A Review of 'Big Data' Variable Selection Procedures For Use in Predictive Modeling

Sarah Papke

Follow this and additional works at: https://dsc.duq.edu/etd

Recommended Citation

This Immediate Access is brought to you for free and open access by Duquesne Scholarship Collection. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Duquesne Scholarship Collection. For more information, please contact phillipsg@duq.edu.
A REVIEW OF ‘BIG DATA’ VARIABLE SELECTION PROCEDURES FOR USE IN PREDICTIVE MODELING

A Thesis
Submitted to the McAnulty College & Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for
the degree of Master of Science

By
Sarah Papke

August 2017
A REVIEW OF ‘BIG DATA’ VARIABLE SELECTION PROCEDURES FOR USE IN PREDICTIVE MODELING

By

Sarah Papke

Approved May 11, 2017.

___________________________________________  ____________________________________________
Dr. Frank D’Amico                                Dr. John Kern
Professor of Statistics                          Professor of Statistics
(Committee Chair)                                (Department Chair)

___________________________________________  ____________________________________________
Mr. Sean Tierney                                 Dr. James Swindal
Professor of Statistics                          Dean, McAnulty College and Graduate
(Committee Member)                               School of Liberal Arts
                                                Professor of Philosophy
ABSTRACT

A REVIEW OF ‘BIG DATA’ VARIABLE SELECTION PROCEDURES FOR USE IN PREDICTIVE MODELING

By
Sarah Papke
August 2017

Thesis supervised by Dr. Frank D’Amico.

Several problems arise when attempting to use traditional predictive modeling techniques on ‘big data.’ For instance, multiple linear regression models cannot be used on datasets with hundreds of variables. However several techniques are becoming common tools for selective inference as the need for analyzing big data increases. Forward selection and penalized regression models (such as LASSO, Ridge Regression, and Elastic Net) are simple modifications of multiple linear regression that can provide some guidance on simplifying a model through variable selection. Dimension reducing techniques, such as Partial Least Squares and Principal Components Analysis, are more complex than regression but have the ability to handle highly correlated independent variables. Each of the afore-
mentioned techniques are valuable in predictive modeling if used properly. This paper provides a mathematical introduction to these developments in selective inference. A sample dataset is used to demonstrate modeling and interpretation. Further, the applications to big data, as well as advantages and disadvantages of each procedure, are discussed.
ACKNOWLEDGEMENT

This paper would not have been possible without the help, guidance and support of several people in my life. First I’d like to thank my advisor, Dr. Frank D’Amico, for sharing his knowledge and enthusiasm throughout the research and writing process. I’d also like to thank my committee, Mr. Sean Tierney and Dr. John Kern, for giving me their time and feedback. My CMPA classmates, especially Michael McKibben, Wei Li, and Clayton Elwood, all deserve my endless gratitude: they have made the last two years truly memorable and enjoyable.

My parents, Rich and Sharon Papke, deserve the biggest thanks of all. They have always been supportive of my academic and employment pursuits, no matter how crazy or far from home they take me, and provided me with so much encouragement in my return to Pittsburgh for my Master’s degree. Last, but certainly not least, I’d like to thank my boyfriend Brian, who has been there to make me laugh and give a reality check every time I stressed too much over the last two years.
# TABLE OF CONTENTS

1 Background and Introduction ................................................................. 1  
   1.1 Background ......................................................................................... 1  
   1.2 Introduction ...................................................................................... 2  
   1.3 Basic Definitions .............................................................................. 3  

2 Statistical Procedures Useful for Selective Inference ......................... 5  
   2.1 Multiple Linear Regression and Forward Selection ......................... 5  
      2.1.1 Multiple Linear Regression ...................................................... 5  
      2.1.2 Forward Selection .................................................................. 6  
   2.2 Principal Components Analysis and Partial Least Squares .............. 8  
      2.2.1 Principal Components Analysis ............................................. 8  
      2.2.2 Partial Least Squares .............................................................. 9  
   2.3 Penalized Regression Models .......................................................... 11  
      2.3.1 LASSO ................................................................................. 11  
      2.3.2 Ridge Regression ................................................................. 13  
      2.3.3 Elastic Net ............................................................................ 13  

3 Illustrative Examples .............................................................................. 16  
   3.1 Data Description .............................................................................. 16  
   3.2 Multiple Linear Regression and Forward Selection ....................... 17  
      3.2.1 Multiple Linear Regression ...................................................... 17  
      3.2.2 Forward Stepwise Selection .................................................. 18  
   3.3 Principal Components Analysis and Partial Least Squares ............ 20  
      3.3.1 Principal Components Analysis ............................................. 20  
      3.3.2 Partial Least Squares .............................................................. 25
List of Figures

2.1 Hui Zou: Two-Dimensional Penalized Regression Comparison [14] 14

3.1 JMP: Variable Correlation Table 16

3.2 JMP: ANOVA table, parameter estimates, and summary statistics for multiple linear regression 17

3.3 JMP: Selection options for forward stepwise regression 19

3.4 JMP: ANOVA table, parameter estimates, and summary statistics for stepwise regression 20

3.5 JMP: Eigenvalues for PCA 21

3.6 JMP: Eigenvectors for PCA 21

3.7 JMP: Loading Plot for PCA 22

3.8 JMP: Scree Plot for PCA 23

3.9 JMP: Principal Components Regression 24

3.10 JMP: Root Mean PRESS Plot 25

3.11 JMP: VIP Plot 27

3.12 JMP: VIP vs Coefficients Plot 28

3.13 JMP: Percent Variation Explained Plots 29

3.14 JMP: Model Coefficients for Centered and Scaled Data 30

3.15 JMP: LASSO Solution Path 31
Chapter 1: Background and Introduction

1.1 Background

In the last 20 years, several ‘predictive modeling’ techniques have emerged to address some of the problems in analyzing ‘big data’. The phrase ‘big data’ can be traced back to John Mashey of Silicon Graphics in 1998. He said, “I was using one label for a range of issues, and I wanted the simplest, shortest phrase to convey that the boundaries of computing keep advancing.” [5] The use of big data encompasses ‘Three V’s’ (Volume, Variety, and Velocity), referring to the amount of data, types of data, and speed of data processing. [1] In 2002, the amount of information worldwide stored digitally surpassed analog information storage and by 2009, it was estimated that every US company with more than 1000 employees had at least 200 terabytes of digitally stored data. [6] The increased opportunities of learning from large datasets were accompanied by increased challenges and issues with analyzing the data correctly.

There are many terms (‘machine learning,’ ‘data mining,’ ‘knowledge discovery,’ etc.) used to describe tools that sift through data, with the intention of finding patterns important to a question and provide answers. The aim of all these tools is the same: to make an accurate prediction. This is commonly referred to as ‘predictive modeling.’

In the paper “Statistical Learning and Selective Inference,” Jonathan Taylor and Robert Tibshirani state, “Most statistical analyses involve some kind of ‘selection’ - searching through the data for the strongest associations. Measuring the strength of the resulting associations is a challenging task, because one must account for the
effects of the selection.” [10] They go on to briefly describe several recent developments for exploring these associations in data and issues with selecting variables for inclusion in the predictive model.

The objective of the thesis is to describe and compare each of the proposed modeling techniques discussed by Taylor and Tibshirani. Each procedure will be defined and demonstrated using an example dataset, with an explanation of common problems and corrections. These same procedures will be discussed in the context of ‘big data’ analysis.

1.2 Introduction

Multiple linear regression is a common technique for predictive modeling, as the set of $\hat{\beta}$ coefficients effectively minimize bias, as well as being easy to interpret. In matrix form, the optimal linear regression solution is $(X'X)^{-1}X'y$, where $(X'X)^{-1}$ is proportional to the covariance matrix of the predictors. [4] This inverse matrix poses a problem, however, if any of the $X$ variables are collinear or if there are more $X$ variables than records in the dataset. Big datasets frequently possess one or both of these problem characteristics. Data can be pre-processed, however additional measures may still be required to produce an accurate predictive model. Techniques such as Principal Components Analysis and Partial Least Squares can be employed to address the collinearity issue and provide model dimension reduction. Alternatively, penalized regression models (LASSO, ridge regression, and elastic net) can be used to shrink $\beta$ parameter estimates, a useful feature when the number of variables exceeds the number of records.[4] These dimension reduction and penalized regression techniques will be explored at length in the following chapters. As a reference, the following section is a list of basic definitions that will used throughout the paper.
1.3 Basic Definitions

AICc Akaike Information Criterion (corrected). A value used to compare different models for the same data, calculated by $AICc = -2\hat{L} + 2k + \frac{2k(k+1)}{n-k-1}$ where $\hat{L}$ is the log likelihood, $k$ is the number of estimated parameters, and $n$ is the number of records in the dataset. Generally, a lower AICc indicates a better model fit.

bias The difference between a parameter’s expected value and its true value, present in penalized regression models.

BIC Bayesian Information Criterion. Like AIC (same equation, except it contains an additional term ‘log N’ in its calculation), a value used to compare different models for the same data.

big data A dataset containing, at a minimum, hundreds of records and variables.

cross-validation A process where a dataset is partitioned into $k$ subsets, with an analysis then performed $k$ times. Each time the analysis is performed, a different subset is reserved for use as a validation set while the remaining sets are the training set. Thus each subset will be used in training $k-1$ times and in validation 1 time. After all analyses have been performed, the results of the $k$ analyses are averaged.

dimension reduction Transforming independent variables to a smaller dimensional space.

model complexity The number of independent variable terms present in a model.

multicollinearity Two or more $X$ variables in a predictive model are highly correlated.
penalty parameter An additional term in a model to constrain the independent variable terms.

$R^2$ A goodness-of-fit statistic; values range between 0 and 1, where 1 indicates that the regression line that perfectly fits the data.

regression Determining a functional relationship between a dependent variable and one or more independent variables.

regularization Introducing a penalty parameter to a statistical model in order to prevent model overfitting.

$RMSE$ Root Mean Square Error; A measure of a predictive model’s accuracy.

selective inference Using correlations and patterns between dependent variables to develop a predictive model.

sparse model A number of regression $\beta$ coefficients are equal to 0.

standardized (normalized) data Data is put into the same scale by subtracting the mean from each element and then dividing those results by the standard deviation.

supervised learning Data has an input and a corresponding output; the dependent variable is predetermined, so analysis is ‘supervised’ at each stage by knowing (observing) the y-variable.

unsupervised learning Data has no output; input is categorized or sorted into groups and is used to study the most important contributors among the independent variables.
Chapter 2: Statistical Procedures Useful for Selective Inference

2.1 Multiple Linear Regression and Forward Selection

This chapter provides an overview of the mathematics behind each procedure. It is divided into three sections. The first section outlines multiple linear regression and forward stepwise regression to provide a foundation for comparison. The second section provides an explanation of the dimension reduction techniques, Principal Components Analysis and Partial Least Squares. The third section provides an explanation of penalized regression models, specifically LASSO, ridge regression, and elastic net. These are the recent new developments in selective inference discussed by Taylor and Tibshirani. [10]

2.1.1 Multiple Linear Regression

Traditionally, the most common technique for predictive modeling is multiple linear regression using least squares. Multiple linear regression is a supervised learning approach where several independent $X$ variables are used to predict a single continuous dependent $Y$ variable. The model is represented by the equation:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k + e = \beta_0 + \sum_{i=1}^{n} \beta_ix_i + e \quad (2.1)$$

where $Y$ is the outcome, $\beta_i$ are coefficient parameter estimates for the independent variables, $X_i$ are the independent variables and $e$ is random error. The linear regression model [13] can be written in matrix form as $y = X\beta + e$, where:
The goal is to choose a model such that the sum of squared deviations between the actual outcome $Y$ and the predicted outcome $\hat{Y}$ is minimized. In other words, regression finds a vector $\beta$ such that:

$$
\min \left[ \sum_{i=1}^{N} (y_i - \hat{y})^2 \right] = \min \left[ \sum_{i=1}^{n} e_i^2 \right] = e' e = (2.3)
$$

$$(y - \hat{y})'(y - \hat{y}) = (y - X\beta)'(y - X\beta)$$

where the fitted regression model is $\hat{y} = X\beta$. The minimized $\beta$ vector can be found by the equation

$$
\beta = (X'X)^{-1}X'y
$$

The equation $\beta = (X'X)^{-1}X'y$ will appear again in a modified form in penalized regression models explanation of Section 2.3. The hypotheses for the multiple linear regression model are:

$$
H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0
$$

$$
H_1 : \beta_i \neq 0 \text{ for at least one } i
$$

2.1.2 Forward Selection

Forward selection follows the same process as multiple linear regression. Statistical computing software uses an automated approach to quickly assess potential
regression models corresponding to various subsets of independent variables from a single dataset. In forward selection, independent variables are added to a regression model one at a time, usually from most significant to least significant. There are other methods (such as stepwise and backward elimination) of determining which variable enters the model next, however statistical significance is the usual default method.

After variables are ranked by significance, there are several ways to select a subset of variables to develop a model. For p-value threshold forward selection, all variables added to the model have p-values below the user-defined threshold value. Any variables with p-values greater than the threshold are not included in the model. Alternatives to the p-value threshold selection are AIC and BIC. Both the AIC and BIC are based on the Likelihood function for the model. For both of these measures, a lower number indicates a better model fit; however a better model fit does not necessarily indicate a more accurate or reliable prediction.
2.2 Principal Components Analysis and Partial Least Squares

2.2.1 Principal Components Analysis

Principal Components Analysis (PCA) is an unsupervised learning process which examines the correlation between independent variables and transforms them into a set of uncorrelated component variables. [3] Although PCA removes the variation common to any two variables, it preserves the unshared variation of the individual variables. Each component can be represented by the following equation:

\[ PC_i = w_{i1}x_1 + w_{i2}x_2 + \cdots + w_{ik}x_k \]  

(2.5)

where \( i \) represents the component and \( w_{ij} \) is the weight corresponding to the \( j^{th} \) variable in the corresponding component. The components are computed by eigenvalue analysis.

Begin by creating the correlation matrix of the standardized (normalized) data, then calculate the eigenvectors and corresponding eigenvalues for this matrix. (An eigenvector is a vector \( \mathbf{v} \) such that \( A\mathbf{v} = \lambda\mathbf{v} \), where \( A \) is a square matrix and \( \lambda \) is a constant multiplier called the eigenvalue. There are several linear algebra techniques for calculating eigenvectors and eigenvalues which will not be discussed in the context of this paper.) The eigenvectors are sorted by eigenvalue, from highest to lowest. The eigenvector with the highest eigenvalue becomes the first principal component, the eigenvector with the second highest eigenvalue becomes the second principal component, etc. Because eigenvectors corresponding to distinct eigenvalues are linearly independent, the eigenvectors and components are uncorrelated with each other. Each component explains a percentage of the variability in the \( X \) data, with the first component accounting for the most and the \( k^{th} \) component explaining the
least. One advantage of the PCA is that it can be used with datasets that have high
correlation or a large number of independent variables.

If the original dataset has a dependent variable, the principal components can
then be used in place of the original independent variables in a regression model,
thereby reducing the number of independent variables (complexity) to a fewer num-
ber. However, it is very difficult to analyze the coefficients of this model because each
component depends on several of the original independent variables. The components
are also developed only considering variation among the independent variables and
there is no guarantee that they can be used to accurately predict Y. These issues will
be further discussed at length in Chapter 4.

2.2.2 Partial Least Squares

Partial Least Squares (PLS) is a sister method to PCA that incorporates the de-
pendent variable in the dimension reduction process. There are several algorithms to
perform PLS regression, including methods to regress with a single dependent vari-
able and methods to regress with multiple dependent variables. Similar to PCA, PLS
can be used with datasets that are highly correlated or have a large number of inde-
pendent X variables. PLS can also be used with datasets that have more independent
variables than records or even when the dataset contains several dependent variables.

One commonly used PLS regression method is the Statistically Inspired Modifi-
cation of Partial Least Squares (SIMPLS). This technique is designed for use with
multiple dependent variables. However, SIMPLS can also be used to perform PLS on
data with one dependent variable; in this instance, the algorithm simply terminates
after one iteration and is equivalent to the Nonlinear Iterative Partial Least Squares
algorithm.

This section will focus on the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm, which can be used to perform regression with a single dependent variable. This algorithm was developed in 1966 by Herman Wold for computing PCA and later modified for PLS. It is the most basic of the partial least squares algorithms and also the most computationally expensive algorithm because it iteratively performs a sequence of ordinary least squares regressions. As with PCA, the algorithm produces a set of scores and a set of loadings to use for dimension reduction.

For $m$ independent variables and $n$ total records composing $m \times n$ matrix $X$ and one dependent variable $y$ composing an $n \times 1$ vector, the algorithm is as follows:

**Algorithm 1** NIPALS for Partial Least Squares

1: find latent variable and principal component:

2: $w = X^T y$ ▷ create vector $w$ orthogonal to $X^T y$

3: $w \leftarrow \frac{w}{\|w\|}$ ▷ convert $w$ to a unit vector

4: $t = Xw$ ▷ $t$ is a latent feature of $X$ and $y$

5: $p = \frac{X^T t}{t^T t}$ ▷ find principal component

6: deflate $X$ and $y$:

7: $X \leftarrow X - tp^T$ ▷ residual in $X$ unaccounted for by $tp^T$

8: $\hat{c} = \frac{t^T y}{t^T t}$ ▷ find regression coefficient of $y$ as a function of $t$

9: $y \leftarrow y - t\hat{c}$ ▷ residual in $y$

10: repeat algorithm using deflated $X$ and $y$ to compute the next feature and component

11: continue to apply the algorithm until convergence

Because PLS uses the dependent variable in dimension reduction, the components derived from the algorithm are guaranteed to provide some percentage of explanation of the variation in $Y$. Moreover, PLS is a reliable method to determine high-importance variables in order to develop multiple linear regression or forward selection models.
2.3 Penalized Regression Models

Penalized regression models shrink the size of multiple linear regression coefficients, thus introducing bias, while attempting to improve the prediction. The additional bias attempts to correct for instances where there are more independent variables than records, when variables are highly correlated, or other situations where traditional linear regression is inappropriate. This section is an overview of three types of penalized regression models: LASSO, ridge regression, and elastic net.

These three models are all modifications of multiple linear regression, each with a penalty parameter added to the \((X'X)^{-1}\). In matrix form, this penalty parameter appears in the form of a tuning parameter (called \(\lambda\)) which is added to the diagonal entries of the multiple linear regression estimation for \(\beta\). The tuning parameter is algorithmically chosen by software to produce the ‘best model’, where the user can select criteria such as \(R^2\), training error, or cross-validation to give the software definition of what the ‘best model’ is.

2.3.1 LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) method is a regression technique that performs variable selection by adding penalty, \(\lambda\), to the diagonal entries of the multiple linear regression estimation \(\beta = (X'X)^{-1}X'y\), as shown in the equation below:

\[
\beta_{\text{LASSO}} = (X'X + \lambda I)^{-1}X'y
\] (2.6)
Rather than minimizing the sum of squares as multiple linear regression does, the LASSO regression minimizes the sum of squares plus the penalty parameter:

\[
\min \left[ \sum_{i=1}^{N} (y_{i} - \hat{y})^2 + \lambda \sum_{j} |\beta_j| \right] = (2.7)
\]

\[
\min \left[ \sum_{i=1}^{N} (y_{i} - \beta_0 - \sum_{j} \beta_j x_{ij})^2 + \lambda \sum_{j} |\beta_j| \right]
\]

In standardized form, \( \beta_0 = 0 \), so the equation can be rewritten more simply as

\[
\min \left[ \sum_{i=1}^{N} (y_{i} - \sum_{j} \beta_j x_{ij})^2 + \lambda \sum_{j} |\beta_j| \right] = (2.8)
\]

The term \( \lambda \sum_{j} |\beta_j| \) is referred to as the \( l_1 \) penalty parameter. The \( \lambda \) (where \( \lambda \geq 0 \)) variable is a tuning parameter, which controls the magnitude of the penalty and the amount of shrinkage of the \( \beta \) estimates. It provides a balance between the ‘goodness of fit’ produced by the sum of squared differences and the ‘model complexity’ produced by the sum of absolute regression coefficients. [10] The \( \lambda \) parameter must be greater than or equal to 0. If \( \lambda = 0 \), the penalty term will not be in the model and a multiple linear regression model remains. As \( \lambda \) increases, the independent variables become more constrained. If \( \hat{\beta}_j^0 \) represents the least squares estimates, a value of \( \lambda < \hat{\beta}_j^0 \) will cause \( \beta \) coefficients to shrink toward 0. In some instances, the \( \beta \) coefficients will equal 0, effectively reducing the complexity of the model through variable selection. [11] There are many ways to select the ‘best model,’ which directly affects the value for \( \lambda \). Cross-validation is the most common method because it can be performed with any number of parameters and records.
2.3.2 Ridge Regression

Originally proposed in 1970 by Arthur E. Hoerl and Robert W. Kennard, ridge regression attempts to correct for overfitting through regularization. It takes the standard multiple linear regression estimation \( \beta = (X'X)^{-1}X'y \) and adds a \( \lambda \) constant parameter to the diagonal entries of the \( X'X \) matrix prior to inversion, providing the new estimate:

\[
\beta_{\text{ridge}} = (X'X + \lambda I)^{-1}X'y
\]  

(2.9)

Ridge regression differs from the LASSO only in the penalty parameter: ridge regression uses the \( l_2 \) penalty parameter, which adds the sum of squared \( \beta \)'s rather than the absolute value. While the \( l_1 \) parameter creates a sparse model (a number of \( \beta \) coefficients are equal to 0) through variable selection, the \( l_2 \) parameter provides regularization (\( l_2 \) shrinks the coefficients but none will ever equal 0). The minimization equation looks as such:

\[
\min \left[ \sum_{i=1}^{N} (y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j \beta_j^2 \right]
\]  

(2.10)

An important difference to note between LASSO and ridge regression is that the solution to ridge regression will bound the size of the \( \beta \) parameters but will never produce a \( \beta \) coefficient equal to 0. Similar to LASSO, cross-validation is one of the most common techniques to help select a value for \( \lambda \). [12]

2.3.3 Elastic Net

Elastic Net regularization adds both the \( l_1 \) and \( l_2 \) penalties to the multiple linear regression model. As with both LASSO and ridge regression, \( \lambda \) is added to the diagonal entries of the multiple linear regression estimation \( \beta = (X'X)^{-1}X'y \):

\[
\beta_{\text{elas.net}} = (X'X + \lambda_1 I + \lambda_2 I)^{-1}X'y
\]  

(2.11)
The $\beta$ parameters are now predicted by finding:

$$\min \left[ \sum_{i=1}^{N} (y_i - \sum_j \beta_j x_{ij})^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_j \beta_j^2 \right]$$

(2.12)

Because Elastic Net uses both penalty parameters, the procedure allows for both the variable selection features present in the LASSO and the stability present in ridge regression. Differences between the three penalized regression models can be seen in the graph below:

![2-Dimensional Illustration $\alpha=0.5$](image)

Figure 2.1: Hui Zou: Two-Dimensional Penalized Regression Comparison [14]

In the figure above, the inner diamond depicts the constraint region of a two-dimensional LASSO. When the regression sum of squares is tangent to one of the
corners, the corresponding variable has been shrunk to zero. The two-dimensional depiction of the elastic net constraints is the rounded diamond (the middle shape of figure 2.1). Though the edges are curved, the region also has corners which will indicate that a variable has been shrunk to zero if the sum of squares is tangent to it. The outer circle in the figure is the constraint region of the ridge regression. There are no corners so no matter the constraint, all variables will still be present in the regression.
Chapter 3: Illustrative Examples

3.1 Data Description

The dataset called Bread, from the book Discovering Partial Least Squares with JMP [2], will be used to demonstrate each of the procedures outlined in Chapter 2. This dataset contains 24 records, where each contains data regarding a variety of bread made by a commercial bakery. Each record contains 67 variables: 7 consumer attributes and 60 expert panel sensory attributes. For the purposes of this chapter, only consumer attributes will be used. The consumer attributes represent a cross-sectional study of 50 participants. Each participant rated each of the attributes from 0 to 9, with 9 being the most desirable, for each bread. The participants scores were then averaged and entered into the data as one observation. The attributes are as follows: Overall Liking, Strong Flavor, Gritty, Succulent, Full Bodied, Strong Aroma. The dataset contains no missing observations and all variables are approximately normally distributed. The independent variables have little correlation with each other, as can be seen in the correlation table in Figure 3.1:

<table>
<thead>
<tr>
<th>Correlations</th>
<th>Overall Liking</th>
<th>Dark</th>
<th>Strong Flavor</th>
<th>Gritty</th>
<th>Succulent</th>
<th>Full Bodied</th>
<th>Strong Aroma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Liking</td>
<td>1.0000</td>
<td>0.0649</td>
<td>0.1640</td>
<td>-0.6764</td>
<td>0.4638</td>
<td>-0.5501</td>
<td>0.2921</td>
</tr>
<tr>
<td>Dark</td>
<td>0.0649</td>
<td>1.0000</td>
<td>-0.3672</td>
<td>0.1437</td>
<td>0.2082</td>
<td>0.0882</td>
<td>0.2228</td>
</tr>
<tr>
<td>Strong Flavor</td>
<td>0.1640</td>
<td>-0.3672</td>
<td>1.0000</td>
<td>-0.1685</td>
<td>-0.1172</td>
<td>-0.1548</td>
<td>-0.0463</td>
</tr>
<tr>
<td>Gritty</td>
<td>-0.6764</td>
<td>0.1437</td>
<td>-0.1685</td>
<td>1.0000</td>
<td>-0.2092</td>
<td>0.3758</td>
<td>-0.2974</td>
</tr>
<tr>
<td>Succulent</td>
<td>0.4638</td>
<td>0.2082</td>
<td>-0.1172</td>
<td>-0.2092</td>
<td>1.0000</td>
<td>-0.4311</td>
<td>-0.1674</td>
</tr>
<tr>
<td>Full Bodied</td>
<td>-0.5501</td>
<td>0.0882</td>
<td>-0.1548</td>
<td>0.3758</td>
<td>-0.4311</td>
<td>1.0000</td>
<td>-0.0895</td>
</tr>
<tr>
<td>Strong Aroma</td>
<td>0.2921</td>
<td>0.2228</td>
<td>-0.0463</td>
<td>-0.2974</td>
<td>-0.1674</td>
<td>-0.0895</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Figure 3.1: JMP: Variable Correlation Table

It is important to note that the models in this chapter were not designed with the intention of producing the best fitting model. Rather, the model options and parameters were kept as similar as possible in order to provide a foundation for comparison.
Also to provide consistency, all independent variables have been standardized by subtracting the mean of each variable and dividing by the standard deviation. All output in this chapter is produced using JMP® statistical software. The full output for each analysis is provided in Appendix B.

### 3.2 Multiple Linear Regression and Forward Selection

#### 3.2.1 Multiple Linear Regression

The multiple linear regression model will serve as the basis for comparison. The dependent variable is the consumer attribute *Overall Liking* and the independent variables are the remaining six consumer variables: *Dark, Strong Flavor, Gritty, Succulent, Full Bodied, Strong Aroma*. Output of multiple linear regression is as follows:

![Figure 3.2: JMP: ANOVA table, parameter estimates, and summary statistics for multiple linear regression](image)

**Figure 3.2:** JMP: ANOVA table, parameter estimates, and summary statistics for multiple linear regression
The model in figure 3.2 uses all six independent variables without selection. The F Ratio of $F_{6,17} = 5.3536$ indicates that the null hypothesis is rejected and at least one of the parameter estimates is not equal to 0. Looking at the $t$ Ratios and $p$-values for the parameter estimates, Gritty is the strongest predictor in the model, while the others provide little contribution. These estimates are computed by Type III sums of squares, the most conservative type, where sums of squares are calculated for each variable as if it were the last to enter the model. Other important statistics to note are the $R^2_{Adj}$ and the Root Mean Square Error (RMSE), as these values will be used to compare subsequent models.

3.2.2 Forward Stepwise Selection

Because many of the variables were non-significant in the multiple linear regression model, forward selection can be used to simplify the model by examining variables in order of decreasing contributions to design a subset of significant variables. Several options are available to create a model using forward selection, such as $R^2$, $RMSE$, $AIC$, or $BIC$. Figure 3.3 below shows the solution path of forward selection using ‘AICc’.
To provide consistency with later models, the variables will be chosen by examining the Akaike Information Criterion (AIC) value. AIC provides a relative indication of model quality, with lower values implying a more accurate model. The lowest AIC in the forward selection scenario occurs when only two variables, *Gritty* and *Succulent* are added to the model.
Figure 3.4: JMP: ANOVA table, parameter estimates, and summary statistics for stepwise regression

The summary statistics from the forward selection model in figure 3.4, specifically $r^2$ and RMSE, differ little from those in the multiple linear regression model. Although there is no improvement in model accuracy, there is considerable improvement in model simplicity.

### 3.3 Principal Components Analysis and Partial Least Squares

#### 3.3.1 Principal Components Analysis

Principal Components Analysis (PCA) provides an alternative to forward selection to simplify the model through dimension reduction of the independent variables. Prior to running PCA, all variables were standardized. In the output, first consider the eigenvalues and cumulative component percent, displayed in the figure below:
The percent column (calculated by dividing each eigenvalue by the total sum of all eigenvalues) indicates what portion of the variance is explained by the corresponding component. The eigenvalues represent the strength of the component; in this instance, the first three components account for nearly 75% of the total variation in $X$. Because the goal of PCA is dimension reduction, ideally at least 90% of the total variation could be explained by a small number of PCA variables. To see which variables contribute to each component, consider the eigenvectors:

*Gritty*, *Full Bodied*, and *Succulent* are the strongest contributors to the first component, which is consistent with the forward selection. Similar information can be determined from the Loading Plot, which plots the first principal component eigenvectors as horizontal coordinates and the second principal component eigenvectors as vertical coordinates.
Variables closer to the outer circle are stronger contributors to the respective component, horizontal for component 1 and vertical for component 2. From the above loading plot, the horizontal orientation of *Gritty* (coordinates: (0.57903, 0.00431)) shows that it contributes highly to component 1 and does not contribute to component 2, whereas *Succulent* (coordinates: (-0.40201, 0.41368)) appears to contribute equally to both components but in opposite directions because of its position in the second quadrant.

The principal components can be used as independent variables in a regression to model *Overall Liking*. The eigenvalues and scree plot, displayed in the figure below, can be used to determine the appropriate number of components to model in Principal Components Regression.

Figure 3.7: JMP: Loading Plot for PCA
An ideal scree plot would have a strong elbow- with early components decreasing quickly followed by latter components decreasing slowly. This strong elbow would show that the least important principal components all have low (and similar) eigenvalues and are good candidates to remove from the analysis. However, the scree plot above does not fit the description of an ideal scree plot, as it is mostly linear and does not have a strong elbow. Thus it is difficult to determine from the scree plot the ideal number of components in the Bread PCA.

*Side Note:* After performing a PCA, the components can be used as the independent variables in a standard multiple linear regression model. This technique is called Principal Components Regression (PCR). To create a model that accounts for at least 90% of the variation in the $X$’s, Figure 3.5 indicates five variables must be included in the model. While it is ideal to reduce the model by more than one variable, it is
not possible in this example because the original Bread dataset independent variables were uncorrelated. The regression output modeling *Overall Liking* is displayed below in figure 3.9:

![Figure 3.9: JMP: Principal Components Regression](image)

The $R^2$, *RMSE*, and parameter estimate for intercept are not much different the multiple linear and forward selection models from earlier. Also, the t-Ratio in the parameter estimates indicates that only the first principal component (Prin1) is significant in the model. Interpreting the regression model is much more difficult than assessing the goodness of fit because each independent variable is now comprised of a combination of the original variables. As mentioned in chapter 2, there are lengthy formulae for converting the components parameter estimates into estimates for the original variables. However, PCA is a better tool for independent variable analysis rather than regression. A dimension reduction substitute for PCR that also considers the dependent variable is the Partial Least Squares regression.
### 3.3.2 Partial Least Squares

Like PCA, the Partial Least Squares (PLS) model is run using standardized variables. The PLS model was created with the NIPALS specification using Leave-One-Out validation method \(^1\) (AIC is not an option for PLS) and maximum initial number of factors (which is equal to the number of independent variables, in this instance, 6). To determine the appropriate number of factors for the final model, the Root Mean PRESS Plot can be used:

![Root Mean PRESS Plot](image)

**Figure 3.10: JMP: Root Mean PRESS Plot**

Root mean PRESS (predicted residual sum of squares) is a plot of the residuals corresponding to each number of factors (from 0 through the number of factors specified prior to running the model). For this data, the best PLS model occurs when only

---

\(^1\) Leave-One-Out validation is a version of K-fold cross-validation, where \(K = n\)
one factor is included, as it gives a root mean PRESS of 0.77577. PLS models with
two or more factors have more than 0.85, as can be seen in the root mean PRESS plot.

Another option for selecting an appropriate number of factors is to use the ‘van
der Voet $T^2$’ in combination with root mean PRESS. The van der Voet $T^s$ number
is a test statistic to determine if two models are differ significantly. [9] The number
of factors with the lowest root mean PRESS has a van der Voet $T^2$ p-value of 1.0.
A model with fewer factors than the one with the lowest root mean PRESS can be
selected if the van der Voet $T^2$ indicates the model is not significantly different (a
common practice is to use a p-value greater than 0.1, as the van der Voet $T^2$ null
hypothesis is not rejected). [9]

After selecting the number of NIPALS factors, there are a number of interesting
and useful plots to use for analysis. One such plot is the Variable Importance Plot
(VIP), shown in figure 3.11 below.
The VIP plot is a depiction of the VIP scores, which are a measure of each variable’s importance in measuring both $X$ and $Y$. Typically 0.8 is used as the threshold for importance. Often, a small coefficient and a small VIP score is a good indication that a variable can be removed from the model. [9] This relationship can be seen in the ‘VIP vs Coefficients’ plot in figure 3.12. The VIP vs Coefficients plot includes the threshold line at 0.8, along with variables plotted as VIP (from figure 3.11) against coefficient value.
Variables below the threshold and near the center of the graph (in the above plot, the variables Dark, Strong Flavor, and Strong Aroma) are weak predictors of the dependent variable. The variables above the threshold can be used to develop multiple linear regression model, either directly or through additional forward selection.

An additional useful statistic is the percent variation explained by the factors and the Y variable response. In this PLS model, one factor explains 28.6% of the variation in X and 64% of the variation in Y, as shown in figure 3.13. In datasets larger than the Bread dataset, the percent variation can be used in the process to decide how many factors to model.

Figure 3.12: JMP: VIP vs Coefficients Plot
After examining the various statistics produced by the PLS, a predictive model can be constructed using variable selection from the PLS output or by using the PLS centered and scaled model coefficients, displayed in figure 3.14.
The centered and scaled output produces an intercept of 0 and comparatively large coefficients for *Gritty*, *Succulent*, and *FullBodied*. Both PCA and PLS provide procedures to produce information to reduce the dimension of a model but at the cost of a somewhat complex interpretation of the results.

### 3.4 Penalized Regression Models

#### 3.4.1 LASSO

As with PCA, LASSO runs on standardized independent variables. Cross-validation is the most common technique for finding the best $\lambda$ value. However, for consistency and to compare results with previous models, this LASSO example was chosen using the AICc validation method. The following graph shows the AICc plotted against the magnitude of scaled parameters present in the model.
The lowest two AICc values occur when the parameters are scaled by a factor of three or by a factor of four. Because the objective is to simplify the model, select the option with three parameters. The estimates are displayed in figure 3.16 below:

In the LASSO model, *Gritty*, *Succulent*, and *Full Bodied* were all included in the model, yet only *Gritty* is significant predictor. The $\beta$ coefficients for *Dark*, *Strong Flavor*, and *Strong Aroma* have all been reduced to $\beta_i = 0$. The intercept is similar to that of other models at approximately 6.4. The model summary in figure 3.16 below
details other summary statistics for the model. The AICc is close to AICc used in the forward selection model and the generalized $r^2$ is also comparable to the $r^2$ from other models. Overall, the LASSO model provides very similar results to previous models with the luxury of fewer variables.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of rows</td>
<td>24</td>
</tr>
<tr>
<td>Sum of Frequencies</td>
<td>24</td>
</tr>
<tr>
<td>-LogLikelihood</td>
<td>22.678369</td>
</tr>
<tr>
<td>BIC</td>
<td>61.247007</td>
</tr>
<tr>
<td>AICc</td>
<td>58.690071</td>
</tr>
<tr>
<td>Generalized RSquare</td>
<td>0.5496047</td>
</tr>
</tbody>
</table>

Figure 3.17: JMP: LASSO Model Summary

3.4.2 Ridge Regression

While LASSO is able to perform selection by shrinking some of the $X$ coefficients to 0, ridge regression is only able to perform shrinkage and not selection. The solution path highlights the ideal solution where the magnitude of the scaled parameters minimizes the negative log likelihood (which is effectively maximizing likelihood). The solution path is displayed in figure 3.18 below:
The parameters corresponding to this solution path are in 3.19. Because ridge regression does not select variables, all six are still present in the model. However comparing the coefficients to the multiple linear regression model, it can be observed that the coefficients for all the parameters, while consistent in direction, are all smaller in magnitude in the ridge model.

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Wald ChiSquare</th>
<th>Prob &gt; ChiSquare</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.4084969</td>
<td>0.1234964</td>
<td>2692.7961</td>
<td>&lt;.0001*</td>
<td>6.166483</td>
<td>6.6505455</td>
</tr>
<tr>
<td>Standardize[Dark]</td>
<td>0.0410358</td>
<td>0.1297863</td>
<td>0.0999696</td>
<td>0.7519</td>
<td>-0.213341</td>
<td>0.295423</td>
</tr>
<tr>
<td>Standardize[Strong Flavor]</td>
<td>0.0636315</td>
<td>0.1577309</td>
<td>0.1627463</td>
<td>0.6866</td>
<td>-0.245515</td>
<td>0.372784</td>
</tr>
<tr>
<td>Standardize[Gritty]</td>
<td>-0.242061</td>
<td>0.3449763</td>
<td>0.4923449</td>
<td>0.4829</td>
<td>-0.918202</td>
<td>0.4340806</td>
</tr>
<tr>
<td>Standardize[Succulent]</td>
<td>0.1588534</td>
<td>0.211529</td>
<td>0.5639668</td>
<td>0.4527</td>
<td>-0.255736</td>
<td>0.573426</td>
</tr>
<tr>
<td>Standardize[Full Bodied]</td>
<td>-0.157796</td>
<td>0.213912</td>
<td>0.544156</td>
<td>0.4607</td>
<td>-0.577056</td>
<td>0.264636</td>
</tr>
<tr>
<td>Standardize[Strong Aroma]</td>
<td>0.1076394</td>
<td>0.1873408</td>
<td>0.331352</td>
<td>0.5649</td>
<td>-0.259342</td>
<td>0.475026</td>
</tr>
<tr>
<td>Scale</td>
<td>0.634028</td>
<td>0.2871325</td>
<td>4.8758714</td>
<td>0.0272*</td>
<td>0.0712587</td>
<td>1.1967972</td>
</tr>
</tbody>
</table>
3.4.3 Elastic Net

The output for Elastic Net regression provides similar graphs and tables to that of the LASSO. The Solution Path in figure 3.20 below shows how the magnitude of scaled parameter estimates affects the AICc, which was the selected validation technique for this model.

![Figure 3.20: JMP: Elastic Net Solution Path](image)

The elastic net regression performs variable selection, and two variables shrink to 0. Looking at the parameter estimates of the remaining four variables, displayed in figure 3.21 below, *Gritty* is the only significant variable. The intercept is also similar to other models, with a value of approximately 6.4.
The model summary in figure 3.22 provides other statistics to compare the elastic net to other models. The AICc is similar to other models, as is the $r^2$ value.

![Parameter Estimates for Original Predictors](image1)

**Figure 3.21: JMP: Elastic Net Parameter Estimates**

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Wald ChiSquare</th>
<th>Prob &gt; ChiSquare</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.3922793</td>
<td>0.1182479</td>
<td>2922.2987</td>
<td>&lt;.0001*</td>
<td>6.1605176</td>
<td>6.6240409</td>
</tr>
<tr>
<td>Standardize[Dark]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Standardize[Strong Flavor]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Standardize[Gritty]</td>
<td>-0.401496</td>
<td>0.1341886</td>
<td>8.9521952</td>
<td>0.0028*</td>
<td>-0.654501</td>
<td>-0.138491</td>
</tr>
<tr>
<td>Standardize[Succulent]</td>
<td>0.1950946</td>
<td>0.133581</td>
<td>2.1330501</td>
<td>0.1442</td>
<td>-0.206719</td>
<td>0.456085</td>
</tr>
<tr>
<td>Standardize[Full Bodied]</td>
<td>-0.184264</td>
<td>0.149783</td>
<td>1.5134063</td>
<td>0.2186</td>
<td>-0.477833</td>
<td>0.109305</td>
</tr>
<tr>
<td>Standardize[Strong Aroma]</td>
<td>0.0836552</td>
<td>0.1766015</td>
<td>0.8843866</td>
<td>0.3537</td>
<td>-0.429878</td>
<td>0.429878</td>
</tr>
<tr>
<td>Scale</td>
<td>0.579294</td>
<td>0.1034042</td>
<td>31.349874</td>
<td>&lt;.0001*</td>
<td>0.376256</td>
<td>0.7819625</td>
</tr>
</tbody>
</table>

![Measure Train](image2)

**Figure 3.22: JMP: Elastic Net Model Summary**

### 3.5 Summary of Illustrative Examples

The different selection and regularization techniques demonstrated in this chapter provide a brief example of how to interpret results. As can be seen from the parameter and model summary outputs, the parameter estimates RMSE, and $r^2$ values are similar. Thus for the **Bread** dataset, none of these techniques are statistically preferable and simplicity should take precedence. However, there are several advantages and disadvantages to each method, discussed in Chapter 4, that can provide some indication as to which method to use for a larger dataset.
Chapter 4: Problems and Issues with Each Approach

4.1 Multiple Linear Regression and Forward Selection

While multiple linear regression has a lengthy list of issues, the most significant in terms of this paper is that it does not translate well to large datasets. It is impractical, and often mathematically impossible with regards to degrees of freedom, to use all independent variables and necessary interactions. A large number of independent variables also increases the risk of overfitting the model, where the model accurately reflects the patterns in the data but cannot accurately predict out-of-sample data. While there are many solutions to the issues with multiple linear regression, the most direct is to use a variable selection technique.

Forward selection models are also prone to overfitting the data. To minimize the risk of overfitting for both multiple linear regression and forward selection, models can be constructed using training and validation sets to assess the model’s predictive power. Another issue with forward selection is that the ‘mathematically best’ model could be due to chance.[7] Forward selection is an automated process that does not involve human judgment, so multicollinearity is another problem that can arise. However if correlations are examined prior to running the forward selection, it is simple to manually bypass any models that include highly correlated X variables.

4.2 PCA and PLS

There are two large problems with performing a Principal Components Analysis. The first is that PCA is an unsupervised learning technique. If the only objective is to strictly examine independent variables, unsupervised learning presents no issue.
However, if PCA will later be used as a predictive model, there is no guarantee that
the components chosen to explain the variance in $X$ will also explain the variance
in $Y$. The second issue that arises with PCA is how to interpret the results of
a regression with the components used as independent variables. The parameter
coefficients output by a PCR do not directly translate back to the original variables.
While formulae can be derived to interpret the coefficients, it is recommended to use
PCA as a technique to examine correlations among the independent variables.

Because PLS is an iterative ordinary least squares sequence, it can be performed
on matrices that have missing data. It can also be used on large datasets, especially
performing well when there are more independent variables than records. However
partial least squares cannot be used on binary, nominal, or ordinal data, nor can it
predict non-linear relationships. [8] PLS has several different algorithms to choose
from, so it requires more initial background knowledge than other predictive modeling
techniques. However, one of the PLS algorithms (SIMPLS) is designed specifically to
predict multiple dependent variables.

### 4.3 Penalized Regression Models

Penalized regression models, especially the LASSO and elastic net as they both
perform variable selection, are suitable to use when the number of independent vari-
ables exceeds the requirement. However, LASSO will only select at most $n$ variables,
where $n$ is equal to the number of records in the dataset. LASSO also fails to do
grouped selection - if variables are grouped, the LASSO will often only select one
variable from the group and ignore the others. Penalized regression models are eas-
ier to interpret than PCA and PLS because they are the most similar to standard
multiple linear regression. However penalized regression also lowers all $\beta$ coefficients
through the use of the penalty parameter, so interpretations must also account for
this resulting bias. The tuning parameter can also create some issues for penalized
regression models. Selecting a tuning parameter that is too small will lead to overfit data, resulting in a model with high variance, and one that is too large will lead to underfit data. For a large dataset, ridge regression is unable to perform variable selection, as it does not shrink variables to 0.
Chapter 5: Discussion

Any of the predictive modeling techniques described in this paper can be applied in the context of big data. However, depending on the raw data and the goal of the analysis, some methods may be more appropriate than others. The most important starting point is to examine the raw data carefully to assist in the decision-making process. Handling missing data and studying the pattern of missing data are essential precursory steps for any modeling. After looking at the data and undergoing initial data preparation steps (such as handling missing data, applying transformations, and standardizing variables), selective inference and dimension reduction techniques can be used independently or as supplements to each other as prediction tools. Table 5.1 briefly highlights each of the procedures.

Cross-validation has not been extensively discussed in the context of this paper. When working with big data, it is useful to partition the data into training and validation sets in order to verify, and if necessary, to improve upon the chosen prediction technique. In the context of selective inference, cross-validation can ensure that a reduced model is providing an accurate prediction because there is true correlation between the variables rather than random chance.

After randomly partitioning data, the training set can be used to develop a model using one of the selective inference techniques discussed previously. Though big datasets often have many issues (missing data, correlation, size, etc.), considering dataset size and correlation can help narrow down selective inference techniques to try in the modeling process.

‘How many potential independent variables are in the dataset?’ Multiple linear
regression, forward selection, or ridge regression are simple in both computation and interpretation and could be reasonable options for predictive modeling. The definition of ‘big data’ implies that there will generally be a large number of independent variables (maybe more independent variables than observations), so these are unlikely options for a final predictive model. A large number of independent variables indicates a model reduction process could be necessary; forward selection, PCA/PLS, or a penalized regression model (specifically LASSO and elastic net) are options to explore.

‘Are the independent variables correlated?’ As seen with the uncorrelated Bread dataset in Chapter 3, PCA and PLS do not provide aid in dimension reduction if the variables are not highly correlated, as they are unable to group correlated variables into a single component. However, correlated independent variables can be examined using PCA. While it does not involve the relationship to the $Y$ variable, PCA is a useful tool to initially explore correlations in more detail. PLS can be used for both study and predictive modeling, as it incorporates the dependent variable into the factor computation.

This thesis illustrated how various (currently popular) predictive modeling techniques can be applied to ‘big datasets’ for variable selection. There are other techniques (neural networks, decision trees, and naive Bayes, to name a few) also gaining support. Some of these procedures provide automatic feature selection, have multiple tuning parameters, or may be more robust to variable selection. The field of ‘data analytics’ is expanding at a great rate, and as more technology is developed, the list of prediction techniques will continue to improve.
### Summary of Advantages and Disadvantages for Each Selective Inference Technique Applied to Big Data Analysis

#### Multiple Linear Regression
- Mathematically simple
- Easy to interpret coefficients
- Impractical for big data
- Does not account for correlation

#### Forward Selection
- Simplified model
- Prone to overfitting

#### Principal Components Analysis
- Correlated variables allowed
- Large datasets possible
- Does not account for $Y$ variable
- Difficult to interpret components

#### Partial Least Squares
- Large datasets or missing data possible
- Multiple dependent variables possible
- Complex interpretation
- No binary, nominal, or ordinal data

#### LASSO Regression
- Performs variable selection
- Easy to interpret $\beta$ coefficients
- Selects at most $n$ variables
- Account for bias in interpretation

#### Ridge Regression
- Provides regularization
- Does not reduce the model

#### Elastic Net Regression
- Balances selection and regularization
- Account for penalties in interpretation

Table 5.1: Summary of Selective Inference Techniques
Appendix A: Original Data and Distributions

<table>
<thead>
<tr>
<th>Overall Liking</th>
<th>Dark</th>
<th>Strong Flavor</th>
<th>Gritty</th>
<th>Succulent</th>
<th>Full Bodied</th>
<th>Strong Aroma</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.098</td>
<td>4.674</td>
<td>5.684</td>
<td>4.457</td>
<td>6.038</td>
<td>5.113</td>
<td>4.841</td>
</tr>
<tr>
<td>6.062</td>
<td>4.703</td>
<td>5.821</td>
<td>3.492</td>
<td>5.116</td>
<td>5.254</td>
<td>4.904</td>
</tr>
<tr>
<td>4.772</td>
<td>5.184</td>
<td>5.770</td>
<td>4.855</td>
<td>5.124</td>
<td>5.471</td>
<td>4.842</td>
</tr>
<tr>
<td>6.150</td>
<td>5.420</td>
<td>6.962</td>
<td>3.446</td>
<td>5.536</td>
<td>4.989</td>
<td>5.200</td>
</tr>
<tr>
<td>7.409</td>
<td>5.733</td>
<td>5.515</td>
<td>2.804</td>
<td>5.649</td>
<td>5.253</td>
<td>5.195</td>
</tr>
<tr>
<td>5.950</td>
<td>6.054</td>
<td>5.524</td>
<td>3.871</td>
<td>5.514</td>
<td>5.939</td>
<td>4.727</td>
</tr>
<tr>
<td>7.020</td>
<td>6.193</td>
<td>5.661</td>
<td>3.137</td>
<td>5.508</td>
<td>5.035</td>
<td>5.304</td>
</tr>
<tr>
<td>6.245</td>
<td>5.191</td>
<td>4.930</td>
<td>3.348</td>
<td>5.459</td>
<td>5.141</td>
<td>4.939</td>
</tr>
<tr>
<td>6.237</td>
<td>6.185</td>
<td>5.159</td>
<td>5.255</td>
<td>5.465</td>
<td>5.495</td>
<td>4.857</td>
</tr>
<tr>
<td>6.205</td>
<td>4.901</td>
<td>5.349</td>
<td>3.748</td>
<td>5.077</td>
<td>5.535</td>
<td>5.117</td>
</tr>
<tr>
<td>8.466</td>
<td>5.251</td>
<td>6.473</td>
<td>2.273</td>
<td>5.144</td>
<td>5.071</td>
<td>5.639</td>
</tr>
<tr>
<td>5.656</td>
<td>6.008</td>
<td>4.565</td>
<td>4.836</td>
<td>5.370</td>
<td>5.095</td>
<td>5.165</td>
</tr>
<tr>
<td>7.908</td>
<td>6.219</td>
<td>5.276</td>
<td>3.802</td>
<td>5.614</td>
<td>4.934</td>
<td>5.810</td>
</tr>
<tr>
<td>6.039</td>
<td>5.560</td>
<td>5.612</td>
<td>3.705</td>
<td>5.766</td>
<td>5.409</td>
<td>5.604</td>
</tr>
<tr>
<td>5.283</td>
<td>6.012</td>
<td>5.578</td>
<td>5.130</td>
<td>4.964</td>
<td>5.321</td>
<td>5.032</td>
</tr>
<tr>
<td>4.926</td>
<td>5.722</td>
<td>5.517</td>
<td>5.747</td>
<td>4.757</td>
<td>5.891</td>
<td>5.111</td>
</tr>
<tr>
<td>6.413</td>
<td>4.948</td>
<td>5.856</td>
<td>3.728</td>
<td>5.195</td>
<td>5.171</td>
<td>4.609</td>
</tr>
<tr>
<td>5.602</td>
<td>5.167</td>
<td>6.308</td>
<td>4.959</td>
<td>4.591</td>
<td>5.482</td>
<td>5.454</td>
</tr>
<tr>
<td>5.473</td>
<td>6.030</td>
<td>6.075</td>
<td>4.063</td>
<td>5.128</td>
<td>5.283</td>
<td>4.621</td>
</tr>
<tr>
<td>6.945</td>
<td>5.506</td>
<td>5.355</td>
<td>3.467</td>
<td>5.552</td>
<td>5.584</td>
<td>5.016</td>
</tr>
</tbody>
</table>

Table A.1: Original Dataset from Discovering Partial Least Squares with JMP [2]
Original Data Distributions

Figure A.1: JMP: Distributions of Original Data

Figure A.2: JMP: Distributions of Original Data
Appendix B: Full JMP Outputs for All Models

Multiple Linear Regression

Figure B.1: JMP: Multiple Linear Regression Report
### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>13.502812</td>
<td>2.25047</td>
<td>5.3536</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>17</td>
<td>7.146263</td>
<td>0.42037</td>
<td></td>
<td>0.0029*</td>
</tr>
<tr>
<td>C. Total</td>
<td>23</td>
<td>20.649075</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Parameter Estimates

| Term            | Estimate | Std Error | t Ratio | Prob>|t| |
|-----------------|----------|-----------|---------|-----|---|
| Intercept       | 6.3922793| 0.132346  | 48.30   | <.0001*|
| Standardize[Dark]| 0.0989273| 0.16008   | 0.62   | 0.5448|
| Standardize[Strong Flavor]| 0.1221643| 0.148605  | 0.82   | 0.4224|
| Standardize[Gritty]| -0.453268| 0.160127  | -2.83  | 0.0115*|
| Standardize[Succulent]| 0.2738622| 0.169716  | 1.61   | 0.1250|
| Standardize[Full Bodied]| -0.208954| 0.163553  | -1.28  | 0.2186|
| Standardize[Strong Aroma]| 0.152713| 0.15871   | 0.96   | 0.3494|

### Sorted Parameter Estimates

| Term            | Estimate | Std Error | t Ratio | Prob>|t| |
|-----------------|----------|-----------|---------|-----|---|
| Standardize[Gritty]| -0.453268| 0.160127  | -2.83  | 0.0115*|
| Standardize[Succulent]| 0.2738622| 0.169716  | 1.61   | 0.1250|
| Standardize[Full Bodied]| -0.208954| 0.163553  | -1.28  | 0.2186|
| Standardize[Strong Aroma]| 0.152713| 0.15871   | 0.96   | 0.3494|
| Standardize[Strong Flavor]| 0.1221643| 0.148605  | 0.82   | 0.4224|
| Standardize[Dark]| 0.0989273| 0.16008   | 0.62   | 0.5448|

Figure B.2: JMP: Multiple Linear Regression Report
Forward Selection

<table>
<thead>
<tr>
<th>Measure</th>
<th>Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of rows</td>
<td>24</td>
</tr>
<tr>
<td>Sum of Frequencies</td>
<td>24</td>
</tr>
<tr>
<td>-LogLikelihood</td>
<td>22.230301</td>
</tr>
<tr>
<td>BIC</td>
<td>57.175918</td>
</tr>
<tr>
<td>AICc</td>
<td>54.059906</td>
</tr>
<tr>
<td>Generalized R²</td>
<td>0.5641155</td>
</tr>
</tbody>
</table>

**Solution Path**

Figure B.3: JMP: Forward Selection Report
### Fit Group

**Response Overall Liking**

**Effect Summary**

<table>
<thead>
<tr>
<th>Source</th>
<th>LogWorth</th>
<th>PValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardize[Gritty]</td>
<td>3.313</td>
<td>0.00049</td>
</tr>
<tr>
<td>Standardize[Succulent]</td>
<td>1.491</td>
<td>0.03226</td>
</tr>
</tbody>
</table>

**Summary of Fit**

- RSquare: 0.566116
- RSquare Adj: 0.524793
- Root Mean Square Error: 0.653172
- Mean of Response: 6.392279
- Observations (or Sum Wgts): 24

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>11.689752</td>
<td>5.84488</td>
<td>13.7000</td>
</tr>
<tr>
<td>Error</td>
<td>21</td>
<td>8.959313</td>
<td>0.42663</td>
<td>Prob &gt; F</td>
</tr>
<tr>
<td>C. Total</td>
<td>23</td>
<td>20.649075</td>
<td>0.42663</td>
<td>0.0002*</td>
</tr>
</tbody>
</table>

**Parameter Estimates**

| Term               | Estimate   | Std Error | t Ratio | Prob > |t| |
|--------------------|------------|-----------|---------|--------|-----|
| Intercept          | 6.3922793  | 0.133328  | 47.94   | <.0001*|
| Standardize[Gritty]| -0.574042  | 0.139278  | -4.12   | 0.0005*|
| Standardize[Succulent]| 0.3193865 | 0.139278  | 2.29    | 0.0323*|

Figure B.4: JMP: Forward Selection Report
Principal Components Analysis

Figure B.5: JMP: Principal Components Analysis Report
Figure B.6: JMP: Principal Components Analysis Report
Partial Least Squares

Figure B.7: JMP: Partial Least Squares Report

Figure B.8: JMP: Partial Least Squares Report
Figure B.9: JMP: Partial Least Squares Report

Figure B.10: JMP: Partial Least Squares Report

Figure B.11: JMP: Partial Least Squares Report
Figure B.12: JMP: Partial Least Squares Report

![Model Coefficients for Original Data](image)

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Overall Liking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.3923</td>
</tr>
<tr>
<td>Standardize[Dunk]</td>
<td>0.0381</td>
</tr>
<tr>
<td>Standardize[Strong Flavor]</td>
<td>0.0911</td>
</tr>
<tr>
<td>Standardize[Gritty]</td>
<td>-0.0378</td>
</tr>
<tr>
<td>Standardize[Succulent]</td>
<td>0.2578</td>
</tr>
<tr>
<td>Standardize[Full Bodied]</td>
<td>-0.3058</td>
</tr>
<tr>
<td>Standardize[Strong Aroma]</td>
<td>0.1624</td>
</tr>
</tbody>
</table>

Figure B.13: JMP: Partial Least Squares Report

![Variable Importance Plot](image)
Figure B.14: JMP: Partial Least Squares Report

Figure B.15: JMP: Partial Least Squares Report
Figure B.16: JMP: Partial Least Squares Report

LASSO Regression

Figure B.17: JMP: LASSO Regression Report
Figure B.18: JMP: LASSO Regression Report

Ridge Regression

Figure B.19: JMP: Ridge Regression Report
Elastic Net Regression
<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Chisquare</th>
<th>Prob $&gt;\text{Chisquare}$</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.3922793</td>
<td>0.1180479</td>
<td>2522.2887</td>
<td>$&lt;.0001^*$</td>
<td>6.1605176</td>
<td>6.6240409</td>
</tr>
<tr>
<td>Standardize[Dark]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Standardize[Strong Flavor]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Standardize[Gritty]</td>
<td>-0.404486</td>
<td>0.1341088</td>
<td>8.9521952</td>
<td>$0.0028^*$</td>
<td>-0.554501</td>
<td>-0.254084</td>
</tr>
<tr>
<td>Standardize[Succulent]</td>
<td>0.1956946</td>
<td>0.133581</td>
<td>2.1330501</td>
<td>0.1442</td>
<td>-0.065719</td>
<td>0.456085</td>
</tr>
<tr>
<td>Standardize[Full Bodied]</td>
<td>-0.184264</td>
<td>0.149763</td>
<td>1.5134063</td>
<td>0.2186</td>
<td>-0.477833</td>
<td>0.1091054</td>
</tr>
<tr>
<td>Standardize[Strong Aroma]</td>
<td>0.0836552</td>
<td>0.1786013</td>
<td>0.2248866</td>
<td>0.6357</td>
<td>-0.262477</td>
<td>0.4297878</td>
</tr>
<tr>
<td>Scale</td>
<td>0.575294</td>
<td>0.1034042</td>
<td>31.384984</td>
<td>$&lt;.0001^*$</td>
<td>0.3766256</td>
<td>0.7819625</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>Df</th>
<th>Chisquare</th>
<th>Prob $&gt;\text{Chisquare}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardize[Gritty]</td>
<td>1</td>
<td>1</td>
<td>8.9521952</td>
<td>$0.0028^*$</td>
</tr>
<tr>
<td>Standardize[Succulent]</td>
<td>1</td>
<td>1</td>
<td>2.1330501</td>
<td>0.1442</td>
</tr>
<tr>
<td>Standardize[Full Bodied]</td>
<td>1</td>
<td>1</td>
<td>1.5134063</td>
<td>0.2186</td>
</tr>
<tr>
<td>Standardize[Strong Aroma]</td>
<td>1</td>
<td>1</td>
<td>0.2248866</td>
<td>0.6357</td>
</tr>
<tr>
<td>Standardize[Dark]</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1.0000 Removed</td>
</tr>
<tr>
<td>Standardize[Strong Flavor]</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1.0000 Removed</td>
</tr>
</tbody>
</table>
References


