hw01

Samuel Swain

2023-01-20

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-6
```

Problem 1

1(a)

```
## MSE:
##  - Average: 21154585
##  - SD 326112.6
```

```
## RMSE:
##  - Average: 163.8434
##  - SD 1.257097
```

As we can see above, the average error is 163.8433881 per prediction with a standard deviation of 1.2570975. For about one fourth of the data, that is the entire cost to insure the customer. The predictive power of this model can definitely be improved or we can explore other models that will do a better job at predicting cost.

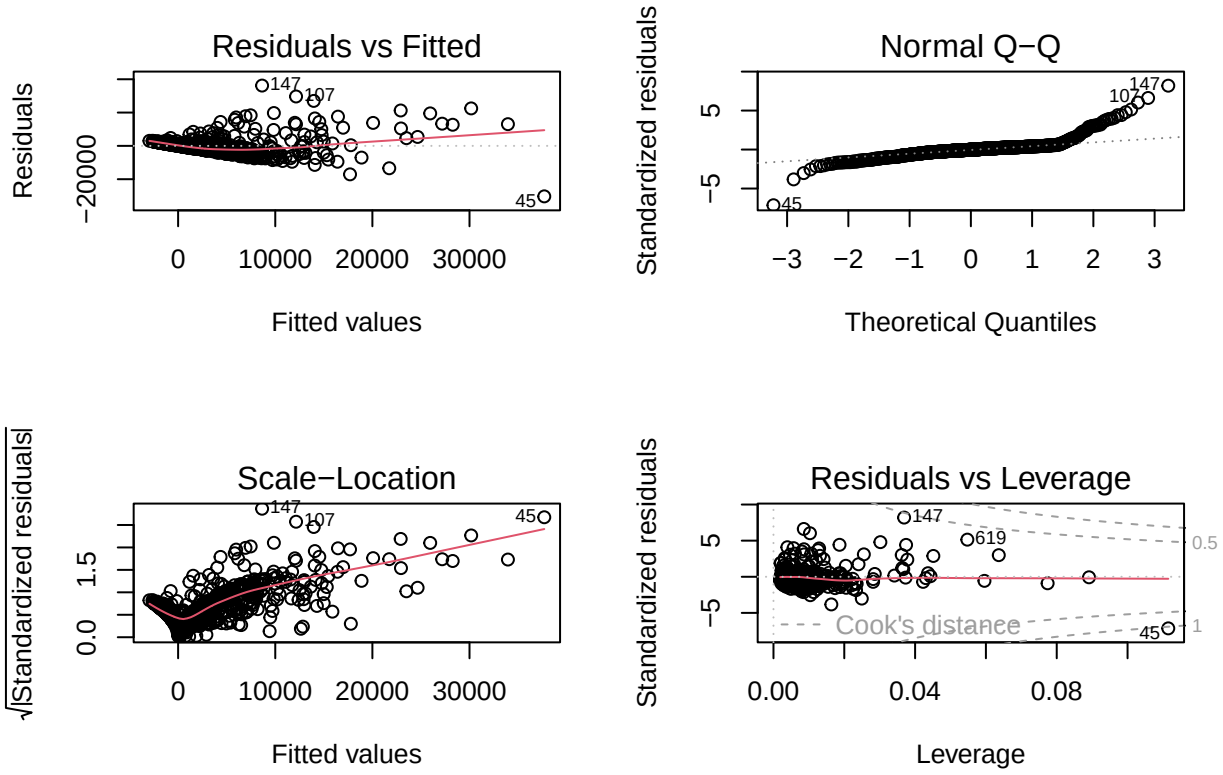
1(b)

```
##
## Call:
## lm(formula = cost ~ age + gend + intvn + drugs + ervis, data = df_1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30308  -1702    -56   1235  36196
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2800.0      160.4  17.457  < 2e-16 ***
## age           -315.6      161.1  -1.959   0.0505 .
## gend          -391.6      161.6  -2.423   0.0156 *
## intvn         4543.3      173.0  26.263  < 2e-16 ***
## drugs         -400.6      189.2  -2.117   0.0346 *
## ervis         1127.9      199.6   5.651 2.23e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4502 on 782 degrees of freedom
## Multiple R-squared:   0.55,  Adjusted R-squared:  0.5471
## F-statistic: 191.2 on 5 and 782 DF,  p-value: < 2.2e-16
```

Total number of interventions and number emergency room visits seem to have the highest effect on the cost of the subscriber. Increasing intvn by only one will increase the cost of the subscriber by about 812.08 whereas increasing ervis by just one will also increase the cost of the subscriber by about 376.53 dollars.

1(c)



As we can see from the plots above, there are a lot of problems with using a linear model to attain the relationship between the features and the cost. The problems are listed below: - Heteroskedasticity based on the Residuals vs Fitted plot - Non-normally distributed residuals based on the Normal Q-Q plot - A few outliers based on the Residuals vs Leverage plot

Problem 2

2(a)

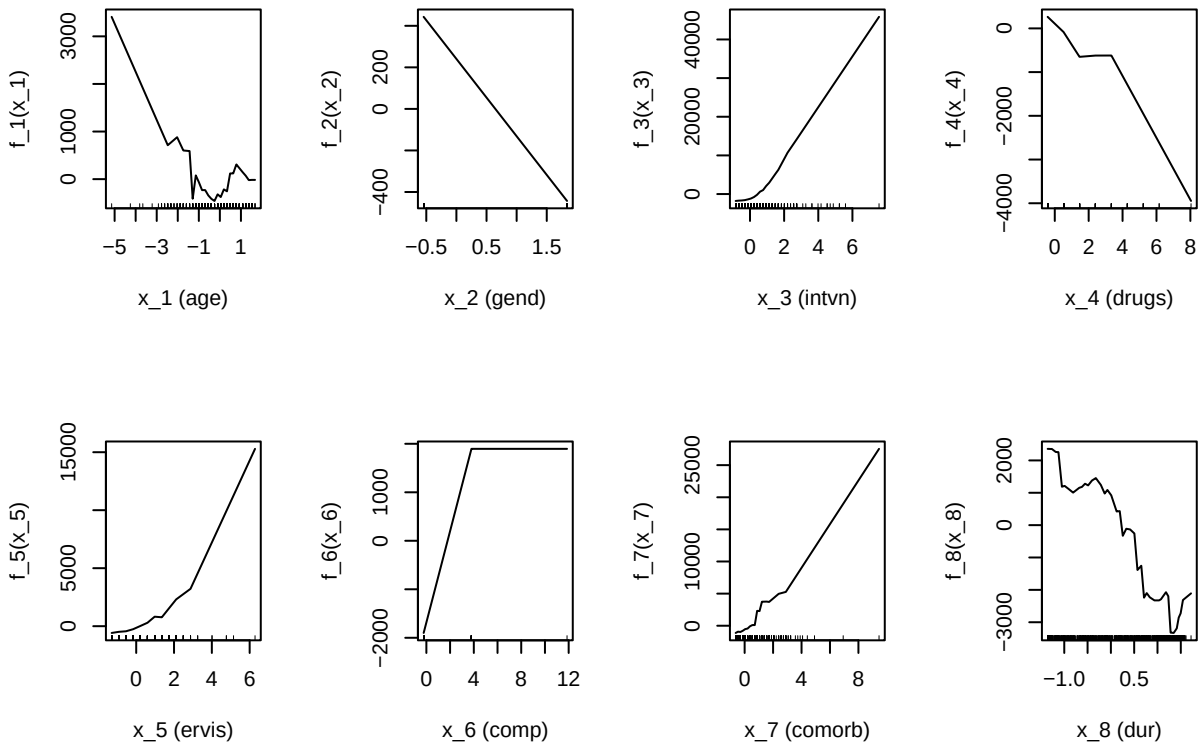
```
## Best NNet:
##  Model: 4
##  Decay: 0.10
```

```
##  Size: 20
```

2(b)

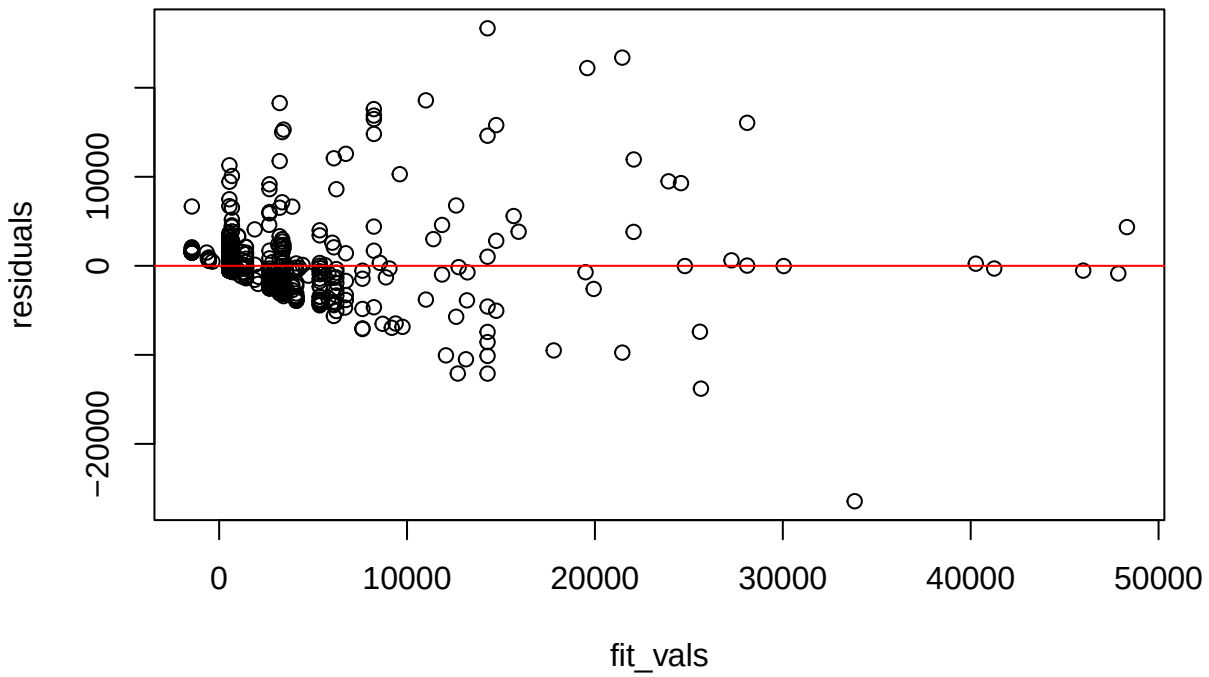
The best Neural Net attained above has a size of 20 and a decay rate of 0.10. In terms of predictive power, I think the model is not very good. With an average CV MSE of 2.8388973×10^7, it performs worse than the linear model with even just two variables in it. The linear model has an average CV MSE of about three hundred thousand.

2(c)



Below are the top three features ranked by strength of effect: - Intervention (intvn): Positive - Complications (comp): Positive for the first 25% and then negative for the last 75% - Drugs (drugs): Negative for the first 20% and then negative for the last 80%

2(d)



Looking at the fitted versus residual plot above, we can see pretty much all of the non linearity has been captured besides the patters going on at the beginning.

Problem 3

3(a)

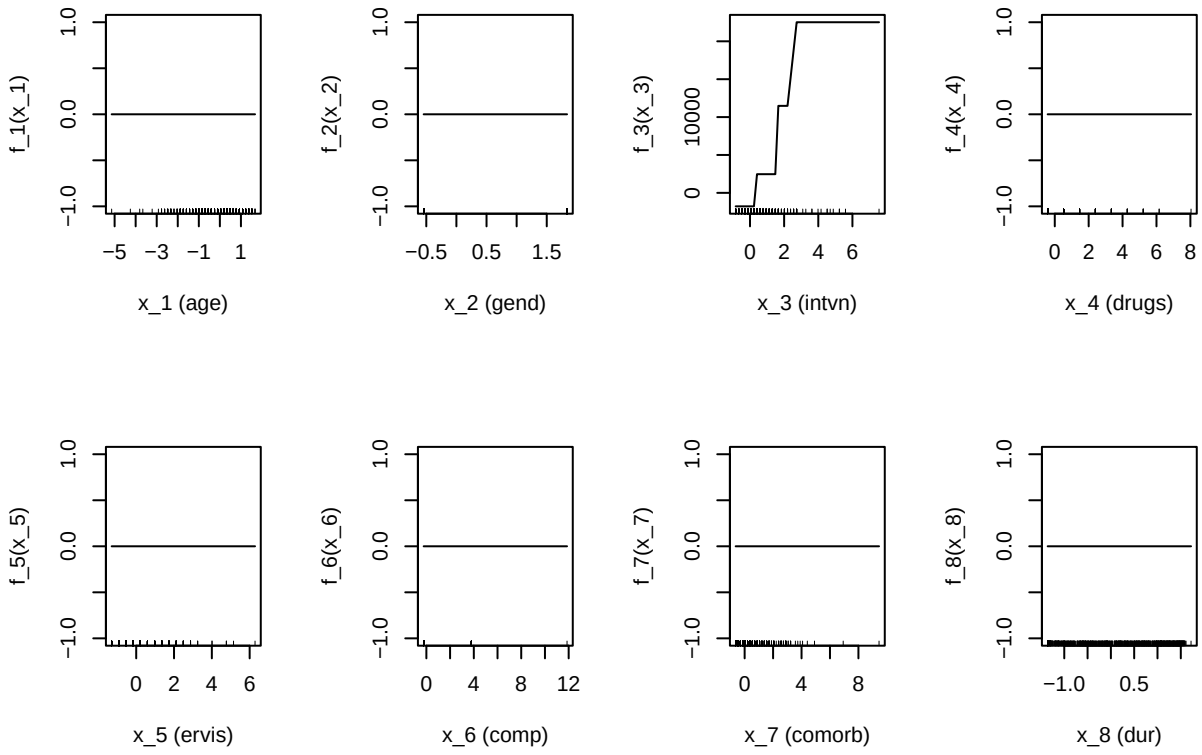
```
## Best Tree:
##  Model: 3
##  Minbucket: 15
##  CP: 0.01
```

3(b)

```
## MSE:
##  - Average: 22172035
##  - SD 921931.5
```

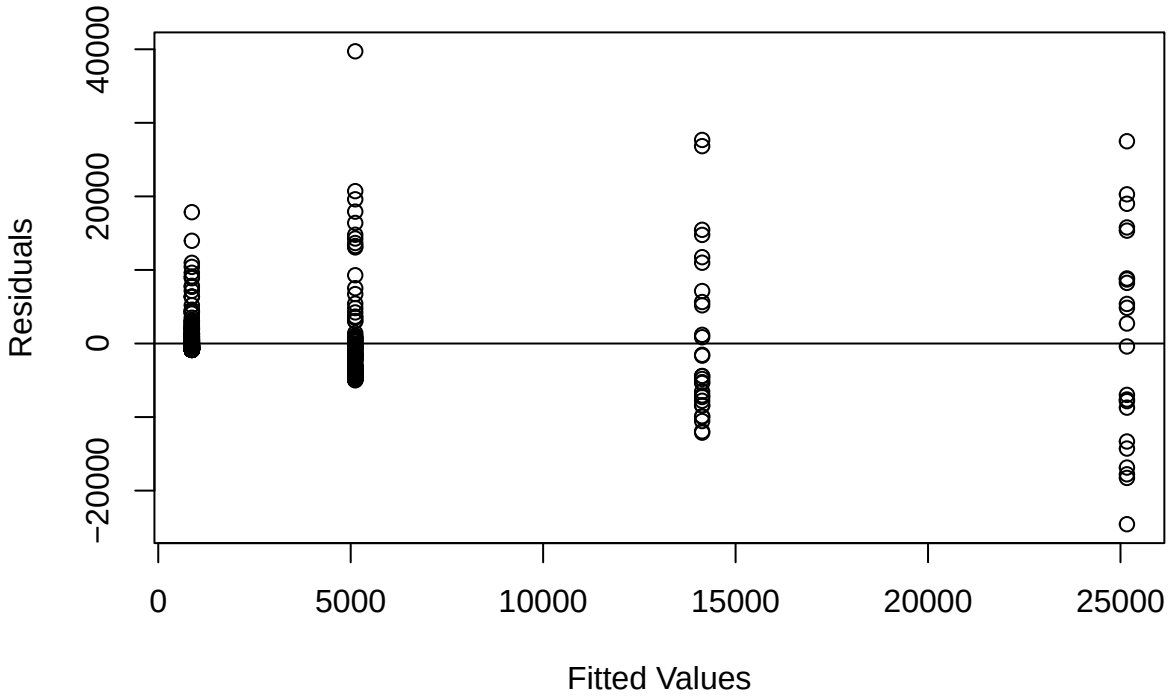
The best Tree attained above has a min-bucket of 15 and a decay rate of 0.01. In terms of predictive power, I think the model is not bad. It's almost better at predicting than the linear and a lot better than the N Net model given average CV MSE of 2.2172035×10^7.

3(c)



As we can see above, intervention (intvn) is the only important variable when fitting the tree. The cost of the subscriber goes up for each additional intervention until two, then the effect levels off.

3(d)



The pattern of this tree Fitted vs Residual plot is similar to the N Net plot. It seems to have captured most of the non linearity. There is no trend down or up as we increase fitted values, only an increase in the variance.

3(e)

I would use the linear model as it produces the lowest MSE out of the three models. As the professor said in class, the assumptions are only for doing fancy statistical proofs with the model. If it predicts well then it is usable! That's why I will go with the linear model. As we were instructed to not alter the variables in the beginning, I'm sure we could fix a good amount of the assumptions as well. This would make the linear model even better than its current state.

Question 4

4(a)

4(b)

4(c)

4(d)

Appendix

Commented out code is the output used to attain the results above. Commended to shorted final submission pdf.