Major Part A: - 19msms26

Question 1:-

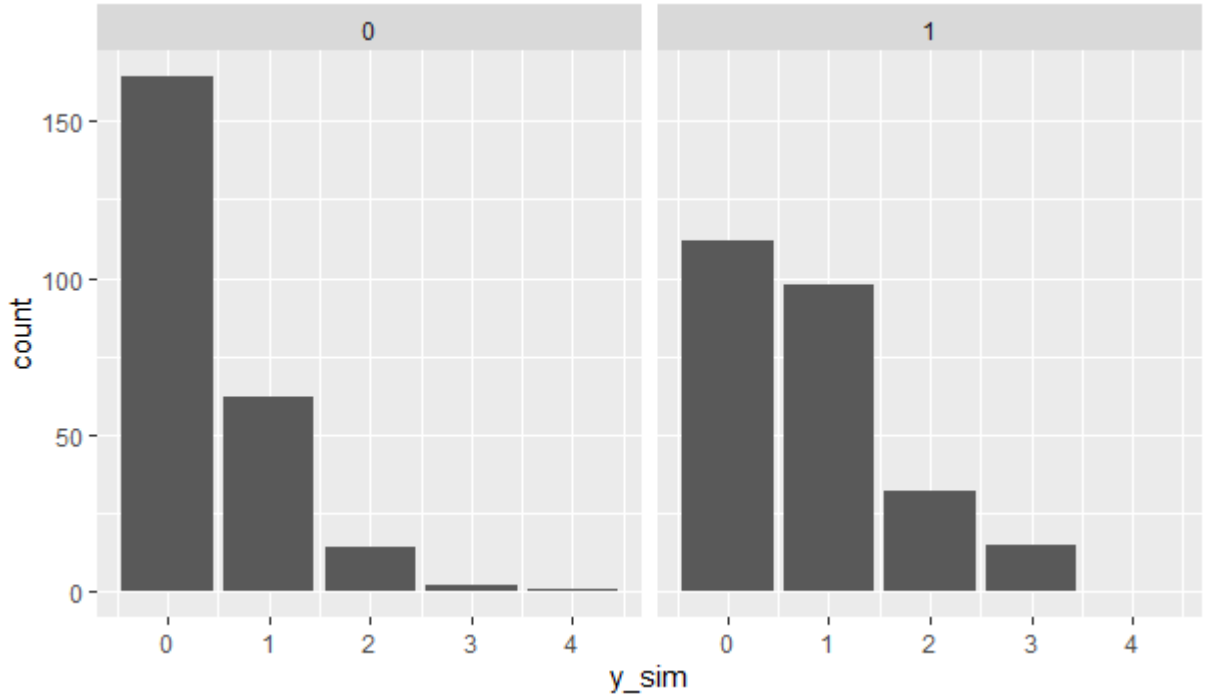
We generated a simple model that generates different counts on the basis of you smoke or not. So for That I have simulated a variable called smoke having 0 and 1, where 1 denotes yes and 0 as not smoking. Using this variable I have simulated my response variable from Poisson distribution having 500 Observations.

My model for generating the response is $\mathbf{log}(\boldsymbol{\mu}) = -\mathbf{1} + \mathbf{0.7} * \boldsymbol{I}(\boldsymbol{smoke} == \mathbf{1})$

True values of parameters are for beta0= -1 and for beta1= 0.7.

Now we will fit a glm model on this simulated data and examine the estimated parameters and compare It with true values but before that we will explore the simulated data. There is a frequency table of our Count generated using the above model with respect to the smoke variable.

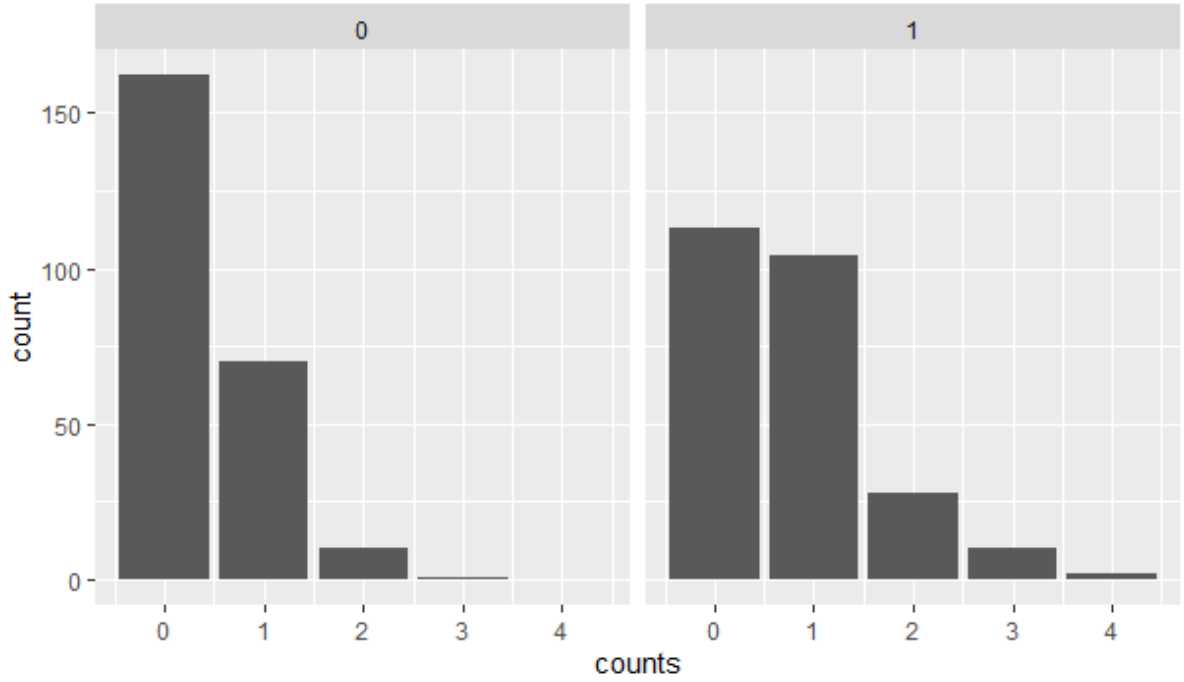
y_sim	smoke 0	smoke 1
0	164	112
1	62	98
2	14	32
3	2	15
4	1	0



The above bar plot is just the graphical version of the table we showed above. So notice that there are 2, 3 and 4 values present as observation in our response created with irrespective of both the level of smoke variable. We will fit the Poisson glm on the simulated data. Although we should check the condition that the mean of our response is same as variance before fitting the Poisson glm, if we do not know how the response is generated. But in this case mean is conditional depending on you smoke the exp (0.7) and if you don't then exp (-1).But since we know how data is generated. So we will proceed and fit the Poisson glm.

```
Call:
glm(formula = y_sim ~ smoke, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2692  -0.9072  -0.9072   0.7738   3.3191

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.88789    0.09999  -8.880  < 2e-16 ***
smoke        0.67153    0.12177   5.515  3.5e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 550.17  on 499  degrees of freedom
Residual deviance: 517.84  on 498  degrees of freedom
AIC: 1016.2

Number of Fisher Scoring iterations: 5
```

From the summary output of the model fitted we can see that the estimates of beta0 is -0.88 and beta1=0.67.Both the estimates of parameters beta0 and beta1 are not too far from true values which were -1 and 0.7 for beta0 and beta1 respectively. And both the estimates are come out as significant. The interpretation of beta 1 is that the expected count is about 1.95 times greater for smokers than non-smokers.
Now we try to generate the counts using the model we have fitted and compare it with the original data. Basically a count model returns the expected count. It will work as lambda in a Poisson model.



So this bar plot does look similar as bar plot of our original data.

Question -2:-

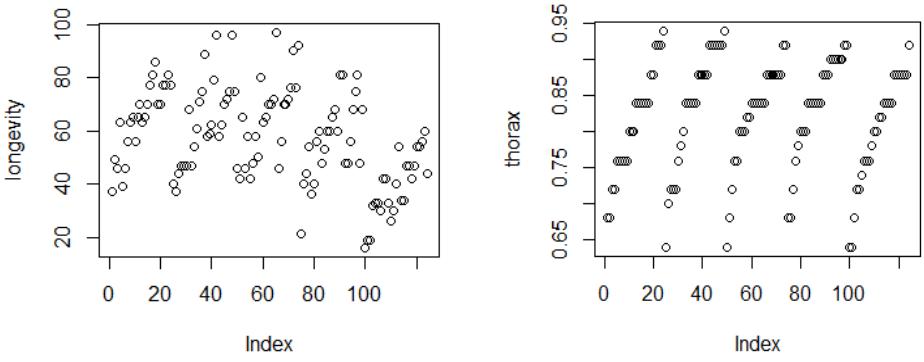
My dataset is fruitfly data.

Description: - Female Drosophila mate only once. In the experiment reported, male flies were caged with different numbers of females. The females were either mated (or pregnant) or virgin. There were 5 groups of 25 male flies each.

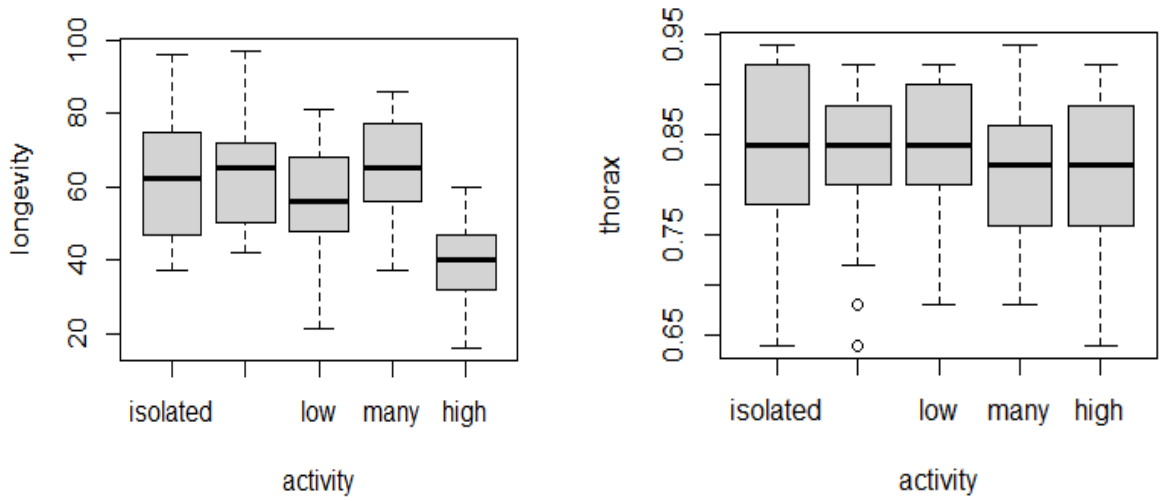
- In the isolated group, each male was kept solitary.
- In the one group, each male was kept with 1 pregnant female.
- In the many group, each male was kept with 8 pregnant females.
- In the low group, each male was kept with 1 virgin female.
- In the high group, each male was kept with 8 virgin females.

Variable Observed	Description
`Thorax(numeric)`	thorax length, ratio scale (mm)
`Longevity(numeric)`	lifetime, ratio scale (d)
`Activity(factor)`	the treatment group, nominal scale

Exploratory data analysis:-Firstly we will plot the scatter plots of thorax and longevity to understand whether there is some pattern visible which might help us in modeling.

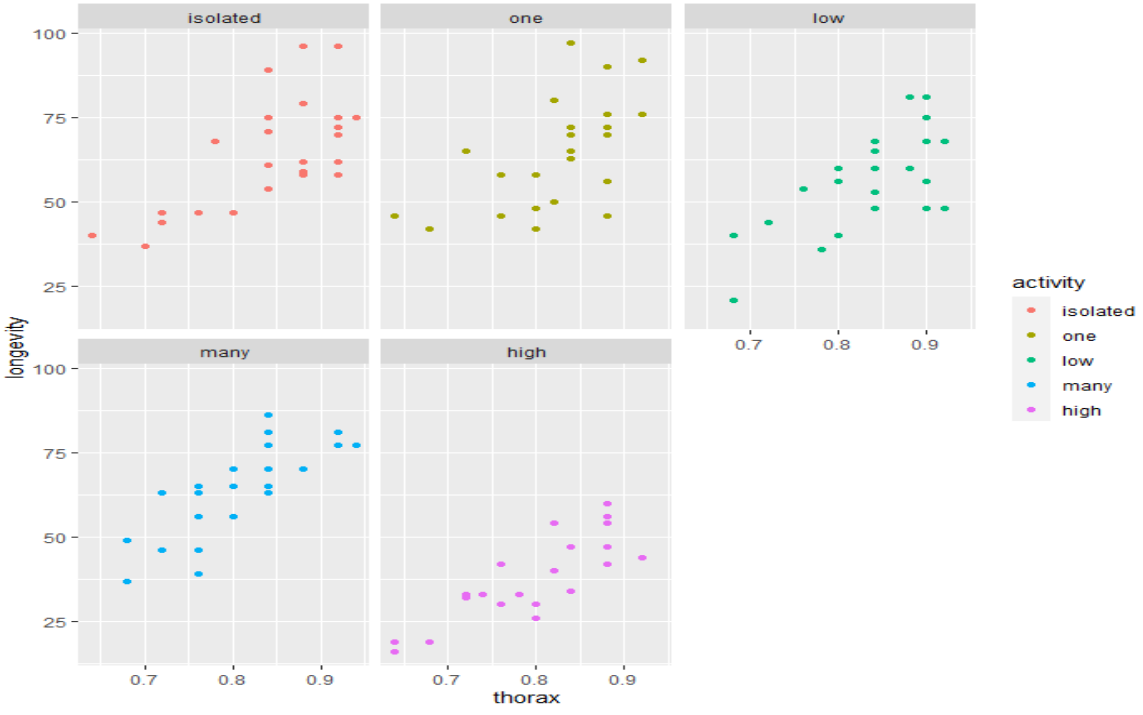


From the scatter plot of longevity we observed there is no visible pattern. But in the scatter plot of thorax we did not observe any recognizable pattern but there seems some kind of grouping and we already know there are five levels of activity variable so there is a linear increase in thorax variable is visible for each of the five groups. We will try to plot boxplot of both the variables to get better understanding of the data.



So from the boxplot of longevity we observed that the mean longevity for treatment groups many and one are highest and lowest for the high group and for high group the boxplot looks symmetrical, where as for thorax mean of isolated group ,one and low is approximately same and is slightly less for groups many and high. We also noticed that there are two observations in for one treatment group of thorax. And boxplots of groups high and one seems symmetrical for thorax.

Now we are interested in the relationship between thorax and longevity for each treatment group through the scatter plot.



So from the above scatter plots we observed that longevity and thorax has an increasing linear relationship.
Below is the basic summary table of the data. We observed that mean thorax length is 0.8224 and mean longevity is 57.62.

thorax	longevity	activity
Min. :0.6400	Min. :16.00	isolated:25
1st Qu.:0.7600	1st Qu.:46.00	one :25
Median :0.8400	Median :58.00	low :25
Mean :0.8224	Mean :57.62	many :24
3rd Qu.:0.8800	3rd Qu.:70.00	high :25
Max. :0.9400	Max. :97.00	

Model fitting and Residual Diagnostics:-

First we will try to fit linear models. Here our variable of interest is longevity which will be our response and the rest will be the regressors.So we will fit two linear models and compare them. In the first model we considered thorax as our only regressor.and in the second model we will add activity as our second repressor. So our population models are-

Model1:- longevity = a+b*thorax+e

Model2:-longevity =a1+b1*thorax+c1*activity+e1

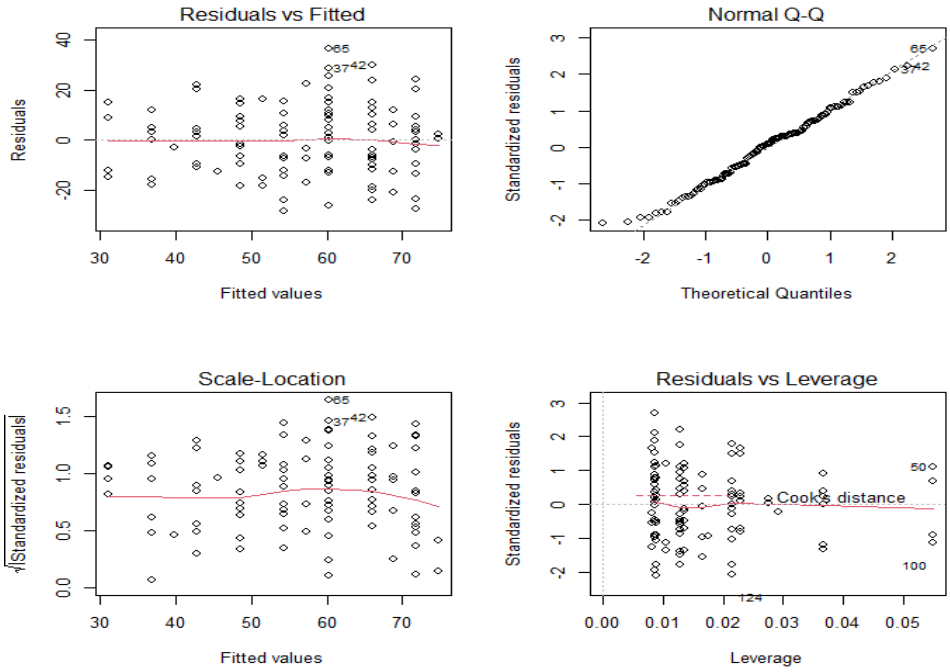
The summary output of model 1 is as follows-

```
Call:
lm(formula = longevity ~ thorax, data = fruitfly)

Residuals:
    Min      1Q  Median      3Q     Max
-28.364  -9.986   1.258   9.264  36.825

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -61.86      13.37  -4.625 9.39e-06 ***
thorax        145.28      16.19   8.971 4.27e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.65 on 122 degrees of freedom
Multiple R-squared:  0.3975,    Adjusted R-squared:  0.3926
F-statistic: 80.49 on 1 and 122 DF,  p-value: 4.275e-15
```

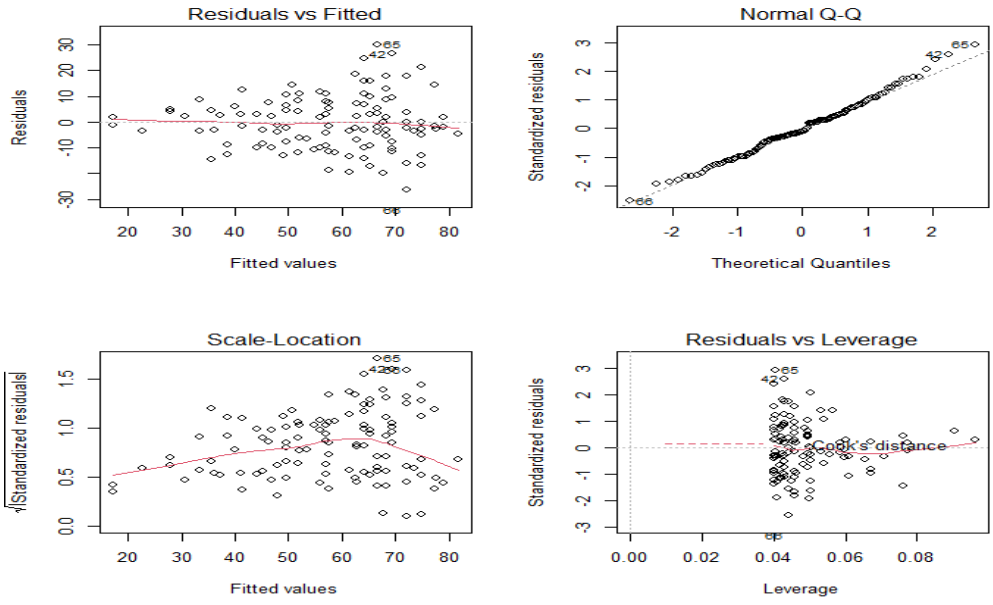
Hence from this output we observed that our intercept and thorax both are significant in the model. Adjusted R-squared is 0.39 which mean our model captured only 39 % variability of the data. And p value is less than 0.05 hence model is significant and estimate of mse is 0.111.
Let’s see the summary output of the model 2-

```
Call:
lm(formula = longevity ~ thorax + activity, data = fruitfly)

Residuals:
    Min      1Q  Median      3Q     Max
-26.108  -7.014  -1.101   6.234  30.265

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -48.749     10.850  -4.493 1.65e-05 ***
thorax        134.341     12.731  10.552  < 2e-16 ***
activityone     2.637      2.984   0.884   0.3786
activitylow    -7.015      2.981  -2.353   0.0203 *
activitymany    4.139      3.027   1.367   0.1741
activityhigh  -20.004      3.016  -6.632 1.05e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.54 on 118 degrees of freedom
Multiple R-squared:  0.6527,    Adjusted R-squared:  0.638
F-statistic: 44.36 on 5 and 118 DF,  p-value: < 2.2e-16
```

From the summary output of model 2 we observed that intercept, thorax, activity high and activity low were significant. And adjusted R-squared is 0.64 which is more than model 1's adjusted R-squared. Model is significant since the p-value is less than 0.05 and estimate of mse here is 0.09.
Here we concluded that "low" survives about 7 days less and the high sexual activity flies "high" survive about 20 days less than the reference group.



model 1

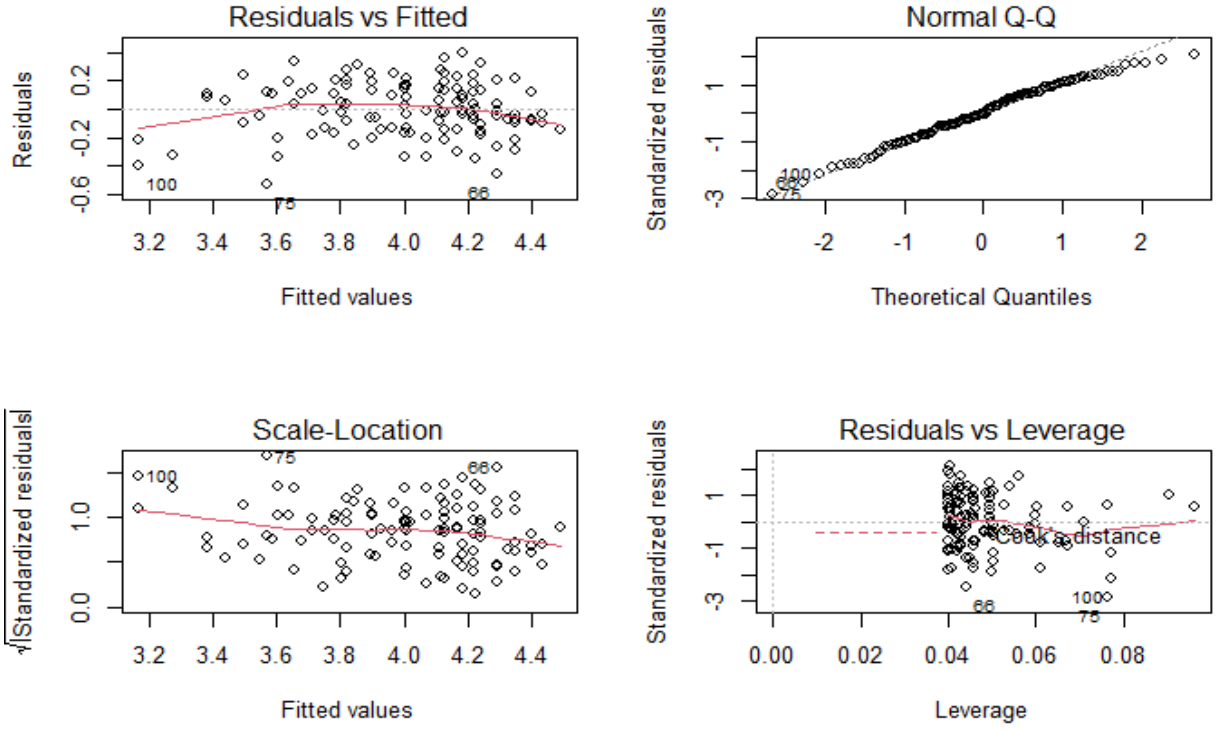


model 2

So from both residual vs fiited and scale –location plots of model1 we observed that the concentration of points are not evenly spread which is sign of voilation of homoscedasticity assumption. Residual seems to follow normality assumption.From the residual vs leverage plot we observed that there are few observation which are having high leverage if these are potential outlier than our estimates can vary significantly once we remove these observations from data.
From the model 2 plots we observed that there is still the sign of voilation of homoscedasticity assumption but the spread of the observations has become dense compare to model1.Residuals still seems to follow normality assumption.In the residual vs leverage plot the observations has become less spreat and more dense.
We will use the log transformation to see solve the problem of heteroscedasticity.
Model3:- log(longevity) =a2+b2*thorax+c2*activity+e2

```
Call:
lm(formula = log(longevity) ~ activity + thorax)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52641 -0.13629 -0.00823  0.13918  0.39273

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.84421    0.19882   9.276 1.04e-15 ***
activityone   0.05174    0.05468   0.946   0.3459
activitylow  -0.12387    0.05463  -2.268   0.0252 *
activitymany  0.08791    0.05546   1.585   0.1156
activityhigh -0.41925    0.05527  -7.586 8.35e-12 ***
thorax        2.72146    0.23329  11.666  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1931 on 118 degrees of freedom
Multiple R-squared:  0.7025,    Adjusted R-squared:  0.6899
F-statistic: 55.72 on 5 and 118 DF,  p-value: < 2.2e-16
```

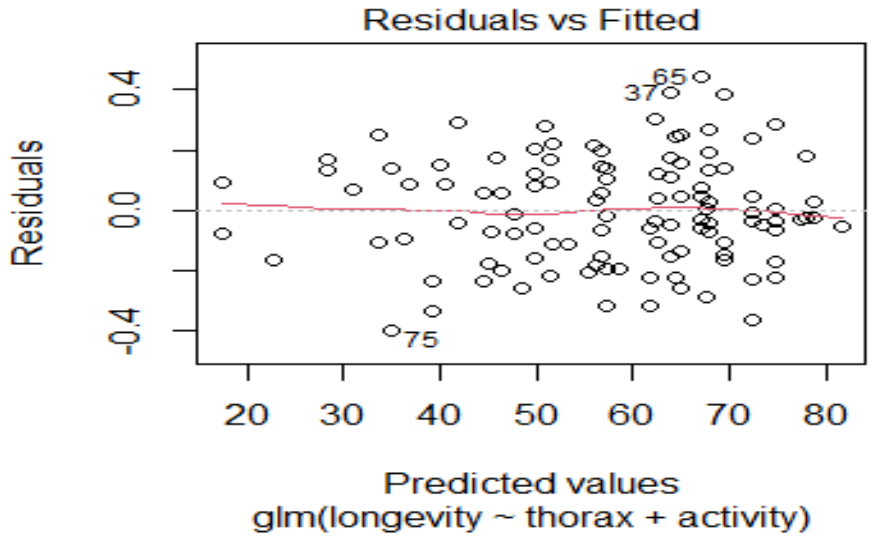
This model has adjusted R-squared 0.69 which is highest among all the models we have fitted till now with the lowest estimate of mse i.e. 0.001636441.Model is also significant.



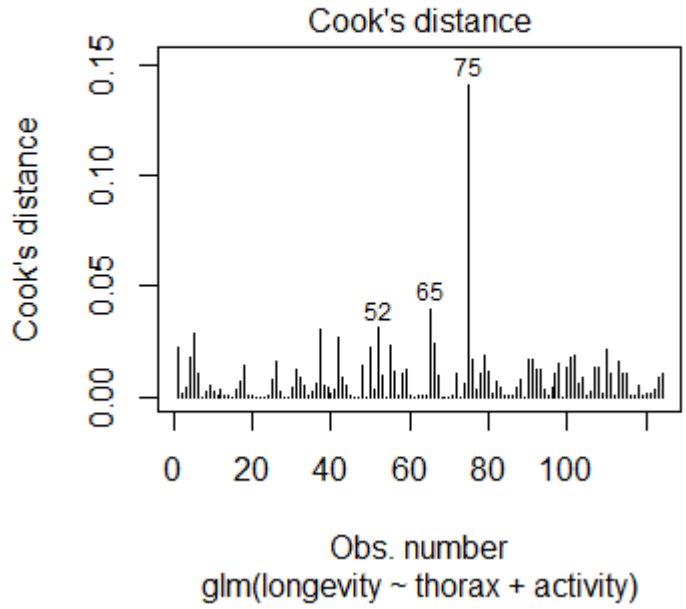
This model seems to do better in terms of following all the assumptions.But one thing we did not keep in mind that our response was not following normal distribution so a linear model will not be an appropriate model. So we will move to generalized linear model and try to fit a Gamma glm after trying all the possible link functions, we found the Gamma model with identity link function performed best in terms of Aic and Residual deviance. Here is the summary output of this model

```
Call:
glm(formula = longevity ~ thorax + activity, family = Gamma(identity))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.47212  -0.12606  -0.01736   0.12734   0.39302

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -49.935      7.886  -6.332 4.55e-09 ***
thorax        135.612      9.500  14.275  < 2e-16 ***
activityone     3.097      3.112   0.995   0.3217
activitylow    -7.227      2.884  -2.506   0.0136 *
activitymany    3.996      3.128   1.277   0.2040
activityhigh  -19.461      2.520  -7.723 4.09e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.03240427)

    Null deviance: 13.2803  on 123  degrees of freedom
Residual deviance:  3.9415  on 118  degrees of freedom
AIC: 931

Number of Fisher Scoring iterations: 5
```

Here notice that the terms which were not significant earlier are still not significant and these are the treatment groups having average longevity greater than others. Thorax, intercept and activity for treatment group high are significant terms. The Aic is 942.29 and aic for the model 1, model 2 were 1004.145, 943.816 respectively which are higher than the aic of our glm model.



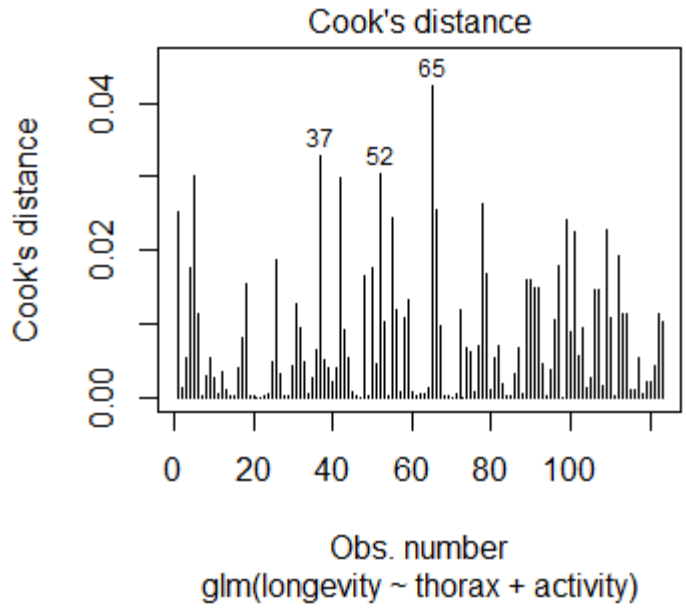
This residual plot seems better among all other residual vs fitted plots.



There seems one outlier which is observation number 75th which belongs to low treatment group having the lowest longevity value in the low treatment group. This was indeed influential observation, after removing this observation and re refitted the glm model.

```
Call:
glm(formula = longevity ~ thorax + activity, family = Gamma(identity),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.41831  -0.12216  -0.01593   0.11948   0.39647

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -46.268      7.934  -5.832 4.97e-08 ***
thorax        131.036      9.560  13.707  < 2e-16 ***
activityone     3.071      3.046   1.008   0.3154
activitylow    -5.734      2.930  -1.957   0.0527 .
activitymany    3.955      3.062   1.291   0.1991
activityhigh  -19.731      2.470  -7.987 1.07e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.03095144)

    Null deviance: 12.5294  on 122  degrees of freedom
Residual deviance:  3.6956  on 117  degrees of freedom
AIC: 918.55

Number of Fisher Scoring iterations: 4
```

Now from the summary output of this model we observed that the Null deviance, Residual deviance and Aic has decreased and lowest among all the previously fitted models. Number of Fisher scoring iterations are also decreased compared to previous glm.But the residual vs fitted plot is same as the plot of previous glm. But the plot of cooks distance has changed.



And if we consider 0.05 as our threshold then there does not seem any potential outlier.

Conclusion:-

So in this analysis we first explored the data from various aspects and gained some insights. Then we fitted 3 linear models and 2 Gamma glm and from the residual diagnostic and on the basis of AIC we concluded that our last glm model seems to perform best among all the models we have fitted so far.