University of Gothenburg

Linear Statistical Models (MSG500-MVE190)

Sarah Alkhateeb & Nilufar Hatamova

Mini-analysis summary

TV dataset

In this minianalysis, we implemented data transformation; using the TV dataset, we examined how variables transformation may change the fit of the model. There are different types of data transformation such as natural logarithm, square root, and inverse.

The TV dataset contained 40 observations with 5 different variables such as life expectancy, people per TV, people per doctor, male and female life expectancy. Our outcome (response) is *people per TV* and the independent covariate is *People per Doctors*. We fitted the model by using raw data. The fitting and residuals plots are shown below:

Data Transformation

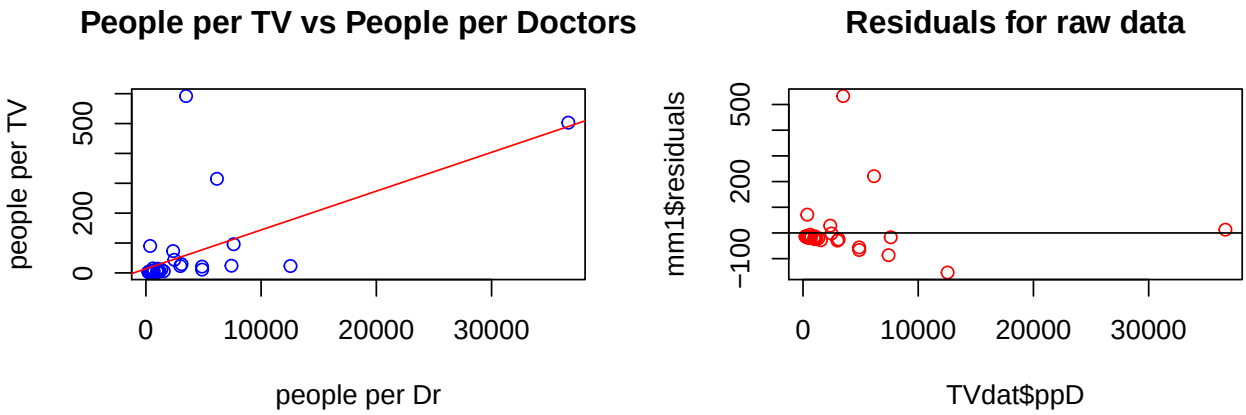

Figure 1: Fitting and residuals for raw data

From the fitting and residuals plots (*Figure 1*) we can see that:

- The model does not describe the data well (left graph).
- Residuals are not distributed symmetrically around 0 (right graph).
- We can see the outliers from both fitting plot and residuals graph.
- Based on the residuals plot we can say that there is non-constant variance.

To make the distribution more symmetric, we used some data transformations such as natural logarithm and inverse and checked basic assumptions to decide whether the least square is a sensible approach for modeling the data.

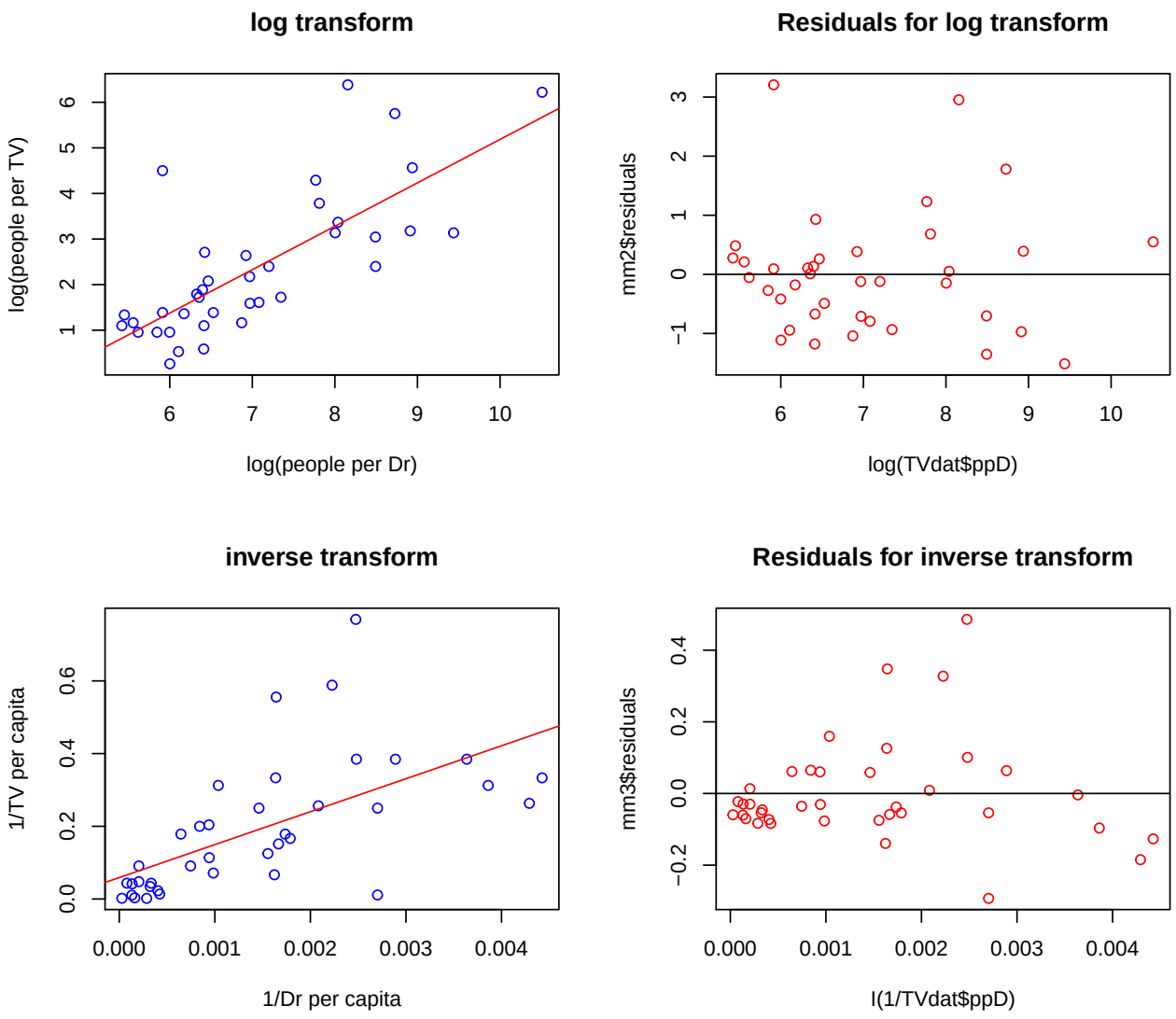

Figure 2: Fitting and residuals after data transformation

Figure 2 shows the fitting and residuals plots after using natural logarithm transformation and inverse transformation. After natural logarithm transformation, we can say that:

- The model still does not describe the data well.
- The residuals are distributed almost symmetrically around zero.
- We have almost constant variance.
- We clearly can see the outliers from both graphs.

Our next data transformation is the inverse transformation. After implementing this transformation:

- The model still not sufficient to describe the data.
- Residuals are not distributed symmetrically around zero.
- Increasing variance (non-constant) can be seen from the residuals graph and we still have outliers.

AirBNB DataSet

Exploring the data & fitting a model

We continued our analysis using a new dataset; AirBnB prices dataset, which contained almost 49,000 observations with 16 different variables such as neighborhood, latitude, longitude, room type and price.

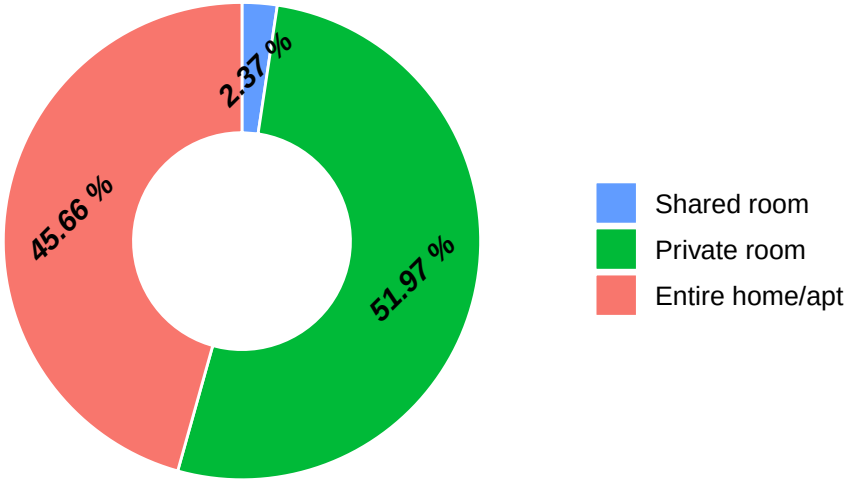

Figure 3: Room Type proportion

Figure 3 shows the room type proportion for the whole dataset. The pie chart tells us that *Private room* has the highest proportion (52%) of all types of accommodations followed by *Entire home/apt* (45.6%). *Shared room* has the smallest proportion which is 2.3% among all types of accommodations.

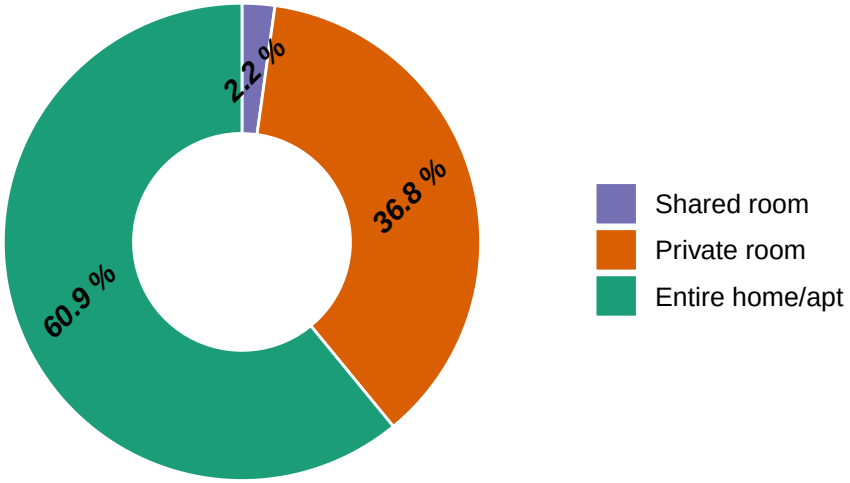

Figure 4: Room Type proportion in Manhattan

Choosing a subset from the data to analyze and see the effect of a specific covariate, we chose a subset including the *neighbourhood_group* Manhattan. *Figure 4* displays that in Manhattan, the Entire home apartment has the highest proportion (60.9 %) followed by private room (36.8%), and least preferred is shared room (2.2%). However, comparing to the whole dataset, Private room has the highest proportion followed by Entire home/apt.

Adding a new covariate to the dataset: The distance. The distance was calculated from the whole dataset from the expensive area (-74.007819, 40.718266) by using longitude and latitude of accommodations, and the radius of the earth. Fitting a simple linear regression model with a response log(price) and the distance covariate.

Table 1: Model summary for log(price)~distance

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.2192387	0.0403665	129.29624	<0.001	***
distance	-0.0000651	0.0000048	-13.52466	<0.001	***

Table 1 shows that as x increases by 1 unit, the value of the expected y decreases by 1% (because 1 - 0.99 = 0.01). In our case, 5000 units increase in distance, decreases the value of the expected price by about 22%.

So:

-How much does it cost to rent an entire apartment compared to a private room in Manhattan?

-How much does it cost to rent a full apartment in the Bronx compared to Manhattan or rent a private room in the Bronx?

-Does adding distance to the model sounds like a good idea?

Using the reference category and coefplot() to compare the significances of the several parameter estimates. First, comparing the rent cost between an entire apartment and a private room.

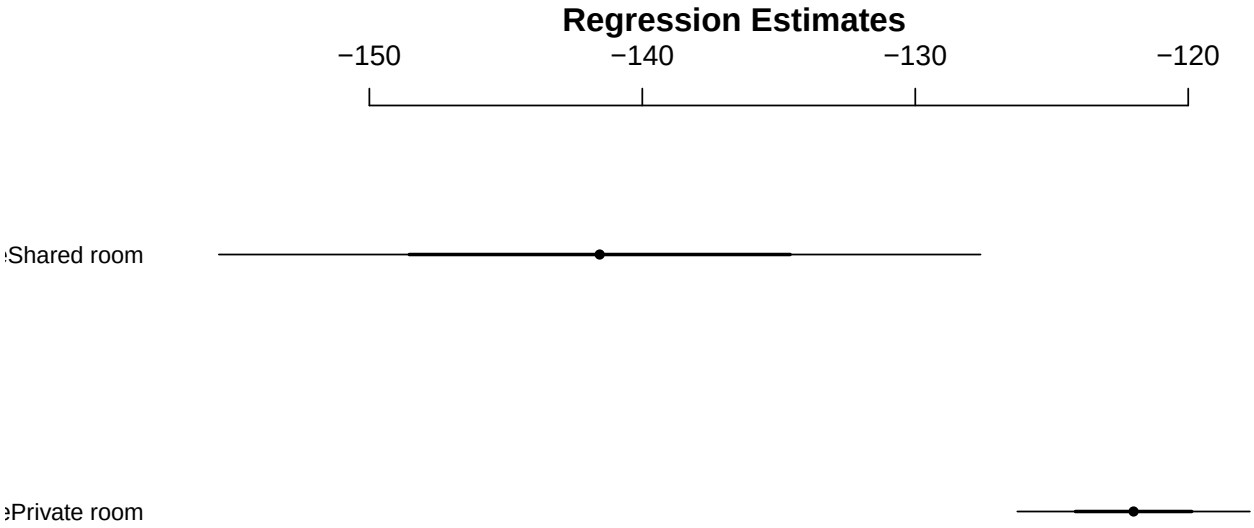

Figure 5: Regression estimates for room types

(*Figure 5*), we can clearly see that the shared room has the least price among them all, and their average price is smaller than the private room. We also can say that the cost for private and shared rooms are significantly lower than the cost for an entire home/apt where the entire home/apt is the reference point.

For every 1 unit increase in private room price, we expect a 122 unit decrease in the entire home/apt price holding all other variables constant.

Next, calculating the variability of a private room and a shared room.

The formula for variability is: $\beta_j \pm SE(\beta_j)$

For a shared room, the variability is (-148.64, -134.7). For a private room, the variability is (-124.14, -119.9)

Regression estimates

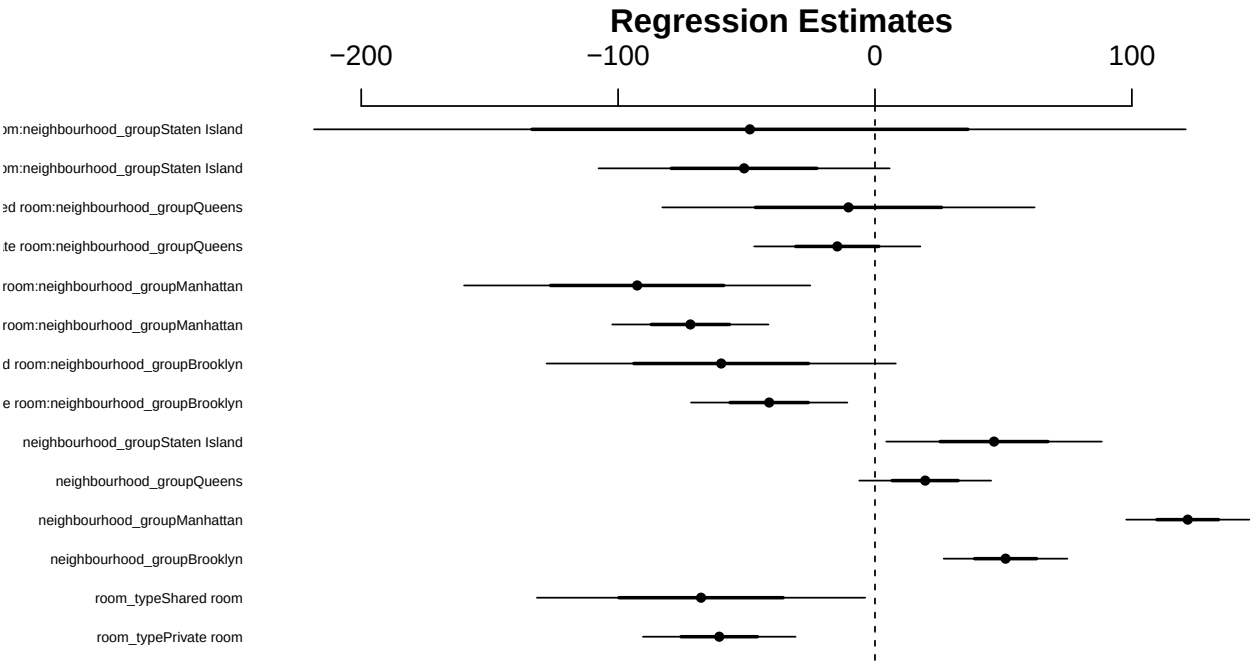

Figure 6: Regression estimates for room types and neighbourhood groups

Compared prices for different room types (private, entire apartment, shared room) for neighborhood groups (*Figure 6*). The reference point is the entire apartment in Bronx.

- Entire home/apt in Manhattan is more expensive than a private room in Manhattan.
- Entire home/apt in Manhattan is more expensive than the entire home/apt in Bronx.
- A private room in Manhattan is more expensive than a private room in Bronx.

The rent prices for the entire apartment in Manhattan, private room in Manhattan and private room in Bronx based on model coefficients. β_0= 127.51 is entire home/apt in Bronx (reference point)

β_1 is entire home/apt in Manhattan: $\beta_0 + \beta_1 = 249.24$

β_2 is private room in Manhattan: $\beta_0 + \beta_2 = 116.78$

β_3 is private room in Bronx: $\beta_0 + \beta_3 = 66.79$

Our next question is will adding distance to the model sounds like a good idea or not?

We selected *neighbourhood_group*, *room_type*, *price*, *distance* as a subset of the full dataset where the price is between 0.1 and 0.9 quantiles.

Table 2: Anova Test

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
38950	4453.905	NA	NA	NA	NA
38949	4154.639	1	299.2662	2805.567	0

First, we fitted the model by using two variables, *room_type* and *neighbourhood_group*. Then we added the *distance* to our model. We performed the partial F-test test to compare our models.

-Null Hypothesis: Full model and Reduced model do not significantly differ.

-Alternative Hypothesis: Full model is significantly better than the reduced model.

The result of the ANOVA test as in *Table 2*. Because the p-value- Pr(>F) is extremely small (< 2.2e-16) we reject the null hypothesis and we can say that the Full model (with distance) is better than the reduced model.

NOTE: In the table p value- Pr(>F) seems equal to 0 because it is too small.

The actual Pr(>F) is (< 2.2e-16)

Diagnostic plot

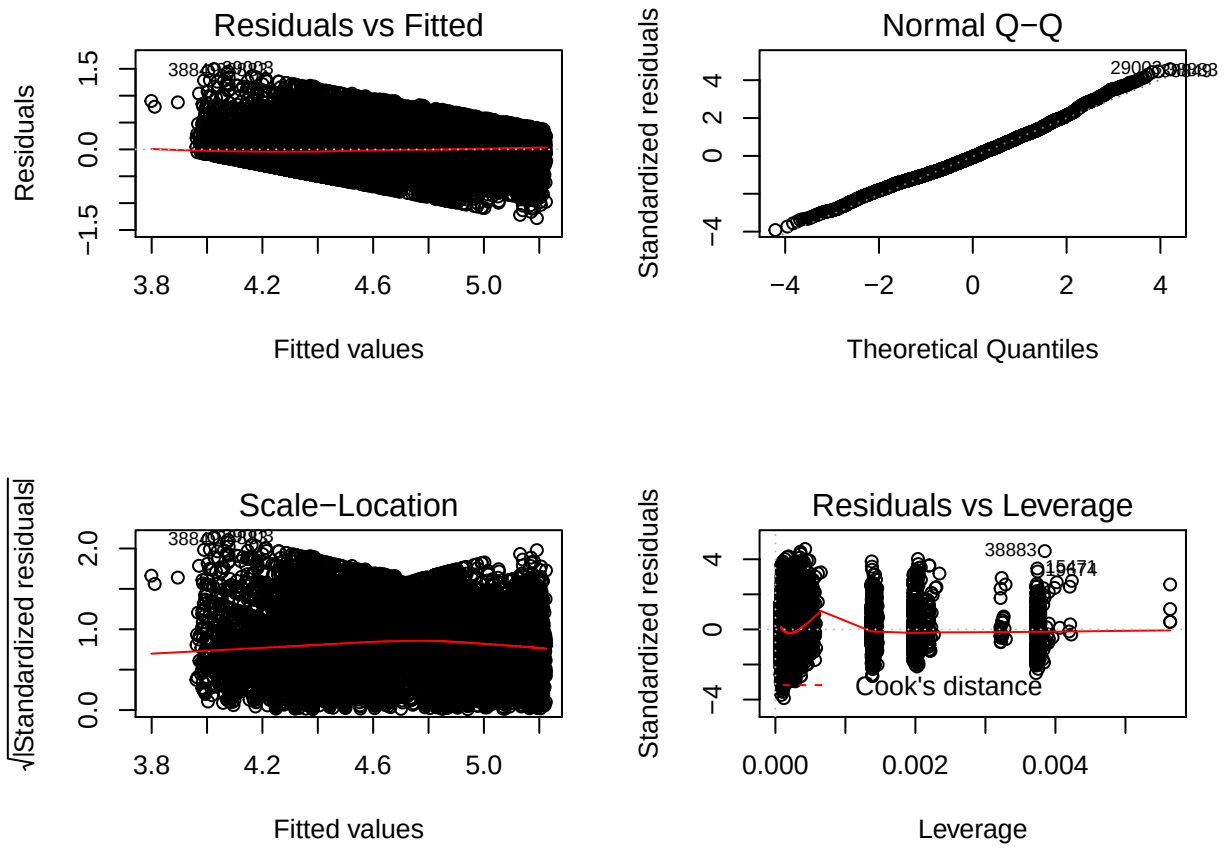

Figure 7: Diagnostic plots

The diagnostic plot is a tool to help to fix problems in our data using *neighbourhood_group*, *room_type*, and *distance* (Figure 7).

The first plot is Residuals vs fitted: From this plot, we can say that there are no patterns and residuals are distributed around zero, but we can see outliers clearly.

The second plot is Normal Q-Q: From this plot, we can say that there is a straight line, which means residuals are almost normally distributed.

The third plot is Scale-location: This is scaled residuals against fitted values. Residuals are distributed around zero, but we can see outliers from this plot also.

The Forth plot is Residuals vs Leverage: From this plot, we can diagnose outliers. Compared to previous plots this plot helps more to identify the outliers.

Training and test data

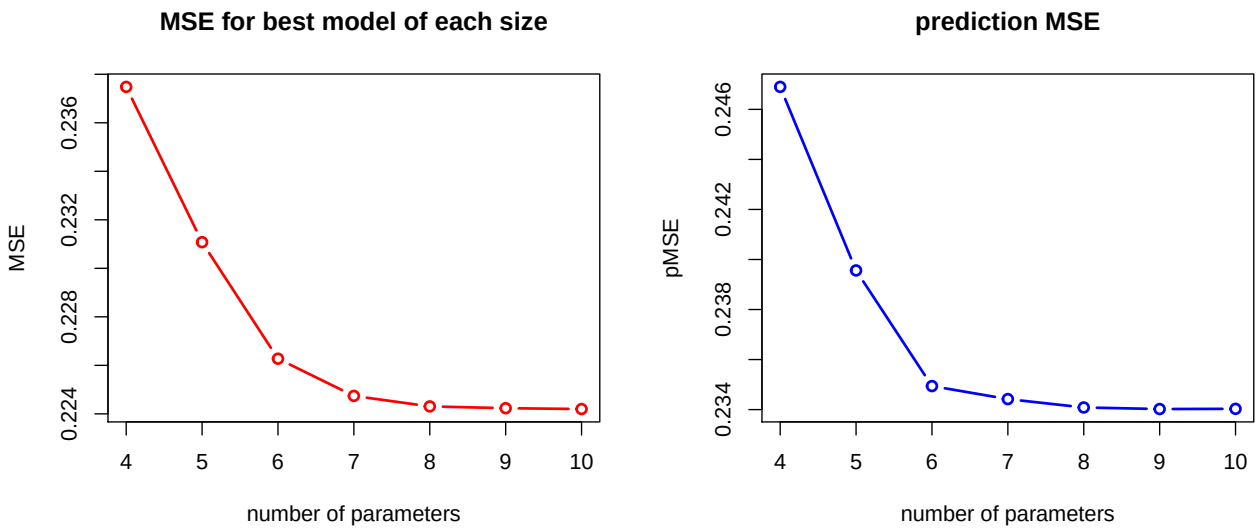

Figure 8: MSE and pMSE curves

To select the best subset we used regsubsets() which choose best with the minimum residual sum of squares (RSS). We fitted all models to the training data and collected the mean squared error- MSE of these models fits. Next, we applied the fitted models to testing data for checking the model ability to predict and collected the predicted mean squared error- pMSE for these models fits. Our first model contained the response (price), 7 numerical variables and 1 categorical variable (room type) and data used 80% as a training set and 20% as a test set.

We checked the best model for MSE and prediction MSE (*Figure 8*). Both curves suggest models with 7 parameters. But, we noticed the different choice of parameters for training MSE and prediction MSE. We didn't choose any best model from this result. Instead, we further experimented with using different models.

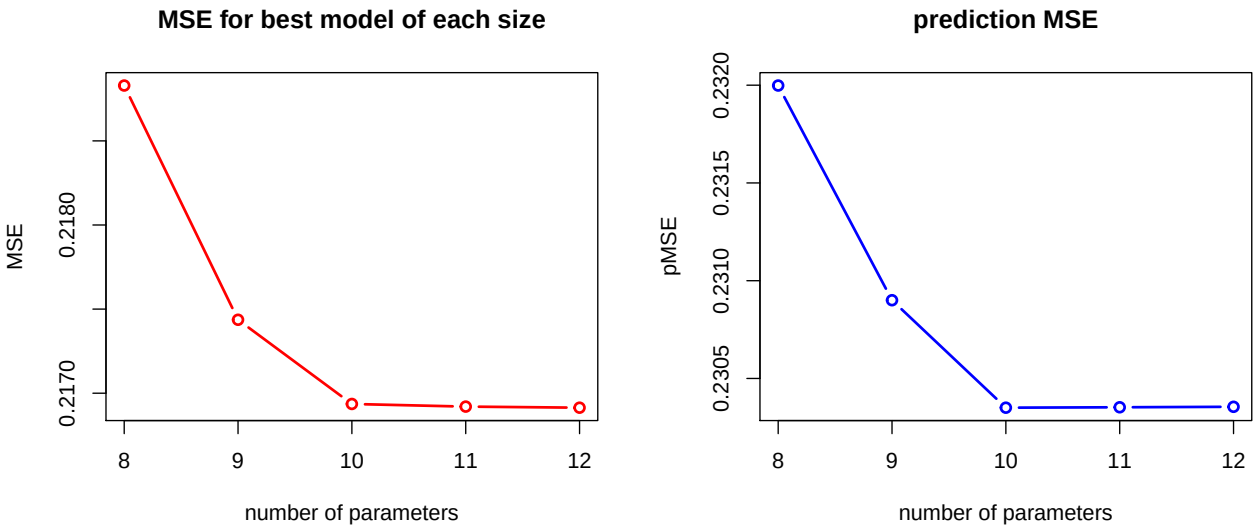

Figure 9: MSE and pMSE curves

In One of the models, we used two categorical variables - *neighbourhood_group* and *room_type* and 5 most important numerical variables: *minimum_nights*, *number_of_reviews*, *reviews_per_month*, *calculated_host_listings_count*, *availability_365*. We changed the fraction of training and test sets to 70% and 30%.

Figure 9 shows the MSE curve and pMSE curve for the best model of each size. We compared the best models that MSE and predicted MSE suggested. In this case, MSE suggests the model with 11 parameters, pMSE suggests the model with 10 parameters. We chose the model with 10 parameters which have the smallest pMSE equals 0.2303505 and this model is better also for model complexity. We changed the fractions to 50-50, 60-40 and checked the min pMSE. In all cases, there is no big difference between pMSE-s of best models. We decided to use the 70-30 fraction.

Stepwise backward selection

We performed the stepwise backward selection for the same full model (the model which has both *neighbourhood_group* and *room_type*). We used different model selection measures; Bayesian information criterion (BIC) and Akaike information criterion (AIC). We got the same results from stepwise with AIC and stepwise with BIC selection which is not the same with pMSE selection result.

Table 3: Compare models

	Residual standard error	R-squared	Adjusted R-squared
pMSE Result	0.4787	0.4798	0.4797
Stepwise Result	0.4701	0.4983	0.4982

We had two different models-stepwise selection results and pMSE selection model. We compared residual standard error, r-squared and adjusted r-squared for them (*Table 3*). Because stepwise selection result has smaller residual standard error, higher r-squared and adjusted r-squared, we chose stepwise selection result-the model with *minimum_nights* ,*number_of_reviews*, *availability_365*, *room_type*, *neighbourhood_group*.

Conclusion

During the mini analysis, we learned about different data transformations, when data transformation is needed and how they affect the fitting model, fitted models, and interpretation of the slope coefficients. We practiced with numerical and categorical variables and use different model selection methods. We also discovered how to validate the fitting model.

The Tv dataset was small and we thought that with more data or information we could have more interesting results. Airbnb dataset was a nice challenge. The data contained several categorical variables that needed some special tools and careful interpretation.

Starting any analysis could be challenging and confusing but, learning more about the variables and different tools that we can use for different types of data and situations make the process more interesting. In addition, putting a plan and staring by asking several questions about the data is an essential part to start an effective analysis where the goal will be addressing these questions and get to the optimal results.

Take-home message: Do not be afraid to explore and try new and different things with your data. One question will lead to more exciting questions that will teach you new things all the way.

Project

Introduction

This project will contain two parts.

Part 1

The first part of the project explores a dataset contains 205 cars that were on the market in the 80's with 26 different characteristics (Car Name, Fuel type, Horsepower, ...) of each car.

This analysis tries to answer some questions related to car price:

1. What factors on which the pricing of cars depends.
2. Which variables are significant in predicting the price of a car?
3. How well these variables/chosen model describe the price of a car?

In this analysis, we modeled the cars prices with different combinations of the available independent variables and tried to show how the prices vary with the independent variables. Our goal was to find a model that is simple and safe to make price prediction as well as easy to interpret.

Different methods were introduced in this analysis such as the backwards step algorithm to reduce our decided full model, training and testing data approach (80-20%), Cross validation and LOOCV.

Data preparation

We checked our dataset for prices equals zero or NAs (missing data). In our dataset, we do not have any missing data or zero prices. Because of the wide range of CarName levels (147 levels), we extracted company names from the Car name variable, and created a new variable CarCompany. We checked also for spelling errors, found some errors in CarCompany, and corrected them. We also removed Car_id and CarName from our dataset. Some of the variables were transformed using natural log transform to display more symmetric distributions and move big data points closer together and spread out the smaller ones.

Exploring the Data

Car Prices

Table 4: Descriptive statistics of price

	Price
Min.	5118.00
1st Qu.	7788.00
Median	10295.00
Mean	13276.71
3rd Qu.	16503.00
Max.	45400.00

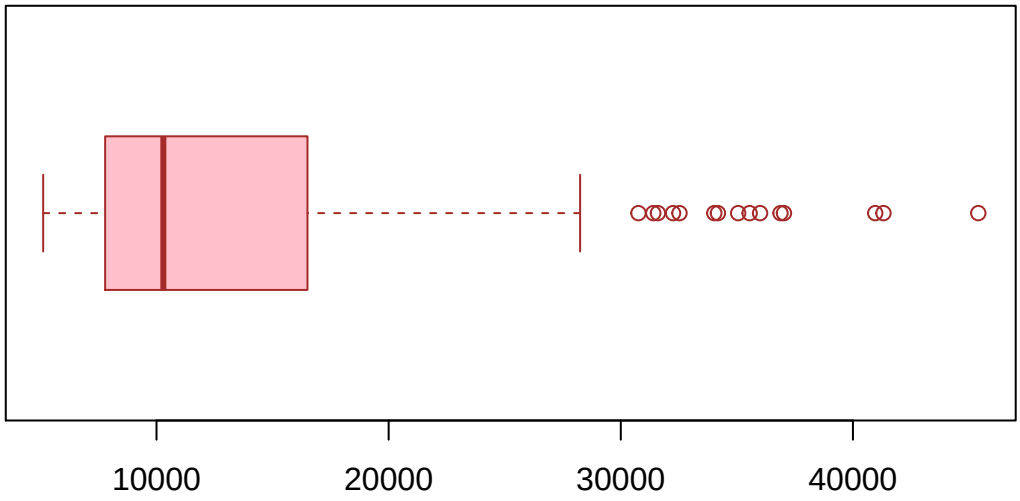

Figure 10: Car price boxplot (US dollar)

The boxplot (*Figure 10*) and *Table 4* show that car prices lie between $7800 and $16500 with the majority of car prices are less than $18000. The average car prices are approximately $13000. There are a few large data points above the maximum with a maximum price reaches $45400.

These outliers make the price distribution right-skewed. Log transformation will be implemented to make the distribution more symmetric.

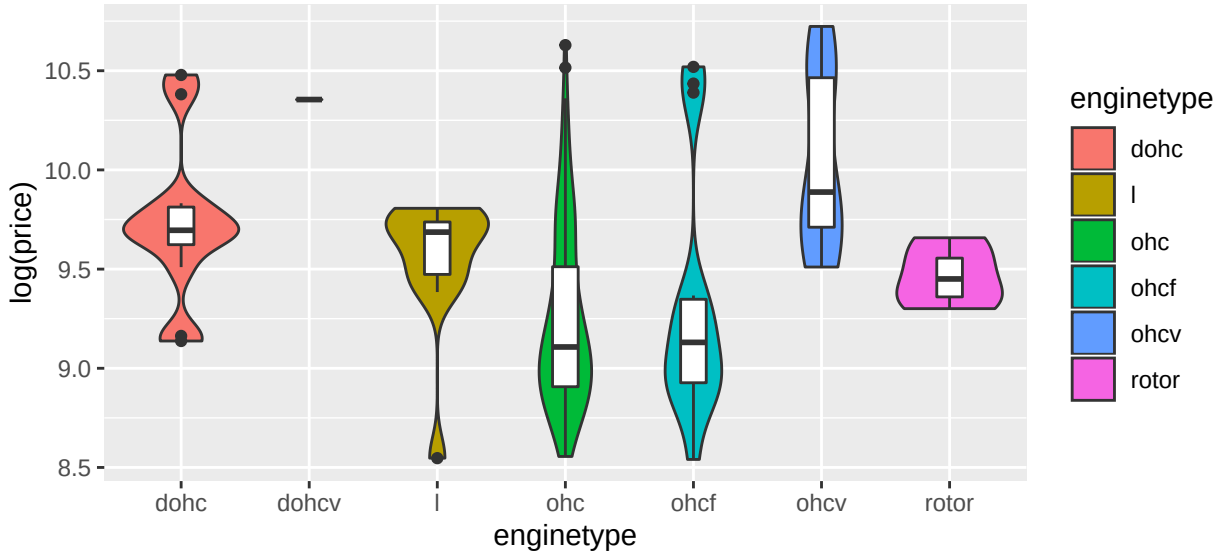

Figure 11: Price by Engine type

Violin and box plot (*Figure 11*) show a couple of things about the distribution of prices.

First, we can see that 'ohcv' engine has the highest range of prices with approximately $19000 (as the median), followed by 'dohc' and 'l' with approximately $16000. 'dohcv' engine has one data point with the

highest price = $31400, 'ohc' and 'ohcf' is the cheapest among them all.

Figure 11 also shows the distribution of engine type's prices. Both 'ohc' and 'ohcf' has a right-skewed distribution with high price outliers while 'l' has a left tail with less prices. The 'rotor' engine prices are concentrated around the median $12745. 'ohcv' displays a uniform distribution.

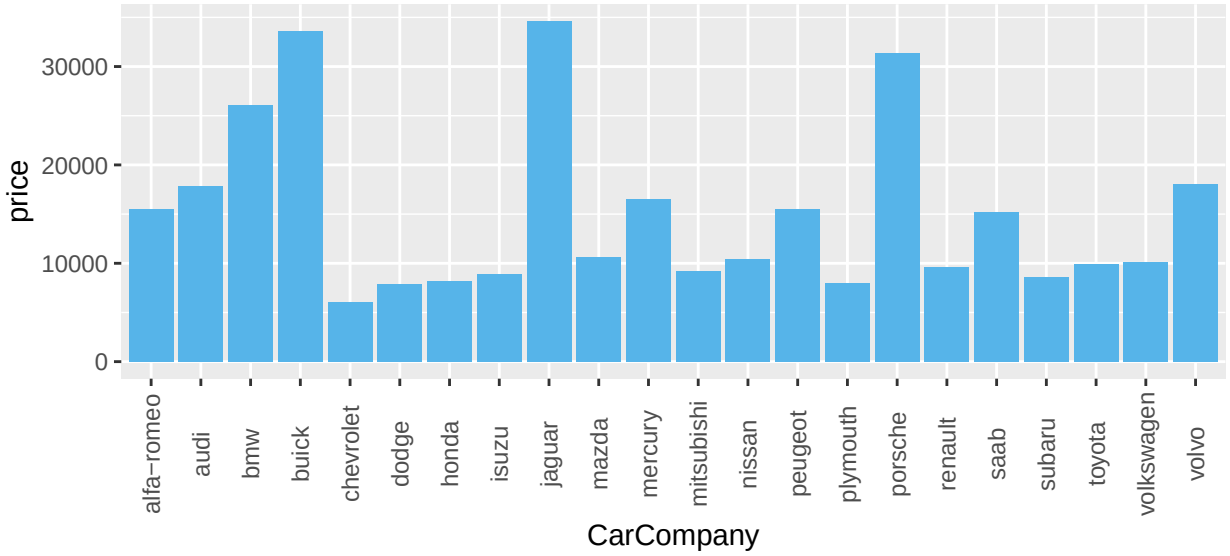

Figure 12: Car Company Prices

From *Figure 12* we can see the average price of cars for each company, with Toyota has average price = $9900, and Mercury has average price = $16500. The company with the highest average price is Jaguar with $34000, and the least average price is Chevrolet with $6000.

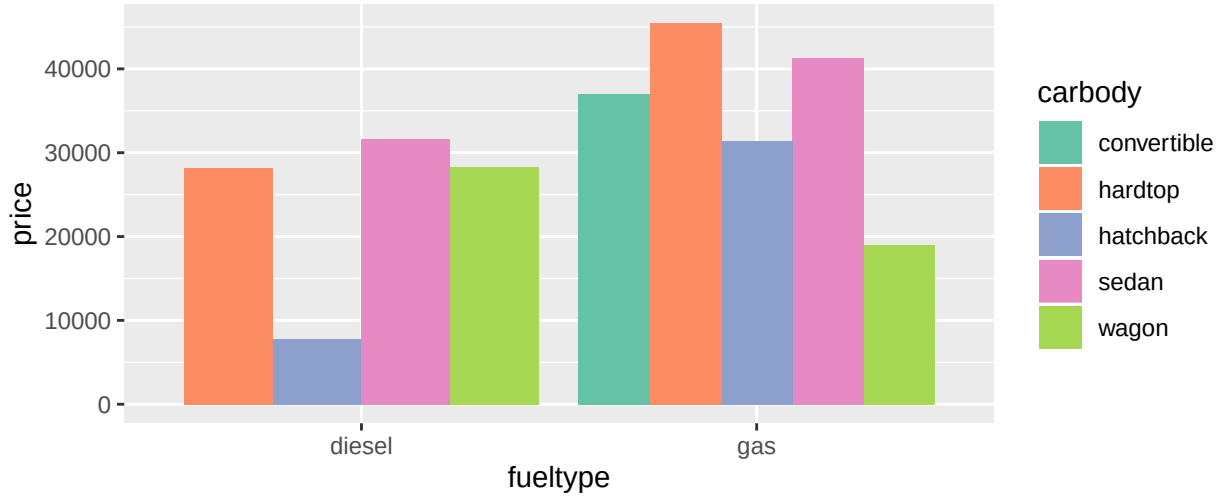

Figure 13: Prices for carbody per fuel type

Figure 13 shows that with the use of gas, hardtop carbody has the highest price = $45400, followed by sedan with price = $41315, while wagon has the lowest price among them with price = $28248.

For diesel fuel, sedan has the highest prices = $31600 followed by wagon with price = $28248, and hatchback has the lowest price = $7788 in this category. We can see that the price for gas fueltype is more expensive than diesel fueltype for all carbodies.

Interaction between horsepower and fuel type

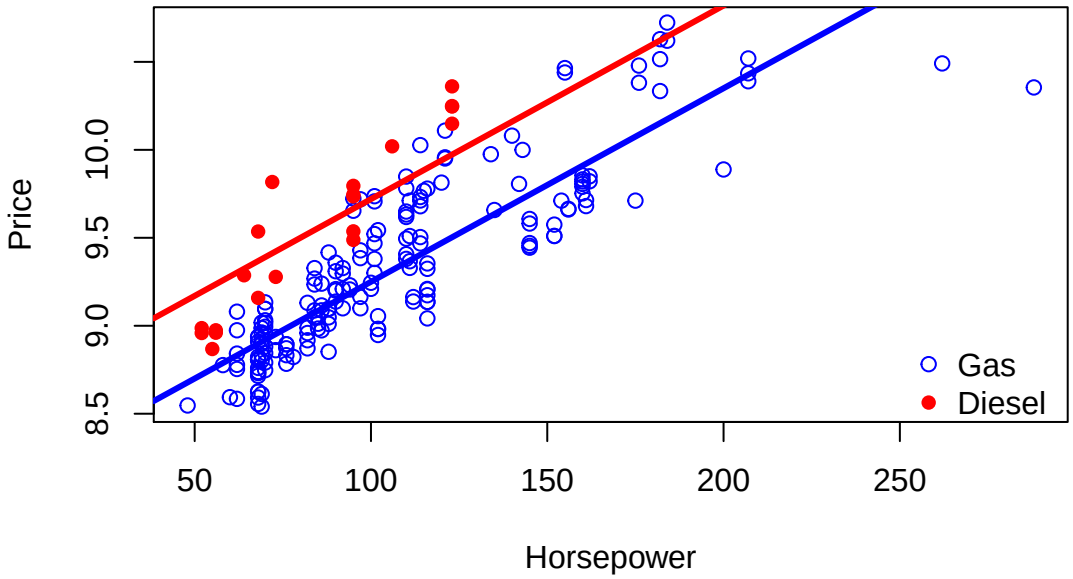

Figure 14: Interaction between horsepower and fuel type

Figure 14 says that horsepower has an effect on the price, as horsepower increases by 1 unit we expect the average price to increase by 0.011 (horsepower coefficient) and the increase assumed to be the same for gas & diesel (the two lines having the same slope) In addition, gas has an effect on the price, for gas, the average price decreased by 0.45 (fuletype coefficient), this effect assumed to be the same for all horsepowers. The effect of horsepower is independent of gas and the effect of gas is independent of horsepower (parallel lines), which means there is no interaction between fueltype and horsepower

ANOVA for categorical variables

Null hypothesis: the mean price is the same for all groups (in a specific covariate)

Using aov() with price and enginetype, we had a p-value that is significant (<0.05) so we reject the null hypothesis:the mean price is the same in all enginetype levels. To check which enginetype may differ from the others we used TukeyHSD test to see all comparisons.

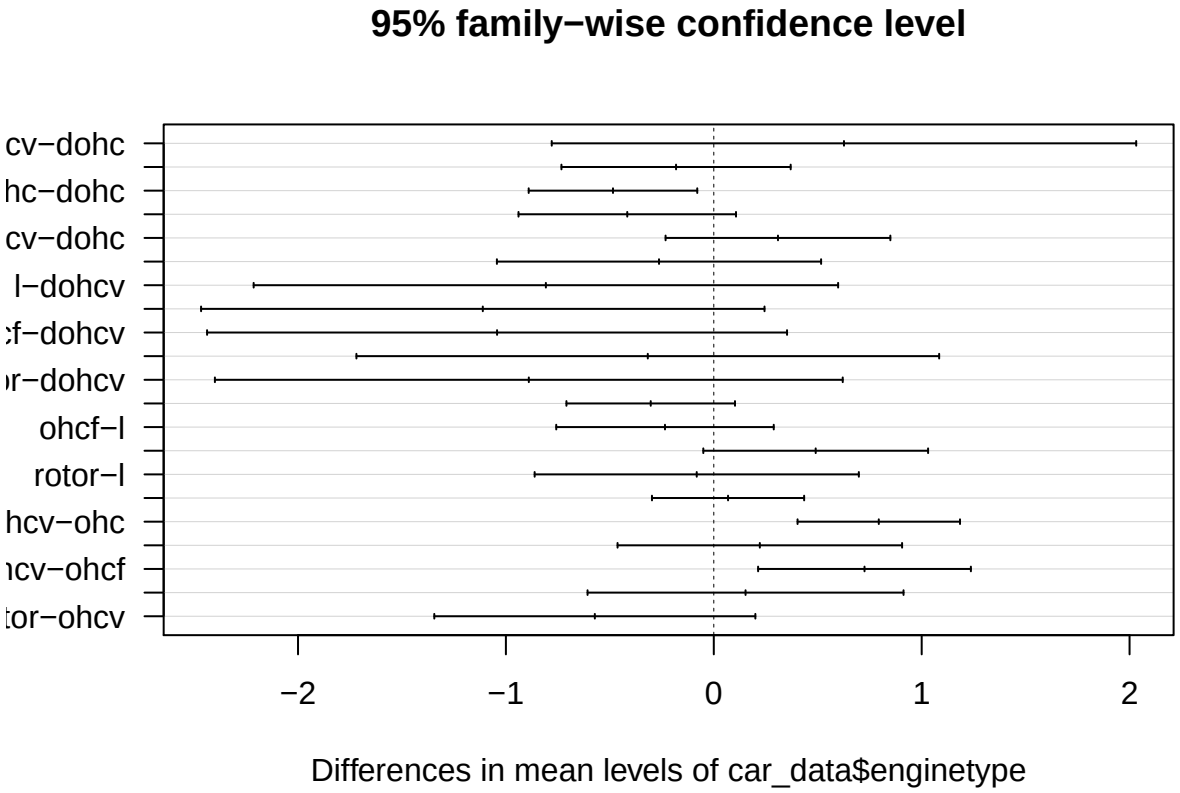

Figure 15: Differences in mean levels of enginetype

Returned overall 95% confidence intervals for the difference in means of all possible pairs for the enginetype covariate. From TukeyHSD() we got that (ohc-dohc & ohcv-ohc & ohcv-ohcf) are significant, which means that the means of ohc and dohc are significantly different (same for ohcv-ohc & ohcv-ohcf). The other group means do not differ (these groups contain 0 in their confidence intervals and thus, have no significant difference). From *Figure 15*, the levels with different means do not contain zero in their intervals

Correlation and Multicollinearity

We used log transformation for price, horsepower, enginesize, citympg, highwaympg to deal with the problem of non-constant error scatter.

carwidth , carlength, curbweight ,enginesize ,horsepower seems to have a strong positive correlation with price (r > 0.6) , carheight does not show any significant pattern with price. citympg and highwaympg seem to have a significant negative correlation with price (r = -0.7)

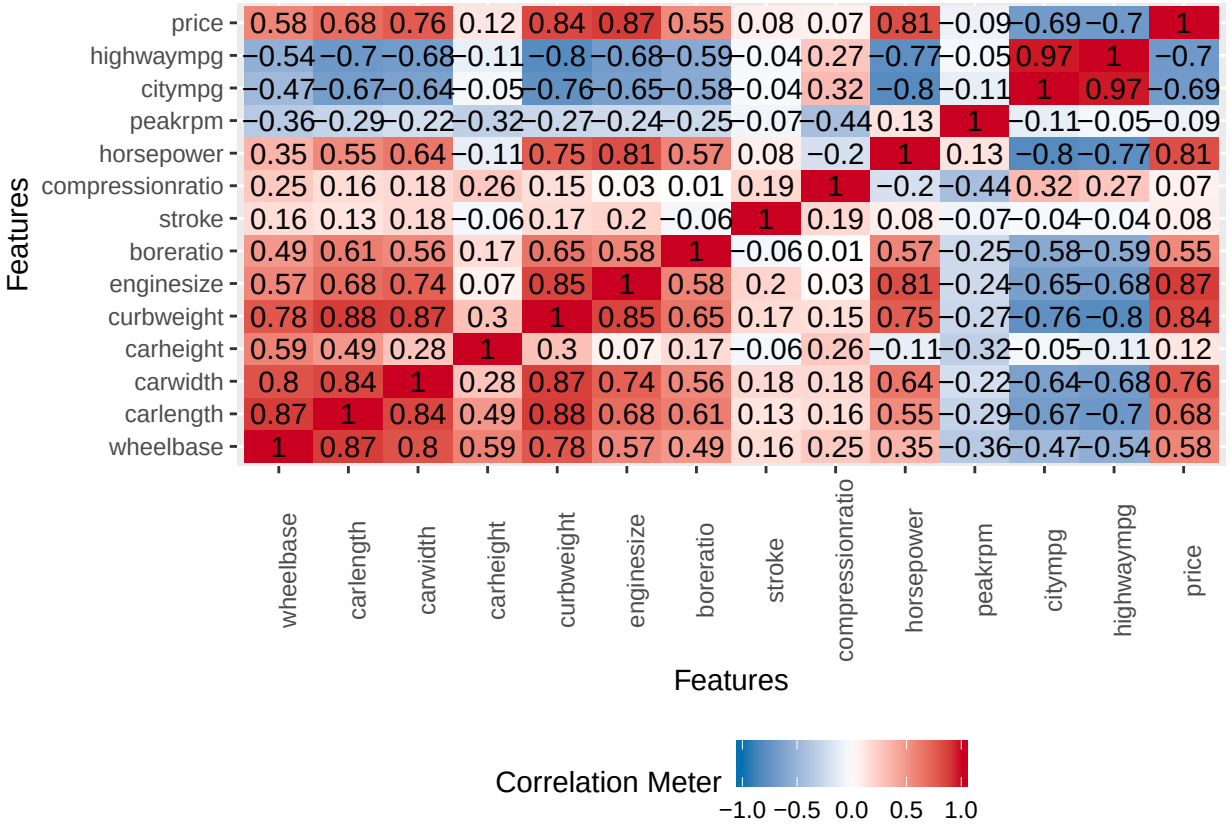

Figure 16: A correlation heatmap for numerical categories

The correlation plot (*Figure 16*) shows that we have correlations between predictor variables. curbweight and highwaympg have a strong negative correlation = -0.8, and highwaympg have a strong positive correlation with citympg = 0.97.

While checking the correlation between numerical covariates, we saw that some variables are highly corellated. We checked the multicollinearity between variables by using Eigensystem analysis and VIF().

Eigensystem analysis

Eigensystem analysis checks how each variable correlate with all other variables not just pairwise. The Eigen values for each predictor values vary in magnitude from 6.7 (large) to 0.02 (small) which means we have much multicollinearity between predictor variables.

Table 5: Eigensystem analysis

Max	Min	Condition number
6.673852	0.0196329	339.932

We quantified the range of Eigen values using the condition number which is the ratio of max and min Eigen values. From *Table 5* we can see that the condition number = 339.932. The rule of thumb is that if the value is on the order of 100 or more then there is significant multicollinearity between covariates. This means we have a multicollinearity problem in the dataset.

Variance Inflation Factor (VIF)

We also used VIF for identifying multicollinearity, which computes the extent of correlation between the predictors in a model. We fitted a model with all numerical variables and checked for multicollinearity between them using VIF(). After performing a linear regression, we looked for the inflation of the increase in the variance for a particular variable that comes about because of the correlation of that variable with the other predictor variables.

Includ all of the variables with high multicollinearity will result in increased standard errors and some of the variables may not end up being significant. The rule of thumb is that if VIF is 5 or larger we worry about it. The VIF of 4 in the standard error for the confidence interval around the coefficient will be twice as big as they would be otherwise. If we have VIF bigger than 10 then we have to do something about it.

Table 6: VIF for numerical covariates

	VIF
wheelbase	7.340949
carlength	9.422999
carwidth	5.586367
carheight	2.205975
curbweight	16.413371
enginesize	6.658982
boreratio	2.103912
stroke	1.195781
compressionratio	2.175511
horsepower	8.247880
peakrpm	2.053763
citympg	27.128671
highwaympg	24.277439

VIF for numerical covariates is shown in *Table 6*. In our analysis, we decided to remove the variables with VIF above 10. We removed *citympg* with VIF equals 26.6 and then checked VIF after removing the variable. We removed *curbweight* with VIF 16.1 and checked VIF again.

Choosing a Model

Stepwise Backward selection by using BIC

After VIF checking between numerical variables, we added some categorical variables to our model. We continued with the model below:

$log(price) \sim wheelbase+carwidth+carheight+log(enginesize)+boreratio+stroke+log(compressionratio)+peakrpm + highwaympg + fueltype + carbody + enginetype + cylindernumber$

After choosing the full model, we implemented different selection methods to choose the best model for prediction. We started by using the stepwise backward selection method based on the Bayesian information criterion (BIC). As a result, step function removed *carlength*, *wheelbase*, *carheight*, *log(compressionratio)*, *peakrpm*, *highwaympg*.

Using Traning and Test data

We split our dataset to 80% (Training set) and 20% (Testing set). We used regsubsets() for selecting best subset, which quantifies best model using the residual sum of squares (RSS).

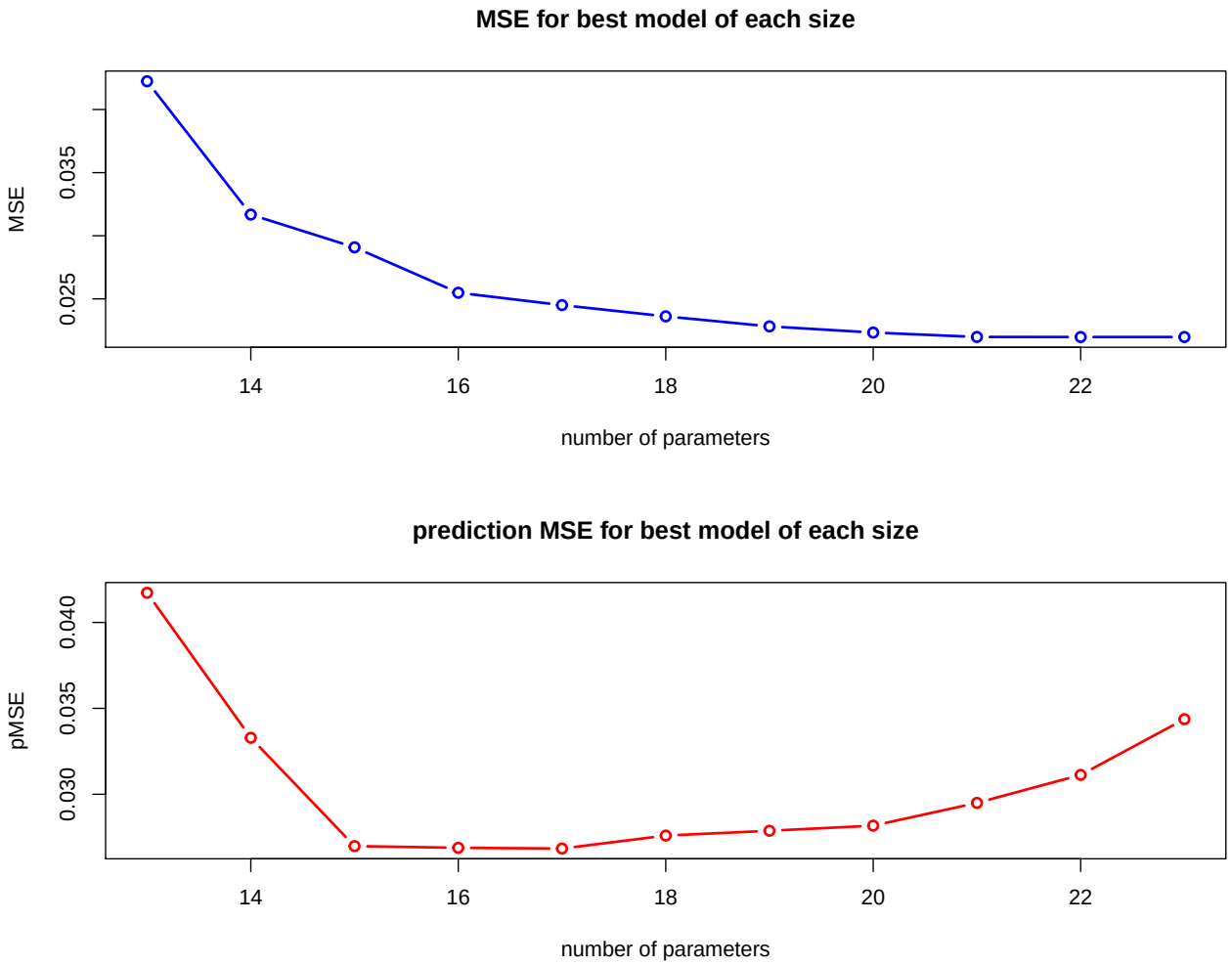

Figure 17: MSE and pMSE curves

We fitted all models to the training data and collected the mean squared error- MSE for these models fits. Next, we applied the fitted models to testing data to check the model ability to predict, and collected the predicted mean squared error- pMSE for these models fits. *Figure 17* shows the MSE curve and pMSE curve for the best model of each size.

In our case, the model with 17 parameters (with factors) gives the smallest pMSE where pMSE equals 0.02684. But we decided to choose a model with 16 parameters (with factors) pMSE equals 0.02688. There is no big difference between model pMSES with 17 and 16 parameters. The model we chose based on pMSE:

$log(price) \sim fueltype + carbody + enginetype + log(horsepower) + log(enginesize) + boreratio + stroke$

Cross-validation and LOOCV

To test how good our model that we chose using prediction error, we performed a cross-validation and LOOCV methods to validate the model from previous tests. We performed the CV and LOOCV using 10 folds. After cross-validation we got the same model as pMSE model. But LOOCV removed *boreratio* and *stroke* from our model.

We ended up with 4 different models:

- Model after stepwise bacward selection- step_model: $log(price) \sim carwidth + log(horsepower) + log(enginesize) + boreratio + stroke + fueltype + carbody + enginetype$
- Model after using prediction error pMSE- pmse_model: $log(price) \sim fueltype + carbody + enginetype + log(horsepower) + log(enginesize) + boreratio + stroke$
- Model after CV is the same with the pMSE model.
- Model after LOOCV- loocv_model: $log(price) \sim fueltype + carbody + enginetype + log(horsepower) + log(enginesize)$

Because our models are nested, we tried ANOVA test to decide which one will be using as the final model.

Table 7: Comparing loocv_model and pmse_model

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
191	6.878451	NA	NA	NA	NA
189	5.707389	2	1.171061	19.38983	0

The ANOVA test has a hypothesis:

- Null hypothesis: there is no significant difference between RSS for the full (pmse_model) and reduced (loocv_model) models
- Alternative hypothesis: The full model (pmse_model) has lower RSS (better) than the reduced model (loocv_model)

  From *Table 7* we can see that RSS for the full (pmse_model) has better RSS compared to reduced model (loocv_model), also based on the p-value of the test = 2.19e-08 (almost zero), we reject the null hypothesis, we believe that pmse_model is better than loocv_model. We rejected LOOCV model and continued with two models. They are nested; we can do the ANOVA test again.

Table 8: Comparing step_model and pmse_model

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
189	5.707389	NA	NA	NA	NA
188	5.045311	1	0.6620781	24.67057	1.5e-06

From *Table 8* we can see that RSS for the full (step_model) is better than the RSS for the reduced model (pmse_model), also based on the p-value of the test = 1.5e-06 (almost zero), we reject the null hypothesis, we believe that step_model is better than pmse_model.

We chose our final model, which is the same as the step_model.

Final model: $log(price) \sim carwidth + log(horsepower) + log(enginesize) + boreratio + stroke + fueltype + carbody + enginetype$

Identifying Outliers

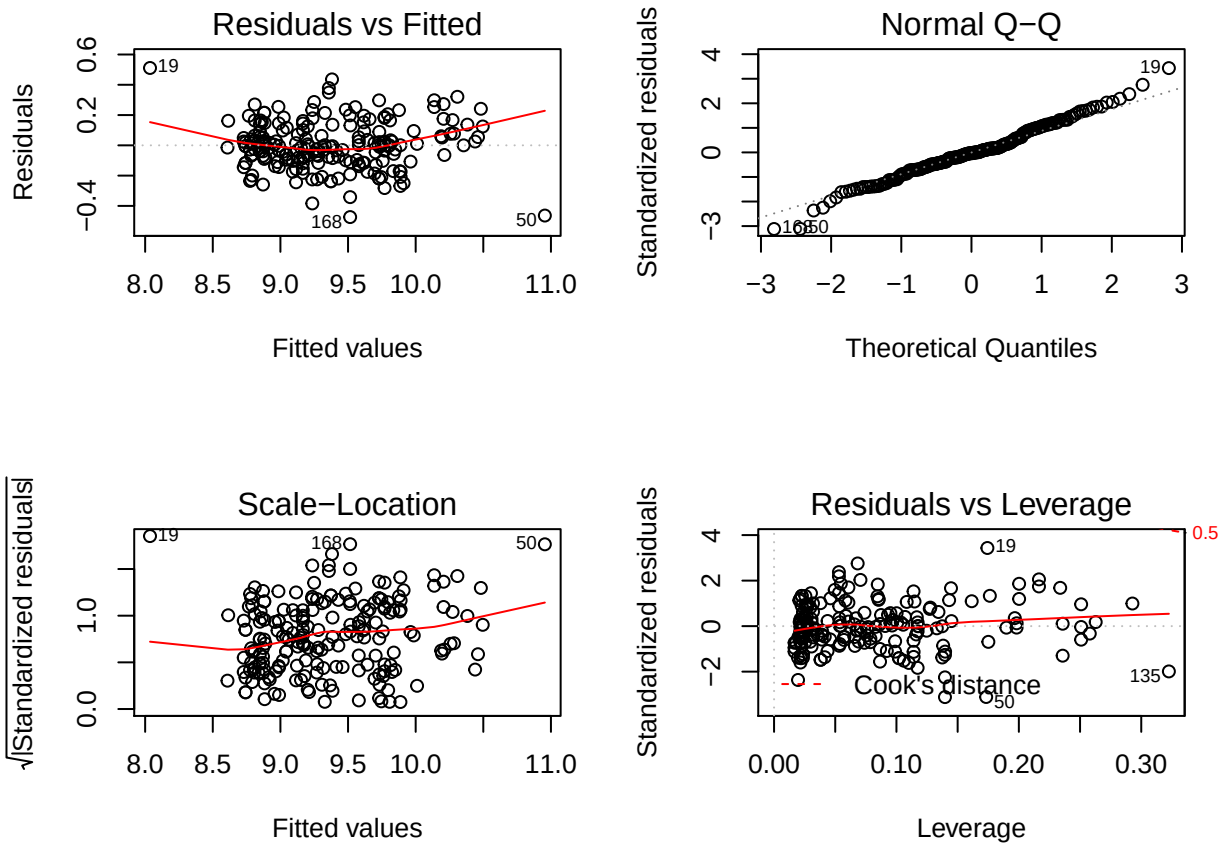

Figure 18: Diagnostic plots for final model

From *Figure 18* we can see that:

- The residual vs fitted plot shows that the linearity assumption is not quite met cause the line is not fairly flat and there is some pattern (curve-nonlinear)
- The Q-Q plot shows that points below and above do not follow the normal line and that is the errors/residuals are not normally distributed
- The scale-location plot shows that the variance varies with the fitted values (the curve is not flat)
- The residuals vs leverage plot show some extremes, there are 3 extremes in this model.

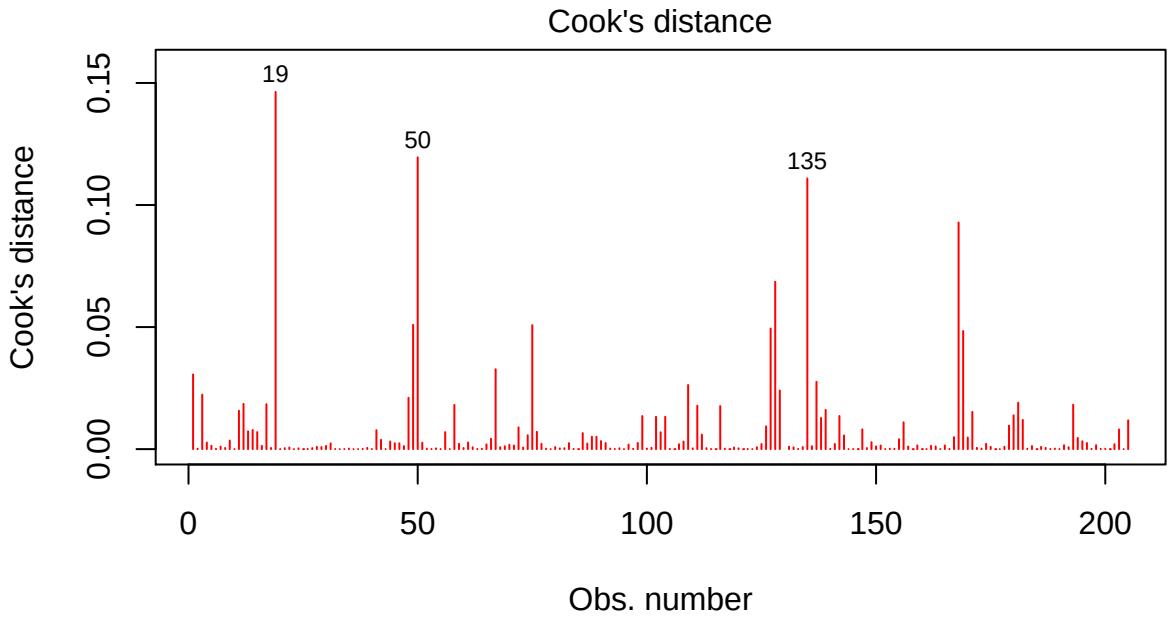

Figure 19: Cook's distance for chosen model

Cook's distance (*Figure 19*) plot shows that we have 3 points with high cook's distance: 19, 50, 135. After gathering all points from top residuals, influence points and standardized residuals, we picked out most frequency observations and extract these data points: *19, 50, 67, 128, 168, 169, 135*. After removing the outliers we checked the Diagnostic plots and saw that we have better fitting.

Model Intrepretation

Table 9: Summary for model parameters

	Estimate	CI (lower)	CI (upper)	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.0995037	-0.7482598	0.9472673	0.4296483	0.2315935	0.817	
carwidth	0.0543837	0.0364504	0.0723170	0.0090887	5.9836898	<0.001	***
log(horsepower)	0.7019990	0.5628373	0.8411608	0.0705274	9.9535585	<0.001	***
log(enginesize)	1.0407357	0.7991150	1.2823565	0.1224539	8.4990012	<0.001	***
boreratio	-0.2735058	-0.4140185	-0.1329931	0.0712121	-3.8407205	<0.001	***
stroke	-0.4045706	-0.5085734	-0.3005677	0.0527088	-7.6755724	<0.001	***
fueltype: gas	-0.2195130	-0.3066798	-0.1323463	0.0441763	-4.9690238	<0.001	***
carbody: hardtop	-0.1350748	-0.3091520	0.0390023	0.0882226	-1.5310679	0.127	
carbody: hatchback	-0.3038519	-0.4327360	-0.1749678	0.0653187	-4.6518361	<0.001	***
carbody: sedan	-0.2030358	-0.3291238	-0.0769479	0.0639016	-3.1773199	0.002	**
carbody: wagon	-0.2153569	-0.3514086	-0.0793051	0.0689513	-3.1233181	0.002	**
enginetype: dohcv	-0.3227101	-0.6293720	-0.0160481	0.1554169	-2.0764157	0.039	*
enginetype: l	0.0732688	-0.0634677	0.2100053	0.0692983	1.0572951	0.292	
enginetype: ohc	0.1653576	0.0683097	0.2624055	0.0491841	3.3620153	0.001	***
enginetype: ohcf	-0.0642034	-0.1978267	0.0694199	0.0677206	-0.9480636	0.344	
enginetype: ohcv	-0.2386120	-0.3672311	-0.1099930	0.0651844	-3.6605690	<0.001	***
enginetype: rotor	0.7963014	0.5803184	1.0122844	0.1094606	7.2747758	<0.001	***

From *Table 9* we can say:

- When enginesize increases by 1%, the car price increases by 1.04% (holding all other variables constant).
- When horsepower increases by 1%, the car price increases by 0.7% (holding all other variables constant).
- Car width is associated with an increase of 5% in the car price per 1 inch increases in the width.
- A car with a sedan body is estimated to have 0.2 less price per $1 than a convertible body (baseline level) holding all other variables constant.

The engine types *dohcv* ,*ohc* ,*ohcv* and *rotor* are statistically significant, while *ohcf* and *l* are not significant. The difference between the baseline level (*dohc*) and the enginetype *l* average price is not statistically significant, which means that there is not enough evidence to show that there is a difference in average prices between the *dohc* enginetype and enginetype *l* assuming all other variables are constant.

Summary for Car price prediction

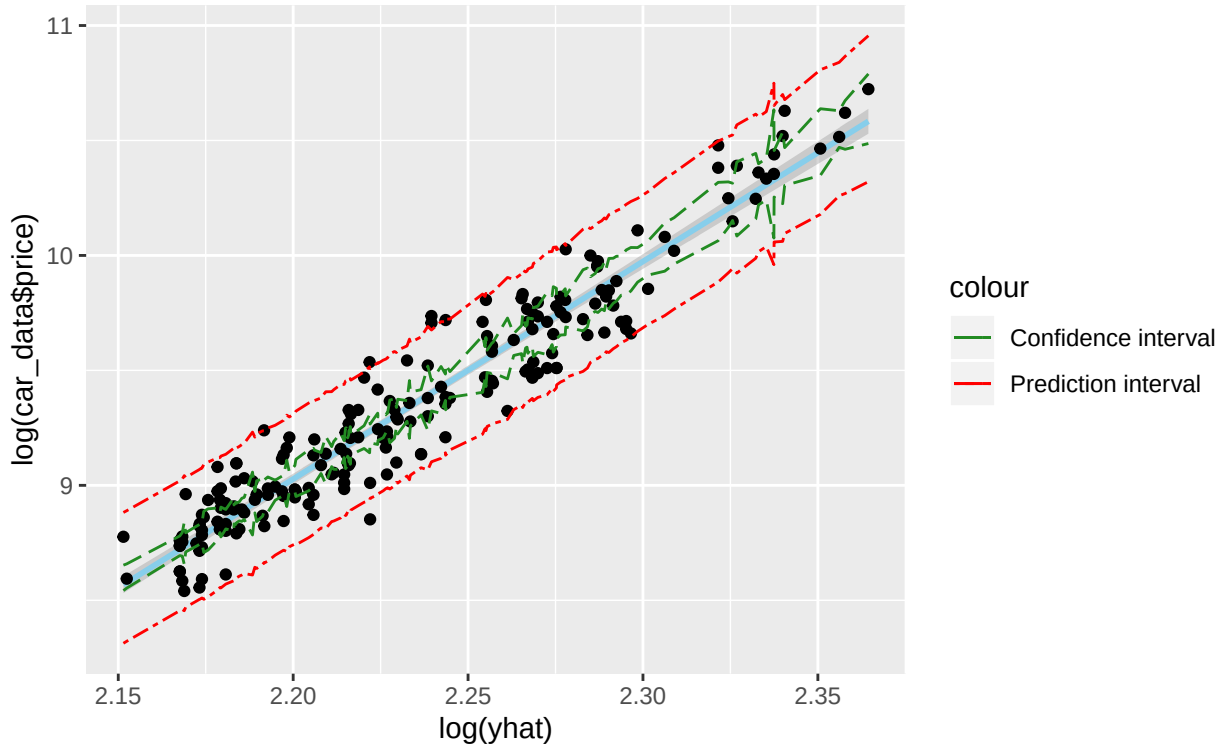

Figure 20: Observed prices vs predicted prices with confidence and prediction intervals

Figure 20 shows that most of the observations lie inside the prediction interval and close confidence interval, which means that our model is good and the prediction interval covered the values.

Table 10: Summary for model

Residual.standard.error	R.squared	Adjusted.R.squared
0.1414279	0.925	0.9183

Conclusion

R-sqaured of the final model = 0.925 and the Adjusted R-squared =0.9183, which means that our model explaines about 92% of variability in the prices. Therefore, our model is good to predict the prices with the chosen predictor variables *Table 10*

Part 2

Introduction

This part explores a dataset contains 46 county with 9 different variables (bird cases, equine cases, human density, ...) for each county. This analysis tries to predict the rate of West Nile Virus (WNV) positive equine in a county.

We modeled the equine cases with independent variables that we think are highly correlated with this variable, and tried to show how equine cases vary with these independent variables.

Data Preparation

In the dataset, the covariates are non-negative integers. We checked data for missing values and created a new variable from our data *equine_rate*:

$$equine\_rate = \frac{equine\_case}{farms}$$

Exploring the Data

To check the correlation between the variables, we used the correlation plot and it showed that there is a strong association between *popul* and *human_dens* (0.96) because of the collinearity between them. The bird_cases has the strongest relationship with equine_cases among other variables. We compared the effect of *bird_cases*, *human_dens* to choose the best model for prediction.

Poisson Model

Table 11: Summary for model parameters

	Estimate	CI (lower)	CI (upper)	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.4580552	-6.8407633	-6.1124374	0.1852890	-34.85396	<0.001	***
bird_cases	0.0442244	0.0306091	0.0568518	0.0066509	6.64936	<0.001	***

From *Table 11*, when bird_cases increases by 1, the equine_cases relatively increases by 1.04 that is a 4% increase (exp(beta1)=exp(0.04)=1.04). This result obtained without any other information from other covariates. Equine_cases may expect to increase according to areas with greater human population densities, because this will increase the probability of people finding dead birds. In addition, more bird cases may result in more human cases.

Estimate of overdispersion: Residual Deviance/ Residual df = 76.968/44 = 1.75

Using dispersiontest() we had a p-value that is not significant (0.06 > 0.05), which means we fail to reject the null hypothesis: mean = variance. The dispersion = 1.9 > 1, this result indicates that we have an overdispersion! Next we check the effect of adding the *human_dens* to our model.

Table 12: Summary for model parameters

	Estimate	CI (lower)	CI (upper)	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.2154666	-6.7439864	-5.7278783	0.2585020	-24.044169	<0.001	***
bird_cases	0.0461387	0.0323743	0.0589397	0.0067290	6.856692	<0.001	***
human_dens	-0.0017186	-0.0046951	0.0008364	0.0014012	-1.226511	0.22	

From *Table 12* we can say that for a fixed bird cases, a unit increase in the human density predicts a 1% decrease in the number of equine cases. (exp(-0.002) = 0.996) –> about 1) When human density increases by 1000 (holding bird cases fixed), we would expect a decrease of equine cases equal to exp(1000*0.01) = 10. Residual deviance is larger than the degrees of freedom, which means that we have overdispersion.

Do we need to add human density when bird cases is in the model?

Likelihood ratio test between the two models

Table 13: Anova test

Resid. Df	Resid. Dev	Df	Deviance
44	76.96821	NA	NA
43	75.31092	1	1.657288

Difference between deviances is 1.66 (*Table 13*), which is smaller than 3.84 (Chi-Squared distribution with 1 degrees of freedom), we failed to reject the null hypothesis=reduced model. We do not need the extra info provided by the human density.

The p-value is significant ($0.02 < 0.05$), which means we reject the null hypothesis: mean = variance. The dispersion = $1.8 > 1$, this results indicates that we have an overdispersion! In this case, it is better to use negative binomial.

We plotted predicted rate of WNV-positive equine against observed rate of WNV-positive equine for our Poisson model. We also plotted y=x line to check whether observed and predicted rates are similar or not.

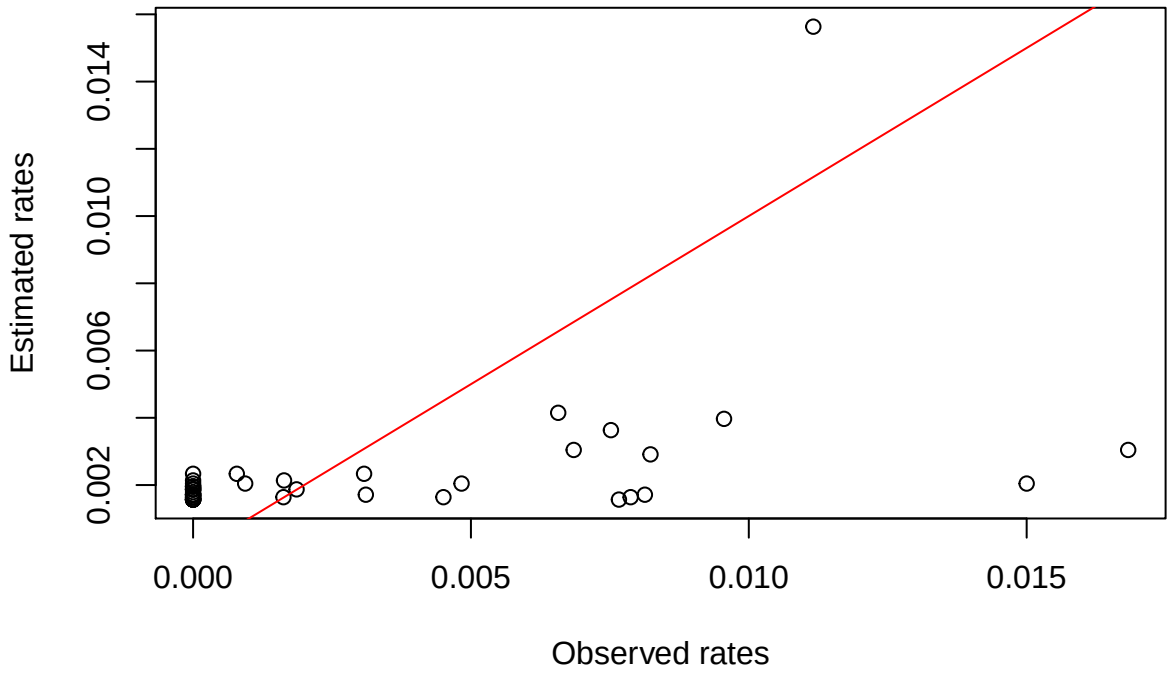

Figure 21: Observed and Predicted rates (Poisson model)

Using slope = 1 shows that observed and predicted rates are not similar (*Figure 21*)

Negative Binomial Model

Table 14: Summary for Negative Binomial parameters

	Estimate	CI (lower)	CI (upper)	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.7029103	-7.365207	-6.098161	0.2766374	-24.229953	<0.001	***
bird_cases	0.0785529	0.032208	0.137296	0.0180477	4.352527	<0.001	***

From *Table 14*, when bird_cases increases by 1, the equine_cases relatively increases by 1.07 that is 7% increase (exp(0.07) = 1.07)

Using the Chi-Square goodness of fit test on the residuals deviance, does not reject the negative binomial fit, which means the negative binomial model is better.

We plotted predicted rate of WNV-positive equine using Negative Binomial model against the actual rate of WNV-positive equine with y=x line to check whether actual and predicted rates are similar or not.

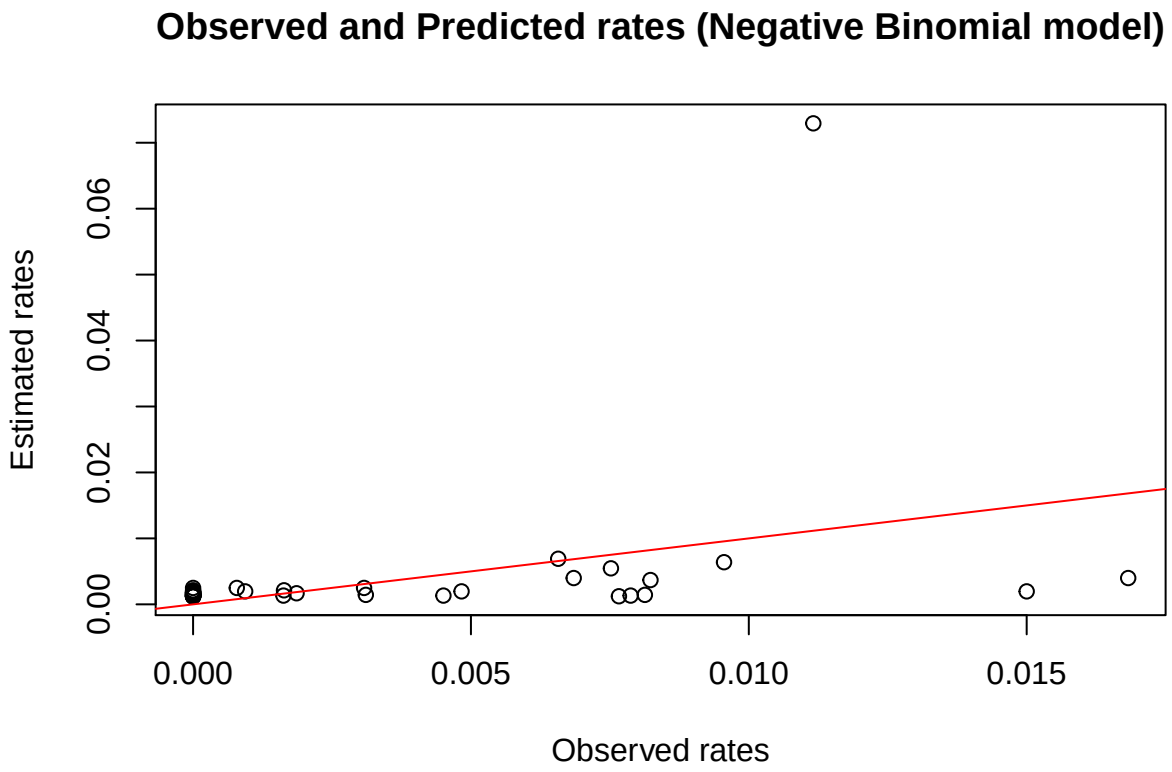

Figure 22: Observed and Predicted rates (Negative Binomial model)

From the graph above, we can see that Negative Binomial model is better and the line is closer to data points comparing to Poisson model (*Figure 22*).

Comparison of Negative Binomial model and Poisson model

Comparing the Negative Binomial model and Poisson model using likelihood ratio test to check the difference between deviances of Negative Binomial model and Poisson model. We calculated 1 degree of freedom (Poisson and NB differ by exactly 1 parameter (Theta)) for the chi-squared distribution (3.84), and compared the result with the difference between deviances (-2*(logLik(g0)[1]-logLik(ng)[1]) = 11). The difference between deviances is larger than chi-squared distribution result and so we reject the Poisson model at 0.05 significance level.

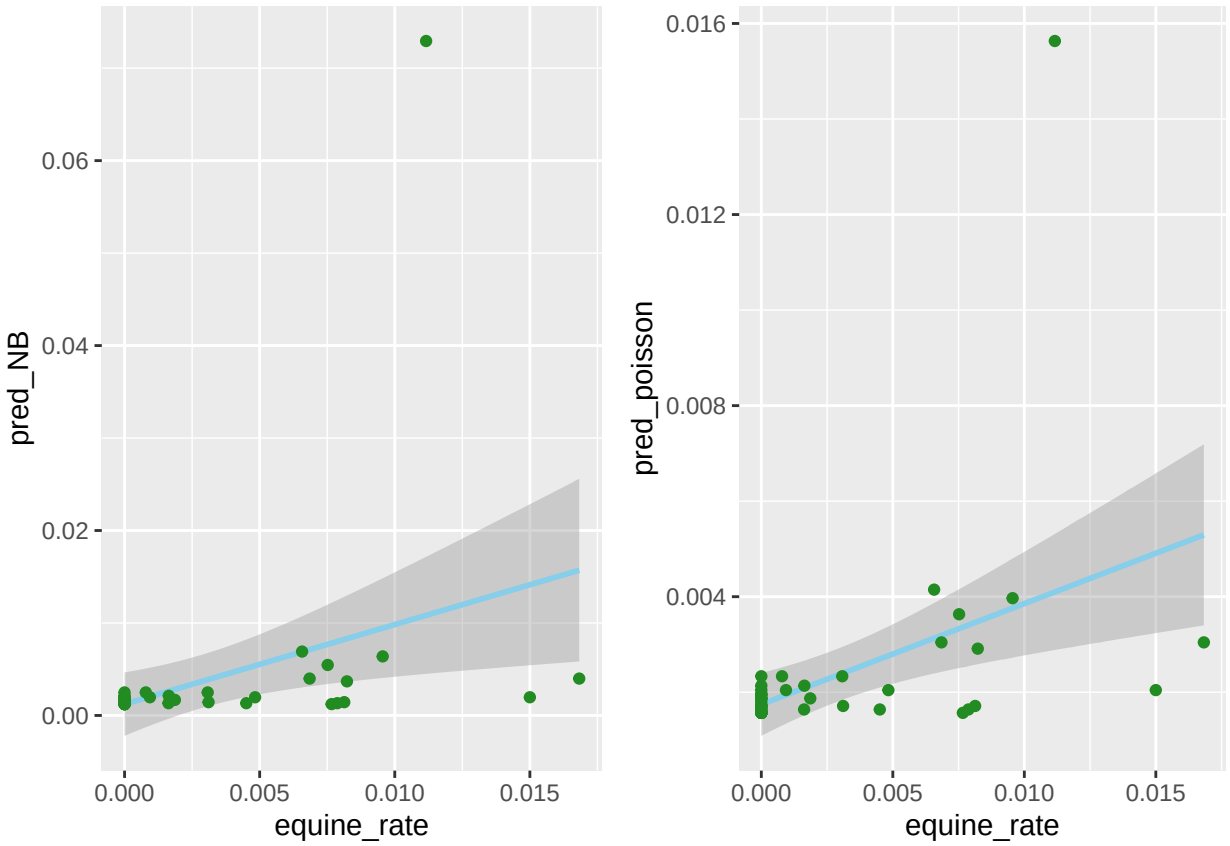

Figure 23: Confidence interval for Poisson and Negative Binomial models

The confidence intervals for both Poisson model and negative binomial model show that the negative binomial model has more capacity and covers more data points (*Figure 23*)

Formal test (lrtest)

Likelihood ratio test is larger than likelihood ratio test critical and so, we reject null hypothesis (Poisson model is better than Negative binomial model). We can say Negative binomial model is better than Poisson model.

Conclusion

The WNV dataset has very small data points, which we think is not enough to use for prediction or having a good model that represents the true relationship between response and predictors. To have more accuracy we need more data to learn more about what affect the equine_cases and test the model for validation.

We can see from graphs that we have some extreme values that affect the line. We decided not to remove any of these values. Removing some of them affected the results significantly, and we are not sure if these points are critical situations meaning they are an important evidence (like Horry county, which have 52 bird cases) for the situation, or it is just a mistake and so, we decided to keep all points and predict based on all data points.