Stat 306 Final report:
Factors affecting a Vehicles Price

Kelvin Zhou
Louis Luo
Ronald Liu
STAT 306 201
Kenny Chiu

May 28, 2025

Abstract

This report aims to evaluate factors that affect the price of vehicles, factors including engine capacity, amount of cylinder, horsepower, top speed, seats and country. This will proivde Insights regarding which variables are most influential to a vehicles pricing. This paper will implement a full linear regression model to evaluate its performance, and implement a subset selection model to see which variables are most influential and compare it with the full model.

1 Introduction

As university students, we need to eventually consider purchasing a vehicle for either commuting or work. The issue is however, students have trouble deciding what car they should purchase as most students are on a budget and would like to spend the least amount of money and get the best car available at that price. This motivation led to our research question:

How does explanatory variables such as engine capacity, number of cylinders, horse power, top speed, number of seats, and the country of the manufacturer influence the cost of the car.

1.1 Data Source

The Cars dataset used in this project was scraped from the YallaMotor. It contains approximately 6,750 rows and 9 columns, with each row representing an individual car listing. The dataset includes variables such as price, brand, engine capacity, horsepower, number of cylinders, top speed, number of seats, and other key specifications. Prices are listed in various Middle Eastern currencies (e.g., SAR, AED, KWD).

1.2 Data Cleaning

The Cars dataset originally contains the price in multiple separate denominations of money, which were all converted to CAD using the conversion rates of the date the report was created. A filter was used to remove all potential faulty data, engine capacity greater than 20, or horsepower greater than 1500 is most likely an error with the data rather than a real observation. We also removed all data points where numerical data such as price or engine capacity contains a categorical value (i.e. price is “following” or engine capacity is “electric”). The final post filtered dataset consisted of 3850 observation each with the columns stated below:

1.3 Data Description

Response Variable:

- Price - The price of the car (CAD)

Explanatory Variable:

- Engine Capacity - The car’s engine capacity (0 to 8)
- Cylinder - The car cylinder power (3 to 16)
- Horse Power - The car horsepower (0 to 1500)
- Top Speed - The car top speed (km/hr)
- Seats - The number of seats in the car (2 to 18)
- Brand - The brand of the car (73 different brand)
- Country Manufacturer - The country where the car’s brand’s manufacturer originated from. (7 different countries)

The covariate ‘Country of Manufacturer’ was not provided by the dataset, but was transformed from the brand of the car to lower the amount of categorical data in the analysis. A decision was made to convert each brand of the car and transform it into the country where the brand was manufactured, which significantly lowers the amount of categorical variables. First we sort every brand into every country, then get the amount of observations per country and combine all the countries who have less than 200 observations into the "Other" category. The final covariate Country contains categorical data China, USA, South Korea, Japan, Germany, UK, and Other.

2 Exploratory Analysis

2.0.1 Exploratory Analysis on the Response Variable

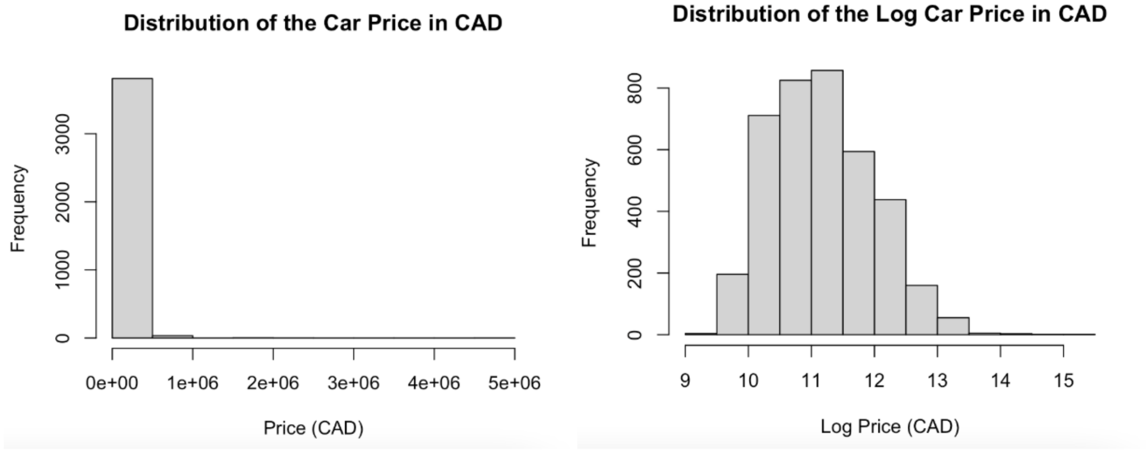

Figure 1: The histogram of car prices in CAD and the histogram of log car prices in CAD.

By observing the histogram of the price of cars, which is the first table shown in Figure 1. We can see that most of the prices are stacked near the lower end. This distribution intuitively makes sense because most cars are priced similarly near the lower end while there are a few brands who are known to make more expensive cars. However, to ensure a proper linear regression model can be created. We decided to take the log of the price to smooth out the distribution, which is the second plot shown in Figure 1. The histogram now is roughly bell shaped and appears to be normally distributed, which has a more stable variance. Therefore it would be better to fit the model based on the log price in CAD instead of just the price in CAD.

2.0.2 Exploratory Analysis on Explanatory Variables

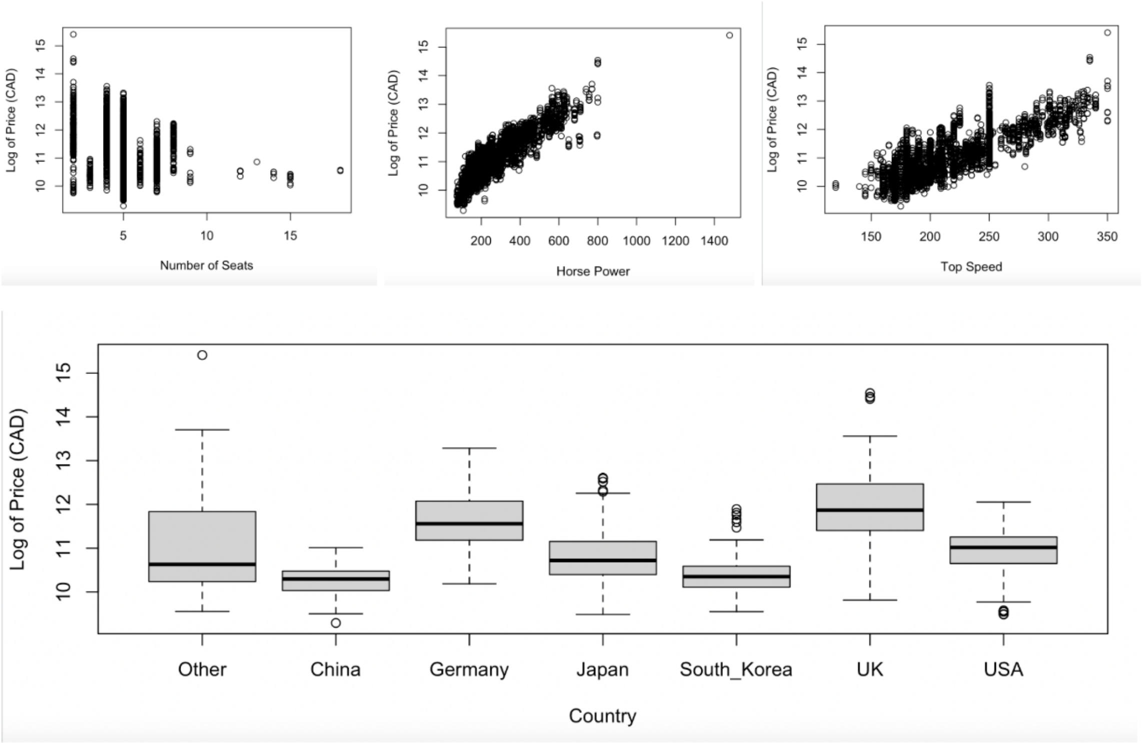

Figure 2: Some of the covariates we chose to plot against the response.

By observing the plots such as engine capacity against log price, horsepower against log price, and top speed against price. We can see a positive linear trend among 2 of the covariates with price. One being horsepower and the other being top speed. We observe a negative trend with the plot number of seats against log price, meaning more car seats might potentially result in a cheaper car. The boxplot of all the manufacturer countries showed relatively similar medians in log price amongst all of them except Germany and UK, which seems to have higher median compared to the rest. The observation from the boxplot could indicate that cars manufactured by German and UK companies tend to be more expensive.

2.1 Full Model Linear Regression

We fit the model using the log price against the additive model of every covariate described in the data set description.

2.1.1 Full Model Interpretation

The R^2 of the full model is 0.8869, meaning that approximately 88.7% of the variability in log car prices is explained by the model's covariates.

The interpretation of the each covariate of the full fitted model:

- **Intercept:** If a car has 0 cylinders, 0 horsepower, 0 km/h top speed, 0 seats, and is from another country that isn't the ones in the categorical variable country, it would be estimated to cost 7828.90.
- **Engine Capacity:** For every 1 unit increase in a car's engine capacity, the price on average is estimated to increase by 2.36%.

- **Number of Cylinders:** For every 1 unit increase cylinder in a car's engine, the price on average is estimated to increase by 4.09%.

- **Horsepower:** For every 1 unit increase in a car's horsepower, the price on average is estimated to increase by 0.30%.

- **Top Speed:** For every 1 km/h increase in a car's top speed, the price on average is estimated to increase by 0.36%.

- **Number of Seats:** For every 1 unit increase in a car's number of seats, the price on average is estimated to increase by 3.42%.

- **China:** If the car was produced by a manufacturer based in China, the price on average is estimated to decrease by 24.02% compared to the countries that aren't the UK, Germany, Japan, the USA or South Korea.

- **Germany:** If the car was produced by a manufacturer based in Germany, the price on average is estimated to increase by 34.57% compared to the countries that aren't the UK, the USA, Japan, China or South Korea.

- **Japan:** If the car was produced by a manufacturer based in Japan, the price on average is estimated to increase by 0.84% compared to the countries that aren't the UK, Germany, the USA, China or South Korea.

- **South Korea:** If the car was produced by a manufacturer based in South Korea, the price on average is estimated to decrease by 21.48% compared to the countries that aren't the UK, Germany, Japan, China or the USA.

- **UK:** If the car was produced by a manufacturer based in the UK, the price on average is estimated to increase by 36.50% compared to the countries that aren't the USA, Germany, Japan, China or South Korea.

- **USA:** If the car was produced by a manufacturer based in the USA, the price on average is estimated to decrease by 17.67% compared to the countries that aren't the UK, Germany, Japan, China or South Korea.

2.1.2 Full Model Goodness of Fit

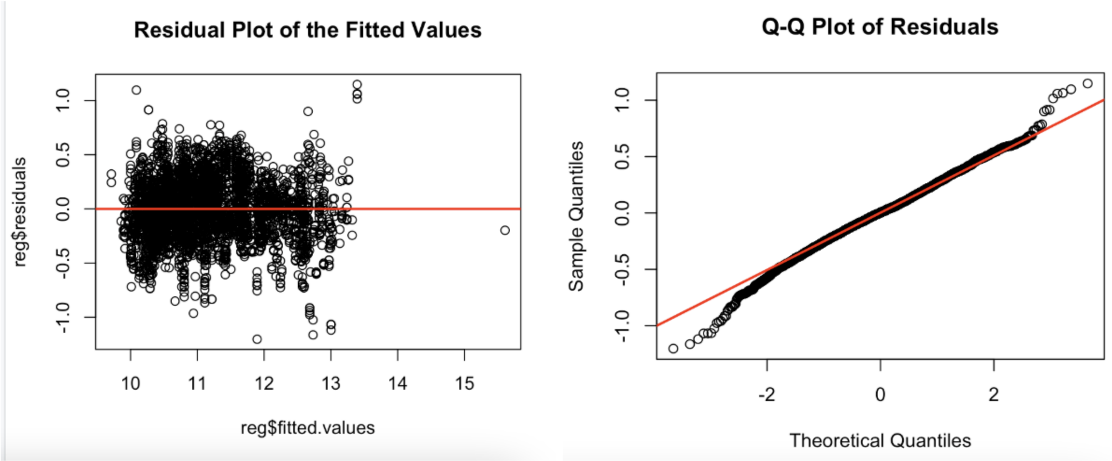

Figure 3: The residual plot and QQ plot with the fitted values

We assume that car listings are independent, since each row represents a different seller and vehicle. However, there may be weak dependence due to reposted listings, sellers benchmarking off one another, or shared pricing patterns among dealers. These effects are likely small relative to the variation explained by our model, but should be acknowledged when interpreting inference results.

The residuals of the fitted values, we can see that the data is quite randomly scattered. The variance is also mostly the same with no residuals branching outwards or shrinking towards the center. Therefore homoscedasticity is most likely not violated.

To assess the normality assumption, a QQ-plot of residuals was used. From Figure 3 we can see that the residuals closely follow the reference line, with minor deviations in the tails.These deviations are small and unlikely to significantly affect the normality of error assumption.

2.2 Multicollinearity analysis

To ensure the covariates aren't linearly related. We use the variance inflation factor with a rule of thumb of 10 to decide whether or not there may be multicollinearity in the explanatory variables.

Covariates	**Engine Capacity**	**# of Cylinders**	**Horsepower**	**Top Speed**	**# of Seats**	**Country**
VIF	8.87	9.08	8.63	4.37	1.37	2.44

Table 1: VIF Table for each covariate

Looking at table 1, we see that none of the covariates have VIF greater than 10, which means multicollinearity within the full model is most likely not violated. Therefore we can reasonably assume that the parameter estimates in the full model are stable and reliable, and that multicollinearity is not inflating the variance of the coefficient estimates.

2.3 Model Selection

Using best subset selection, which was chosen because our full model doesn't contain many covariates and it won't be computationally expensive to calculate every possible subset.

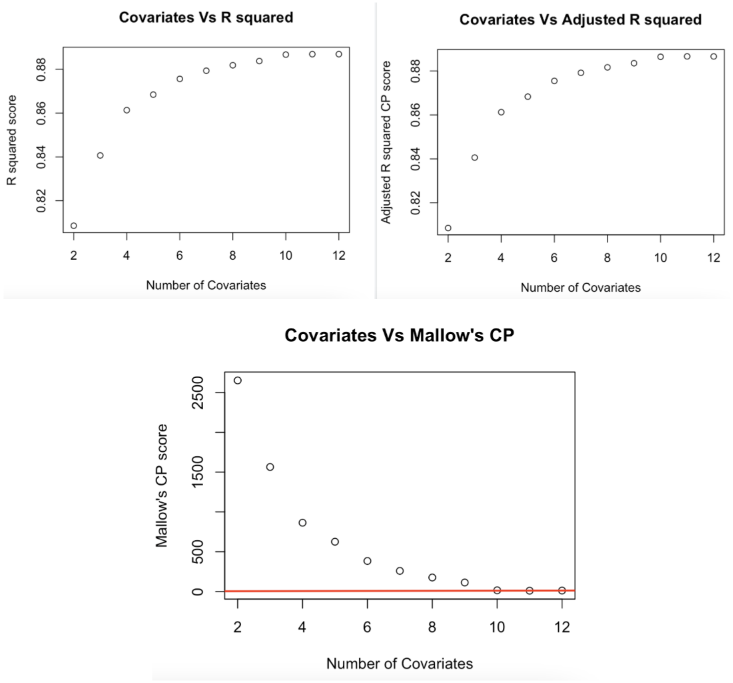

Figure 4: The R squared, Adjust R squared, and the Mallow's Cp plotted against the number of covariates selected by best subset selection.

Looking at figure 4, the R squared plot with every best covariate subset, we can see that at with 6 covariates, the R squared curves shows a diminishing return after including around 6 covariates. This suggests that models with 6 to 8 predictors capture most of the explainable variance without excessive complexity.

Similar to the R squared value, the adjusted R squared value also stops drastically increasing at 6 covariates. Therefore based on the adjusted R squared, the best model is the model with 6 covariates.

The Mallow's Cp suggests that model 8 is the best model with it having one of the closest score compared to the amount of covariates plus the intercept. The models with more covariates do yield a better Mallow's Cp score, but we decided values are due to the result of being very close to the full model and it will still most likely overfit.

Considering all three metrics—$\mathbf{R^2}$, adjusted $\mathbf{R^2}$, and Mallow's Cp—along with the principle of parsimony, we selected the model with 6 covariates as the best balance between simplicity and performance.

Model	**R-squared**	**Adjusted R-Squared**	**Mallow's Cp**
Full Model	0.8870	0.8866	12.00
Best Selection Model	0.8794	0.8792	258.94

Table 2: Model Comparison Table

2.4 Interpretation of the Best Model

Assuming all other variables are held constant, the following coefficients are significant at a 1% significance level. The interpretation of the each covariate of the best fitted model:

- **Intercept:** A car would be expected on average to cost 7882.45 CAD if all of the predictors are 0.
- **Horsepower:** For every 1 unit increase in a car's horsepower, the price on average is estimated to increase by 0.30%, assuming all other covariates are constant.
- **Top Speed:** For every 1 unit increase in a car's top speed in km/h, the price on average is estimated to increase by 0.36% assuming all other covariates are constant.

- **Number of Cylinders:** For every 1 extra cylinder in a car, the price on average is estimated to increase by 5.54% assuming all other covariates are constant.
- **Germany:** If the car was produced by a manufacturer based in Germany, the price on average is estimated to increase by 50.76% compared to the countries that aren't the UK or Japan assuming all other covariates are constant.
- **Japan:** If the car was produced by a manufacturer based in the Japan, the price on average is estimated to increase by 18.82% compared to the countries that aren't Germany or the UK assuming all other covariates are constant.
- **UK:** if the car was produced by a manufacturer based in the UK, the price on average is estimated to increase by 36.50% compared to the countries that aren't Germany or Japan assuming all other covariates are constant.

3 Conclusion

3.1 Results and Findings

Our analysis aimed to explore how various car specifications and manufacturer origin influence vehicle price. After cleaning and transforming the dataset, we used multiple linear regression on the log-transformed price to stabilize variance and better satisfy model assumptions.

Our analysis shows that car specifications such as engine capacity, number of cylinders, horsepower, top speed, and number of seats are **positively associated** with car price. That is, as these values increase, the log-transformed price of the car tends to increase as well.

In terms of manufacturer origin, cars from **Germany and the UK** are significantly more expensive on average, while those from **China, South Korea, and the USA** tend to be cheaper, even after accounting for other features. This suggests that **brand origin plays a meaningful role** in pricing, potentially due to brand reputation or import market positioning.

Overall, both technical specifications and country of origin have statistically significant and interpretable effects on car price, helping explain around **88% of the variation** in log price using a multiple linear regression model.

3.2 limitations

- Residual behavior: While the updated model shows improved residual diagnostics, there is still mild deviation from normality in the upper tail of the Q-Q plot. This could affect inference for extreme cases but is unlikely to impact overall model validity.
- Data quality: The dataset was scraped from an online marketplace and may contain unknown inconsistencies or input errors despite our efforts to filter and clean it.
- Generalizability: As students based in Canada, we are using a dataset of car listings primarily from the Middle East. Therefore, model conclusions may not directly extrapolate to the Canadian car market due to differences in brand availability, pricing structure, and consumer preferences.
- Model assumptions: The model assumes linear relationships between predictors and log-transformed price. Although this assumption generally holds, any unmodeled nonlinear relationships could reduce predictive accuracy.
- Data transformation: A limitation of transforming the brand column to countries is that, the brand of a car heavily affects a cars price. For example a lamboroghini and a ferrari would be much more expensive than a mercedes benz vehicle, however due to the amount of categories we would have to consider we simply reduce it to just country of origin.

4 References

Ahmed, Mahmoud. "YallaMotors Cars Dataset." Kaggle, https://www.kaggle.com/datasets/mahmoudahmed6/yallamotors-cars-dataset. Accessed 15 Apr. 2025.

5 Appendix

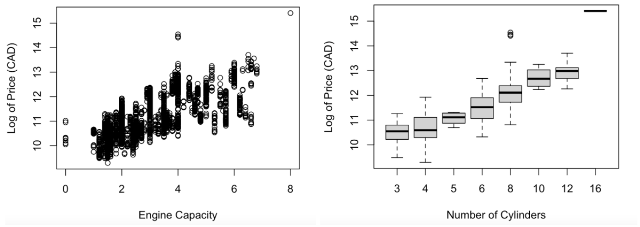

Figure 5: Some of the covariates we chose to plot against the response.

Kelvin Zhou submitted the Final group report
Louis Luo Submitted the R files and Dataset to canvas