ANALYSIS OF THE CAR PRICE DATA

Indian Statistical Institute, Kolkata

Fall, 2023

Swapnaneel Bhattacharyya
Srijan Chattopadhyay
Sevantee Basu
Soumava Mondal

Introduction

Introduction

- Multiple linear regression is a statistical technique that aims to establish a linear dependence between one response variable and a set of predictor variables.
- For the simple interpretation and less computational cost, it is very much preferred as a statistical technique in modern decision sciences such as machine learning, economics, and econometrics...
- In this project we analyze the car price dataset by fitting a multiple linear regression model.
- In the context of this data analysis task, we also describe a couple of new methods.

The Goal of this Project

- The car price dataset contains values of 205 cars along with the values of 23 different factors associated with the quality of the cars.
- We want to know
  - Which variables are significant in predicting the price of a car
  - How well those variables describe the price of a car
- To answer these questions we have to fit a suitable model so that the car companies can accordingly manipulate the design of the cars, the business strategy, etc. to meet certain price levels.
- Further, the model will be a good way for management to understand the pricing dynamics of a new market.

A brief description of our work

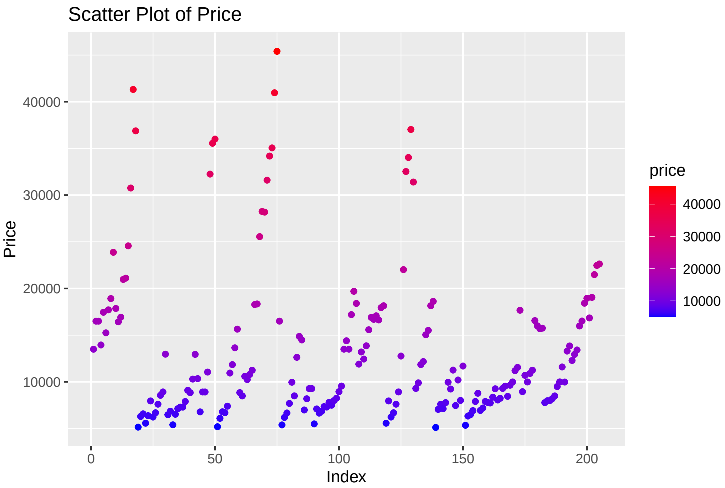
- We fit a multiple linear regression model into the car price dataset.
- We discuss a new variable selection criteria for regression.
- Besides the whole model, we also fit 3 different submodels using a reduced number of predictors.
- We discuss a new method for comparing the predictive power of different models.
- We compare the predictive powers of each model.
- We interpret the interesting results as well.

A look into the data

- The data is taken from Kaggle.
- It has 205 rows and 23 predictors.
- The response variable is the price of cars with different features.
- Some of the covariates are
  - Car fuel type i.e gas or diesel (Categorical)
  - Number of doors in a car (Categorical)
  - Wheelbase of the car (Numeric)
  - Width of the car (Numeric)
  - Size of car (Numeric)
  - Mileage in the city (Numeric)
  - Mileage on highway (Numeric)

Visualising the Data

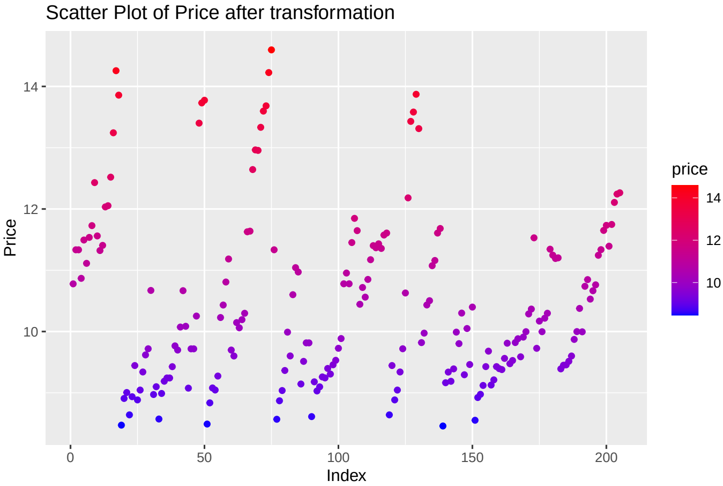
Scatter Plot of the Response variable (Car Price)



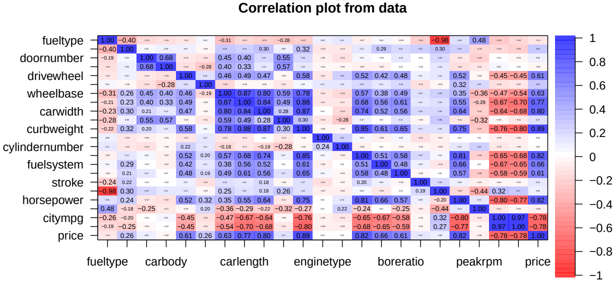
Observations and necessary transformations

- The car prices have a pretty high variance which can be seen from the scatter plot.
- So in order to stabilize the variance, we make a transformation of $g(y) = y^{1/4}$ into the response variable.
- We will fit all the models into the changed value of the response variable keeping the values of the remaining predictor variables the same.

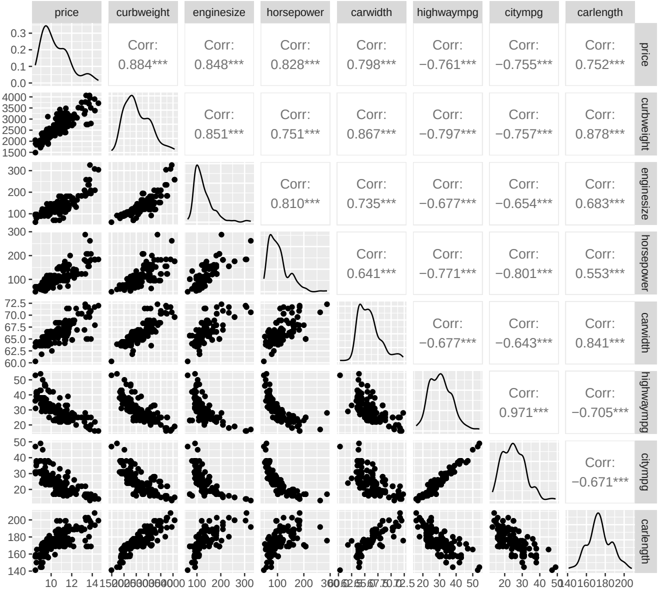
Stabilizing the variance



Correlation Plot



Pair plot of a subset

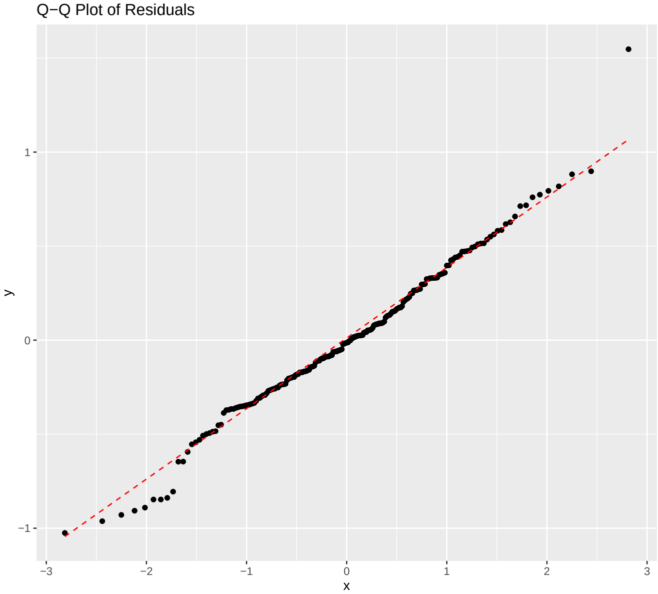


Main Results

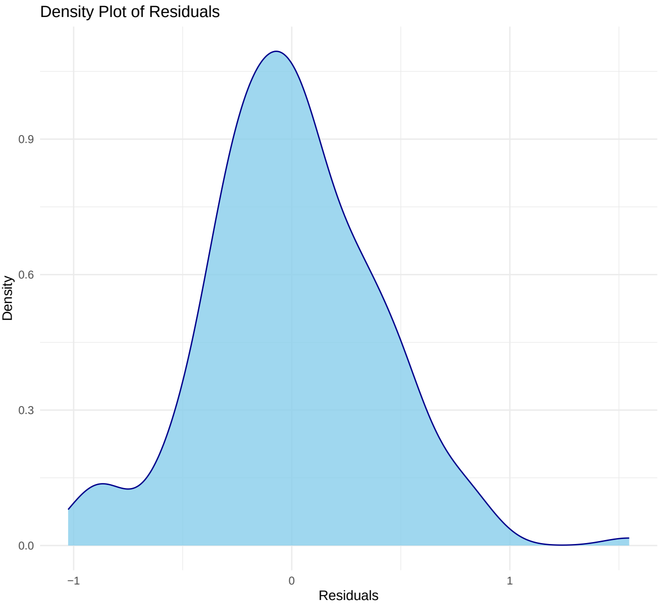
Fitting the full model

- We fit a Gaussian multiple linear regression model in the car price data taking the variable car price as the response variable and the other as predictor variable.
- We first check the standard requirements of the Gaussian model i.e.
    - Assumption of Normality
    - Assumption of homoscedasticity
    - Assumption of independence
    - Checking for Multicollinearity

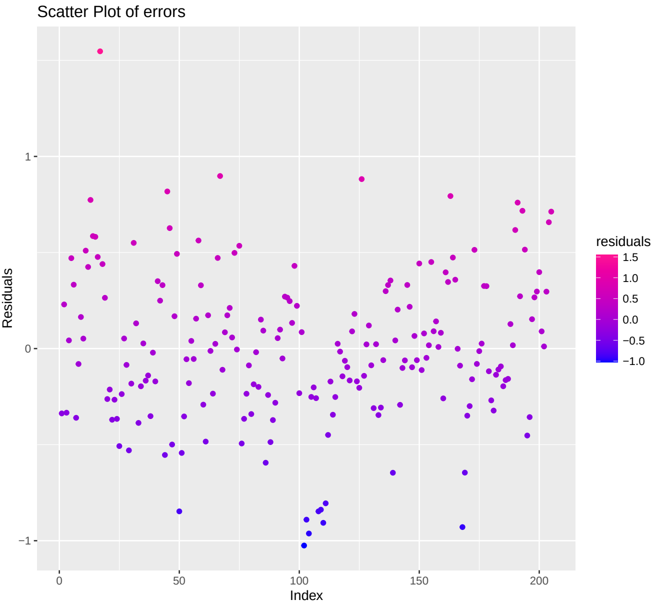
QQplot of the residuals



Density of the residuals



Scatter Plot of the residuals



Observations

- From the QQ plot and density plot, we see assumptions of the normality of the residuals held in our dataset.
- From the scatter plot we see the assumption of homoscedasticity of the residuals is valid in our context since there is no pattern in the scatterplot.
- From the pair plot, we see the presence of multicollinearity in the predictor variables.

Output of the Full Model

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.8004645	3.4080801	-1.995	0.04751*
Symboling	0.0024547	0.0395091	0.062	0.95053
CarName	-0.0057896	0.0008888	-6.514	7.10e-10***
Fueltype	0.9952707	1.0082270	0.987	0.32489
Aspiration	0.1214066	0.1384471	0.877	0.38170
Doornumber	0.1741367	0.0990351	1.758	0.08039 .
Carbody	-0.1546054	0.0568883	-2.718	0.00722**
Drivewheel	0.1771178	0.0843200	2.101	0.03707*
Enginelocati	1.5430098	0.3214051	4.801	3.31e-06***
Wheelbase	0.0071442	0.0160294	0.446	0.65635
Carlength	0.0040693	0.0081552	0.499	0.61840
Carwidth	0.0893677	0.0407973	2.191	0.02977*
Carheight	0.0462252	0.0211097	2.190	0.02983*
Curbweight	0.0007119	0.0002453	2.903	0.00416**
Enginetype	0.0152322	0.0352291	0.432	0.66599

Output of the Full Model

Coefficients	Estimate	SE	t value	Pr(>|t|)
Cylindernumber	0.0360657	0.0551827	0.654	0.51422
Enginesize	0.0061144	0.0026322	2.323	0.02130*
Fuelsystem	0.0667239	0.0234690	2.843	0.00498**
Boreratio	-0.0299916	0.1706684	-0.176	0.86070
Stroke	-0.3026401	0.1174077	-2.578	0.01075*
Compressionrat	0.1022376	0.0722984	1.414	0.15906
Horsepower	0.0058690	0.0028991	2.024	0.04441*
Peakrpm	0.0001408	0.0001016	1.386	0.16751
Citympg	-0.0373715	0.0256182	-1.459	0.14637
Highwaympg	0.0257845	0.0225882	1.142	0.25518

Residual standard error: 0.4176 on 180 degrees of freedom
Multiple R-squared: 0.9185, Adjusted R-squared: 0.9077
F-statistic: 84.55 on 24 and 180 DF, p-value: $<$ 2.2e-16

Inference from the Full Model

- From the model summary, it turns out **Car Name, Carbody, Location of car engine, Curb weight, Fuel system** are the most significant predictors in the model.
- The adjusted R^2 is 0.9, so the model can account for a significant portion of the total variance in the data. So the independent predictor variables provide valuable information for predicting the car price.
- The p-value for the F-test of the null model vs full model is very low, which signifies some of the predictors are pretty much significant.

Importance of a submodel

- From the summary of the full model, it can be seen that most of the predictors are not very significant.
- Also multicollinearity is present in the predictor variables.
- But the adjusted R^2 is 0.9077. So, maybe a subset of this model can explain the variability in the data very well.
- So it turns out that all the variables are not significant though there are some really good predictors.
- Therefore, we try to find a better submodel for the data.

Splitting of the dataset

- To fit the submodels we first split the dataset into two parts: The test set and the Train set.
- The size of the train set is 80 % of the original dataset which is 164.
- The test dataset has 41 rows.
- The training set will be used for fitting the submodels.
- The test dataset will be used to check the predictive power of the model i.e. given the features how accurately can our model predict the price of a car?

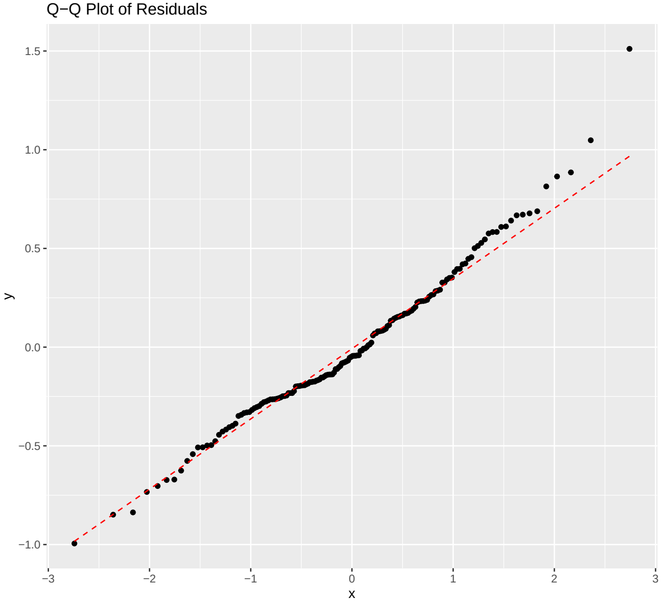
Fitting different submodels

- We fit different submodels using the stepwise selection method based on p-values, AIC, and Mallows' C_p.
- All of the above-mentioned methods are iterative. We propose a simpler ANOVA-based method to choose a significant submodel that has the same predictive power as the submodels obtained using standard methods.

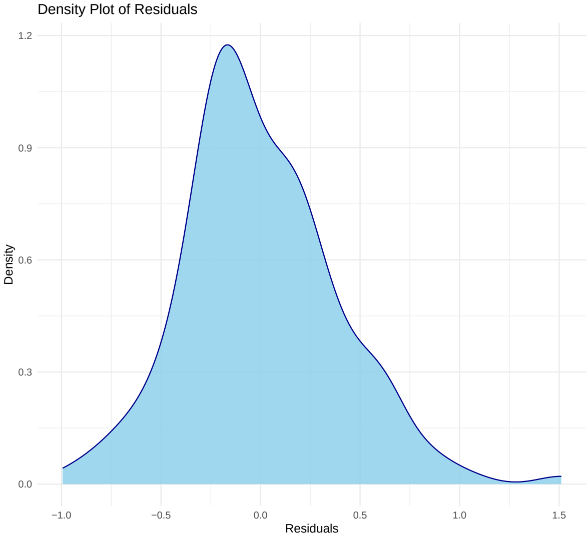
Submodel 1 : $\mathcal{M}_1$

- The first submodel is based on the significant predictors chosen using a stepwise selection method based on p-values.

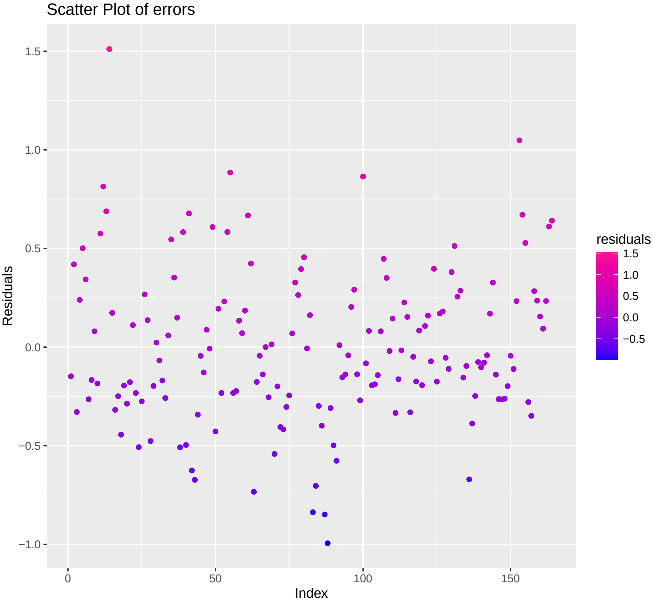
QQplot for the residuals of model $\mathcal{M}_1$



Density plot of the residuals of the model $\mathcal{M}_1$



Scatter plot of the residuals of model $\mathcal{M}_1$



Observations

- From the QQ plot and density plot, we see assumptions of the normality of the residuals held in model $\mathcal{M}_1$ as well.
- From the scatter plot we see the assumption of homoscedasticity of the residuals is valid in our context since there is no particular pattern.

Submodel 1 : $\mathcal{M}_1$

Coefficients	Estimate	Std. Error	t value	$\text{Pr}(>\vert t\vert)$
(Intercept)	-5.2348521	2.1220195	-2.467	0.01475*
Curbweight	0.0008615	0.0002074	4.153	5.47e-05***
Enginelocati	1.8095390	0.3156435	5.733	5.20e-08***
Horsepower	0.0063755	0.0023180	2.750	0.00668**
CarName	-0.0050875	0.0009259	-5.495	1.63e-07***
Carwidth	0.1064444	0.0322181	3.304	0.00119**
Drivewheel	0.1825197	0.0769326	2.372	0.01893*
Carheight	0.0465606	0.0198570	2.345	0.02034*
Fuelsystem	0.0740454	0.0245061	3.022	0.00296**
Enginesize	0.0056964	0.0020452	2.785	0.00603**
Compressionr	0.0301075	0.0118102	2.549	0.01179*
Stroke	-0.2551354	0.1175531	-2.170	0.03154*
Peakrpm	0.0001806	0.0001007	1.793	0.07497*

Submodel 1: $\mathcal{M}_1$

Residuals:

Min	1Q	Median	3Q	Max
-1.05499	-0.27201	-0.02124	0.23627	1.41947

Residual standard error: 0.4293 on 151 degrees of freedom Multiple R-squared: 0.9175, Adjusted R-squared: 0.9109 F-statistic: 139.9 on 12 and 151 DF, p-value: $<$ 2.2e-16

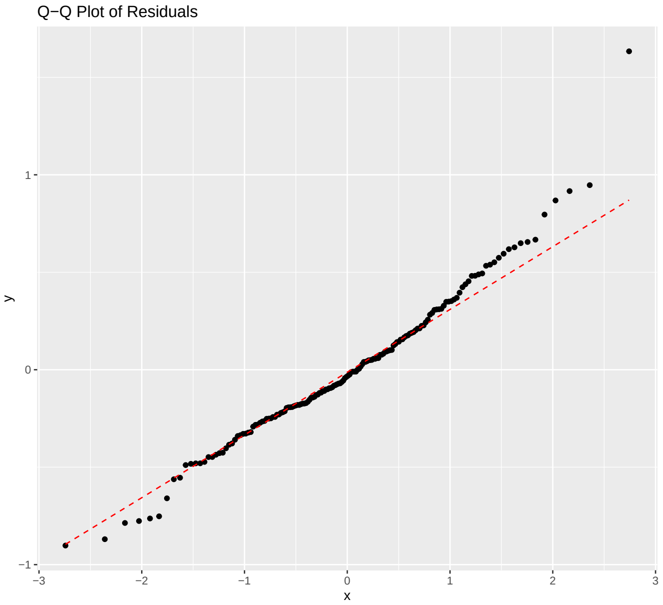
Inference and discussion on model $\mathcal{M}_1$

- In the reduced model $\mathcal{M}_1$, all the predictors are pretty much significant which can be inferred from the p-values of the table.
- There are a couple of predictors whose significance is more reflected in the reduced model. That's because those predictors collectively explain the variability of the response variable in a better way than in the full model.
- This also justifies the issue of overfitting in the whole model.
- A possible reason for this can be that the multicollinearity might be confounding the effects of other variables in the full model. By removing the correlated variables, we unveil the true, significant relationship of the variable that was previously masked by the multicollinearity.

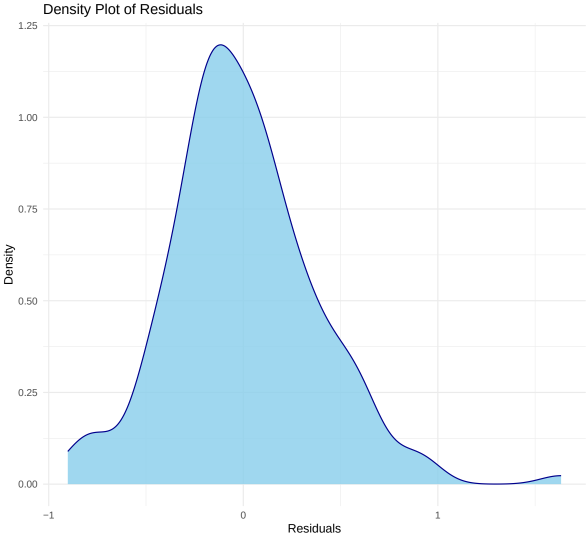
Submodel 2 : $\mathcal{M}_2$

- The second submodel is based on the significant predictors chosen using a stepwise selection method based on AIC.

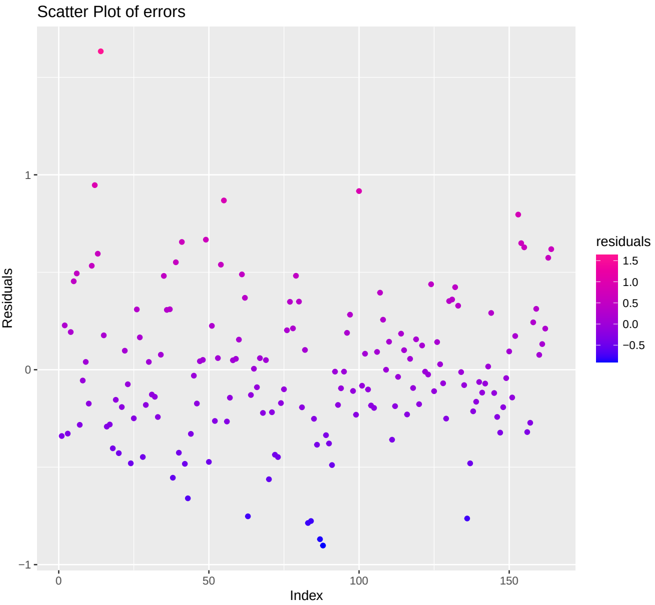
QQplot for the residuals of model $\mathcal{M}_2$



Density plot of the residuals of the model $\mathcal{M}_2$



Scatter plot of the residuals of model $\mathcal{M}_2$



Observations

- From the QQ plot and density plot, we see assumptions of the normality of the residuals held in model $\mathcal{M}_2$ as well.
- From the scatter plot we see the assumption of homoscedasticity of the residuals is valid in our context since there is no particular pattern.

Submodel 2: $\mathcal{M}_2$

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.2348521	2.1220195	-2.467	0.01475*
Curbweight	0.0008615	0.0002074	4.153	5.47e-05***
Enginelocati	1.8095390	0.3156435	5.733	5.20e-08***
Horsepower	0.0063755	0.0023180	2.750	0.00668**
CarName	-0.0050875	0.0009259	-5.495	1.63e-07***
Carwidth	0.1064444	0.0322181	3.304	0.00119**
Drivewheel	0.1825197	0.0769326	2.372	0.01893*
Carheight	0.0465606	0.0198570	2.345	0.02034*
Fuelsystem	0.0740454	0.0245061	3.022	0.00296**
Enginesize	0.0056964	0.0020452	2.785	0.00603**
Compressionr	0.0301075	0.0118102	2.549	0.01179*
Stroke	-0.2551354	0.1175531	-2.170	0.03154*
Peakrpm	0.0001806	0.0001007	1.793	0.07497*

Submodel 2: $\mathcal{M}_2$

Min	1Q	Median	Col4	Col5
-1.05499	-0.27201	-0.02124	0.23627	1.41947

Residual standard error: 0.4293 on 151 degrees of freedom Multiple R-squared: 0.9175, Adjusted R-squared: 0.9109 F-statistic: 139.9 on 12 and 151 DF, p-value: $<$ 2.2e-16

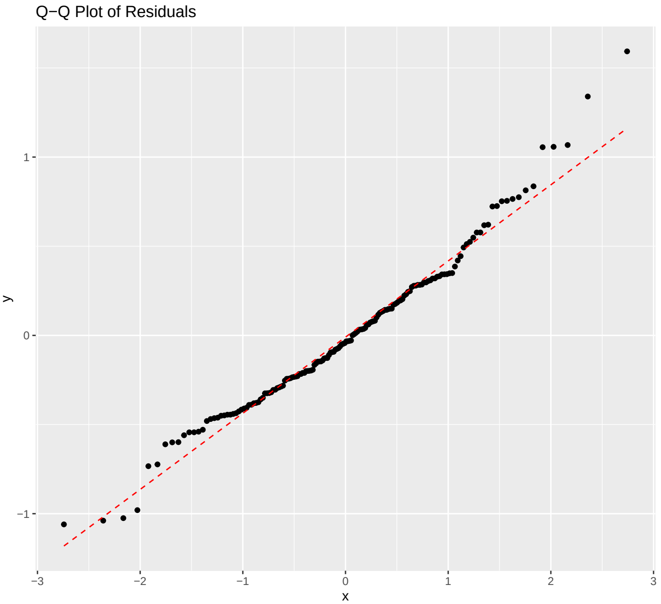
Inference and discussion on model $\mathcal{M}_2$

- The model is $\mathcal{M}_2$ is quite similar to model $\mathcal{M}_1$, therefore the key inferences are the same.
- Model $\mathcal{M}_2$ also resolves the potential issue of multicollinearity and overfitting in the whole model.

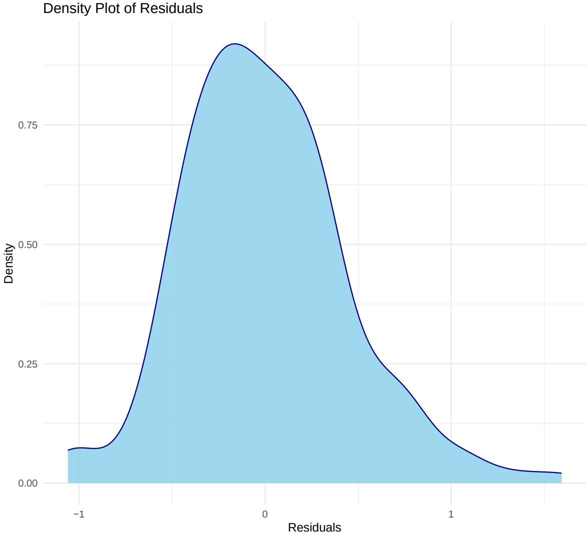
Submodel 3 : $\mathcal{M}_3$

- The third submodel is based on the significant predictors chosen using a stepwise selection method based on Mallows' C_p

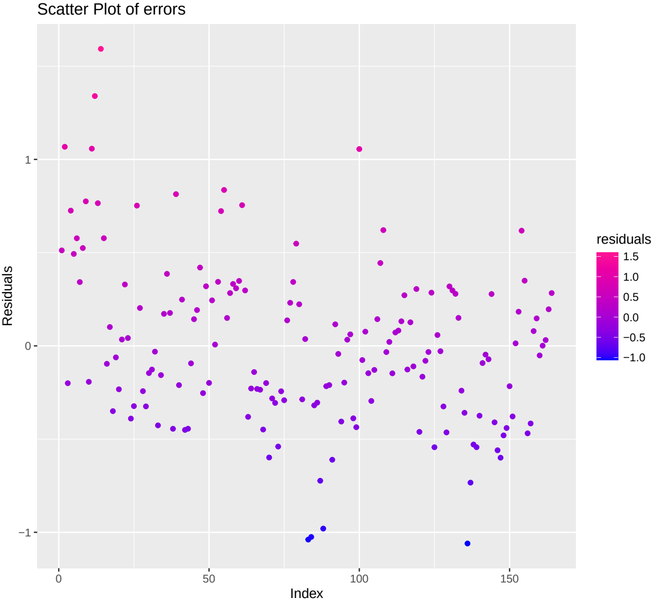
QQplot for the residuals of model $\mathcal{M}_3$



Density plot of the residuals of the model $\mathcal{M}_3$



Scatter plot of the residuals of model $\mathcal{M}_3$



Observations

- From the QQ plot and density plot of the residuals, we see there is a slight deviation from the normality of the residuals.
    - The residual distribution is slightly heavy-tailed.
    - From the density plot, it turns out the residuals are slightly positively skewed.
- From the scatter plot we see the assumption of homoscedasticity of the residuals is valid in our context since there is no particular pattern.

Submodel 3: $\mathcal{M}_3$

Coefficients	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	3.1385230	3.5167368	0.892	0.373609
Symboling	0.0105828	0.0428236	0.247	0.805156
Fueltype	-1.2358651	1.0870129	-1.137	0.257415
Aspiration	-0.2551765	0.1563517	-1.632	0.104806
Carbody	-0.1353495	0.0617434	-2.192	0.029943*
Drivewheel	0.0368744	0.0934597	0.395	0.693748
Enginelocati	1.4719933	0.4151975	3.545	0.000526***
Wheelbase	0.0348869	0.0141487	2.466	0.014824*
Carheight	0.0058838	0.0233360	0.252	0.801291
Curbweight	0.0013185	0.0002524	5.224	5.90e-07***
Enginetype	0.0667447	0.0386547	1.727	0.086324*
Fuelsystem	0.0120762	0.0267590	0.451	0.652443
Boreratio	-0.1369251	0.1935987	-0.707	0.480523
Compressionr	-0.0435542	0.0765058	-0.569	0.570026
Horsepower	0.0122247	0.0027647	4.422	1.89e-05***
Peakrpm	0.0002155	0.0001105	1.950	0.053054
Highwaympg	-0.0025499	0.0130873	-0.195	0.845790

Submodel 3: $\mathcal{M}_3$

Min	1Q	Median	Col4	Col5
-1.05993	-0.29786	-0.03831	0.27829	1.59284

Residual standard error: 0.4636 on 147 degrees of freedom Multiple R-squared: 0.8894, Adjusted R-squared: 0.8773 F-statistic: 73.85 on 16 and 147 DF, p-value: $<$ 2.2e-16

Inference and discussion on model $\mathcal{M}_3$

- In submodel $\mathcal{M}_3$, the most significant predictors are **Engine Location, Curb weight, Horse power**.
- So this model puts more importance on less number of predictors.
- Due to a significant reduction in model complexity, this particular model experiences a noticeable decrease in the R^2 compared to the two previous models.

What Next

- We now discuss **a new ANOVA-based variable selection criteria**.
- As we will see this method is
  - Simpler
  - (Will show later on) Gives a submodel that has the same predictive power as the previous submodels.

A New Variable Selection Criteria

- The key idea is, in the regression model consider only those variables that have a **significant** impact on the **change of the price level** of cars.
- Now to obtain such variables we do the following:
- At first we divide the price level into 5 categories : very high, high, moderate, low, very low
- Now considering these to be factor levels, we fit a one-factor ANOVA into the predictor values for each predictor.

Fitting an one-factor ANOVA

- We fit the one-factor ANOVA for a particular predictor X_K as follows:
- Suppose the values of the predictor X_k are $X_{1k}, \cdots, X_{nk}$, where n is the total no of observations which is 205 here.
- For $1 \leqslant i \leqslant 5$, let A_i be the set of indices that corresponds to the values of the response variable which are in ith group.
- For $1 \leqslant i \leqslant 5$, consider $\{X_{jk} : j \in A_i\}$ as the replicates of the ith factor level.
- So under this one way classification, we fit the one-factor ANOVA into $X_{1k}, \cdots, X_{nk}$.

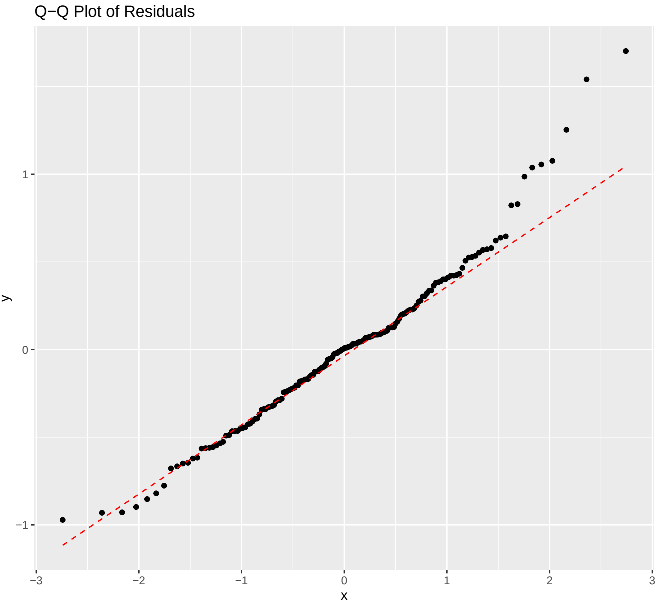
Our variable selection criteria

- Using the F test for testing equality of factor level means for one factor ANOVA, for each predictor we can decide whether that particular predictor has any impact on the change of price levels.
- We consider the predictor in our model if the hypothesis of equality of factor level means is rejected at 5 % level of significance.

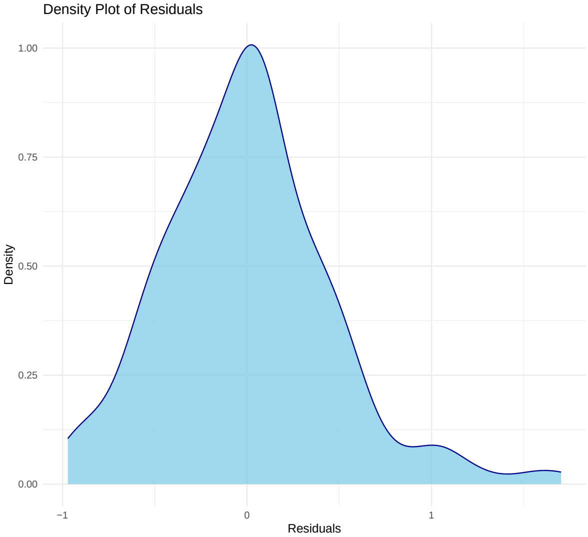
Submodel 4 : $\mathcal{M}_4$

- The predictors for the fourth submodel are obtained using our ANOVA-based variable selection criteria.
- The predictors which are removed are

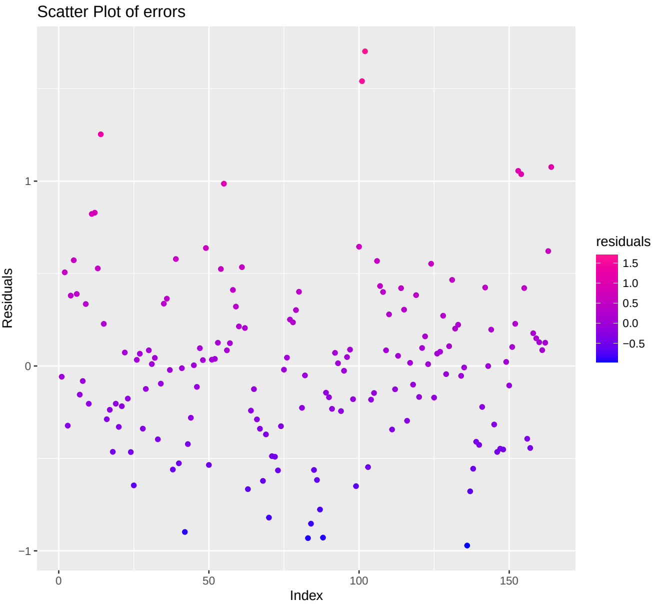
QQplot for the residuals of model $\mathcal{M}_4$



Density plot of the residuals of the model $\mathcal{M}_4$



Scatter plot of the residuals of model $\mathcal{M}_4$



Observations

- From the QQ plot and density plot, we see in model $\mathcal{M}_4$ as well there is a deviation from normality:
    - The distribution of the residuals is heavy-tailed. (In fact, it looks to be heavier-tailed than model $\mathcal{M}_3$.
- From the scatter plot we see the assumption of homoscedasticity of the residuals is valid in our context since there is no particular pattern.

Submodel 4: $\mathcal{M}_4$

Coefficients	Estimate	Std. Error	t value	Pr($> \vert t \vert$)
(Intercept)	-2.0406538	2.2871683	-0.892	0.373705
CarName	-0.0046395	0.0010582	-4.384	2.18e-05***
Aspiration	-0.0646579	0.1338930	-0.483	0.629866
Drivewheel	0.1792145	0.0855618	2.095	0.037891*
Wheelbase	-0.0023491	0.0158782	-0.148	0.882587
Carlength	0.0035871	0.0096257	0.373	0.709925
Carwidth	0.1012476	0.0410156	2.469	0.014690*
Curbweight	0.0011623	0.0002761	4.210	4.38e-05***
Enginesize	0.0013181	0.0028455	0.463	0.643867
Fuelsystem	0.0276593	0.0285049	0.970	0.333441
Boreratio	0.1572141	0.1961117	0.802	0.424020
Horsepower	0.0107937	0.0031159	3.464	0.000694***
Citympg	-0.0142145	0.0293158	-0.485	0.628473
Highwaympg	0.0287841	0.0265880	1.083	0.280725

Submodel 4: $\mathcal{M}_4$

Min	1Q	Median	3Q	Max
-0.97126	-0.30136	0.00695	0.23025	1.70223

Residual standard error: 0.4777 on 150 degrees of freedom
Multiple R-squared: 0.8801, Adjusted R-squared: 0.8697
F-statistic: 84.68 on 13 and 150 DF, p-value: $<$ 2.2e-16

Inference and discussion on model $\mathcal{M}_4$

- In model $\mathcal{M}_4$, the most significant predictors are **Curb weight, Horse power**.
- This model looks to be a little bit different than the previous ones in the sense it puts more importance on a smaller number of predictors.

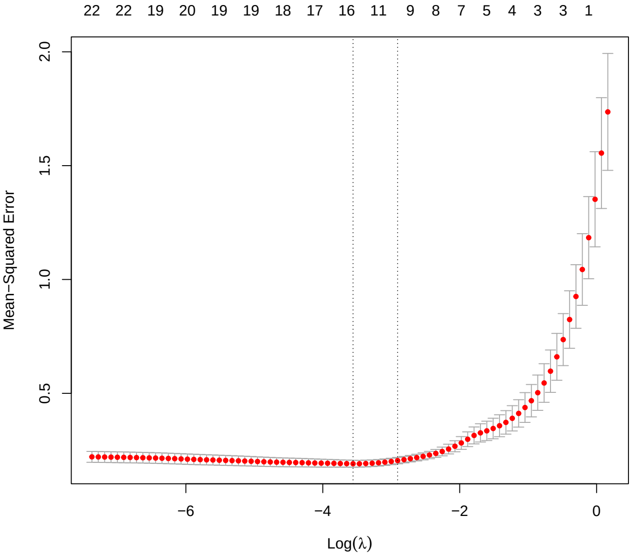
Introducing regularization - Importance in our context

- In the whole model, there is potential overfitting and multicollinearity.
- To resolve these issues, when stepwise selection criteria using MAllows' C_p and our ANOVA-based approach were applied, that made the model oversimplified which resulted in a noticeable decrease of the adjusted R^2.
- Hence to settle the overfitting and multicollinearity issue we use the idea of regularization via Lasso and Ridge regression.

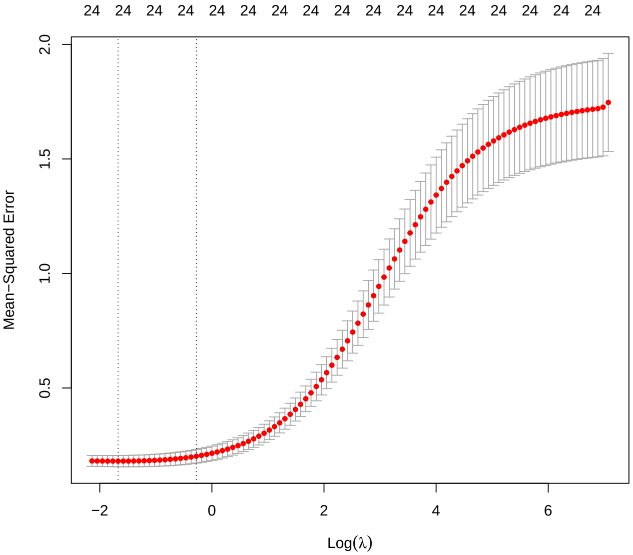
An attempt to perform Lasso and Ridge Regression to the data

- By λ, we denote the lasso and ridge penalty.
- We find the value of the penalty which minimizes the prediction error of the model.
- We plot the prediction error vs penalty.

Lasso Penalty



Ridge Penalty



Lasso and Ridge penalties are not very significant here

- We get the minimum prediction error occurs at $\lambda = 0.02850035$ (for Lasso) and at $\lambda = 0.1875133$ (For Ridge).
- Both of the penalties are very close to 0.
- So the penalties won't give any extra benefit to our prediction for both of the models.
- **So we will not be considering these models for our data.**

Comparison of different models

Comparison of different models - The methods

- Since we have different models for the same dataset, we can now compare the models to check which one fits better for the data.
- Comparisons between different models can be done based on
  - Information criteria like AIC, BIC, Mallows' C_p.
  - Predictive power using cross-validation score.
- Besides the above standard methods, we propose an ANOVA-based method to compare the predictive power of different models.

Comparison of models using AIC, BIC and adjusted R^2

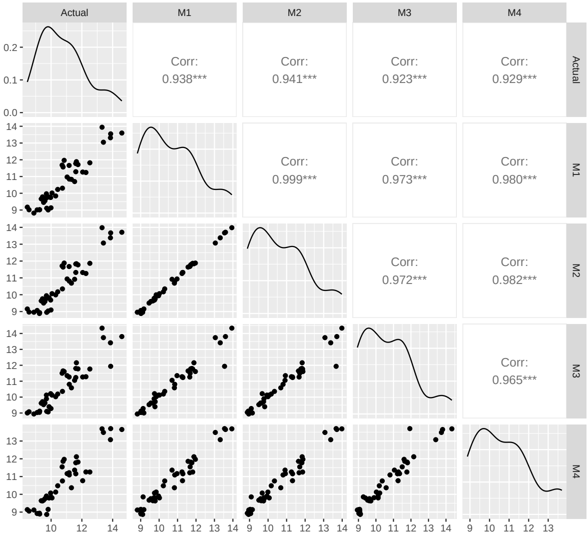
Model	Adj R^2	AIC	BIC
$\mathcal{M}_1$	0.9109	175.6027	212.8011
$\mathcal{M}_2$	0.9109	172.7497	216.1478
$\mathcal{M}_3$	0.8783	238.4781	284.9761
$\mathcal{M}_4$	0.8801	231.2857	287.0833

Table 11: Adj R^2, AIC and BIC of the models

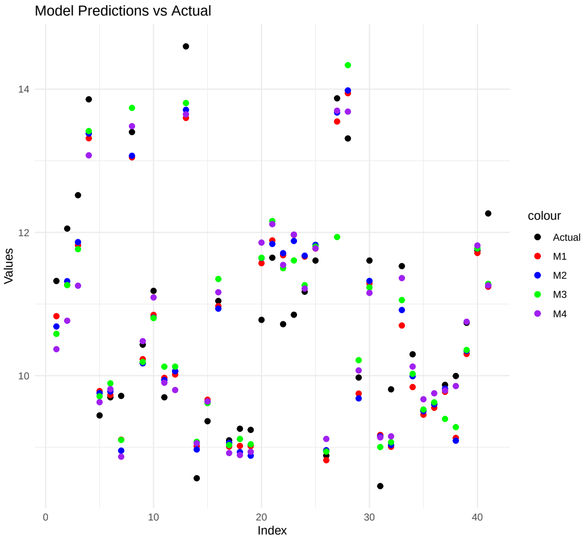
Comparison using predictive power

- Comparison between different models can also be done on the basis of their predictive power.
- One of the most natural ways to do this is to compare the empirical prediction error of the models.
- But we do it in a different way.
- First, for a visual justification, for each model, we plot the predicted values vs the actual values of the response variable from the test dataset and also find a correlation between them.

Plot of the predictions of the test data using all Submodels



Plot of the predictions of the test data using all Submodels



A new method for comparison of predictive power of different models

- We now discuss an ANOVA-based approach for the relative comparison of the predictive power of different models.
- Till now we have models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$.
- For each model, we find the value of the response variable (i.e. car price) for the test dataset.
- To be precise, let k be the size of the test dataset. ($k = 41$ here).
- Let $j_1, \cdots, j_k$ be the indices corresponding to test dataset.

ANOVA-based method of comparison of predictive powers

- For $i \in \{1, 2, 3, 4\}$, let $\{\hat{Y}_{j_r}^{\mathcal{M}_i}\}_{1 \leq r \leq k}$ be the set of predicted values for the test dataset using model $\mathcal{M}_i$.
- Now fit a one-factor ANOVA into the data

$$\begin{bmatrix} \hat{Y}_{j_1}^{\mathcal{M}_1} & \dots & \hat{Y}_{j_k}^{\mathcal{M}_1} \\ \vdots & \ddots & \vdots \\ \hat{Y}_{j_1}^{\mathcal{M}_4} & \dots & \hat{Y}_{j_k}^{\mathcal{M}_4} \end{bmatrix}_{4 \times k} \tag{1}$$

considering each row to correspond to a different factor level (so there are a total of 4 factors) and all the observations in a row represent the replicates of the response variable at that factor level.

ANOVA-based method of comparison of predictive powers

- Now using the one-factor ANOVA F-test for equality of factor level means, we can test whether there is any significant effect of the factor level (i.e. the model) in the response variable (i.e. the predicted car prices).
- If the above F-test is accepted at 5% significance level, we can conclude that $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$ have same predictive power.

ANOVA Table

Source of Variation	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Factor	3	0.0342	0.0114	0.005	0.999
Residuals	160	336.6	2.1040		

Table 12: ANOVA Table

Inference from the ANOVA table

- The $p-$value of the ANOVA F test of testing equality of factor level means (here equality of prediction power for four models) is very close to 1.
- So the null hypothesis is accepted with very strong evidence.
- So all of the 4 models have the same predictive power.

Conclusion and Remarks

- We notice that though model $\mathcal{M}_3, \mathcal{M}_4$ has less adjusted R^2 than model $\mathcal{M}_1, \mathcal{M}_2$, the prediction power of the four models are the same. So **decrease in R^2 does not necessarily imply a decrease of prediction power.**
- In practice, without using any prior knowledge of the predictor variables, since all of the models have similar predictive power, it is reasonable to choose the simplest model.
-

References

References

That's all!

Thank You :)