# Cars Linear regression project

Aiden Tan, Brody, Steve

2022-12-05

## Introduction

Cars have become an indispensable means of transportation in our lives. My team members mentioned a idea. Nowadays, many people like to live in cities with relatively low prices, so they can buy vehicles there. My team and I decided to do a project on the topic of car prices, and we are about to create some linear models to explore this topic.

Our main data collection comes from Autotrader("Autotrader Dataset Generator — Myslu.stlawu.edu"). We collected two datasets, which come from two different states, one is NEWYORK State and the other is PENNSYLVANIA State. The main research targets of the two datasets are Nissan model, and each dataset Both have three variables which are year, price and mileage. The report mainly involves three research questions. These questions can all be summed up in one question, that is, will the price of the same car model be different in different states if consider some variables?

We merged the two different state data into a complete data, and we added a "location" variable so that we can use them easily.

## Research Question 1

### Is there a difference in price between the locations?

Our research question basically compare price of car between two places,It is different? It is better to start with one variable which is locations, since locations only have two possible outcomes, one is NEWYORK, another one is PENNSYLVANIA We gonna use binary variable in our model, NEWYORK will be represented as 0, and PENNSYLVANIA will be represented as 1. one variable linear model should be $y = \beta_0 + \beta_1 x_1$. If we plug 0 or 1 in our model:

$$\text{MODEL FOR PENNSYLVANIA:} y = \beta_0 + \beta_1$$

$$\text{MODEL FOR NEWYORK: } y = \beta_0$$

$\beta_1$ is only thing we can test, we can express these two hypotheses as the following models:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

## Exploratory data analysis

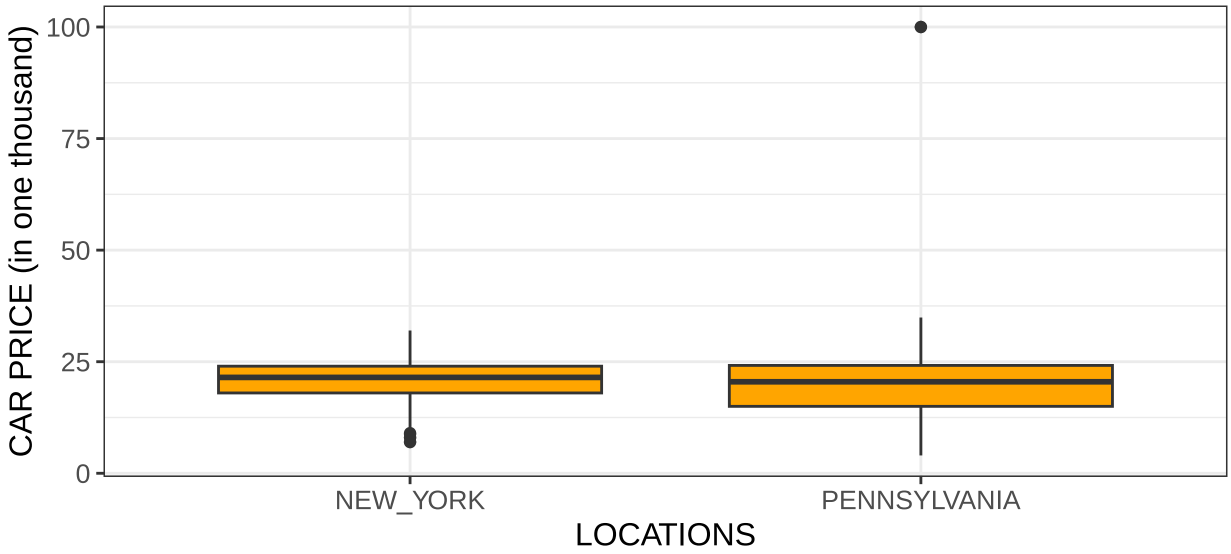

Figure 1: Box Plots of Car price in two locations

Firgure 1 shows the box plot of two locations,look at the box plot of New York, its variability of price seem less than Pennsylvania,since its length of box plot is shorter. Both box plots have outliers, The small dots that appear at the end of each box plot are what we call the maximum or minimum values and are generally called outliers.There is a black line in the middle of each box plot, and that black line represents the median, and we generally use it to compare the mean between two box plots. We think there is no much difference on mean of price between these two locations. In order to further verify our conclusion, we need to test the two models.

### Data Table

Table 1: Summary Statistics for Price

| | count | Min | Q1 | Median | Q3 | Max | Mean | SD |
|---|---|---|---|---|---|---|---|---|
| NEW_YORK | 300 | 7 | 18 | 21.48 | 24.00 | 32 | 20.87 | 4.61 |
| PENNSYLVANIA | 213 | 4 | 15 | 20.50 | 24.17 | 100 | 20.80 | 11.28 |

Table 1 is showing some statistics of car price in two places, the count is telling us the sample size,we have 3oo cars in New York data set and 213 cars in Pennsylvania data set. The minimum value of price in New York is 7,000 dollars, which is larger than minimum value of price in Pennsylvania. The maximum value of price in Pennsylvania is relatively high, it is $100,000.The lower quartile first quartile (Q1) tells us the price which 25% of data points are found when they are arranged in increasing order. The higher quartile first quartile (Q3) tells us the price which 75% of data points are found when they are arranged in increasing order.The "SD" in table is represent standard deviation.This value can tell us how the data spread. The standard deviation of New York is samller, It verifies what we said in the last part, The length of box plot shorter for New york is shorter.

The mean is important to our research question, The mean of price for New York is $20.87$,the mean of price for Pennsylvania is $20.80$. The difference between mean is $0.07$, The difference is so small that we can now guess that locations do not affect car price.

### Fit model

Table 2: Linear Model Table for Car price

| Parameter | Coefficient | SE | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | 20.8723 | 0.4663 | 44.7581 | 0.000 |
| locationPENNSYLVANIA | -0.0746 | 0.7237 | -0.1030 | 0.918 |

Table 2 shows a table for linear model,for this research question, we are suing simple t-test. restate our hypotheses which is:$H_0 : \beta_1 = 0$ AND $H_a : \beta_1 \neq 0$. Our t-value of $\beta_1$ is $-0.1030$, that means our p-value will be very large. The p-value is $0.918$ is larger than significant level $0.05$, then we could say we fail to reject our $H_0$, we have no significant evidence to say there is a difference in price between the locations.

## Research Question 2

### Is there a difference in price between the locations after count mileage?

First research questions, we only consider locations, if we consider mileage and location together,we add mileage to our model,Is it result same as first research question? That is exactly what we want to do in this part. NEWYORK will be represented as $0$, and PENNSYLVANIA will be represented as $1$,we use $x_2$ represent mileage of car, then two variables linear model should be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

$$\text{MODEL FOR PENNSYLVANIA:} y = (\beta_0 + \beta_1) + \beta_2 x_2$$

$$\text{MODEL FOR NEWYORK: } y = \beta_0 + \beta_2 x_2$$

$\beta_1$ is still our test goal, we can express these two hypotheses as the following models:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

### Exploratory data analysis

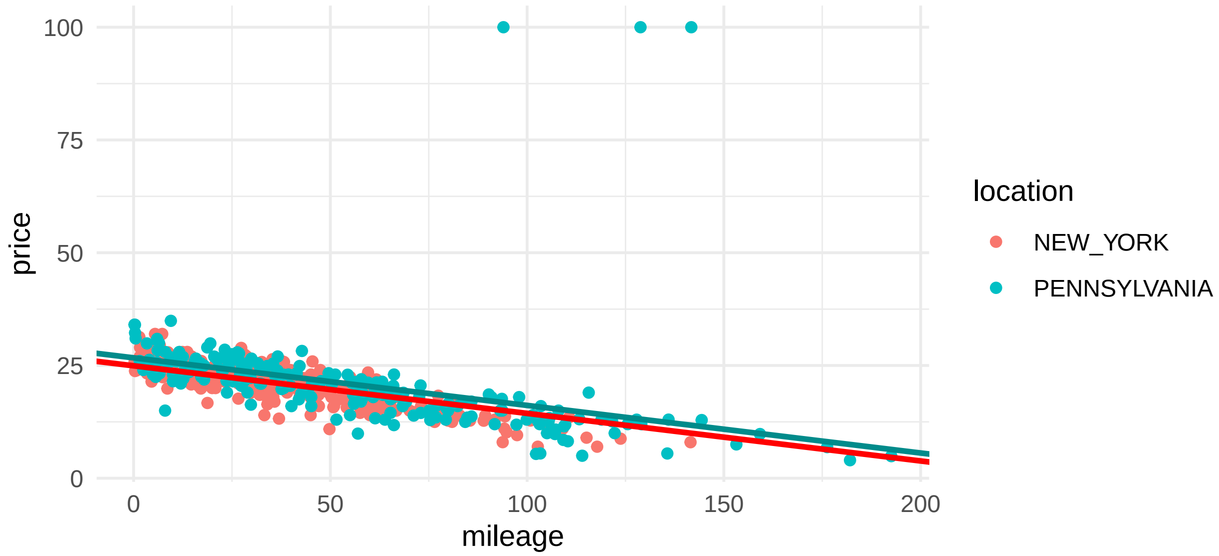

Figure 2: Dot plot of Car price by two locations

Figure 2 shows a dot plot between price and mileage by two locations.The cyan dots are represent Pennsylvania, the red are New York.The distribution of dots is very dense, and it is difficult for us to see whether the prices of two places are different from the dots. There are two lines in dot plot, The cyan line is model for New York, red line is Pennsylvania's model, they look like parallel, that makes sense, since we stated two models at beginning, that have same slope $\beta_2$.The gap between the two lines is actually $\beta_1$, since the difference between the two models is $\beta_1$. The following task is to test whether this $\beta_1$ is significant.

### Fit model

Table 3: Linear Model Table for Car price with mileage

| Parameter | Coefficient | SE | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | 24.9296 | 0.5675 | 43.9299 | 0.0000 |
| locationPENNSYLVANIA | 1.7681 | 0.6771 | 2.6115 | 0.0093 |
| mileage | -0.1055 | 0.0099 | -10.6906 | 0.0000 |

Table 3 shows a table for linear model after consider mileage, there are two p-values we should focus on, first the p-value is $0$ for mileage, if we are testing $H_0 : \beta_2 = 0$, this p-value let us reject $H_0$,then we would say mileage is good predictor to predict car price. But our goal is test $\beta_1$, the p-value for location is $0.0093$ less than significant level 0.05, we would say reject our null hypothesis, we have strong significant evidence to show there is difference in price between the locations after count mileage in our model.

## Research Question 3

### Is there a difference in price between the locations after count mileage and age?

Now! we are not only consider mileage, but also consider age. We will add three variables in the model together, NEWYORK still represent as $0$, and PENNSYLVANIA represent as $1$,we use $x_2$ represent mileage of car, $x_3$ represent age.The linear model with three variables should be: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

$$\text{MODEL FOR PENNSYLVANIA:} y = (\beta_0 + \beta_1) + \beta_2 x_2 + \beta_3 x_3$$

$$\text{MODEL FOR NEWYORK: } y = \beta_0 + \beta_2 x_2 + \beta_3 x_3$$

$\beta_1$ is still our test goal, we can express these two hypotheses as the following models:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

### Fit model

Table 4: Linear Model Table for Car price with mileage and age

| Parameter | Coefficient | SE | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | 25.2109 | 0.5733 | 43.9782 | 0.0000 |
| locationPENNSYLVANIA | 1.9046 | 0.6747 | 2.8230 | 0.0049 |
| mileage | -0.0797 | 0.0136 | -5.8449 | 0.0000 |
| age | -0.4463 | 0.1635 | -2.7304 | 0.0065 |

Table 4 is showing every term is significant predictors, since each p-value is smaller than significant level $0.05$, but our goal for this research question just looking for p-value of $\beta_1$, the p-value of $\beta_1$ is $0.0049$, we can reject$H_0$,we could say we have strong evidence to show there is difference in price between the locations after count mileage and age or there is difference in price between the locations when after consider mileage and age.

## Best model

So far we total have three models for our study, but we did not which one is best to predict price,in other word, what variables should we use to predict price in most effective way. We will use a statistical tool which is call "Mallows' Cp",Best Subsets is a statistical technique which compares all possible multiple regression models for a set of predictor variables. The output displays the best-fitting models containing one predictor, two predictors, and so on. The result is a number of possible regression models and their summary statistics. Mallows' Cp helps you choose between these multiple regression models.Usually, you should look for models where Mallows' Cp is small("Mallows Statistic (C-p) Definition — Isixsigma.com").

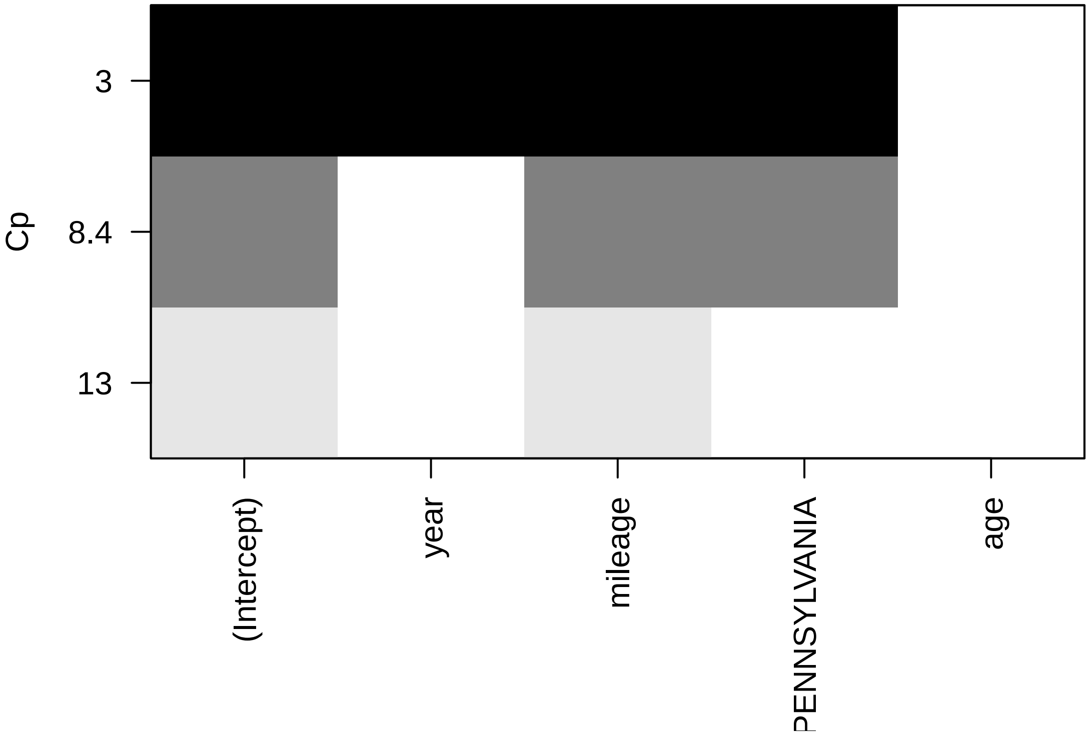

Figure 3: Cp plot

Figure 3 is showing the Cp value for each model, we can see the smallest Cp value is 3, that corresponds to model with three variables;year, mileage and location. Among them we can use age instead of year, they are exactly same thing. So the best model to predict car price is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 . x_1$ is mileage, $x_2$ is location and $x_3$ is age.

## Summary

We can make a conclusion now, if we only consider location variable, there is no difference in price between New York and Pennsylvania. There is difference in price between New York and Pennsylvania when we consider mileage or mileage and age.If you would wondering how difference in price, you may use our best model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ to predict car price between two place.

# Reference

“Autotrader Dataset Generator — Myslu.stlawu.edu.” http://myslu.stlawu.edu/~clee/dataset/autotrader/.

“Mallows Statistic (C-p) Definition — Isixsigma.com.” https://www.isixsigma.com/dictionary/mallows-statistic-c-p/.

## Code Appendix

```r
# Set some options
knitr::opts_chunk$set(
  echo = FALSE,
  warning = FALSE,
  message=FALSE,
  fig.align = "center"
  )

# Load Packages
library(mosaic)
library(ggformula)
library(tidyverse)
library(car)
library(tinytex)
library(Stat2Data)
library(HH)
library(leaps)
library(olsrr)
library(kableExtra)
library(data.table)
library(parameters)

#wrangling data
NY_raw_data <- read_csv("~/Documents/GitHub/FP_Aiden_Brody_Steve/Nissan 11355.csv")
NY_raw_data$location <-"NEW_YORK"  # add "location" variable
PA_raw_data <- read_csv("~/Documents/GitHub/FP_Aiden_Brody_Steve/Nissan 16801.csv")
PA_raw_data$location <- "PENNSYLVANIA" # add "location" variable
car_data <- (rbind(NY_raw_data,PA_raw_data)%>%
             filter(mileage >0)%>%
             filter(price >0)
             )
car_data$age <- 2022-car_data$year # create new variable called age that tell us how ol
#Create a boxplot of price
ggplot(
  data = car_data,
  mapping = aes(x = location, y = price)
  )+
  geom_boxplot(fill = "orange") +
  labs(x = "LOCATIONS", y = "CAR PRICE (in one thousand)") +
  theme_bw()+
  theme(
    text=element_text(size=12)
  )

# Get summary statistics by locations of price
Stats <-psych::describeBy(
  x = car_data$price,
  group = car_data$location,
  na.rm = TRUE,
  range = TRUE,
  quant = c(0.25,0.75),
  IQR = TRUE,
  mat = TRUE,
  digits = 4
)

# Set two location's name and select we want elements
Stats <- Stats%>%
  tibble::remove_rownames()%>%
  tibble::column_to_rownames(
    var = "group1"
  )%>%
  dplyr::select(
    n,min,Q0.25, median, Q0.75, max, mean,sd
  )

# Create and improved table
Stats %>%
kable(
  digits = 2,
  col.names = c("count","Min","Q1","Median","Q3","Max","Mean","SD"),
format.args = list(big.mark = ","),
caption = "Summary Statistics for Price",
booktabs = TRUE,
align = "c") %>%
kable_classic(
font_size = 12,
latex_options = c("HOLD_position", "scale_down")
)

# Fit the linear model
model1 <-lm(price ~ location, data=car_data)%>%
  parameters()

# Create a table linear model statistics
model1 <- model1%>%
  dplyr::select(
    Parameter,Coefficient,SE,t,p
  )

model1 %>%
kable(
  digits = 4,
  col.names = c("Parameter","Coefficient","SE","t-value","p-value"),
format.args = list(big.mark = ","),
caption = "Linear Model Table for Car price",
booktabs = TRUE,
align = "c") %>%
kable_classic(
font_size = 12,
latex_options = c("HOLD_position", "scale_down")
)
```

```
# fit linear model for 2 variables
model2 <- lm(price~ location +mileage, data = car_data)

#Create a dot plot
ggplot(
  data=car_data,
  mapping = aes(x = mileage, y = price, colour = location)
) +
  geom_point(shape = "circle", size = 1.5) +
  scale_color_hue(direction = 1) +
  theme_minimal()+
  geom_abline(slope=model2$coefficients[3],
              intercept=model2$coefficients[1]+model2$coefficients[2],
              color="cyan4",
              size=1)+
  geom_abline(slope=model2$coefficients[3],
              intercept=model2$coefficients[1],
              color="red",
              size=1)


# Fit the linear model and collect statistics
model2 <- lm(price~ location + mileage, data = car_data)%>%
  parameters()

# Create a table linear model statistics
model2 <- model2%>%
  dplyr::select(
    Parameter,Coefficient,SE,t,p
  )

model2 %>%
kable(
  digits = 4,
  col.names = c("Parameter","Coefficient","SE","t-value","p-value"),
format.args = list(big.mark = ","),
caption = "Linear Model Table for Car price with mileage",
booktabs = TRUE,
align = "c") %>%
kable_classic(
font_size = 12,
latex_options = c("HOLD_position", "scale_down")
)
# Fit the linear model and collect statistics
model3 <- lm(price~ location + mileage+age, data = car_data)%>%
  parameters()

# Create a table linear model statistics
model3 <- model3%>%
  dplyr::select(
    Parameter,Coefficient,SE,t,p
  )
```

```r
model3 %>%
kable(
  digits = 4,
  col.names = c("Parameter","Coefficient","SE","t-value","p-value"),
format.args = list(big.mark = ","),
caption = "Linear Model Table for Car price with mileage and age",
booktabs = TRUE,
align = "c") %>%
kable_classic(
font_size = 12,
latex_options = c("HOLD_position", "scale_down")
)
# create Cp plot
all <- regsubsets(price ~.,data = car_data, nbest=1)
plot(all,scal="Cp")
```