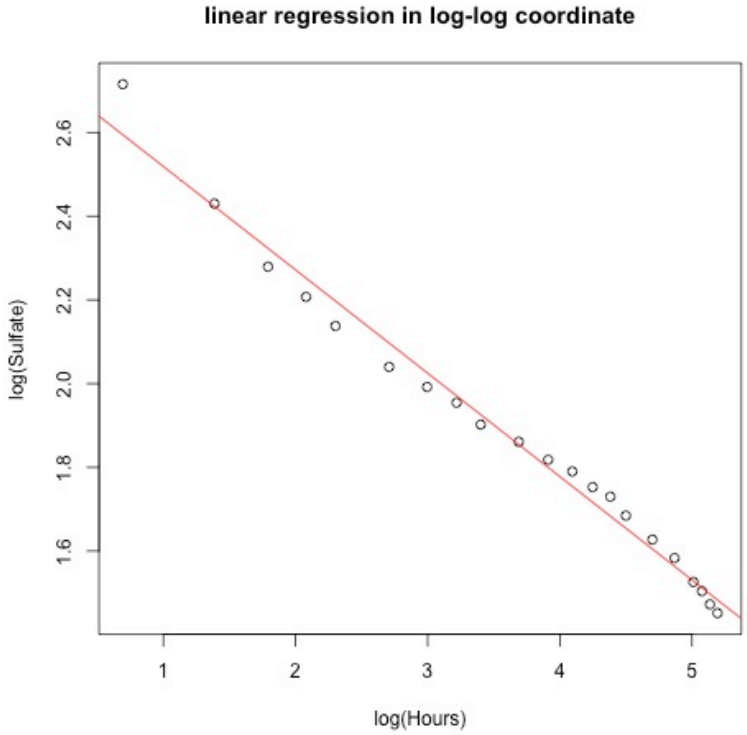
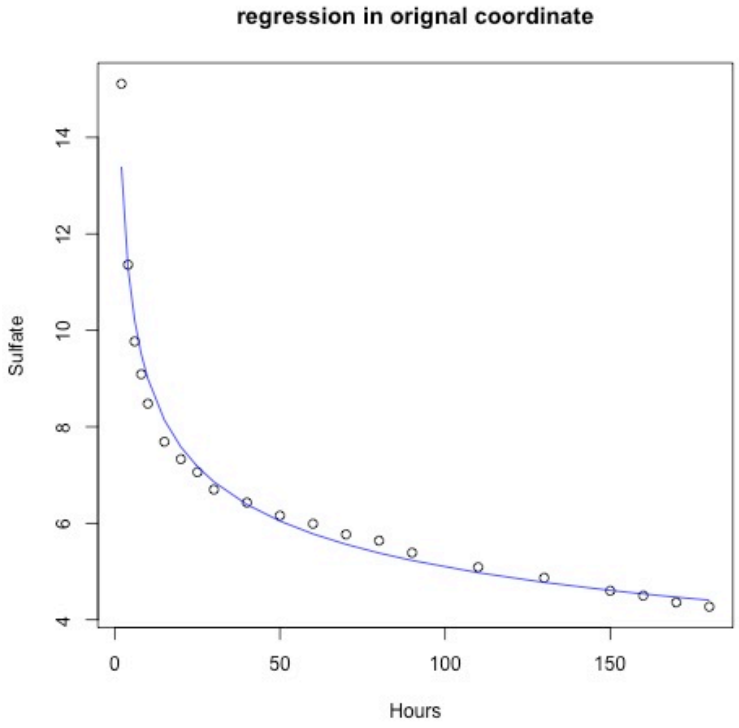
Homework 5 Report

7.9

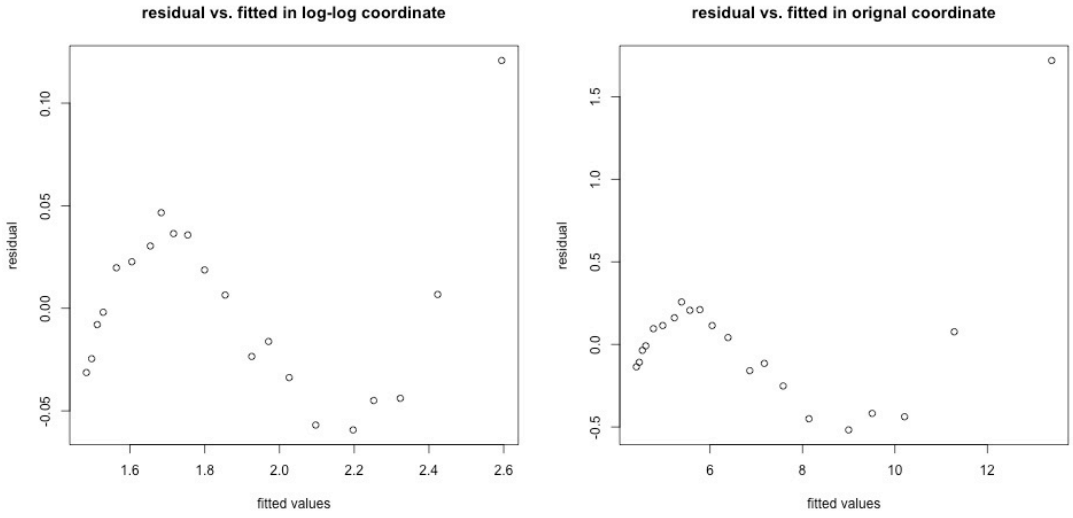
(a)



(b)



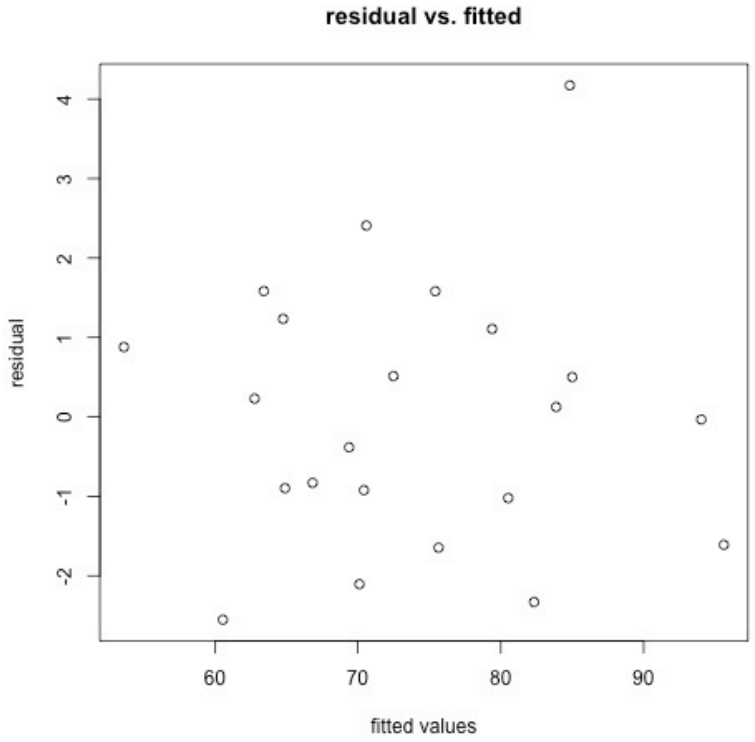
(c)



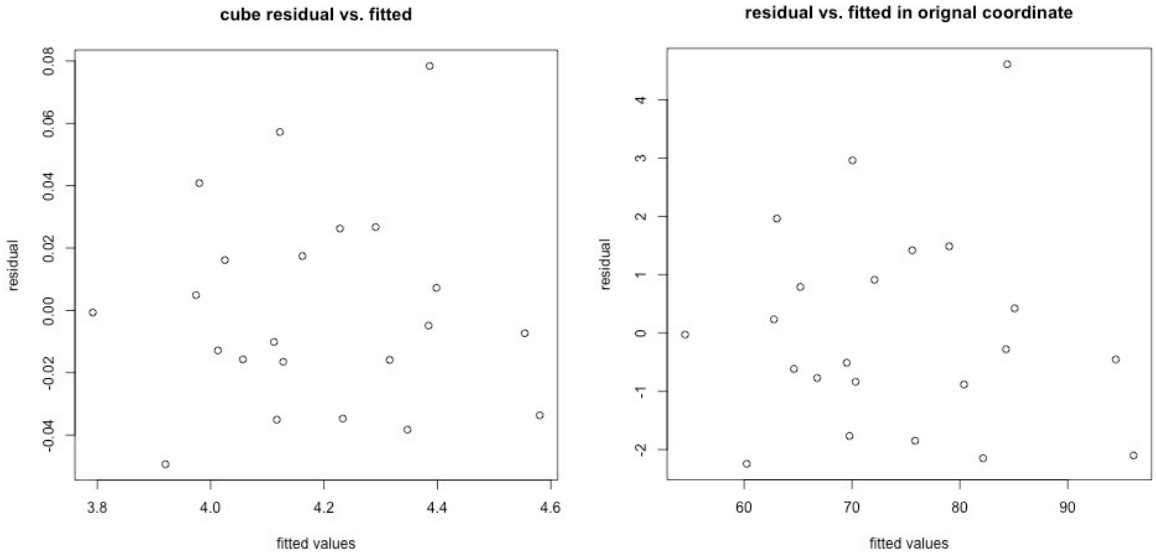
(d) In my opinion the regression model is not very good. Even though the line seems fit those data point, however the plots of residual vs. fitted value shows fluctuations with pattern, indicates that a non-linear regression might be a better model for this dataset.

7.10

(a)



(b)



(c) It is hard to tell which regression is better solely depend on the residual vs. fitted value plots because they are all very similar. Coefficients showing below indicates that regression on original Mass might be slightly better because coefficients are minimal on more variables, means a simpler model.

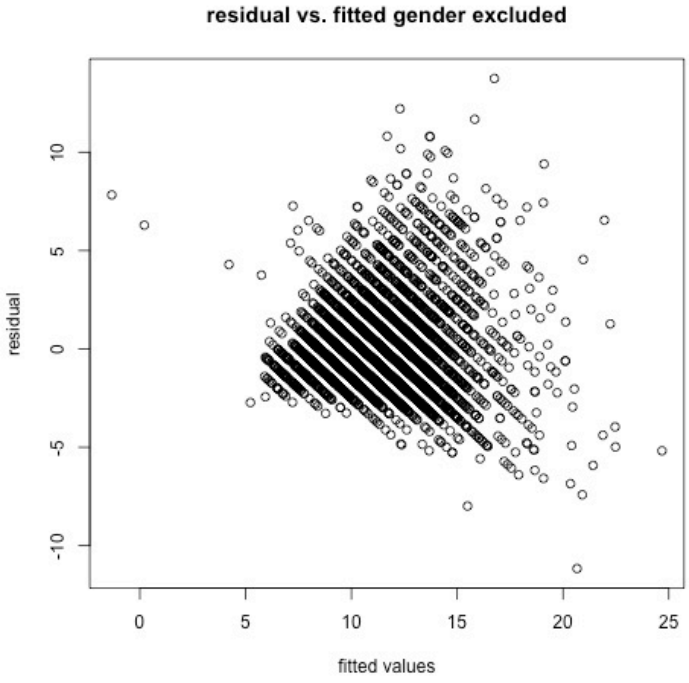
Linear regression coefficients of original Mass:

(Intercept) Fore Bicep Chest Neck Shoulder
1.119228782 0.027971894 0.004143713 0.001051887 -0.002532061 0.000810027
Waist Height Calf Thigh Head
0.011152328 0.005773831 0.010656455 0.007918841 -0.012452012

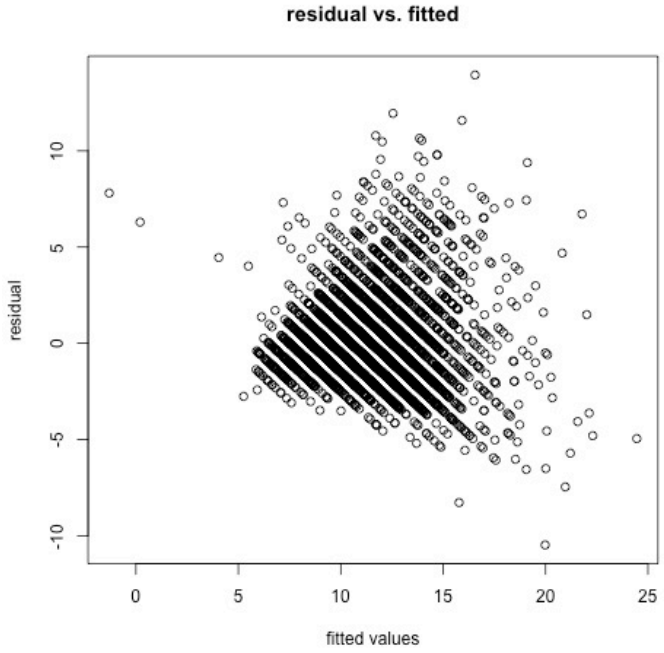
Linear regression coefficients of cubed Mass:
(Intercept) Fore Bicep Chest Neck Shoulder
-69.51713512 1.78181867 0.15509040 0.18913544 -0.48183705 -0.02931235
Waist Height Calf Thigh Head
0.66144124 0.31784645 0.44589018 0.29721231 -0.91956267

7.11

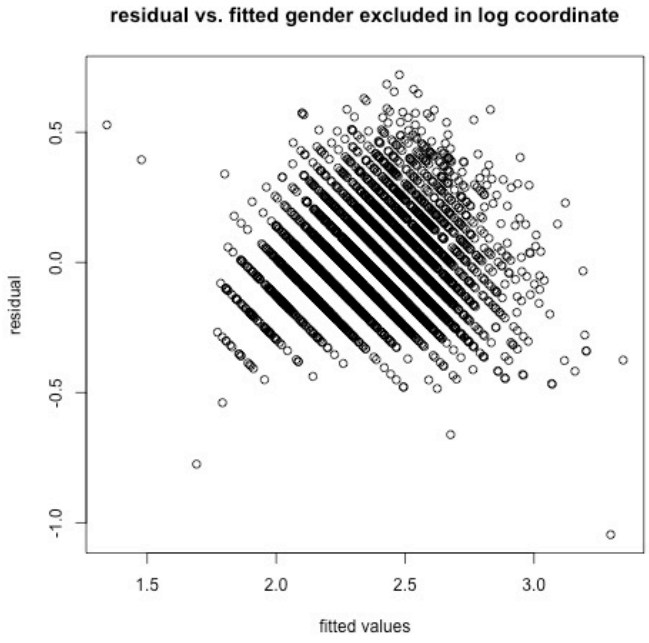
(a)



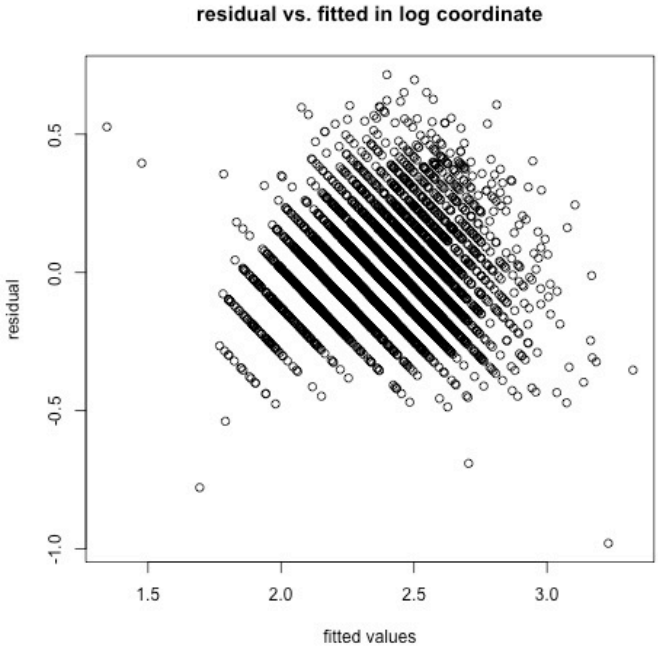
(b)


residual vs. fitted
residual
fitted values
10
5
0
-5
-10
0
5
10
15
20
25

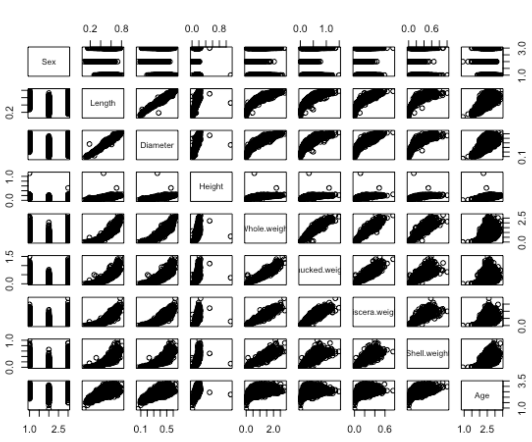
(c)

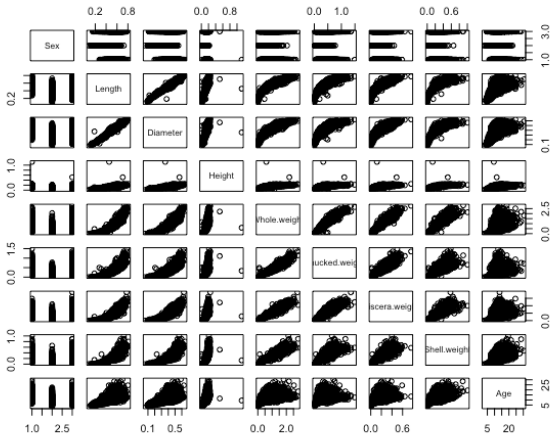

residual vs. fitted gender excluded in log coordinate
residual
fitted values
0.5
0.0
-0.5
-1.0
1.5
2.0
2.5
3.0

(d)

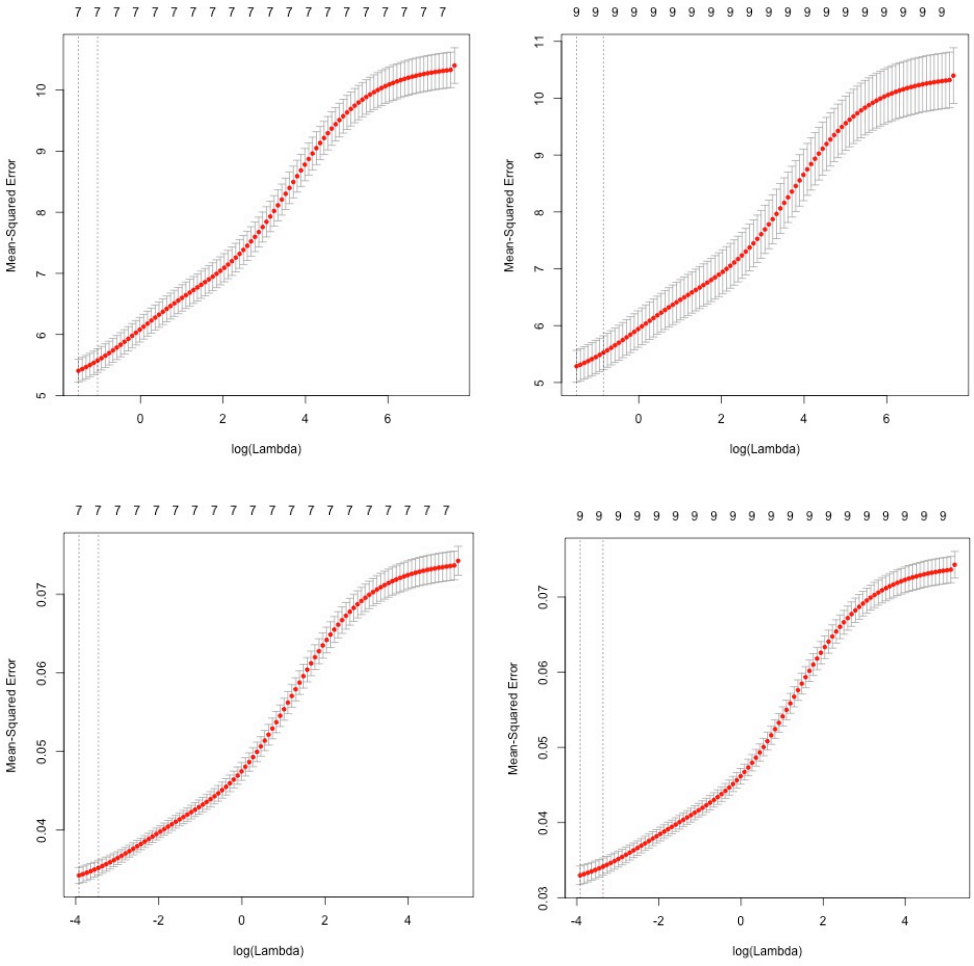


(e) For this dataset, I would prefer using linear regression model in (A). Mainly for 3 reasons:
1. residual plots of model (A) and (B) shows more balanced distribution, whereas model (C) and (D), more data points have higher residuals especially when fitted values are larger.
2. residual plots of (A) and (B) are very similar, means that gender variable does not contribute much to the linear model, therefore I think (A) is better than (B) as a simpler model.
3. I looked at log(Age) against other variables(left) and Age against other variables(right), the right plot shows Age is more linear related to other variables than log(Age), this is in congruence with residual plots observation stated in my first point.





(f) The Mean-Squared Error plots for the 4 models show that a regularizer might not be able to improve the regression model.



Reference

1. example code of transform log-log coordinates back to original
https://stackoverflow.com/questions/46392683/how-to-plot-transformed-regression-back-on-original-scale
2. explanation of residual plots
https://onlinecourses.science.psu.edu/stat501/node/279
3. glmnet documentation
4. Slack discussions