Olin Yoder

Mid Term Exam

BSAN 450 (Spring 2023)

This Midterm is due on March 2, 2023 at 2:00 PM Central. The total points possible are 100. You can discuss the exam with your friends or group members but you will submit your own solutions to Canvas, either in word or pdf. You are expected to write your own codes and produce your own report. Copying other student's code or report is not allowed and will constitute cheating. Considerations in Grading:

I am not just interested in the final output and R code, but also in the process you go through to arrive at your model. A significant amount of the grade for the exam will be determined by the process you use to come up with your final model. If you have a poor process or do not clearly describe that process your grade will be reduced significantly.

A financial institution is interested in understanding factors that can help them in the early prediction of corporate losses. These losses may occur for a variety of reasons, such as lawsuits, regulatory fines, etc. The institution has data on 300 losses that have been collected over a period of 12 months. For each loss, the financial institution also has information on 20 other variables which they believe are important factors that may help predict such losses in the future. Our goal is to build a model for predicting the loss dollar amount, loss, using the 20 variables. The first 6 variables in the data are described below:

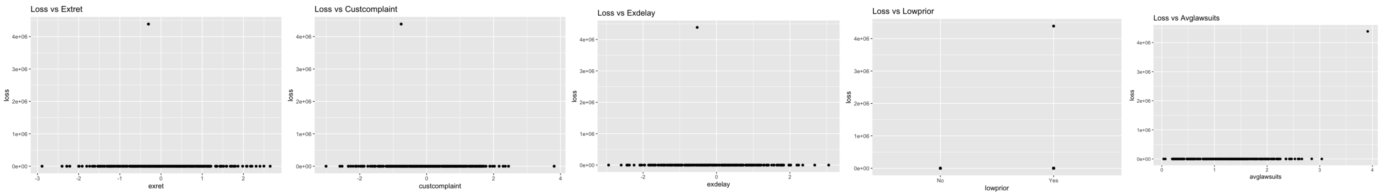
- loss: the dollar loss amount
- exret: the excess return on company stock
- custcomplaint: change in customer complaint
- exdelay: how many days in excess of 7 days did the complaint last
- lowprior: was the complaint low priority (Yes/No)
- avglawsuits: average number of lawsuits associated with the loss

The remaining variables have been anonymized due to privacy concerns. They are still available in the data but we will not know what they actually mean.

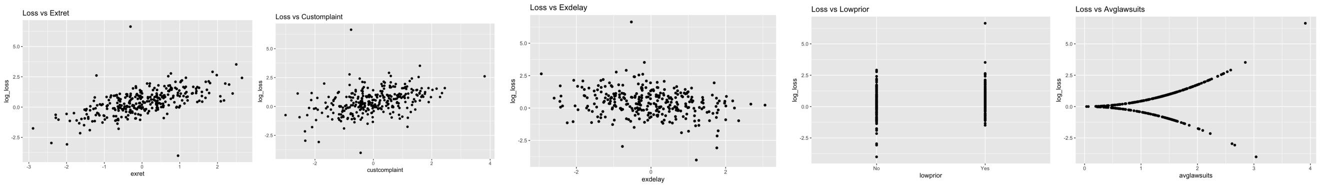
1. Do a preliminary graphical analysis of the dependent variable loss and the first five independent variables. Does it appear that loss may have a linear relationship with either exret, custcomplaint, exdelay, lowprior or avglawsuits? If not, can you think of a transformation of the dependent variable that may help in at least showing graphically that the transformed loss variable may have a linear relationship with the five aforementioned independent variables? *(You may try log(loss) or square root (loss) and pick* one that seems the best to you.)

Olin Yoder

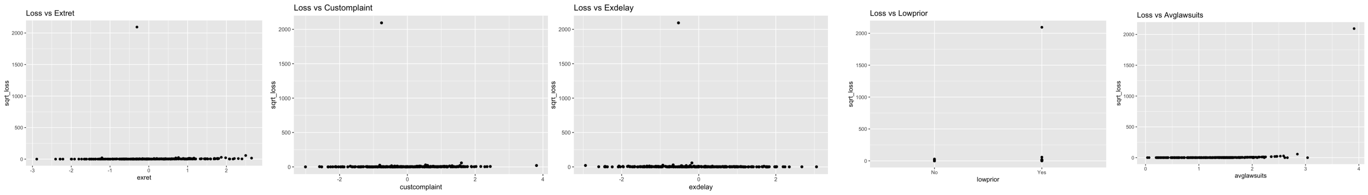
Loss does not appear to have a linear relationship with any of the variables prior to a transformation (although the outlier makes visualizing the relationship more difficult).



After a log transformation:



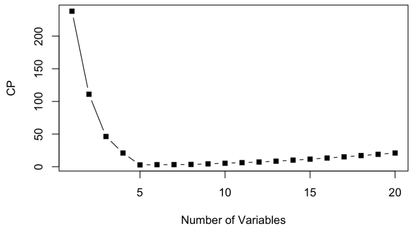
After a sqrt transformation:



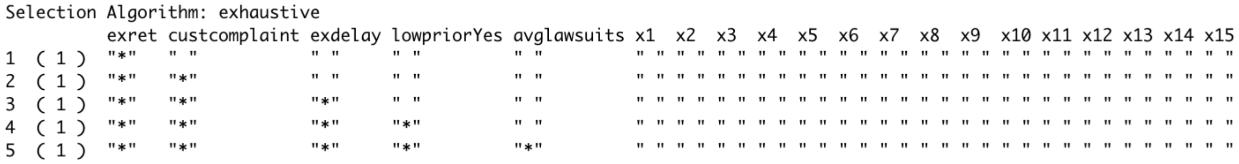
- The log transformation creates a linear relationship between loss and most of the independent variables (avglawsuits looks more like x = abs(y)). As such, we will transform loss to log10(loss) moving forward.

2. Since we have 20 variables, you will need to use some technique for variable selection. Implement the Best Subset Regression technique on this data for selecting a subset of these 20 independent variables that may be useful for predicting your dependent variable.(From part (1) above, if you think a transformation for loss is needed then your *dependent variable is the transformed loss.)* Use C_p for selecting the final model. Remember to show the output, the plots and your justification for selecting the final model. Let us call your final model: **model a**.

Using Best Subset and Cp for selecting the final model, the best model contains 5 variables. Those variables are exret, custcomplaint, exdelay, lowprior, and avglawsuits.



Olin Yoder



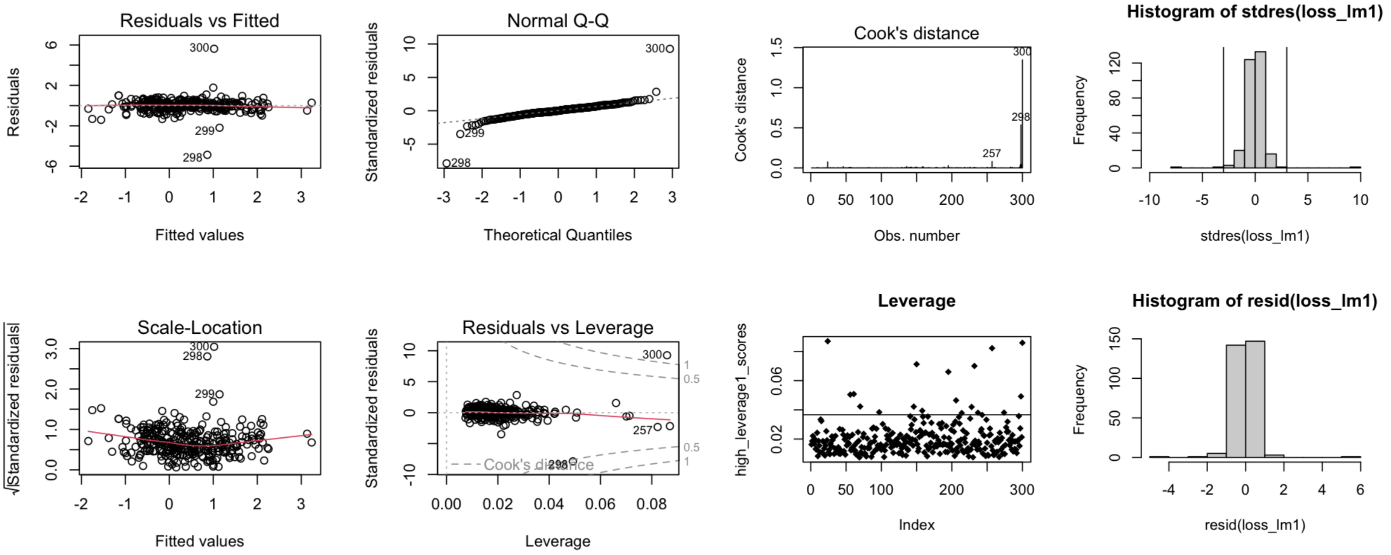
```
Selection Algorithm: exhaustive
         exret custcomplaint exdelay lowpriorYes avglawsuits x1  x2  x3  x4  x5  x6  x7  x8  x9  x10 x11 x12 x13 x14 x15
1  ( 1 ) "*"   " "           " "     " "         " "         " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
2  ( 1 ) "*"   "*"           " "     " "         " "         " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
3  ( 1 ) "*"   "*"           "*"     " "         " "         " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
4  ( 1 ) "*"   "*"           "*"     "*"         " "         " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
5  ( 1 ) "*"   "*"           "*"     "*"         "*"         " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
```

Model A:

log(loss) = -0.15487 + 0.62118(exret) + 0.37337(custcomplaint) - 0.25446(exdelay) + 0.35991(lowpriorYes) + 0.29504(avglawsuits)

- The intercept term is not significant

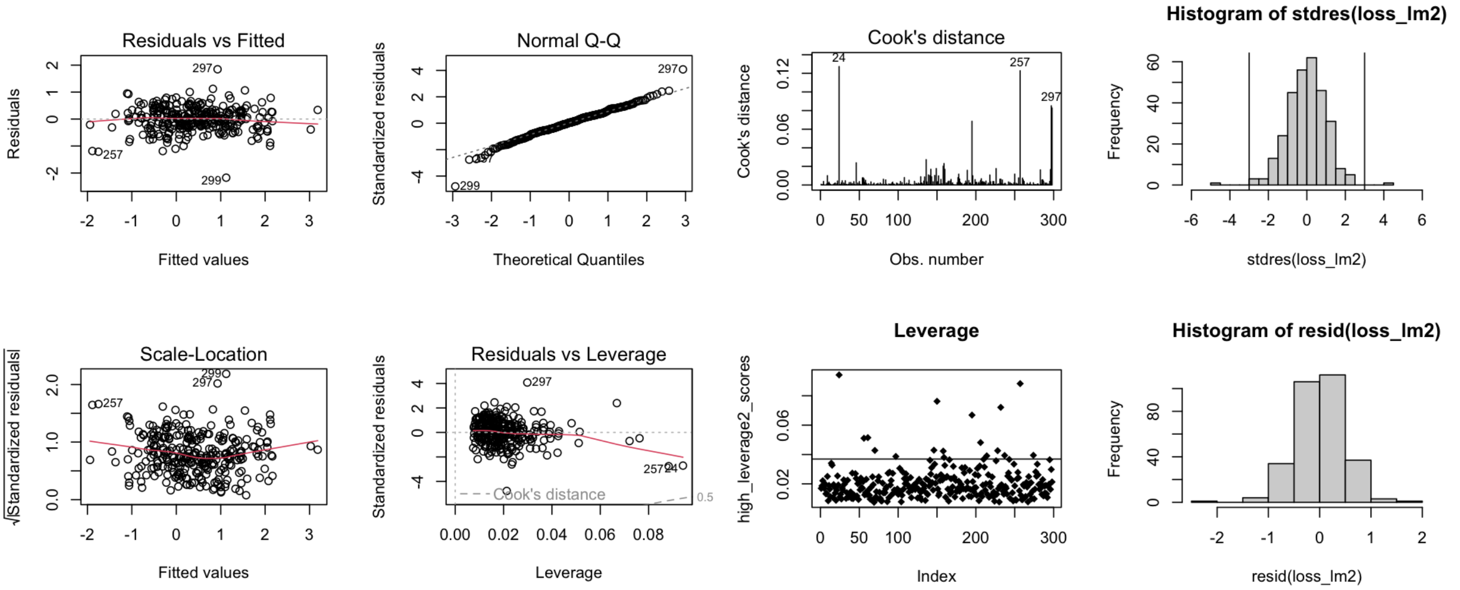
3. Once you have **model a**, run the diagnostic checks on this model. Do the diagnostic plots suggest any problems with model assumptions, presence of outliers and / or influential observations? If so, perform the required analysis and provide all the details in your report on how you went about correcting **model a** diagnostic plots and the issue with outliers, if any.



- The major issue in the diagnostic plots are influential points.
  - Point 300 has a Cook's Distance of 1.347722 and a standardized residual of 9.274369.
  - Point 298 has a Cook's Distance of 0.5355962 and a standardized residual of -7.878461.
  - Point 299 has a standardized residual of -3.493083, but a Cook's Distance of only 0.0442329 (I'm not going to remove it yet, but will keep a close eye on it).
  - There are numerous high leverage points, too.
- The influential points are also likely causing issues with non-constant variance and normality (bptest p-value: 1.516e-08, shapiro.test p-value: < 2.2e-16). Graphically, the residuals vs fitted and histogram of residuals look fine, though.

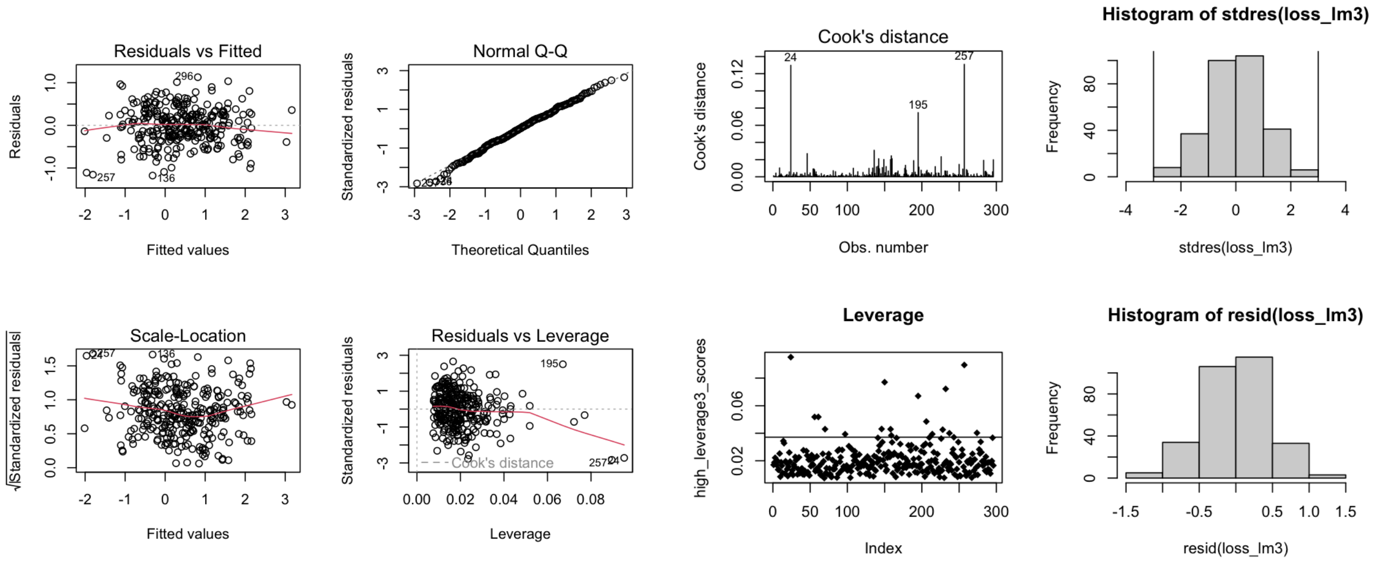
Olin Yoder

First, I am only going to remove Points 298 and 300 and refit the model.



- There are no points with a Cook's Distance > .5, however, there are some points with a standardized residual > 3 or < -3.
  - Point 299 (using the same indexing as before) once again has a large standardized residual, -4.779687. Its Cook's Distance is only .08, though.
  - Point 297 has a standardized residual of 4.078336 (was 2.840247 prior to first removal).
  - Again, there are numerous high leverage points.
- Once again, these points are also likely causing issues with non-constant variance and normality (bptest p-value: 7.11e-06, shapiro.test p-value: 0.000729) despite the graphs looking fine.

Remove Points 297 and 299 and refit the model.



- There are no points with a Cook's Distance > .5 or with a standardized residual > 3 or < -3.
- There are some high leverage points, but none so extreme that warrant removal from the model.
- The residuals are normally distributed (shapiro.test p-value: 0.6914).
- Graphically, there are no issues with the residuals vs fitted values plot. However, the bptest returns a p-value of 5.642e-08 which suggests otherwise. The gvlma function returns a p-value for heteroscedasticity of 0.120989, which contradicts the bptest. Ultimately, though, I am not too concerned about non-constant variance in this model.

Model B:

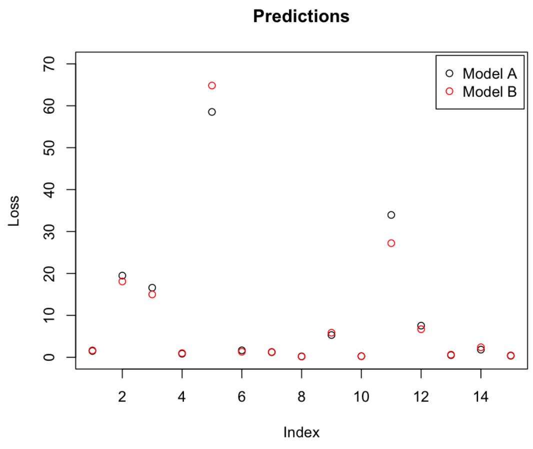
log(loss) = 0.001433 + 0.651793(exret) + 0.402685(custcomplaint) - 0.246645(exdelay) + 0.292233(lowpriorYes) + 0.195209(avglawsuits)

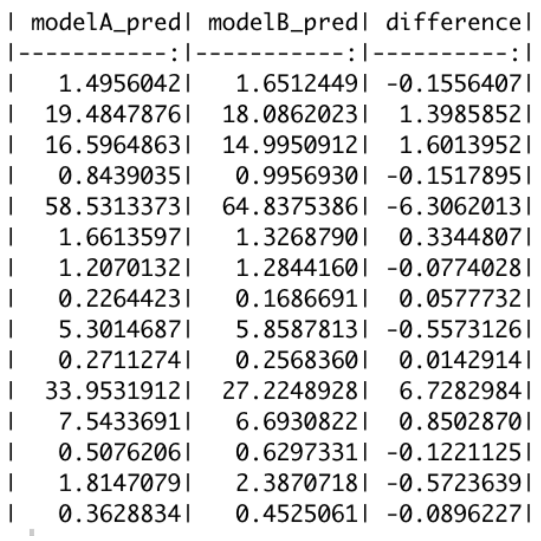
- The intercept term is not significant

4. At this stage, you should have a model that is **your final model**, henceforth **model b**. The diagnostic plots for **model b** should not show any problems. Recall **model a** from part (3) above. Use **model a** and **model b** to predict the **loss** for the data available in 'newdata.csv'.

   How do the predictions from **model a** and **model b** compare?

   The predictions are relatively similar, but Model A tends to predict a bit higher. On average, Model A predicts 0.1968443 higher than Model B. This is due to the removal of extremely high losses in Model B.

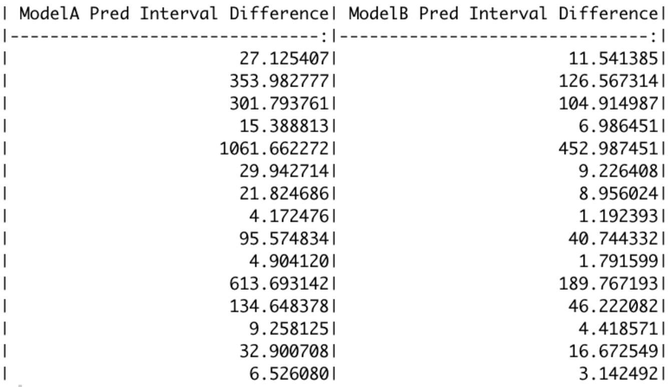




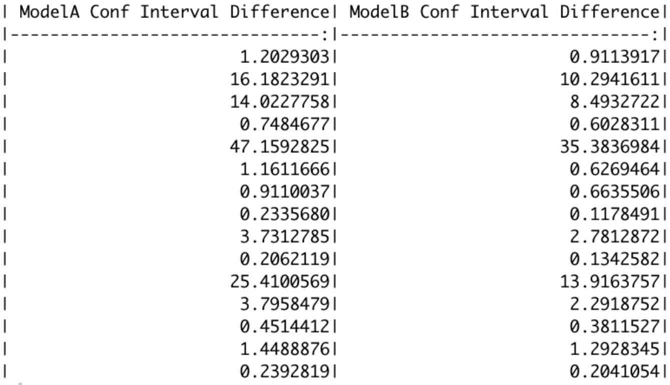
```
| modelA_pred| modelB_pred| difference|
|-----------:|-----------:|----------:|
|   1.4956042|   1.6512449| -0.1556407|
|  19.4847876|  18.0862023|  1.3985852|
|  16.5964863|  14.9950912|  1.6013952|
|   0.8439035|   0.9956930| -0.1517895|
|  58.5313373|  64.8375386| -6.3062013|
|   1.6613597|   1.3268790|  0.3344807|
|   1.2070132|   1.2844160| -0.0774028|
|   0.2264423|   0.1686691|  0.0577732|
|   5.3014687|   5.8587813| -0.5573126|
|   0.2711274|   0.2568360|  0.0142914|
|  33.9531912|  27.2248928|  6.7282984|
|   7.5433691|   6.6930822|  0.8502870|
|   0.5076206|   0.6297331| -0.1221125|
|   1.8147079|   2.3870718| -0.5723639|
|   0.3628834|   0.4525061| -0.0896227|
```

Additionally, if we look at the prediction and confidence intervals between the two models, Model B consistently has a narrower (smaller) interval.

Olin Yoder



```
| ModelA Pred Interval Difference| ModelB Pred Interval Difference|
|-------------------------------:|-------------------------------:|
|                       27.125407|                       11.541385|
|                      353.982777|                      126.567314|
|                      301.793761|                      104.914987|
|                       15.388813|                        6.986451|
|                     1061.662272|                      452.987451|
|                       29.942714|                        9.226408|
|                       21.824686|                        8.956024|
|                        4.172476|                        1.192393|
|                       95.574834|                       40.744332|
|                        4.904120|                        1.791599|
|                      613.693142|                      189.767193|
|                      134.648378|                       46.222082|
|                        9.258125|                        4.418571|
|                       32.900708|                       16.672549|
|                        6.526080|                        3.142492|
```



```
| ModelA Conf Interval Difference| ModelB Conf Interval Difference|
|-------------------------------:|-------------------------------:|
|                       1.2029303|                       0.9113917|
|                      16.1823291|                      10.2941611|
|                      14.0227758|                       8.4932722|
|                       0.7484677|                       0.6028311|
|                      47.1592825|                      35.3836984|
|                       1.1611666|                       0.6269464|
|                       0.9110037|                       0.6635506|
|                       0.2335680|                       0.1178491|
|                       3.7312785|                       2.7812872|
|                       0.2062119|                       0.1342582|
|                      25.4100569|                      13.9163757|
|                       3.7958479|                       2.2918752|
|                       0.4514412|                       0.3811527|
|                       1.4488876|                       1.2928345|
|                       0.2392819|                       0.2041054|
```

5. The financial institution had hired an external consultant a few months back who went about building a similar model. The consultant's model, henceforth **model c**, predicts log(loss) as a linear function of exret, custcomplaint, exdelay and lowprior. Use these four independent variables to fit **model c** in R. From **model c** summary, what are the estimates of the four coefficients pertaining to exret, custcomplaint, exdelay and lowprior. *(You do not have to worry about the diagnostics for **model c** as you are*

   going to just use the consultant's model for comparison in part (6) below.)

   log(loss) = 0.20615 + 0.66503(exret) + 0.40238(custcomplaint) - 0.27833(exdelay) + 0.38575(lowpriorYes)

6. Use 10 fold cross validation to compute the cross validated estimate of mean squared error for **model b** and **model c**. Which model has a lower mean squared error estimate?
   Using 10 fold CV, the RMSE for Model B is 0.4393621 (MSE: 0.1930) while the RMSE for Model C is 0.65491 (MSE: 0.42891). Since Model B has a lower RMSE, it is likely the better model.