Linear and Non-linear Regression: Application to Competitor's Gasoline Volume Estimation

Haomiao Zhai
Supervisor: Magnus Wiktorsson

January 25, 2015

Abstract

This paper is dedicated to give a better estimation for competitor's gasoline volume on the behalf of Kalibrate Technologies and also find a better way to improve business and sales performance. To achieve this goal, both linear regression model and non-linear regression model have been used. The comparison between different models is discussed in the paper. The best model provides the most accurate prediction result based on statistical criterion, whose accuracy is greatly enhanced compared with existing models used in Kalibrate Technologies.
Keywords: Multiple Linear Regression, Ridge Regression, LASSO, Partial Least Square Regression, Supporting Vector Regression, Random Forrest Regression

1 Introduction

Kalibrate Technologies provide petroleum retailers with intelligence tools, software and services to analyze their operating data as the basis for defining and managing strategies to deliver on performance goals. It is very important to help its clients get information about their competitor volumes. And therefore clients could adjust their strategy and tactics for business and production. Also, they could pinpoint optimal locations for new outlets. In order to provide competitor's volume, Kalibrate Technologies dispatch surveyors to collect information onsite, such as number of pump islands, number of bypass lanes, gasoline volume, etc. Some of data are relatively objective and accurate. Like number of pump islands, the surveyor can just count it onsite, however, data such as gasoline volume (monthly) is invisible. Surveyor has to estimate this volume base on the information which they have collected, also relying on their experience and knowledge, which means that different surveyor will give different estimation. This leads to the inaccuracy of estimation. In addition, the result cannot be verified.

Considering the shortcomings of survey estimating gasoline volume and also the big cost since Kalibrate Technologies must hire more educated surveyors to complete the survey, they want to come up with a new method to estimate gasoline volume. Previously, they integrated and weighted the survey result as scores, whose formulas are proposed by market experts based on their understanding to the market. In this way, they have decreased factors to five and just used these five factors to build regression model. Even though they simplified variables within a large extent, the weights of survey result in the score formula are determined artificially. The accuracy of scored estimation is not very well, at least did not reach the desired accuracy. To improve the estimation, raw data (survey results) are used directly to fit models. The fundamental idea is to train and test data from Kalibrate Technologies' client (as proxy for competitors) and predict competitor's gasoline volume. Section 2 talks about data analysis and methods used to select variables and then gives the final list of variables for modeling. Section 3 and 4 will introduce all the regression models applied to the data and their comparison. The criteria for best model are also discussed here. Section 5 gives results of estimation. Based on the result, Section 6 and 7 talk about conclusion and recommendation.

2 Data Analysis

For each outlet, surveyors have collected data as shown in survey table, Appendix 8.1. There are 176 variables in total from the survey form. In the database, we picked up three regions. Table 1 gives a short summary from these regions. For example, surveyors investigated 291 outlets in Front

Table 1: Summary of data

Region	No. of Outlet	Mean of Volume (in thousand)	Survey Date
Front Range	291	76.99	Apr, 2013
Kansas City	223	75.76	Mar, 2013
St. Louis	235	104.25	Mar, 2013

Range. For each outlet, they computed the average monthly gasoline volume in the past year from survey date. All the collected information was entered into the database in 2013 Apr. Average monthly gasoline volume of these 291 outlets is 76,990 gallon. All the outlets belong to Kalibrate Technologies' clients. These data are used for data analysis, modeling and estimation. The model will be used as proxy for client's competitors. Data except gasoline volume can be collected to predict competitor's gasoline volume. Clearly, the objective is to build optimal model to estimate gasoline volume for individual

outlet. The model will provide insights of contribution of attribute to the volume of gasoline, helping clients make business and production decision. Attributes of outlet in the survey form are input and gasoline volume is output.

As mentioned above, each region comes with 176 variables, however, not all of them can be used for modeling. The number of variables has to be decreased to some extent in order to build appropriate regression models. Following methods are operated to do the variable reduction.

2.1 Preliminary Analysis

In preliminary analysis, attributes with no statistical meaning such as Address, Phone Numbers, etc. are removed firstly. In addition, attributes like enrollment count, auto-count, multi-dwell count, and pedestrian count have same value for all outlets, hence they are removed as they do not make any difference to the model. To achieve variable reduction, combination of original attributes are applied to create new ones since they provide similar information, for example, combining weekday and weekend hours of operation to create total number of hours of operation. By the combination, significance of variables are intensified for modeling. A complete approach of new variables created is provided in the Appendix 8.2.

2.2 Variable Selection and Modification

After preliminary analysis, there are still a bunch of variables. More techniques should be employed to explore the relationship between these variables and gasoline volume. Only the strong relationship should be considered into the model since they make contributions to the prediction of volume. For regression model, Pearson's correlation is a good indicator to show linear relationship between continuous variables. To explore other relationships, such as quadratic, scatter plot is considered here. As for categorical variables, chi-square, box plot and analysis of variance are useful to do the selection. Except relationship between explanatory variable and gasoline volume tested, relationship among explanatory variables are also examined to explore how much one variables can be explained by another one. In addition, the assumption of normality of volume should be also tested in preparation for building regression model.

- Pearson's correlation analysis: A measure of the linear correlation between two variables, which gives a value between $+1$ and -1. Variables with correlation coefficient greater than 0.3 are considered to be correlated. ((Lentner and Bishop (1993)))

- Scatter plot: as a complement to Pearson correlation analysis in dealing with non-linear relationship.

- Chi-Square test: used to check if there exists correlation between categorical attributes. (Greenwood and Nikulin (1996))

- Box plot: used to compare distributions of categorical variables. The dependence of volume on categorical attributes is also checked using box plot.

- Analysis of Variance (AOV): used to test if there is any significant difference among different levels of categorical variables with respect to volume. P-value from the test is compared with 0.05 significance level to derive conclusions. (Lentner and Bishop (1993))

- Box-Cox Transformation: used to normalize data in order to improve the validity of Pearson correlation. At the same time, this transformation eliminates skewness and other distributional features which complicate analysis. It is followed by the formula as below:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$

  (Box and Cox (1964))

2.3 Result

By statistical analysis above, it shows:

- Both Scatter plot and Pearson's correlation analysis are used to find all correlated continuous attributes, including the subset of attributes which are correlated to gasoline volume. Attributes with correlation coefficient of 0.3 or higher are selected. 11 continuous attributes are found to be correlated with gasoline volume and most of these attributes are correlated to each other, shown in the Table 2. For the purpose of practical use, model built for different regions should start with the same candidate variables. For example, Diesel Volume shows weak correlation in Kansas City but strong in Front Range and St. Louis and therefore this variable is still counted as candidate. The meaning of these variables are explained in detail in the Appendix 8.3.

Table 2: Correlation with Gasoline Volume

Variable	Front Range	Kansas City	St. Louis
Lot Size	0.4270	0.4654	0.4607
Fueling Positions	0.5930	0.6366	0.6228
# Pump Islands	0.4503	0.5513	0.4538
# Bypass Lane	0.3739	0.2786	0.4456
Total Hours	0.5171	0.4243	0.3624
# C-store Parking Spaces	0.4619	0.3653	0.3526
# C-store Cooler doors	0.4230	0.3382	0.2559
C-store Volume	0.6866	0.6647	0.5694
Car Wash Volume	0.1738	0.3444	0.3648
Diesel Volume	0.3148	0.1197	0.4596
Register	0.4544	0.4080	0.4152

- Box plot is used to compare the distribution of categorical attributes. Analysis of Variance along with Box plot are also used to check if there exists dependence between gasoline volume and categorical variables. The result is summarized in Table 3. 13 categorical attributes are correlated with gasoline volume and Chi-square test showed that most of these attributes are correlated with each other.

Table 3: Analysis of Variance Test

Categorical Variable	FR	KC	SL
Type Location	0	0	0.01
Site Location	0.05	0	0.14
Type Operation	0	0	0
Outlet Landscaping	0	0	0
Visibility	0	0	0.01
Barrier Code	0	0	0.01
Forecourt Layout	0	0	0.02
Building Size(Sq. Ft.)	0	0	0
Outlet Condition	0	0	0
C-store Products	0	0	0.04
Inside Appearance Rating	0	0	0
Car Wash	0	0	0
Gasoline Brand Name	0	0.01	0.08

- The test of Box-Cox transformation shows that the gasoline volume should be computed in the form of logarithm, Figure 1.

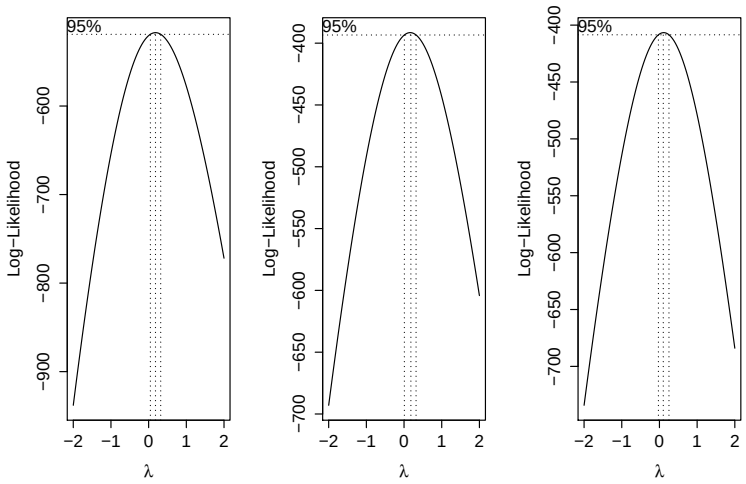

Figure 1: Plots of Box-Cox in FR, KC and SL

Figure 2 shows the heatmap of correlation among final 12 continuous variables in Front Range (For the other regions, see Appendix 8.4). Obviously, besides strong correlation between gasoline volume, the strong correlated relationship is also found between other variables, which is an important concern when building regression model later.

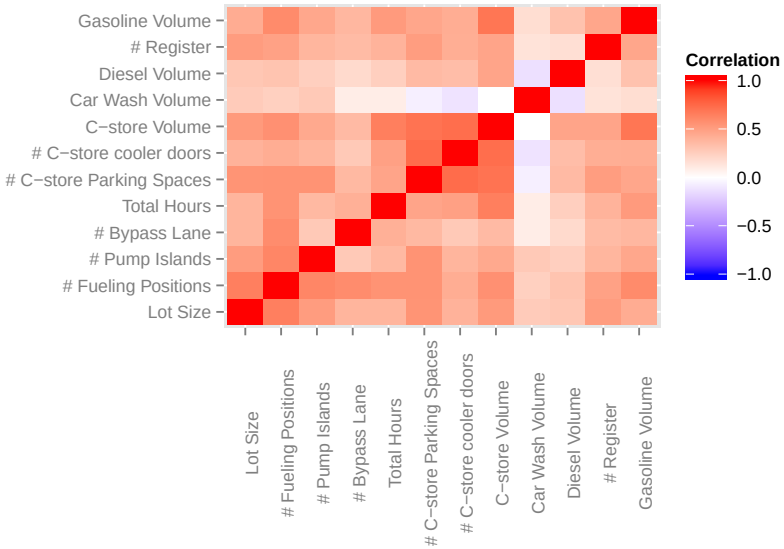

Figure 2: Heatmap of correlation among variables in Front Range

Finally, after preliminary analysis and variable selection, the final list of variables for model building has been confirmed. There are 13 categorical variables and 12 continuous variables, which will be used as candidate for modeling and prediction, shown in Table 4. In this section, we reduce the factor model remarkably to a tolerable amount which could be easy to start with building regression model. In every region, the factor will be chosen independently according to the result from modeling. In Table 2, there are some significant difference w.r.t one variable in different regions, such as weak correlation of Diesel Volume in Kansas City but not in Front Range and St. Louis, which is decided by the demand from the local. More discussion will be included in the following section.

Table 4: Final list of candidate variables

Categorical Variable	Continuous Variable
Type Location	Lot Size
Site Location	Fueling Positions
Type Operation	# Pump Islands
Outlet Landscaping	# Bypass Lane
Visibility	Total Hours
Barrier Code	# C-store Parking Spaces
Forecourt Layout	# C-store cooler doors
Building Size(Sq. Ft.)	C-store Volume
Outlet.Condition	Car Wash Volume
C-store Products	Diesel Volume
Inside Appearance Rating	# Register
Car Wash	Gasoline Volume
Gasoline Brand Name	

3 Methodology

In this section, different regression models, including linear and non-linear, are discussed based on their strength and weakness. It starts with linear regression model with ordinary least squares. More advanced models and techniques will be discussed afterwards in order to solve different problems.

3.1 Multiple Linear Regression Model (MLR)

Ordinary least squares is used to minimize the sum of squared residual between the observed responses in the dataset and the responses predicted by the linear approximation. Mathematically it solves a problem of the

form:

$$\min_{\beta} ||\mathbf{X}\beta - \mathbf{Y}||_2^2 \tag{3.1}$$

where $\beta = (\beta_1, ..., \beta_p)$ is coefficient for the linear model:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{3.2}$$

where $\mathbf{Y}$–response variable, is an $n \times 1$ vector of observations. $\mathbf{X}$–predictor variable, is a matrix $(n \times p)$ of rank p $(n > p)$. ϵ is an $n \times 1$ vector of errors with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 \mathbf{I}_n$.

However, the primary assumption behind ordinary least squares is that explanatory variables in $\mathbf{X}$ must all be linear independent. Otherwise, the matrix $\mathbf{X}$ becomes close to singular and as a result, the least-squares estimate becomes highly sensitive to random errors in the observed response, producing a large variance. And, the formula for coefficient β is violated.

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.3}$$

A plausible method to tackle with this problem is to remove factors to break the correlation relationship among the $\mathbf{X}$, nevertheless, one cannot perform optimization over the estimated predictor.

- Pros: simple to implement and interpret, can take both continuous and categorical variables into the model.
- Cons: rely on strong assumption of linearity, normality, homogeneity and independence. Also sensitive to outliers.

3.2 Ridge Regression (RR)

Hoerl and Kennard (1970) introduced the ridge regression - a biased estimation for nonorthogonal problems - to solve the multicollinearity among explanatory variables, where ordinary least squares(OLS) estimator is invalid. As one of L_1 regularization methods, the ridge regression solves (in Lagrange form):

$$\min_{\beta} ||\mathbf{X}\beta - Y||_2^2 + \lambda||\beta||_2^2 \tag{3.4}$$

The main method is to add small positive quantities to the diagonal of $\mathbf{X'X}$ in order to obtain biased estimates with smaller mean square error. Compared with equation (3.3), the modification has been made to eliminate nonorthogonality of $\mathbf{X}$.

$$\begin{aligned} \hat{\beta}^* &= [\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p]^{-1}\mathbf{X}^T\mathbf{Y},\ k \geq 0 \\ &= [\mathbf{I}_p + k(\mathbf{X'X})^{-1}]^{-1}\hat{\beta} \\ &= (\mathbf{I}_p - k(\mathbf{X'X} + k\mathbf{I}_p)^{-1})\hat{\beta} \end{aligned} \tag{3.5}$$

The choice of k plays a vital role in ridge regression. To simplify the question, it is assumed that columns of $\mathbf{X}$ and $\mathbf{Y}$ are standardized such that $\mathbf{X}'\mathbf{X}$ is a autocorrelation matrix and $\mathbf{X}'\mathbf{Y}$ is the correlation matrix between $\mathbf{X}$ and $\mathbf{Y}$. Taking the eigendecomposition of $\mathbf{X}'\mathbf{X}$:

$$\mathbf{X}'\mathbf{X} = Q\Lambda Q' \tag{3.6}$$

Columns of Q are the eigenvectors of $\mathbf{X}'\mathbf{X}$ and $Q'Q = QQ' = \mathbf{I}_p$ because $\mathbf{X}'\mathbf{X}$ is $p \times p$ real symmetric matrix. λ_i is the eigenvalues of $\mathbf{X}'\mathbf{X}$ corresponding to ith column from Q as eigenvector. Eigenvalues form the diagonal matrix $\Lambda = diag(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0)$. The equivalent model can be built with Q:

$$\mathbf{Y} = \mathbf{Z}\alpha + \epsilon \tag{3.7}$$

where $\mathbf{Z} = \mathbf{X} * Q$ and $\alpha = Q'\beta$. Then the OLS estimator for α is

$$\hat{\alpha} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} = \Lambda^{-1}\mathbf{Z}'\mathbf{Y} \tag{3.8}$$

Combined with (3.3) and (3.6), the relationship is built between OLS estimator $\hat{\beta}$ and $\hat{\alpha}$:

$$\hat{\beta} = Q\hat{\alpha} \tag{3.9}$$

Based on (3.5), the following equation for ridge regression estimator is realized by exchange $\mathbf{X}$ with $\mathbf{Z}$

$$\hat{\alpha}^* = (\mathbf{I}_p - k(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}_p)^{-1})\hat{\alpha} \tag{3.10}$$

Hence, the ordinary ridge regression estimator of β is

$$\hat{\beta}^* = Q\hat{\alpha}^* = Q(\mathbf{I}_p - k(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}_p)^{-1})\hat{\alpha} \tag{3.11}$$

Using mean square error as a criterion to evaluate bias and efficiency between the ordinary least squares estimator and ordinary ridge estimator:

$$MSE = Variance + Bias^2 \tag{3.12}$$

$$MSE(\hat{\alpha}) = \hat{\sigma}^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{3.13}$$

$$MSE(\hat{\alpha}^*) = \hat{\sigma}^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{(\lambda_i + k)^2} \tag{3.14}$$

The special case is when $k = 0$, MSE of these two methods are equal. Hoerl and Kennard (1975) pointed out that a small k will lead to mean square error of ridge estimator less than the mean square error of OLS estimator

and they also provided an approach to set the value for k:

$$k_1 = \frac{p\hat{\sigma}^2}{\hat{\alpha}'\hat{\alpha}},\ \textit{Hoerl and Kennard (1975)} \tag{3.15}$$

The modification has been made by adding p(the rank of $\mathbf{X}$) to the formula compared with their previous study (Hoerl and Kennard (1970)), increasing the penalty for the number of estimators.

$$k_2 = \frac{\hat{\sigma}^2}{\hat{\alpha}'\hat{\alpha}},\ \textit{Hoerl and Kennard (1970)} \tag{3.16}$$

Considering the pervasive multicollinearity in the dataset, it is better to use penalty for the dependence. Dorugade and Kashid (2010) proposed an alternative method for choosing ridge estimator with including variance inflation factor:

$$k_3 = \max\ \Big(0,\ \frac{\hat{\sigma}^2}{\hat{\alpha}'\hat{\alpha}} - \frac{1}{n(VIF_j)_{max}}\Big),\ \textit{Dorugade and Kashid (2010)} \tag{3.17}$$

where $VIF_j = \frac{1}{1-R_j^2},\ j = 1,\ 2, \ldots,\ p.$
In addition, Dorugade (2014) reviewed previous research on the estimation of ridge regression and proposed a new method for ordinary ridge estimator:

$$k_4 = Harmonic\ Mean = \frac{2p}{\lambda_{max}} \sum_{i=1}^{p} \frac{\hat{\sigma}^2}{\hat{\alpha}^2} \tag{3.18}$$

To build the ridge regression model for competitor's volume, we will use these four ordinary ridge estimators. The result is that the estimators $\hat{\beta}_j$ from (3.5), are shrunk towards zero and yet they can not reach zero, which means it is not possible to decrease the number of estimators with ridge regression. It is because we use L_2 norm that it decreases the sum sharply when β goes to 0. Considering that we have a big predictor pool within competitor's volume estimation, we had better choose a model with sparse coefficients.

- Pros: handle multicollinearity, customize parameter to adjust model for different requirements and can reach the closed-form solution.
- Cons: cannot do variable reduction, be inaccurate for large-size regression model

3.3 Least Absolute Shrinkage and Selection Operator (LASSO)

To reach the goal of variable reduction, we think of lasso regression (Tibshirani (1994))–L_1 regularization, as its capability to reduce $\hat{\beta}_j$ to zero, which

solves (in Lagrange form):

$$\min_{\beta} \frac{1}{2}||\mathbf{X}\beta - Y||_2^2 + \lambda||\beta||_1 \tag{3.19}$$

Because of the absolute value operation in (3.19), the objective function is not differentiable except an orthonormal design matrix $\mathbf{X}$, which is not the usual case and as a result there is no closed form solutions for the lasso. However, we could tackle the problem by least angle regression algorithm. The usual lasso has its own limits, such that it gives the same penalty to all β_j and fails to deal with the situation with more categorical predictors, which is the right case for the predictors in regression model. Therefore, we extend our research to more general model. We mentioned that Ridge Regression cannot do variable shrinkage while LASSO can. The reason is regularization. In LASSO, the penalty towards β is less than Ridge Regression. When decreasing β, Ridge Regression can reach the minimum more quickly. So, LASSO will force some of parameters to zero in order to get the minimum. Without loss of generality, we consider the two-factor model, shown in the Figure 3.

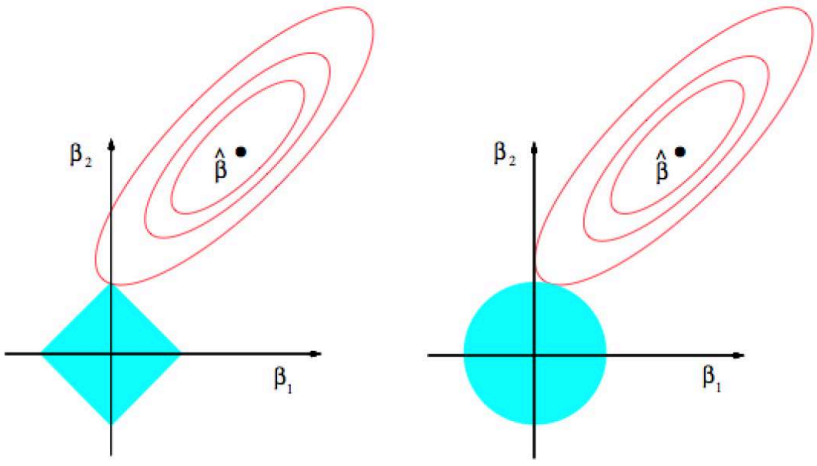

Figure 3: Comparison between Ridge Regression and LASSO

Estimation picture fro the LASSO(left) and Ridge Regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function. Note that LASSO can reach the optimal on the y-axis, forcing β_1 to zero and then do the variable shrinkage; however, Ridge Regression only get the optimal on the contour of the circles instead of axis. To choose the optimal

λ in the problem of optimization 3.19, we compute the mean square error in the Figure 4.

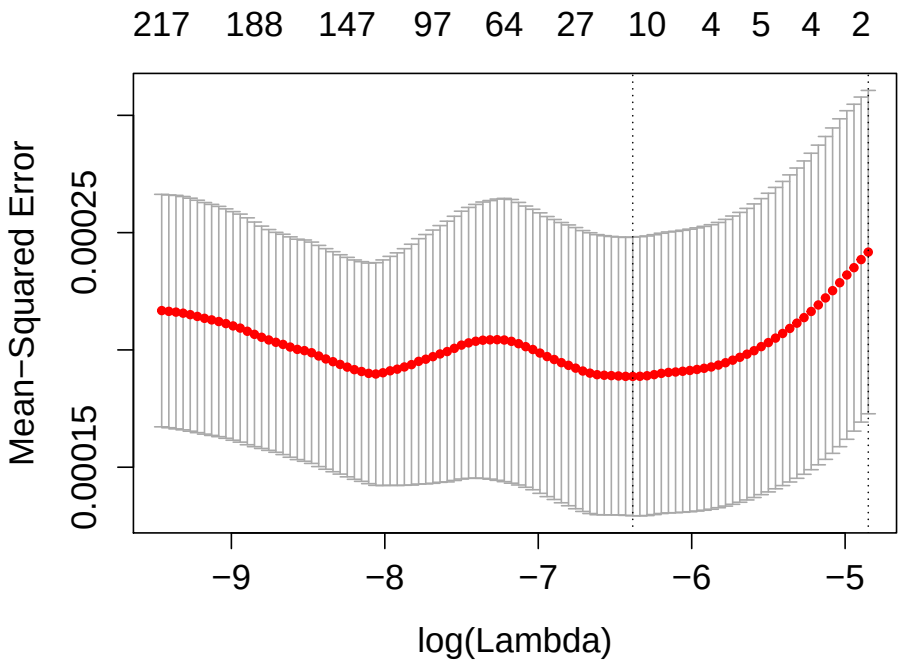

Figure 4: The choice of optimal λ

The numbers across the top of the plot give the number of nonzero $\hat{\beta}_{ij}$ for each value of λ (shown on the log-scale along the horizontal axis). The dotted vertical line on the left marks the point where the MSE is minimized. But, the MSE is only estimated by cross-validation. In recognition of this, another dotted line is drawn at the point which is within one standard error of the minimum MSE; it is typical to use this second λ as optimal choice.

- Pros: can do variable reduction and handle multicollinearity.
- Cons: numerical algorithm has to be used to build the model since L_2 Regularization method is not differentiable.

3.4 Partial Least Squares Regression (PLSR)

The problem with using the least square error is that $\mathbf{X}'\mathbf{X}$ can be singular in the presence of large number of explanatory variables or multicollinearity. PLSR (Hoskuldsson (1988)) finds the solution by decomposing $\mathbf{X}$ in to orthogonal scores and loadings. Underlying Model: Compo-

nents/Scores/Latent Variables are obtained iteratively.

$$S = \mathbf{X}'\mathbf{Y} \tag{3.20}$$

where the variation in both $\mathbf{X}$ and $\mathbf{Y}$ are included. Score t is obtained by using ω as weight of vectors $\mathbf{X}$.

$$t = \mathbf{X}\omega \tag{3.21}$$

The loadings p and q are given by

$$\begin{aligned} p &= \mathbf{X}'t \\ q &= \mathbf{Y}'t \end{aligned} \tag{3.22}$$

w, t, p and q are updated after each iteration in matrices $\mathbf{W}$, $\mathbf{T}$, $\mathbf{P}$ and $\mathbf{Q}$. A different way of expressing the weights so that they all related to the original explanatory variables matrix $\mathbf{X}$ is

$$R = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1} \tag{3.23}$$

Finally instead of regressing $\mathbf{Y}$ on the original variables, the regression is done on Scores $\mathbf{T}$. The regression coefficients are given by:

$$\beta = R(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y} = R\mathbf{Q}' \tag{3.24}$$

- Pros: solve multicollinearity and do variable reduction whose subset selection produces a model that is interpretable and possibly lower prediction error than the full model (bias-variance trade off). Both continuous and categorical variables can be taken in to the model.
- Cons: can not see the effect of original variables and rely on the assumptions of linearity and homogeneity.

3.5 Support Vector Regression (SVR)

Underlying Model: For linear functions of the form:

$$f(x) = \beta_0 + x'\beta, \ where \ \beta_0 \in R, \ \beta \in R^p \tag{3.25}$$

In SVR, the optimal regression function is given by the following problem (Vapnik (1995)):

$$\begin{aligned} \min_{\beta} \quad & \tfrac{1}{2}||\beta||^2 \\ \text{s.t.} \quad & y_{i-}(\beta_0 + x_i'\beta) \leq \epsilon \\ & (\beta_0 + x_i'\beta) - y_i \leq \epsilon \end{aligned} \tag{3.26}$$

In cases of allowing some errors, Vapnik (1995) proposed relaxed variables ξ can be introduced into the problem:

$$\begin{array}{lccl} \min\limits_{\beta} & \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}(\xi_{i+} + \xi_{i-}) & & \\ \text{s.t.} & y_{i-}(\beta_0 + x_i'\beta) & \leq & \epsilon + \xi_{i+} \\ & (\beta_0 + x_i'\beta) - y_i & \leq & \epsilon + \xi_{i-} \\ & \xi_{i+} \geq 0,\ \xi_{i-} \geq 0 & & \end{array} \tag{3.27}$$

where C is a regularization parameter. The Lagrangian function is

$$\begin{aligned} L(\beta_0, \beta, \xi_{\pm}, \gamma_{\pm}) =& \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}(\xi_{\imath+} + \xi_{i-}) - \sum_{i=1}^{N}(\eta_{i+}\xi_{i+} + \eta_{i-}\xi_{i-}) \\ &+ \sum_{i=1}^{N}\gamma_{i+}(y_i - \beta_0 - x'\beta - \epsilon - \xi_{i+}) + \sum_{i=1}^{N}\gamma_{i-}(\beta_0 + x'\beta - y_i - \epsilon - \xi_{i-}) \end{aligned} \tag{3.28}$$

where η_{i+}, η_{i-}, γ_{i+} and γ_{i-} are Lagrange multipliers.

- Pros: no distribution assumption for errors, no limit on the number of variables, map non-linear data using kernels and take both categorical and continuous variables in to the model.
- Cons: models can be sensitive to the choice of kernel which may lead to different results and computationally can be costly and intense.

3.6 Random Forest Regression (RFR)

In Breiman (2001), given random vector Θ and the tree predictor $h(x, \Theta)$ take on numerical values and the training set is independently drawn from distribution of the random vector $\mathbf{Y}$ and $\mathbf{X}$, the mean-squared generalization error for numerical predictor $h(x)$ is

$$E_{\mathbf{X},\mathbf{Y}}(\mathbf{Y} - h(\mathbf{X}))^2 \tag{3.29}$$

There are two stages for building the model. In the first stage, an algorithm is used to grow the tree model. In the second stage, the full tree is pruned back to get the best choice for the model. The algorithm for RFR is summarized below:

1. Draw sample of size n from the training data.
2. For reach of the samples in (1), grow an un-pruned regression tree by randomly taking samples of predictors and choose the best split from those predictors.
3. Output all the trees T_1, T_2, ..., T_M, where M is the number of trees.

4. The prediction for RFR is found by averaging the prediction from all M trees.

The strength and weakness of Random Forest Regression are as followings:

- Pros: better prediction, can deal with multicollinearity and handle large number of predictors when the sample size is small, more stable since it is a combination of many trees.
- Cons: a single predictor variable can show up in several different trees as a result it is difficult to see the size and direction of predictor effects. Also take long computing time.

4 Model Criteria

After building different statistical models, a few statistical comparison creteria are used to compare the models to select the best model.

4.1 Adjusted R-Squared

Adjusted R-squared is a statistical measure that explains how well a model fits a data. It indicates the proportion of total variation in the response variable that is explained by the predictor variables. In addition, this measure is adjusted to the number of explanatory variables. The bigger the adjusted R-Square the better the model.

4.2 Root-mean-square error

Root-mean-square error (RMSE) is used to measure differences between values predicted by a model and observed values. It is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent. A model with small RMSE is preferred.

4.3 Mean Absolute Percentage Error and Mean Absolute Error

Mean Absolute Percentage Error (MAPE) and (MAE) are measures of prediction error. They are the average absolute percent error and average absolute error respectively. Models with small MAPE or MAE are preferred.

4.4 AIC and BIC

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used to select a model from a set of models. The model chosen using AIC and BIC should be the one that minimizes the distance between

the fitted value and the observed value. Simply AIC and BIC are criteria that find a model that has a good fit to the observed values by using penalty for additional terms. The difference between AIC and BIC remains in the penalty term-BIC has larger penalty. These criteria are only applicable to models that use likelihood function to estimate their parameters. Models with small AIC and BIC are preferred.

5 Result

Six main models mentioned above and their sub-models derived from the main models are implemented for all the regions. 80% of the data is used to train the models and 20% is used for testing.

- MLR model, there are 4 sub-models. Stepwise regressions with AIC and BIC as variable selection criteria are used to select the predictors which are listed in the final model. To solve multicollinearity, interaction between variables whose correlation are larger than 0.5 are also considered. Likewise, final models are determined by stepwise regression with both AIC and BIC.

- RR model, there are 4 sub-models. They have different algorithms for configuring parameter k, as mentioned above. The choice of k will influence the accuracy of model. Since ridge regression includes all the parameters, some levels in the categorical variable with only 1 data have to be deleted.

- LASSO model can do variable reduction and therefore simplify the model.

- PLSR model, there are 6 sub-models. Several scores (components) with small root mean square error for prediction and with large percent of variation explained are tried to get the best model, using n to denote number of scores, such as $PLSR_{n=3}$. The variables selected using AIC ($PLSR_{AIC}$) and BIC ($PLSR_{BIC}$) are also used to build PLSR sub models.

- SVR model, there are 3 sub-models. Different combinations of cost and epsilon are cross validated to select the best SVR model. Since it does not do variable reduction, the variables selected using AIC (SVR_{AIC}) and BIC (SVR_{BIC}) above are also used to build SVR sub-models.

- RFR model, the optimal number of variables to split at each nodes are searched with respect to Out-of-Bag error estimates.

For each model, different sub-models are tried, and results are reported below. The models are compared from two different perspectives: prediction

error (RMSE) and goodness of fit (Adjusted R-Square) - how well the models fit the data. RMSE are used to select the best model since RMSE has the same unit as the data and also can capture the presence of big errors, which is a more conservative way for prediction.

5.1 Front Range

Model	Adj. R^2	RMSE	MAPE%	MAE	AIC	BIC
AIC	58.48	26.09	29.19	19.76	249.39	311.43
BIC	57.59	31.67	30.54	21.34	256.31	287.33
$AIC_{interaction}$	66.53	26.31	30.35	18.30	235.90	318.52
$BIC_{interaction}$	62.18	32.18	37.72	22.89	250.19	281.17

Table 5: Result of MLR models in Front Range

From Table 5, it is clear that adding interaction into the model does not improve either fitness or prediction. RMSE, MAPE and MAE in final models selected by stepwise AIC and BIC are smaller than $AIC_{interaction}$ and $BIC_{interaction}$, respectively.

Model	Adj. R^2	RMSE	MAPE%	MAE
RR_{k_1}	59.7897	29.5845	34.9107	20.0080
RR_{k_2}	63.9844	28.2931	32.0103	19.4123
RR_{k_3}	59.7903	29.5844	34.9103	20.0078
RR_{k_4}	40.3333	46.3849	69.4385	34.6330
LASSO	56.64	28.31	31.74	19.37
$PLSR_{n=3}$	46.97	37.06	36.85	25.44
$PLSR_{n=6}$	49.54	33.41	28.90	22.04
$PLSR_{n=7}$	56.19	33.50	31.08	22.95
$PLSR_{n=8}$	56.48	32.02	30.41	22.08
$PLSR_{AIC}$	59.61	31.19	29.91	21.38
$PLSR_{BIC}$	55.80	31.44	30.71	21.24
SVR_{AIC}	70.60	33.67	32.19	22.89
SVR_{BIC}	80.01	30.11	34.18	22.89
SVR_{full_model}	67.21	29.82	28.34	20.89
RFR	90.58	28.62	30.89	20.71

Table 6: Result of other models in Front Range

In Table 6, RR_{k_2} excels above other models based on RMSE, however, it does beat MLR model, which implies, multicollinearity may not be a big problem in Front Range. The result of PLSR depends on the number of scores. In this case, larger score gives better result. The best model for Front Range is MLR.

5.2 Kansas City

Model	Adj. R^2	RMSE	MAPE%	MAE	AIC	BIC
AIC	56.80	36.75	33.52	26.26	236.67	297.12
BIC	53.27	32.39	29.92	22.39	239.53	261.80
$AIC_{interaction}$	56.25	48.28	37.32	30.18	244.13	323.67
$BIC_{interaction}$	52.01	35.22	32.46	23.75	243.27	262.36

Table 7: Result of MLR models in Kansas City

From Table 7, it is even worse to add interaction into the model. In Kansas City, final model using stepwise with BIC gives smaller values for RMSE and MAPE, which is inverse in Front Range. It indicates that more penalty for the number of parameters provides a better prediction.

Model	Adj. R^2	RMSE	MAPE%	MAE
RR_{k_1}	44.98	31.9504	29.3189	22.8416
RR_{k_2}	53.18	33.6514	30.4375	24.3022
RR_{k_3}	44.98	31.9497	29.3184	22.8411
RR_{k_4}	40.23	53.1735	53.9850	37.4569
LASSO	54.77	38.31	31.95	26.01
$PLSR_{n=3}$	45.30	36.43	35.46	24.76
$PLSR_{n=5}$	53.00	35.13	30.89	24.66
$PLSR_{n=7}$	52.50	32.77	29.61	22.87
$PLSR_{n=8}$	52.94	33.47	29.66	22.37
$PLSR_{AIC}$	51.88	35.63	30.55	22.39
$PLSR_{BIC}$	56.17	32.06	29.91	22.25
SVR_{AIC}	46.24	36.40	33.96	25.62
SVR_{BIC}	67.21	37.22	32.43	25.73
SVR_{full_model}	39.09	36.89	33.01	25.42
RFR	89.96	32.38	34.95	22.84

Table 8: Result of other models in Kansas City

In Table 8, RR_{k_3} gives more accurate result compared with other models. The optimal number of scores is 7, however, the model does not win over $PLSR_{BIC}$, which means that variables from final model selected by BIC have included information from other variables well. The best model in Kansas City is RR_{k_3}.

5.3 St. Louis

Model	Adj. R^2	RMSE	MAPE%	MAE	AIC	BIC
AIC	60.36	58.56	31.14	37.28	228.1671	285.74
BIC	59.29	59.99	28.45	36.33	234.8134	266.80
$AIC_{interaction}$	64.83	58.81	32.68	39.43	235.90	318.52
$BIC_{interaction}$	61.12	61.64	30.86	39.00	250.19	281.17

Table 9: Result of MLR models in St. Louis

Just like Front Range and Kansas City, from Table 9, it shows weak prediction capacity after adding interaction into the model. The final model selected by AIC criteria gives better result based on RMSE. Smaller values of MPAE and MAE in model selected by BIC indicates that there are some large values which cannot be predicted well by this model (BIC).

Model	Adj. R^2	RMSE	MAPE%	MAE
RR_{k_1}	56.26	59.0834	30.6887	37.7358
RR_{k_2}	59.81	68.5172	31.4954	39.1128
RR_{k_3}	56.27	59.0833	30.6878	37.7362
RR_{k_4}	50.23	68.7863	49.0375	49.6624
LASSO	59.21	60.32	31.70	39.29
$PLSR_{n=3}$	46.03	65.29	40.77	42.43
$PLSR_{n=5}$	48.44	61.61	30.88	36.36
$PLSR_{n=6}$	57.59	76.92	34.51	43.31
$PLSR_{n=8}$	57.67	66.71	31.74	39.46
$PLSR_{AIC}$	62.19	56.75	29.84	35.47
$PLSR_{BIC}$	59.15	59.42	29.06	36.73
SVR_{AIC}	56.58	57.76	30.50	35.50
SVR_{BIC}	71.29	54.97	31.04	34.59
SVR_{full_model}	98.37	58.66	31.61	37.05
RFR	90.41	56.12	37.43	37.41

Table 10: Result of other models in St. Louis

In Table 10, better prediction result is from RR_{k_3} among ridge regression models. The best model is SVR_{BIC}.

5.4 Result Analysis

By comparison of these six tables above, one can find common properties among them.

- For MLR model, adding interaction is not a good method to improve the model's capacity of prediction. Even though there indeed exists

multicollinearity among variables, which is also verified by the strong correlation between them, stepwise cannot end up with better result of model with interaction. One possible reason is that the improvement by adding interaction is concealed by more variables to model.

- For RR model, it is not hard to see that RR_{k_4} is better than RR_{k_1} in all the regions. This proves that the modification by VIF in computing parameter k takes effect. However, k_4, modified by harmonic mean, always gives a very bad result. This may imply that harmonic mean is not a good choice for model with categorical variables. When facing multi-level in categorical variables, ridge regression needs to break them up into dummy variables (0/1). Considering that ridge regression does not do variable reduction, the size of model will be potentially large.

- For PLSR model, the optimal number of scores changes based on the dataset but it is feasible to use a sub-set of variables to improve the prediction capacity of the model, which is derived from final model using stepwise by AIC and BIC criteria.

- For SVR model, it is similar as PLSR model. Since it does not do variable reduction, manually simplified model is useful in this case.

- For RFR model, there exists over-fitting issues. Adjusted R squared is surprisingly high all the time, compared with other variables, however, the prediction (from the value of RMSE and MAPE) is not good accordingly. This over-fitting problem comes from the algorithm of random forest regression. Therefore, adjusted R squared is not a good indicator when dealing with models of random forest regression.

6 Conclusion

The best model is selected from each region and compared with result of survey estimation and scored model estimation, as shown in Table 11.

Region	Best Model	MAPE(%)	Score-MAPE (%)	Survey-MAPE (%)
Front Range	MLR	29.19	36.00	90.23
Kansas City	RR_{k_3}	29.32	40.54	71.18
St. Louis	SVR_{BIC}	31.04	41.85	64.85

Table 11: Comparison of MAPE among different methods

All the best models using raw data give better accuracy compared with other methods in Front Range, St. Louis and Kansas City. It is a huge improvement beyond survey estimation. Even though, scored models do

better prediction than survey ones, they are still 8% in average less than model above. One possible reason is that scored data which derived from raw data is defined mannualy and objectively. It usually reflects the experience and practice of production and industry, however, it may not have the statistical meaning. Models with raw data are dedicated to find the basic properties and relationships among variables, which is statistical reasonable.

7 Recommendation

Considering the requirement of practical use– Kalibrate Technologies are trying to offer up-to-data estimation result to its clients and also integrate with other software, the final model should be less human-depending. In these models, PLSR and LASSO may be not appropriate. When choosing number of scores, the pick-up has to be determined subjectively. This algorithm cannot finish the process by itself. Similarly, in LASSO, the computation needs to choose indexing path (Efron, Hastie, Johnstone, and Tibshirani (2004)) manually. RFR has the risk of over-fitting. Therefore, it is recommended to use MLR (by AIC and BIC), RR_{k_2}, RR_{k_3} and SVR (variables from final model MLR).

8 Appendix

8.1 Survey Form

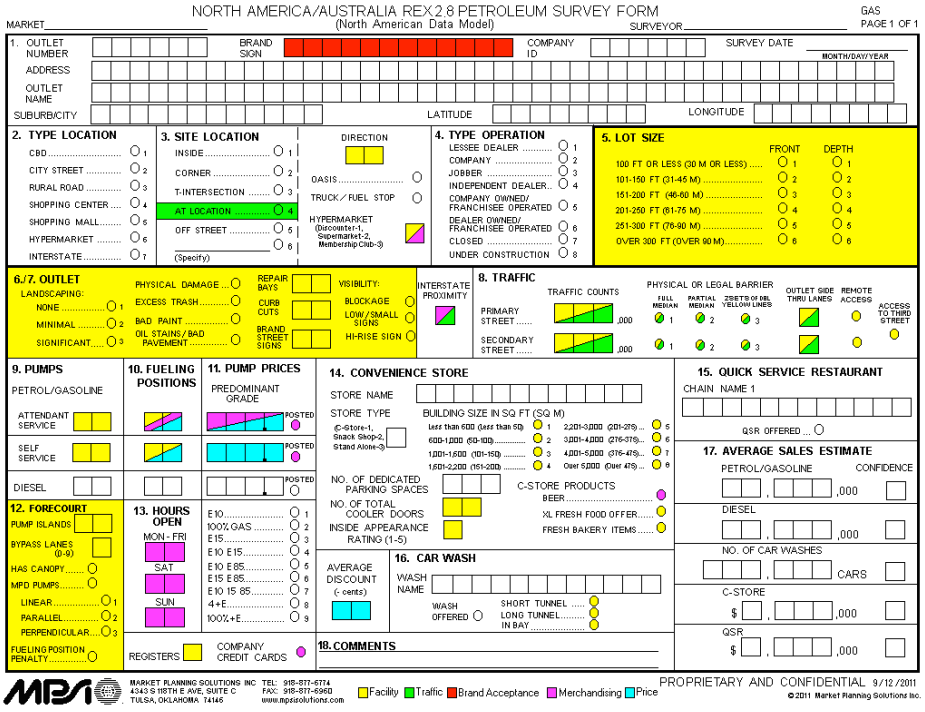

Figure 5: Survey Form

8.2 Newly Created Variables

Outlet Condition: if an outlet has any damage. EXCESS_ TRASH_ ANS+PHYSICAL_ DAMAGE_ ANS+BAD_ PAINT_ ANS+OILSTAINS_ BAD_ PAVEMENT_ ANS

Visibility: if a site has visibility problem. HI-RISE_ SIGN_ ANS+ BLOCKAGE_ ANS+LOW_ SMALL_ SIGN_ ANS

Barrier Code: if there is any physical or legal barrier in the primary or secondary street. ADDR1_ BARR_ LK+ADDR2_ BARR_ LK

C-store Product: if C- store sells at least one of these products (beer, fresh food or fresh bakery items) in the convenient store. CST_ BAKERY_ ANS+CST_ COLD_ BEER_ ANS+CST_ DELI_ ANS

Car Wash: combine short and long tunnel with bay to check if a site has car wash. WSH_ LONG_ TUNNEL_ ANS+ WSH_ OTHER_ ANS+ WSH_ SHORT_ TUNNEL_ ANS

Fueling Positions: total number of fueling positions at attendant service and

self-service. GAS_ AT_ FPOS_ CNT+GAS_ SF_ FPOS_ CNT
Total Hours: total number of hours for weekdays and weekends. MON_ FRI_ HRS+SAT_ HRS+SUN_ HRS
Lot Size: Width∗Height. Width=mid-point of each width range, Height=mid-point of each height range.

8.3 Explanations on Variable

Lot size(5): Size of Front and Depth of the property developed or perceived by the customer as being used for the sale of petroleum and other related products such as auto repair, car wash, convenience food, or truck plaza. Here, lot size is computed as the product of Front and Depth.
Fueling Positions(10): Number of fueling positions for each service type under the appropriate canopy situation.
Pump Prices(11): The price of gasoline. ($\times$ 100)
 # Pump Islands(12): Number of pump islands at the outlet.
 # Bypass Lane(12): Number of bypass lanes at this outlet. A bypass lane is a passing lane between pumps, between pumps and a building, or between pumps and a curb. Bypass lanes will allow a motorist to pull between occupied fueling positions to get to another fueling position or to enter or exit the outlet.
Total Hours(13): The total operation hours in one week.
 # C-store Parking Spaces(14): Number of parking spaces in front of convenient store.
 # C-store Cooler doors(14): Total number of cooler doors set against or within the walls. Free-standing coolers in the middle of the c-store are not counted.
Gasoline Volume(17): Average monthly gasoline volume, not including commercial business.
C-store Volume(17): Average monthly c-store volume for gasoline outlets.
QSR Volume(17): Sales of quick serve resturant.
Car Wash Volume(17): Number of car washes estimate per month.
Diesel Volume(17): Average monthly diesel volume.
Register(13): Number of electronic cash registers used for gasoline, c-store, and quick service restaurant sales at this facility. Every outlet has at least one register, and many may have multiple registers.
Type Location (2): Description of the area in which the outlet is located. They are CBD (Central Business District), City St, Rural Rd, Shopping Cntr, Shopping Mall, Hypermarket, Interstate(Limited Access, Highway) and Oasis.
Site Location(3): Description of the site at which the outlet is located. They are inside–access from one street only, corner–access from two streets, T-Intersection–access from one or two streets depending on the exact lo-

cation, at location–an outlet located at the intersection of two streets but accessible from only one of the two streets, off street–No direct street access, such as an outlet on the grounds of a shopping center or shopping mall and other.
Type Operation(4): If an outlet has more than one brand name displayed (such as a convenience food store that displays both the store name and the gasoline brand name), type of operation is evaluated from the viewpoint of the brand name entered in the Gasoline Brand Sign field. They are Lessee Dealer, Company, Jobber, Contract Dealer, Company Owned/Franchisee Operated, Dealer Owned/Franchisee Operated, Closed and Under Construction.
Outlet Landscaping(6/7): Whether there is a scene point nearby, categorized by none, minimal and significant.
Building Size(Sq. Ft.)(14): Estimated size of the outside dimensions of the c-store building. Includes all the area of the building that is involved with convenience food retailing, such as cooler backup space and storage areas.
Forecourt Layout(12): Focus on the primary gas offering, such as self or automat only, when determining forecourt layout. Ignore half pumps that are not back to back and odd gas pumps on the lot, unless it helps you fit into a category. Two half pumps can equal one full pump when necessary, categorized by linear parallel and perpendicular.
Inside Appearance Rating(14): The convenience storeÕs inside appearance rated using from 1-5. The general housekeeping, cleanliness, and attractiveness of the sales area (decor and equipment) are considered.
Rating Description:

1. Poorly maintained, unpolished floors, spills on counters, floors, or equipment
2. An outdated facility that is well maintained
3. An average clean facility or a new facility that is poorly maintained
4. A modern facility that is well maintained
5. A facility that is superior in cleanliness and appeal, by all industry standards

8.4 Figures and Tables

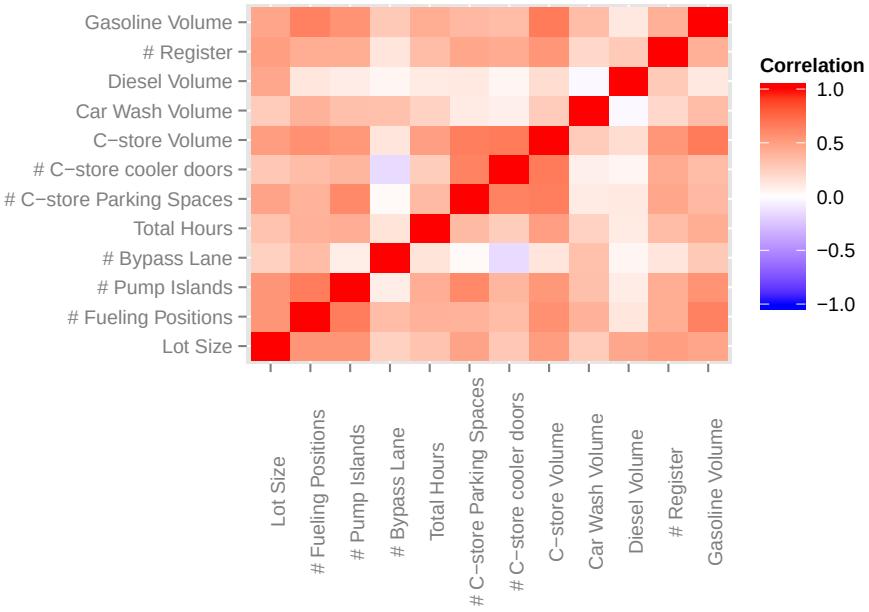

Figure 6: Heatmap of correlation among variables in Kansas City

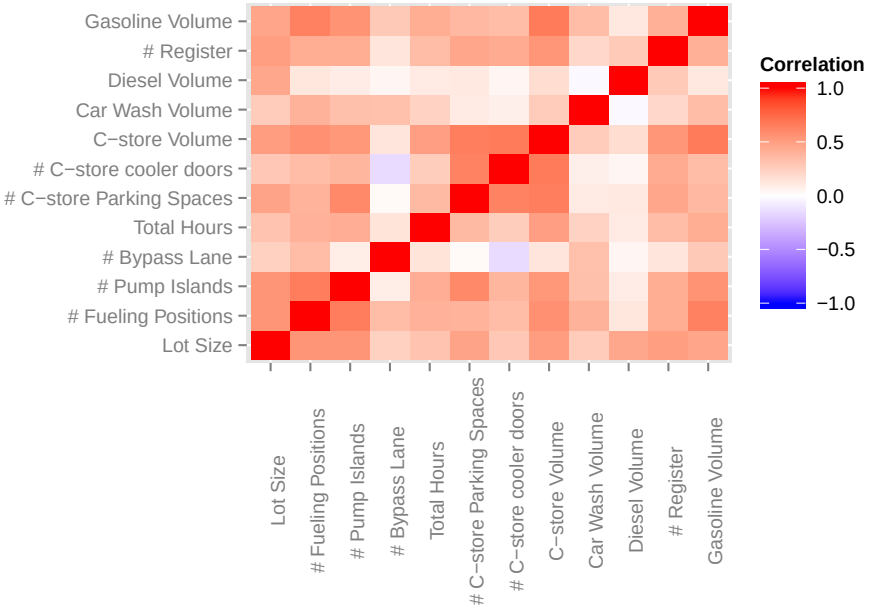

Figure 7: Heatmap of correlation among variables in St. Louis

References

Box, G. E. P. and Cox, D. R. An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26(2):pp. 211–252, 1964. ISSN 00359246.

Breiman, Leo. Random forests. Machine learning, 45(1):5–32, 2001.

Dorugade, A. V. New ridge parameters for ridge regression. Journal of the Association of Arab Universities for Basic and Applied Sciences, 15:94–99, April 2014.

Dorugade, A. V. and Kashid, D. N. Alternative method for choosing ridge parameter for regression. Applied Mathematical Sciences, 4(9):447–456, 2010.

Efron, Bradley; Hastie, Trevor; Johnstone, Iain, and Tibshirani, Robert. Least angle regression. The Annals of Statistics, 32(2):407–499, 04 2004.

Greenwood, P.E. and Nikulin, M.S. A guide to chi-squared testing. Wiley, New York, 1996. ISBN 0-471-55779-X.

Hoerl, Arthur E. and Kennard, Robert W. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67, 1970.

Hoerl, Arthur E. and Kennard, Robert W. Ridge regression:some simulations. Communications in Statistics, 4:105–123, 1975.

Hoskuldsson, Agnar. Pls regression methods. Journal of Chemometrics, 2 (3):211–228, 1988.

Lentner, Marvin and Bishop, Thomas. Experimental design and analysis (Second ed.). Valley Book Company, 1993. ISBN 0-9616255-2-X.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58:267–288, 1994.

Vapnik, Vladimir N. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.