Project 1
Xinyi (Anny) Cui

Why do we use multiple linear regression model in real life? Just imagine a real scenario. In modern life, our daily travel is almost inseparable from the use of cars. So how do we judge how many miles a gallon of fuel will get a motor trend car going? This is determined by multiple factors, such as the number of carburetors, the engine, etc. This paper explores how to construct and fit a multiple linear regression model in a real-world scenario to explain the relationship between a response variable and multiple explanatory variables.

The background part of this report will introduce the data content and explain the method and role of building the model. The analysis section includes data exploration, detailed model building and evaluation with diagnostic plots, and explanations of prediction and confidence intervals. The conclusion section provides brief results, re-analyzes the scope of inference and causality, and discusses the strengths and weaknesses of the model. The paper will also end with a discussion of inference scalability based on the described multiple linear regression model and its applicability.

1 Background

We use the data which comes from the 1974 *Motor Trend* US magazine to build a multiple linear regression model with fuel consumption as the response. This dataset contains data for 32 car models. For each car, there are 11 elements, expressed in different units (US units). Tables below provides the description of the data format.

1.1 Data Context

Data Variables	
Name	Description
mpg	Miles per gallon
cyl	Number of cylinders
disp	Displacement, in cubic inches
hp	Gross Horsepower
drat	Rear axle ratio.
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	Engine (0 = V-shaped, 1 = straight)
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

Table 1: The name and description of the dataset. Given each variable name corresponding explanation by intuition.

The next table provides the first six rows of the dataset.

mtcars Model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Table 2: The example of the data format. The first column is the names of each selected motor trend car model and the other eleven columns represent the values of each variable.

1.2 Method

The main role of a multiple linear regression is to make predictions. Modeling is used to fit the data we have collected, and we perform a parameter estimation after fitting the collected data. The purpose of parameter estimation is mainly to estimate the value of the partial regression coefficient of explanatory variables of the model. After we make the estimation, we can use the data of the new explanatory variables to make predictions.

Specifically, the multiple linear regression analysis in this paper mainly follows the following process:

1. Carry out variable analysis. Explore whether the data needs to be adjusted to better build the model.

2. Determine whether there is a correlation between several specific variables. The model should be formed according to the linear formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

3. Fit and assess the most suitable model.

4. Select the value of several variables, predict the value of the response, and know what kind of accuracy this prediction or confidence interval can achieve.

The more explanatory variables included in a linear regression model, the better it should be able to reflect reality, but the more difficult it is to interpret. We need to pay attention to the rationality of the interpretation of the model, whether it is consistent with the preset intuition, whether it supports our hypothesis, etc. Therefore, we pursue a model that is as reduced as possible and a better interpretation of variation. Often focus on model R-squared, linearity and regression diagnostic issues.

2 Analysis

2.1 Data Exploration

Before constructing the model and do the hypothesis test, we need to explore our data for checking any skewness or ourlier to see whether needs a transformation. After doing densityplots, the graph of gross horsepower seems being slightly left-skewed, so we can do a log transformation for hp values (see Figure 1). Then the new graph displays a better normally distributed densityplot (see Figure 2). Also, the densityplot of quarter mile time seems conclude an outlier (see Figure 3), so we remove it and re-plot the graph (see Figure 4). Other variables have no problems and we do not have to adjust them so we can use other variables directly for modeling.

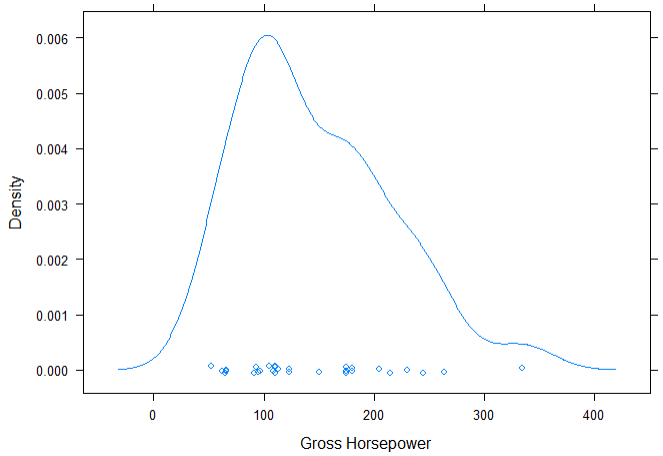

Figure 1: Densityplot of gross horsepower with original data values. The data is slightly left-skewed.

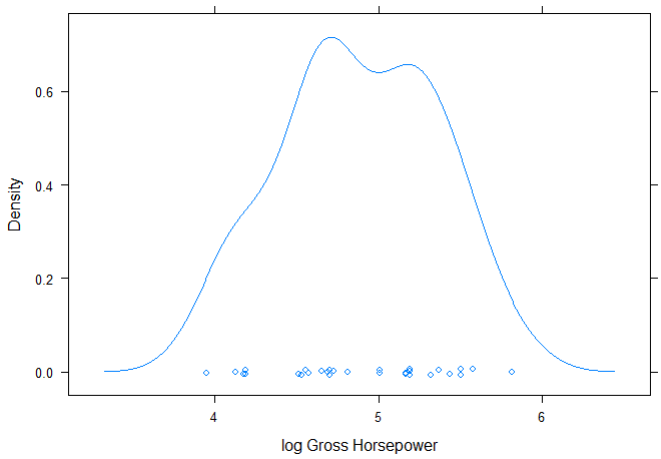

Figure 2: The densityplot of gross horsepower after doing the log transformation on the data values.

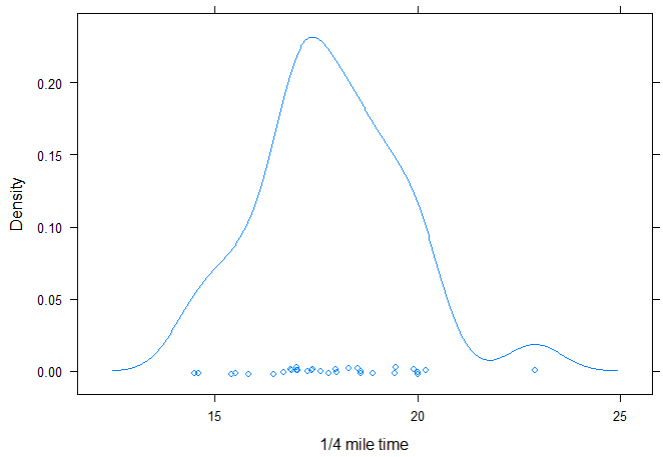

Figure 3: The densityplot of $\frac{1}{4}$ mile time with the original data values. Seems an outlier outside the range of 22.

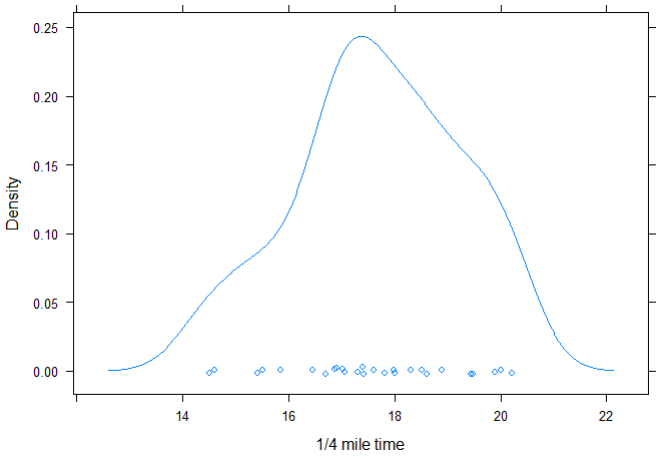

Figure 4: The densityplot of $\frac{1}{4}$ mile time after removing the ourlier.

To see how the response variable *mpg*, mile per gallon, is affected by other variables, we create a matrix scatterplot to show their linear relationships (see Figure 5). The matrix of scatterplots includes variables displayed on the diagonal and a variety of fitted lines. This graph shows that all other variables have a kind of correlation with the *mpg* response variable. However, we cannot decide to use all of them as the explanatory variable in our final fixed multiple linear regression model because some pairs of them have clear linear relationship to each other, such as *cyl*, number of cylinders and *disp*, displacement. Thus, we need to do more comparisons among our models to decide which variables are not necessarily needed for our fixed model. Then, we can select the best one with appropriate explanatory variables.

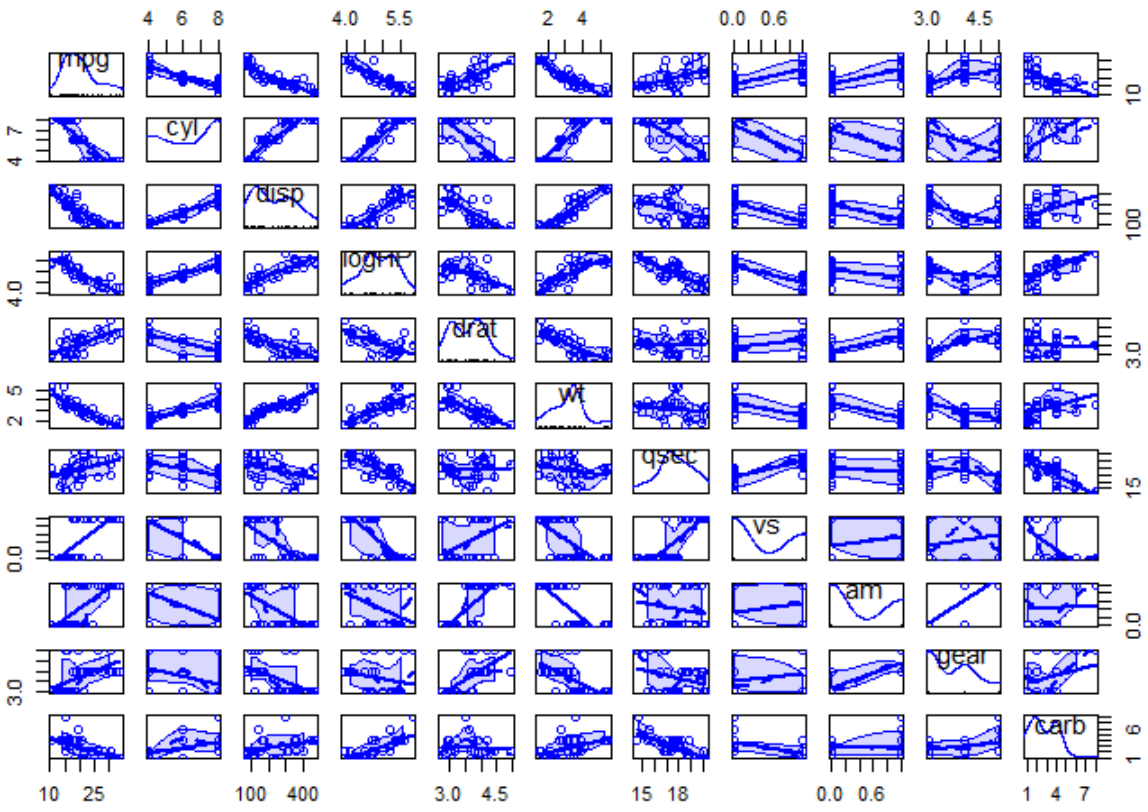

Figure 5: The scatterplot matrix displays the linearity between each two variables

2.2 Modeling

To test whether there is an explanatory variable should be dropped for the model, we firstly use all other ten variables beside *mpg* as the independent variables in the model. This completed multiple linear regression model provides a well-defined p-value of $2.651e-07$ and the R-Squared value is 0.8848, which means this model can explain 88.48% variation in *mile per gallon* values. However, it should be noticed that in the multiple linear regression model, if the R-Squared is too large and the fit is good, there may be multicollinearity, and multicollinearity is the main reason for the error of the multiple regression model. There may be some approximate linear relationships between some pairs of explanatory variables. We can test for multicollinearity with the value of variance inflation value (VIF), a measure of the amount of multicollinearity in regression analysis. Continuing testing the model until dropping all variables with a VIF value higher than 5. VIF exceeding 5 indicates high multicollinearity between an explanatory variable and the other.

After fixing the problem of multicollinearity, we still need to seek for the most appropriate reduced version of the model for an easier interpretation of the model. Using the analysis of variance and checking for the p-values to see if there exist the evidence to choose the reduced model. If the p-value is larger than 0.05, there will be more reason to go with the reduced model because it is not statistically significant to keep using the full model.

2.2.1 Fit and Assessment

Our final fixed model is

$$mpg = 20.705729 - 0.018749(disp) + 1.582855(drat) + 0.895256(vs) + 3.396180(am) - 1.338036(carb)$$

To test the fitting assumptions of this model, we can construct some diagnostic plots(see Figures 6, 7, and 8). The graph of residuals versus fitted values provides a zero-mean value with random spread of appeared points, the densityplot shows a well normal distribution, and the quantile graph displays a linear increasing line. These graphs mean the model is well fitted. The p-value of this model is $3.946e-09$, which means the multiple linear regression model is considered to be statistically significant for us. The F-statistic give 25.87 on 5 and 25 DF, which is a good value to show the a statistically significance of coefficients in the linear regression model. Also, the coefficient of determination (R^2 value) is 0.838, which means 83.8% variation in mile per gallon can be explained by the selected explanatory variables.

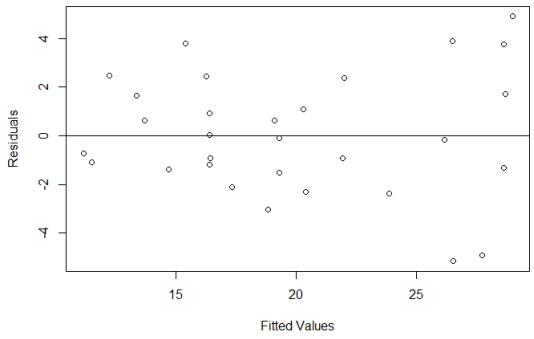

Figure 6: Graph of residuals versus fitted values of the fitted multiple linear regression model.

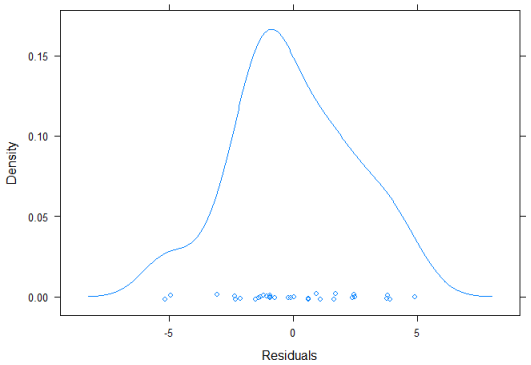

Figure 7: Densityplot of residuals of the fitted multiple linear regression model.

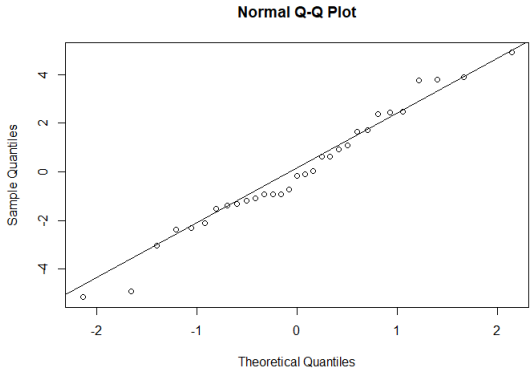

Figure 8: Normal quantile-quantile graph of the fitted multiple linear regression model.

2.2.2 Interpretation of Intervals

For motor trend cars with similar structures provided in the collected dataset, we predict the mean of mile per gallon to be 18.6722 mpg (95% CI: [15.72464, 21.61976], PI: [12.39307, 24.95132]) with given values of explanatory variables in the model (200 in^3 displacement, 3.9 rear axle ratio, straight engine, automatic transmission, 4 carburetors). The values of explanatory variables in the model are selected randomly for predicting the average value of miles a motor trend car can be driven with one gallon fuel consumption.

The 95% confidence interval gives a range of 15.72464 to 21.61976. This means we are 95% sure that the value of mean is between 15.72464 and 21.61976. If doing fitted model computations many times with sample values from the same population, we would like to expect 95% of these computations containing the true population mean in the confidence interval, where 15.72464 is the lower limit and 21.61976 is the upper limit. The prediction interval predicts the range where a single observation will fall in the future and it expresses the uncertainty in forecasting. The prediction interval must consider the uncertainty in estimating the mean value and the random variation in individual values. Thus, the prediction interval is always wider than the confidence interval, which gives the range of 12.39307 to 24.95132 under the given case.

3 Conclusion

3.1 Inference for Regression

We use analysis of variance to select the most appropriate explanatory variables for fitting the multiple linear regression model of mile per gallon. We then conduct the t-test on the null hypothesis that *mpg* does not vary statistically significant with *disp*, *drat*, *vs*, *am*, and *carb*. The p-value less than 0.05 tells use the null hypothesis should be rejected and we need to accept the alternative hypothesis. Also, a high R^2 value expresses that a high variation in *mpg* can be explained by the given model. We can explain that the *mpg* model will vary with a constant intercept value, decreasing *disp* and *carb*, and increasing *drat*, *vs*, am, approaching the sample mean.

3.2 Strength and Weakness

Based on our analysis, the fitted model can do a good job on predicting how many miles a motor trend car in 1974 can be driven with one gallon fuel consumption. However, we still need to consider that our model is not perfect for predicting a mean value for a broader population or within samples of other kinds of vehicles. The data collection is randomly sampled and only 32 different automobiles are selected. The observations are less and the data year is old, in 1973-74. The structure of cars in modern years will be much more different from the old models, so the variables will be less trusted. Thus, our results cannot be extended to a broader population or used in modern year car samples. The causal relationship should not be inferred and we need more data and information to see the correlation between mile per gallon and other explanatory variables.

4 Bibliography

Data description: mtcars datasets
Version: R-4.2.1
Packages: lattice, car