Question 1 - Report

Introduction

A statistical analysis on a real life dataset was undertaken with the aim to find a model which enables prediction of the dependent variable (vitamin D) given some explanatory variables. The data is available for 30 people and contains information about: vitamin D level (ng/ml), body mass index (BMI) , average time spent outside and average time spent exercising (both in hours per week). There is also a categorical variable (supply) indicating if the person takes a vitamin D supplement, this splits the patients into two groups of 15.

Exploratory Analysis

Exploratory data analysis is an essential step for model formulation, where one can familiarise with the data, visualise and inspect values. In this spirit, the entire data was first plotted using scatterplot (pairs command) as well as histograms of individual variables, to visualise their distribution (refer to Figures 1 and 2 at the end of the section). A summary of the data statistics was also produced and can be seen below. The associated standard deviations are approx.: 87.7 for vit.D, 3.1 for BMI, 2.8 for outside and 1.4 for exercise.

```
      vitd             bmi           outside          suppl        exercise
 Min.   :  0.0   Min.   :18.90   Min.   : 1.200   Min.   :0.0   Min.   :0.190
 1st Qu.:154.5   1st Qu.:22.68   1st Qu.: 6.175   1st Qu.:0.0   1st Qu.:3.305
 Median :199.5   Median :24.60   Median : 7.000   Median :0.5   Median :4.045
 Mean   :212.7   Mean   :24.81   Mean   : 7.760   Mean   :0.5   Mean   :3.758
 3rd Qu.:264.6   3rd Qu.:26.70   3rd Qu.: 9.625   3rd Qu.:1.0   3rd Qu.:4.710
 Max.   :396.4   Max.   :32.60   Max.   :13.500   Max.   :1.0   Max.   :6.140
```

From the above information and associated figures it can be observed that vit.D levels range between 0 and ~ 400ng/ml, they are concentrated below the mean and speared out above the mean. BMI ranges between ~19 (underweight) and ~33 (obese), most values are concerted in the centre of the distribution but not at the mean/median, some extreme values can also be seen. This variable could potentially be transformed to better resemble a Gaussian distribution. The 'Outside' variable has large variability; many extremely high and low values indicating some people spend virtually no time outdoors whereas others spend a lot, this could be driven by factors such as their occupation. Unsurprisingly, as vitamin D is naturally produced when skin is exposed to sunlight, a strong linear relationship exists between the amount of vitamin D and time spent outdoors. The exercise variable is concentrated around the median, with relatively small quantile range, indicating that on aggregate most people exercise similar amounts of time. Supply is a control variable, which could affect the remaining explanatory variables i.e. people with and without treatment may exhibit different distributions of BMI, outside, exercise and of course, also vit.D levels. To better understand the relationship between supply and the remaining variables, data was split by supply and plots were made to visualise the difference (Figures 3 and 4).

From Figures 3 and 4 its can be seen that those supplied with vit.D have in general higher levels of vit.D (higher mean and interquartile range). A possible anomaly exists in the group with no treatment, where one patient has exceptionally high levels of vit.D. The BMI of treated group is generally better than that of untreated (less variability and lower BMI values). A slight negative linear relationship can be noticed in the BMI variable. Taking a log transformation of bmi produces a distribution closer to the Gaussian, here this may be a viable transform (Figure 1). There may be significant interaction (impact) between supply and bmi. Exercise seams to follow no pattern, indicating this variable may not be significant in the model.

From box plots (Figure 3) the treated group spends more time outdoors and exercises more then the untreated group. Thus, the treated group seams to be living a healthier lifestyle. This rises questions about how the treatment group was selected and data was collected. Was vitamin D administered to patients diagnosed with its lack? or on the contrary, is it health-conscious patients, aware of benefits of taking vitamin D supplements, that were used to make up the treatment group? Looking at exercise, time spent outside and BMI it is reasonable to claim those treated are health-conscious individuals, possibly skewing the analysis. Thus, if this analysis were to be taken further, with the aim to generalise results to a broad population, a closer investigation of the data collection process should be undertaken.

Model Fitting and Selection

Findings from the exploratory analysis section were used to find the best model to predict vitamin D levels basis the explanatory variables. The best model is that which achieves good vitamin.D predictions whilst being as simple as possible. In order to find the best model, 3 variations of a simple model were examined, the best one was selected and then backward elimination was used to 'cut-down' the least significant elements of the model until satisfactory results were reached. Changes were made to one element of the model at a time so as to keep track of their effect on the results. The first 3 models are visualised below:

```
> model1 <- lm( vitd ~ bmi + outside + supply + exercise)
> model2 <- lm( vitd ~ log(bmi) + outside + supply + exercise)
> model3 <- lm( vitd ~ I(log(bmi))+supply+I(log(bmi)):supply + outside + exercise )
```

Model 1 is the simplest additive linear model including all exploratory variables. The summary of this model reveals that the constant intercept and variables 'exercise', 'bmi' are not statistical significant (p_value>5%). Model 2 explores the possibility of taking the log transform of the bmi variable, which reveals log(bmi) is still not significant. Various other transformations (log, square, square root) were also tested at this stage, however the thus resulting models were not satisfactory. From exploratory analysis it was also noted that the interaction of supply and BMI may be significant so this is explored in Model 3. This achieves a slightly improved R^2 metric as compared to the previous models. It can be observed that exercise is least significant, thus, applying backward elimination in Model 4 this variable is removed. Model 4 summary and annova (looking at F test value) indicate that supply is the next least significant variable and it is therefore removed, whilst keeping the interaction term (log(bmi): supply)(Model 5). Further from this, log(bmi) is insignificant and is also removed in Model 6. This highly simplified model yields good results with a na improved R^2. The intercept term is not statistically significant but will be kept in the model for interpretability. The final model (Model 6) formulation, summary and analysis of variance outputs can be seen below:

```
> model6<- lm( vitd ~ I(log(bmi)):supply + outside )
> summary(model6)

Call:
lm(formula = vitd ~ I(log(bmi)):supply + outside)

Residuals:
     Min       1Q   Median       3Q      Max
-167.717   -8.295    4.359   14.377   40.350

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -17.797     20.306  -0.876   0.3885
outside              27.689      2.363  11.718 4.27e-12 ***
I(log(bmi)):supply    9.813      4.115   2.385   0.0244 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.95 on 27 degrees of freedom
Multiple R-squared:  0.8435,    Adjusted R-squared:  0.832
F-statistic: 72.79 on 2 and 27 DF,  p-value: 1.331e-11
```

```
Analysis of Variance Table

Model 1: vitd ~ I(log(bmi)):supply + outside
Model 2: vitd ~ I(log(bmi)):supply + I(log(bmi)) + outside
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     27 34900
2     26 31586  1    3314.5 2.7284 0.1106
```

Model 6 is a satisfactory fit, this takes the interaction term of log(bmi) and supply together with the outside variable. The exercise variable is eliminated completely, perhaps it is not significant also due to the fact that it is indirectly linked to BMI (those who exercise tend to have lower bmi). Model 6 is simple yet it effectively predicts vitamin D levels. Goodness of fit analysis was performed on this model and results are visualised in Figure 5.

Refinement: Final Model

The goodies of fit analysis on Model 6 points to 3 potential outlier data instances (rows: 5, 15, 22). Patient 22 has the largest absolute residual deviance from the remaining patients, and a Cook's distance 3 times the magnitude of the other outlier points. Further inspection reveals that this patient has zero levels of vitamin D, be it by measurement error or severe illness. This is not a usual observation and its origination should be further examined. For the purpose of this exercise, it is reasonable to consider this measurements as erroneous and exclude it from the model. Thus,

Model6 was re-fitted to a corrected dataset (excl. point 22). The summary of results is presented below and the accompanying goodness of fit analysis can be seen in Figure 6. Unsurprisingly, the model improves upon the removal of the outlier, with significantly better ..... .

The residual vs fitted values, Cook's distance and QQ plots all point to a good fit. With the exception of points 5 and 15, in the QQ plot, data points at the extremes of the quantile range do not tail off in a predominant direction but rather stay aligned along the line, which is indicative of a good fit. There is evidence not to exclude points 5 and 15 as after re-fitting the model this points do not significantly affect the QQ plot, similar to the way point 22 did.

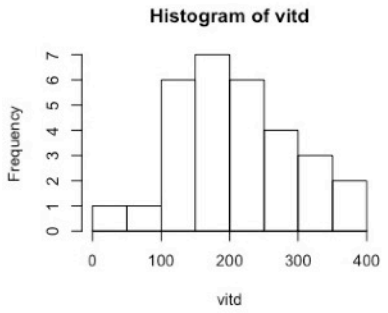
```
Residuals:
    Min      1Q  Median      3Q     Max
-25.129  -8.551  -1.169   6.442  32.647

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.1580     7.6655  -0.021 0.983714
outside            26.9526     0.8798  30.636  < 2e-16 ***
I(log(bmi)):suppl   6.1043     1.5554   3.925 0.000569 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.36 on 26 degrees of freedom
Multiple R-squared:  0.9737,    Adjusted R-squared:  0.9717
F-statistic: 480.9 on 2 and 26 DF,  p-value: < 2.2e-16
```

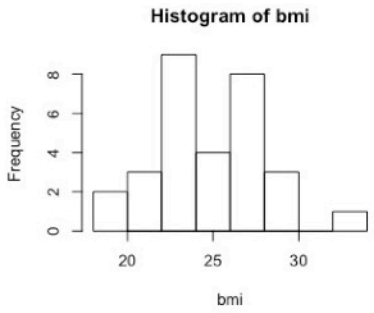
Conclusion

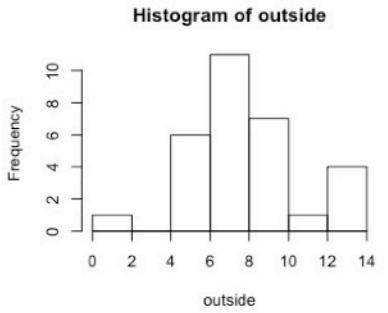
After undertaking a detailed exploratory analysis and careful model fitting a satisfactory model was found to model the dependance of vitamin D levels on the exploratory variable (bmi, supply, time spent outside and exercising). It was found that two terms described vitamin D sufficiently well: the interaction of log(bmi) and supply term and an additional term capturing time spent outside. The formulation of this model is printed below.

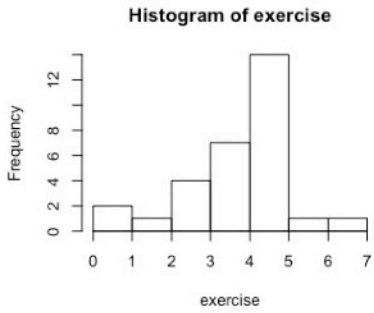
$$\alpha + \beta_1(outside) + \beta_2(log(bmi) * supply)) = vit.D$$

It should be however note that in order to generalise results, and for the purpose of further epidemiological studies, certain questions about the data should be answered. Firstly, the inspection of the anomalous data point 22 should be undertaken. Secondly, further insights into the data collection process is desirable, as there is evidence that the patients taking vitamin D also live a healthier lifestyle and it is uncertain whether this could affect their behaviour when choosing to take a vitamin D supplement.









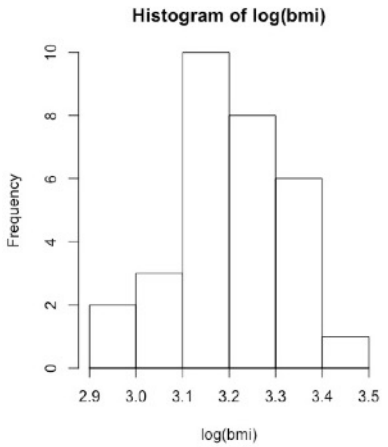


FIGURE 1. Histograms of dependent and exploratory variables (supply excluded). log(bmi) also included

vita denotes the dependent variable- level of vitamin D

bmi is the body mass index

outside is the average time spent outside and exercise is the average time spent exercising (both in hours per week

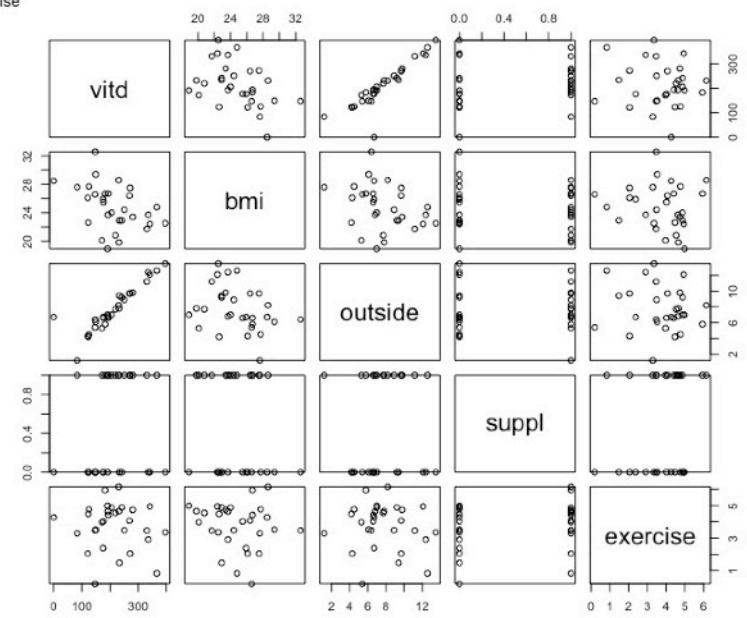
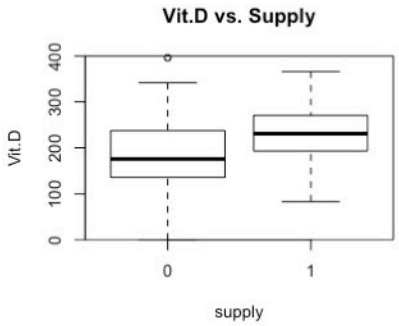
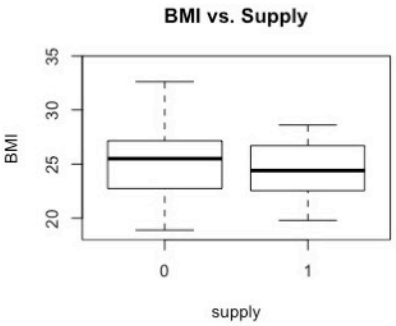
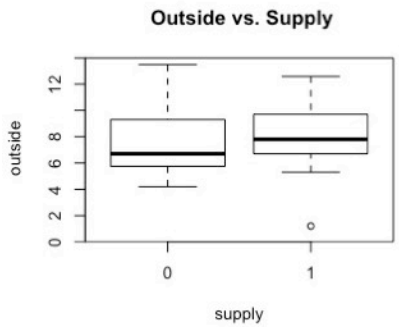


FIGURE 2. Scatter plot of all variables







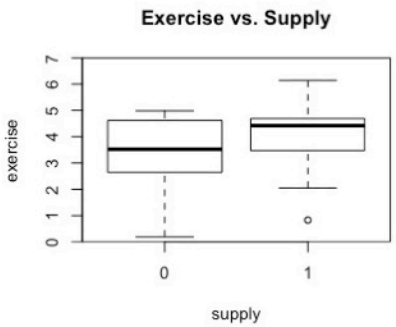
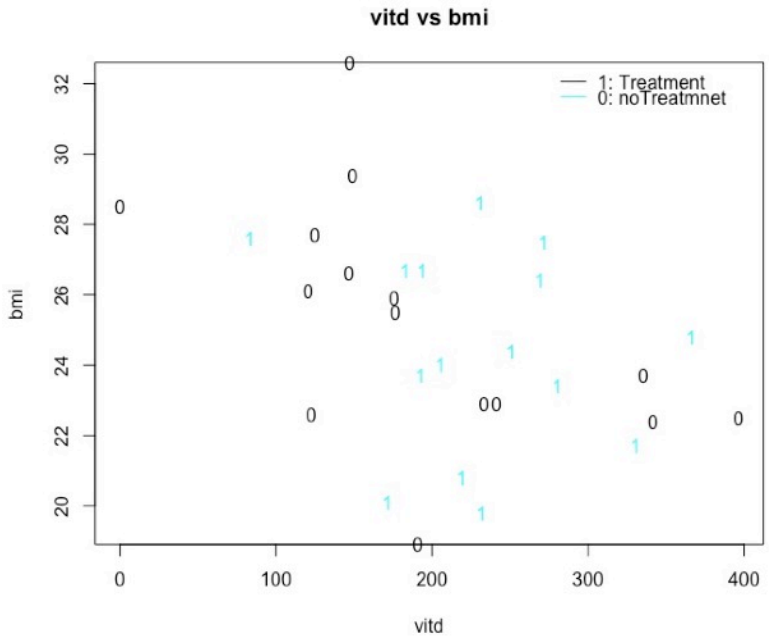
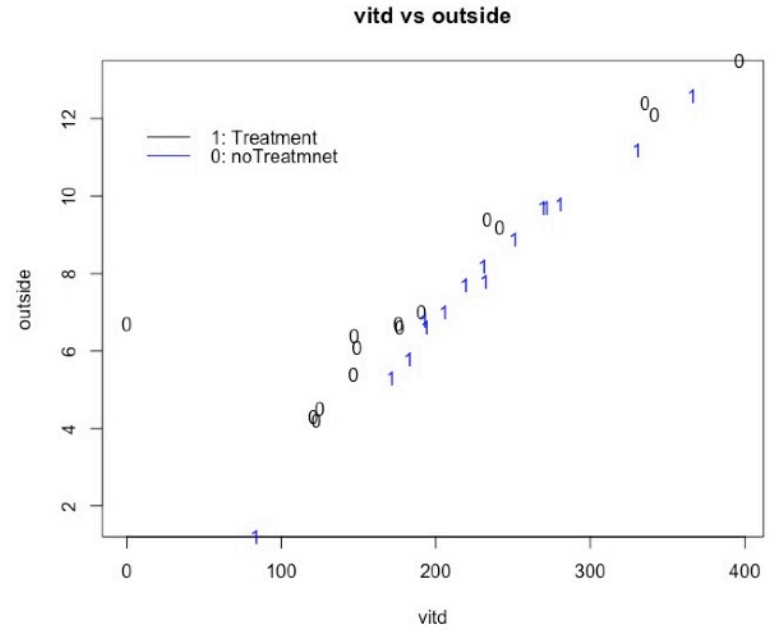


FIGURE 3. Box plots of dependent and exploratory variables split by the control variable (vitamin D. supply).





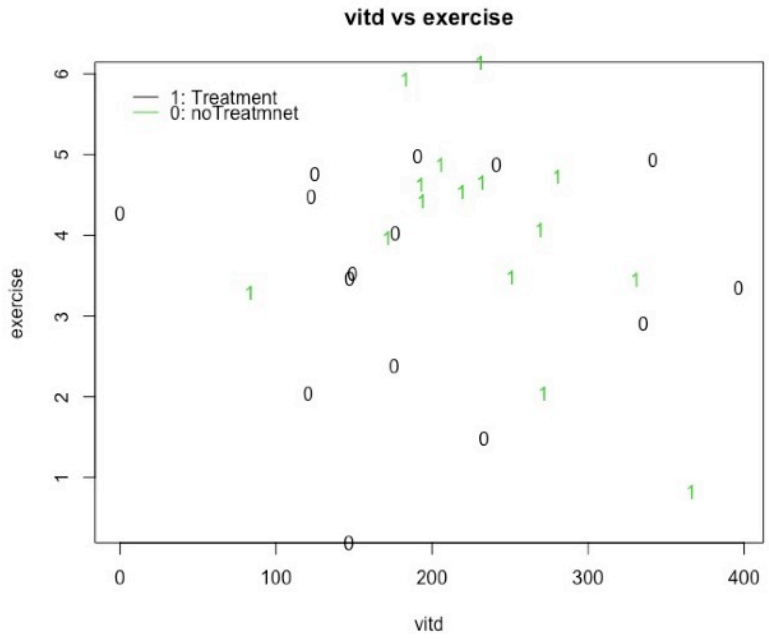


FIGURE 4. Scatter plots of each exploratory variable split by the control variable (vitamin D. supply).

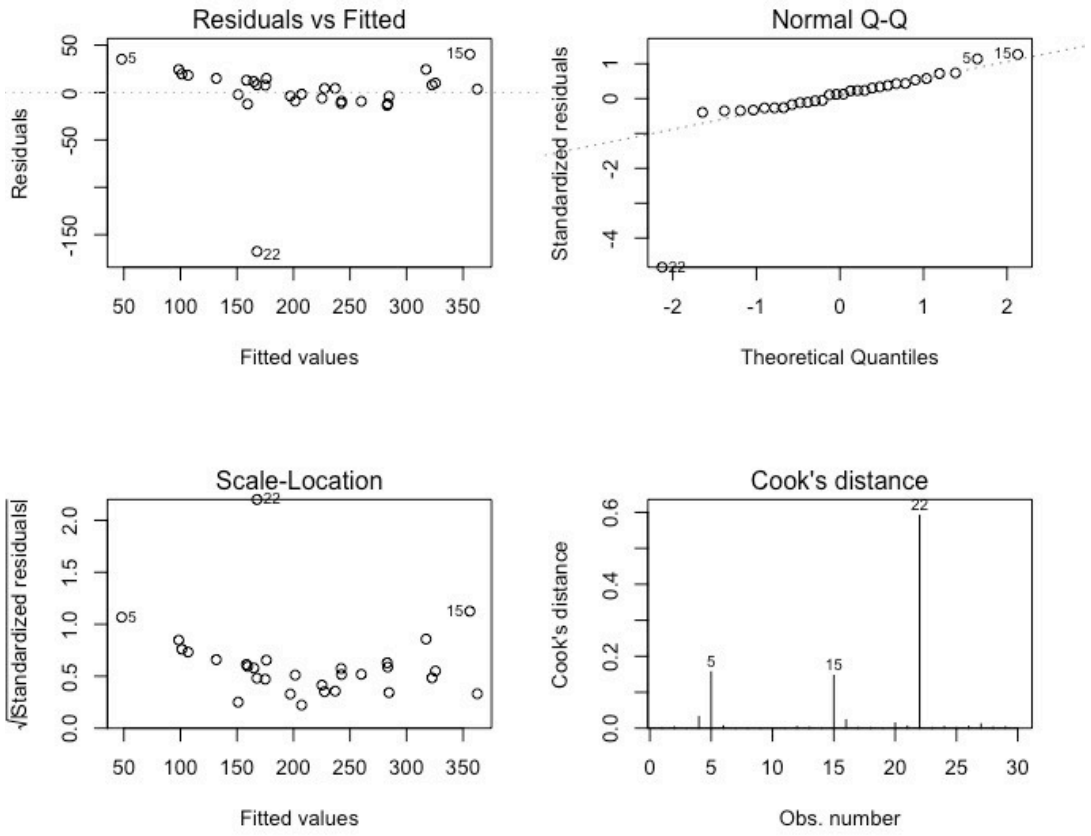


FIGURE 5. First fitted model Summary

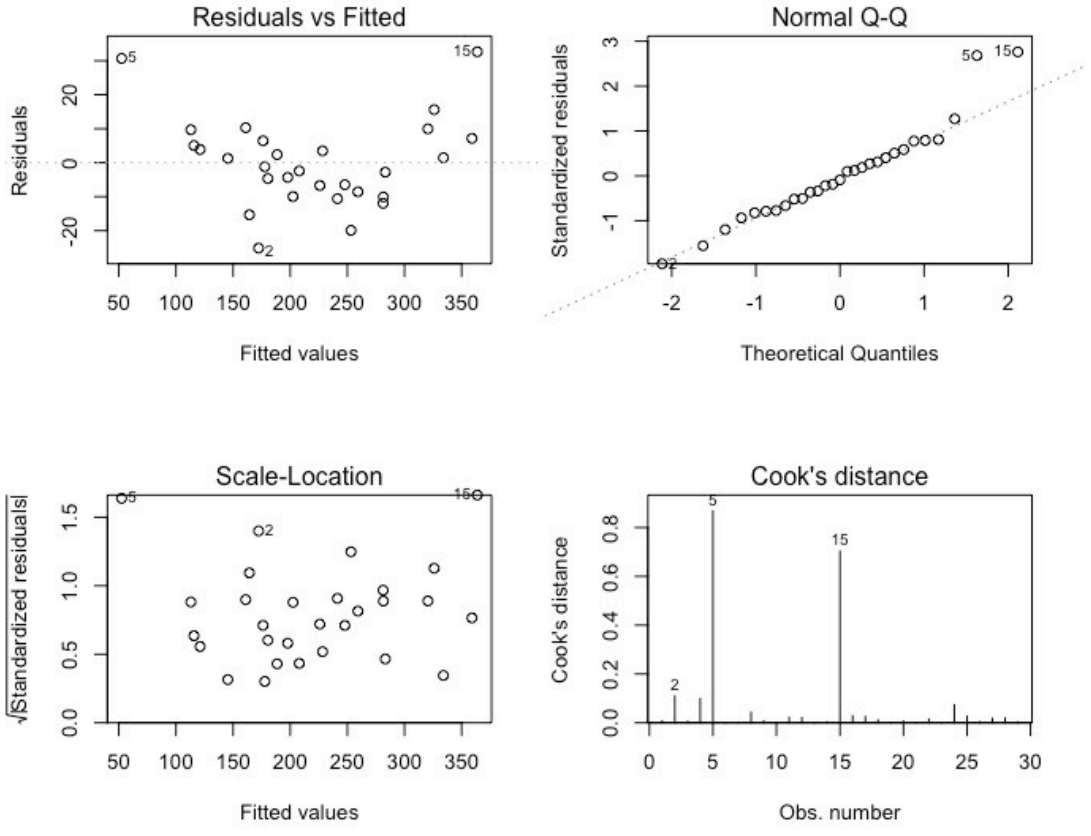


FIGURE 6. Final fitted model

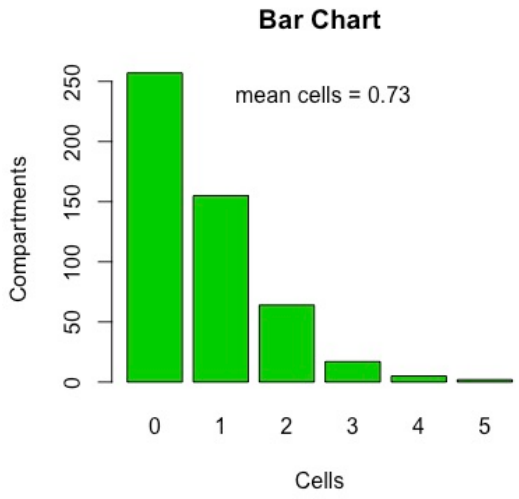
Question 2 - Summary

Problem description

In this question cells.data is load an visualised (Parts A and B). The user-defined function *negbinll* is then implemented to calculate the negative log likelihood (Part C). This in turn is used to calculate the maximum likelihood estimates (MLE) of the parameters mu and tau (Part E), with the corresponding standard errors (Part F).

R output:

```
Part A and B output:
There are  364  cells, containing a total of 500 compartments.
Thus the mean number of cells per compartment is approx.  0.73

Part C and D output:
Summary of log-likelihood for different parameters
       tau  mu Log.Lik
Case 1 0.1 0.1   909.8
Case 2 0.1 0.5   778.8
Case 3 0.5 0.5   624.9
Case 4 1.0 0.7   588.3
Case 5 1.3 1.0   595.1

Part E output:
The maximum likelihood estimates, which minimise the negative log likelihood, are:
Tau parameter:  3.9868  and mu parameter:  0.728

Part F output:
The associated standard error on Tau is approx. 1.807 and on Mu approx. 0.0415
```



Results of nlm() :

```
$minimum
[1] 576.448

$estimate
[1] 3.9868459 0.7279995

$gradient
[1] -3.632873e-05 -3.751666e-06

$hessian
             [,1]         [,2]
[1,] 0.306246820 1.112104e-03
[2,] 0.001112104 5.805820e+02

$code
[1] 1

$iterations
[1] 17
```

Commentary:

Working on part E - the optimisation problem, the output of nlm() allows for the further exploration of the optimisation results. The minimum value of negbinll is found at approx. 577, and since this function defines the negative log likelihood it is equivalent to finding the maximum likelihood. $estimate gives the estimated tau and mu parameters, whereas $gradient gives the slope of the function at the $minimum point. $hessian is a matrix of second order partial derivatives of the function, which was useful to compute the standard error of the estimates in part F. $code =1 indicates that the iteration has stopped because the gradient reach a value very close to zero. Exploration of different initialisation parameters encouraged in part D prompted the discovery that only sensibly initialised parameters converge to the required minimum.