Regression Model to Predict Viscosity of a Blend from its Composition

Mohammed Quazi

PhD Candidate – Statistics
Department of Mathematics & Statistics
University of New Mexico
mquazi@unm.edu
https://math.unm.edu/~mquazi/
https://github.com/mquazi

Overview

1. Introduction
   - Viscosity Data Exercise
2. Data Exploration
   - Preliminary Data Analysis
3. Model Selection
4. MLR – Model Assumptions
5. Final Model
   - Conclusion

Viscosity Data Exercise

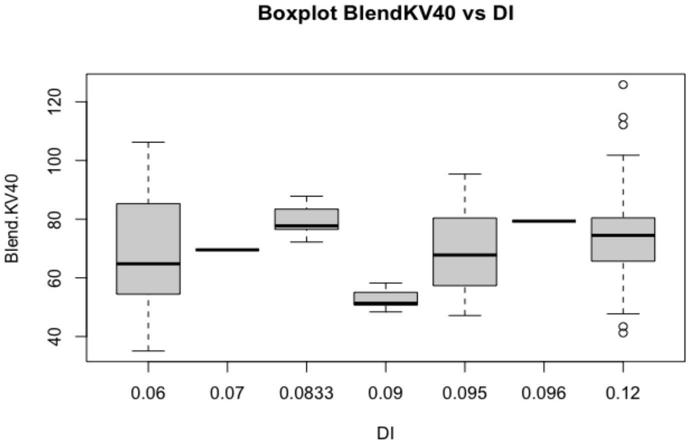
- Files to execute this study are here: https://github.com/mquazi/viscproblem
- Main question: Build a predictive model to predict Blend KV40
- Viscosity of a blend is the response variable – **Blend KV40**
- Predictor variables considered are 6 – performance package **(DI)**, viscosity modifier **(VM)**, base stock density **(BS Density)**, base stock KV40 **(BS KV40)**, base stock KV100 **(BS KV100)**, base stock total **(BS total)**
- If the prediction model is not accurate enough, need to include the individual base stocks

Preliminary Data Analysis

- No NAs or missing data points
- 86 rows and 20 columns
- Correlation between Blend KV40 and BS KV100 is 0.708 (good)
- Correlation between Blend KV40 and BS KV40 is 0.718 (good)
- Correlation between BS KV40 and BS KV100 is 0.9888 (bad)

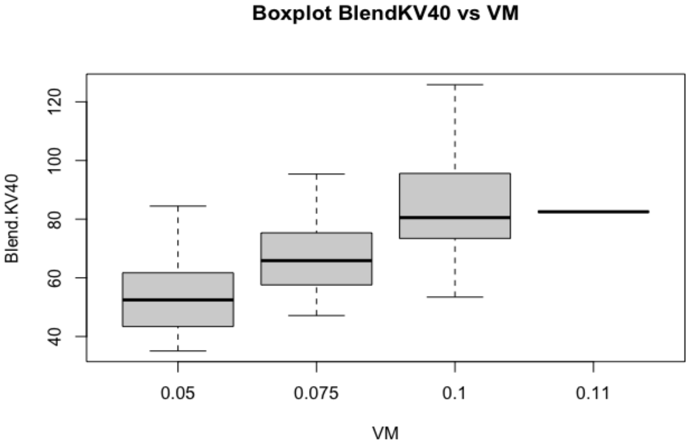
Boxplot DI

- No real takeaways, DI median Blend KV40 levels do not really differ

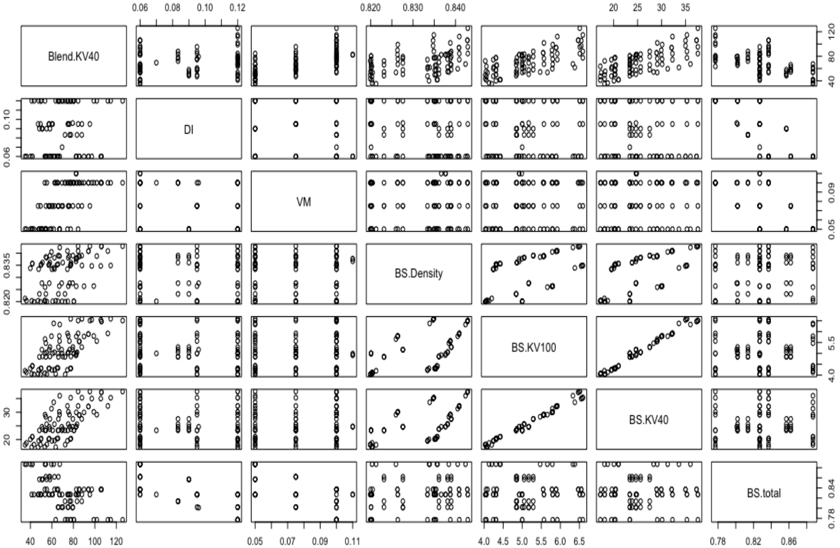


Boxplot VM

- Upward trend, 0.11's Blend KV40 median to lookout for



Pairplot


Blend.KV40
DI
VM
BS.Density
BS.KV100
BS.KV40
BS.total
0.06 0.08 0.10 0.12
0.820 0.830 0.840
20 25 30 35
120 80 40
0.10 0.06
0.09 0.05
0.835 0.820
5.5 4.0
30 20
0.84 0.78
40 60 80 100 120
0.05 0.07 0.09 0.11
4.0 4.5 5.0 5.5 6.0 6.5
0.78 0.82 0.86

Multiple Linear Regression

- After constructing the required interactions
- Model form is:

$$Y = \beta_0 + \sum_{i=1}^{6} \beta_i X_i + \sum_{i=4}^{5} \delta_i X_2 X_i + \epsilon \tag{1}$$

$\beta_0,\ \beta_i,\ \&\ \delta_i$ are constants

ϵ iid Normal$(0, \sigma^2)$

- But the interaction terms have correlations with other predictors

- I built another additive model and the dropped interactions are considered later using added variable plots
- Model form is:

$$Y = \beta_0 + \sum_{i=1}^{6} \beta_i X_i + \epsilon \tag{2}$$

$\beta_0,\ \beta_i, \&\ \delta_i$ are constants

ϵ iid Normal$(0, \sigma^2)$

Model Selection

- Backward elimination and best subsets criteria based on adjusted R^2, R^2, C_p, BIC suggested models are

Procedure	**Variables included**	R^2	**Adj**R^2	$C_p(p+1)$	**BIC(lowest)**
Backward elimination	X_1, X_2, X_3, X_5	0.98	0.98	4.5	-320
Best subsets	X_1, X_2, X_3, X_4, X_5	0.98	0.98	5	-320
Best subsets	X_1, X_3, X_4, X_5, X_6	0.98	0.98	5	-320
Best subsets	X_1, X_2, X_3, X_5	0.98	0.98	4.5	-320

Model Selection

- Final model selected by me by striking a balance between a simpler model and good model attributes for further analysis is

$$Y = \beta_0 + \sum_{i=1}^{3} \beta_i X_i + \beta_5 X_5 + \epsilon \tag{3}$$

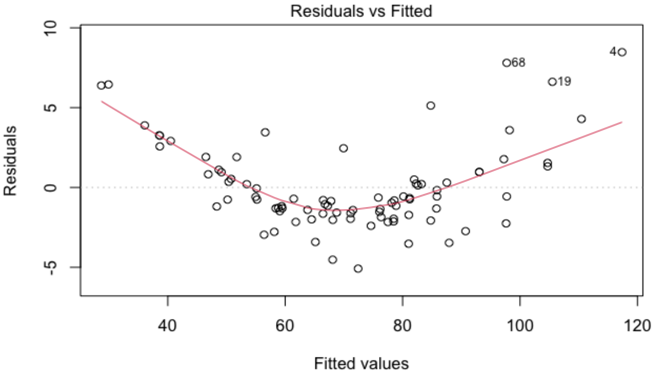
$$\beta_0,\ \beta_i,\ \&\ \beta_5 \text{ are constants}$$

$$\epsilon \text{ iid Normal}(0, \sigma^2)$$

- Retained variables are: DI, VM, BS Density and BS KV40

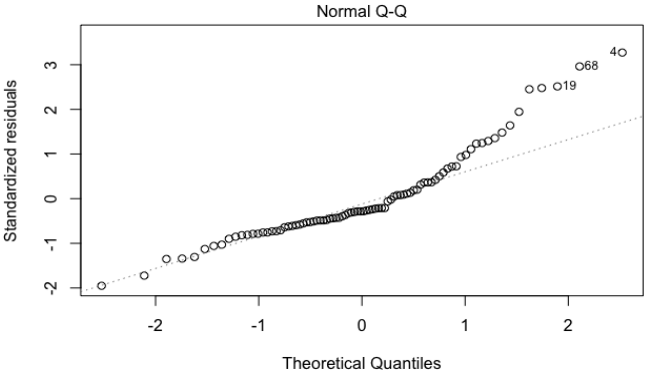
Residual Analysis

- **Linearity** assumption is not in danger, the fit curve is not too erratic
- **Homoscedascity** is clearly violated with an obvious curvature. However, Breusch-Pagan test P-value is 0.2545



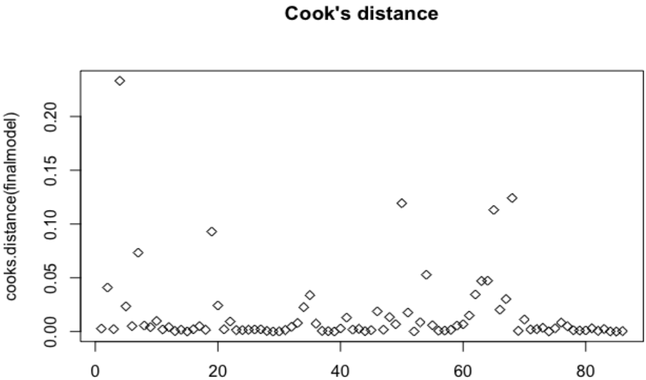
Residual Analysis

- **Normality** assumption regarding variances is violated. Plot shows points deviating too much from the straight line
- Shapiro-Wilks test yields the same result



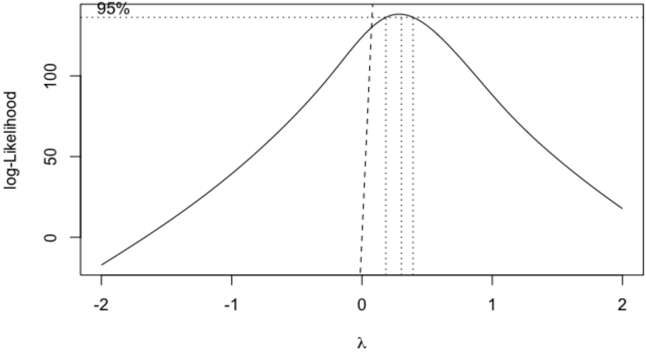
Residual Analysis

- **Multicollinearity** is not an issue at all, since all the VIFs are well within the critical value of 5
- After considering the Bonferroni limit, DFFITS and Cook's distances, leverage points, no case is particularly alarming as an **outlier**



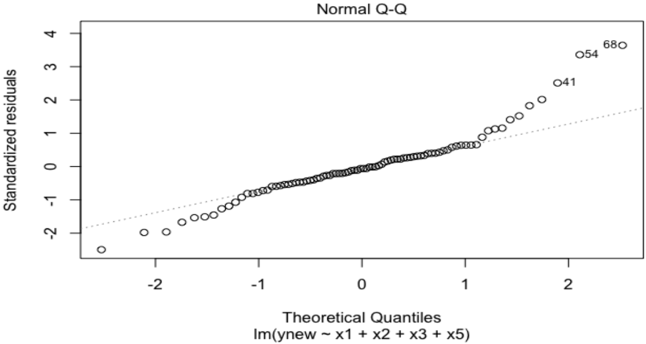
Residual Analysis

- To fix the issue of **non-constant variances**, from Box-Cox procedure, square root transformation of the Blend KV40 variable looks reasonable



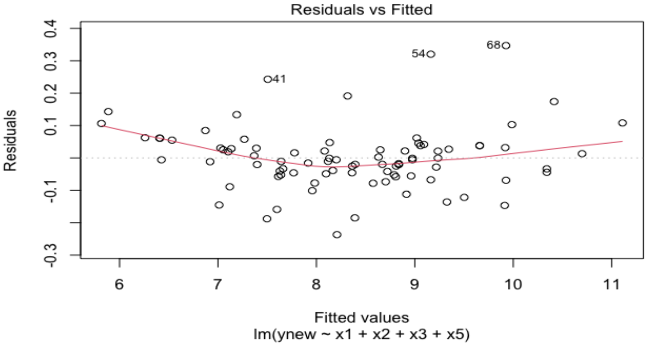
Residual Analysis

- After transformation, **linearity** and **independence** assumptions still hold **Multicollinearity** is not a serious issue, as all VIFs are well within the limit
- **Normality** of error terms is still not satisfied, but greatly improved



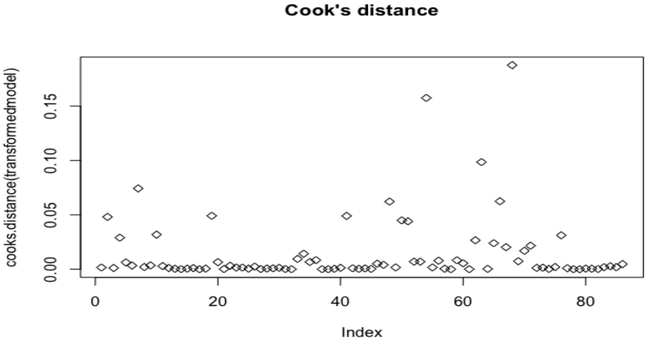
Residual Analysis

- **Homoscedasticity** is still not satisfied but greatly improved from previous plot. Curvature has weakened



Residual Analysis

- From Bonferroni limit, DFFITS and Cook's distances, and leverage points, no case is particularly alarming as an **outlier**



Final Model for future predictions of Blend KV40

- The interaction terms dropped just before variable selection were checked again using added variable plots, but none could have improved the model. Final transformed model is

$$Y' = \beta_0 + \sum_{i=1}^{3} \beta_i X_i + \beta_5 X_5 + \epsilon \tag{4}$$

$Y'$is the square root transformation of the Blend KV40 (response)

β_0, β_i, & β_5 are constants

ϵ iid Normal$(0, \sigma^2)$

- Retained variables are: DI, VM, BS Density and BS KV40

Final Model for future predictions of Blend KV40

Table: ANOVA Table for the Final Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1(DI)	1	1.334	1.334	137.238	0
x2(VM)	1	54.382	54.382	5, 595.461	0
x3(BS Density)	1	24.667	24.667	2, 538.046	0
x5(BS KV40)	1	30.015	30.015	3, 088.242	0
Residuals	81	0.787	0.010		

Final Model for future predictions of Blend KV40

Table: Coefficients and SEs – All predictors are significant at $\alpha = 0.05$

	Dependent variable:
	ynew
x1(DI)	12.212***
	(0.423)
x2(VM)	35.681***
	(0.496)
x3(BS Density)	15.657***
	(1.712)
x5(BS KV40)	0.125***
	(0.002)
Constant	$-$11.825***
	(1.399)
Observations	86
R^2	0.993
Adjusted R^2	0.993
Residual Std. Error	0.099 (df = 81)
F Statistic	2,839.747*** (df = 4; 81)
Note:	*p<0.1; **p<0.05; ***p<0.01

Thank You!