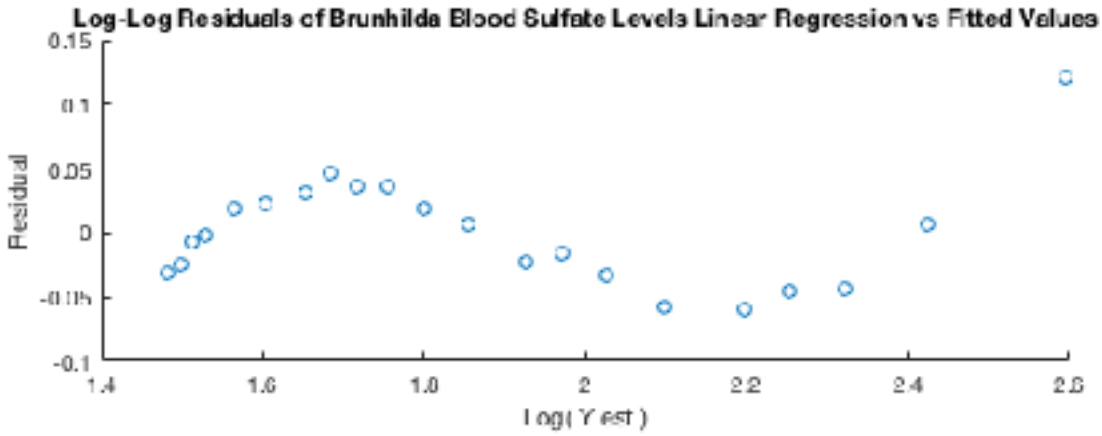
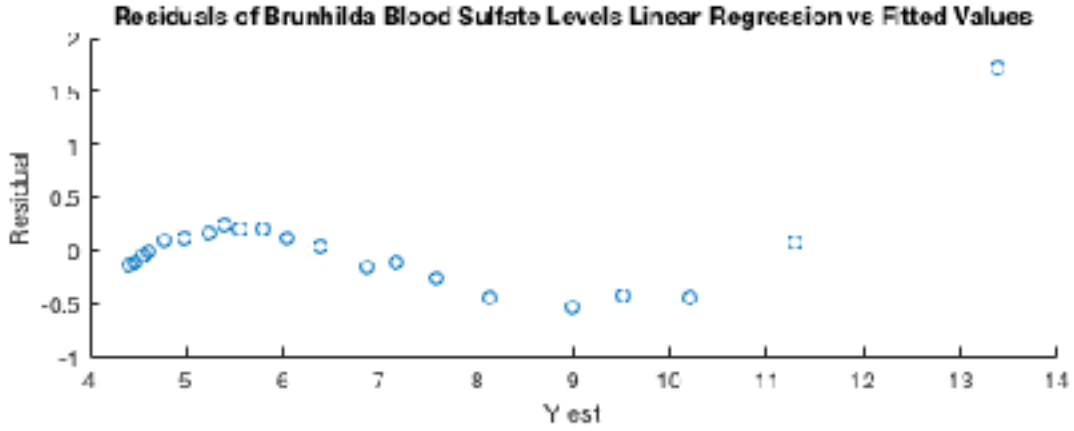
CS 498 AML: Homework 4

Rauhul Varma (rvarma2), Tadas Aleksonis (alekson2)

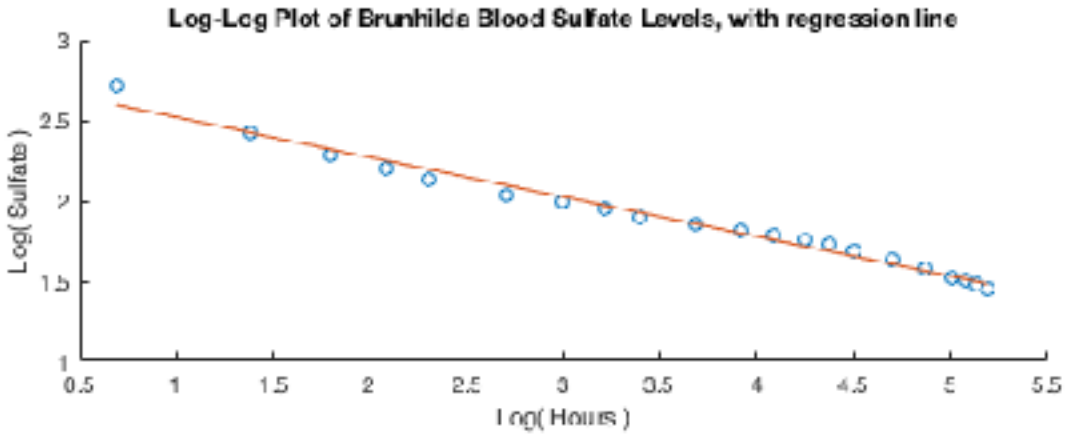
Problem 7.9

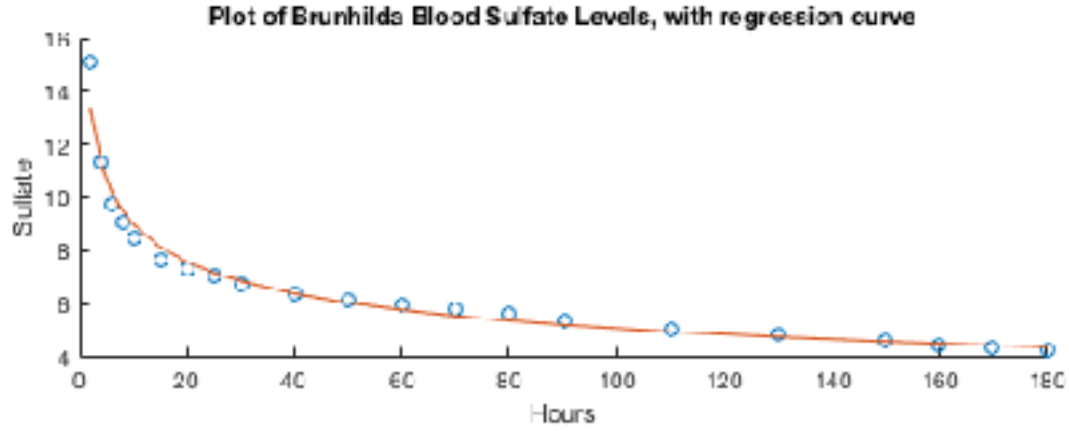
a. Prepare a plot showing (a) the data points and (b) the regression line in log-log coordinates
b. Prepare a plot showing (a) the data points and (b) the regression curve in the original coordinates





c. Plot the residual against the fitted values in log-log and in original coordinates



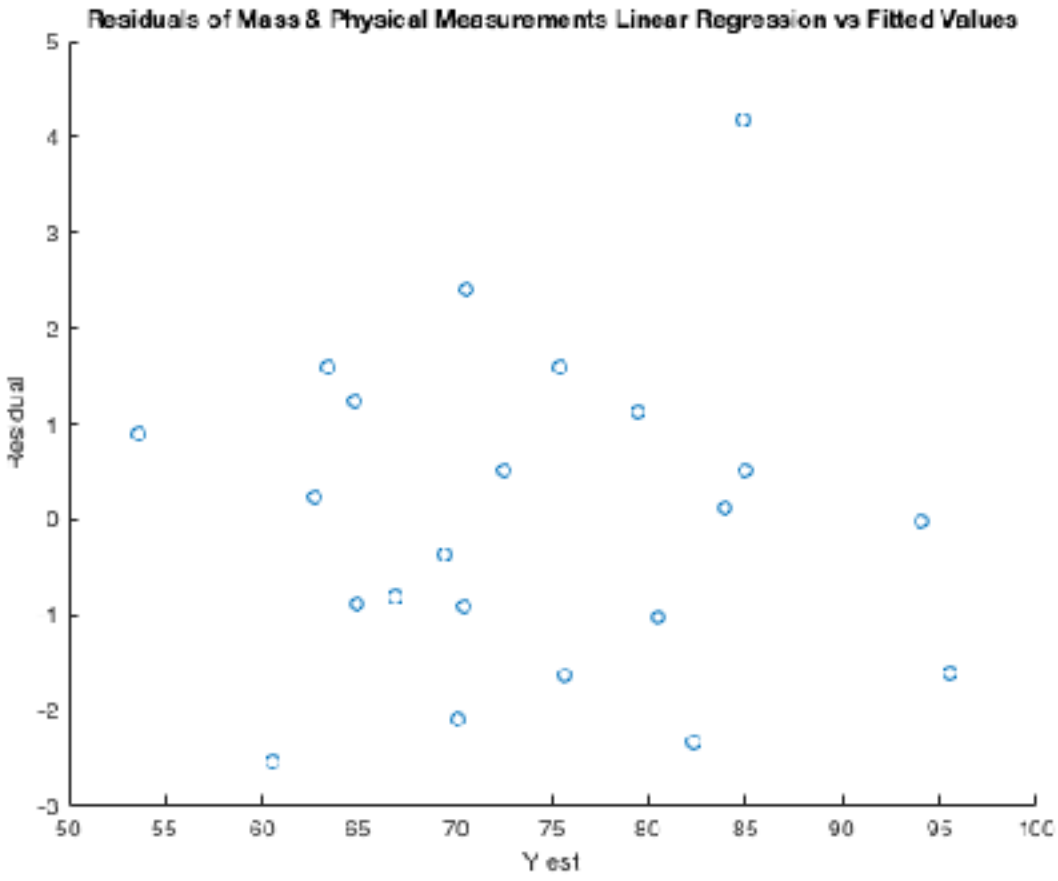


d. Use your plots to explain whether your regression is good or bad and why

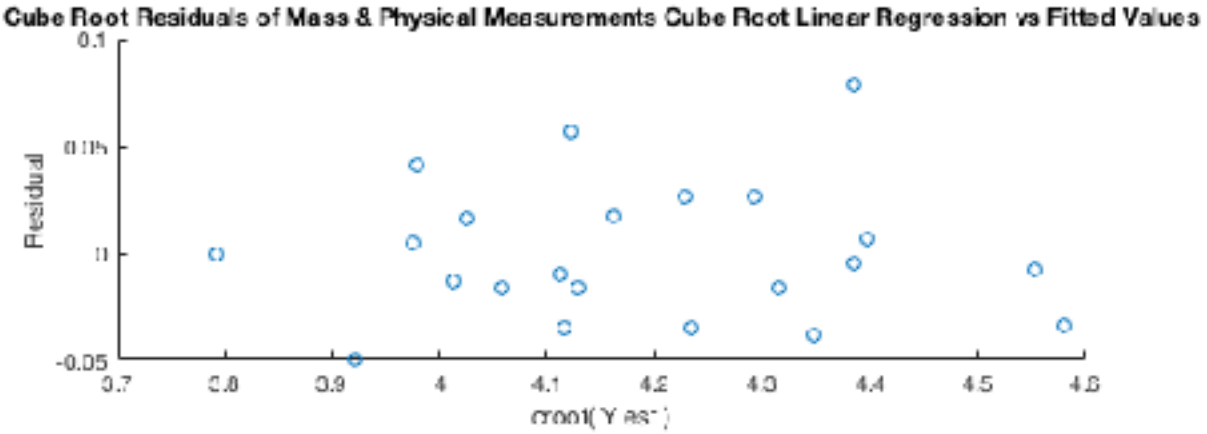
Our calculated regressions are not good because as shown in the residual, the error is not random. There is some sort of polynomial curve that is not captured by our regression.

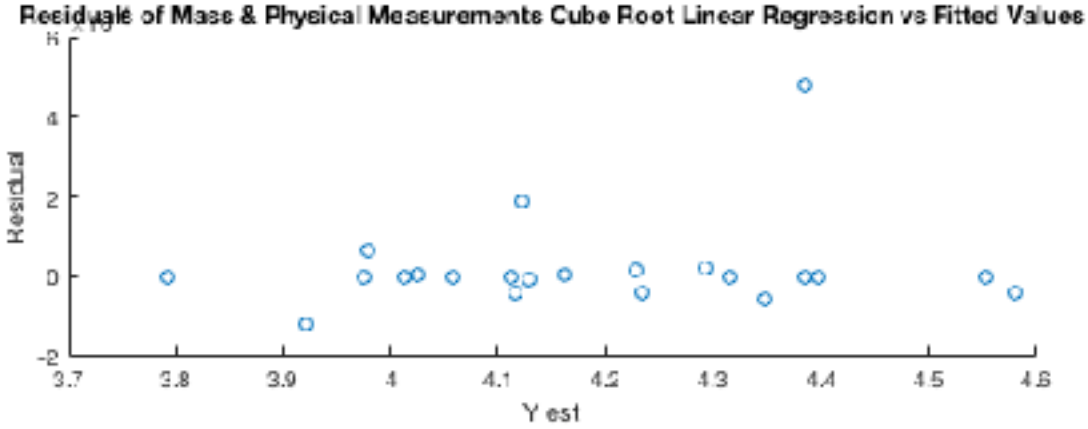
Problem 7.10

a. Plot the residual against the fitted values for your regression.



b. Now regress the cube root of mass against these diameters. Plot the residual against the fitted values in both these cube root coordinates and in the original coordinates.



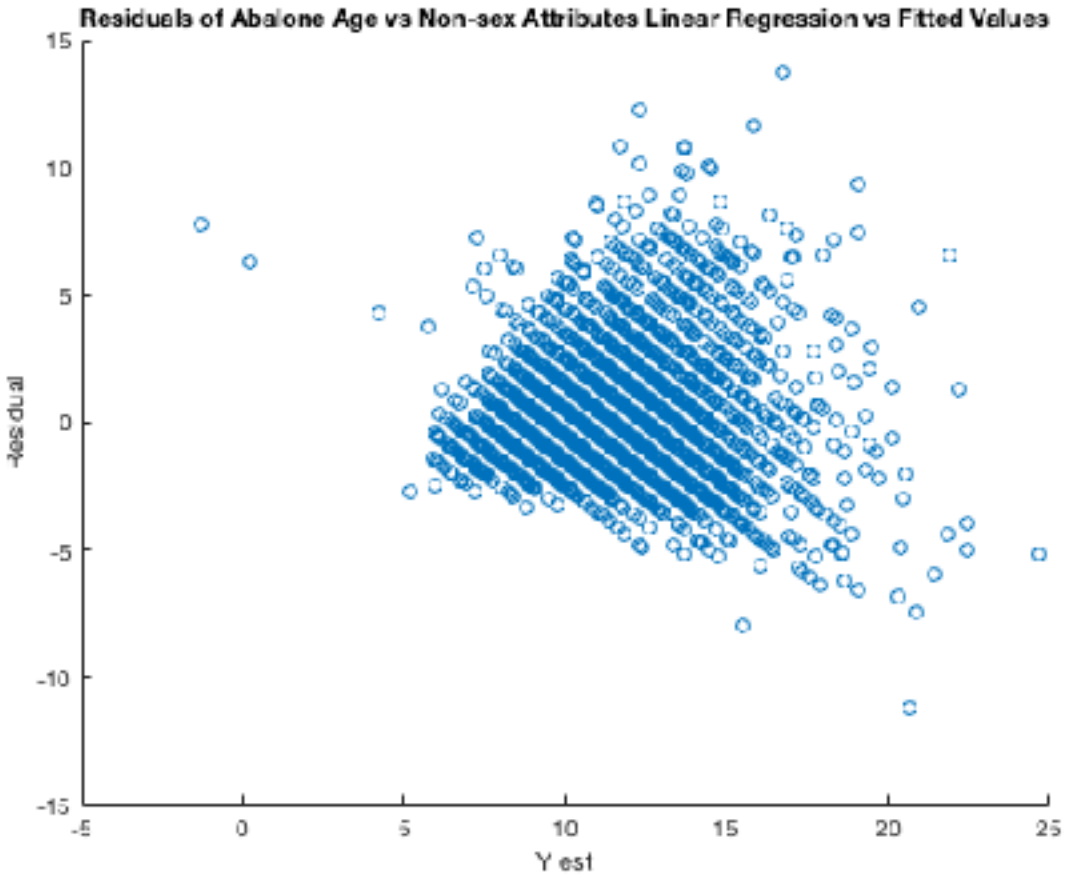


c. Use your plots to explain which regression is better.

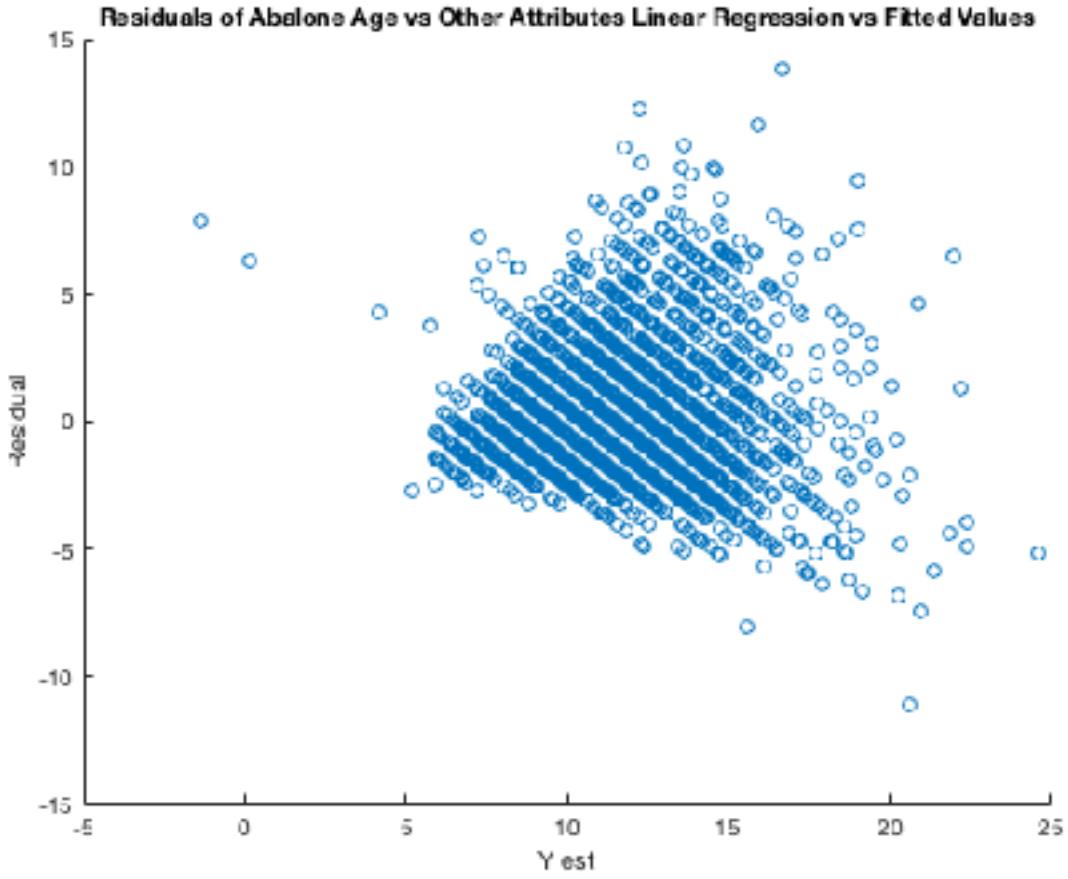
Both regressions are good as they both have random error in their residuals, but the cube root residual is better because it minimizes the total error. This fact is seen by the smaller overall variance of the residual error in the cube root linear regression.

Problem 7.11

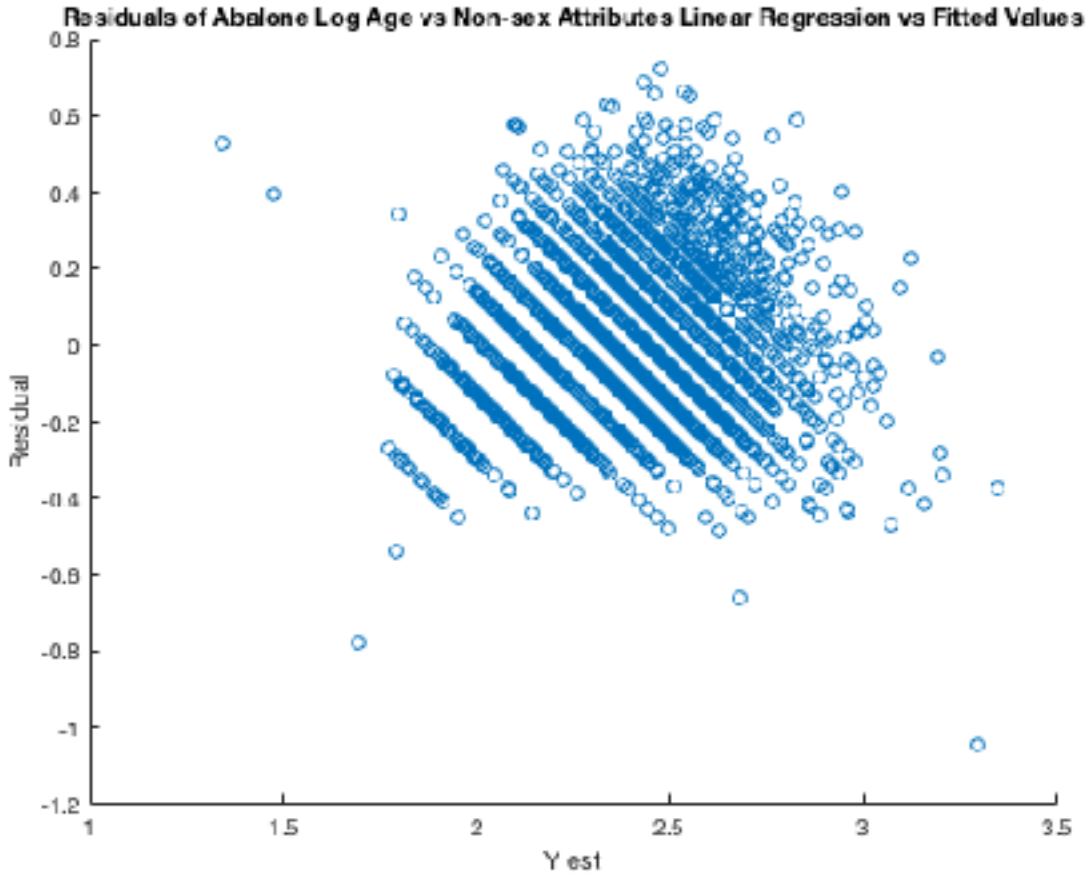
a. Build a linear regression predicting the age from the measurements, ignoring gender. Plot the residual against the fitted values.



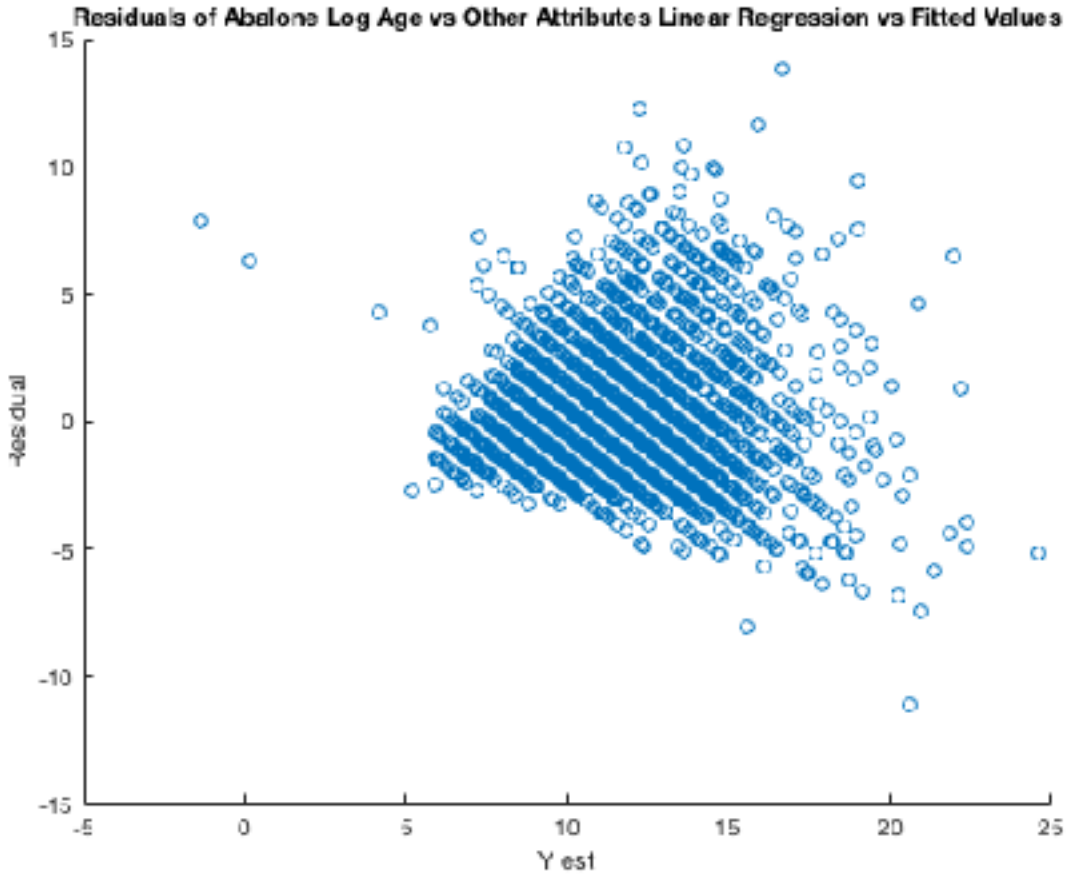
b. Build a linear regression predicting the age from the measurements, including gender. There are three levels for gender. Plot the residual against the fitted values.



c. Now build a linear regression predicting the log of age from the measurements, ignoring gender. Plot the residual against the fitted values.



d. Now build a linear regression predicting the log age from the measurements, including gender, represented as above. Plot the residual against the fitted values.



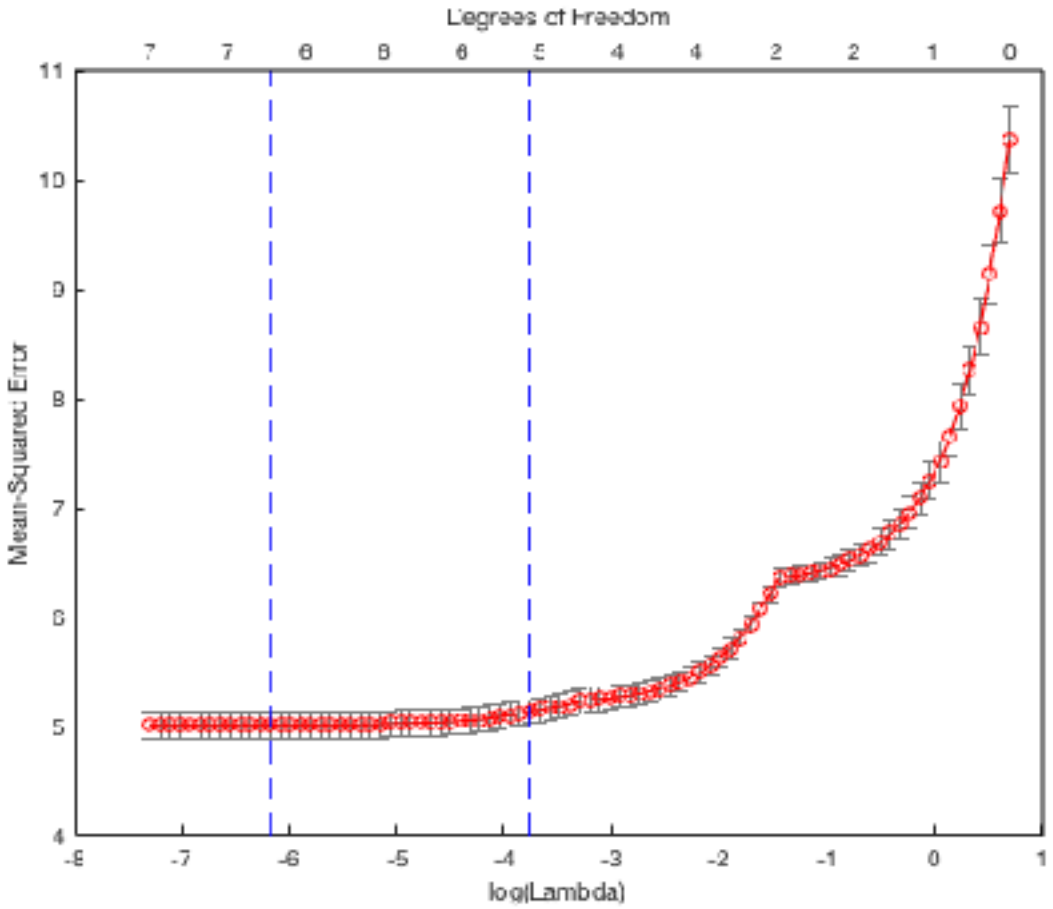
e. It turns out that determining the age of an abalone is possible, but difficult (you section the shell, and count rings). Use your plots to explain which regression you would use to replace this procedure, and why.

Out of the four regressions, the best regression to use is the Age vs Non-Sex Attributes of the Abalone regression, as it has the smallest variance in the residuals and mean closest to zero. However, without a more sophisticated method, most differences between the regressions cannot be noticed by eye.

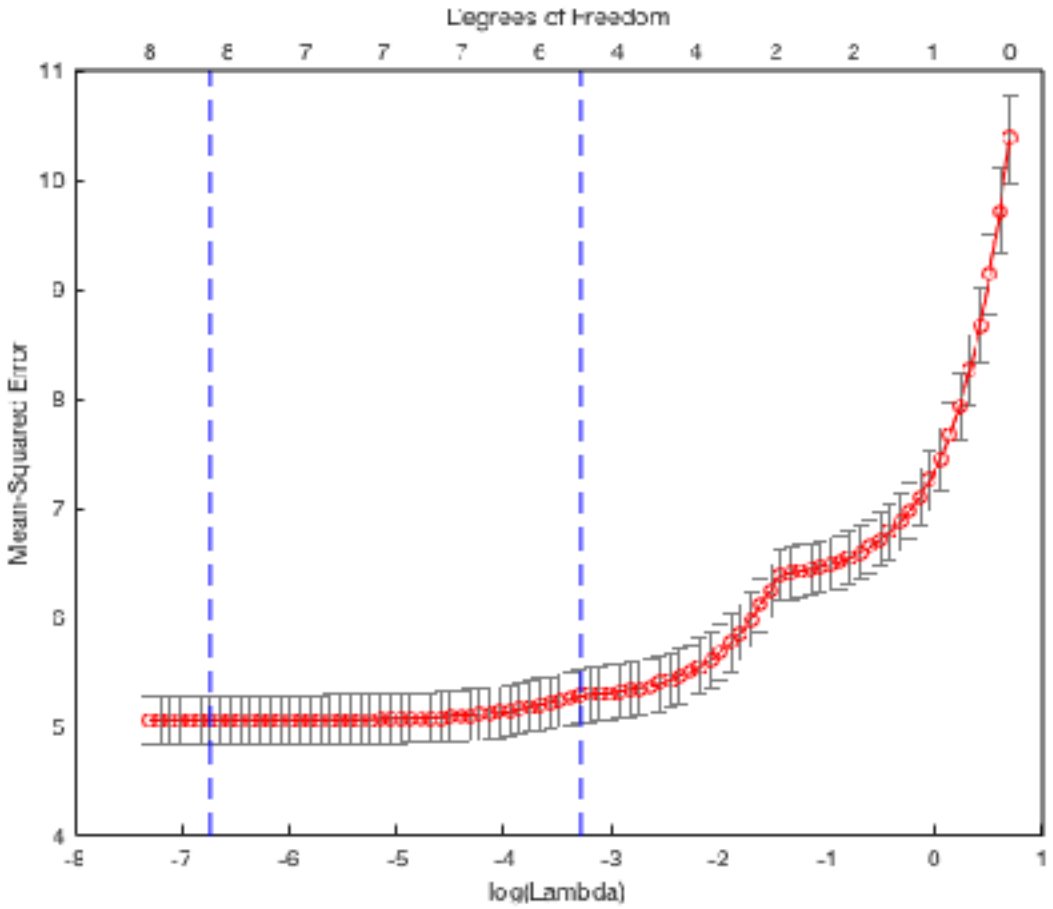
f. Can you improve these regressions by using a regularizer? Use glmnet to obtain plots of the cross-validated prediction error.

The following graphs are each of the four regressions through the glmnet regularizer package. The best regression uses the Log(age) vs Non-Sex attributes as is seen by the orders of magnitude smaller mean-squared error.

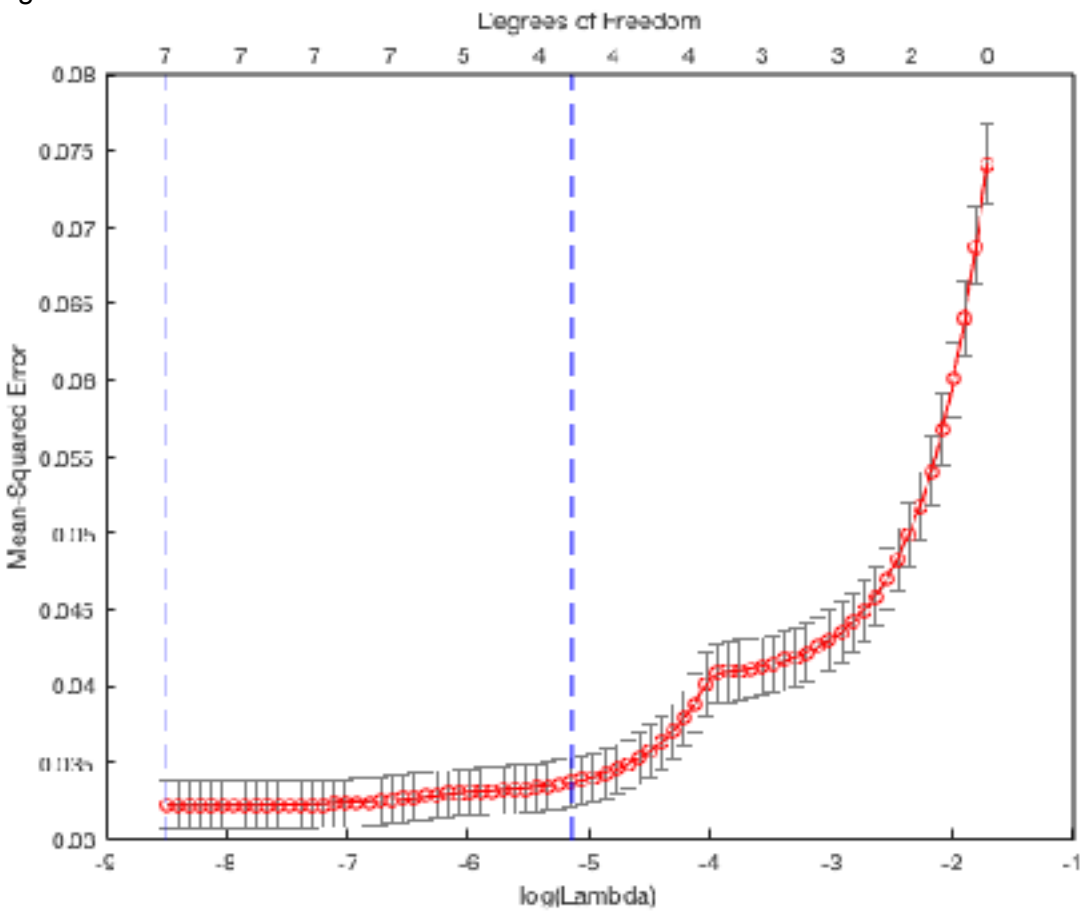
Age vs Non Sex Attributes



Age vs All Attributes



Log Age vs Non Sex Attributes



Log Age vs All Attributes