Homework 1

Marta Font 1604517 and Nora Alquézar 1671727

March 18, 2024

1

1.1 Provide basic information about variables in the dataset

First, we read all the data provided by Cars.xlsx using the following code:

```
data <- read.xlsx("cars.xlsx")
```

Then, we display the first few rows of the data:

```
head(data)
```

```
> head(data)
  year   fuel seller_type transmission               owner     brand km_driven selling_price
1 2012 Diesel  Individual       Manual         First Owner   Hyundai       100           600
2 2014 Diesel  Individual       Manual Second & Above Owner     Honda       141           450
3 2016 Petrol  Individual       Manual         First Owner   Hyundai        25           550
4 2015 Petrol  Individual       Manual         First Owner   Hyundai        25           850
5 2015 Petrol  Individual       Manual         First Owner Chevrolet        35           260
6 2018 Petrol      Dealer    Automatic         First Owner    Toyota        25          1650
```

Lastly, we provide summary statistics for the data:

```
summary(data)
```

```
> summary(data)
      year          fuel            seller_type        transmission          owner              brand             km_driven       selling_price
 Min.   :1998   Length:1775        Length:1775        Length:1775        Length:1775        Length:1775        Min.   :  1.00   Min.   :  20.0
 1st Qu.:2011   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.: 36.00   1st Qu.: 250.0
 Median :2014   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 60.00   Median : 400.0
 Mean   :2013                                                                                                  Mean   : 69.04   Mean   : 558.4
 3rd Qu.:2016                                                                                                  3rd Qu.: 90.00   3rd Qu.: 650.0
 Max.   :2020                                                                                                  Max.   :400.00   Max.   :8900.0
```

2

2.1 Multiple linear regression with selling price as the response

All the other variables except brand as predictors.

```
my_model  <- lm(selling_price ~ year + fuel + seller_type+
transmission + owner + km_driven, data)
my_model
```

```
> my_model <- lm(selling_price ~ year + fuel + seller_type + transmission + owner + km_driven, data)
> my_model

Call:
lm(formula = selling_price ~ year + fuel + seller_type + transmission +
    owner + km_driven, data = data)

Coefficients:
           (Intercept)                    year            fuelElectric                 fuelLPG              fuelPetrol
            -8.962e+04               4.528e+01              -8.259e+02              -2.892e+02              -2.851e+02
 seller_typeIndividual      transmissionManual  ownerSecond & Above Owner              km_driven
            -3.790e+01              -8.721e+02              -3.624e+01              -8.112e-01
```

2.2 Model inequation form

$$\text{selling\_price} = \beta_0+\beta_1\cdot\text{year}+\beta_2\cdot\text{fuel}+\beta_3\cdot\text{seller\_type}+\beta_4\cdot\text{transmission}+\beta_5\cdot\text{owner}+\beta_6\cdot\text{km\_driven}+\epsilon$$

Having:
β_0: as the intercept, β_1:as the coefficient for year, β_2: as the coefficient for fuel, β_3: as the coefficient for seller type, β_4: as the coefficient for transmission, β_5: as the coefficient for owner, β_6: as the coefficient for km driven, ϵ: as the error term.

2.3 Design Matrix

```
X <- model.matrix(my_model)
X
```

```
> X <- model.matrix(my_model)
> X
   (Intercept) year fuelElectric fuelLPG fuelPetrol seller_typeIndividual transmissionManual ownerSecond & Above Owner km_driven
1            1 2012            0       0          0                     1                  1                         0   100.000
2            1 2014            0       0          0                     1                  1                         1   141.000
3            1 2016            0       0          1                     1                  1                         0    25.000
4            1 2015            0       0          1                     1                  1                         0    25.000
5            1 2015            0       0          1                     1                  1                         0    35.000
6            1 2018            0       0          1                     0                  0                         0    25.000
7            1 2019            0       0          0                     0                  1                         0     5.000
8            1 2013            0       0          0                     1                  1                         1    33.000
9            1 2014            0       0          0                     0                  0                         0    28.000
10           1 2013            0       0          0                     0                  0                         0    59.000
11           1 2011            0       0          0                     0                  0                         0   175.900
12           1 2018            0       0          1                     0                  1                         0    14.500
13           1 2013            0       0          0                     0                  0                         0    50.000
14           1 2012            0       0          0                     0                  0                         1    33.800
15           1 2011            0       0          0                     0                  0                         1   130.400
16           1 2016            0       0          1                     1                  0                         0    50.000
17           1 2015            0       0          0                     1                  1                         0    80.000
18           1 2019            0       0          1                     1                  1                         0    10.000
19           1 2010            0       0          1                     0                  0                         0   119.000
20           1 2014            0       0          1                     1                  1                         1    60.000
21           1 2013            0       0          0                     0                  0                         1    75.800
22           1 2009            0       0          0                     0                  0                         1    78.000
23           1 2012            0       0          0                     1                  1                         0    40.000
24           1 2014            0       0          0                     1                  1                         1    74.000
25           1 2009            0       0          1                     1                  1                         1    79.000
26           1 2019            0       0          1                     1                  1                         0    15.000
27           1 2018            0       0          0                     0                  1                         0    29.000
28           1 2014            0       0          0                     0                  1                         0    70.000
29           1 2014            0       0          0                     0                  1                         0    90.000
30           1 2014            0       0          1                     0                  1                         1    73.300
31           1 2014            0       0          0                     0                  1                         0    92.000
32           1 2010            0       0          0                     1                  1                         1   350.000
33           1 2011            0       0          0                     1                  1                         0   230.000
34           1 2018            0       0          1                     1                  1                         0    31.000
35           1 2009            0       0          0                     1                  1                         1   120.000
36           1 2017            0       0          0                     1                  1                         0    35.000
37           1 2007            0       0          1                     1                  0                         1    54.000
38           1 2010            0       0          1                     1                  1                         1    63.000
39           1 2014            0       0          0                     1                  1                         1   120.000
40           1 2005            0       0          1                     1                  1                         1   120.000
41           1 2014            0       0          1                     1                  1                         1    76.000
```

[Omitted 1664 rows]

2.4 Interpretation of each coefficient in the model

The intercept (β_0) represents the expected selling price when all predictor variables are set to zero.

The coefficient for the variable "Year" (β_1) indicates how the expected value of the selling price changes for each unit increase in the year of manufacture of the car, while holding all other variables constant in the model.

For the fuel-related coefficients (Electric, LPG, Petrol) (β_2), they indicate the effect of each fuel type on the expected selling price compared to a reference fuel type (which is not specified in your data). A positive coefficient suggests that the associated fuel type tends to increase the expected selling price compared to the reference fuel type.

The coefficient for the "Seller Type" variable (β_3)represents how the expected selling price changes when the seller is an individual seller compared to another type of seller.

Similarly, the coefficient for the "Transmission" variable (β_4) indicates how the expected selling price changes when the car's transmission is manual compared to another type of transmission.

The coefficient for the "Second Hand or Above Owner" variable (β_5) indicates how the expected selling price changes when the car owner is second-hand or above compared to another type of owner, presumably a first-hand owner.

Lastly, the coefficient for the "Kilometers Driven" variable (β_6) indicates how much the expected selling price changes for each unit increase in the number of kilometers driven, while holding all other variables constant in the model.

2.5 Comment on the output of the summary() function

```
summary(my_model)
```

```
> summary(my_model)

Call:
lm(formula = selling_price ~ year + fuel + seller_type + transmission +
    owner + km_driven, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1251.0  -186.1   -39.2   125.7  7520.4

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -8.962e+04  7.598e+03 -11.795  < 2e-16 ***
year                      4.528e+01  3.766e+00  12.024  < 2e-16 ***
fuelElectric             -8.259e+02  4.737e+02  -1.743  0.08143 .
fuelLPG                  -2.892e+02  1.581e+02  -1.829  0.06756 .
fuelPetrol               -2.851e+02  2.489e+01 -11.453  < 2e-16 ***
seller_typeIndividual    -3.790e+01  2.663e+01  -1.423  0.15479
transmissionManual       -8.721e+02  3.393e+01 -25.700  < 2e-16 ***
ownerSecond & Above Owner -3.624e+01  2.710e+01  -1.338  0.18120
km_driven                -8.112e-01  3.119e-01  -2.601  0.00937 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 471.3 on 1766 degrees of freedom
Multiple R-squared:  0.4273,    Adjusted R-squared:  0.4247
F-statistic: 164.7 on 8 and 1766 DF,  p-value: < 2.2e-16
```

This summery enables us to assess the factors that affect the selling prices, determining wether they increase or decrease the cost.

Based on the collected data, the average person in the sample would receive a reduction in the selling cost (represented by the Intercept estimate).

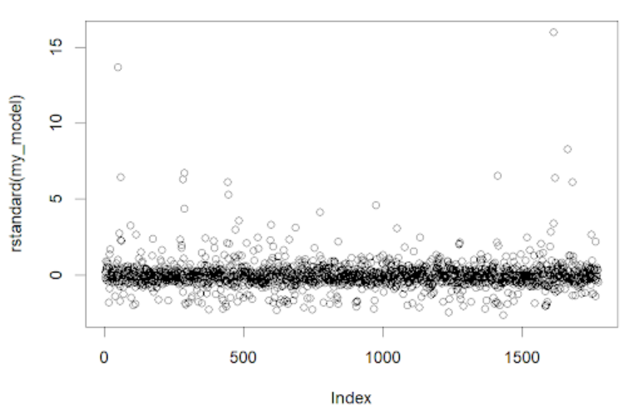
When analyzing each variable's effect on selling prices, we found that the year tend to increase the cost. In contrast, all the other predictors tend to decrease the cost.

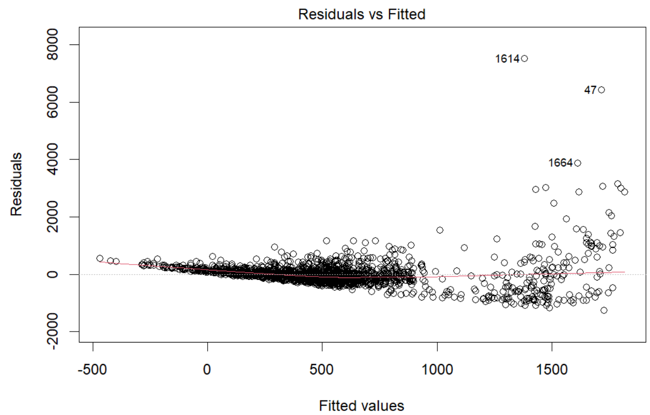
If we look at the p-value of each coefficient, those containing *** (intercept, year, fuelPetrol, ownerSecond) are the most significant ones, meaning that the value associated with the coefficient is lower than a predetermined significance level (0.05). Therefore, the coefficient is considered significant, and thus, we can reject the null hypothesis. Hence, there is sufficient evidence in the data to suggest a significant relationship between the corresponding predictor variables and the response variable. However, for the values fuelElectric and fuelLPG, there would not be enough evidence to reject the null hypothesis, and the coefficient is considered not significant. Thus, we cannot conclude that there is a relationship between the variables. Then, if we look at the Residual standard error: 471.3, on average, the actual values of the response variable can deviate approximately 471.3 units from the model predictions. Multiple R-squared: 0.4273, (proportion of variability in the response variable) 42.73% of the variability in the response variable can be explained by the set of predictor variables.

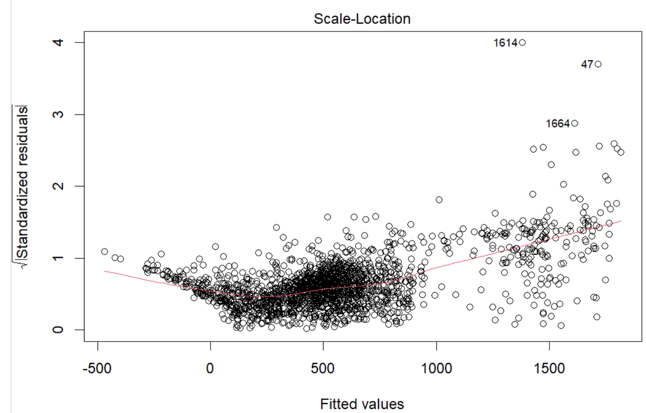
F-statistic: 164.7 on 8 and 1766 DF: Dividing the variance explained by the model by the unexplained variance by the model, when it is larger, it indicates more evidence that at least one of the independent variables is significantly different from zero in predicting the response variable.

2.6 Plot of the residuals of the model

```
plot(my_model)
```







3

3.1 Use a logarithmic transformation in the multiple linear regression model

```
model_original <- lm(selling_price ~ year + fuel + seller_type
+ transmission + owner+ km_driven, data = data)
data$log_selling_price <- log(data$selling_price)
model_transformed <- lm(log_selling_price ~ year + fuel + seller_type +
transmission + owner+ km_driven, data = data)
summary(model_original)
summary(model_transformed)
```

```
> summary(model_original)

Call:
lm(formula = selling_price ~ year + fuel + seller_type + transmission +
    owner + km_driven, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1251.0  -186.1   -39.2   125.7  7520.4

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -8.962e+04  7.598e+03 -11.795  < 2e-16 ***
year                       4.528e+01  3.766e+00  12.024  < 2e-16 ***
fuelElectric              -8.259e+02  4.737e+02  -1.743  0.08143 .
fuelLPG                   -2.892e+02  1.581e+02  -1.829  0.06756 .
fuelPetrol                -2.851e+02  2.489e+01 -11.453  < 2e-16 ***
seller_typeIndividual     -3.790e+01  2.663e+01  -1.423  0.15479
transmissionManual        -8.721e+02  3.393e+01 -25.700  < 2e-16 ***
ownerSecond & Above Owner -3.624e+01  2.710e+01  -1.338  0.18120
km_driven                 -8.112e-01  3.119e-01  -2.601  0.00937 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 471.3 on 1766 degrees of freedom
Multiple R-squared:  0.4273,    Adjusted R-squared:  0.4247
F-statistic: 164.7 on 8 and 1766 DF,  p-value: < 2.2e-16
```

```
> summary(model_transformed)

Call:
lm(formula = log_selling_price ~ year + fuel + seller_type +
    transmission + owner + km_driven, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.60364 -0.30256 -0.01285  0.29380  2.29119

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -2.307e+02  7.563e+00 -30.511  < 2e-16 ***
year                       1.181e-01  3.749e-03  31.507  < 2e-16 ***
fuelElectric              -4.037e-01  4.715e-01  -0.856 0.391942
fuelLPG                   -5.681e-01  1.574e-01  -3.610 0.000315 ***
fuelPetrol                -4.531e-01  2.478e-02 -18.287  < 2e-16 ***
seller_typeIndividual     -1.182e-01  2.650e-02  -4.461 8.66e-06 ***
transmissionManual        -8.411e-01  3.377e-02 -24.905  < 2e-16 ***
ownerSecond & Above Owner -5.158e-02  2.697e-02  -1.913 0.055940 .
km_driven                  2.816e-04  3.104e-04   0.907 0.364383
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4691 on 1766 degrees of freedom
Multiple R-squared:  0.6431,    Adjusted R-squared:  0.6415
F-statistic: 397.8 on 8 and 1766 DF,  p-value: < 2.2e-16
```

Upon comparing the two models, the transformed model demonstrates a slightly higher Multiple R-squared value of 0.6431 and an Adjusted R-squared value of 0.6415. In contrast, the original model shows lower values, with Multiple R-squared at 0.4273 and Adjusted R-squared at 0.4247.

4

4.1 Fit a selling price as a function of year using a second order polynomial

```
model_poly <- lm(selling_price ~ poly(year, 2), data = data)
summary(model_poly)
new_data <- data.frame(year = c(2007, 2017))
predictions <- predict(model_poly, newdata = new_data, interval = "confidence",
level = 0.95)
predictions
```

```
> predictions
       fit      lwr      upr
1 201.1051 148.6829 253.5272
2 825.7023 786.0500 865.3546
```

From the provided predictions and their associated 95% confidence intervals:

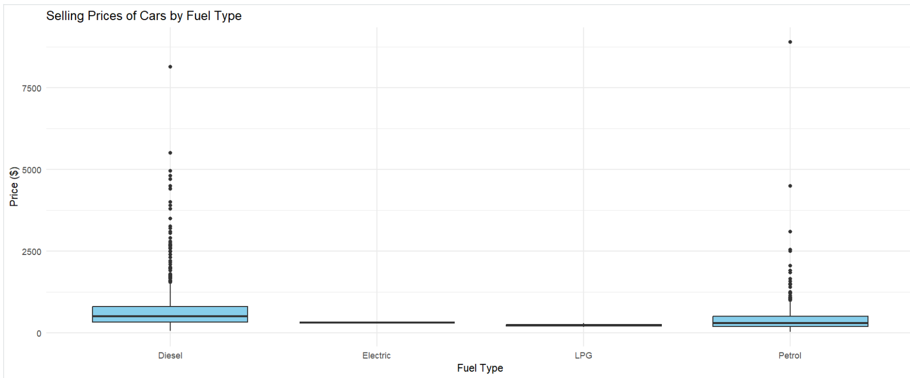
- For a car from 2007, the predicted selling price is approximately $201.11 thousand, with a 95% confidence interval spanning from approximately $148.68 thousand to $253.53 thousand.
- For a car from 2017, the predicted selling price is approximately $825.70 thousand, accompanied by a 95% confidence interval ranging from approximately $786.05 thousand to $865.35 thousand.

5

5.1 Boxplot

To generate the boxplot:

```
boxplot(selling_price ~ fuel, data = data,
        main = "Boxplot of Selling Prices by Fuel Type",
        xlab = "Fuel Type", ylab = "Selling Price")
```



Comment: We see a clear difference between LPG and Electric fuels compared to Diesel and Petrol. Clearly, Diesel has the highest prices.

We realize the ANOVA test:

```
anova_result <- aov(selling_price ~ fuel, data = data)
summary(anova_result)
```

```
> summary(anova_result)
              Df    Sum Sq  Mean Sq F value Pr(>F)    
fuel           3  43291791 14430597   39.84 <2e-16 ***
Residuals   1771 641528511   362241                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```