an exploratory factor and Thorax as a continuous covariate. Produce a plot of the residuals against the fitted values for this model. Are there any obvious problems with this plot? **(1 mark)**



The above residual shows fairly obviously increasing variance (heteroscedasticity). Possible remedies, discussed in the prerequisite Regression Modelling course, include variance weights (proportional to Thorax or some function of Thorax) and/or variance stabilising transformation(s) to either or both of the continuous variables (the response Longevity and the covariate Thorax). We will investigate the second of these solutions in part (b).

Question 1 continued

(b) Refit the model in part (a), applying a log() transformation to the response variable (using the default in R, which are logarithms to base *e*). 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. Does the log transformation appear to have to have corrected any problems identified in part (a)? Identify unusual observations on your plot and discuss any other interesting features of the plot. **(3 marks)**



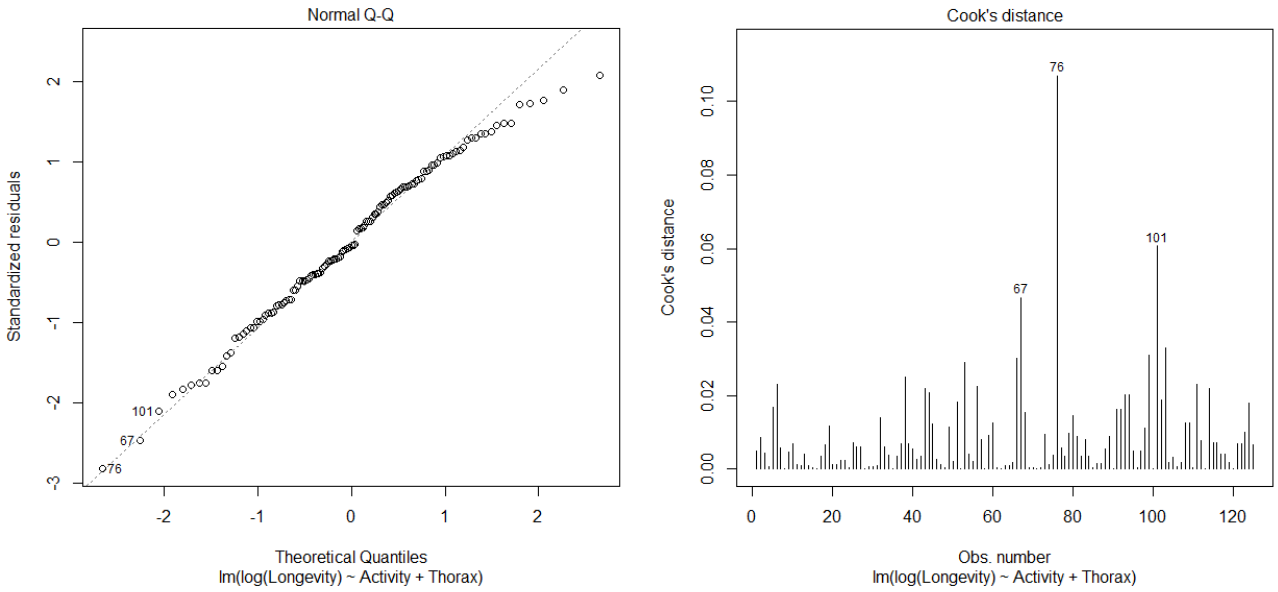
The log transformation to the response variable does appear to have successfully corrected the problem with non-constant variance.

The three observations (numbers 101, 102 and 103) with the smallest Thorax values in group E (the group with "high" levels of sexual Activity) form a distinct cluster in the lower left of this plot.

Note that 4 observations lie more outside the ±2 limits which I have included on the plot, however, this only represents 4/125 or 3.2% of the observations (within the expected 5%).

(c) Produce a normal quantile plot of the residuals from the model in part (b). Also produce a bar plot of Cook's 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. Do NOT refit the model to exclude any outliers. **(3 marks)**

Question 1, part (c) continued



The normal quantile plot is reasonable, showing only a slight departure from normality (slightly too short in the positive tail). The plot of Cook's distance has one stand-out, observation 76, which has a still reasonable Cook's distance value of just over 0.1. We can further investigate all of the observations that had Standardized residual values greater than ±2, which were also identified in part (b):

```
> correct.df <- fruitfly.lm$df.residual - 1
> c(qt(0.025, correct.df), qt(0.975, correct.df))
[1] -1.980272  1.980272
> rstudent(fruitfly.lm)[c(66,67,76,101)]
       66        67        76       101
 2.109336 -2.517160 -2.904932 -2.136116
```

The externally Studentised residuals are calculated by completely excluding the current observation (and so involve one less residual degree of freedom than the model). All four of the rstudent() values lie outside the suggested interval (particularly observations 76 and 67) suggesting that these observations are "mean-shift" or vertical outliers.

```
> range(hatvalues(fruitfly.lm))
[1] 0.04000000 0.09381236
> 2*length(coef(fruitfly.lm))/length(Longevity)
[1] 0.096
> hatvalues(fruitfly.lm)[c(66,67,76,101)]
        66         67         76        101
0.04029047 0.04414541 0.07479224 0.07585997
```

However, none of the observations (including the four of interest) exceed the arbitrary $2p/n$ cut-off, so there are no observations that have more than twice the average leverage, suggesting that no observation has been particularly influential in determining how the model fits that data.

For this reason, I have chosen not to exclude any observations as being influential outliers. The stated research question in the article in Nature is fairly general: "does increased sexual activity adversely affect longevity in male fruit flies?" The model including the outliers should be sufficient to address this question. The vertical outliers may have perturbed the size of the estimated coefficients of this model, so we may have to deal with the outliers, if we want a predictive model to address a more specific research question, such as quantifying the amount by which longevity is affected

Question 1 continued

(d) 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. Is this an example of an ANOVA model or an ANCOVA model? **(2 marks)**

$$\text{log(Longevity)}_{ij} = \beta_0 + \tau_j + \beta_1 \text{Thorax}_{ij} + \varepsilon_{ij} \quad \varepsilon_{ij} \sim i.i.d.\ N\left(0, \sigma^2\right)$$

Where j represents the 5 different levels of Activity : j = {"A", "B" "C", "D", "E"}, and i = 1, 2, …, 25 (equivalent to the different values of the ID variable) represents the observations within each of the five different Activity groups. In the R code, I have not specified any special contrasts, so the default treatment contrasts will have been used with Activity group "A" as the reference level, so the constraint applied is $\tau_A = 0$.

This is a "parallel lines" analysis of covariance (ANCOVA) model with Activity as an experimental factor and Thorax as a continuous covariate. As will we see in part (f), the model consists of five parallel lines (with different intercepts, but the same slope) for the 5 different Activity groups on a graph of Longevity (on the log) scale against Thorax, which become a series of proportional curves when we "back-transform" Longevity to the scale used in the original data.

(e) 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. **(2 marks)**

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

$$\text{log(Longevity)}_{ij} = \beta_0 + \tau_j + \beta_1 \text{Thorax}_{ij} + \gamma_j \text{Thorax}_{ij} + \varepsilon_{ij} \quad \varepsilon_{ij} \sim i.i.d.\ N\left(0, \sigma^2\right) \quad \gamma_A = 0$$

```
> 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
```

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

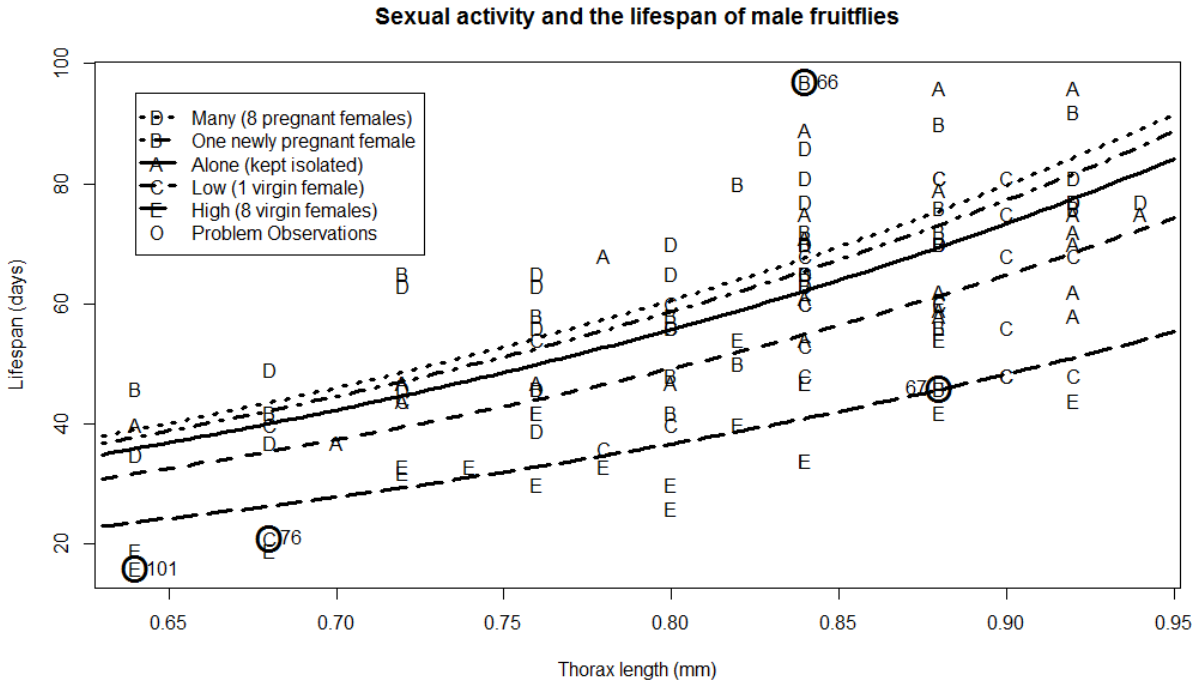
$$\text{H}_0 : \frac{\sigma^2_{Addition}}{\sigma^2_{Error}} = 1 \quad \text{H}_\text{A} : \frac{\sigma^2_{Addition}}{\sigma^2_{Error}} > 1 \quad \equiv \quad \text{H}_0 : \gamma_A = \gamma_B = \gamma_C = \gamma_D = \gamma_E = 0 \quad \text{H}_\text{A} : \text{not all } \gamma_j = 0$$

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

So, as p = 0.1894 is not less than α = 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.

Question 1 continued

(f) 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). **(2 marks)**



(g) 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. **(2 marks)**

```
> 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
```

The above significant F tests (p-values << $\alpha = 0.05$) indicate that we do need different intercepts (for the parallel lines on the log scale), for at least some of the five different levels of Activity AND that there is a significant relationship between Thorax length and log(Longevity), which implies there will also be a significant relationship between Thorax length and Longevity, as log is an order preserving transformation.

Judging by the above plot, the curves for the two control groups that involve some company, but no sexual activity (B and D), are slightly higher than the main control group (A) which involves no company or sexual activity. This suggests that there might be some positive effects on longevity associated with increased competition for food (hence lower food intake) and/or some positive effects from simply having company.

The two groups that involve sexual activity (C and E) are both located below the control groups, suggesting that increased sexual activity does reduce longevity.

Question 1, part (g) continued

```
> 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
```

The positive coefficients in the above table of coefficients for groups B and D confirm that both of these groups have slightly higher curves than the reference group (A), but as the corresponding t-tests are not significant, there is not strong evidence for any positive effect on longevity due to company.

The larger negative coefficients associated with the two groups that involve sexual activity (C and E) and the significant corresponding t-tests confirm that both of these groups have slightly lower curves than the reference group, so there is evidence that increased sexual activity does reduce longevity. Note that as group D (lots of company but no sexual activity) is above the reference group, whilst group E (same amount of company and presumably lots of sex) is significantly below the reference group, that the comparison between these groups will also be significant. Similarly, the comparison between group B (1 female for company but no sex) and group C (1 female for both company and sex) will also be significant.

Finally the coefficient of Thorax and the "slopes" of the curve in the plot are positive (and significant), so increased Thorax length leads to increased Longevity, i.e. bigger fruitflies tend to live longer.

(h) 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. (Hint – you may need to investigate possible relationships between ID and the other variables). **(3 marks)**

A little exploratory data analysis (see the appendix of R commands for details) reveals there is a very strong relationship between ID and Thorax. This relationship is so close that ID can be viewed as simply a 25 category summary of Thorax. I suspect that the researchers ordered the experimental units (the male fruitflies) into 25 size categories with 5 units in each ID category and then randomly assigned one member of each category to each of the 5 different experimental treatments (the 5 different levels of Activity. ID is effectively a blocking factor and appropriately treated as a random effect.

Question 1, part (h) continued

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

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

Where *i, j* are as before (as is the constraint applied to the fixed factor $\tau_A = 0$), however, the variance model now has two independent components:

$\delta_i \sim i.i.d.\ N\left(0, \sigma_\delta^2\right)$ and $\varepsilon_{ij} \sim i.i.d.\ N\left(0, \sigma_\varepsilon^2\right)$.

Using the lme() function from the nlme library to fit this model:

```
> fruitfly.lme <- lme(log(Longevity) ~ Thorax + Activity, random=~1|factor(ID))
> fruitfly.lme
Linear mixed-effects model fit by REML
  Data: NULL
  Log-restricted-likelihood: 19.27973
  Fixed: log(Longevity) ~ Thorax + Activity
(Intercept)      Thorax   ActivityB   ActivityC   ActivityD   ActivityE
 1.82122873  2.74894752  0.05203020 -0.12391127  0.08400508 -0.41826248

Random effects:
 Formula: ~1 | factor(ID)
        (Intercept) Residual
StdDev: 6.486293e-06 0.192596

Number of Observations: 125
Number of Groups: 25
```

Note that using the other function discussed in lectures (the lmer() *function from the* lme4 *library) produces slightly different results – see the* R *appendix for details.*

(i) Present and examine the summary output (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. **(3 marks)**

```
> anova(fruitfly.lme)
            numDF denDF  F-value p-value
(Intercept)     1    95 53831.34  <.0001
Thorax          1    95   173.23  <.0001
Activity        4    95    27.97  <.0001
>
> summary(fruitfly.lme)
Linear mixed-effects model fit by REML
 Data: NULL
        AIC        BIC   logLik
  -22.55945 -0.3264626 19.27973
```

Question 1, part (e) continued

```
Random effects:
 Formula: ~1 | factor(ID)
         (Intercept) Residual
StdDev: 6.486293e-06 0.192596

Fixed effects: log(Longevity) ~ Thorax + Activity
                 Value  Std.Error DF   t-value p-value
(Intercept)  1.8212287 0.19441704 95  9.367640  0.0000
Thorax       2.7489475 0.22794617 95 12.059635  0.0000
ActivityB    0.0520302 0.05452594 95  0.954228  0.3424
ActivityC   -0.1239113 0.05447560 95 -2.274619  0.0252
ActivityD    0.0840051 0.05491336 95  1.529775  0.1294
ActivityE   -0.4182625 0.05508900 95 -7.592486  0.0000
 Correlation:
          (Intr) Thorax ActvtB ActvtC ActvtD
Thorax    -0.980
ActivityB -0.183  0.043
ActivityC -0.134 -0.007  0.499
ActivityD -0.263  0.126  0.501  0.495
ActivityE -0.285  0.149  0.500  0.493  0.509

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.71074813 -0.69873756 -0.04147324  0.71686540  2.03709352

Number of Observations: 125
Number of Groups: 25
```

There has been almost NO real change from the model in part (g), the variance component associated with the additional random effects term is almost 0 and the residual standard error is unchanged (to 4 decimal places) at 0.1926.

The intraclass correlation coefficient is calculated as follows:

$$\frac{\hat{\sigma}_\delta^2}{\hat{\sigma}_\delta^2+\hat{\sigma}_\varepsilon^2}=\frac{\left(6.486293\times10^{-6}\right)^2}{\left(6.486293\times10^{-6}\right)^2+\left(0.192596\right)^2}=\frac{4.2072\times10^{-11}}{4.2072\times10^{-11}+0.03709322}$$

$$=\frac{4.2072\times10^{-11}}{4.2072\times10^{-11}+0.03709322}=1.134223\times10^{-9}\approx0$$

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

At this stage, many students will realise that given the strong relationship between them, it is not a good idea to include both ID *and* Thorax *as explanatory variables in the same model. The sensible choice is to either: include* Thorax *as a continuous covariate in an ANCOVA model – the model fit in part (b) and described in part (d); or include* ID *as a random blocking effect, but exclude* Thorax*. If you exclude* Thorax *from the model, but include* ID *as a random effect (or as a fixed effect, which will give similar results), as part of a model for a balanced randomised block design, then you will get an intraclass correlation coefficient of just over* 61%*, suggesting that* 61% *of the variation in* log(Longevity) *is due to differences in* Thorax *length – see the* R *appendix for details.*

Given a choice between the two variables (and the two models), I would prefer to use Thorax *as a continuous covariate in an ANCOVA model, as* ID *is only a categorical summary of the more detailed information available in* Thorax*.*