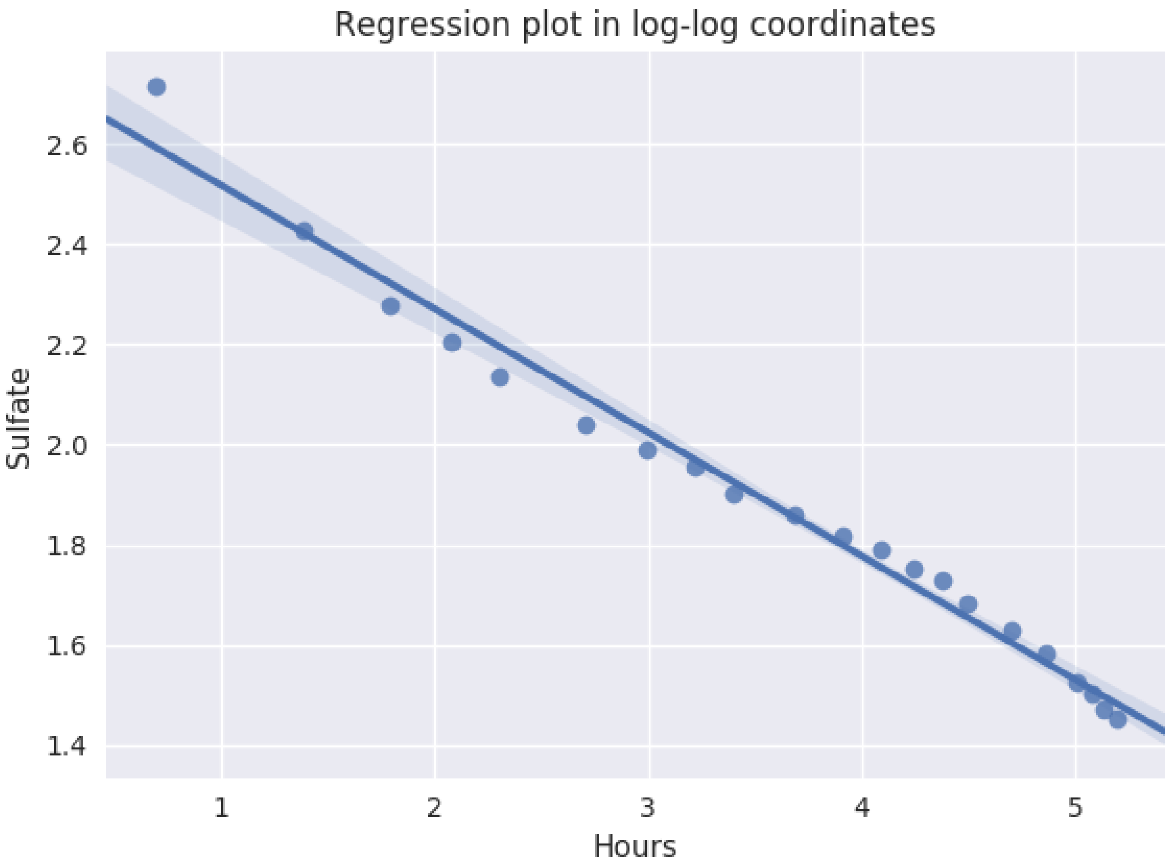
HOMEWORK 5

Problem 7.9:

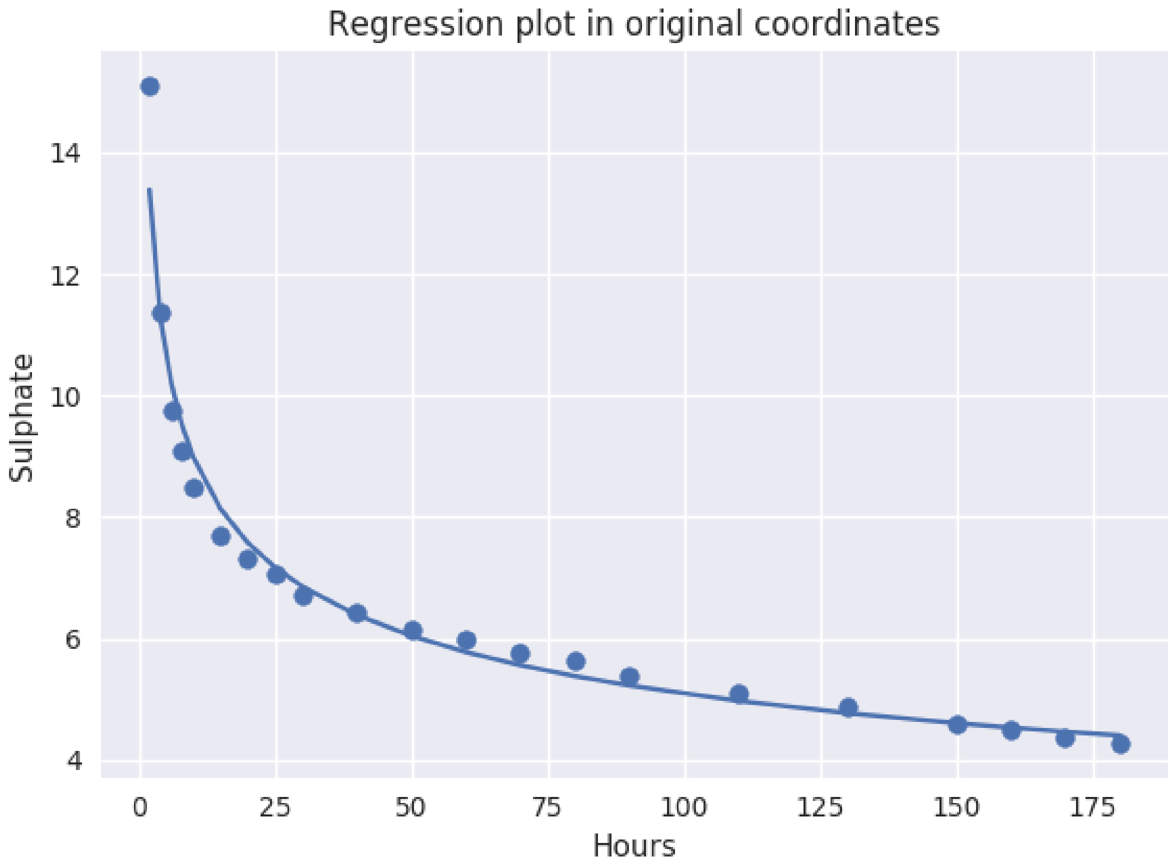
The dataset we used measured the concentration of sulphate in the blood of the baboon as a function of time. We took the log of the coordinates of this dataset and fit a linear regression model such that it takes no of hours as input and predicted the concentration of sulphate in blood.

a) Following is the plot that shows the data point and regression line in log-log coordinate:



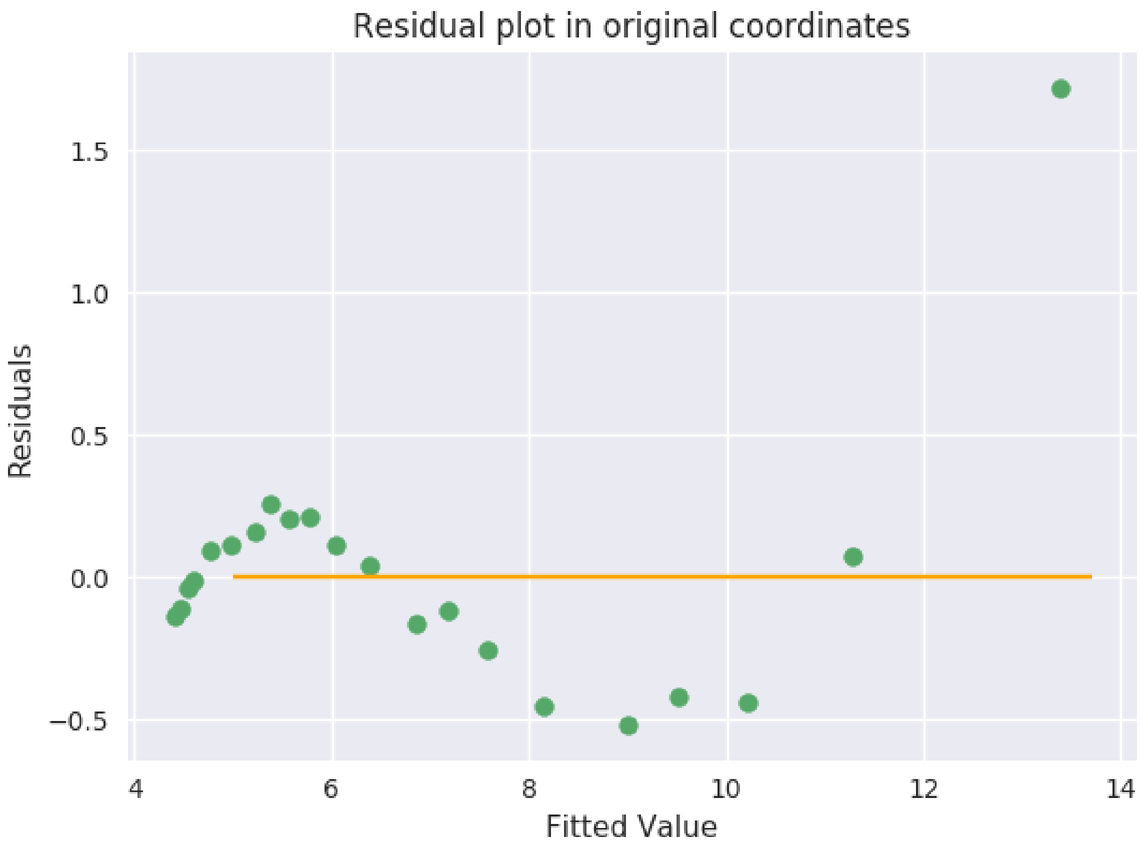
The regression seems to fit the data quite good and that too in a linear fashion. The **r2** value obtained is **0.9839** which indicates how well is the regression.

b) Following is the plot that shows the data point and regression line in original coordinate:



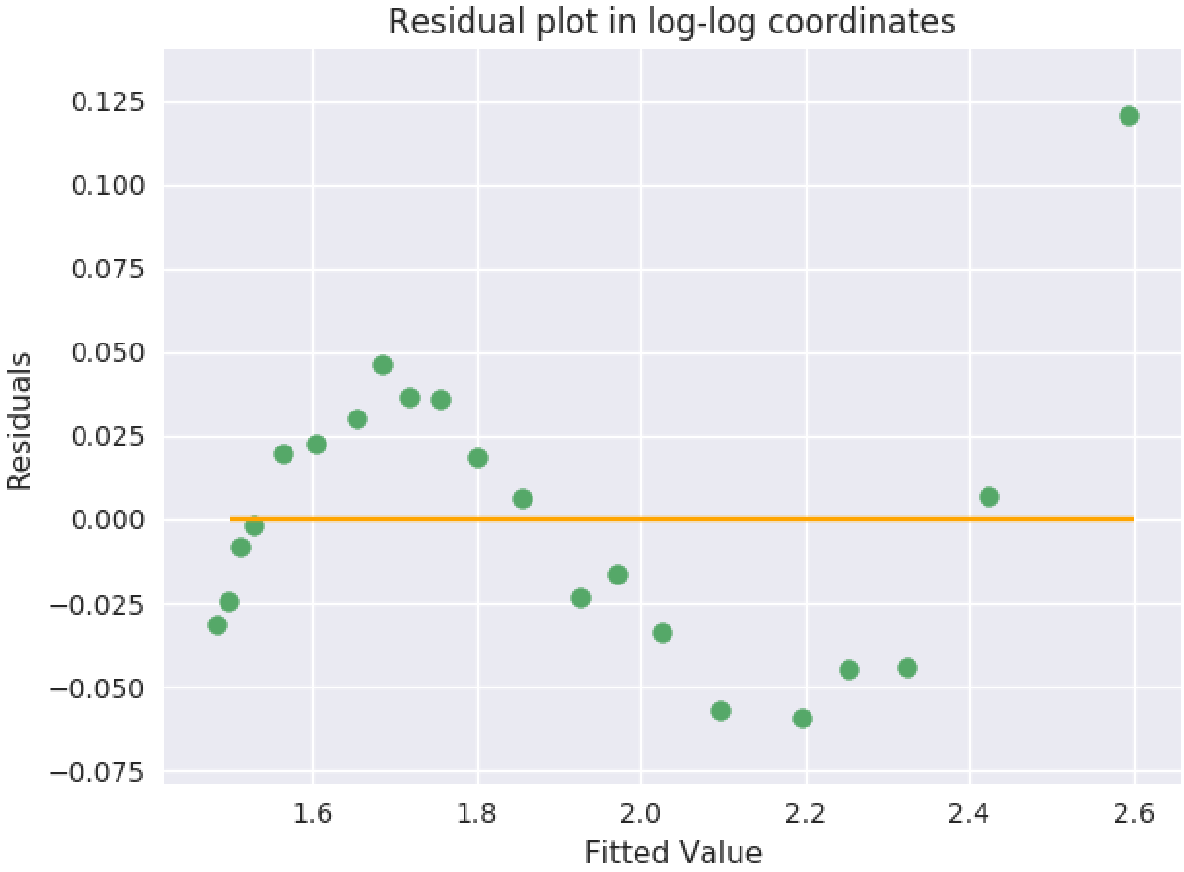
The following regression is also fitting the data nicely but in a nonlinear fashion the curve is exponential and the **r2** value for this is **0.9708.**

C) Following is the plot that shows the residual and fitted values in original coordinate:



Following plot shows the fitted value against the residuals, and here we see the points are not very randomly spread across the space, the residual values are very small. Even though there isn't any loud pattern depicted the high r2 value confirms that the model works fine.

Following is the plot that shows the residual and fitted values in log-log coordinate:



Here also we don't see much randomness in the points, the residuals values are quite small.

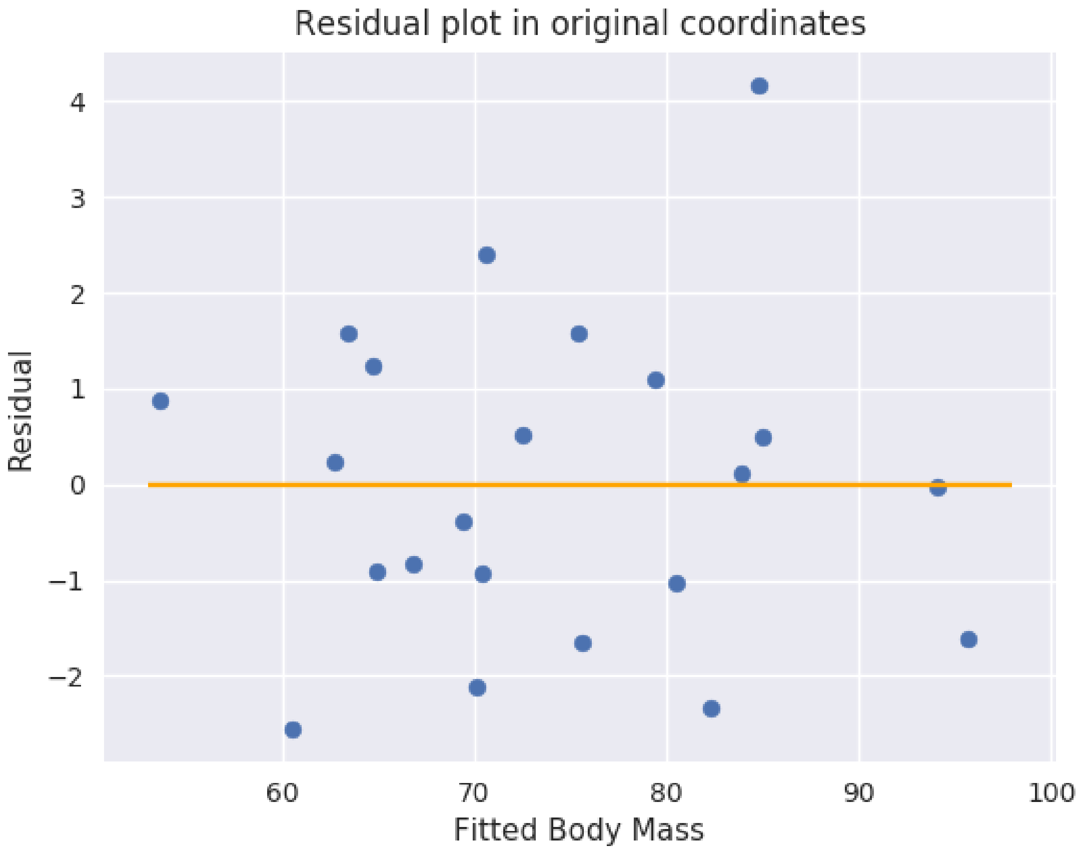
d.) The above plots shows that the regression is working quite good, having the regression build in log-log coordinate fits the data really well in both the original and log-log coordinate, the residual plots does not depict much randomness but the r2 values shown above are adhering to the fact that the following regression is doing really well. We could also see some outliers but those are not impacting much.

The stated behavior is confirmed by **plotting** the regression, studying the **residual plot** and calculating the **r2 value**.

Problem 7.10:

The dataset used contains the measurement that include the body mass along with other diameters. The aim is to fit a linear regression that could be used to predict the mass given these measurements.

a.) The following plot of residuals against the fitted value:

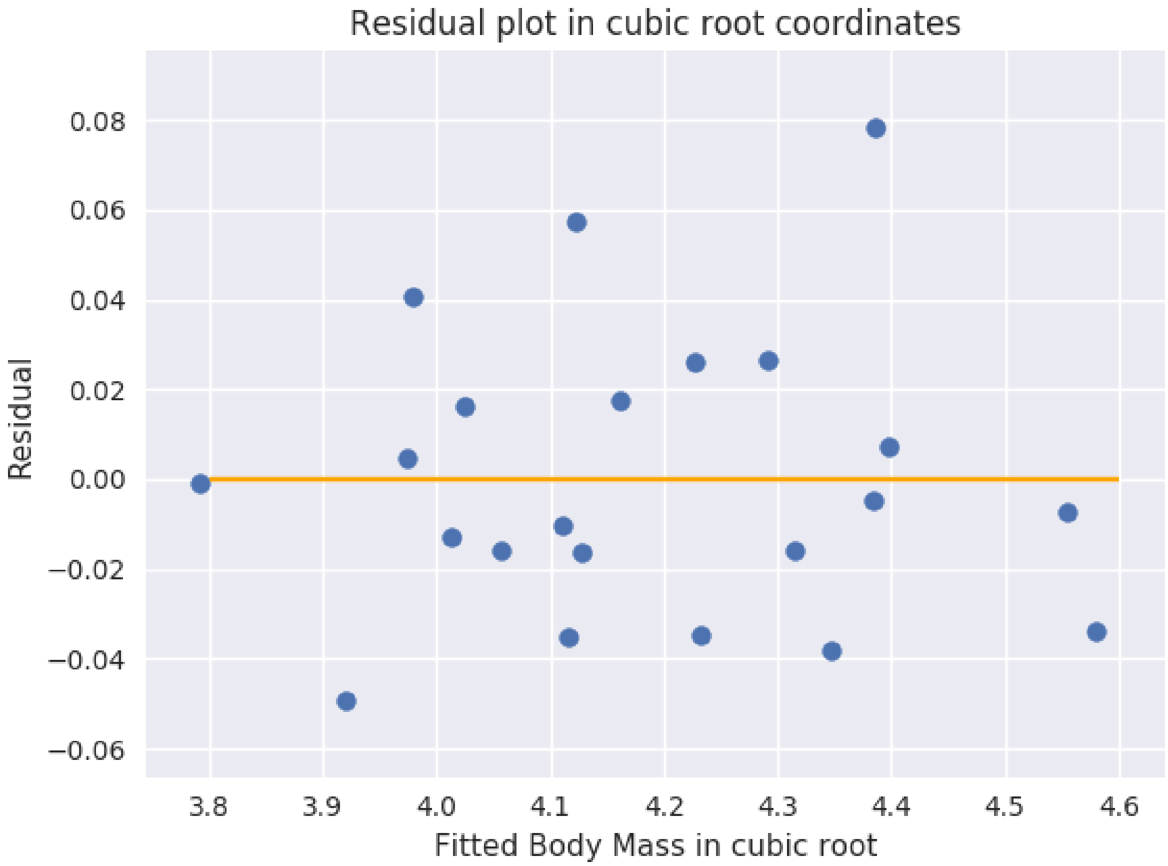


The plot clearly shows the randomness in the points and how they equally vary both above and below the 0 reference point. This might give us a hint that the model is working good.

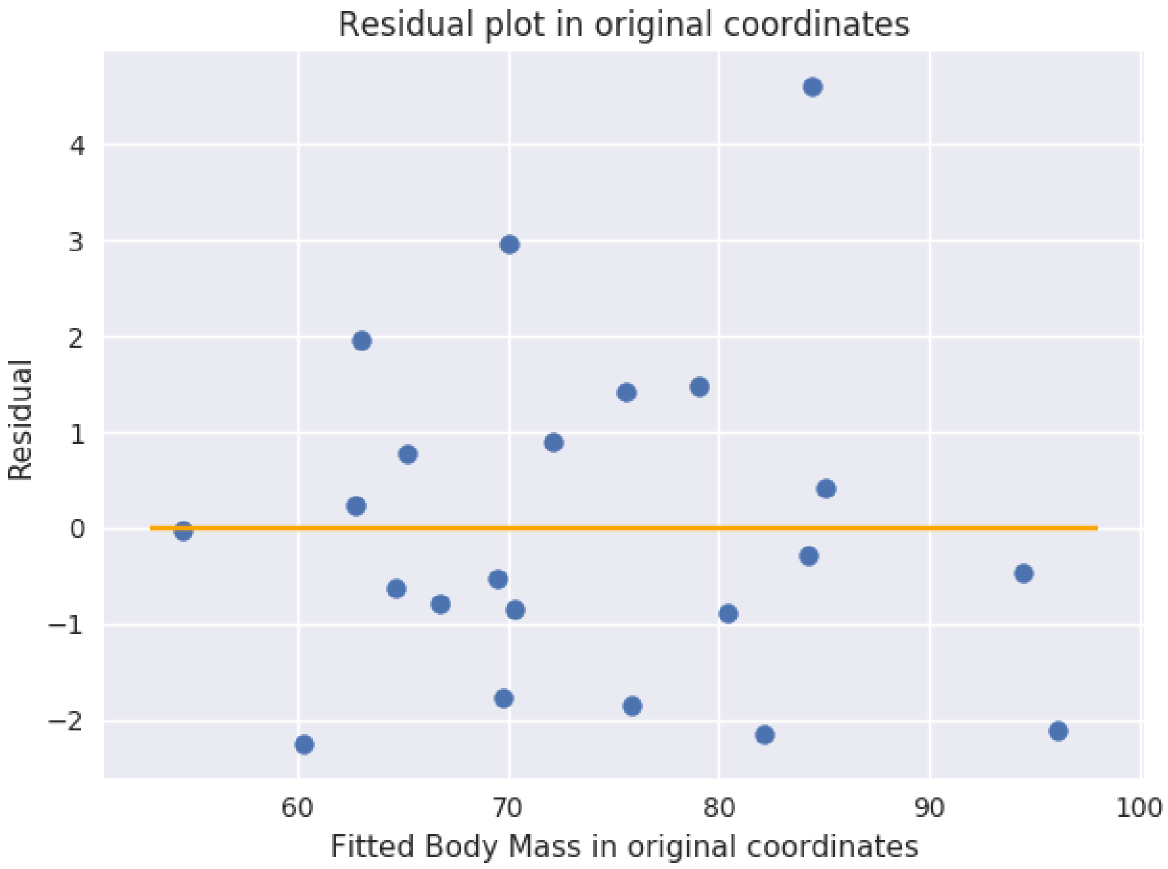
For this model the **r2 value** reported is **0.9772.**

b.) Now we take the cube root of the dependent variable(mass) and regress it against the diameter, following are the residual plots in both original and cube root coordinates.

i) Cube root coordinates:



ii.) Original coordinates:



The above 2 plots show the same behavior of randomness in the points indicating a good regression, the difference in the values of residuals are due to the fact that they are placed in different coordinate system.

For this model the **r2 value** reported is **0.974.**

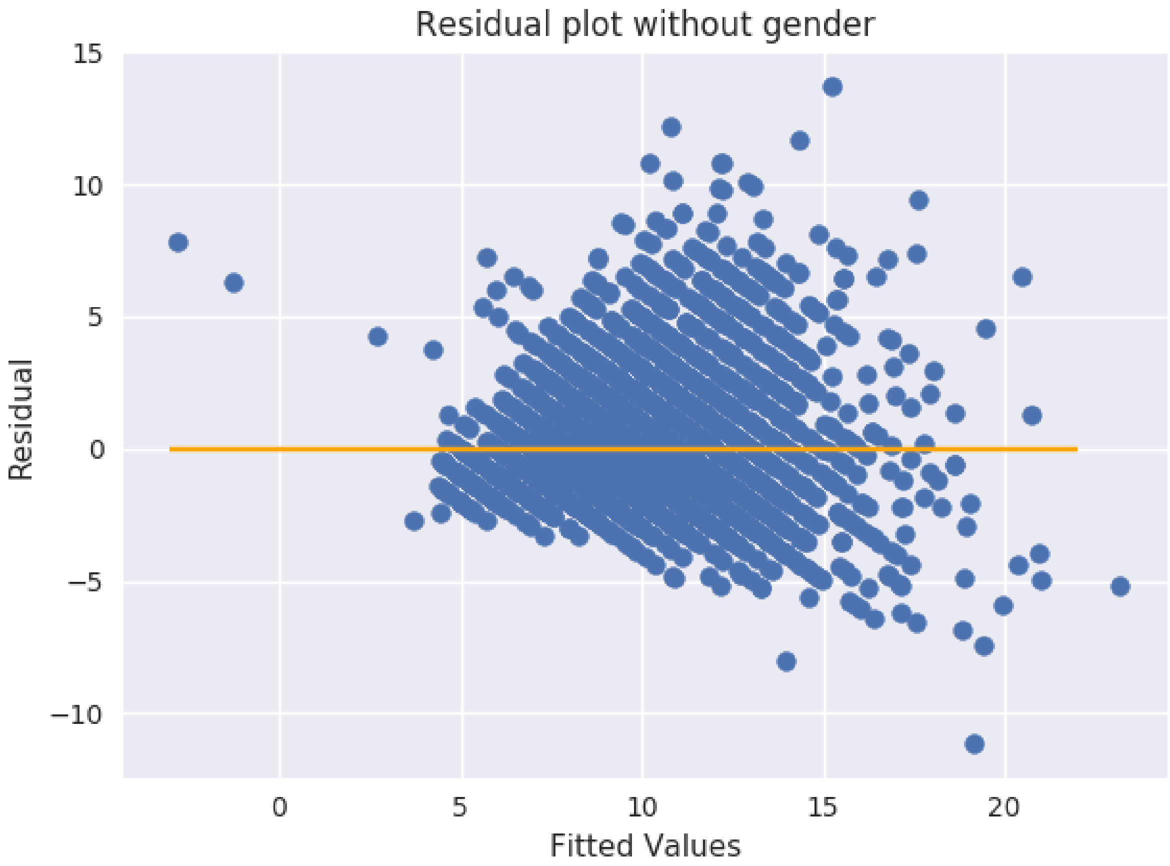
Such a high value confirms our understanding about the fitness of our model. The model is fitting the data quite well in both the original and cubic root coordinates.

c.) Looking at the above plots and r2 values it seems that both the models are doing really well having a r2 score of around **0.97** adheres to the applicability of the result. Even though there are outliers in the data but it isn't affecting the model much.

Problem 7.11:

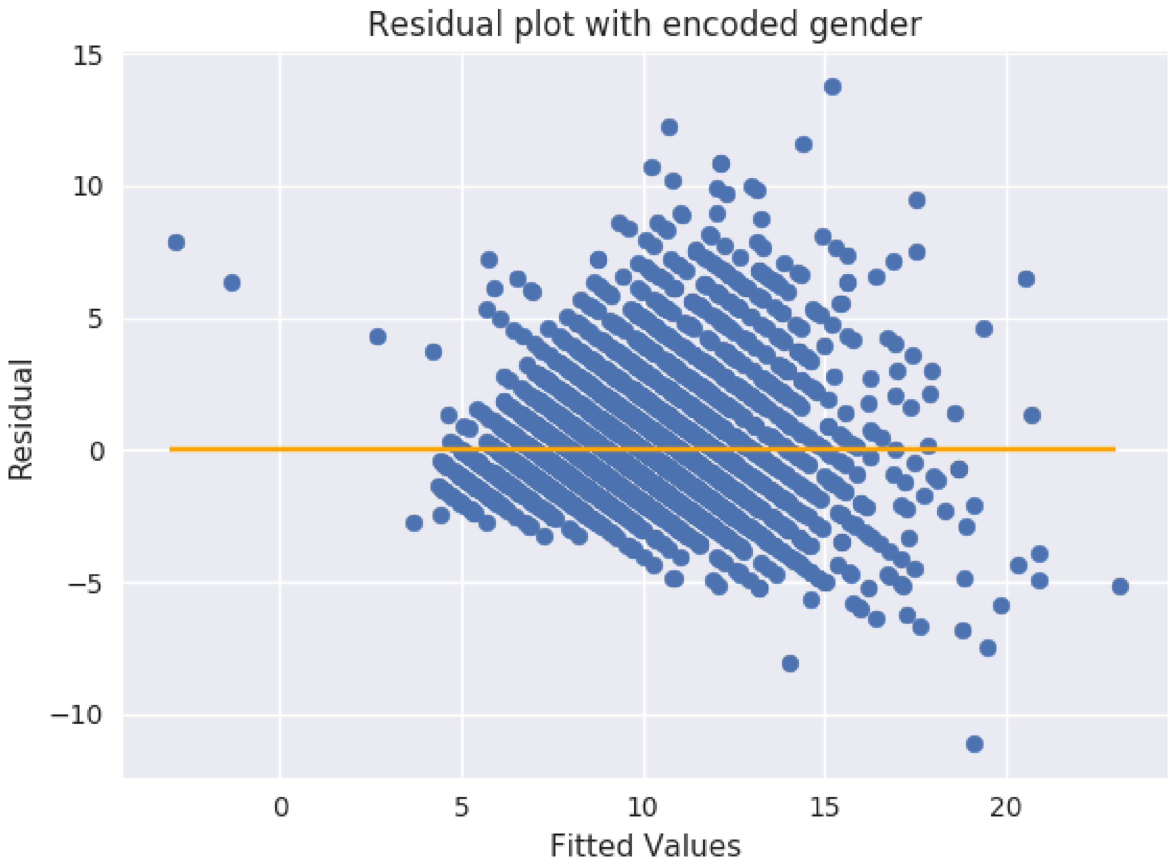
The data set measures the age from a variety of measurements and we build different models as per follow:

a.) We build a model predicting age from all the other measurements ignoring the gender variable. The residual plot is as follow:



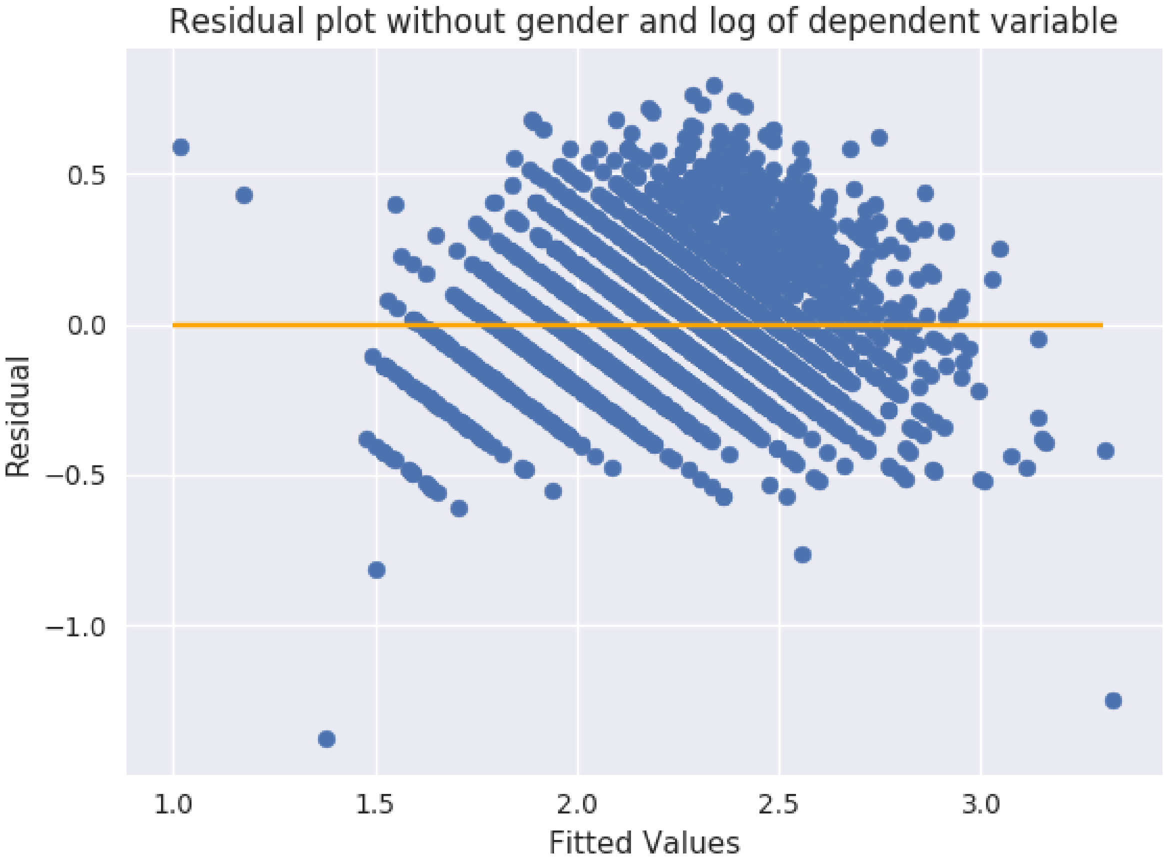
There is some randomness in the data points but there exists lot more values in the upper half as compared to lower half. The residual values are quite high as well and might be an indication of having some outliers in the dataset. The r2 value for this model is **0.527,** which is certainly not the best.

b.) We build a model predicting the age from the other measurements with the gender feature included and encoded. The residual plot is as follow:



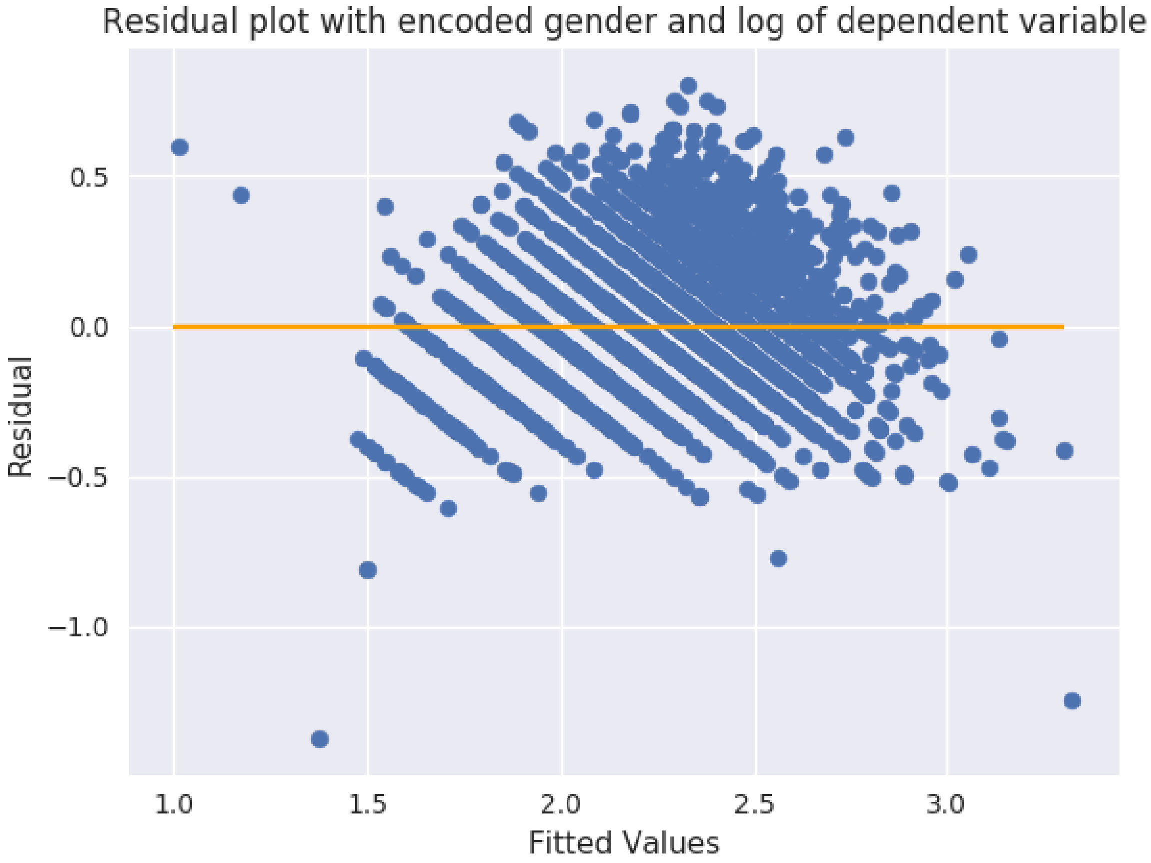
There isn’t much change in the plot as compared with the previous one, the randomness is low with more values in upper half and residuals are high. The r2 value reported for this model is **0.527.**

c.) We fit a model that predicts the log of age from the measurements ignoring the age. Following is the residual plot:



The plot starts showing some improvement in the randomness as compared to previous 2, also the residuals have reduced intensity. The r2 value reported is **0.5854**.

d.)We build a model that predicts the log of age from the measurements including the gender. Following is the residual plot:



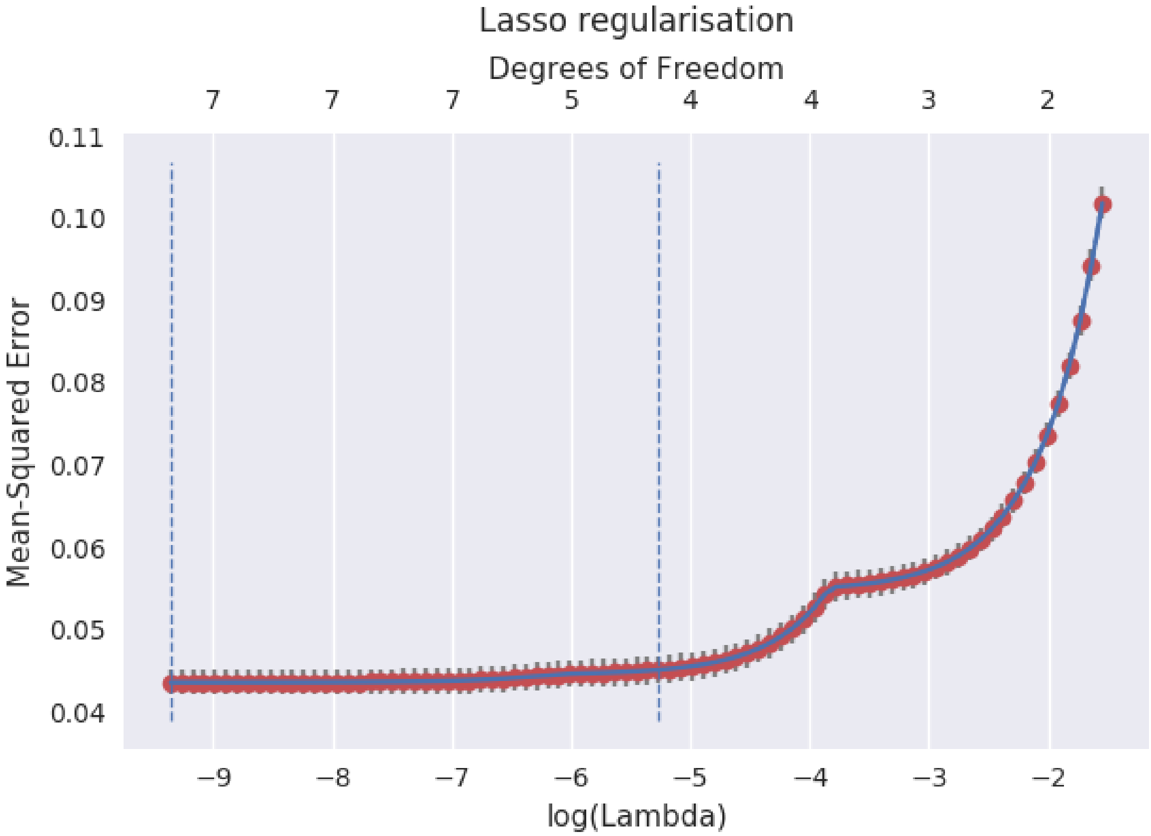
So there does not seem to be any effect on having or not having gender in the model and the plots looks about the same as previous one. The r2 value is **0.585.**

e.) From the above four plots I would say that the model we constructed for C that is predicting the log of the age excluding gender against the other measurement is working best when compared with other three. The gender feature doesn't contribute much in regression which is explained by residual plot and r2 score. The residual plot shows randomness in data points, the residuals are low and the r2 score is fairly significant 0.585.

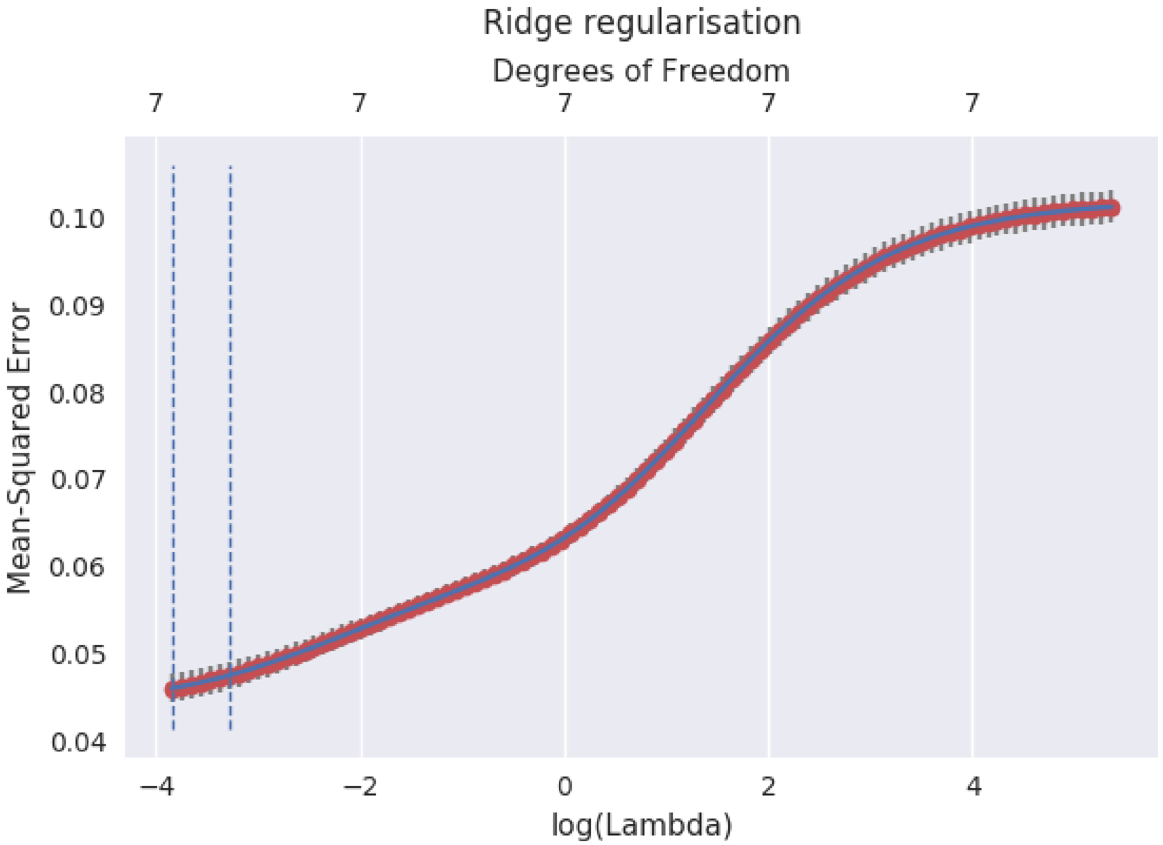
Even though this might show good value for predicting log of the age it might not be the best one to predict the age as we saw in the above plots, we may need to look for the outliers and try removing them as this property is indicative from the residual plot. We can also consider scaling the features so that we reduce the within data variance and have most of the variance being explained by the model.

f.) We used glmnet with cross validated prediction error to find a value of lambda that can act as a good regularization constant and help us improve our model. Following are the plots in ridge and lasso:

i.) Lasso:


Lasso regularisation
Degrees of Freedom
7
7
7
5
4
4
3
2
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
Mean-Squared Error
−9
−8
−7
−6
−5
−4
−3
−2
log(Lambda)

ii.) Ridge:

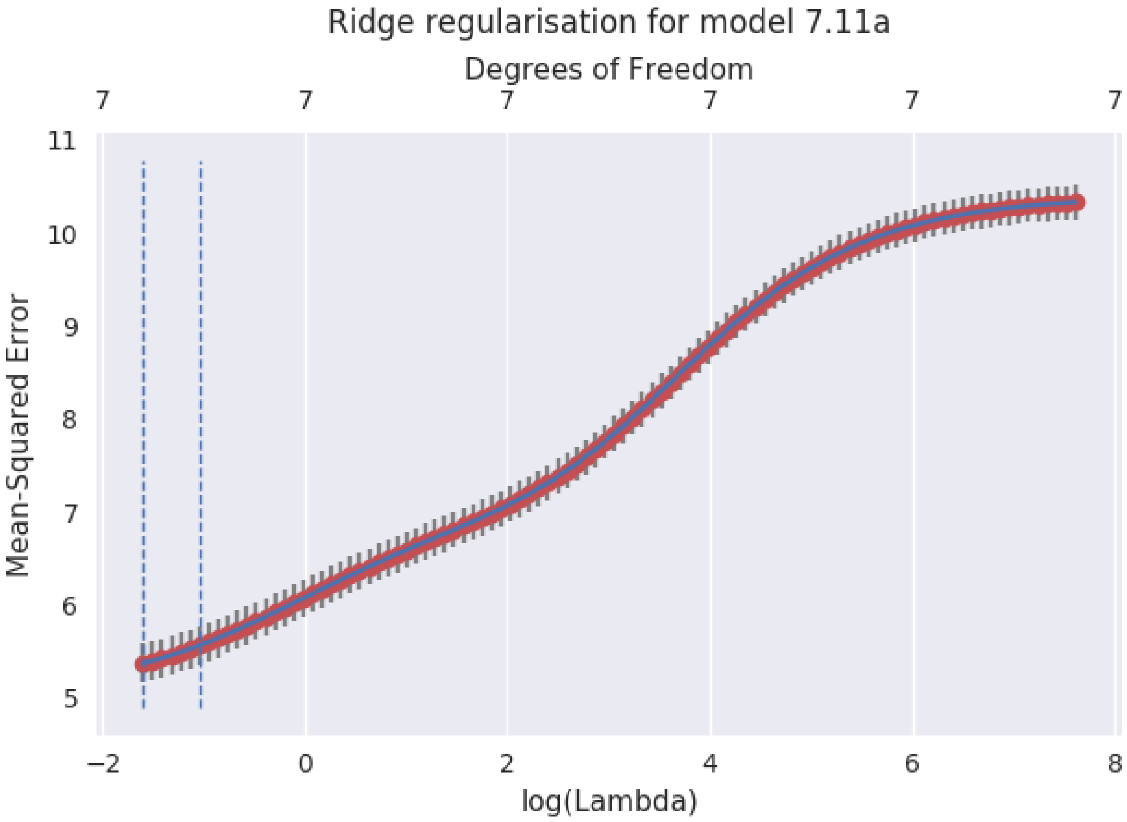


The above plots show 0.021 for ridge and 0.000094 for lasso as a good value of lambda that could be used as a regularization constant.

We tried using the above findings to do predictions and calculate r2 score and it came out to be **0.585 for lasso and 0.5597 for ridge.** We used the model predicting log of the age against the other features and added this regularization constant to that. For lasso we got similar r2 score for ridge we got a lower r2 score hence concluding that the regularization won't help that much.

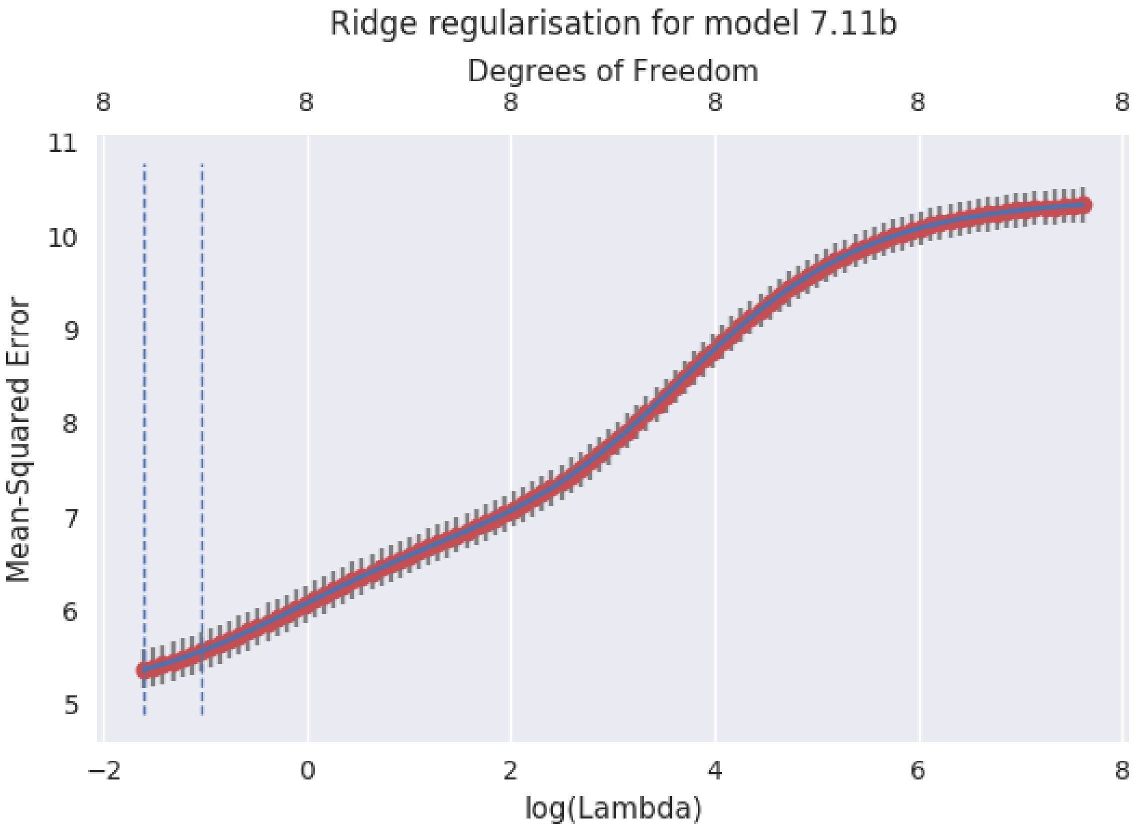
We then performed the ridge regularization on each of the model created from a-d and got the following output:

I.) For model constructed in 7.11a:



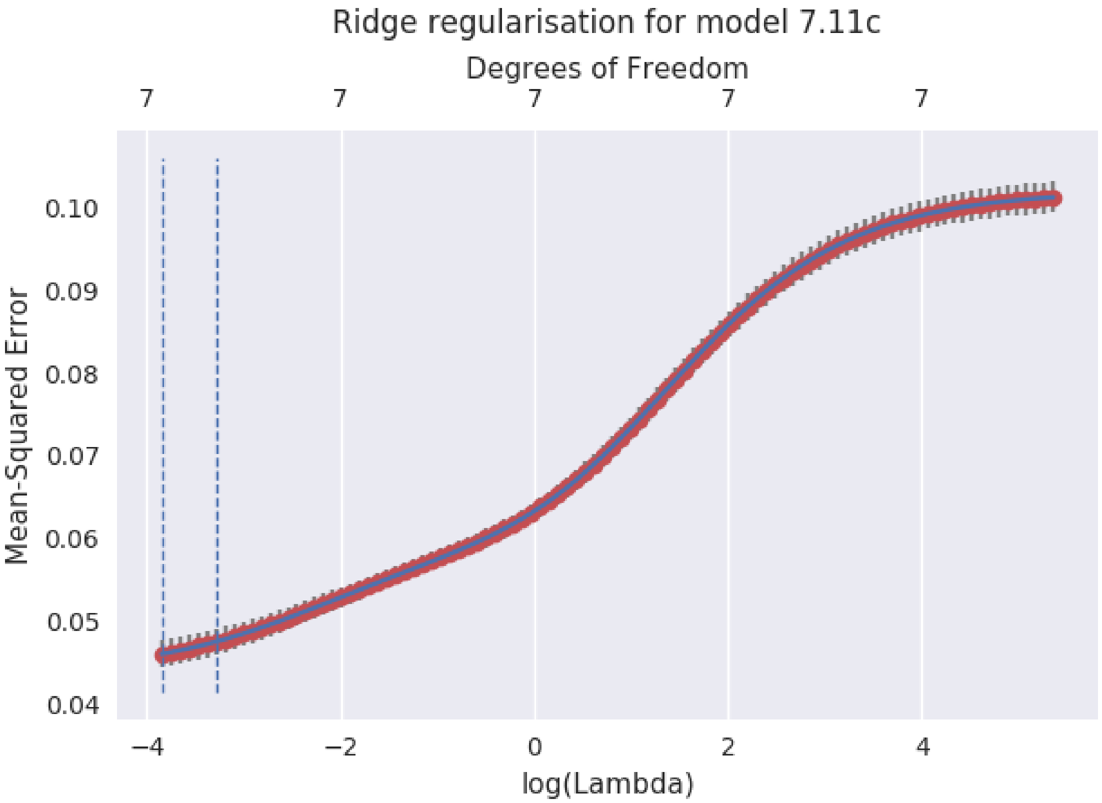
r2 score is 0.491635924617

2.) For model constructed in 7.11b:



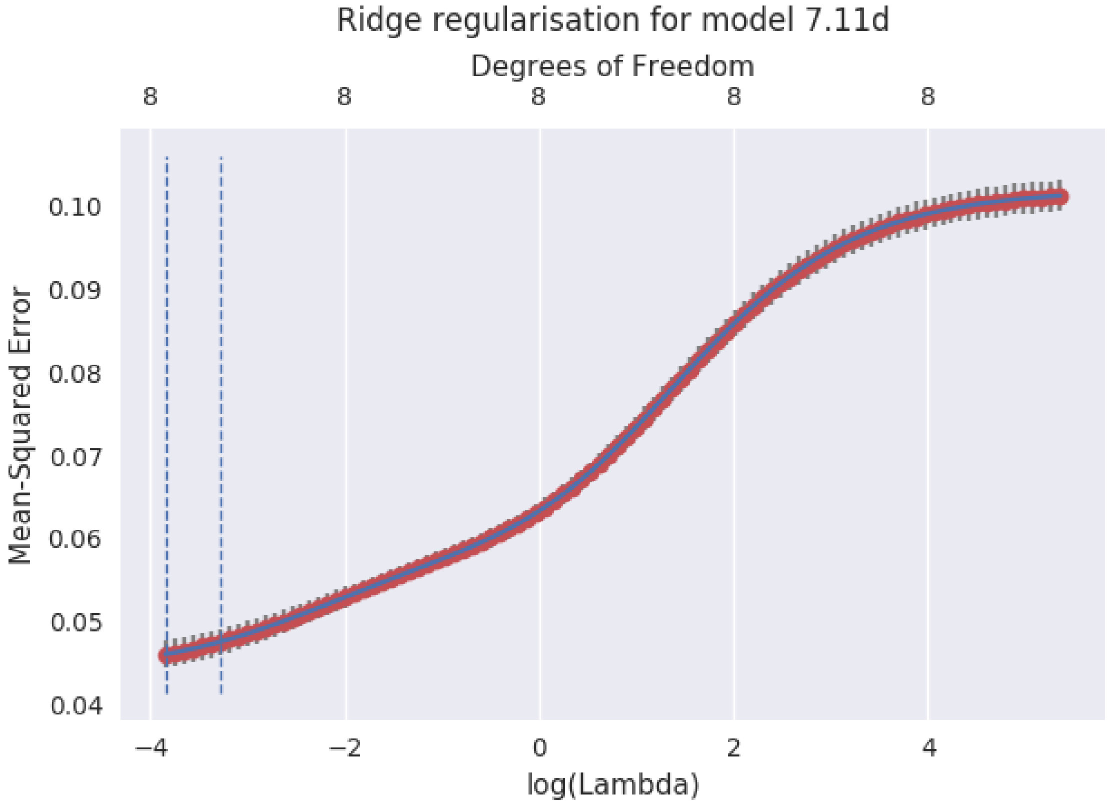
r2 score is 0.491716379542

3.) For model constructed in 7.11c:



r2 score is 0.559738742298

4.) For model constructed in 7.11d:



r2 score is 0.559968459886

Looking at the above plots and r2 score it's clear that the regularization isn't working in favor of the model.

<u>**Citation**</u>: We looked at some of the stack overflow resource and glmnet documentation to see how glmnet works.

-http://glmnet-python.readthedocs.io/en/latest/glmnet_vignette.html

-http://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

-https://stackoverflow.com/questions/31287552/logarithmic-returns-in-pandas-dataframe