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

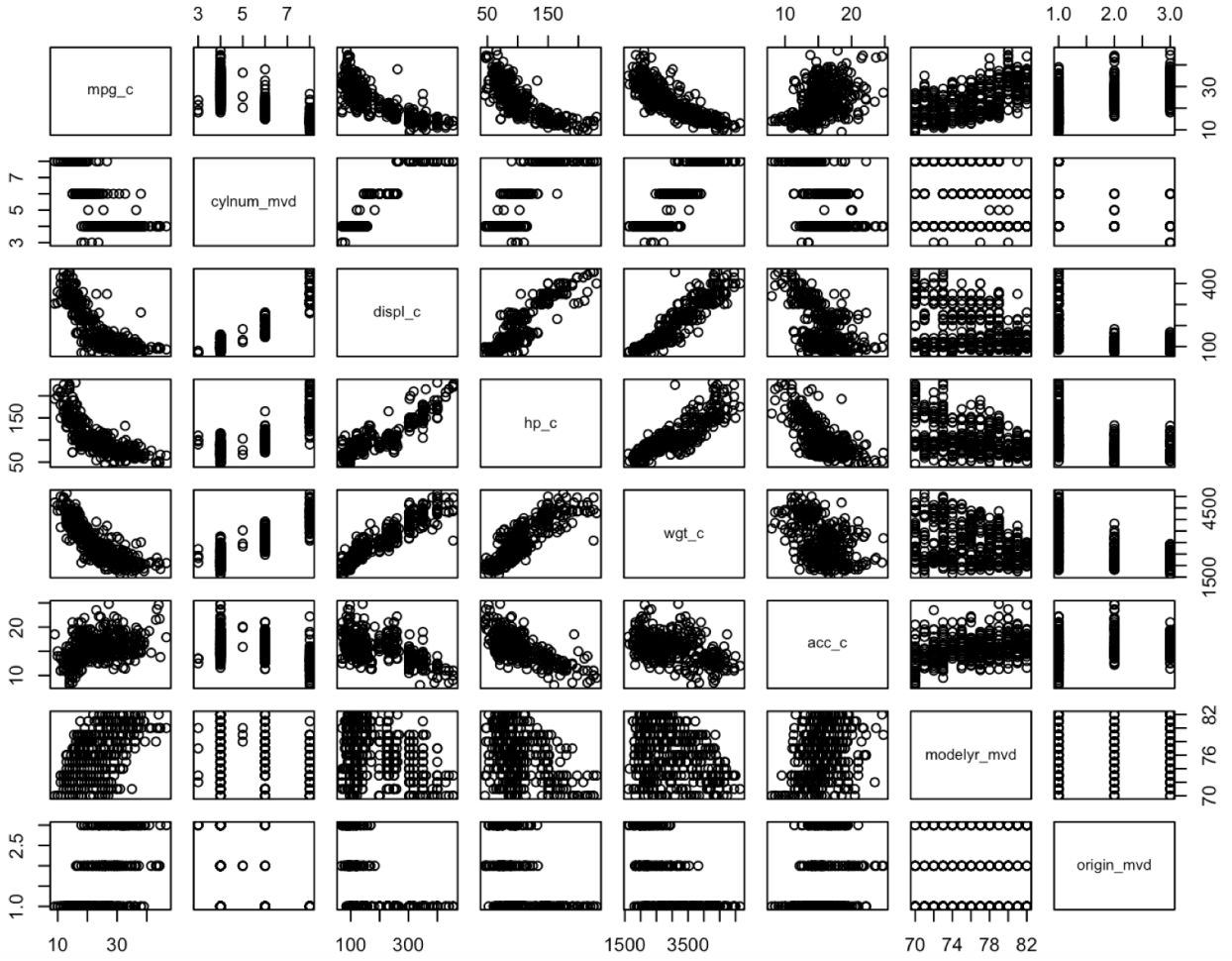

Figure 1: A scatterplot matrix between all the variables in the automobile data set

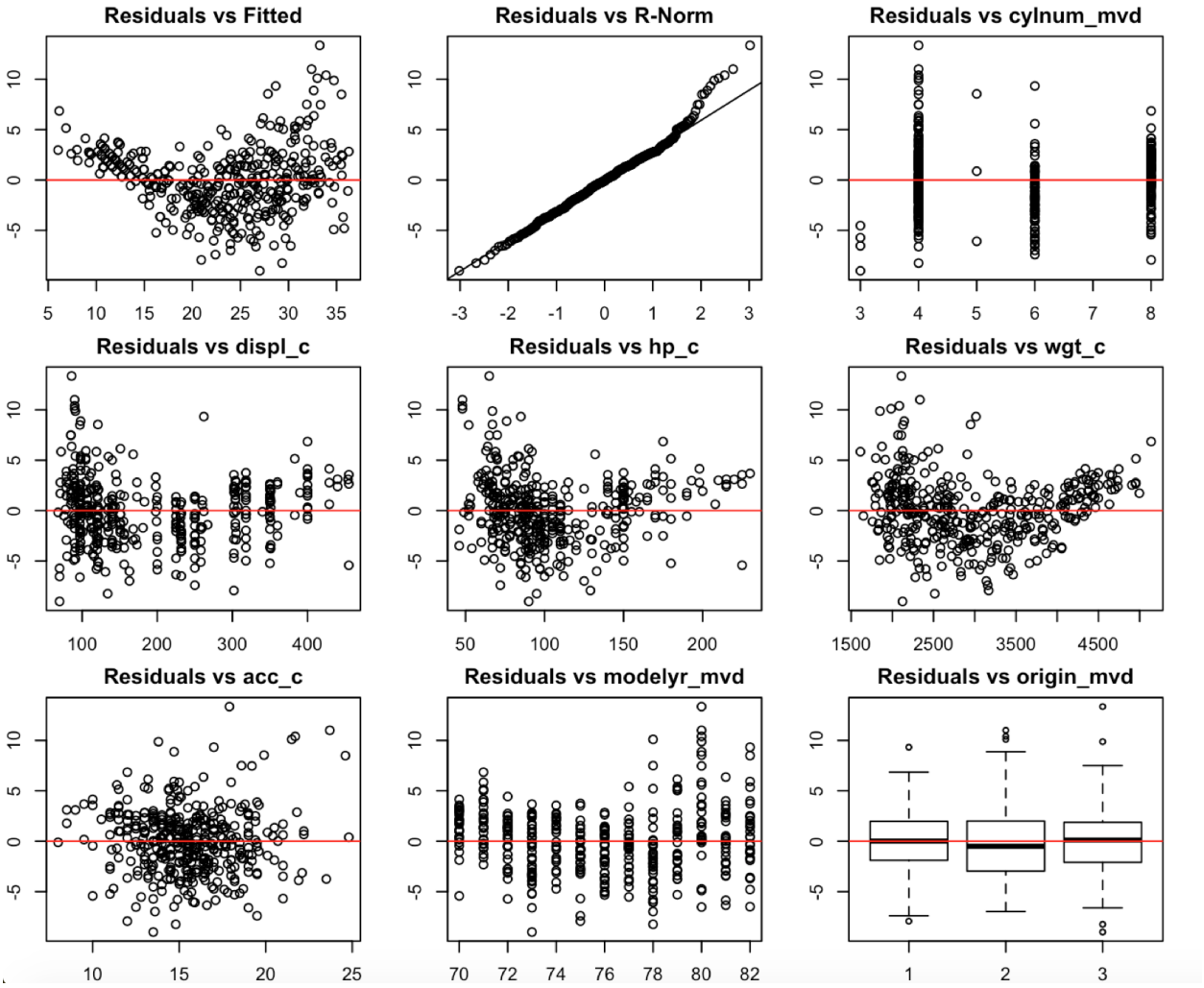

Figure 2: A Residual vs Fitted, a Residual vs R-Norm, and Residual vs Regressors plots of the Original Full Model

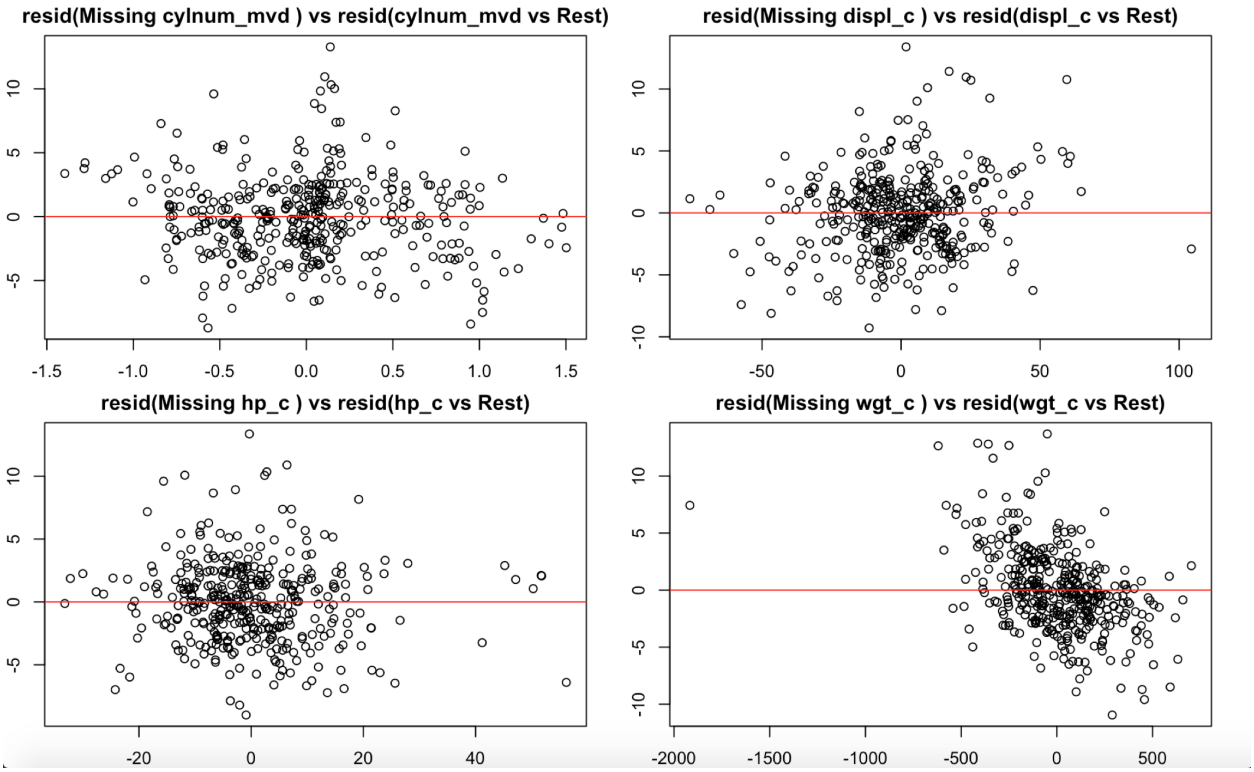

Figure 3: Partial regression plots on each of the regressors with high VIF scores

	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

A full multiple linear regression was fitted on the data set with mileage (mpg) as the response variable and number of cylinders (cylnum), displacement (displ), horsepower (hp), weight (wgt), acceleration (acc), model year (modelyr), and origin (origin) as the predictor variables. The following five key assumptions were examined:

a. Normality of errors: Based on figure 2, it appears that the residuals are mostly normally distributed with a light tail.

b. Independence of errors: Based on figure 2, our errors might be nontrivially dependent due to the quadratic shape of the Residuals vs Fitted plot.

c. Constant variance of errors: Based on figure 2's Residuals vs Fitted plot, it appears that, due to the inconsistency of the spread of the residuals, constant variance is not upheld.

d. An expected error value of 0: Based on figure 2's Residuals vs Fitted plot, it is likely that this is upheld (due to half the data being roughly above 0 and the other half being below).

e. Linearity: Based on the significance of multiple coefficient estimates for each regressor in Table 1, it appears that our response variable is linearly related to our predictor variables.

Correlation among variables was also examined. Figure 1 and Table 3 indicate a strong relationship between hp, wgt, displ, and cylnum. Further inspection was done with partial regression plots and a partial f test for each of said variables. At this point, due to the significance of some variables and lack of significance of others from the partial f tests, it was concluded that a log transformation should be done on the response variable (mpg), due to the quadratic pattern in the Residual vs Fitted plot, and possibly an addition of an interaction term.

Transformed Full Model Analysis

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

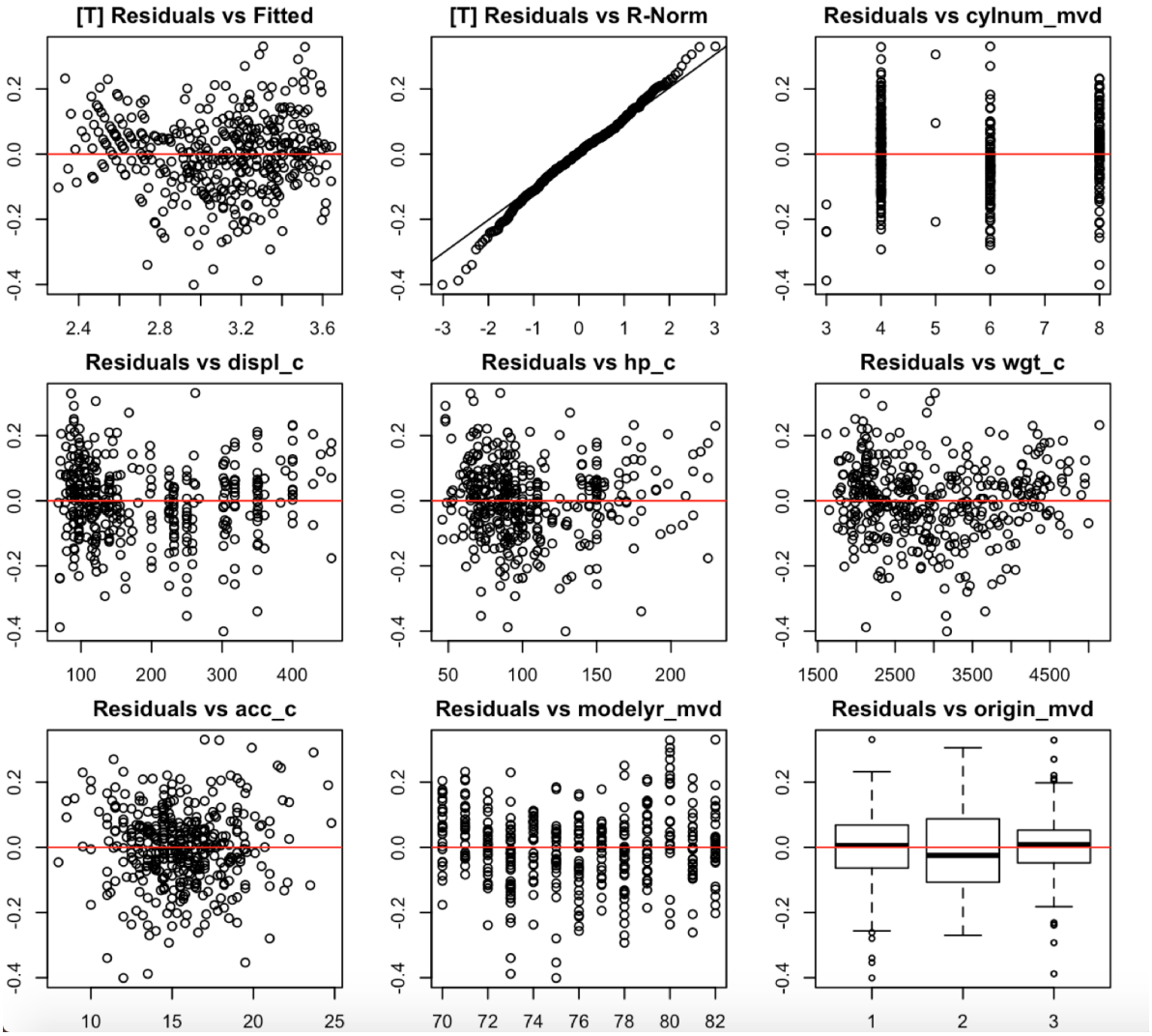

Figure 4: A Residual vs Fitted, a Residual vs R-Norm, and Residual vs Regressors plots of the Transformed Full Model

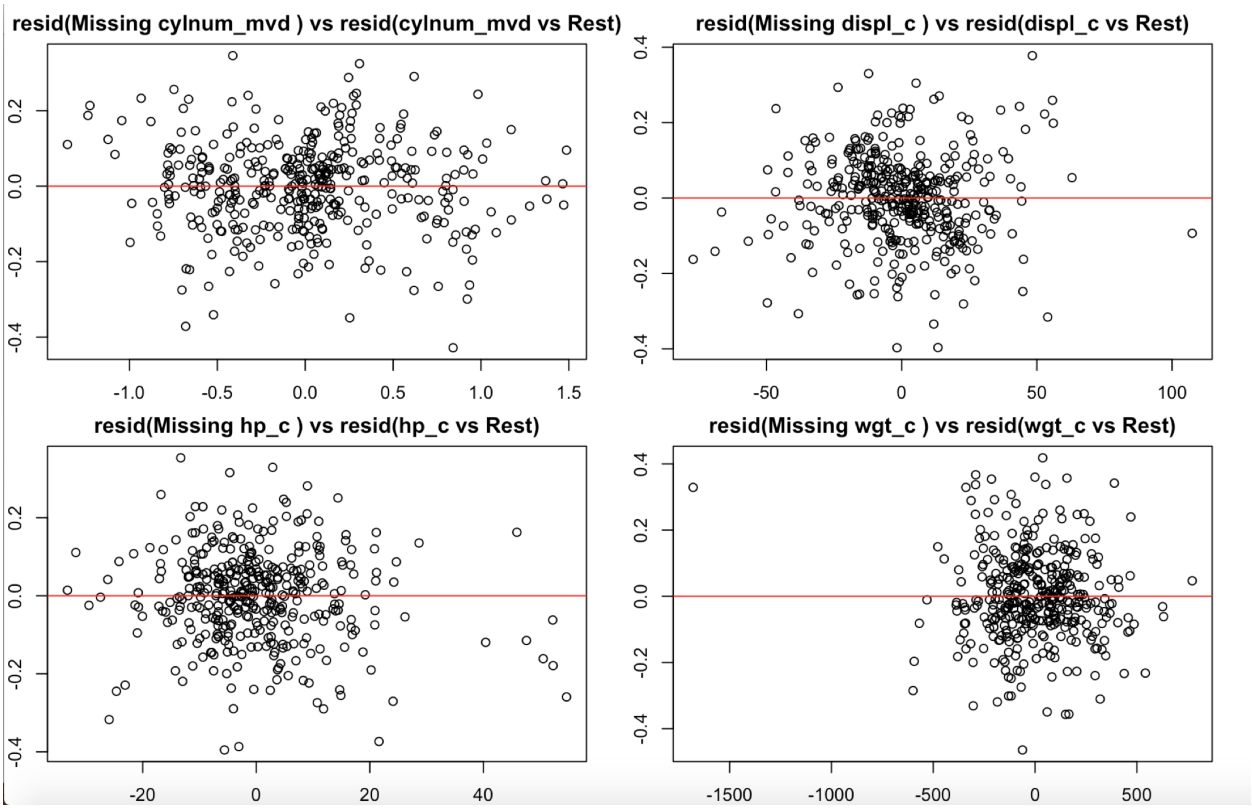

Figure 5: Partial regression plots on each of the regressors with high VIF scores on the Transformed Full Model

	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

A new model was created with a log transformation on the response variable and then examined. Fortunately, the transformation dampened the quadratic pattern (increasing the independence of the errors) and stabilized the variance of the residuals.

However, the partial regression plots and partial f tests were redone with the same 4 regressors with mixed results. The transformation made the insignificant regressors (hp and cylnum) more significant (**, and *, respectively), however, the partial regression plots showed that the explanatory power of these two regressors was not significant.

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 was suspected, due to the somewhat significant difference between the partial f tests and partial regression plots of the transformed full model, that there was possible multicollinearity, so it was decided that an interaction term should be added. Due to the number of possible interaction terms, it was decided that all regressors plus all possible combinations of the regressors with high VIFs should be modeled and then compared to the original full model. The addition of these interaction terms resulted in a significant increase in R^2. A plain full model with these interaction terms was also compared to a transformed full model with the same interaction terms, and it was concluded that, given that interaction terms were going to be included, the transformed model was better based on its R^2. In the next section, other model metrics will be considered.

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

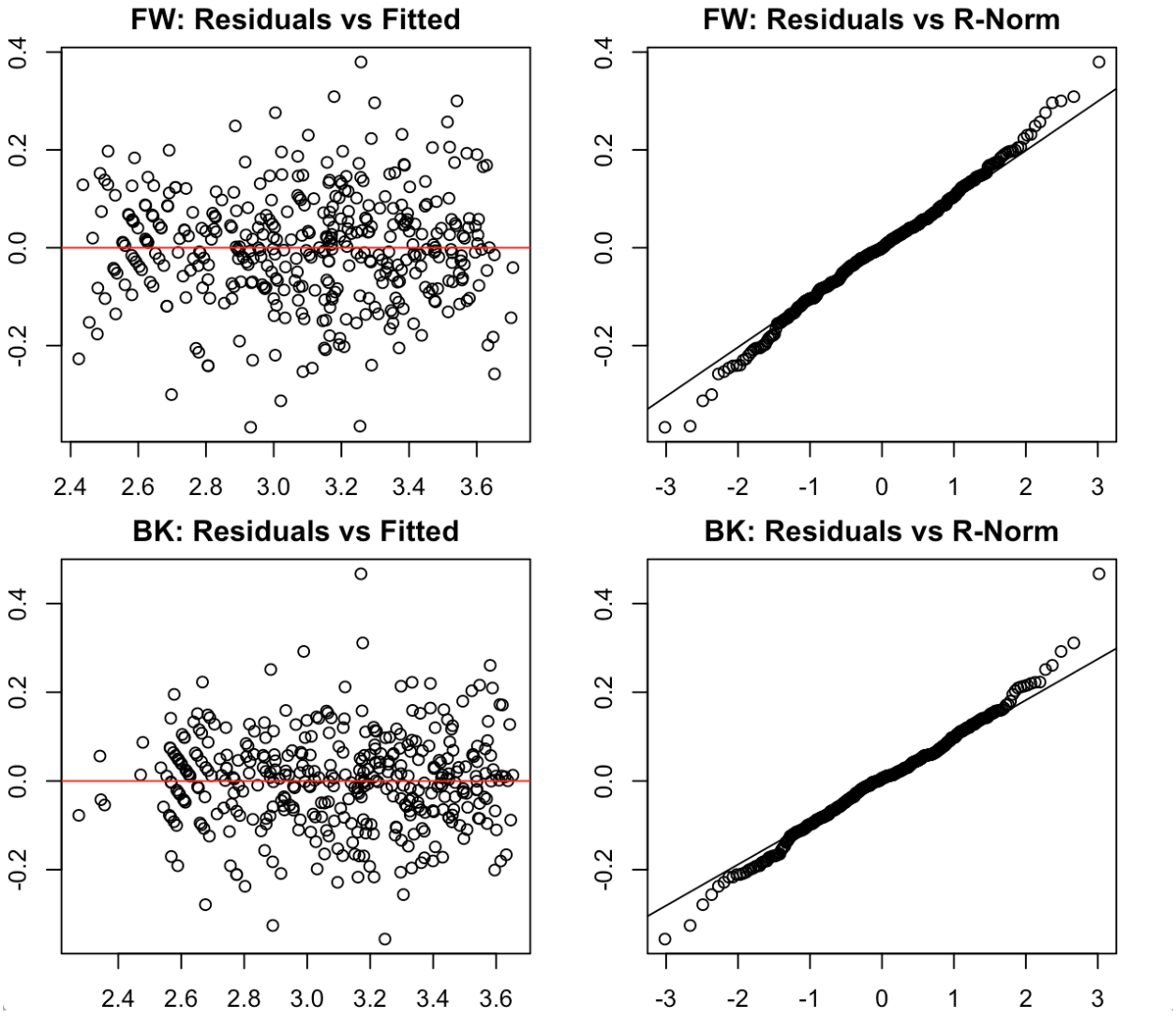

Figure 6: Residuals vs Fitted and vs Random Normal plots for Forward, Backward, and Stepwise Models

In order to select the most significant variables, several tests were applied: Forward, Backwards and Stepwise (Table 7).

The Forward and Stepwise selection algorithms both selected a model with 6 variables (5 original variables and 1 interaction term). The Backwards selection algorithm selected a model with 16 variables (mostly interaction terms).

The Backward selection model has a slightly higher R^2, but with slightly less degrees of freedom (due to the increased parameters) and possibly more complexity.

The VIF values of each regressor of each model were also examined. As expected, the VIF for the known correlated variables was still high, but, more importantly, the other variables showed little correlation.

The 5 key assumptions for each model were also checked (see Figure 6) and no major problems were observed. Outliers and influential points are examined in the next section.

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 was observed that the Backward model had almost twice as many points deemed influential as the Forward model, indicating that the Forward model might be either more resilient to influential points or a better model of the data in general, since both models have a roughly equal R^2 and $MS_{residuals}$.

Final Model Choice

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

Many model combinations were considered, but the simplicity, high R^2, low $MS_{residuals}$, and resilience to influential points of the Forward model made it the number one choice. Tables 12 and 13 indicate the R Summary and R ANOVA of the chosen model.

Conclusion

A model that mostly satisfies the 5 key assumptions of multiple linear regression and also explains 89% of the variance in the response variable (mpg) was found, which, by most metrics is a good job. It does appear that there is a significant linear relationship between the gas mileage of a car and other attributes like its weight and model year.

The initial model appeared to violate 2 of the 5 key assumptions (independent errors and constant variance of residuals) and seemed to have some multicollinearity problems among regressors. A log transformation on the response variable eased the 2 key assumption violations (as well as increased the predictive ability of the model), and the addition of an interaction term compensated for the correlation of two of the variables (as well as increased the predictive ability of the model. Since the models generated by the selection algorithms had similar predictive power, the one that had the least influential points was chosen.

References

The data was acquired from UC Irvine's Machine Learning Repository at the following web address: https://archive.ics.uci.edu/ml/datasets/auto+mpg