Assignment 5

CS498 Applied Machine Learning

Jianshu Wang , Hari Manan, Jasdeep Duggal- March 12, 2018

Introduction

Problem7.9

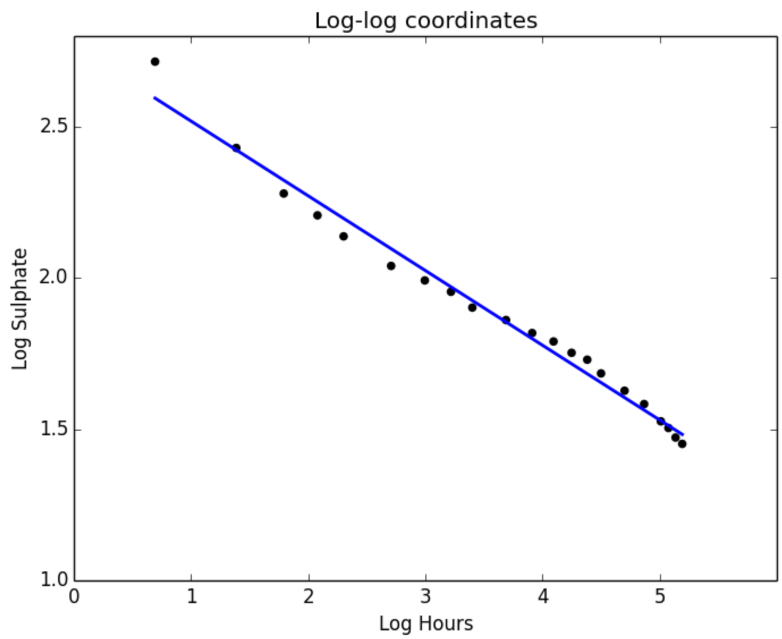

Figure 1 (A) : Linear Regression in log-log Coordinates

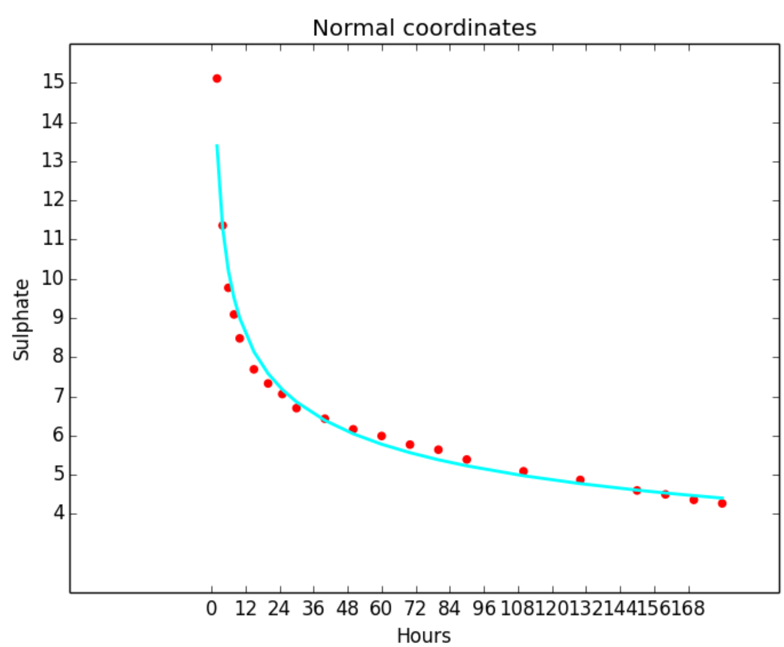

Figure 2 (B) : Curve Line in Normal Coordinates

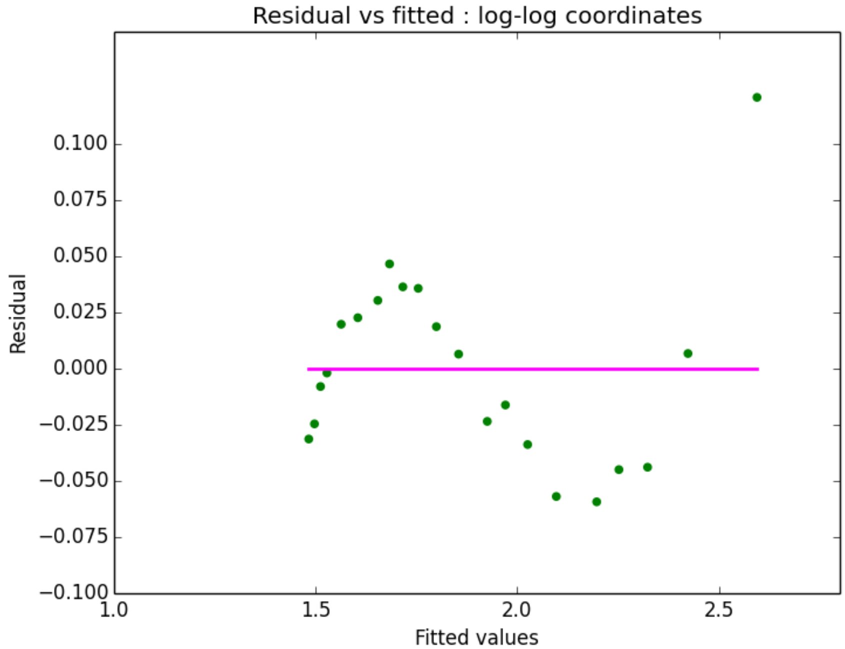

Figure 3 (C – A): Residual againt Fitted Value in log-log Coordinates

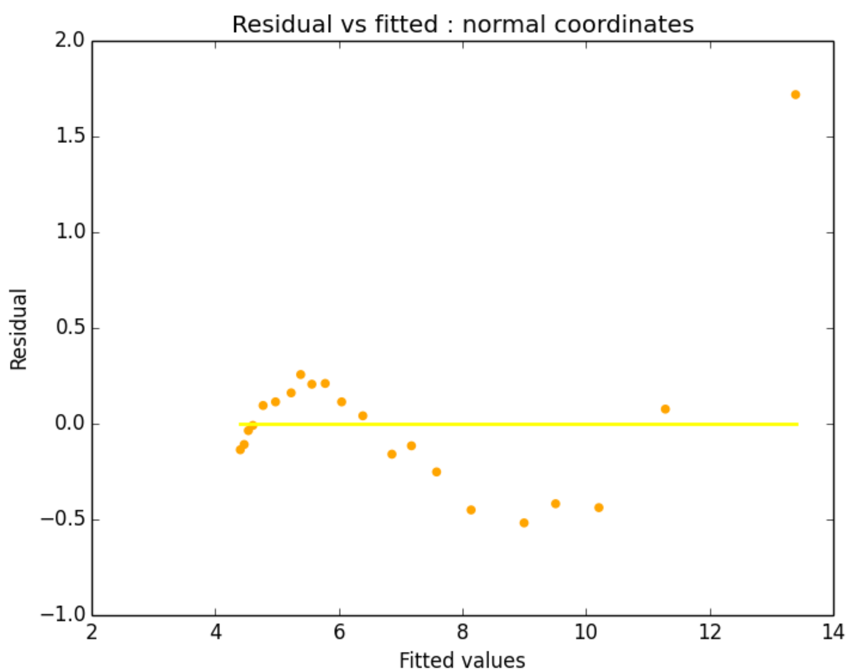

Figure 4 (C -B): Residual against Fitted Value in Normal Coordinates

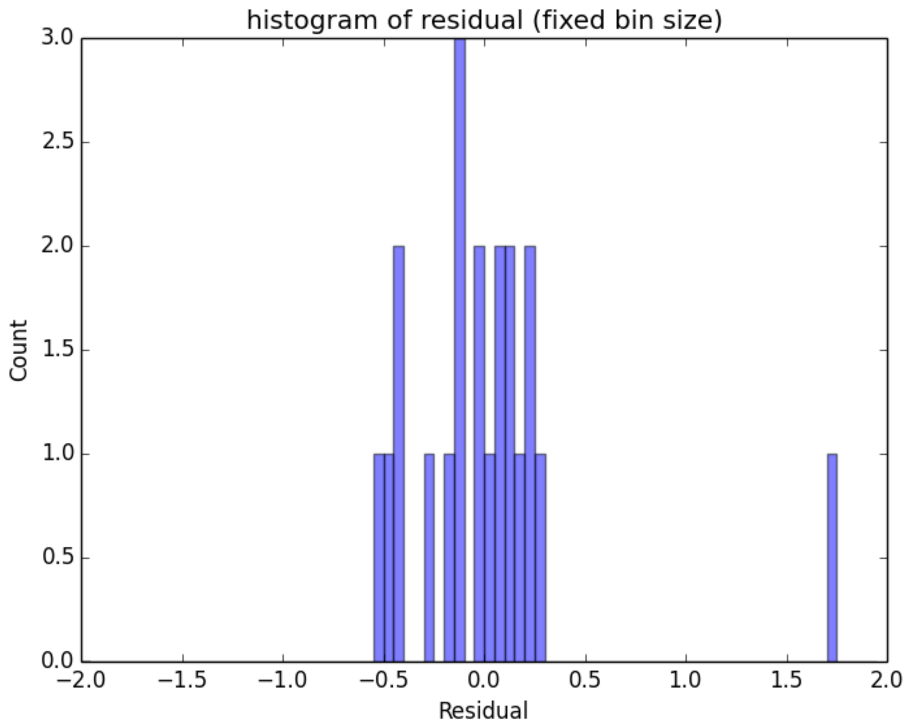

Figure 5 (C Extra): Histogram of Residual Like a Normal Distribution

Let's us calculate variance code of the model , the variance value for our model is 0. 9839250931007382 .The best values is 1 , our model is 0.9839250931007382 which is very close to best And also, if we look at the figure 2 , the values of hours vs . Sulphate in normal coordinates , we can infer that the plots almost coincide with the scattered values, showing that it is a good regression.

Residual plots are considered to be good if they follow these 3 rules.

Figure 3 and 4 , the residual values in both the log and normal coordinates follow these rules.

(1) they're pretty symmetrically distributed, tending to cluster towards the middle of the plot

(2) they're clustered around the lower single digits of the y-axis (e.g., 0.5 or 1.5, not 30 or 150)

(3) in general there aren't clear patterns

Source :

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/

Problem 7.10

Part a)

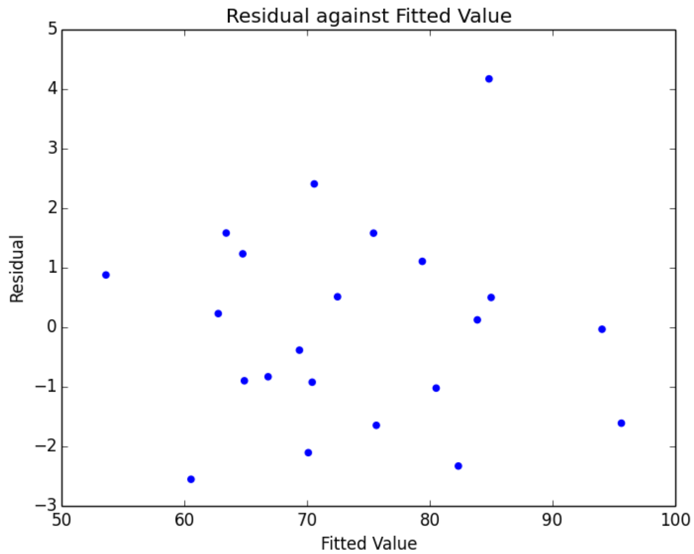

Figure 6: Residual against Fitted Value

Part b)

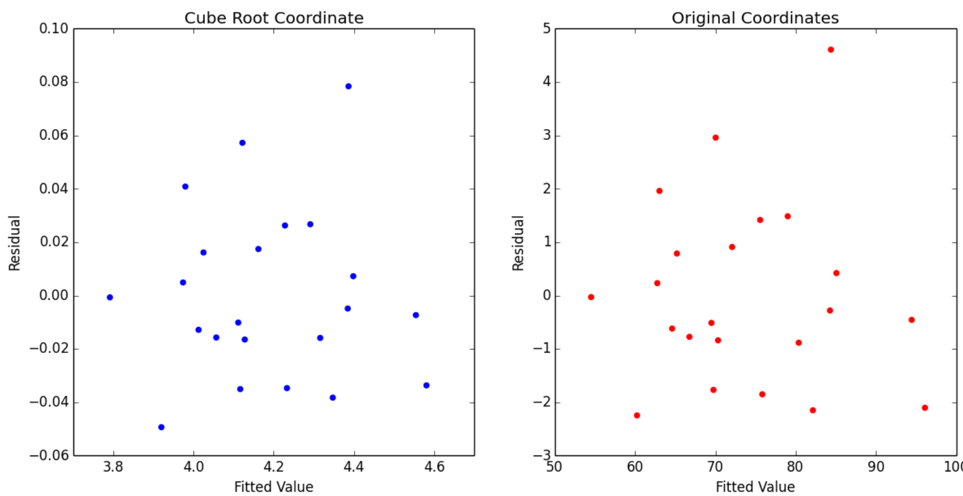

Figure 7: Cube Root Residual in Cube Root Coordinates and Original Coordinates

Part c) R squares are 0.977210661741 for normal vs 0.984234958257 for cube root, so I consider cube root to be a better one

Problem 7.11

(a) Linear regression predicting the age from the measurements, ignoring Gender

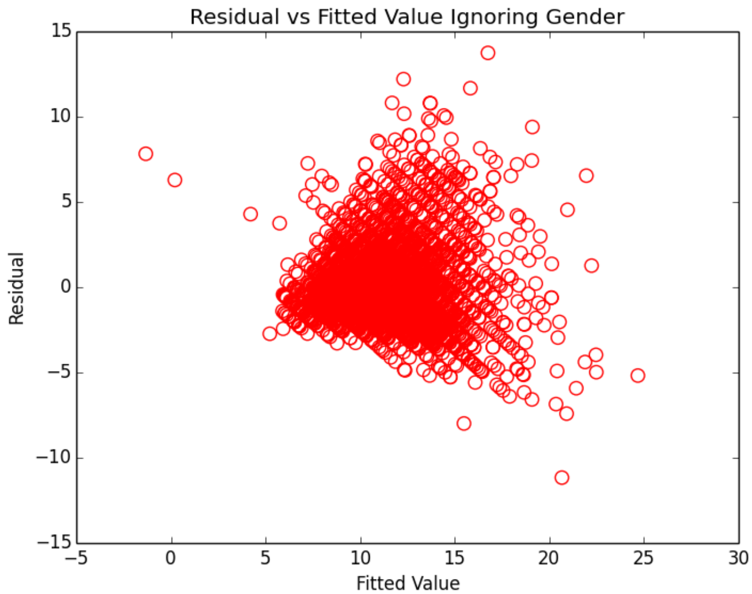

Figure 8: Residual against Fitted Value with no Gender

b) Linear regression predicting the age from the measurements, including gender.

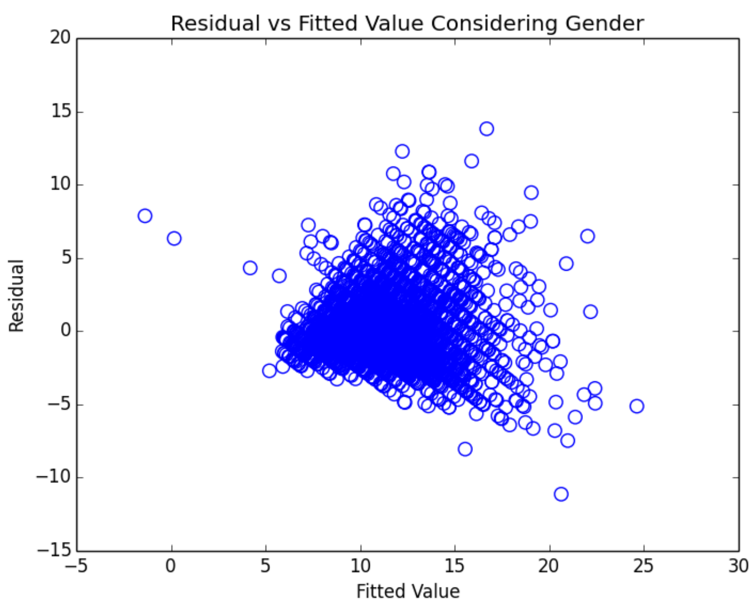

Figure9: Residual vs Fitted Value with Gender

c) Linear regression predicting the log of age from the measurements, ignoring gender.

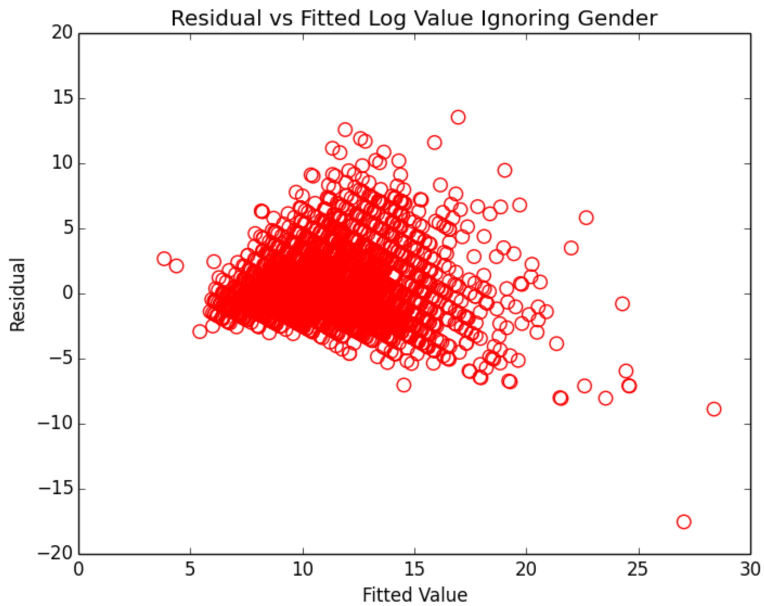

Figure 10: Log of age Residual against Fitted Value without Gender

d) Linear regression predicting the log age from the measurements, including gender,

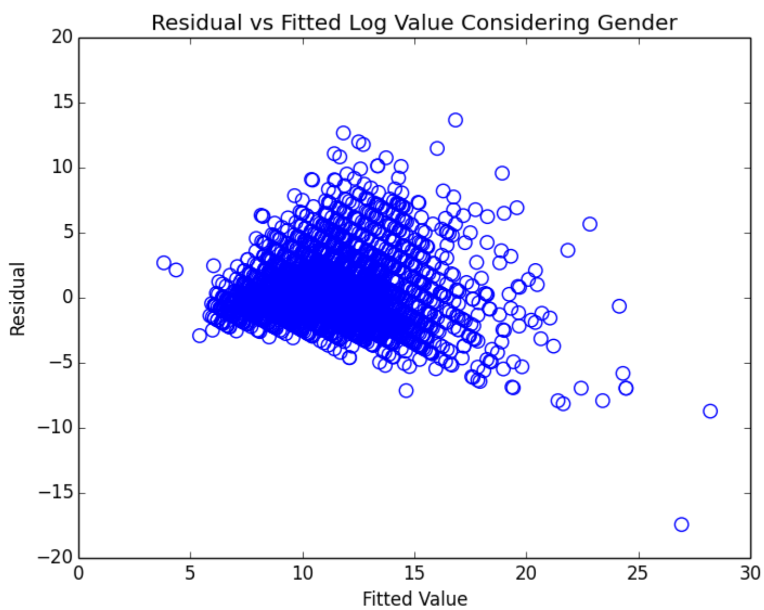

Figure 11: Log of age Residual against Fitted Value with Gender

e) Use your plots to explain which regression you would use to replace this procedure, and why.

Calculating R Square:

The R square values for the above plots are as follows:

a. R square : 0. 527629939992

b. R square : 0. 527890935736

c. R square : 0. 544422151851

d. R square : 0. 544301293342

As the highest value of R square is for the case c, "Linear regression predicting the log age from the measurements, ignoring gender," I will choose this.

f) Can you improve these regressions by using a regularizer?

Here I used sklearn.linear_model ElasticNet library and sklearn.linear_model regressionlibrary

We tried to compare regular Regression vs Regularization results when alpha = 0 and following is our observation:

- Decreasing lambda value makes the Regularization result approach Regression results, but we could not get a better result. This does not necessarily mean that a better result is not possible. It can also mean that the good lambda range is too narrow that we cannot catch it.

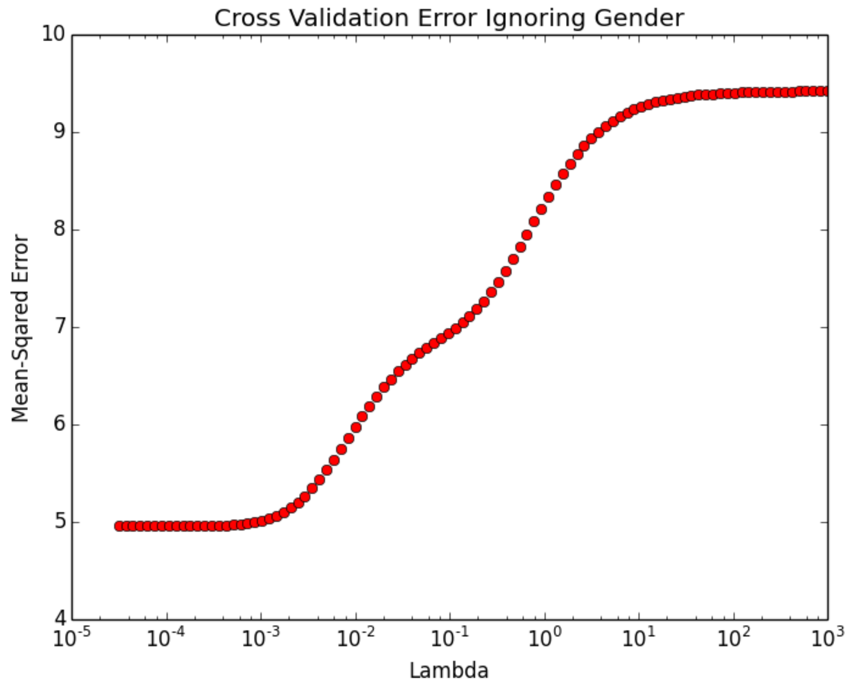

Figure 12: Cross Validation of Regularization without Gender

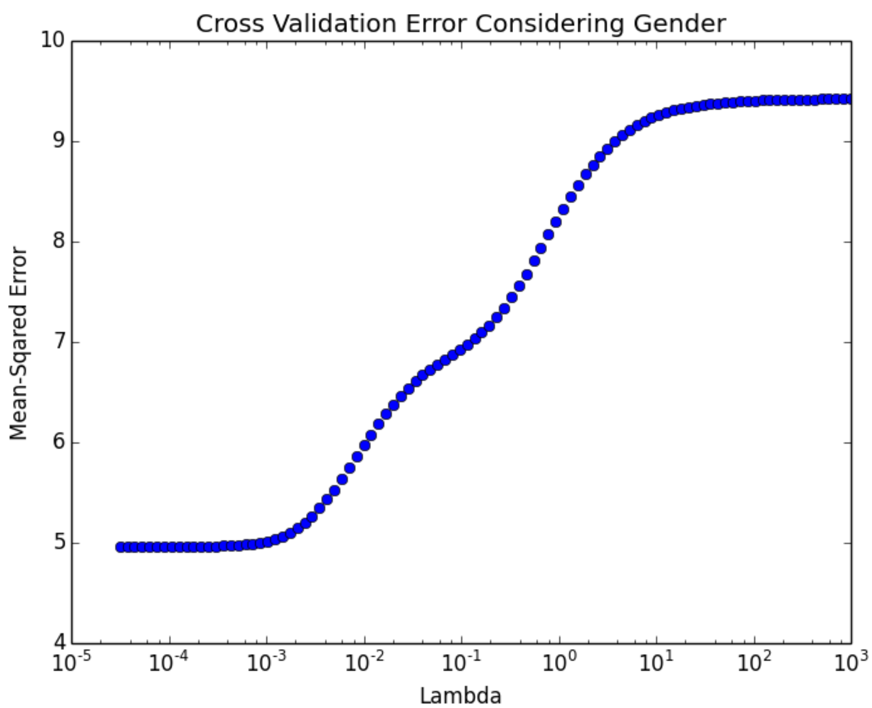

Figure13: Cross Validation of Regularization with Gender

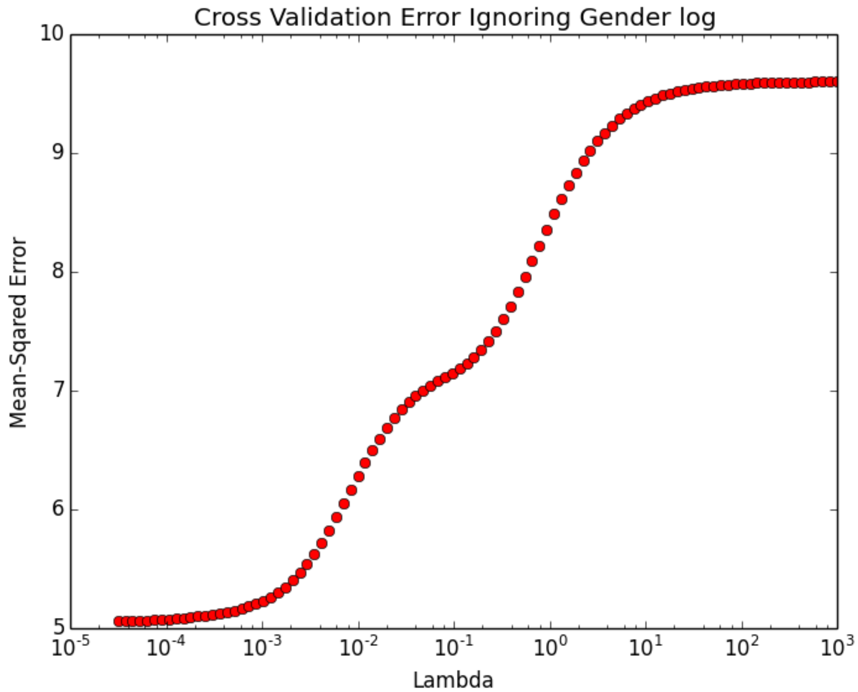

Figure 14: Cross Validation of Regularization without Gender and log of age

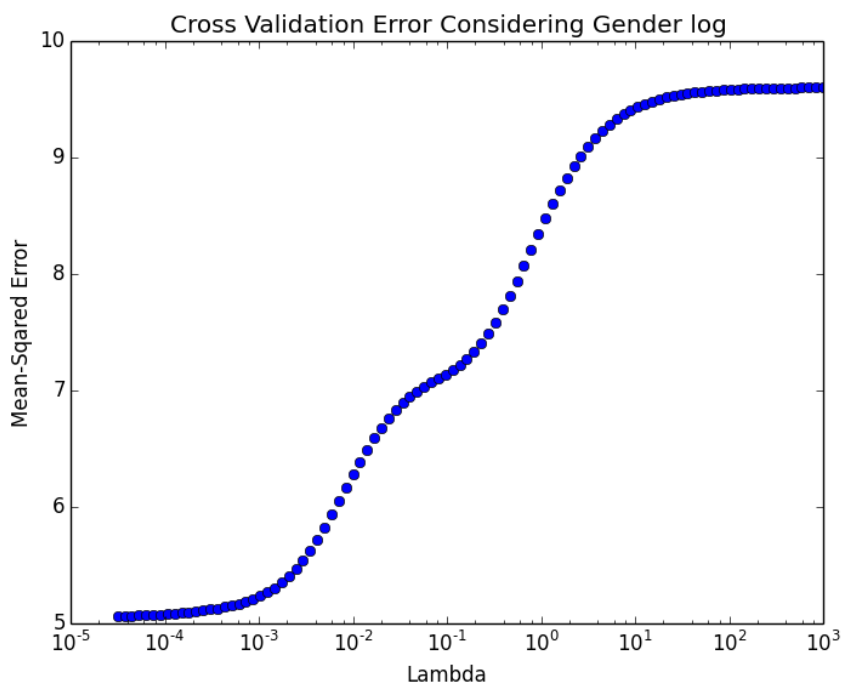

Figure 15: Cross Validation of Regularization with Gender and log of age