Team Essay 1

Sahil, Ariel, Ilya, Lekha, Anthony, Trang

2/25/2021

Introduction

Over the years, there have been tremendous technological advancements in the functioning, developement, and even the presentation of various modes of transportation. Accordingly, there has been a significant economic growth observed in the motor industry, contributing to the overall rise in car prices today, including the prices of cars manufactured by the Mazda Motor Corporation, as the models get older and older. Keeping in mind the future of the motor industry and Mazda's business, it is important to consider how the company's car prices have fluctuated in the past with time, and study the nature of this fluctuation in order to predict the possible state of car prices in the future.

In practical support of this, our team has built a simple linear regression model that evaluates the impact that time imposes on the prices of Mazda cars. Using the necessary programming functions and mathematical formulas required, we have taken a dataset of 124 Mazda cars, as taken from the classified section of the Melbourne Age during 1991.

To construct the model, we considered the predictor variable, X, to be the time (in years), and the response variable, Y, to be the price of the car.We've also given graphical representations depicting the relationship between price and the time (in years), taking both from our original dataset and from the lens of our constructed model. In these graphs, our axis scales are as follows: + x-axis scale: time (in years) + y-axis scale: price (in dollars)

Formula and Basics

Linear regression: `y = b0 + b1*x + e`

- *b0* is the intercept of regression line: predicted value when $x = 0$.
- *b1* is the slope of the regression line.
- *e* is the error term, the part of y that can be explained through the regression model.

Loading required R packages

readxl: to read data from xl sheet | tidyverse: for data manipulation and visualization

```
library("readxl")
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data Description

Examples of data and problem

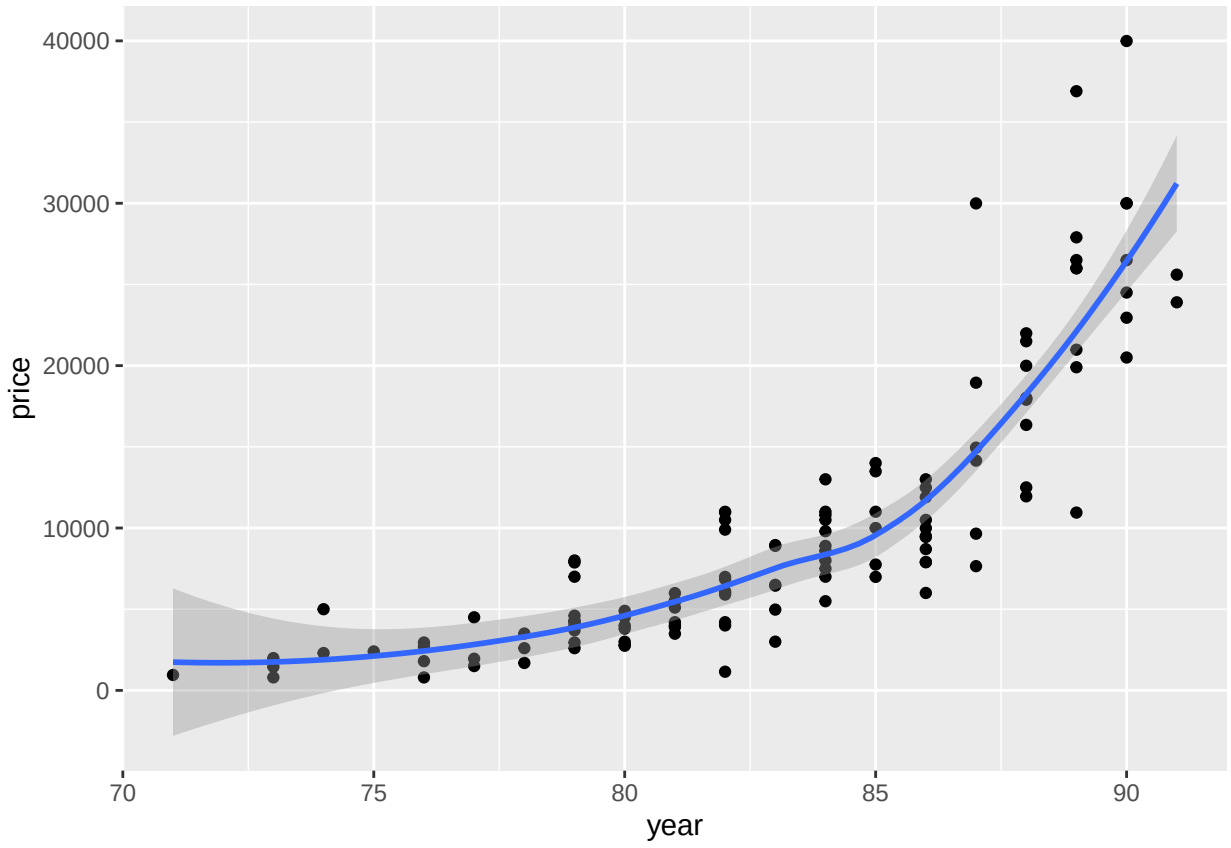
```
mazda_data <- read_excel("Mazda.xlsx")
mazda_data <- mazda_data %>%
  rename(
    "year" = "Year",
    "price" = "Price")
mazda_data <- mazda_data[1:122, ]
mazda_data
```

```
## # A tibble: 122 x 2
##     year price
##    <dbl> <dbl>
##  1    79  2950
##  2    82  5900
##  3    83  2999
##  4    88 11950
##  5    82  6100
##  6    90 26500
##  7    82  6850
##  8    81  5490
##  9    86 12999
## 10    82  9900
## # ... with 112 more rows
```

Visualization

```
ggplot(mazda_data, aes(x = year, y = price)) +
  geom_point() +
  stat_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
cor(mazda_data$price, mazda_data$year)
```

```
## [1] 0.8112544
```

The correlation coefficient measures the level of the association between two variables x and y. Its value ranges between -1 (perfect negative correlation: when x increases, y decreases) and +1 (perfect positive correlation: when x increases, y increases). In our example, the correlation coefficient is large enough, so we can continue by building a linear model of y as a function of x.

Analysis

Computation

- model1: `price = b0 + b1 * year`

```
model1 <- lm(price ~ year, data = mazda_data)
model1
```

```
##
## Call:
## lm(formula = price ~ year, data = mazda_data)
##
## Coefficients:
## (Intercept)         year
##     -112466         1476
```

The results show the intercept and the beta coefficient for the Year variable.

Interpretation of Model

From the output above for model1:

We can see that the resulting regression equation is:

- price = -112466 + (1476) * year
- b0 = -112466
- b1 = 1476

Using this model we can predict that the price of a new Mazda vehicle will increase about $1476 each year.

Model1: `price = -112466 + (1476) * year`

1. Year = 90

- Estimate of Price: 20374
- Actual: 22950

2. Year = 90

- Estimate of Price: 20374
- Actual: 24500

3. Year = 91

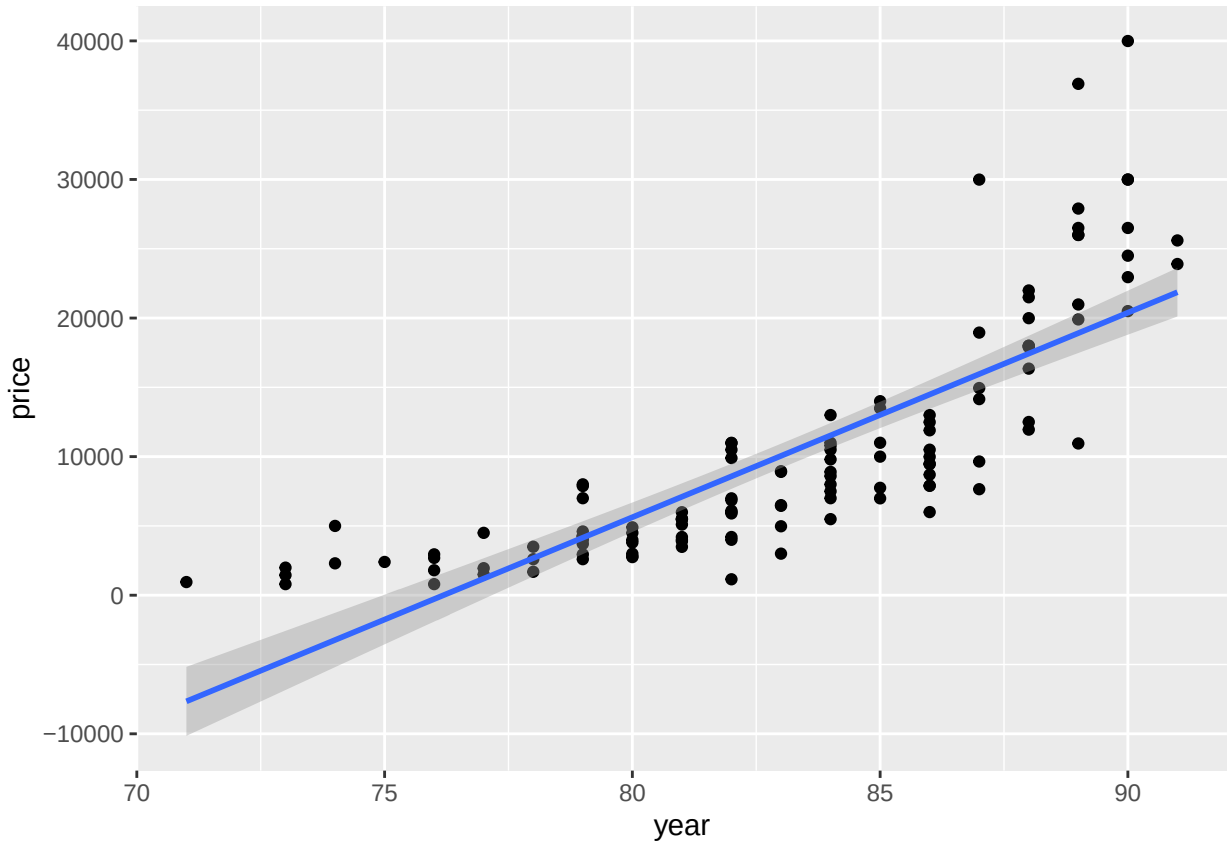
- Estimate of Price: 21850
- Actual: 23900

Model Evaluation

Regression Line

```
ggplot(mazda_data, aes(x = year, y = price)) +
  geom_point() +
  stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ year, data = mazda_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8483  -2923  -1020   2413  19603
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -112465.66    8084.31  -13.91   <2e-16 ***
## year           1476.14      97.12   15.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4903 on 120 degrees of freedom
## Multiple R-squared:  0.6581, Adjusted R-squared:  0.6553
## F-statistic:   231 on 1 and 120 DF,  p-value: < 2.2e-16
```

```
summ_model1 <- summary(model1)

confint(model1, level = 0.95)
```

```
##                   2.5 %     97.5 %
```

```
## (Intercept) -128472.038 -96459.283
## year           1283.854   1668.436
```

```
percent_error <-sigma(model1)*100/mean(mazda_data$price)
print("This is the percent error of our model:")
```

```
## [1] "This is the percent error of our model:"
```

```
percent_error
```

```
## [1] 47.95486
```

```
print("This is the coefficient of determination of our model")
```

```
## [1] "This is the coefficient of determination of our model"
```

```
summary(model1)$r.squared
```

```
## [1] 0.6581337
```

```
print("This is the mean squared error of our model")
```

```
## [1] "This is the mean squared error of our model"
```

```
mean(summ_model1$residuals^2)
```

```
## [1] 23643224
```

Model Assesment

Notice that despite the car model's year being a significant predictor of the price of the car, the regression line does not fit well at all with a very high MSE of 24346553 and the price of each car deviates from the regression line by an average of $96.75. Also, only around 64% of the price's variance is explained by the car model's year in this regression model. What makes this model insufficient may be the outliers which are the more recently made mazda cars that cost much higher than any car preceding it.

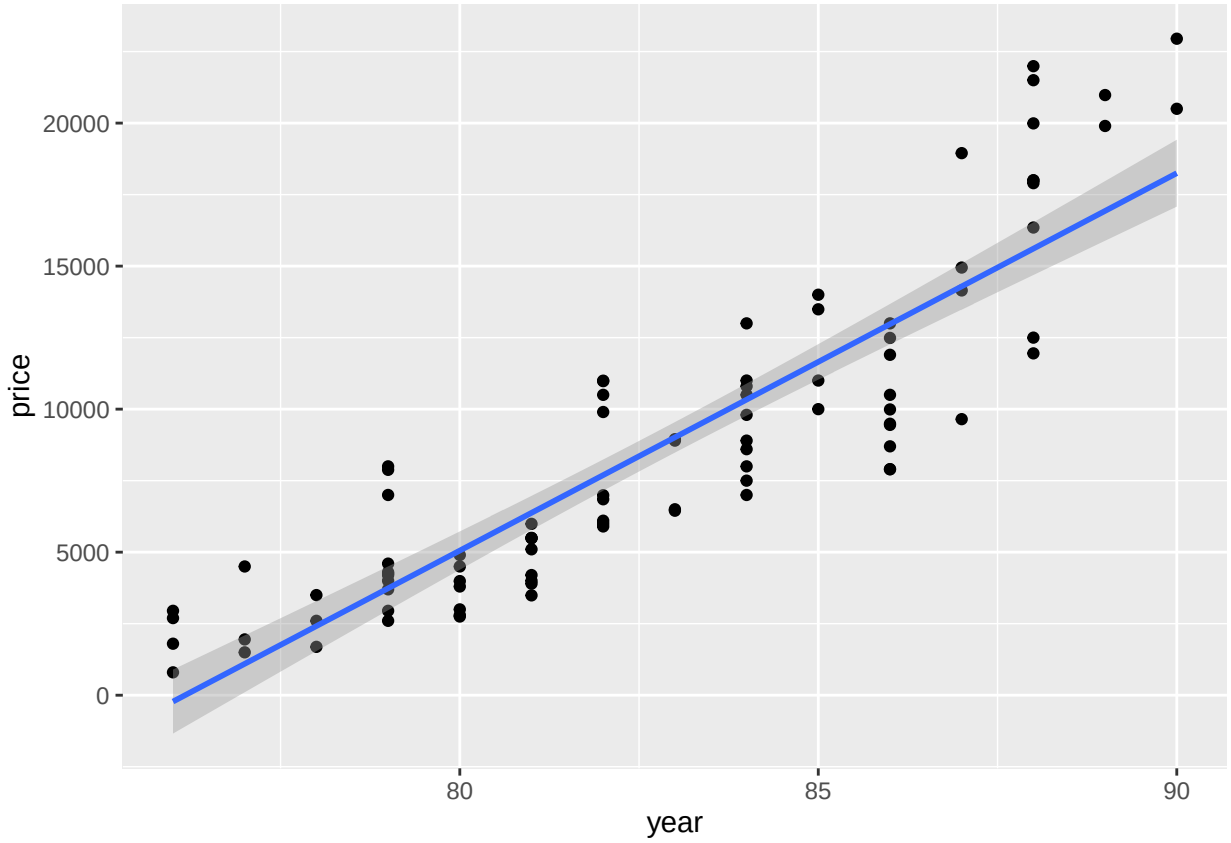
Model Adjustment

The removal of a portion of outliers from the Mazda dataset resulted in significant improvements to our model. The new Mazda dataset saw positive changes in areas such as Residual Standard Error, R-squared, t-statistic, F-statistic and MSE.

```
mazda_data2 <- read_excel("Mazda2.xlsx")
mazda_data2 <- mazda_data2 %>%
  rename(
    "year" = "Year",
    "price" = "Price")
mazda_data2 <- mazda_data2[1:91, ]
model2 <- lm(price ~ year, data = mazda_data2)

ggplot(mazda_data2, aes(x = year, y = price)) +
  geom_point() +
  stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
print("This is the summary of our second model:")
```

```
## [1] "This is the summary of our second model:"
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = price ~ year, data = mazda_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5071.8 -1763.6  -155.5  2113.9  6379.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -100495.11    6035.12  -16.65   <2e-16 ***
## year           1319.38      72.81   18.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2558 on 89 degrees of freedom
## Multiple R-squared:  0.7867, Adjusted R-squared:  0.7843
## F-statistic: 328.3 on 1 and 89 DF,  p-value: < 2.2e-16
```

```r
confint(model2, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) -112486.766 -88503.450
## year           1174.703   1464.062
```

```r
percent_error2 <-sigma(model2)*100/mean(mazda_data2$price)
print("This is the percent error of our second model:")
```

```
## [1] "This is the percent error of our second model:"
```

```r
percent_error2
```

```
## [1] 29.22869
```

```r
print("This is the coefficient of determination of our second model")
```

```
## [1] "This is the coefficient of determination of our second model"
```

```r
summary(model2)$r.squared
```

```
## [1] 0.7867399
```

```r
summ_model2 <- summary(model2)
print("This is the MSE of our model:")
```

```
## [1] "This is the MSE of our model:"
```

```r
mean(summ_model2$residuals^2)
```

```
## [1] 6401007
```

Model with New Dataset Assesment

- RSE: Eliminating some outliers also benefited the Residual Standard Error (RSE).
  - In the model containing all the outliers we were seeing: RSE = 4920
  - Compared to the RSE of the second model without some outliers: RSE = 2558

From this we can see that our model predicted incorrectly by an average value of $2641 and a $2,334 decrease from the previous model. In general a lower RSE indicates that the model is more fitted to the data when compared to a higher RSE, thus showing that the new model is better.

- t-statistic and F-statistic: In terms of t-statistic and F-statistic, we saw an increase in both of these categories. For the t-statistic the new model resulted in a 4.36 unit increase:
  - In the old model: t-statistic = 14.91
  - In the new model: t-statistic = 18.12

  While the F-statistic has less significance on simple linear models like these, It still saw increase of 148.2 units

  - In the original model: F-statistic = 222.3
  - In the new model: F-statistic = 328.3

  Since both models had the same P-value, the increase in t-statistic demonstrates that the newer model is a better predictor than the older model. The Increase in F-statistic is positive in a general sense because it shows a high significance. It also still helps indicate that the new model would be more useful than the last if more predictors are added.

- R-squared: In the new model we saw an increase in the R-squared

  - R-squared of the old model: R^2 = 0.6494476
  - R-squared of the new model: R^2 = 0.7867399

  R-squared serves as an indicator of how well a model fits the data. In the case of the Mazda dataset, it serves as the percentage of the variance in price explained by the year each car was made. An increase of around 16% in R-squared shows that the new model aligns with the data much more than the older model did.

- MSE: The mean squared error of the new model is where the largest amount of change was seen.

  - The MSE of the old Mazda model: 23805591
  - The MSE of the new Mazda model: 6401007

  MSE is used to determine the quality of fit of a model. In most cases, lower MSE translates into a better model or a better quality fit. The decrease of 17,523,667 by the new model illustrates a positive change as well as how absurd the MSE was in the original model.

Prediction and Model Accuracy

- Model2: `price = -100495 + (1319) * year`

  In model2, we can see that the intercept went down, thus the slope fell as well. Using this model, we can predict that the price of a new Mazda vehicle will increase by $1319.

1. Year = 90

- Estimate of Price: 18215
- Actual: 22950

2. Year = 90

- Estimate of Price: 18215
- Actual: 24500

3. Year = 91

- Estimate of Price: 19534
- Actual: 23900

Conclusion

Summary

The removal of a portion of outliers from the Mazda dataset resulted in significant improvements to our model. The new Mazda dataset saw positive changes in areas such as Residual Standard Error, R-squared, t-statistic, F-statistic and MSE. These metrics give the overall quality of the model.

For model1,

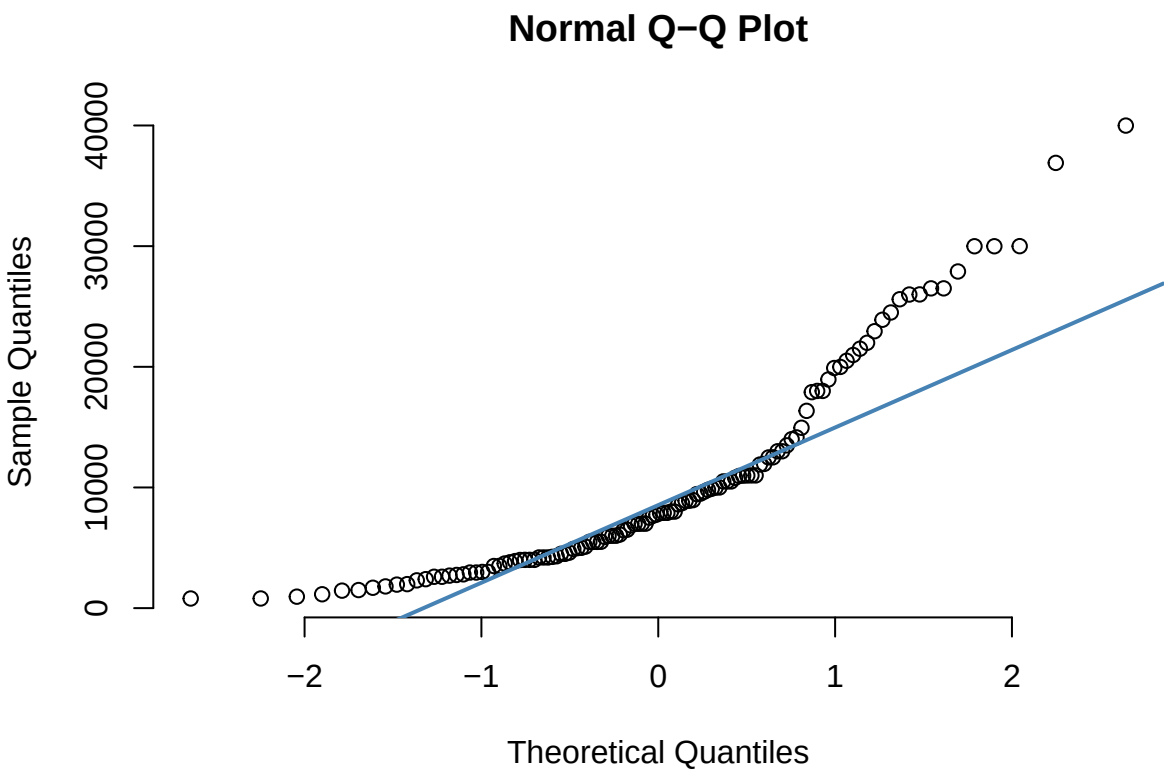
RSE = 4903 on 120 degrees of freedom. R-Squared = 0.6581 F-Statistic = 231 on 1 and 120 DF t-Statistic = 15.20 MSE = 23643224

For model2,

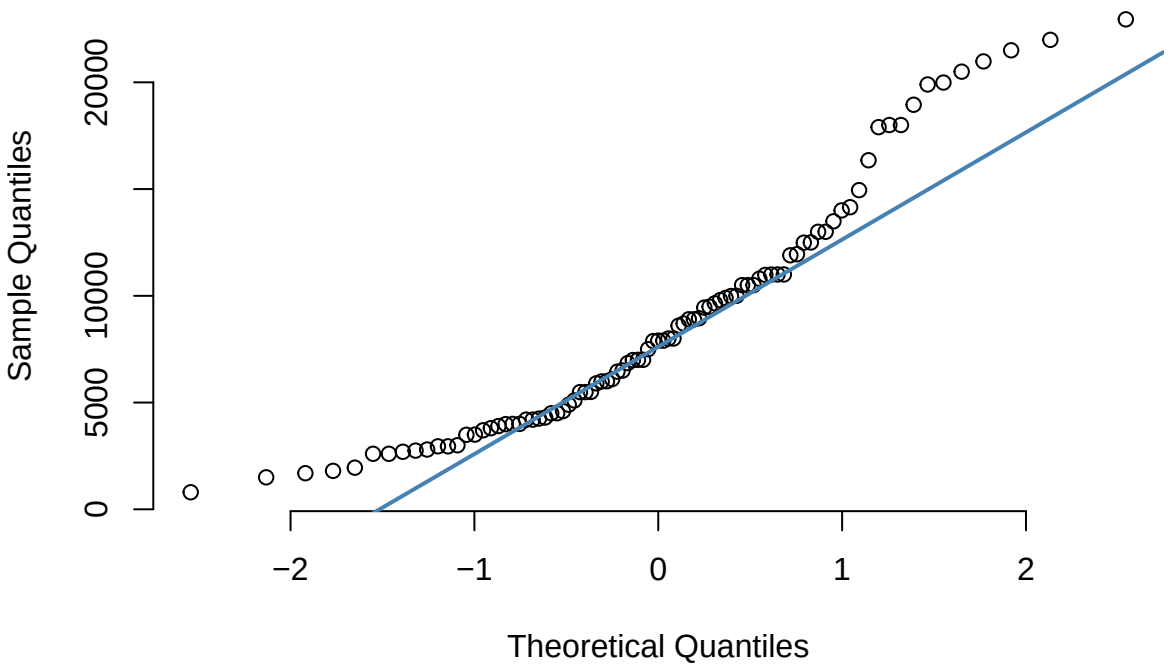
RSE = 2641 on 91 degrees of freedom. R-Squared = 0.8016 F-Statistic = 367.7 on 1 and 91 DF t-Statistic = 19.18 MSE = 6822886

Thus, we can clearly see that model2 is a better estimator than model1 since it fits the criteria for RSE being closer to zero, R-Squared being higher, F-Statistic being higher, t-Statistic being higher and MSE for model2 being very much lower than model1.

```
qqnorm(mazda_data$price, pch = 1, frame = FALSE)
qqline(mazda_data$price, col = "steelblue", lwd = 2)
```




```
qqnorm(mazda_data2$price, pch = 1, frame = FALSE)
qqline(mazda_data2$price, col = "steelblue", lwd = 2)
```



To summarize, the model2 shows a positive correlation between the number of years and Mazda cars, that is as the number of years increases, the price of Mazda cars also increases. We can say that per year, the price increases by 1319, as b1 = 1319.

Furthermore, if you look at the qqplot produced above we see that the errors/residuals are normally distributed in our model because the expected residual data points fall on the diagonal line and show less variance. This suggests that our linear regression model is a good fit for our data, but can be better.

Reference

"http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r"