The report

Tie Ma

2024-02-23

Part one: The data evaluation process.

#At here I skiped their name which does not impact the model training, #their ticket number with is not important and their cabin which has limit number

```
## PassengerId    Survived      Pclass         Sex         Age       SibSp
##           0           0           0           0         177           0
##       Parch        Fare    Embarked
##           0           0           2
```

The trainning data missing 177 Age value with 2 Embarked value. Consider the size of the trainning data set are only 881 rows. Remove all the missing Age variable will further reduce avaiable data for the model. Therefore, I decided to using the average age to fill the missing 177 value. However I do need to check the age average between the people who surviied and people does not to aovid introduce the biased the training data set.

The average age by survival as following.

```
##   Survived      Age
## 1        0 30.62618
## 2        1 28.34369
```

The average age of the entire data set

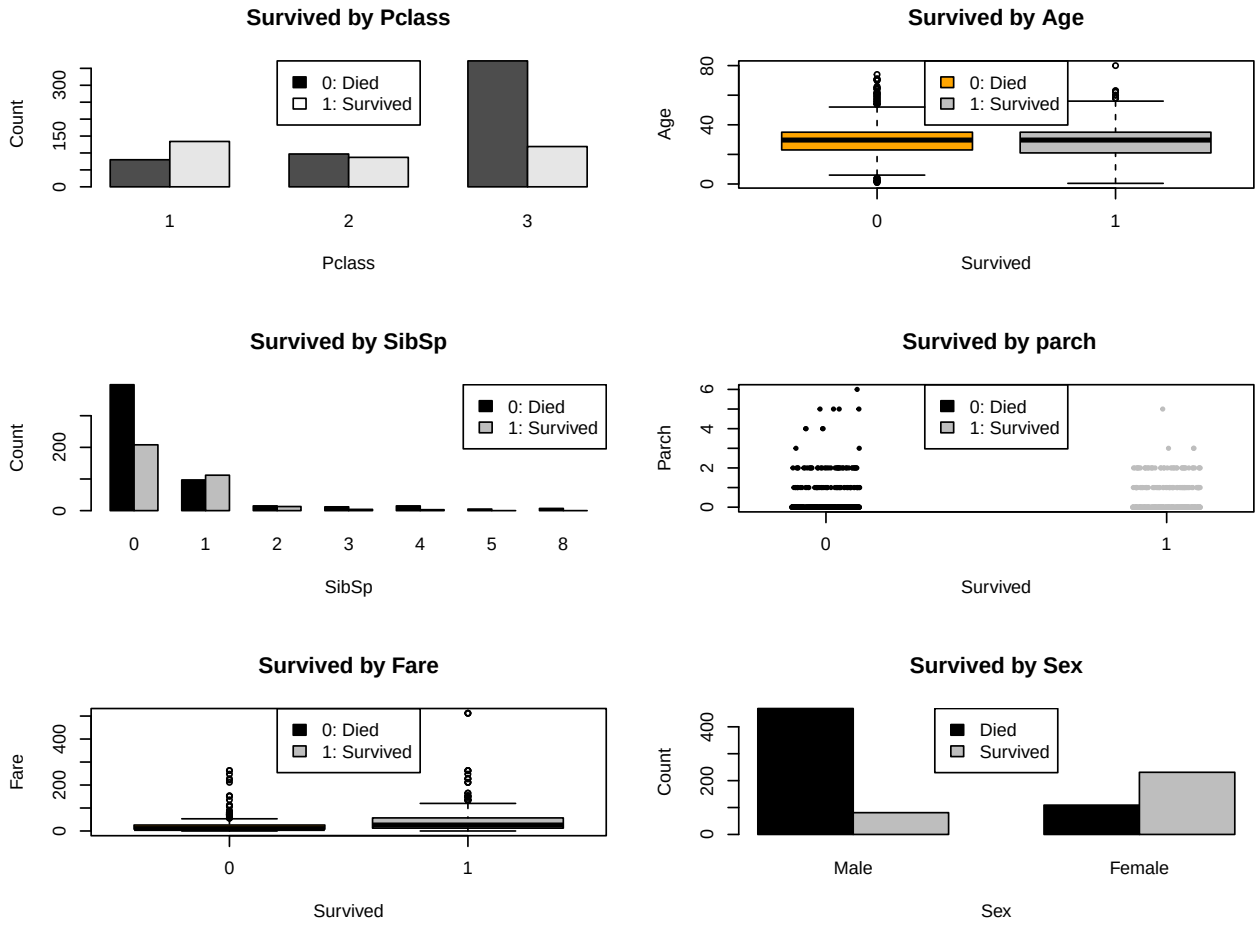
```
## [1] 29.69912
```

The average age of the entire training data set are 29.69 which is right in between the average age for people who surivied and who does not. Therefore using the average age will not introduce significant bias toward the training data.

Further more, After fill all the missing age value, I removed the 2 missing row of the embarked value from the training data set and transfer the "Sex" columns into the dummy variable. It increasing the accuracy of Linear regression model from 76.794% to the 77.033% which also the second highest of all my model only 0.718% less than the regression forest model.

Step two: check the data.

For a better understanding of the possible patterns in the data and to determine potential variables for the linear regression and logit models, I examined graphs comparing the impact of different variables on passenger survival. From these graphs, we can observe that Passenger Class (Pclass), SibSp, and gender significantly influence the survival of passengers. Furthermore, the fare, representing the ticket price, shows that passengers who paid higher fares were more likely to survive. This observation aligns with the conclusion drawn from the Passenger Class graphs. Furthermore, the rest of the variables do not provide a visible impact on survival according to the analysis.



For better perform of the linear regression, I check the heteroscedasticity and the Multicollinearity of the training data set.

For heteroscedasticity test, I am using BP test.

```
##
##  studentized Breusch-Pagan test
##
## data:  data_test_1
## BP = 9.8992, df = 7, p-value = 0.1944
```

Since the p value is 0.19 which is greater than 0.05, we fail to reject the null hypothesis. we state that there is not enough statisticallty significant evidence of heteroskedasticity exist in the data set.

For the Multicollinearity, I am using the VIF test.

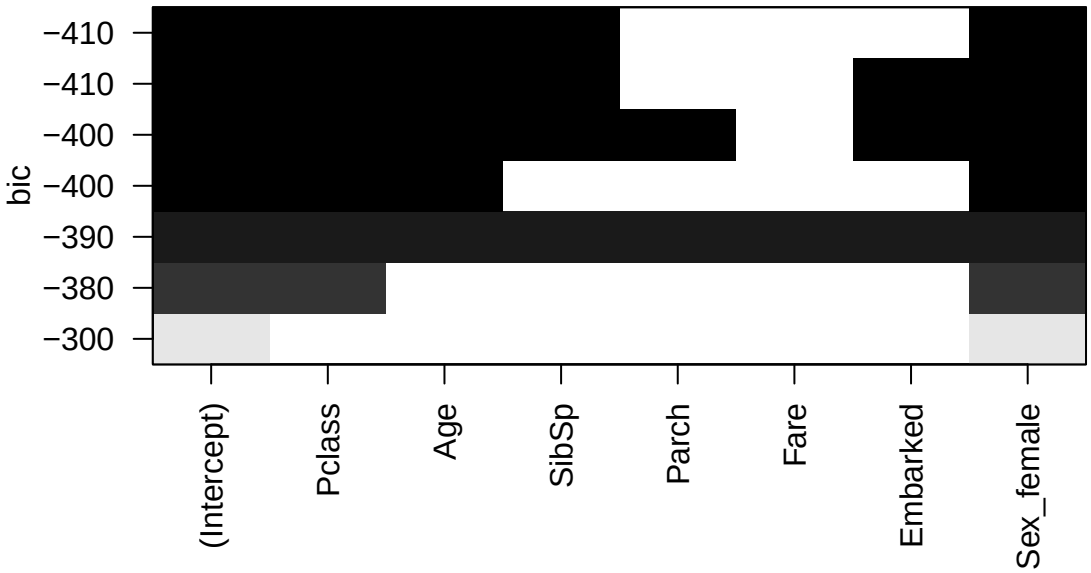
```
##     Pclass        Age      SibSp      Parch       Fare   Embarked Sex_female
##   1.664130   1.203003   1.281804   1.323066   1.645957   1.081159   1.109890
```

All the VIF values for the variable are between 1 and 1.6. This indicates that the training data show moderate correlation, which is common for real-life examples, and will not raise a red flag.

Step Three: choose the variable.

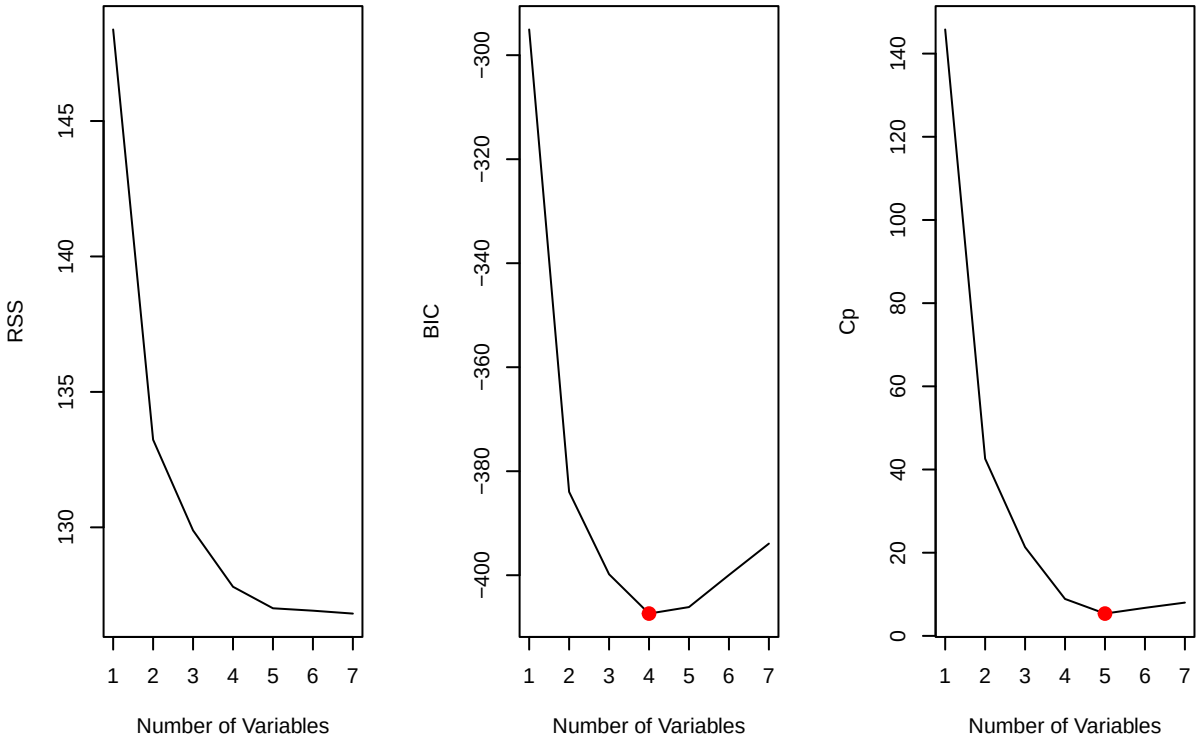
In order to select the optimal variable for constructing the linear regression model without risking overfitting, I will use the best subset selection and Ridge regression to explore all possible linear combinations of variables

First, I will using the best subset selection to search all the possible combination and compare their BIC vlaue



From the graph, we can see that the combination of Pclass, Age, SibSp, and sex_female has the lowest BIC, which means this model is likely the most efficient among the tested combinations in terms of balancing model complexity and goodness of fit.

Next, we determine the number of variable.



The BIC graphy suggest with the four variable will reach the lowest BIC same as the early grpahy point out. As CP(AIC) graph suggest that the optimal variable number is 5.

According to the best subset selection, I should choose the two group with the following variables.

- The first group includes:
  - Pclass, Age, SibSp, Sex_female
- The second group includes:

– Pclass, Age, SibSp, Embarked, Sex_female"

Now, let's proform the Ridge regression.

```
##              Variable           OLS         Ridge
## Pclass         Pclass -0.1688667659 -0.1564689820
## SibSp           SibSp -0.0411820740 -0.0388136461
## Embarked     Embarked -0.0353433928 -0.0358988675
## Parch           Parch -0.0168729693 -0.0138132743
## Age               Age -0.0058566464 -0.0053474253
## Fare             Fare  0.0002829035  0.0003878972
## Sex_female Sex_female  0.5059113514  0.4799392450
```