Does Price Depend on Years, Mileage, and Title-Status?

Muliple Linear Regression Analysis

Michael Gurge

5/10/2021

Introduction

Research Question:

The question of whether price of a car depends on year, mileage, and status people entails discovering whether year, mileage, and title status is significant when explaining the price of a car. In this report the nature of the relationship between price based on year, mileage, and title status will be analyzed to determine the significant of the relationship. First, there would presumably be a relationship because usually when a car is older, unless it is a sports car, muscle car, or rare model of a car, the price is less. Title status, whether it is a salvage title may contribute as a factor because a car with a salvage title can entail that the car was not maintained, therefore, decreasing the price of what the car is worth. The population being studied are new and used cars for sale in the United States. The variables that will be used to answer this question are: Year (measured in years), mileage of the car (measured in miles on the odometer of the car), and title status of the car (clean title or salvage insurance). insurance. The response variable price is quantitative, while predictor variables year and mileage are quantitative and title status is categorical (clean or salvaged).Title status is broken down into 2 levels, a clean title or salvaged.

Methods

Data Collection

The **US Cars Dataset** was collected on April 20, 2020 by scraping a website called Auction Export.com. For further information about this data click the link below. The sampling schema used is simple random sampling with all of the cars clean and used scraped from the website.

Design

The hypotheses that will be investigated are whether or not year, mileage, and title status as a model is significant in explaining car prices and whether or not the variables mentioned previously are significant individually in explaining car prices. The null hypothesis for the significance of the model is that there is no differences in the slope parameters for the model to be significant and the alternative hypothesis is that at least one of the variables are significant in explaining car prices making the model significant. The second null hypothesis is that there is no difference in the slope parameters and the alternative hypothesis is that there are differences in the significance of the slope parameters individually, meaning that that each slope parameter in the model is significant in explaining car prices. I think the year, mileage, and title status will be useful in explaining car prices and each variable will be significant in explaining car prices. The statistical model that will be used is multiple linear regression since the response variable price is quantitative and more than one explanatory variable is present in the model. Lastly, 5% will be the significance level used in the model.

Results

The sample size consists of 2498 with the minimum year of a car being 1973 and the maximum year being 2020. The range of mileage is 0 miles to 1,017,936 miles and 2336 cars in the sample have a clean title status and 163 have a salvage insurance status. The price of cars in the sample range from $0 to $84,900 in US dollars (Appendix 1). Next, the VIF's or variance inflation factors associated with year, mileage and title status are under 5, meaning that there was no issue in multicollinearity being present in the model(Appendix 2).The assumptions for the model consist of linearity, independence, equal variance, and normality. The

linearity assumption was not met since the residuals do not have a constant spread across the zero residual line (Appendix 3). The independence assumption was met since there is no particular ordering to the observations (e.g. it is not based on time) and the random sample of 2498 cars of various years, mileages, and title status,so it is unlikely that there is a violation of independence (Appendix 3). The equal variance assumption was also not met since there is not a constant vertical spread across the residual line (Appendix 3).The normality assumption was met with the the histogram not showing any severe skewness or extreme outliers and the normal QQ plot of residuals having values fall close to the line (Appendix 3).
For inferential methods the F-test the F-statistics was deemed significant with a $p - value < 2.2 \times 10^{-16}$ with a 0.05 significance level as α, which means there is evidence to accept the alternative hypothesis that at least one of the predictor variables (year, mileage, and title status) is significant in explaining car prices (Appendix 4). When using the t-test for each slope parameter year($p - value < 2.2 \times 10^{-16}$), mileage ($p - value < 2.2 \times 10^{-16}$), and title status($p - value = 5.58 \times 10^{-13}$).The ($pvalues < 0.05$) indicate that each variable is significant in explaining the price of cars as it relates to this dataset (Appendix 5). The $R^2 = 0.2261$ indicates that 22.6% of the variation in car prices can be explained by the linear regression model that contains the explanatory variables year, mileage, and title status whether it is a clean title or salvage insurance (Appendix 5).

Discussion/Conclusion

The results from statistical analysis of the multiple linear regression model of price based on year, mileage, and title status of vehicles in the sample consisting of 2498 vehicles is that the model is significant in explaining how price relates to year, mileage, and title status. Secondly, the statistical analysis indicates that the slope parameters year, mileage, and title status are individually significant in explaining car prices. Since the model is used to understand and describe price in relation to the explanatory variables previously stated, no casual conclusions can be implied. The results were as expected with the explanatory variables being significant in relation to price. Lastly, it would be intriguing to analyze title status as a categorical response varied based on price using a one way anova test or determine whether interaction in the model is present between year,title status,and title status.

Appendix

Data Preparation and Numerical Summary

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(DescTools)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:DescTools':
##
##     Recode
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(readr)
library(ISLR)
library(leaps)
```

Determining the Statistical Method and Variable Selection (Appendix 1)

```
CarDataset<- read.csv(
  file = "USA_cars_datasets.csv", header = TRUE
)
CarDataset_noNA<- na.omit(CarDataset)

CarDataset2<- select(CarDataset_noNA,-X, -brand,
                           -model, -color, -vin, -lot,
                           -state, -country, -condition)
CarDataset2$title_status<- factor(CarDataset2$title_status)
summary(CarDataset2)
```

```
##      price            year                 title_status     mileage
##  Min.   :    0   Min.   :1973   clean vehicle    :2336   Min.   :      0
##  1st Qu.:10200   1st Qu.:2016   salvage insurance: 163   1st Qu.:  21467
##  Median :16900   Median :2018                            Median :  35365
##  Mean   :18768   Mean   :2017                            Mean   :  52299
##  3rd Qu.:25556   3rd Qu.:2019                            3rd Qu.:  63473
##  Max.   :84900   Max.   :2020                            Max.   :1017936
```

```
lm_info_car<- lm(price~ year+mileage+title_status, data= CarDataset2)
```

Fitting a Model and Assessing Multicollinearity (Appendix 2)

```
summary(lm_info_car)
```

```
##
## Call:
## lm(formula = price ~ year + mileage + title_status, data = CarDataset2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20865  -7270  -1911   5269  64299
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1.498e+06  1.678e+05  -8.928  < 2e-16 ***
## year                           7.534e+02  8.313e+01   9.064  < 2e-16 ***
## mileage                       -4.177e-02  4.519e-03  -9.243  < 2e-16 ***
## title_statussalvage insurance -7.525e+03  1.038e+03  -7.249 5.58e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 10660 on 2495 degrees of freedom
## Multiple R-squared:  0.227,  Adjusted R-squared:  0.2261
## F-statistic: 244.3 on 3 and 2495 DF,  p-value: < 2.2e-16
```

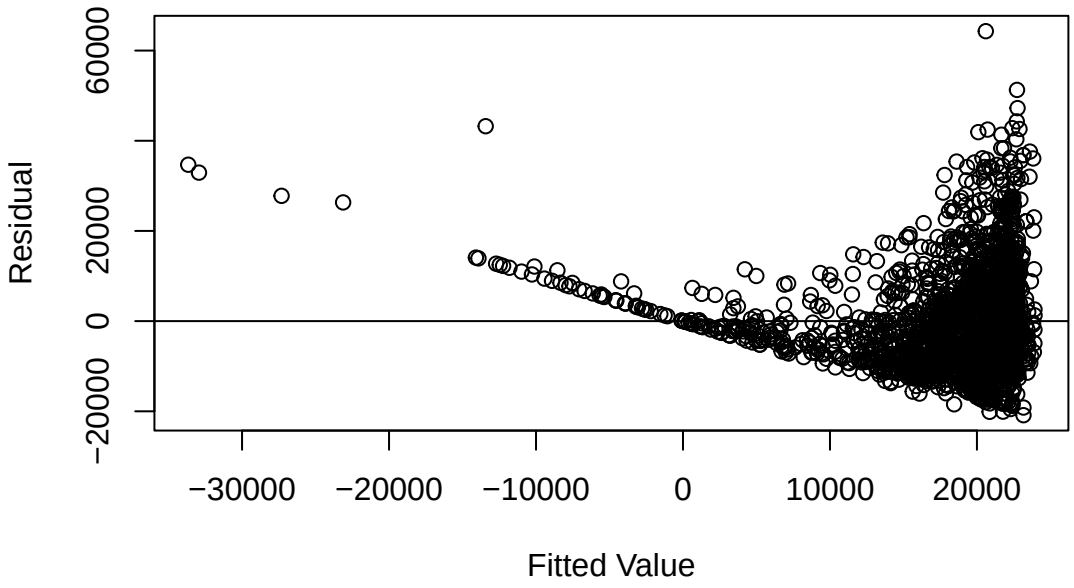
$\hat{Price} = -1498000 + 753.4_{Year_i} - 0.04177_{mileage_i} - 7525_{titlestatus_i}$

```
vif(lm_info_car)
```

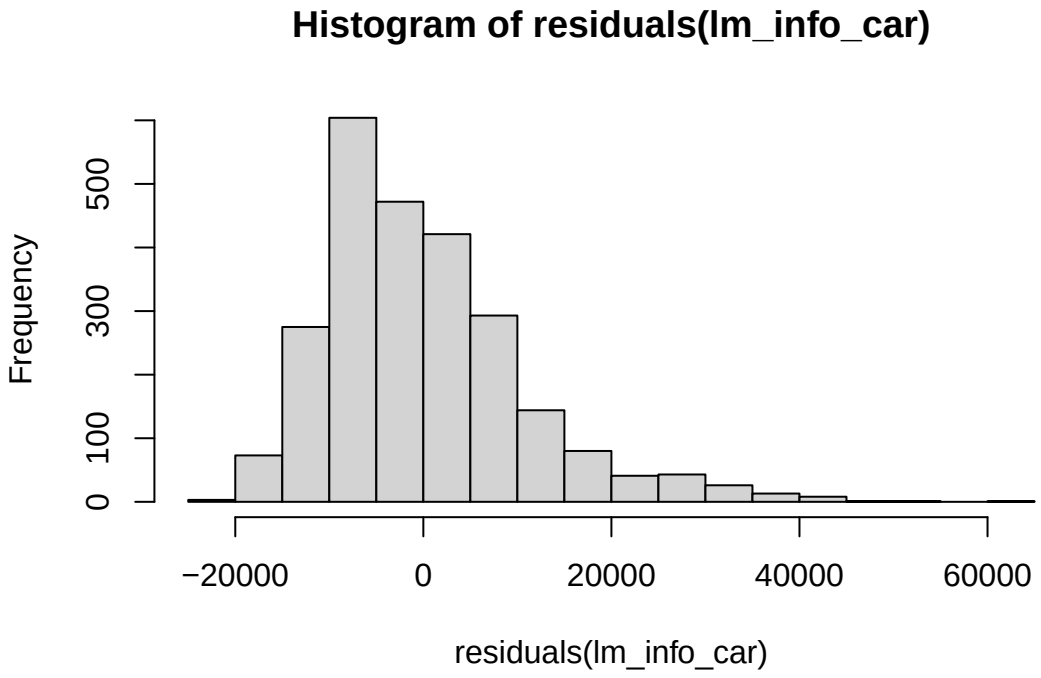
```
##         year      mileage title_status
##     1.800746     1.600567     1.445434
```

Checking Assumptions of the Procedure (Appendix 3)

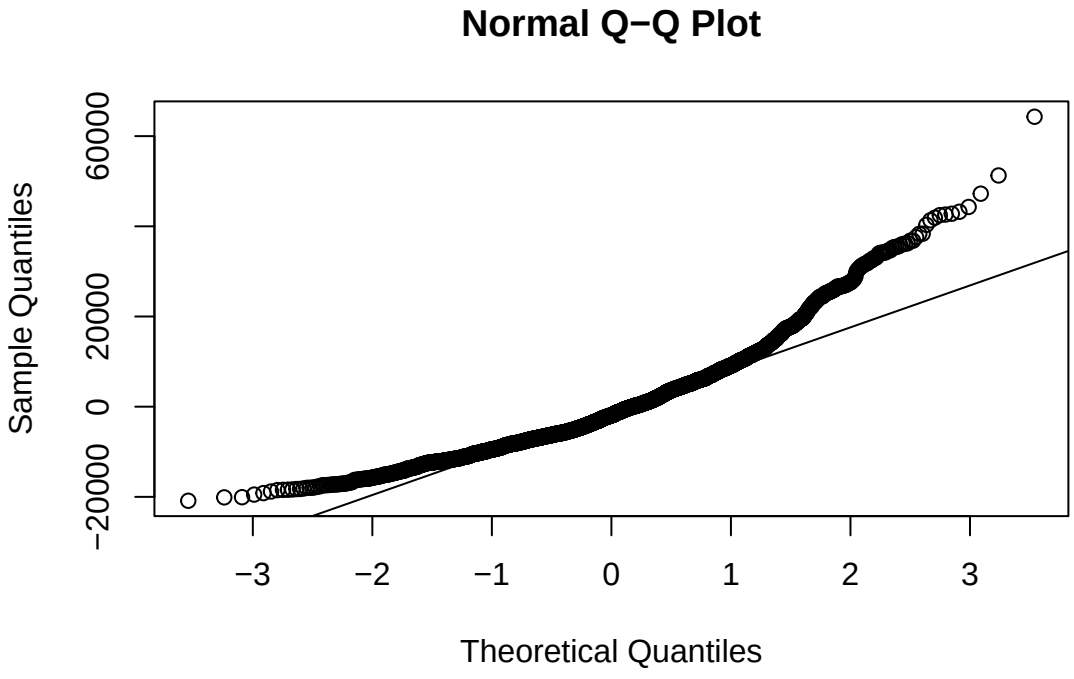
```
plot(x = lm_info_car$fitted.values, y = (lm_info_car$residuals),
xlab = "Fitted Value", ylab = "Residual"
)
abline(a=0,b=0)
```




```
hist(residuals(lm_info_car))
```




```
qqnorm(lm_info_car$residuals)
qqline(lm_info_car$residuals)
```



F test to determine whether the model is useful in explaining car prices(Appendix 4)

```
summary(lm_info_car)
```

```
##
## Call:
## lm(formula = price ~ year + mileage + title_status, data = CarDataset2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20865  -7270  -1911   5269  64299
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1.498e+06  1.678e+05  -8.928  < 2e-16 ***
## year                           7.534e+02  8.313e+01   9.064  < 2e-16 ***
## mileage                       -4.177e-02  4.519e-03  -9.243  < 2e-16 ***
## title_statussalvage insurance -7.525e+03  1.038e+03  -7.249 5.58e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10660 on 2495 degrees of freedom
## Multiple R-squared:  0.227,  Adjusted R-squared:  0.2261
## F-statistic: 244.3 on 3 and 2495 DF,  p-value: < 2.2e-16
```

Determining whether Year, Mileage, and Title Status are related to the car prices by testing slope parameters(Appendix 5) and R^2 value of the model

```
summary(lm_info_car)
```

```
##
## Call:
## lm(formula = price ~ year + mileage + title_status, data = CarDataset2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20865  -7270  -1911   5269  64299
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1.498e+06  1.678e+05  -8.928  < 2e-16 ***
## year                           7.534e+02  8.313e+01   9.064  < 2e-16 ***
## mileage                       -4.177e-02  4.519e-03  -9.243  < 2e-16 ***
## title_statussalvage insurance -7.525e+03  1.038e+03  -7.249 5.58e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10660 on 2495 degrees of freedom
## Multiple R-squared:  0.227,  Adjusted R-squared:  0.2261
## F-statistic: 244.3 on 3 and 2495 DF,  p-value: < 2.2e-16
```