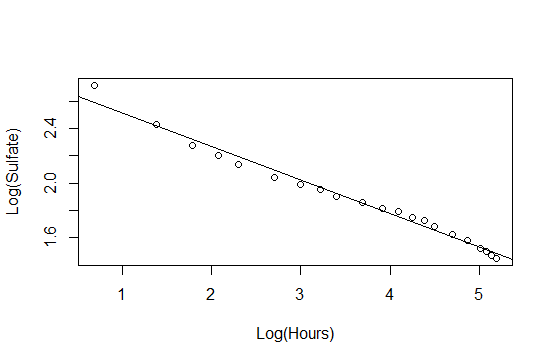
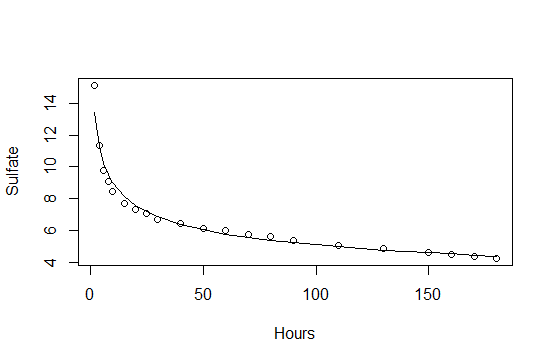
**Question 1**

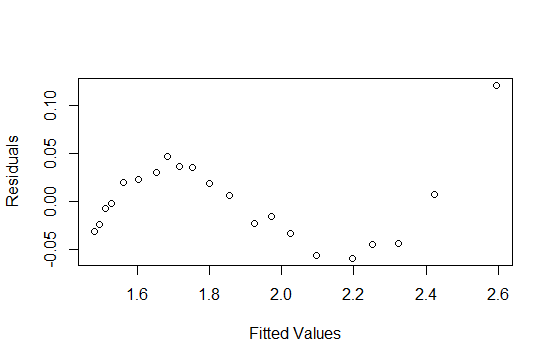
a- Plotting the regression in log/log coordinates



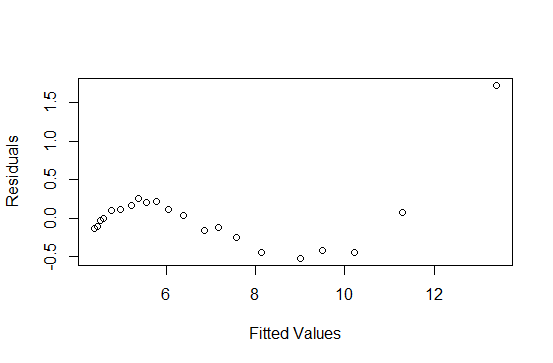
b- Plotting the regression in original coordinates



c- Residuals against fitted values in log-log coordinates



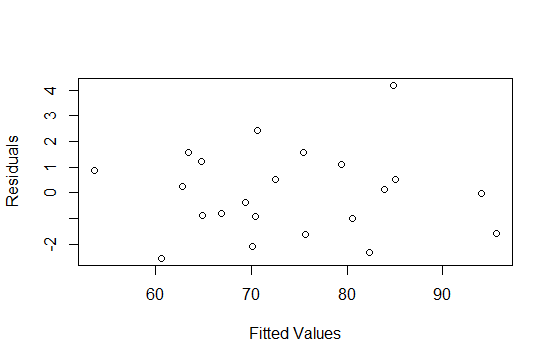
Residuals against fitted values in original coordinates



d- The regression looks good as it pretty much fits the points, although residuals are not entirely random, but I think that’s because there are very few points in general, however, their means looks like 0.

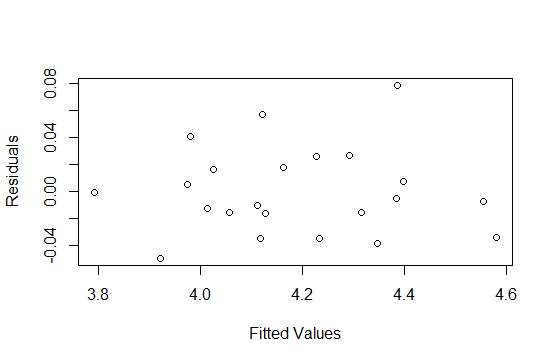
**Problem 2**

a- Residual against fitted values for the mass against diameters

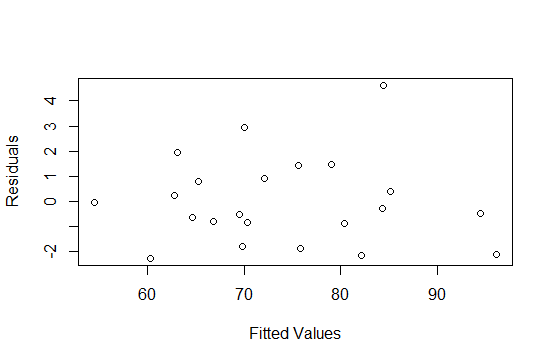


b- Residual against fitted values for the cube root of mass against the variables

1- In Cube root coordinates



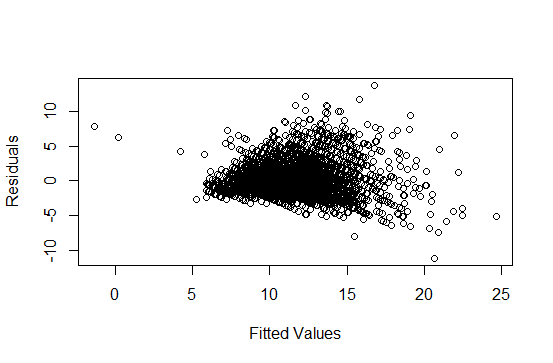
2- in Original coordinates



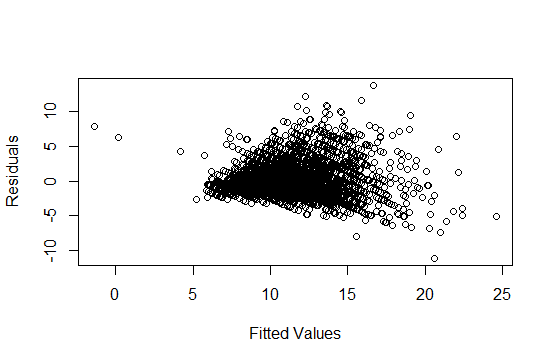
c- The two regressions looks pretty similar, so I would go with the simpler model.

**Problem 3**

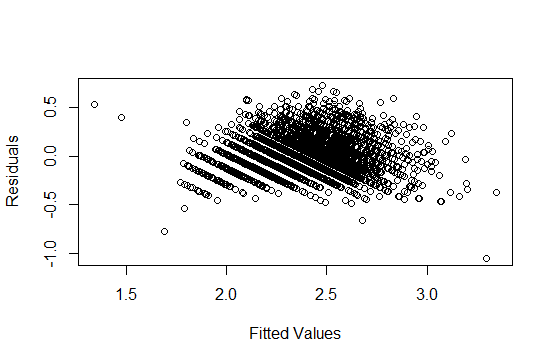
a- Residual against fitted values for regressing the age from the measurements excluding gender



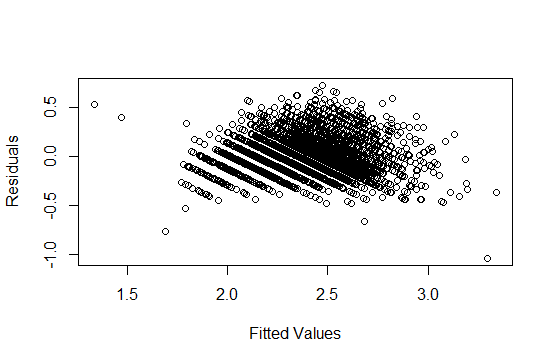
b- Residual against fitted values for regressing the age from the measurements including gender



c-Residual against fitted values for regressing the log of age from the measurements excluding gender



d-Residual against fitted values for regressing the log of age from the measurements including gender

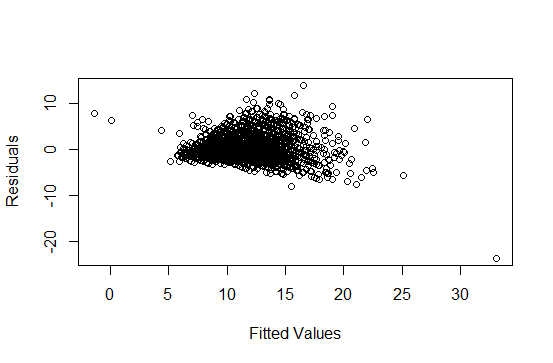


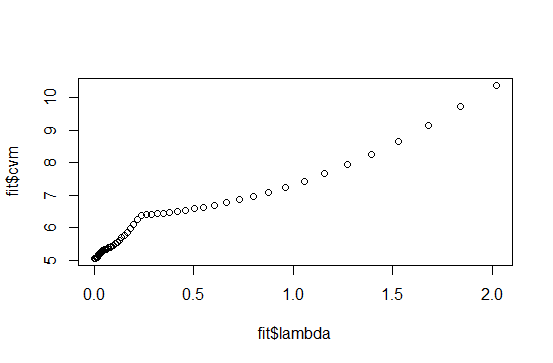
e- Out of the four regressions I would use the last one, predicting the log of age from all the variables including gender, its plot is more scattered and thus represents noise better in compare to using the age as is (a and b models), if we consider this as a random variable, its mean is around 0.

In compare to the third model (log of age against variables without gender), including gender made the points a bit more centered around 0, this was also verified in the R-Squared measure.

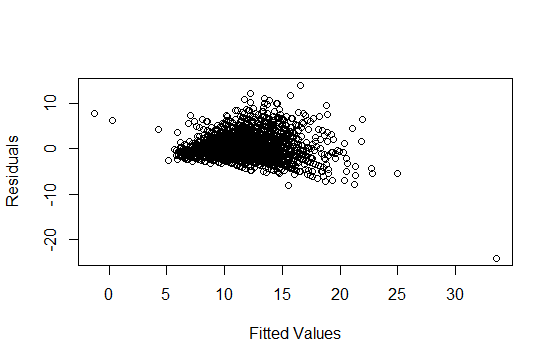
f- The following plots are using cv.glmnet to predict the same regressors like the previous questions but with regularization and cross validation error. Will shows two plots (residuals against fitted values and cross validation error against lambda values respectively) for each of the four models

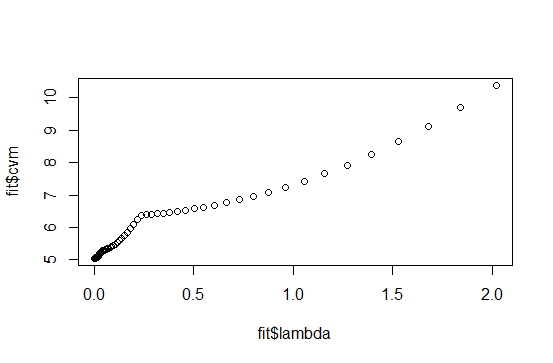
f.a- Residual against fitted values for regressing the age from the measurements excluding gender (regularization included)



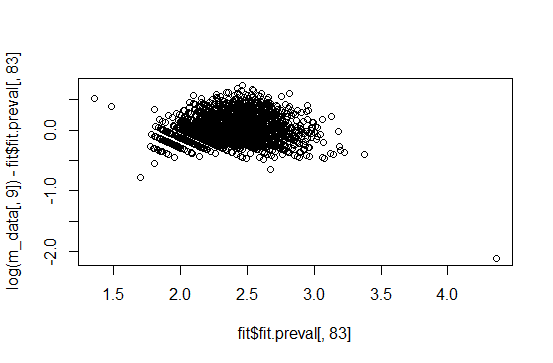


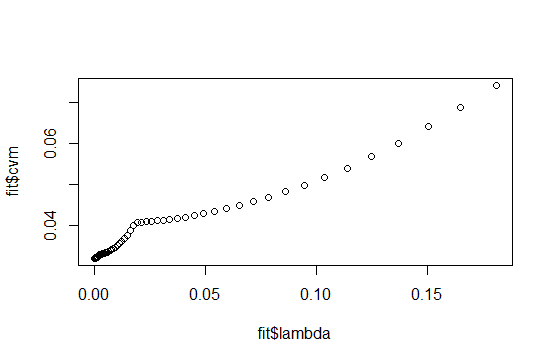
f.b- Residual against fitted values for regressing the age from the measurements including gender (regularization included)



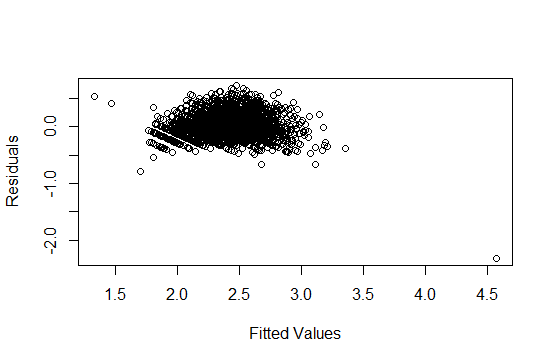


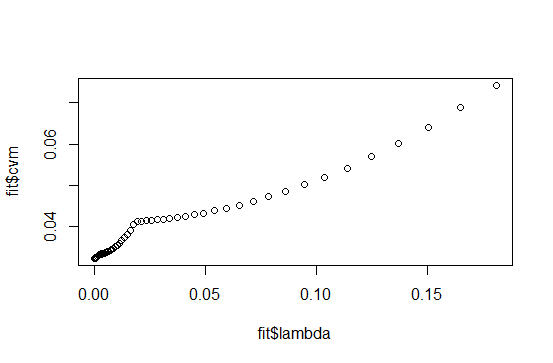
f.c- Residual against fitted values for regressing the log of age from the measurements excluding gender, and then a plot for the cross-validation error against different lambda values





f.d- Residual against fitted values for regressing the log of age from the measurements including gender, and then a plot for the cross-validation error against different lambda values





**Problem 3 Conclusion**

Using regularization made slight improvement for all the regressions, I used the cross-validation error against the lambda to pick the best lambda, and then plotted the residuals against fitted values to compare with the originals. Error looks closer to noise than the original plots, but more importantly, I calculated R-Squared measure for all the regressions with and without regularization, and regularization is always giving a higher value which is better.

|  |  |  |
| --- | --- | --- |
| **Regressor** | **R-Squared with regularization** | **R-Squared with regularization** |
| Age against variables without gender | 0.5276299 | 0.5343709 |
| Age against variables with gender | 0.5278909 | 0.5346378 |
| Log of age against variables without gender | 0.5796935 | 0.5866443 |
| Log of age against variables with gender | 0.5801268 | 0.5866369 |