7.11. At https://archive.ics.uci.edu/ml/datasets/Abalone, you will find a dataset of measurements by W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn and W. B. Ford, made in 1992. These are a variety of measurements of blacklip abalone (Haliotis rubra; delicious by repute) of various ages and genders.

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

R^2 = 0.5276299

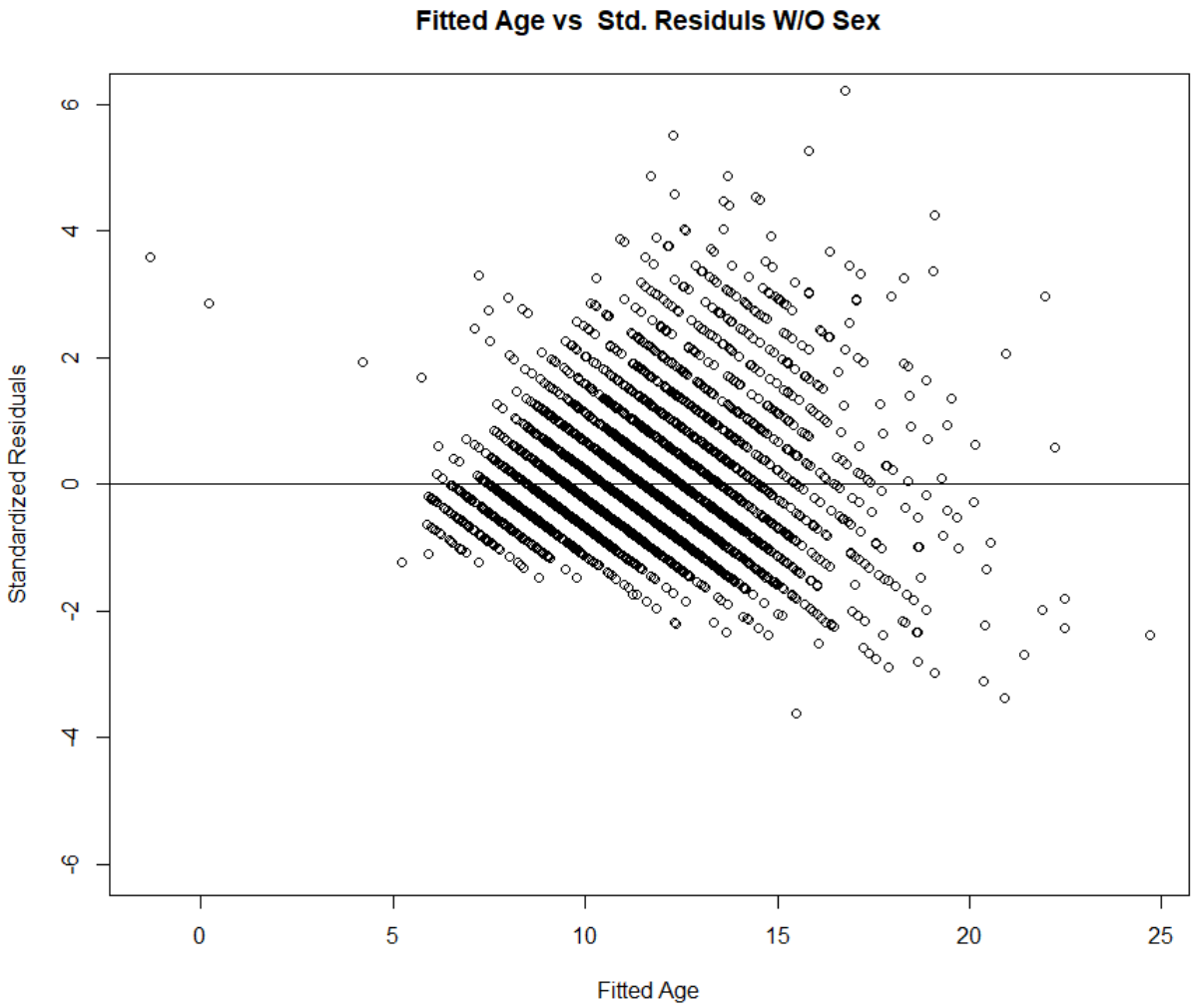

Fig 1. Standardized residuals in original system plotted against predicted Age (rings + 1.5) values using all features but sex, all points

(b) Build a linear regression predicting the age from the measurements, including gender. There are three levels for gender; I'm not sure whether this has to do with abalone biology or difficulty in determining gender. You can represent gender numerically by choosing 1 for one level, 0 for another, and -1 for the third. Plot the residual against the fitted values.

$R^2 = 0.5278909$

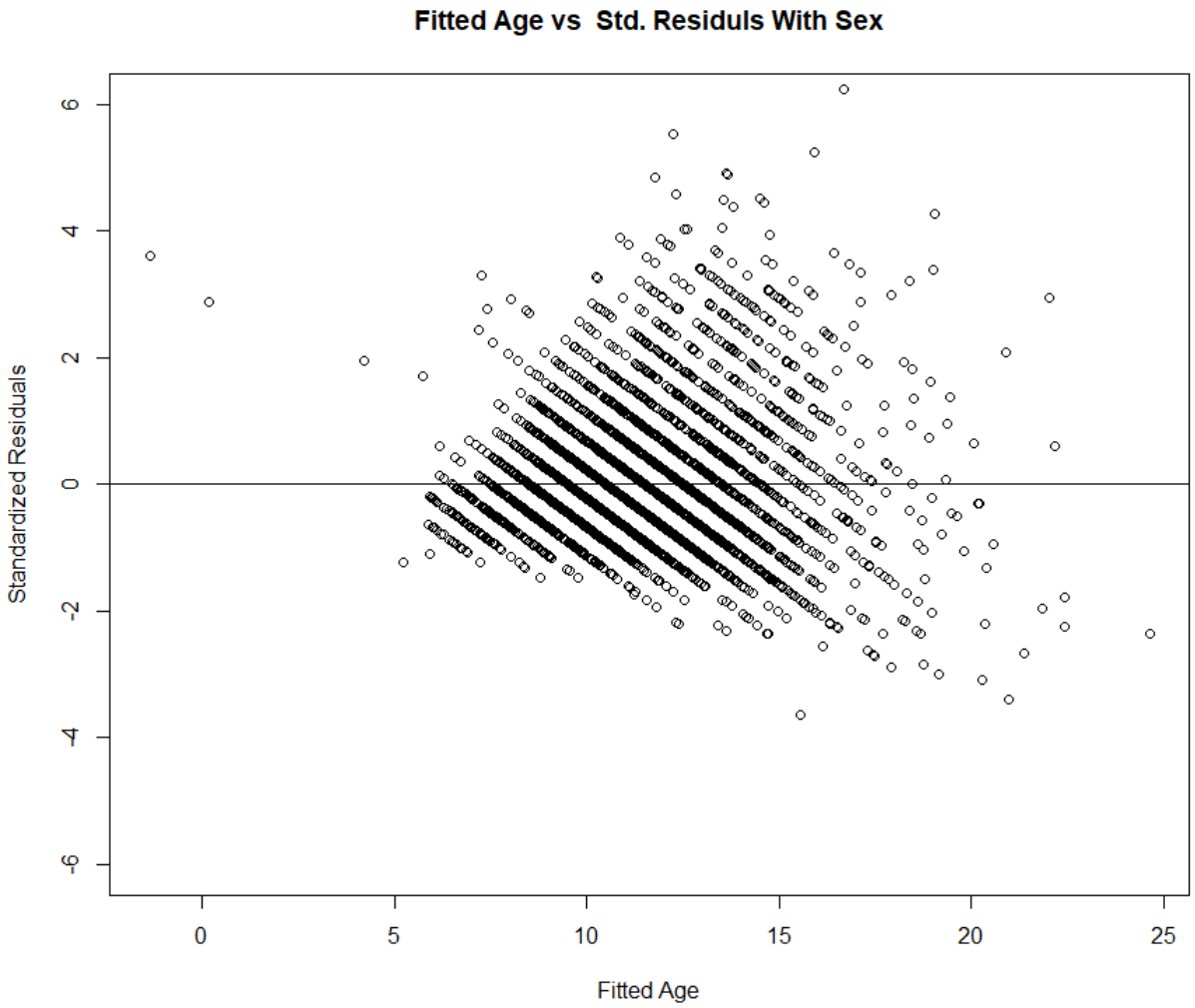

Fig 2. Standardized residuals in original system plotted against predicted Age (rings + 1.5) values using all features, all points

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

R^2 = 0.5796935

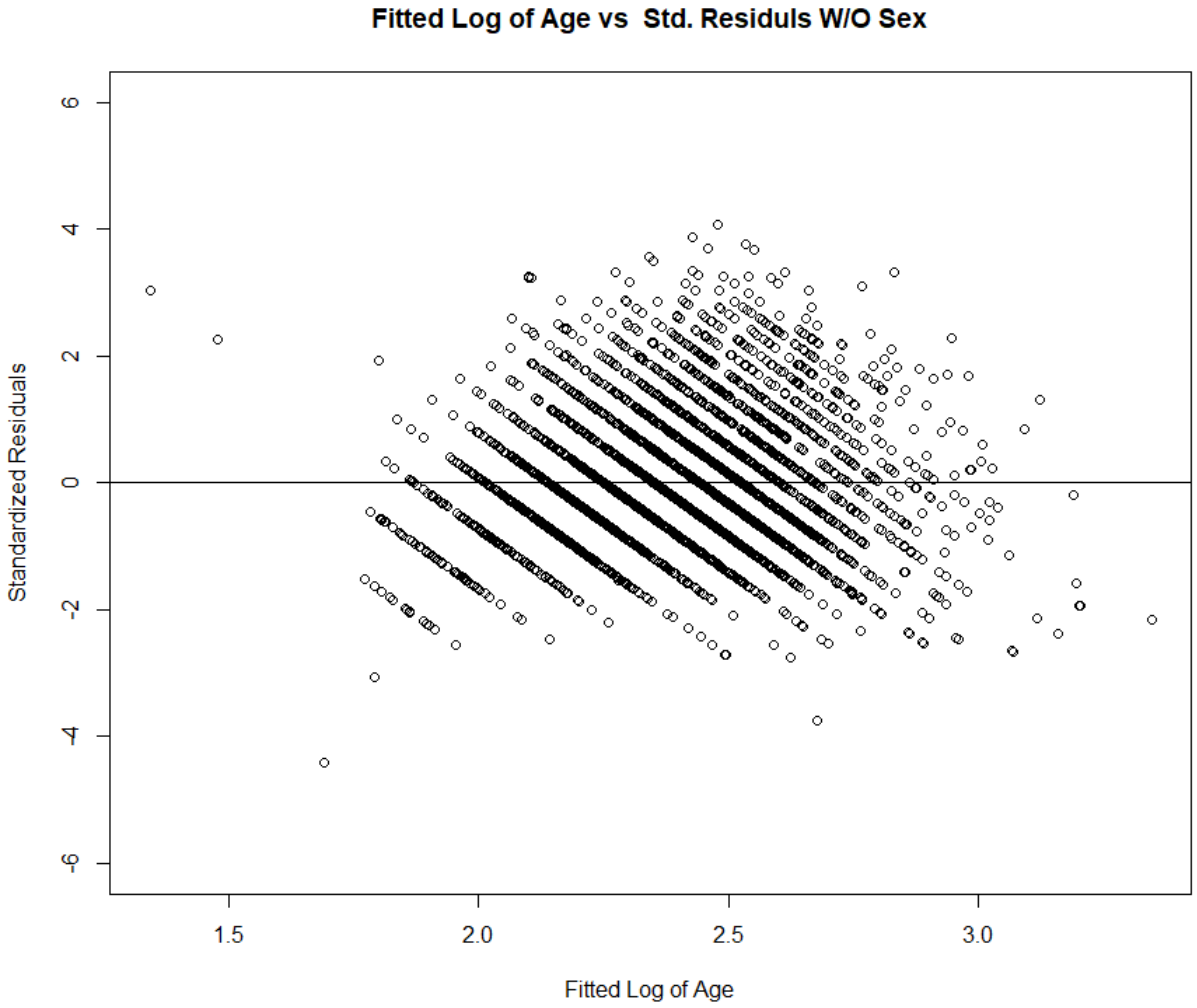

Fig 3. Standardized residuals obtained from a linear regression model predicting the log of age (residuals in log system) plotted against predicted Age (rings + 1.5) values using all features but sex, all points

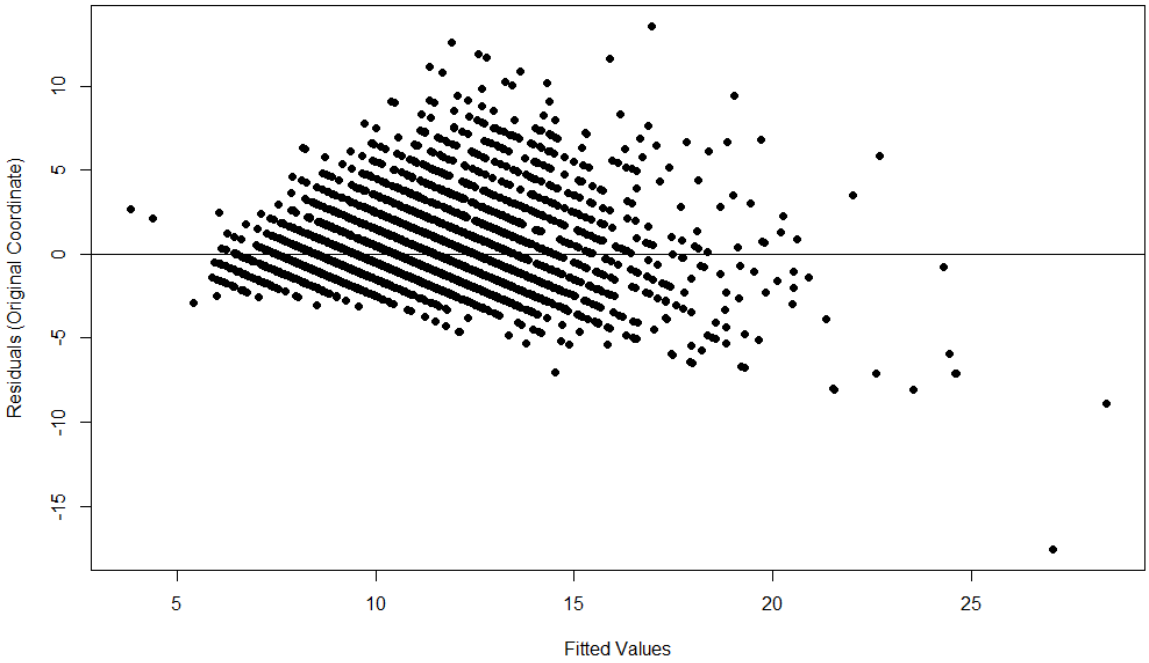

Fig 4. Residuals obtained from a linear regression predicting the log of age converted back to original system plotted against predicted Age (rings + 1.5) values using all features but sex, all points

(*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.*

R^2 = 0.5801268

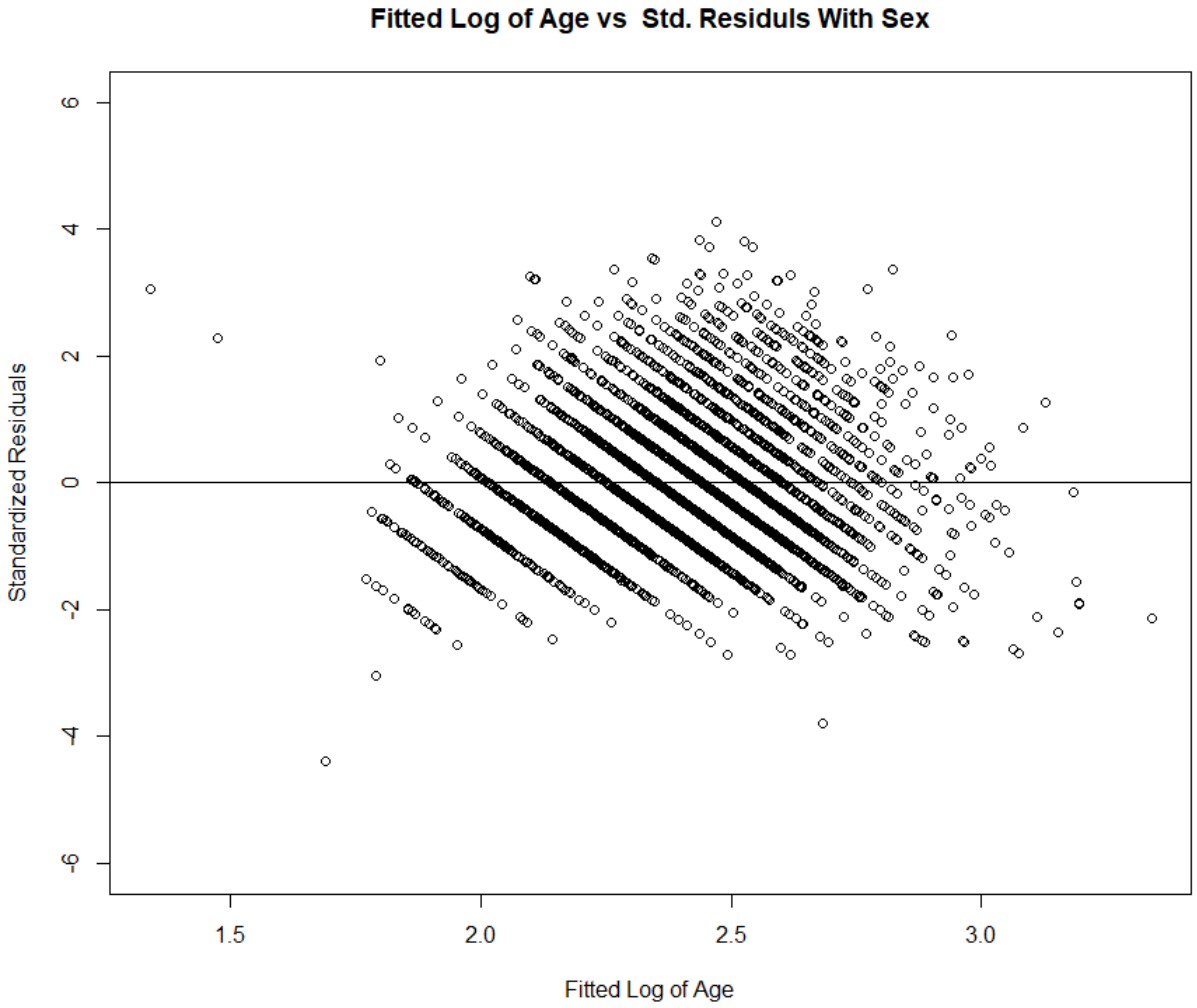

Fig 5. Standardized residuals obtained from a linear regression model predicting the log of age (residuals in log system) plotted against predicted Age (rings + 1.5) values using all features, all points

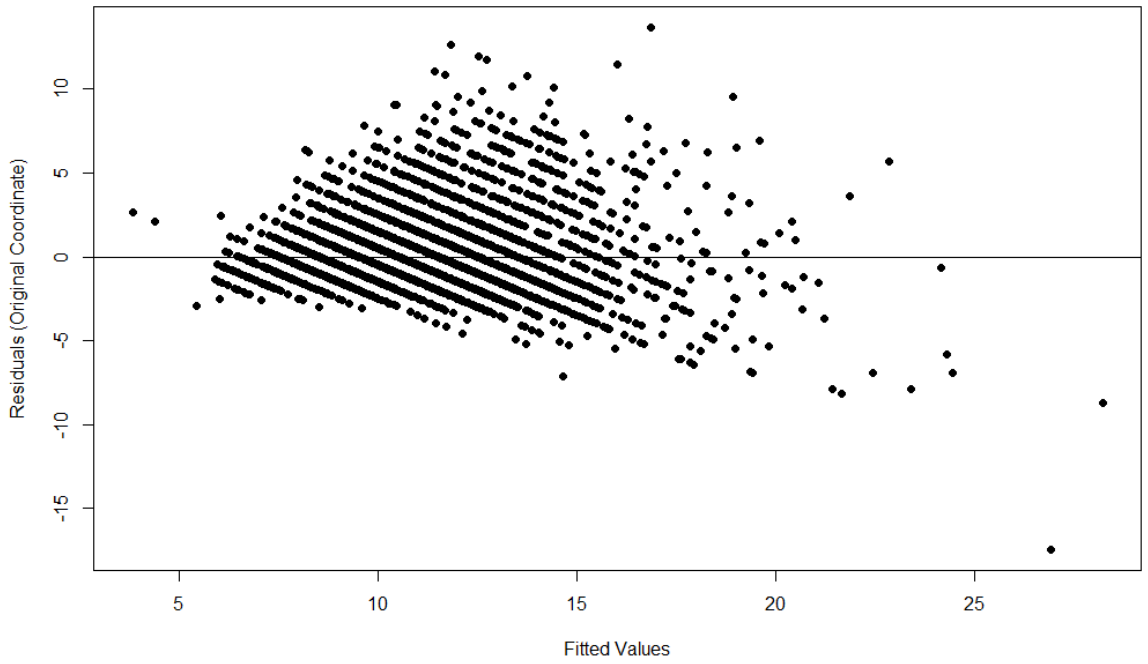

Fig 6. Residuals obtained from a linear regression predicting the log of age converted back to original system plotted against predicted Age (rings + 1.5) values using all features, all points

(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.

Comparing the obtained R-squared indices, it can be concluded that the linear regression models developed based on the log of the age from all measurements is a more accurate model for determining age of an abalone.

Figs 1 and 2 show the standardized residuals plotted against fitted values in original system. Figs 3 and 5, show the standardized residuals of log of ages plotted against fitted values (log system). In case of Figs 4 and 6, linear models are developed using log of ages, fitted values of log of ages are converted back to original unit system (ages and not log of ages) and then residuals are calculated and plotted against the fitted values. In case of all these figures, the mean of the residual seems to be zero. However, in case of Figs 1 and 2, residuals are partially correlated with the fitted values and the variance of the residual seems to be dependent on the predicted value. Where in case of Figs 3 to 6, the residuals are uncorrelated to the predicted and the variance of the residuals are much less dependent on the fitted values.

All, these suggest that the linear regression models developed based on the log of the ages are more fitting and expected to yield more accurate predictions.

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

Using Glmnet() command of r and assuming alpha = 0 (ridge regression) a generalized linear model is fitted. Figs. 7 to 10 below show Mean-squared error for cross validated regularization of the linear models developed in Sections a to d. The best lambda value obtained from these figures is 0.22 for linear models predicting ages in original system (not log) and 0.02 for linear models predicting log of ages. The lambda values, specifically for liner model of log of ages, are very close to zero and consequently the regularization does not have significant effect on correcting and improving the linear model. Also, the R-squared indices are calculated for regularized linear models predicting ages and log of ages and obtained R-squared indices has not improved. Therefore, in case of linear models predicting ages and log of ages of an abalone tree, regularization will not improve regressions.

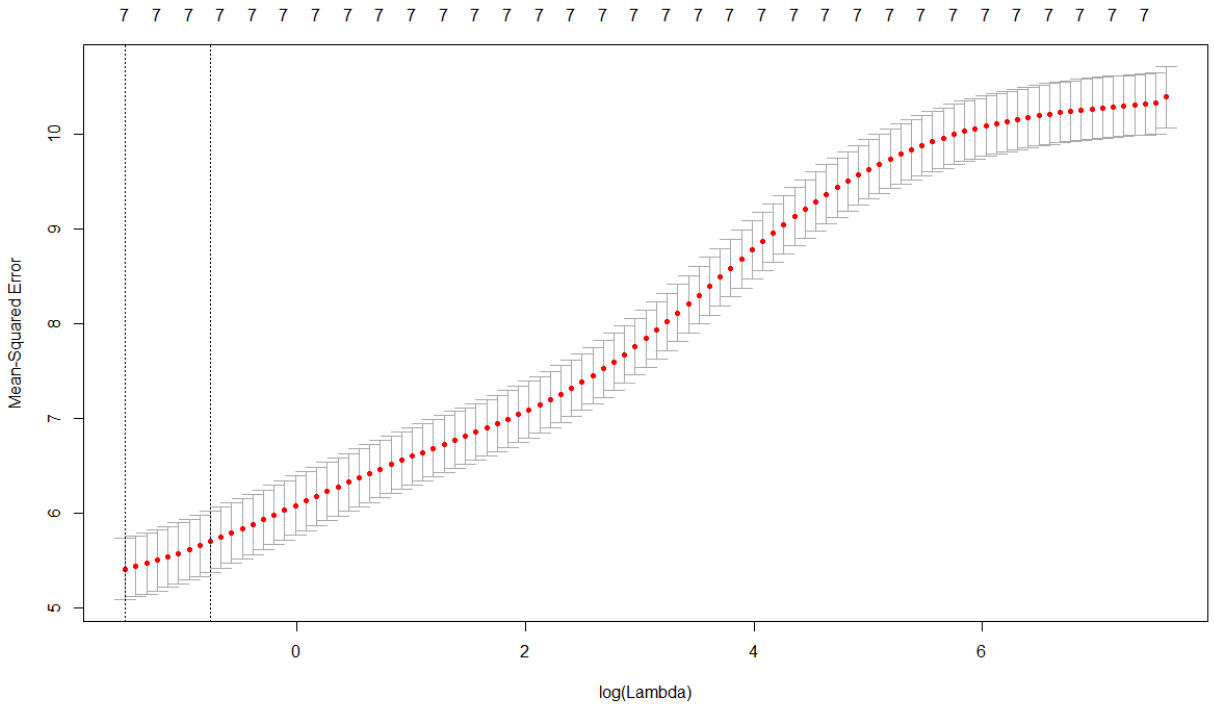

Fig 7. Mean-squared error for various Lambdas, Age (rings + 1.5) against all features other than age, all points

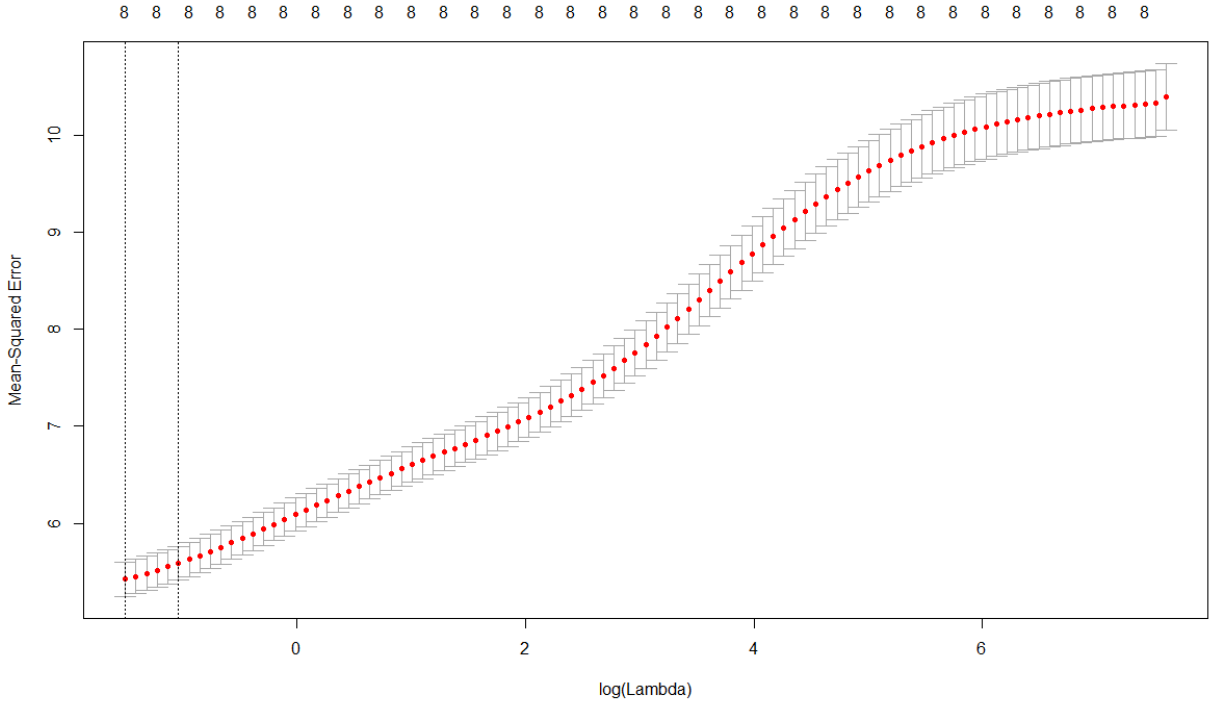

Fig 8. Mean-squared error for various Lambdas, Age (rings + 1.5) against all features, all points

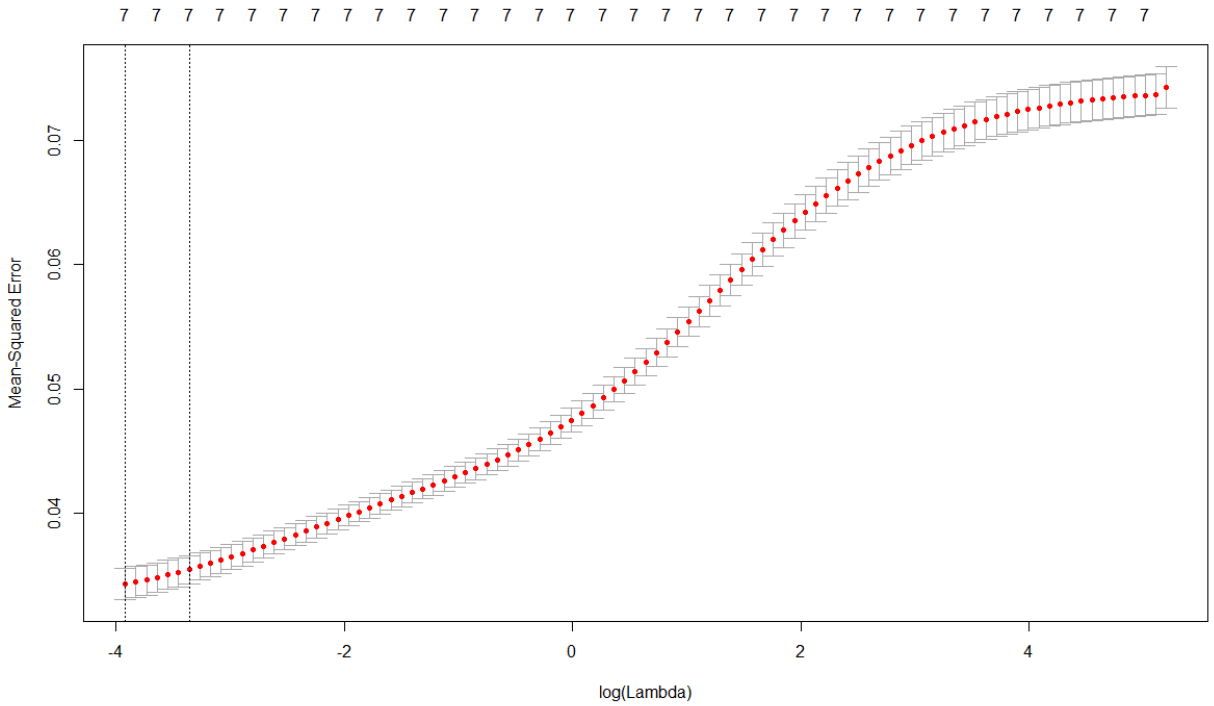

Fig 9. Mean-squared error for various Lambdas, Log of age (rings + 1.5) against all features other than age, all points

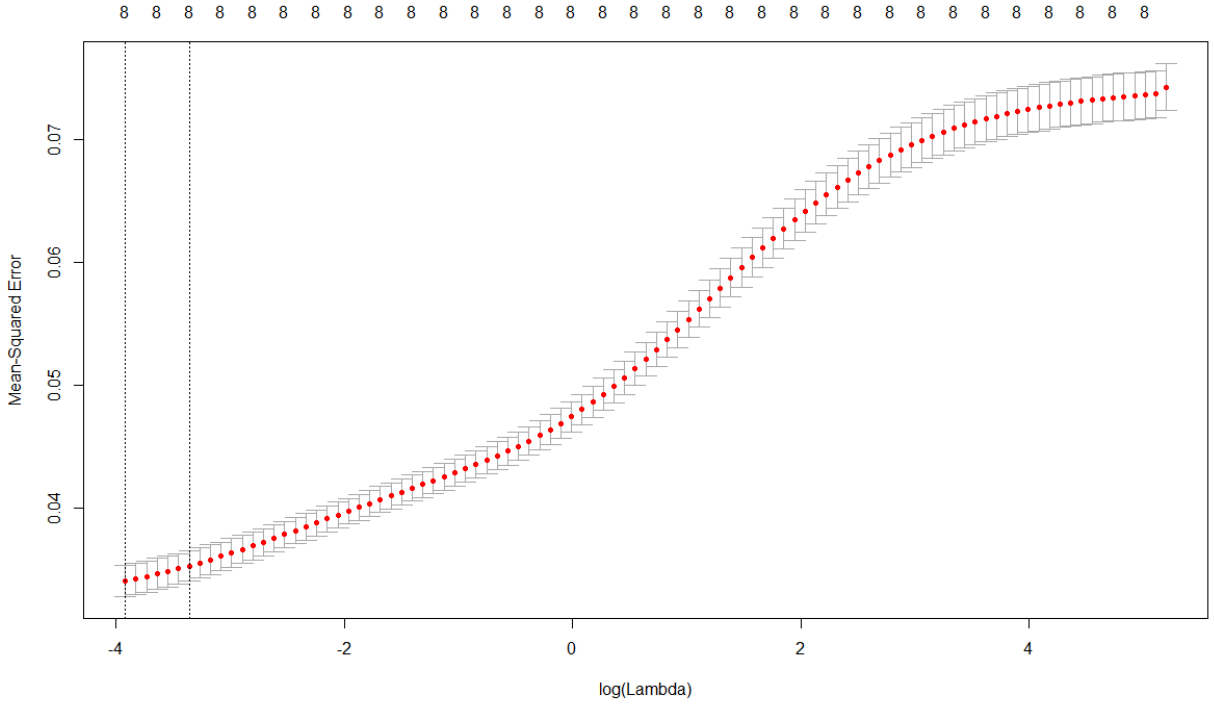

Fig 10. Mean-squared error for various Lambdas, Log of age (rings + 1.5) against all features, all points