CS498 Homework 4

Xinchen Pan, Fangxiaozhi Yu, Jiefei Shi

March 2, 2017

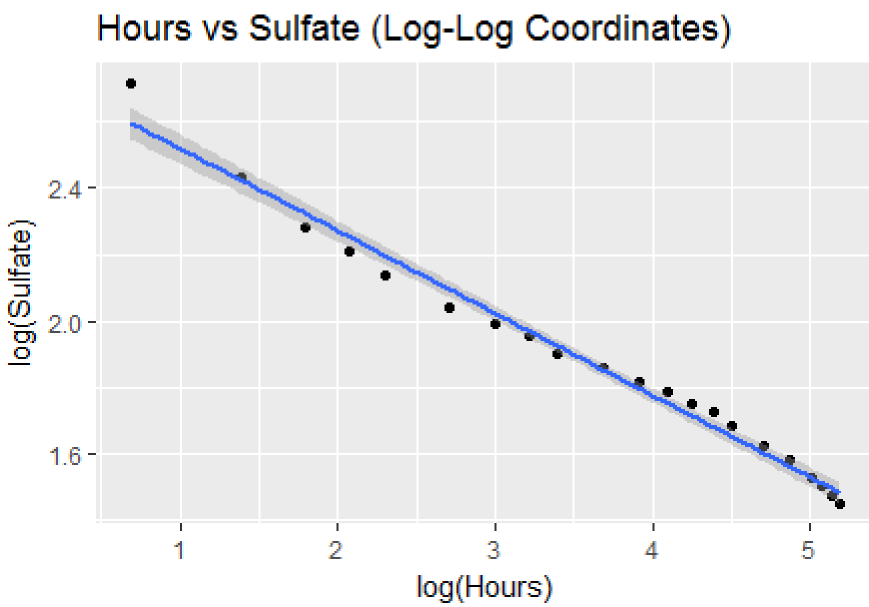
7.9

The dataset we used has 21 observations and two variables. `Sulfate` is our response variable and `hours` is the predictor. We built a simple linear regression model of the log of the `Sulfate` against the log of `Hours`.

```
##
## Call:
## lm(formula = log(Sulfate) ~ log(Hours), data = dt)
##
## Coefficients:
## (Intercept)   log(Hours)
##       2.766       -0.247
```

(a)

Below is the plot showing the data points and the regression line in log-log coordinates.



(b)

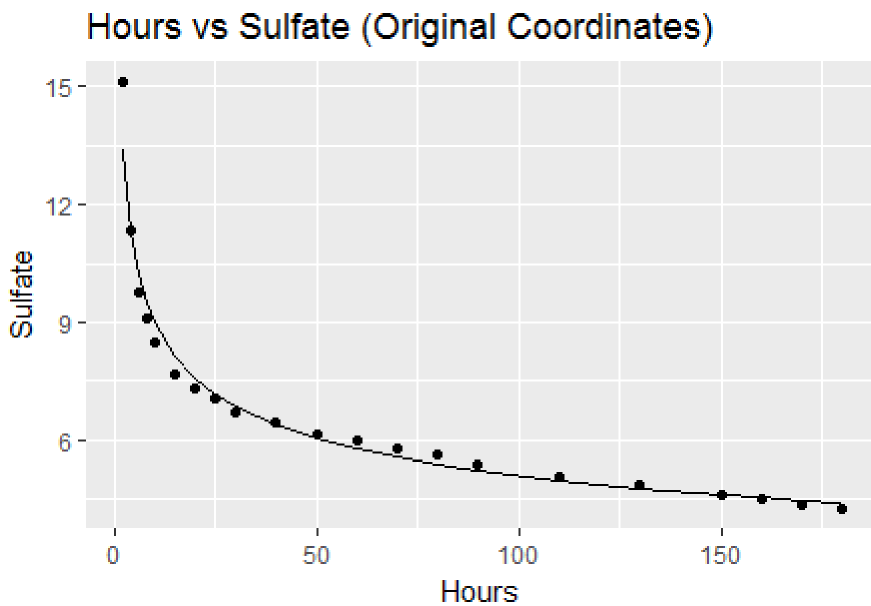
To show the data points and regression curve in the original coordinates, I did the following transformation.

$$\widehat{\log(\text{Sulfate})} = -0.247 * \log(\text{Hours}) + 2.766$$

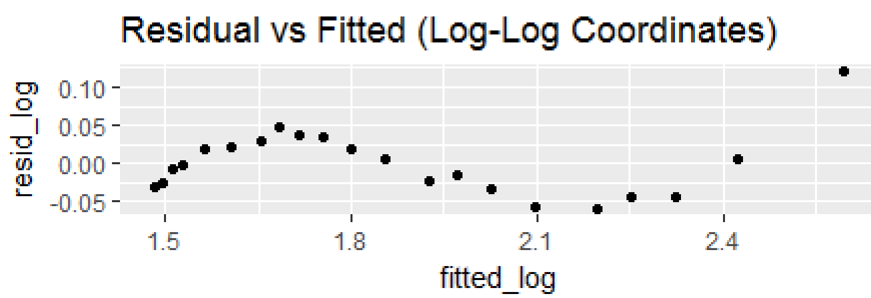
$$\exp(\widehat{\log(\text{Sulfate})} = \exp(-0.247 * \log(\text{Hours}) + 2.766)$$

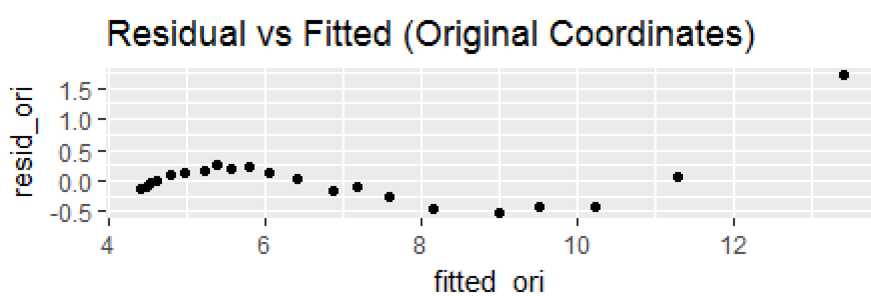
$$\widehat{\text{Sulfate}} = \text{Hours}^{-0.247} * \exp(2.766)$$

Below is the plot showing the data points and the regression curve in original coordinates.



(c) Below is the residual against the fitted values plot in log-log and in original coordinates.





(d)

From the first plot, we can see that the data points are almost all on the regression line. It means that the regression model does a good job in predicting. Checking the second plot, we find that in original coordinates, the data points are also almost all on the regression curve. However, by checking both plots of the **Residual vs Fitted** plot, we can observe a very obvious **pattern** or **trend**. As the fitted value gets larger, the residual will be larger. Thus the model will not do a good job in predicting large values. Thus, our model might **not** be a very good model.

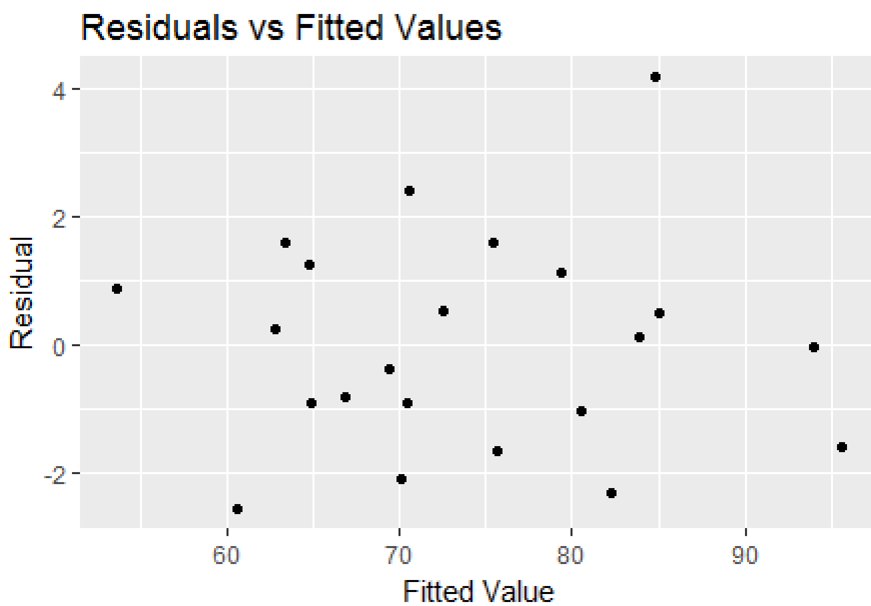
7.10

For this problem, the dataset we used has 22 observations and 11 variables. We were predicting `Mass` using all other predictors.

```
##
## Call:
## lm(formula = Mass ~ ., data = dt2)
##
## Coefficients:
## (Intercept)         Fore        Bicep        Chest         Neck
##   -69.51714      1.78182      0.15509      0.18914     -0.48184
##    Shoulder        Waist       Height         Calf        Thigh
##    -0.02931      0.66144      0.31785      0.44589      0.29721
##        Head
##    -0.91956
```

(a)

Below is the residual against the fitted values plot for the regression model we made.



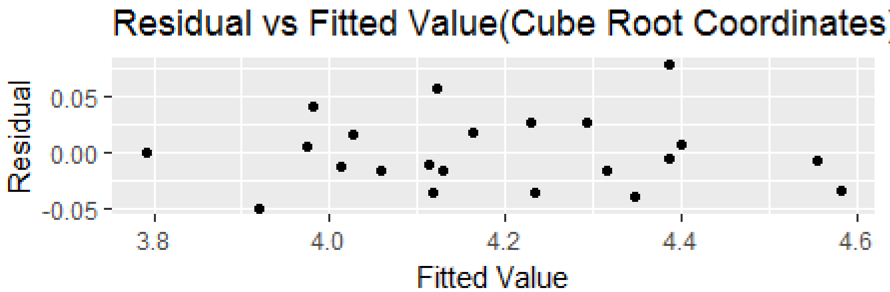
(b)

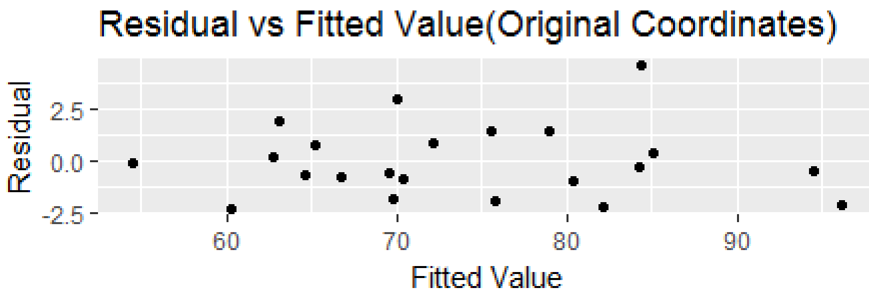
We firstly made the model of the cube root of mass against other parameters.

```
##
## Call:
## lm(formula = Mass^(1/3) ~ ., data = dt2)
##
## Coefficients:
## (Intercept)         Fore        Bicep        Chest         Neck
##    1.119229     0.027972     0.004144     0.001052    -0.002532
##    Shoulder        Waist       Height         Calf        Thigh
##    0.000810     0.011152     0.005774     0.010656     0.007919
##        Head
##   -0.012452
```

To get the residuals in the original coordinates, we did something like this. We cubed the fitted values first, then we used the original values to minus these values.

Below is the residuals again the fitteed values plot in both cube root coordinates and original coordinates.





(c)

Both regression models are good from the **Residual vs Fitted** plots. We can see that the trends in both plots are roughly flat with equal vertical spread. The mean of the residual seems to be 0 in both plots. Thus we believe both models are good and can make good predictions.

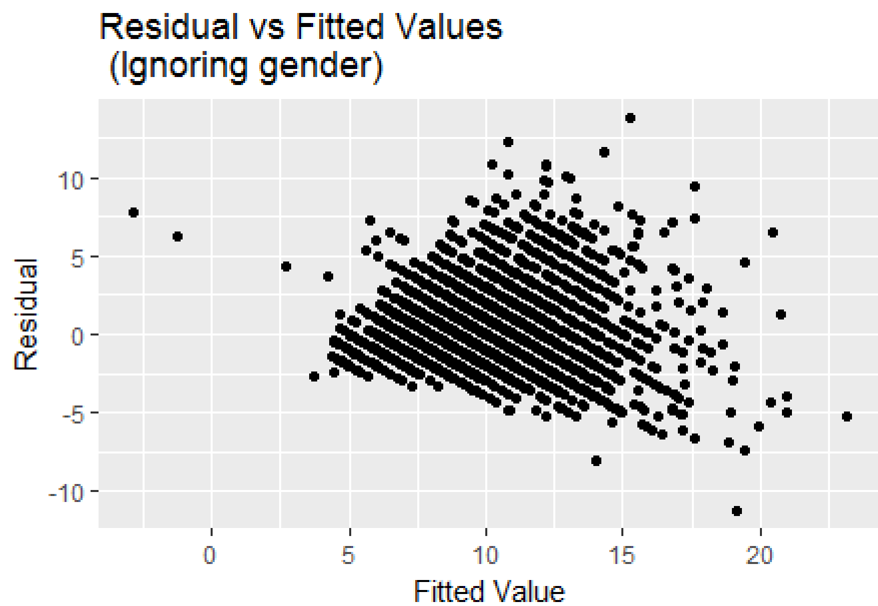
7.11

The dataset we used for this problem has 4177 observations and 9 variables. What we wanted to predict is the age of abalone. It was represented as number of `rings` in the data.

(a)

Age vs other measures, ignoring gender

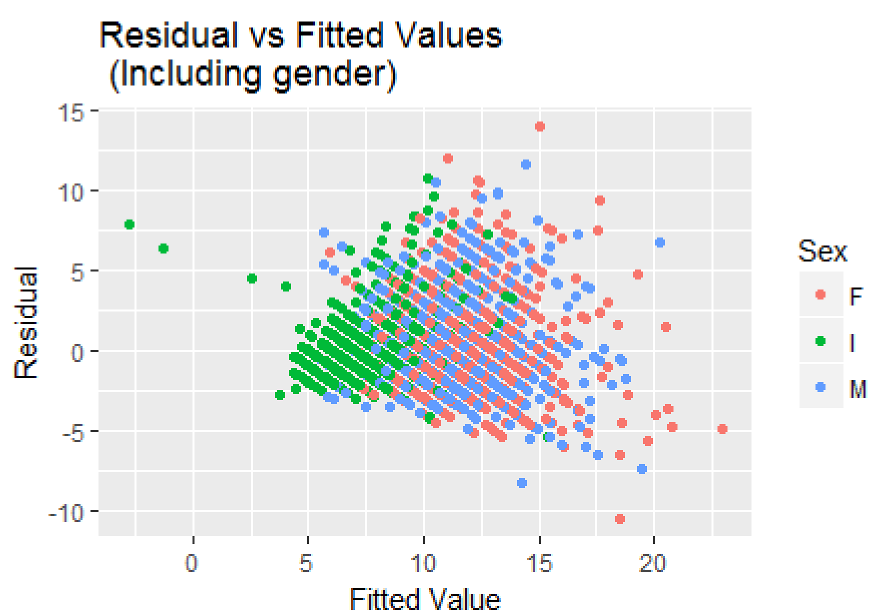
```
##
## Call:
## lm(formula = rings ~ . - Sex, data = dt3)
##
## Coefficients:
##    (Intercept)          Length        Diameter          Height
##          2.985          -1.572          13.361          11.826
##   Whole_weight   Sucked_weight  Viscera_weight    Shell_weight
##          9.247         -20.214          -9.830           8.576
```



(b)

Age vs other measures, including gender

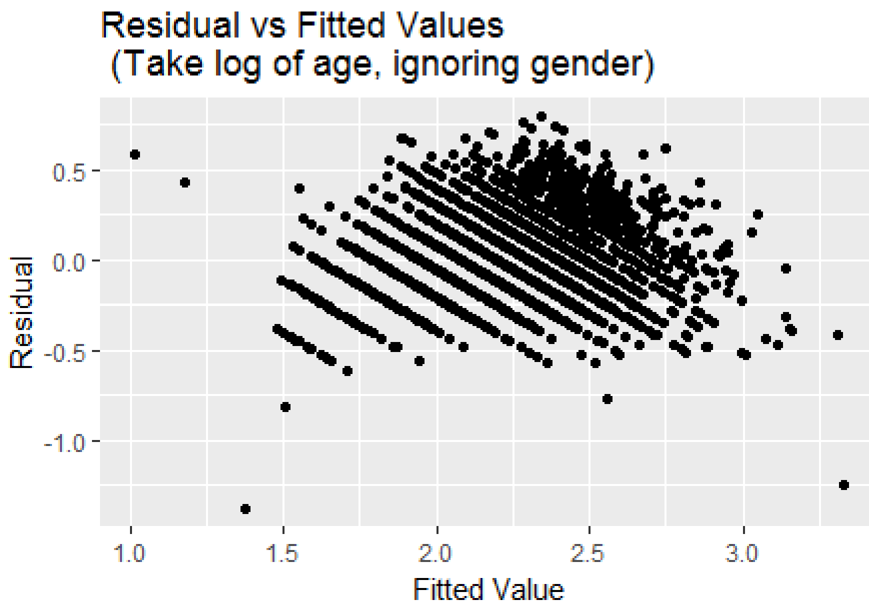
```
##
## Call:
## lm(formula = rings ~ ., data = dt3)
##
## Coefficients:
##     (Intercept)            SexI            SexM          Length
##         3.89464        -0.82488         0.05772        -0.45834
##        Diameter          Height    Whole_weight   Sucked_weight
##        11.07510        10.76154         8.97544       -19.78687
## Viscera_weight    Shell_weight
##       -10.58183         8.74181
```



(c)

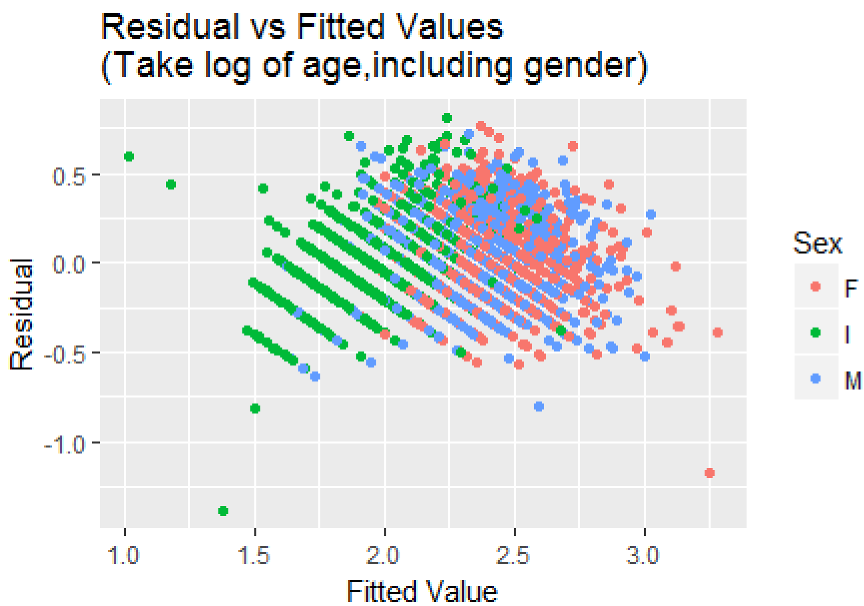
log(Age) vs other measures, ignoring gender

```
##
## Call:
## lm(formula = log(rings) ~ . - Sex, data = dt3)
##
## Coefficients:
##    (Intercept)          Length        Diameter          Height
##         1.2395          0.4067          1.6820          1.3268
##   Whole_weight   Sucked_weight  Viscera_weight    Shell_weight
##         0.6391         -1.7043         -0.7514          0.5879
```



(d) log(Age) vs other measures, including gender

```
##
## Call:
## lm(formula = log(rings) ~ ., data = dt3)
##
## Coefficients:
##    (Intercept)            SexI            SexM          Length
##       1.341185       -0.092485        0.008926        0.533049
##       Diameter          Height    Whole_weight   Sucked_weight
##       1.423575        1.206625        0.608252       -1.657046
## Viscera_weight    Shell_weight
##      -0.835499        0.606814
```



(e)

From the four plots above, we can see that basically gender does **not** have an effect in predicting the ages. The reason is that the plots do not change much after ignoring the `gender` predictor. However, transformation of the `rings` make a difference. After taking the log for the response variable, the **Residual vs Fitted** plots have a flat trend and vertical evenly spread. Before log transformation, the pattern looks like a funnel towards right. That is saying for larger fitted values, the residuals will be larger too.

Thus we would choose the regression model from **c** or **d**. We believe they will do a good job in prediction.

(f)

We made a total of four models for questiosn (a)-(d). For this question, we tried two regularization methods, `ridge` and `lasso`, to check if we could improve the regression models. We made 8 models in total and produced 8 plots.

In the `glmnet` package, the default regularization method is ridge. The parameter is `alpha = 0`. Setting `alpha = 1`, the method changes to lasso. We used `cv.glmnet` to do cross validations. The default fold number is 10. The regularization value λ is determined automatically by the function. It tested 100(default) *lamada* values and choose the one with lowest cross-validation error.

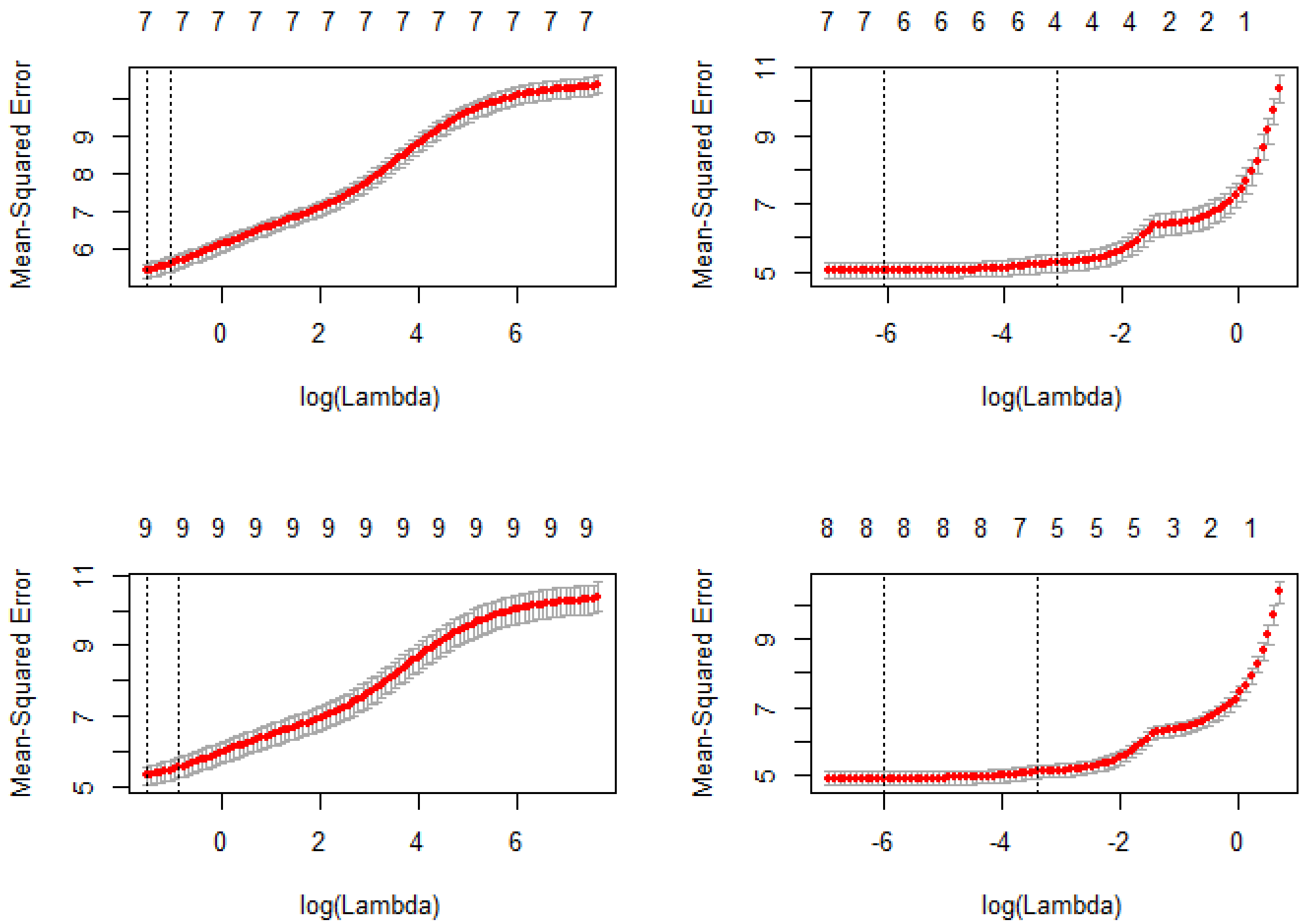
From the table below, we can see that regularzation method **does not** improve the regressions. Lasso and Ridge are often used for the purpose of control over-fitting and deal with multicollinearty among the predictors. In our case, there are basically no collinearity among predictors. As we can see from the plots that none of the variables was dropped for lasso. Thus regularation methods are not helpful here. We also noticed Lasso method behaves better than Ridge regression. **In summation of above, the models without regularizers are better**.

Table 1: MSE Table

Models	Ridge	Lasso	No Regularizer
Age, ignoring sex	5.4044720	5.0528580	4.9092370
Age, including sex	5.3166590	4.9226560	4.8026640
log(Age), ignoring sex	0.0463197	0.0433879	0.0423100
log(Age), including sex	0.0445046	0.0421886	0.0409233

The following plots are log(Lambda) vs Mean-Squared Error plots

Upper left: Age vs other variables, ignoring gender (Ridge regression)
Upper right: Age vs other variables, ignoring gender(Lasso regression)
Bottom left: Age vs other variables, including gender (Ridge regression)
Bottom right: Age vs other variables, including gender (Lasso regression)



The following plots are log(Lambda) vs Mean-Squared Error plots

Upper left: Log(Age) vs other variables, ignoring gender (Ridge regression)
Upper right: Log(Age) vs other variables, ignoring gender(Lasso regression)
Bottom left: Log(Age) vs other variables, including gender (Ridge regression)
Bottom right: Log(Age) vs other variables, including gender (Lasso regression)