MATH 5345 / Regression Analysis: Final Report

Dr. Sun

Gabriela Lara and Joshua Mitchell

Contents

List of Tables

List of Figures

Abstract

A multiple linear regression was calculated to predict the gas mileage of a car (its MPG) based on the number of cylinders, displacement, horsepower, weight, acceleration, model year, and origin of the car. The initial model had problems with non-constant variance and multicollinearity. A transformation on the response variable and an additional interaction regressor was added to improve the homoscedasticity and the explanatory power of the model, resulting in an R^2 value of 0.89. Only a subset of the seven regressors and an interaction term turned out to be significant.

Introduction

The Auto MPG data set, from the UC Irvine Machine Learning Repository, contains 397 samples of various models of vehicles. The goal of the analysis is to learn to predict a vehicle's gas mileage (mpg) given other attributes of the vehicle (e.g. its weight, model year, origin, etc).

The data contains a total of nine attributes. Three are discrete (origin, model year, and number of cylinders), one is alphanumeric (model name), and the rest are continuous. The model name column was thrown out, and model year and number of cylinders were interpreted as continuous variables.

The data also had some missing values (6), but they constituted a negligible amount of the data, so they were removed.

Discussion of Models and Analysis Results

Original Full Model Analysis

Original Full Model:
mpg (c) ~ wgt (c) + modelyr (mvd) + origin (mvd) + hp (c) + displ (c) + cylnum (mvd) + acc (c)

	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	-18.3106	4.6933	-3.90	0.0001	***
wgt_c	-0.0067	0.0007	-10.23	0.0000	***
modelyr_mvd	0.7805	0.0519	15.03	0.0000	***
origin_mvd2	2.6340	0.5665	4.65	0.0000	***
origin_mvd3	2.8557	0.5528	5.17	0.0000	***
hp_c	-0.0174	0.0137	-1.27	0.2056	
displ_c	0.0241	0.0077	3.14	0.0018	**
cylnum_mvd	-0.5123	0.3222	-1.59	0.1126	
acc_c	0.0845	0.0984	0.86	0.3913	

Table 1: R Summary of original full model (relating mpg to the rest)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Significance
wgt_c	1	16470.05	16470.05	1505.92	0.0000	***
modelyr_mvd	1	2756.94	2756.94	252.08	0.0000	***
origin_mvd	2	261.23	130.61	11.94	0.0000	***
hp_c	1	8.96	8.96	0.82	0.3659	
displ_c	1	77.03	77.03	7.04	0.0083	**
cylnum_mvd	1	29.10	29.10	2.66	0.1037	
acc_c	1	8.06	8.06	0.74	0.3913	
Residuals	382	4177.89	10.94			

Table 2: R ANOVA of original full model (relating mpg to the rest)

	GVIF	Df	GVIF^(1/(2*Df))
wgt_c	11.07	1.00	3.33
modelyr_mvd	1.30	1.00	1.14
origin_mvd	2.09	2.00	1.20
hp_c	9.98	1.00	3.16
displ_c	22.87	1.00	4.78
cylnum_mvd	10.74	1.00	3.28
acc_c	2.62	1.00	1.62

Table 3: VIF of each regressor in the Full Original Model

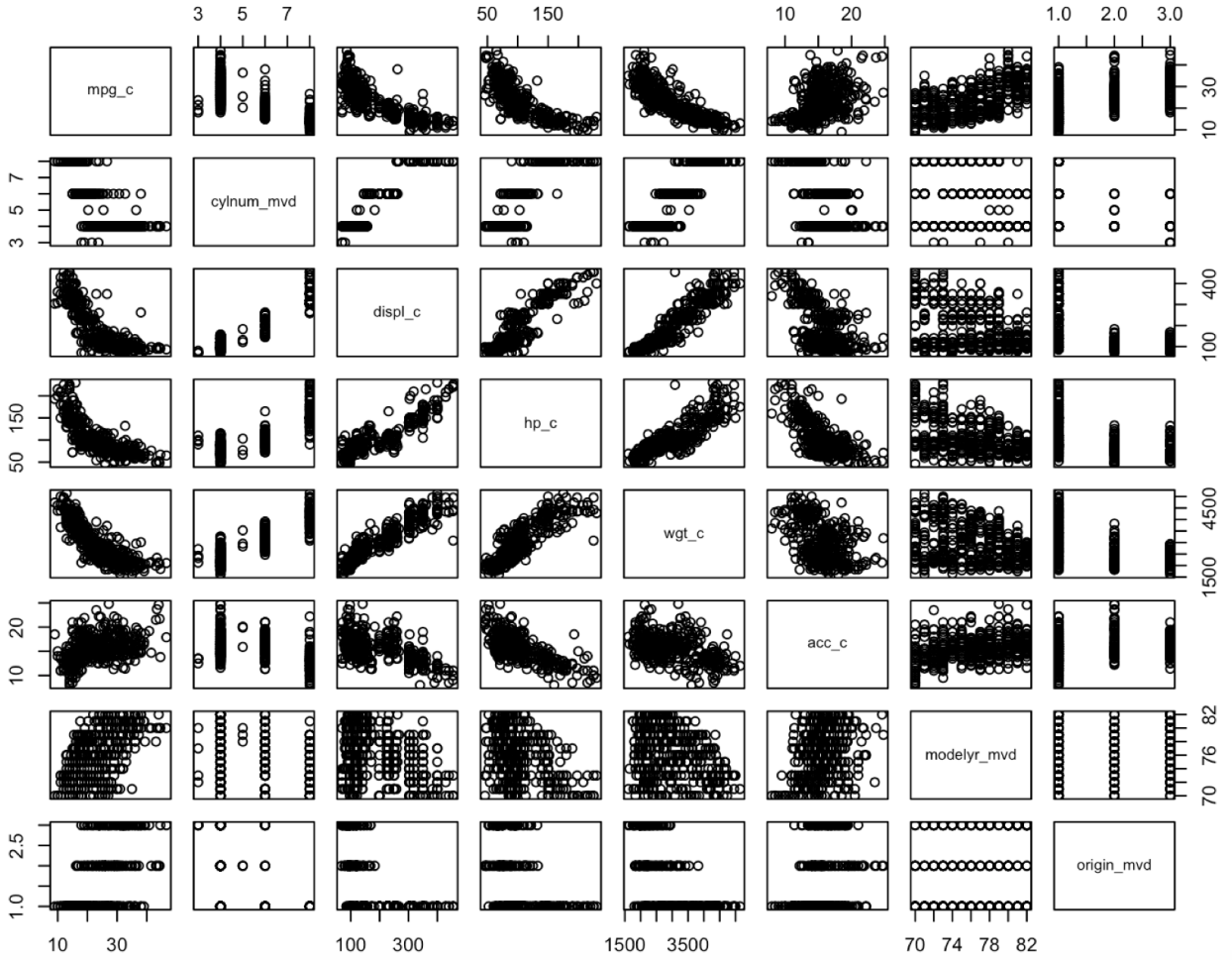

Figure 1: A scatterplot matrix between all the variables in the automobile data set

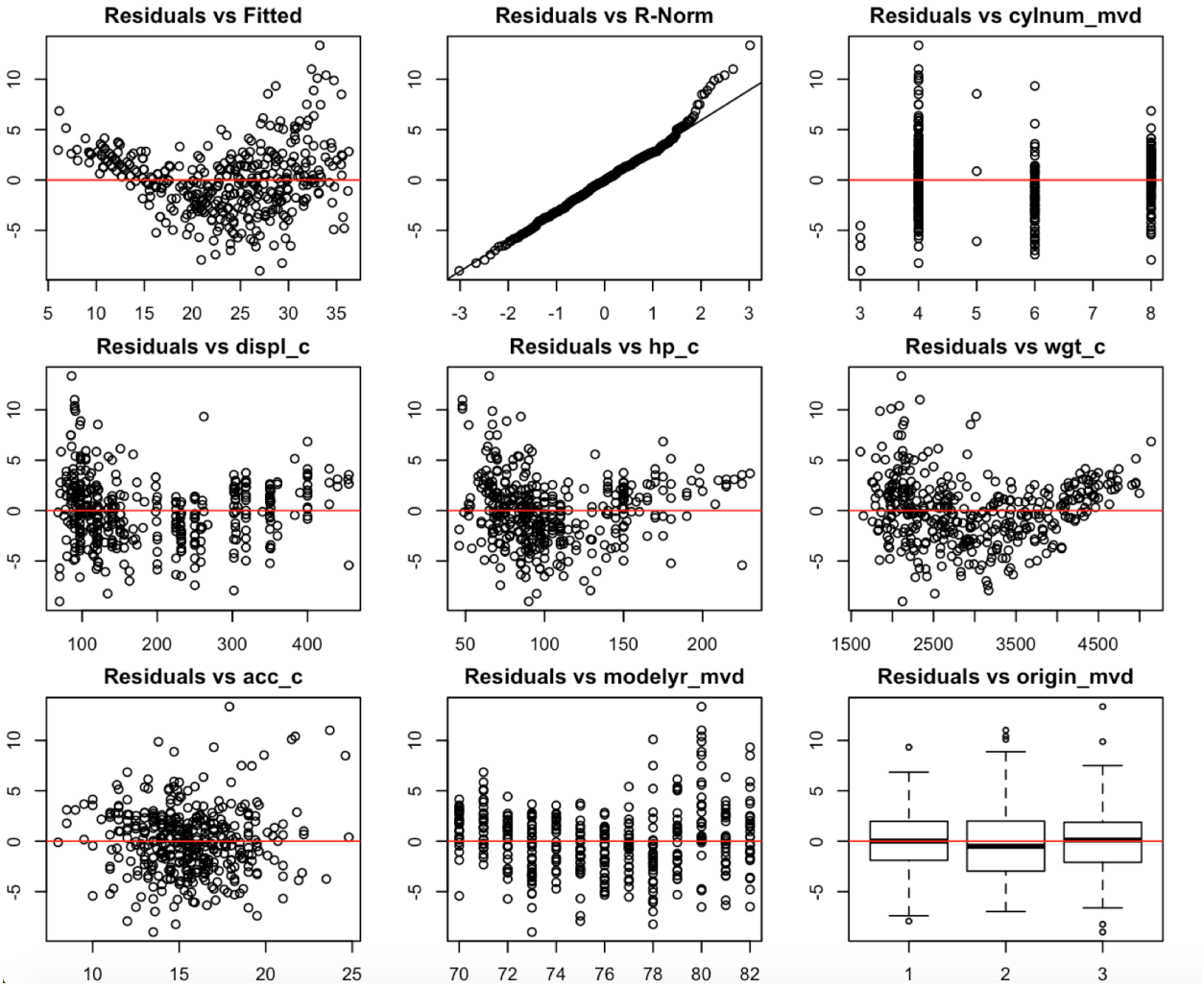

Figure 2: A Residual vs Fitted, a Residual vs R-Norm, and Residual vs Regressors plots of the Original Full Model

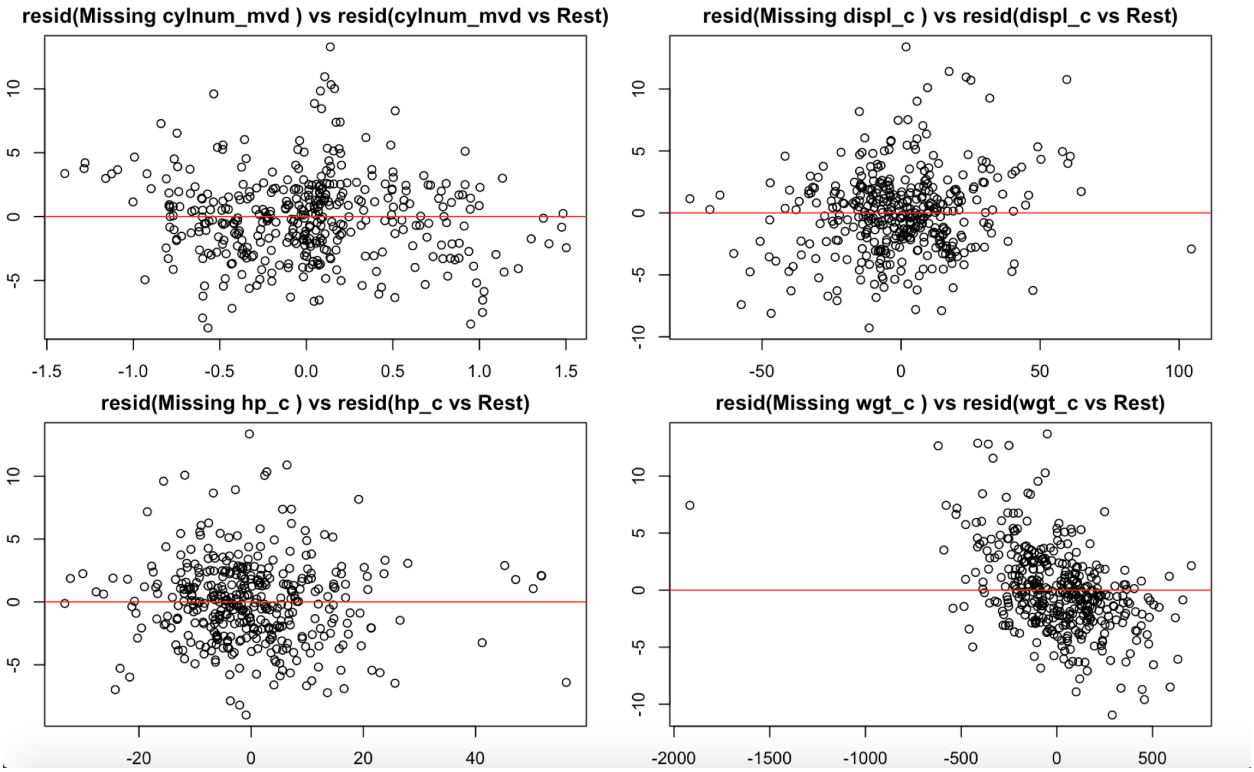

Figure 3: Partial regression plots on each of the regressors with high VIF scores

	Regressor	F_Statistic	P_Value	Significance
1	Displacement	9.89	0.00	**
2	Weight	104.63	0.00	***
3	HP	1.61	0.21	none
4	Cylinder Num	2.53	0.11	none

Table 4: A partial F test on each of the regressors with high VIF scores

Original Model

A multiple linear regression was fitted on miles per gallon number of cylinders (as a factor), displacement, horsepower, weight, acceleration, model year, and origen (as a factor); and the five requirements to realize a regression analysis were checked.

Normality:

Notice that the Residual vs. Normal plot behaves mostly normal except at the tails. Where the tails seem lighter than the textbook ideal. Yet, this will not bring major issues to the analysis. Therefore, normality can be concluded for this purpose.

Independence, Linearity and Constant Variance:

The Residual vs. Fitted Values plot for this regression, makes a shoe horse shape plot and, most of the data points are below the abline. It indicates lack of constant variance, non-zero mean error, and a non-linear distribution. This will bring major problems to the predictions, and analysis. The issue can be addressed in a simple manner; and will be discussed in the next section. Also, notice how all of the residuals are bounded, this is an indication of uncorrelated errors.

Figure 2 and Table 3 gives an insight for correlation among variables. Since a lot of the plots show a linear relationship (with a nonzero slope), multicollinearity is suspected. Further evidence that there exists a dependence among variables are the VIF values of cylinder number, displacement, horsepower and weight; hence it will be needed to include some interaction variables.

Transformed Full Model Analysis

Transformed Full Model:
log(mpg) (c) ~ wgt (c) + modelyr (mvd) + origin (mvd) + hp (c) + displ (c) + cylnum (mvd) + acc (c)

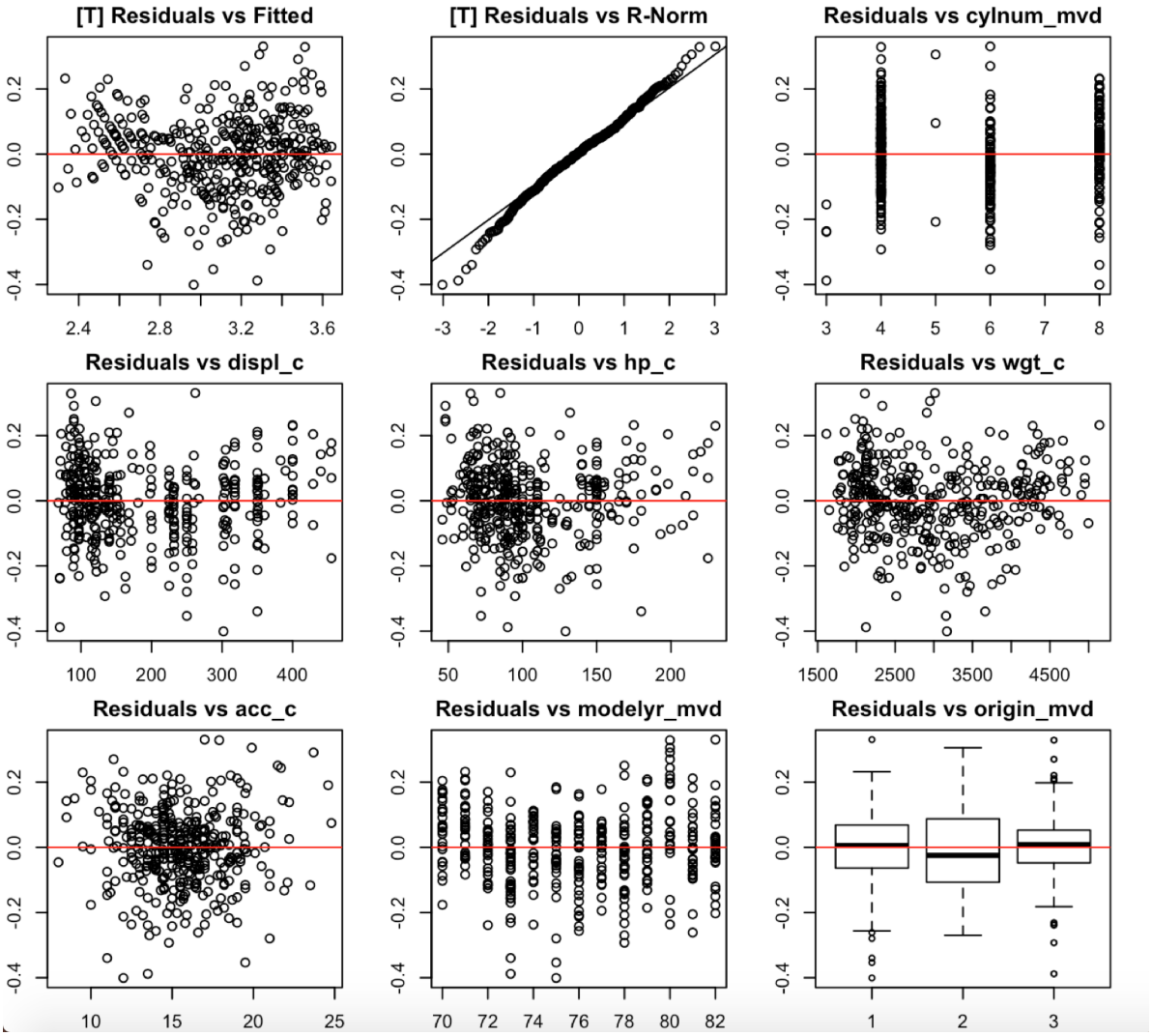

Figure 4: A Residual vs Fitted, a Residual vs R-Norm, and Residual vs Regressors plots of the Transformed Full Model

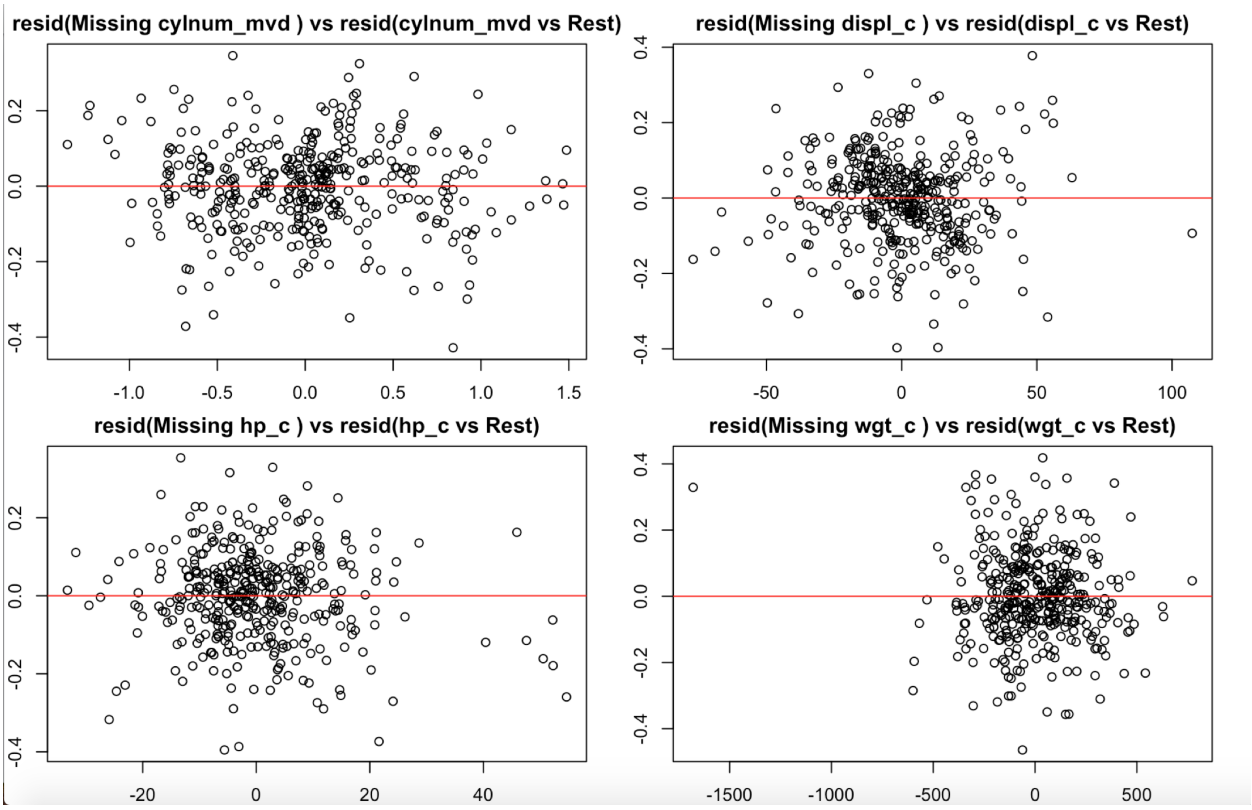

Figure 5: Partial regression plots on each of the regressors with high VIF scores on the Transformed Full Model

	Regressor	F_Statistic	P_Value	Significance
1	Displacement	8.39	0.00	**
2	Weight	126.68	0.00	***
3	HP	9.10	0.00	**
4	Cylinder Num	6.35	0.01	*

Table 5: A partial F test on each of the regressors with high VIF scores on the Transformed Full Model

Several issues, not very major, have been found in the original model. In order to address them a transformation has to be applied into the fit: Since the Residuals vs. Fitted plot showed a horseshoe pattern, based on previous encounters, a log transformation in the predictor was selected:
Insert transformed regression here
Which fixed most of the problems, with regard of the assumptions:
Figure 4, Residuals vs. Fitted shows that the new model is linear, has mean zero error, and independent .
Notice, in Figure 4 Residuals vs. R-Norm, that a new problem has arised in the tails have become lighter, compared to the original. This could be caused by interaction among variables and/or influential points. It is worth to remark that in this model, so far cylinder number, horsepower, and weight do not contribute much towards the model: based on the partial residual plots (Figure 5), there is no apparent pattern on said variables. But, the partial F-Test indicates otherwise, which is a stronger sign of interaction.
The remainder of the plots, show a constant band of residuals, which is a sign of constant variance. Thus, based on all of the previews evidence there must be variables interacting with each other, and should be included in the model.

Interaction Terms Analysis

	Model	R_Sq	AR_Sq	MS_res
1	Interaction	0.88	0.87	7.90
2	Transformed + Interaction	0.90	0.90	0.01

Table 6: A chart comparing the Untransformed Full Model with all combinations of interaction terms for the high VIF regressors against the same model with a log transformation on the response variable (mpg)

It has been discovered during the analysis that the model contains variable interaction. This might have caused the violations towards the assumptions. Therefore, a comparison among a regression with interaction and a transformed transformation with interaction has to be made.
Note, that the transformed regression with interaction has the highest R^2 (0.9), and the lowest mean square errors (0.01). This model may be the most appropriate one for the data. Yet, all of the insignificant variables have to be vanish, and all of the assumptions have to be revised again.

Model Selection Analysis

	Selection_Method	Num_Regressors	R_Sq	Adj_R_Sq	MS_res
1	Forward	6.00	0.89	0.89	0.01
2	Backward	16.00	0.90	0.90	0.01
3	Stepwise	6.00	0.89	0.89	0.01

Table 7: Statistics about the models outputted from Forward, Backward, and Stepwise Selection algorithms in R (note that the model selected by Forward and Stepwise selection is identical, so just the Forward model will be considered in further sections)

	GVIF	Df	GVIF^(1/(2*Df))
wgt_c	13.83	1.00	3.72
modelyr_mvd	1.27	1.00	1.13
origin_mvd	1.74	2.00	1.15
hp_c	37.47	1.00	6.12
acc_c	2.61	1.00	1.62
wgt_c:hp_c	58.06	1.00	7.62

Table 8: VIF of each regressor in the Forward Model

	GVIF	Df	GVIF^(1/(2*Df))
wgt_c	3110.02	1.00	55.77
modelyr_mvd	1.44	1.00	1.20
origin_mvd	3.01	2.00	1.32
hp_c	2806.06	1.00	52.97
displ_c	10568.29	1.00	102.80
cylnum_mvd	975.92	1.00	31.24
acc_c	3.64	1.00	1.91
wgt_c:hp_c	27058.36	1.00	164.49
hp_c:displ_c	39680.75	1.00	199.20
wgt_c:displ_c	11724.72	1.00	108.28
wgt_c:cylnum_mvd	9069.37	1.00	95.23
hp_c:cylnum_mvd	9125.25	1.00	95.53
displ_c:cylnum_mvd	19861.83	1.00	140.93
wgt_c:hp_c:cylnum_mvd	41867.56	1.00	204.62
wgt_c:displ_c:cylnum_mvd	15077.88	1.00	122.79
hp_c:displ_c:cylnum_mvd	44842.29	1.00	211.76

Table 9: VIF of each regressor in the Backward Model

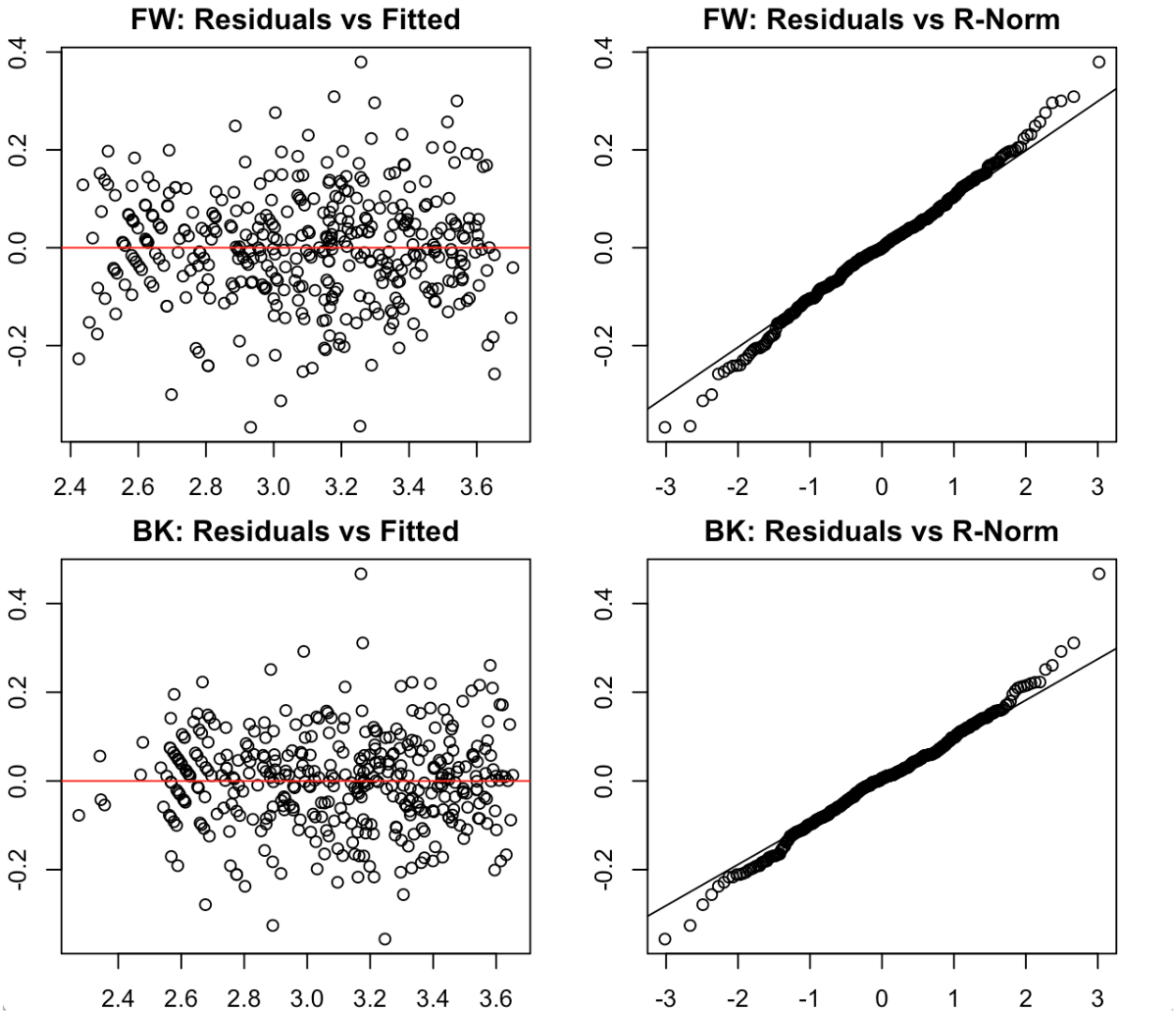

Figure 6: Residuals vs Fitted and vs Random Normal plots for Forward, Backward, and Stepwise Models

In order to select the most significant variables, several tests were applied: Forward, Backwards and Stepwise (Table 7).
Although, the Backward Test has the largest R^2 and R^2_{adj}, Forward and Stepwise have the least number of regressors, in fact the exact same regressors were selected. Another reason to select the Forward model over the Backward, are the VIF values. On the Forward all of the VIF values are relatively small and consistent compared to the Backwards; where we have values that reached the 45,000.
Note that all of the assumptions hold for both (Figure 6); maybe with exception that for both variable selections, the Residual vs R-Norm plot still has light tails (Figure 6); this might be an issue of influential points.

Influential Points Analysis

	Model	Num_Infl_Pnts	Percent_Infl_Pnts	Common_Infl_Pnts
1	Forward	20.00	5.12%	14.00
2	Backward	36.00	9.21%	14.00

Table 10: Influential point comparison of Forward Model vs Backward Model

	Model	R_Sq	AR_Sq	MS_res
1	Forward w/o Infl	0.91	0.91	0.01
2	Backward w/o Infl	0.90	0.90	0.01

Table 11: Forward Model with no influential points vs Backward Model with no influential points

It may be beneficial to our model to identify influential points, without deleting a good portion of the data. Thru Backward variable selection a 9.21% of the data will have to be vanish; whereas thru Forward variable selection there will only be a 5.12

Final Model Choice

FINAL MODEL: log(mpg) (c) ~ modelyr (mvd) + origin (mvd) + hp (c) + acc (c) + wgt (c) * hp (c)

	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	2.1373	0.1735	12.32	0.00001	***
wgt_c	-0.0004	0.0000	-14.76	0.00001	***
modelyr_mvd	0.0309	0.0018	17.59	0.00001	***
origin_mvd2	0.0558	0.0177	3.14	0.00180	**
origin_mvd3	0.0455	0.0180	2.52	0.01210	*
hp_c	-0.0064	0.0009	-7.06	0.00001	***
acc_c	-0.0053	0.0034	-1.59	0.11180	
wgt_c:hp_c	0.0000013	0.0000002	6.71	0.00001	***

Table 12: R Summary of the final model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Significance
wgt_c	1	34.62	34.62	2714.64	0.0000	***
modelyr_mvd	1	4.72	4.72	369.94	0.0000	***
origin_mvd	2	0.25	0.12	9.78	0.0001	***
hp_c	1	0.11	0.11	8.87	0.0031	**
acc_c	1	0.001	0.00	0.26	0.6078	
wgt_c:hp_c	1	0.57	0.57	45.04	0.0000	***
Residuals	383	4.88	0.01			

Table 13: R ANOVA of the final model

Up to this point, the model suggested by the Forward variable selection, has had the lowest VIF values, the highest R^2 and R^2_{adj} (Table 11), and requires to delete a smallest portion of data, due to influential points. Hence, it will be recommended to proceed with the model the Forward test selected:

```
lm(log(mpg_c) \tl wgt_c + modelyr_mvd + origin_mvd + hp_c + acc_c + wgt_c:hp_c, data = autodata)
```

The summary indicates high significance of the variables selected to the model (Table 12)

Conclusion

References

https://archive.ics.uci.edu/ml/datasets/auto+mpg