Milestone_8

Charles James

2023-11-05

```
titanic = read.csv("titanic.csv")
```

Time to make a simple linear regression model. The only quantitative variables in this dataset is Age and Fare. I will first change the Sex and Pclass columns to integers. This model will attempt to predict the Age of the passenger given the amount of the their Fare, the Sex, where they Embarked and the Pclass of their cabin. I want to do this because I have a lot of missing values in the Age column. If I can create a model that can accurately predict the Age of the passenger, I can then use that model to add that predicted Age, instead of removing that row. Which will increase the amount of the data I have to make future models. First I will remove the rows with missing Age data then create, train and test the model.

```
titanic = na.omit(titanic)
# Convert 'Embarked' column to integers
titanic$Sex <- as.integer(factor(titanic$Sex))
# Convert 'Embarked' column to integers
titanic$Embarked <- as.integer(factor(titanic$Embarked))


titanic_lm = lm(Age ~ Fare + Sex + Embarked+ Pclass, data= titanic)
titanic_lm
```

```
##
## Call:
## lm(formula = Age ~ Fare + Sex + Embarked + Pclass, data = titanic)
##
## Coefficients:
## (Intercept)         Fare          Sex     Embarked       Pclass
##    41.01658     -0.03643      4.20049      0.35457     -8.13313
```

R-Squared = .1727

This R-Score is relatively low. This means that Fare, Sex, Embarked, and Pclass are not sufficient in making a model that can explain the variance of the Ages. This model is therefore not that useful in predicting the missing Age values.

```
png("titanic_hist_resids.png", width = 800, height = 600)
hist(resid(titanic_lm),
     main='Residuals of Titanic LM Model',
     xlab='Residuals',
     ylab='Frequency',
)
dev.off()
```

```
## pdf
##   2
```

```
png("titanic_scatter_age_resid.png", width = 800, height = 600)

plot(titanic$Age, resid(titanic_lm),
     main='Actual Age vs Residuals',
     xlab='True Age',
     ylab='Residual',
)
dev.off()
```

```
## pdf
##   2
```

The equation of the model: y = 41.01658(X0) + -0.03643(X1) + 4.20049(X2) + 0.35457(X3) + -8.13313(X4)

Where y = Age X1 = Fare X2 = Sex X3 = Embarked X4 = Pclass