Selected Projects

Stephen G. Merrill

A project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Dr. Brian Hartman, Chair
Dr. C. Shane Reese
Dr. Scott D. Grimshaw

Department of Statistics

Brigham Young University

April 2017

ACKNOWLEDGMENTS

I would like to thank several parties that were integral in helping make this accomplishment possible. First and foremost, my parents, who offered me loving and tireless support of every kind. Second, those in and around the statistics department. Faculty, students, and especially friends combined to be a vital resource for me as I gained knowledge and made my way through the program. Thank you to all that played a part in this achievement. Heaven knows I needed the help, and you all selflessly provided it.

Table of Contents

Part I

Defended Projects

Modeling NBA Shooting Efficiency

Stephen Merrill

Brigham Young University

Abstract. In the ultra-competitive zero sum game otherwise known as the National Basketball Association, advantage lies in maximizing offensive efficiency. This analysis proposes a Bayesian hierarchical latent variable probit model in order to model the efficiency of NBA players, using shot log data from the 2014-2015 season. The resulting posterior distributions allow for comparisons between players and selected covariates. Further, the model identifies factors that affect individual player efficiency and emphasize universal trends such as the continued importance of the three point shot in the NBA.

Keywords: Hierarchical Bayes, Latent Variable, Gibbs Sampler

1 Introduction

As players in the National Basketball Association continue to increase their shooting abilities and push the boundary of the definition of a "good" shot, the league is evolving to be ever more shooter-centric. Additionally, the advent of the basketball data revolution has ushered in a new era of emphasis on analytics as teams vie for the smallest edge on their competition. As a byproduct of both of these trends, there is a clear demand for advanced shooting-ability metrics. However, current evaluation of shooting centers around simple percentages and shot charts or ventures into the realm of subjective opinion. This chapter proposes a model to evaluate players' shooting ability that includes game and play-specific situational variables that have an obvious effect on shooting but are not commonly included in analyses. Including these additional factors will allow for conclusions to be made on their effect on individual player ability and show how teams can best exploit the unique advantages afforded to them by their players' abilities.

2 Methods

This analysis was done using data on over 128,000 shots taken during the 2014-2015 NBA season. These shots were segmented by each player, with a total of nine well-known

players[1] considered. These players were selected as a representative sample of various skill sets found in NBA players. The response was the result of each shot, an ordinal variable with categories for 0, 2, or 3 points scored. Free throws were not considered. There were five covariates of interest, given below. Figure 1 and 2 display an exploratory glance at the data from one player.

- **Location**: Whether the game was played at home or away, from the perspective of the shooting player
- **Shot Clock**: Time remaining on the shot clock when the shot was taken
- **Dribbles**: Number of times the ball was dribbled by the player prior to the shot being taken
- **Touch Time**: Amount of time the player has possessed the ball prior to the shot being taken
- **Closest Defender Distance**: Distance from the nearest defender to the player taking the shot

Jimmer Fredette's 2014-15 Shot Data

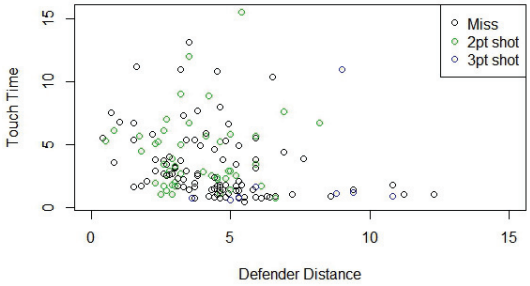


Fig. 1: There are more made shots as distance from the nearest defender increases and touch time decreases, with almost all the three point makes coming from "catch and shoot" situations.

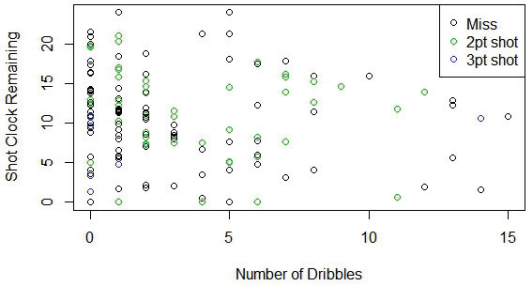


Fig. 2: There are more attempted and a higher ratio of made shots as the number of dribbles decreases. More shots are attempted in the middle of the shot clock, but it is difficult to tell where more are made.

[1] Anthony Davis, Gordon Hayward, James Harden, Jimmer Fredette, Klay Thompson, Kobe Bryant, LeBron James, Russell Westbrook, Stephen Curry

2.1 Hierarchical Latent Variable Probit Model

The ordinal data is modeled with a latent variable Z, which uses a probit link to model the underlying distribution of the response on a standardized scale[2]. The selection of this model was motivated by a desire to generate β effect sizes for the covariates and to fit a hierarchical structure that allows for comparison between players. This model also results in the derivation of posterior distributions from which prediction and meaningful statements of probability can be generated.

$$\mathbf{Z}_{ik} \sim \mathcal{N}(\mathbf{X}_k'\boldsymbol{\beta}_k, 1)$$

$$\mathbf{Y}_{ik} = j, \quad \text{if} \quad \gamma_{j-1} \leq \mathbf{Z}_{ik} \leq \gamma_j$$

$$j = \{0, 2, 3\}$$

$$k = 1, \ldots, n_{players}$$

$$i = 1, \ldots, n_{shots}$$

$$\boldsymbol{\beta}_k \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} \sim MVN(\mathbf{m}, \mathbf{V})$$

$$\boldsymbol{\Sigma} \sim InverseWishart(w, \mathbf{I})$$

There are two γ cutpoints, which are fixed at 2 and 3 as those are natural points to cut. In order to assign uninformative priors, values for the prior on $\boldsymbol{\mu}$ are set as $\mathbf{m} = \mathbf{0}$ and $\mathbf{V} = \mathbf{I}$, a 5x5 identity matrix. Values are the prior on $\boldsymbol{\Sigma}$ are set as w = 1 + $n_{players}$ and $\mathbf{I}$ = a 5x5 identity matrix. The model is implemented by deriving complete conditionals (see Appendix) and using a Gibbs sampling technique to obtain draws from the posterior distributions. The results here are from a run of 10,000 draws with a burn-in of 500.

[2] This latent variable model is taken from Albert and Chib (1993) and extended with a hierarchical structure.

3 Results

Examining the β posterior distributions answers the research question of the effect of different factors on individual player ability. As this is a probit model, the effect sizes are interpreted as the change in Z-score across the scale of the ordinal response (from missed shot to two points to three points). A few interesting observations from these results are highlighted below.

Figure 3 shows an effect when playing at home that varies by player but is generally positive. Shot clock time remaining in Figure 4 is an interesting covariate because of unique behavior for different types of players at the beginning and end of the shot clock. However, the results are able to bear out an overall trend. Number of dribbles and touch time in Figures 5 and 6 are obviously correlated, but both were included in the regression model in hopes that different players would have different correlation relationships, which the results partially supported. Closest defender distance in Figure 7 shows interesting results, but this variable does not account for a player's ability to create space from a defender on his own, which discounts a player like Harden, who utilizes a dynamic step-back move. In addition, as a way of explanation, Fredette's distribution has much more variability due to the comparatively low number of shot attempts he recorded.

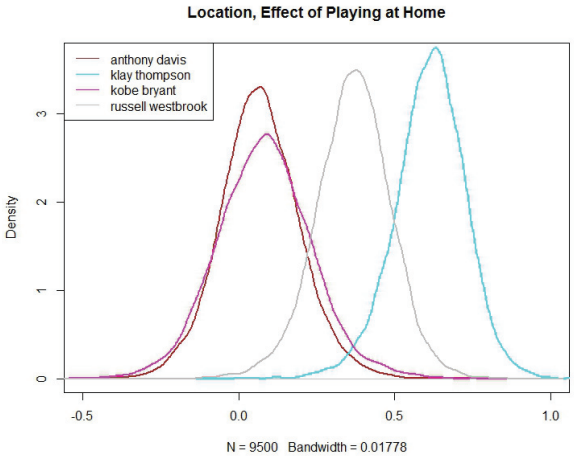


Fig. 3: Thompson and Westbrook are significantly more efficient when at home, in contrast to Bryant and Davis, who are hardly effected.

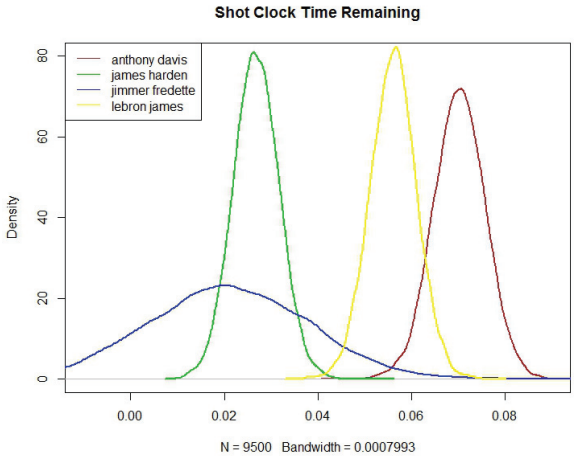


Fig. 4: Davis and James are more efficient with more time on the shot clock, in contrast to Fredette and Harden, who are much less effected.

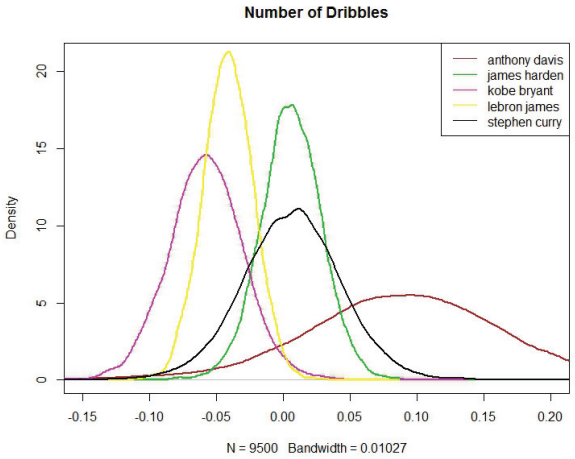


Fig. 5: Davis is more efficient as he takes more dribbles, in contrast to the other players, who are very slightly less efficient bordering on no effect.

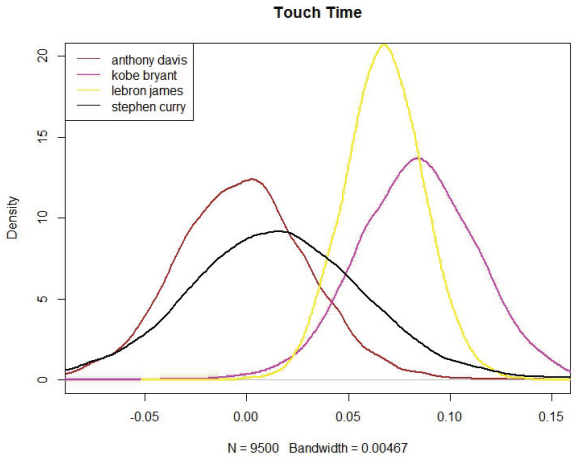


Fig. 6: Bryant and James are more efficient with longer touch times times, in contrast to Curry and Davis, who are hardly effected.

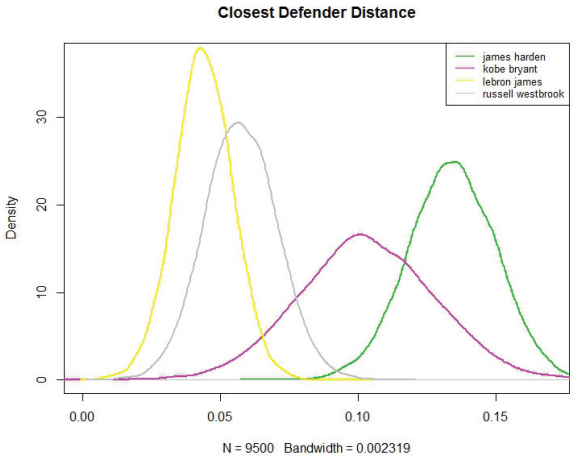


Fig. 7: Harden is significantly more efficient when a defender is further away, in contrast to James and Westbrook, who are hardly effected.

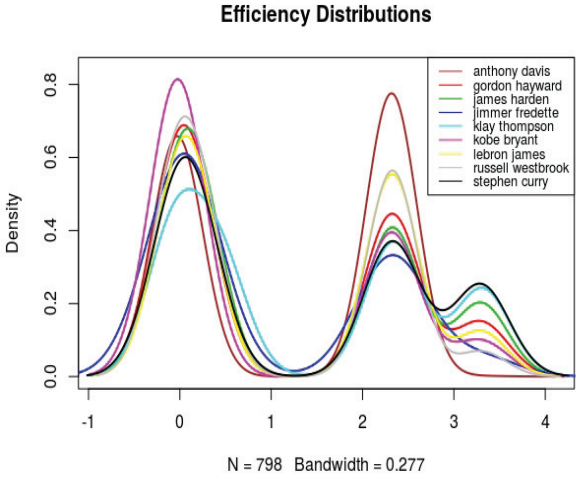


Fig. 8: Bryant is much more likely to miss, Davis excels at two point shots, and Curry and Thompson top the three point category.

Figure 8 shows a cumulative distribution of player efficiency, obtained from the density of the means of the draws of each player's Z-distributions. This distribution shows a player's ability to either miss or make a two point shot or three point shot. To answer the research question of direct player ability evaluation, the cursory glance at player ability shown by this plot can be formalized by extension to a posterior predictive distribution. This distribution

reflects the probability of a player's next shot being either a miss, two point make, or three point make and is seen in Table 1.

	Miss	2 Points	3 Points	Composite Score
Stephen Curry	0.4716	0.4524	0.0760	1.1328
Anthony Davis	0.4908	0.4912	0.0180	1.0364
LeBron James	0.5264	0.4212	0.0524	0.9996
Gordon Hayward	0.5276	0.4236	0.0488	0.9936
James Harden	0.5416	0.4112	0.0472	0.9640
Klay Thompson	0.5572	0.3740	0.0688	0.9544
Russell Westbrook	0.5620	0.3932	0.0448	0.9208
Kobe Bryant	0.5952	0.3700	0.0348	0.8444
Jimmer Fredette	0.6168	0.3440	0.0392	0.8056

Table 1: The posterior predictive distribution is reflected in columns one through three. This distribution reflects the discrete multinomial data with π_j probabilities of each player's next shot of being in category j. The composite score ranks player ability by multiplying each π_j by its corresponding 0, 2, 3 point value and summing across each player.

These results make intuitive sense to any knowledgeable NBA fan. It is no surprise that Curry, the league MVP, received the highest efficiency rating. Despite being a smaller guard, during the 2014-2015 season Curry improved his ability to finish inside and enhanced his already deadly outside shooting. Interestingly, Anthony Davis, who rarely attempts three point shots, is ranked second. This suggests that there is still a place in the league for players that are able to operate closer to the basket. LeBron James and Gordon Hayward receive almost identical ratings in every category, which highlights a shortcoming in this analysis. James is clearly a more dynamic player, considered by some to be the greatest of all-time. However this model only considers shooting and fails to consider other factors that would separate James from Hayward.

The lower half of the rankings include more players with unique skill sets that the results help identify. James Harden and Russell Westbrook are two other players that are hurt by the exclusion of factors other than shooting because the strengths of their game lie in other areas. Neither Harden's ability to get to the free throw line, nor Westbrook's raw athleticism that earns him triple-doubles are included in the model. However, the posterior

predictive distribution does accurately profile Klay Thompson's three point expertise and two point deficiency. Finally, the results show that both Kobe Bryant and Jimmer Fredette struggled in their final year in the league. Bryant was never able to regain his elite level of play after his Achilles injury and Fredette's shooting ability mysteriously abandoned him during his time with the New Orleans Pelicans.

4 Conclusion

This analysis set out to explore factors that influence shooting ability and rank players according to that ability. However, once analysis was underway it became clear that the ability being explored was one of efficiency, not pure shooting. This is due to the decision to classify and model the data as ordinal when in fact there are relationships between misses, two point makes and three point makes that are more complex than a simple ordinal model can account for. However, this model was able to represent a metric of efficiency, which was taken to mean a player's ability to produce three point field goals over two point field goals, and both over missed shots. If shooting ability was going to be modeled, a nominal model that differentiated misses into two point attempts and three point attempts would be better able to bear out worthwhile results.

Despite this unanticipated deviation, interesting, interpretable results that make intuitive sense to those well-versed in the NBA were still obtained. It makes sense that, for example, cold-blooded Kobe is less affected by road games and LeBron is closely guarded on his many crashes down the lane. There is plenty of room for future studies in player efficiency, especially for those that are able to include additional factors, like spatial data, fouls drawn and free throws (the exclusion of which certainly hurt James Harden here), and other commonplace statistical categories like rebounds and assists.

5 Appendix

5.1 Convergence

Trace plots from the Gibbs Sampler with 10,000 draws and a burn-in of 500 are shown below. They converge quickly and exhibit good mixing behavior.

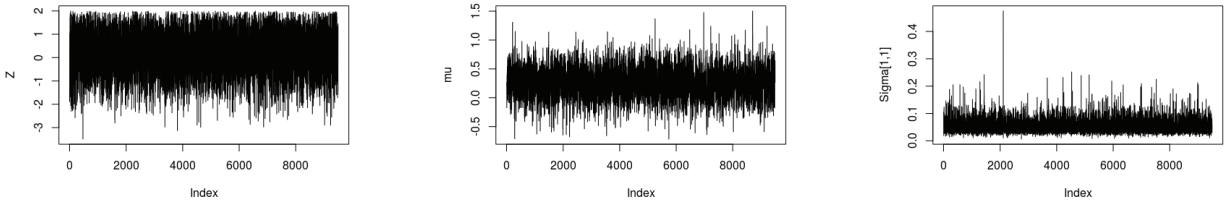


Fig. 9: Trace plots for Z, μ, and the first entry of Σ

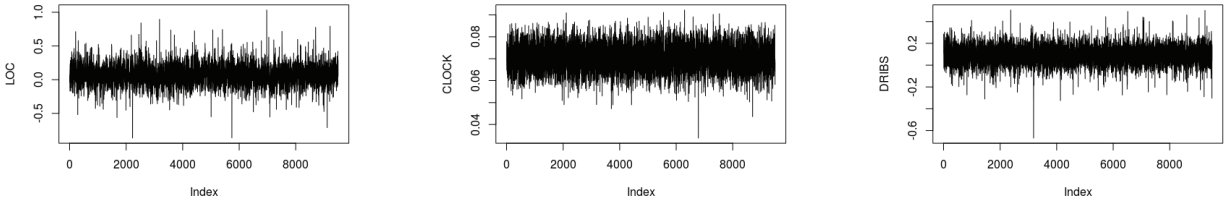


Fig. 10: Trace plots for the first three β coefficients corresponding to location, shot clock time, and number of dribbles

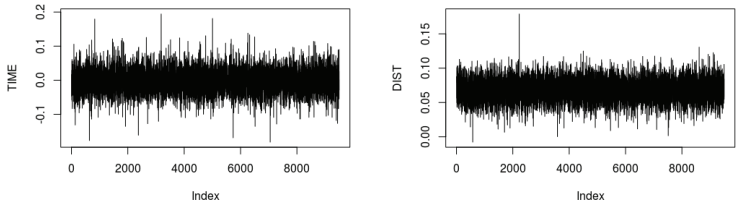


Fig. 11: Trace plots for the β coefficients corresponding to touch time and closest defender distance

5.2 Complete Conditionals

The complete conditionals used in the Gibbs step were drawn from three different sources, and are given below.

$$\boldsymbol{\beta}_k|* \sim MVN(\Sigma^*(\Sigma^{-1}\mu + X_k'Z_k), (\Sigma^{-1} + X_k'X_k)^{-1}) \tag{1}$$

$$\Sigma^* = (\Sigma^{-1} + x_k'x_k)^{-1}$$

$$\boldsymbol{\mu}|* \sim MVN(V^*(V^{-1}\mu + \Sigma^{-1}\bar{\boldsymbol{\beta}}_k), (V^{-1} + \Sigma^{-1})^{-1})$$

$$V^* = (V^{-1} + \Sigma^{-1})^{-1}$$

Peter D. Hoff, *Introduction to Bayesian Statistics for the Social Sciences* (2006)

$$\boldsymbol{\Sigma}|* \sim InverseWishart(w + n_{players}, (I + \sum_{k=1}^{n_{players}} (\beta_k - \mu)(\beta_k - \mu)')^{-1})$$

Peter D. Hoff, *A First Course in Bayesian Statistical Methods* (2009)

$$\mathbf{Z}_{ik}|*, Y_{ik} = j \sim \mathcal{TN}(\mathbf{X}_k'\boldsymbol{\beta}_k, 1, \gamma_{j-1}, \gamma_j)$$

Candace Berrett, *Generalized Linear Models* (2016)

5.3 Code

```
y<-shots$PTS


X.list <- list()
for(i in 1:n.players) {
  X.list[[i]] <- shots[shots$player_name==unique(shots$player_name)[i],
  c("LOCATION","SHOT_CLOCK","DRIBBLES","TOUCH_TIME","CLOSE_DEF_DIST")]
}


beta<-matrix(0,ncol(X.list[[1]]),nrow=n.players)
```

```
beta.save<-array(dim = c(n.players,ncol(X.list[[1]]),(burn+length)))
beta.save[,,1]<-beta

mu<-matrix(0,ncol=ncol(X.list[[1]]),nrow=(burn+length))
mu[1,]<-rep(0,ncol(X.list[[1]]))

m<-rep(0,ncol(X.list[[1]]))
V<-1*diag(ncol(X.list[[1]]))

w<-1+n.players
I<-diag(ncol(X.list[[1]]))
Sigma<-array(dim = c(ncol(X.list[[1]]),ncol(X.list[[1]]),(burn+length)))
Sigma[,,1]<-I

Z.list <- list()
for(i in 1:n.players) {
  Z.list[[i]] <- matrix(0,nrow=(length+burn),ncol=n.player.shots[i])
}

gamma<-c(2,3)

for(d in 2:(length+burn)){
  # update beta
  for (k in 1:n.players){
    sigstar <- solve( solve(Sigma[,,(d-1)]) + t(X.list[[k]])%*%
        as.matrix(X.list[[k]]) )
    mustar <- sigstar %*% (solve(Sigma[,,(d-1)])%*%mu[(d-1),] +
        t(X.list[[k]])%*%Z.list[[k]][(d-1),])
    beta.save[k,,d] <- mvrnorm(1,mustar,sigstar)
  }
```

```
# update mu
Vstar <- solve( solve(V) +  solve(Sigma[,,(d-1)]) )
mstar <- Vstar %*% (solve(V)%*%m + solve(Sigma[,,(d-1)])%*%
      colMeans(beta.save[,,d]))
mu[d,] <- mvrnorm(1,mstar,Vstar)

# update Sigma
wstar<-w+n.players
S.mu<-matrix(0,nrow = ncol(X.list[[1]]),ncol = ncol(X.list[[1]]))
for(k in 1:n.players){
  S.mu<-S.mu+(beta.save[k,,d]-mu[d,])%*%t((beta.save[k,,d]-mu[d,]))
}
Istar<-solve(I+S.mu)
Sigma[,,d] <- riwish(wstar,Istar)

# update Z
index <- 1
for(k in 1:n.players){
  for (i in 1:n.player.shots[k]){
    Z.list[[k]][d,i] <-rtnorm(1,t(X.list[[k]][i,])%*%
      beta.save[k,,d],1,gamma[2],Inf)
    if(y[index]==2)
    {
      Z.list[[k]][d,i] <-rtnorm(1,t(X.list[[k]][i,])%*%
            beta.save[k,,d],1,gamma[1],gamma[2])
    }
    if(y[index]==0)
    {
      Z.list[[k]][d,i] <-rtnorm(1,t(X.list[[k]][i,])%*%
```

```
                    beta.save[k,,d],1,-Inf,gamma[1])
      }
      index <- index+1
    }
  }
}

#posterior predictive
z.pred<-matrix(0,nrow = length,ncol = n.players)

X.samp<-array(dim = c(n.players,5,length))
for(j in 1:5){
  for(k in 1:n.players){
    X.samp[k,j,]<-sample(X.list[[k]][,j],length,replace = TRUE)
  }
}

for(k in 1:n.players){
  index<-sample((burn+1):M,length,replace = T)
  for(d in 1:length){
    z.pred[d,k]<-rnorm(1,t(X.samp[k,,d])%*%beta.save[k,,index[d]],1)
  }
}

prob.pred<-matrix(0,nrow = n.players,ncol = 3)

for(k in 1:n.players){
  prob.pred[k,]<-hist(z.pred[,k],plot = F,breaks=
   c(-Inf,2,3,Inf))$count/length
}
```

```
pred.avg.score<-numeric(9)
values<-matrix(c(0,2,3),ncol = 1)
pred.avg.score<-(prob.pred%*%values)
```

Predicting U.S. Ozone Levels Using CMAQ and EPA Station Data

Stephen Merrill

Brigham Young University

Abstract. Accurate representation of the distribution of ground-level ozone has crucial application to the burning issue of climate change. Ozone levels are currently measured at EPA stations across the country and modeled through computer simulations known as CMAQ. However, these systems do not produce identical results. This analysis uses a spatial gaussian process regression model in order to combine these two sources of data and model ozone levels across the United States. The model improves upon both systems and produces results that align with the station measurements and also predict ozone levels at more locations across the country.

Keywords: Spatial statistics, Gaussian process, Principal components

1 Introduction

The increasing quantity of ground-level ozone (O_3) is a major concern to environmentalists. O_3 is the main component of smog and breathing high concentrations can cause medical conditions such as asthma. Because of this, O_3 levels are monitored by the EPA. In this study, they have provided 800 station measurements of maximum O_3 levels in an eight hour period on May 22, 2005. Another attempt to monitor O_3 is carried out the Community Multi Scale Air-Quality Model (CMAQ), which uses an algorithm to project O_3 levels. However, these projections are not in line with the actual observed O_3 levels from the EPA. There is some relationship, but neither method provides accurate coverage of the full range of locations scientists are interested in. It is therefore valuable to understand the relationship between CMAQ simulations and the EPA station observations and be able to predict ground-level ozone at 2,834 locations that have been identified by scientists. To this end, this analysis will use both sets of data in order to build a spatial model of the United States that satisfies these research goals.

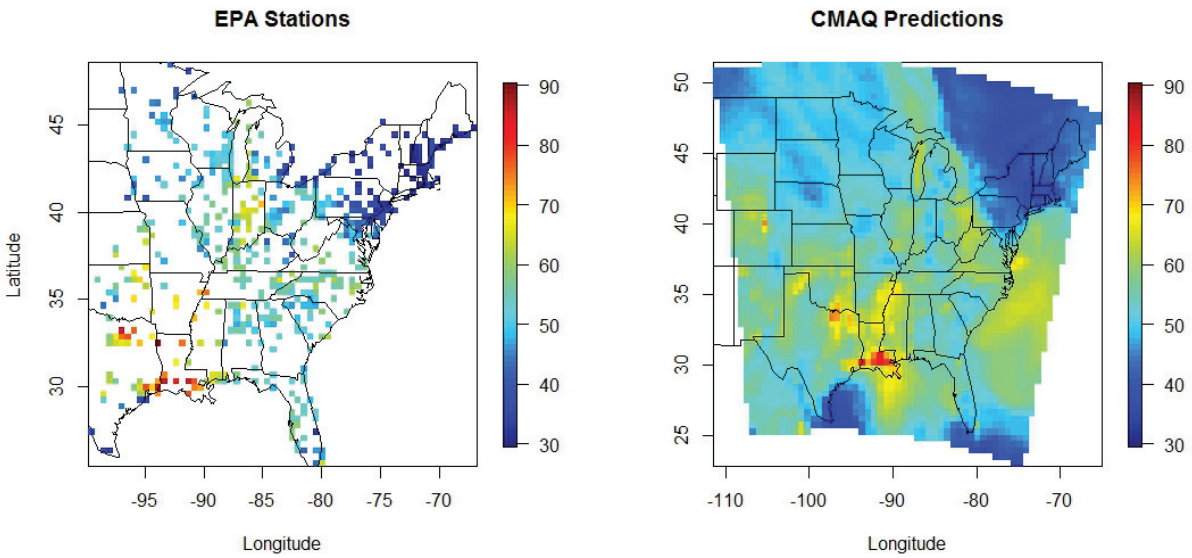


Fig. 1: This analysis is centered around combining the accuracy of the stations with the spatial coverage of the CMAQ projections.

1.1 Data Exploration

Figure 1 below shows the difference between the two O_3 measurements. Although CMAQ predictions cover more area, they are inaccurate in that they do not accurately represent the observed station levels. The analysis is complicated by the clear non-linear relationship between the O_3 levels and location on the map. This relationship between O_3 and latitude and longitude can be seen further in Figure 2. It is also necessary to consider the high dimensionality of the CMAQ data, which contains predictions at 66,960 different locations. Comparatively, the EPA has observations from only 800 stations. Finally, because this is spatial data it is of note that the O_3 level at one location depends on the levels elsewhere and independence cannot be assumed.

2 Methods

2.1 Principal Components Regression

Before modeling the O_3 levels on a spatial grid it is necessary to define a set of predictors. In order to do so, the relationship between the 66,960 calculated CMAQ values and the 800 observed values at EPA stations must be determined. Initially, the CMAQ covariates are defined as a list ordered by distance to the associated station observation. However, the

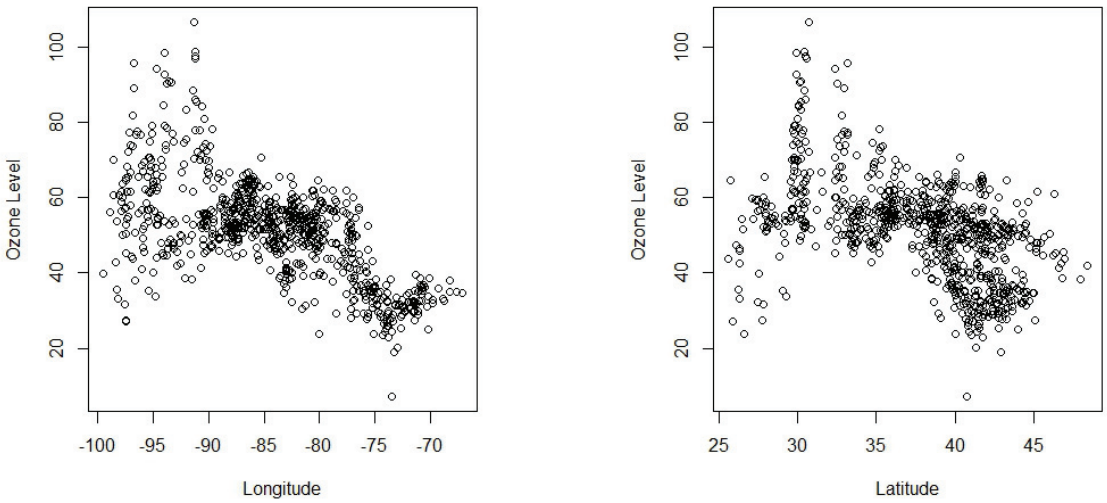


Fig. 2: Latitude and longitude have a non-linear relationship with ground-level ozone, which satisfies intuition about the spatial relation between O_3 levels and location around the country.

resulting matrix is 800 x 66960, so the problem of high dimensionality present in the data must be addressed. High dimensionality refers to having more predictors than observations, which creates problems with overfitting, potentially overreacting to small bumps in the data and creating false positive relationships with variables that are not related.

An initial approach to dimension reduction is to arbitrarily select an amount of CMAQ locations that are closest to each EPA station and only include those in the data matrix, based on the logical assumption that O_3 level at one location can be determined by nearby concentrations. However, this is less than ideal because using these surrounding locations induces collinearity among the covariates and inflates the variance. This analysis uses principal components regression in order to solve this problem. This technique represents the 66960 potential variables with far fewer predictors by reducing the dimension to fit the directions of the data in which the observations vary the most. These directions are orthogonal combinations of the original variables and are referred to as principal components. The subset of these principal components to include can be determined through the regression model below.

$$Y(x_i) = \beta_0 + \sum_{i=1}^{N} z_{ij}\theta_i + \epsilon_j$$

Subject to the constraint $\beta = (\Psi^{-1})'\theta$

- Ψ is the matrix of the eigenvectors of the covariance of $\mathbf{X}$
- z_{ij} is the ordered principal component value for location j
- θ is the loading for the associated z
- $\epsilon_j \overset{iid}{\sim} N(0, \sigma^2)$
- x_{ij} is the ordered CMAQ value for location j

In order to implement this model, the $\mathbf{X}$ was defined as the closest 20 locations. This covers a radius of about 20 miles and was determined to be sufficient in representing the location effect and eased the otherwise burdensome computation. Figure 3 displays iterations of the model, each including different numbers of principal components, by showing the amount in RMSE in each. The lowest RMSE was obtained by the model with a single component, and that component which has a strong linear relationship with the actual O_3 levels.

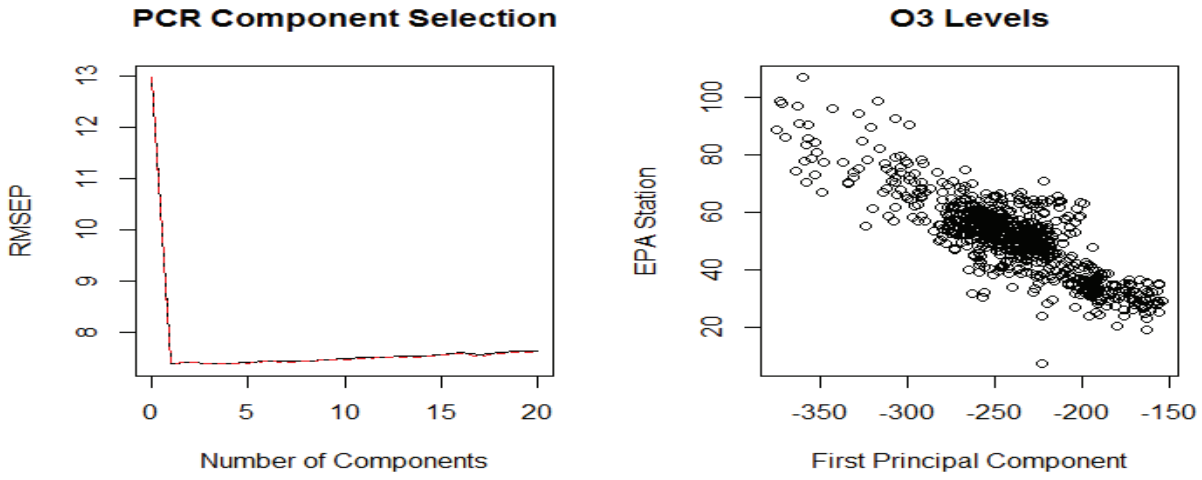


Fig. 3: PCR reduces the dimensionality down to only one linear variable.

2.2 Gaussian Process Regression

Now that predictors have been defined, a spatial statistics model will be used to formalize the relationship between the EPA station measurements and CMAQ simulations over a spatial field and use this data to predict O_3 levels at new locations. The use of a spatial statistics model is a viable method for analyzing data with linear and/or non-linear relationships with observations that are correlated over space. In this analysis, using a spatial model allows for a combination of the non-linear latitude and longitude relationships and linear CMAQ distance relationships.

This spatial statistics model relies heavily on gaussian process regression. Gaussian process regression is a viable method for analyzing data that is non-linear with correlated observations. A gaussian process is a type of stochastic process, which refers to a collection of random variables over a specified interval where only a finite collection of random variables are observed. In this case, we observe 800 observations of O_3 levels corresponding to latitude and longitude. A gaussian process is a stochastic process where any finite collection of random variables follow a multivariate normal distribution. Therefore, the observed ground-level ozone at various locations need a smooth function that minimizes residuals. This function will be considered random at each location so it can be modeled with a distribution as follows.

If O_3 levels, $y(s_1), \ldots, y(s_N)$ and the covariates $x(s_1), \ldots, x(s_N)$ are observed at N distinct spatial locations $s_1, \ldots, s_N$ in some spatial region D,

$$\boldsymbol{Y} = \begin{bmatrix} y(s_1) \\ \vdots \\ y(s_N) \end{bmatrix} \sim N(\boldsymbol{Z\theta}, \sigma^2 \boldsymbol{R} + \tau^2 \boldsymbol{I}_N)$$

- τ^2 is the variance of the distribution of O_3 given the random variables that describe the non-linear relationship between O_3 and latitude and longitude.
- σ^2 is the variance of the collection of those random variables.
- $R_{ij} = \text{Matern}\ (\mid x_i - x_j \mid, \nu, \alpha)$ is the correlation between O_3 levels based on the difference between locations.

This gaussian process model uses a Matern covariance structure, notated as $\boldsymbol{R}$. This function accounts for the effect of non-linear covariates and satisfies the requirements of being positive definite and strongly correlated over small distances. The Matern function includes parameters α and ν which govern decay and smoothness, respectably. Decay can be considered the correlation between points. As α increases, the correlation decreases and the function tends to jump more. Smoothness determines how cohesively the function moves from value to value. As ν increases, the function becomes more smooth. Values for these parameters are chosen separately. The decay parameter, α, is chosen using maximum likelihood estimation. However, the smoothing parameter, ν, is not easily chosen. It is usually selected by visually examining different values; however that is not easily done in this setting, so a default value of $\nu = 1.5$ is used.

2.3 Prediction Using GPR

This model will fulfill both goals of the analysis by allowing for inference to be made on the relationship between CMAQ simulations and EPA measurements and also making predictions of ground-level ozone at new locations. The EPA is interested in prediction at 2834 specified locations. This can be done by using the properties of the MVN to obtain the conditional distribution of the new locations given the other data, $\boldsymbol{Y^*} \mid \boldsymbol{Y}$. This is done by building a vector of the predicted values $\boldsymbol{Y}^* = (y(x_1^*), ..., y(x_k^*))'$ stacked on the observed O_3 values. The partitioned matrix is also distributed multivariate normal and is detailed below.

$$\begin{bmatrix} \mathbf{Y}^* \\ \mathbf{Y} \end{bmatrix} \sim \begin{bmatrix} y(x_1^*) \\ \vdots \\ y(x_k^*) \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{Z}_k\boldsymbol{\theta}, \sigma^2 \begin{bmatrix} \boldsymbol{R}_{Y^*} & \boldsymbol{R}_{Y^*,Y} \\ \boldsymbol{R}'_{Y^*,Y} & \boldsymbol{R}_{\boldsymbol{Y}} \end{bmatrix} + \tau^2 \boldsymbol{I}_{N+k})$$

The predicted O_3 values are the expected value of the conditional distribution of the predicted values given the observed values. Therefore, the observed O_3 values are being used to make predictions for unknown values. Prediction is made subject to the following mean and variance equations, through which estimates and uncertainty bounds can be formed.

$$\mathrm{E}(\boldsymbol{Y}^* \mid \boldsymbol{Y}) = \boldsymbol{Z}_k\beta + \sigma^2\boldsymbol{R}_{Y^*,Y}(\sigma^2\boldsymbol{R}_{\boldsymbol{Y}} + \tau^2\boldsymbol{I}_N)^{-1}(\boldsymbol{Y} - \boldsymbol{Z}_n\beta)$$

$$\mathrm{Var}(\boldsymbol{Y}^* \mid \boldsymbol{Y}) = (\sigma^2\boldsymbol{R}_{\boldsymbol{Y}^*} + \tau^2\boldsymbol{I}_N) - [\sigma^2\boldsymbol{R}_{Y^*,Y}](\sigma^2\boldsymbol{R}_{\boldsymbol{Y}} + \tau^2\boldsymbol{I}_N)^{-1}[\sigma^2\boldsymbol{R}'_{Y^*,Y}]$$

2.4 Model Justification

This spatial statistics model does not carry the same assumptions as linear regression. It is not restricted to model only linear relationships, so linearity is only an assumption if that covariate is being included in the $\boldsymbol{X}$ matrix, as defined above. The method also predicts relationships when observations are not independent. This indicates that dependence will be present and the variance will be relative to the amount of data. Therefore, the only assumption that is made with this model is that the residuals are normally distributed. However, this assumption is not easily verified as there is only have one draw from the multivariate normal. The analysis moves forward assuming that the data comes from a multivariate normal distribution with the knowledge that the distribution is robust to misclassification.

3 Results

3.1 Inference

Inference on the relationship of the CMAQ simulations on the EPA station observations is made by interpreting the results of the gaussian process regression model. However, the **Z** principal component cannot be interpreted concisely. In order to obtain interpretable β coefficients, the principal components must be back-transformed through the following identity.

$$\hat{\beta} = \Psi\hat{\theta}$$

Uncertainty estimates and confidence intervals can also be obtained through the following equations.

$$\mathrm{Var}(\hat{\beta}) = \Psi\mathrm{Var}(\hat{\theta})\Psi' = \sigma^2\Psi(Z'Z)^{-1}\Psi'$$

$$\text{CI: } \hat{\beta} \pm t^*_{n-M+1}\mathrm{SE}(\hat{\beta})$$

$$\text{Where M = number of selected components = 1}$$

This back-transformation returns the inference from one principal component effect to the original 20 closest CMAQ locations. A selection of these β coefficient values are included in Table 1. However, because of the previously highlighted collinearity issues, there is little value in interpreting these effect sizes. However, the intercept, β_0, can be usefully interpreted to mean that if the surrounding CMAQ simulation values were zero, the actual O_3 level would read -3.0042 Dobson Units, which is not a supported value. This indicates that the CMAQ values are overestimating the actual O_3 levels.

Table 1: Beta Estimates

	Estimate	SE	95% CI
β_0	-3.0042	0.1380	(-3.2752, -2.7333)
β_1	0.0516	0.0026	(0.0464, 0.0567)
β_2	0.0512	0.0025	(0.0464, 0.0561)
β_3	0.0498	0.0025	(0.0449, 0.0548)
β_4	0.0508	0.0024	(0.0460, 0.0556)

The model also yields estimates of the covariance parameters, presented below in Table 2. σ^2 and τ^2 are variances in multivariate normal distributions that make up the model. α, which governs the correlation of the regression line, is the decay parameter in the Matern covariance function.

Table 2: Parameter Estimates

σ^2	τ^2	α
66.742	23.585	0.579

3.2 Prediction

This analysis addresses the prediction goal of study by producing estimates (with uncertainty intervals) of O_3 levels across the United States. Figure 4 displays renderings of predicted O_3 levels at the desired locations that are both in line with the EPA station observations and inclusive of the simulated CMAQ data.

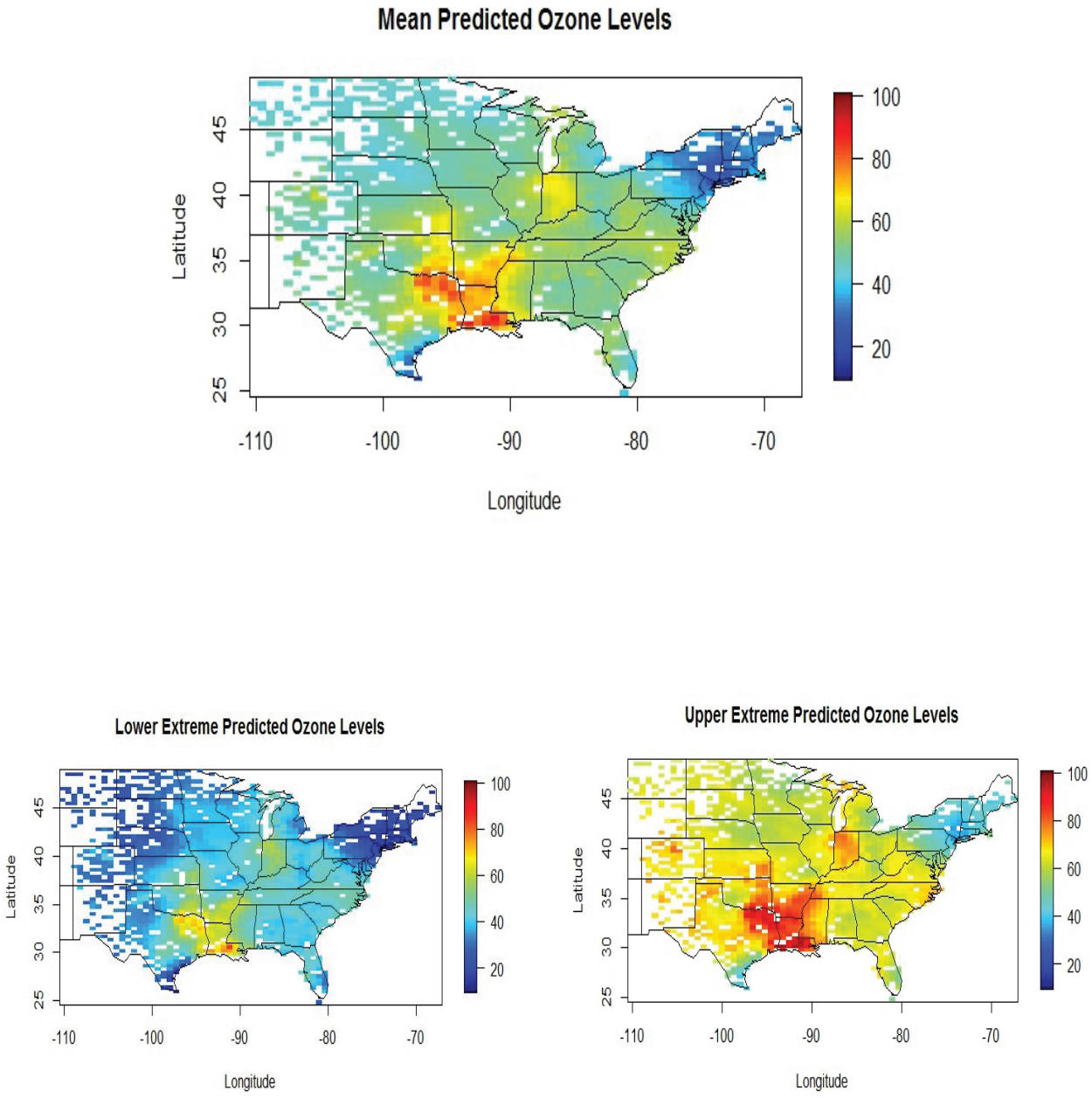


Fig. 4: O_3 level predictions at 2834 locations across the United States, along with 5% and 95% prediction intervals

The predictive accuracy was validated through a simulation using a 10-fold cross validation algorithm. Table 3 displays these results, of which bias and coverage are particularly notable. The low bias value indicates that the overall predictive ability of the model is remarkably unbiased. This is due the properties of the multivariate normal distribution, which is incorporated into the model. The coverage statistic also supports the validity of the model in that it approaches 0.95, which is the theoretical value given an error rate α of 0.05.

Table 3: Prediction Diagnostics

	Estimate
Bias	0.0109
RMSE	5.2135
Coverage	0.956
Interval Width	20.808

4 Conclusion

In order for the EPA to accurately understand ground-level ozone, it is of interest to be able to predict O_3 levels at locations without monitoring stations and understand the relationship between CMAQ projections and actual measurements. This proposed gaussian process regression model meets these goals by detailing the connection between the CMAQ simulations and EPA measurements and then making unbiased predictions on the overall spatial relationship. Analysis of the prediction results shows this was an effective approach to modeling the desired relationship.

One shortcoming of the study is the lack of a time element in the data. In order for the results to be more useful, the EPA will want to be able to project into the future. If more data were to be gathered over time the results would potentially be more interesting.

Fréchet Distribution Parameter Estimation

Stephen Merrill

Brigham Young University

Abstract. The Fréchet distribution is part of a family of distributions that are used to model extreme events with applications in engineering and finance such as snowfall, flooding, and market crashes. This analysis uses the two-parameter Fréchet Distribution to detail the derivation, use, and performance of three parameter estimation techniques: maximum likelihood, method of moments, and Bayes estimation. Each estimator is evaluated through a simulation study, against actual data, and under model misspecification. The results are then used to make conclusions about the robustness of the estimators in context of an extreme value setting.

Keywords: Maximum likelihood, Bayes estimators, Method of moments

1 Introduction

Motivation Estimation of the true but unknown values of the parameters that characterize a mathematical model of some phenomenon is a tool integral to the study of statistics. However, this ability is only useful if the estimates are accurate. To this end, several estimation techniques have been developed by statisticians. This chapter evaluates the performance of three of these estimation techniques: method of moments, maximum likelihood and Bayes estimation. Consideration of the latter two estimators bears particular interest as they originate from the frequentist and bayesian schools of thought respectably, the two largest statistical paradigms. The analysis is carried out in the context of a two parameter Fréchet distribution, which has application in modeling extreme events.

History The Fréchet Distribution was developed by Maurice René Fréchet in 1927. However, his paper was published in a remote journal and received little attention. In 1928, Fisher and Tippett developed the Generalized Extreme Value Distribution (GEV), and incorporated the Fréchet Distribution as the type II GEV Distribution. Fréchet's work was also influential in the development of the Weibull (1951) and Gumbel (1958) Distributions.

1.1 Distribution Description

Distributions are typically characterized by the form of their probability density function (PDF), cumulative density function (CDF), and first and second moments. These formulations are given below.

PDF:

$$f(x|\alpha,\beta) = \frac{\alpha}{\beta}\left(\frac{\beta}{x}\right)^{\alpha+1} exp\Big[-\left(\frac{\beta}{x}\right)^{\alpha}\Big],\ x>0,\ \alpha>0,\ \beta>0$$

CDF:

$$F(x|\alpha,\beta) = exp\Big[-\left(\frac{\beta}{x}\right)^{\alpha}\Big],\ x>0,\ \alpha>0,\ \beta>0$$

Mean, exists when $\alpha > 1$:

$$\beta\Big(\Gamma\big(1-\frac{1}{\alpha}\big)\Big)$$

Variance, exists when $\alpha > 2$:

$$\beta^2\Big(\Gamma\big(1-\frac{2}{\alpha}\big) - \big(\Gamma\big(1-\frac{1}{\alpha}\big)\big)^2\Big)$$

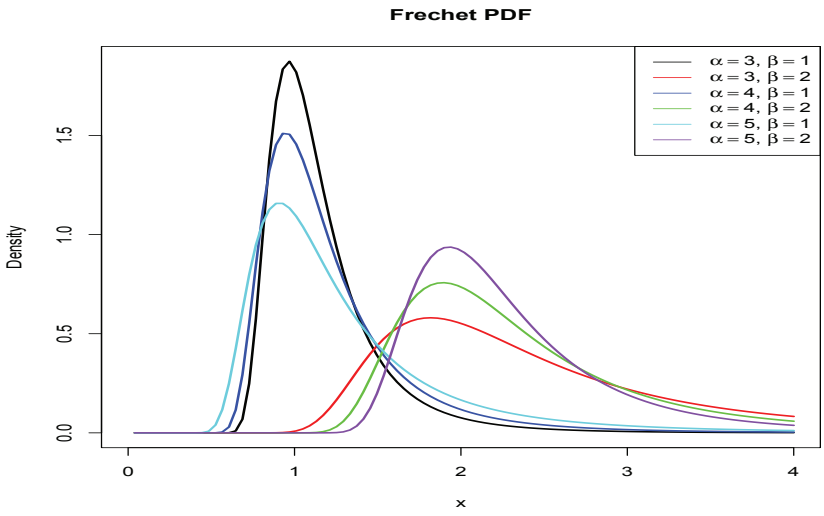


Fig. 1: This visualization of the density for several different parameter value combinations shows that the Fréchet is often right-skewed.

2 Methods

2.1 Maximum Likelihood Estimator

It is not possible to derive an analytical solution for the MLE. In these situations, the Newton-Raphson algorithm is commonly used to estimate the MLE. This algorithm uses the gradient, a vector of partial derivatives, and the Hessian, a matrix of the second order partial derivatives. The steps of the algorithm and some derivations are included below.

Newton-Raphson Algorithm

- Let $\theta_0 =$ an initial guess of the MLE values
- Let i = 0
- While $|\nabla f(\theta_i)| > \epsilon$
  - i = i + 1.
  - $\theta_i = \theta_{i-1} - [D^2 f(\theta_{i-1})]^{-1}[\nabla f(\theta_{i-1})]$
- If $D^2 f(\theta_{i-1})$ is negative definate, $\hat{\theta}_{MLE} = \theta_i$

Where $D^2 f(\theta)$ is defined as the Hessian matrix for $f(\theta)$ and $\nabla f(\theta)$ is the gradient vector. In this case, the Hessian is a 2x2 matrix containing equations (4) and (6) on the diagonal, and two elements of equation (5) on the off-diagonals, which are equal through Clairaut's Theorem. The gradient vector contains equations (2) and (3).

$$L(\alpha, \beta) = \alpha^n \beta^{n\alpha} \prod_{i=1}^{n} (x_i)^{-(\alpha+1)} \, exp\Big[- \sum_{i=1}^{n} \left(\frac{\beta}{x_i} \right)^{\alpha} \Big] \tag{1}$$

$$\frac{\partial \ln L}{\partial \alpha} = \frac{n}{\alpha} + n \ln(\beta) \; - \sum_{i=1}^{n} \ln(x_i) - \sum_{i=1}^{n} \left(\frac{\beta}{x_i} \right)^{\alpha} \ln \left(\frac{\beta}{x_i} \right) \tag{2}$$

$$\frac{\partial \ln L}{\partial \beta} = \frac{\alpha(n - \beta^{\alpha})}{\beta} \sum_{i=1}^{n} x_i^{-\alpha} \tag{3}$$

$$\frac{\partial^2 \ln L}{\partial^2 \alpha} = -\frac{n}{\alpha^2} - \sum_{i=1}^{n} \left(\frac{\beta}{x_i} \right)^{\alpha} \ln \left(\frac{\beta}{x_i} \right)^2 \tag{4}$$

$$\frac{\partial^2 \ln L}{\partial \alpha \partial \beta} = \frac{\partial^2 \ln L}{\partial \beta \partial \alpha} = \frac{n}{\beta} - \beta^{\alpha - 1} (\alpha \sum_{i=1}^{n} x_i^{-\alpha} \ln \left(\frac{\beta}{x_i} \right) + \sum_{i=1}^{n} x_i^{-\alpha}) \tag{5}$$

$$\frac{\partial^2 \ln L}{\partial^2 \beta} = -\frac{\alpha n}{\beta^2} - \alpha(\alpha-1)\beta^{\alpha-2}\sum_{i=1}^{n} x_i^{-\alpha} \tag{6}$$

2.2 Method of Moments

The method of moments arrives at estimators through the simple algorithm of equating the sample moments to the theoretical moments as in equations (7) and (8) and solving. In this case, a closed form solution is not possible so estimators are obtained by solving for $\hat{\alpha}$ in equation (10) using an iterative guess and check method and then using that result in equation (9) to solve for $\hat{\beta}$.

$$\bar{x} = \beta\Big(\Gamma\Big(1-\frac{1}{\alpha}\Big)\Big) \tag{7}$$

$$s^2 = \beta^2\Big(\Gamma\big(1-\frac{2}{\alpha}\big) - \big(\Gamma\big(1-\frac{1}{\alpha}\big)\big)^2\Big) \tag{8}$$

$$\hat{\beta} = \frac{\bar{x}}{\Gamma\Big(1-\frac{1}{\hat{\alpha}}\Big)} \tag{9}$$

$$\frac{s^2}{\bar{x}^2} + 1 = \frac{\Gamma\big(1-\frac{2}{\hat{\alpha}}\big)}{\big(\Gamma(1-\frac{1}{\hat{\alpha}})\big)^2} \tag{10}$$

2.3 Bayes Estimator

The Bayes estimator is the posterior mean taken under squared error loss. That is, the posterior distributions for α and β are found using an MCMC algorithm and $\hat{\alpha}$ and $\hat{\beta}$ are the respective posterior means. The Bayesian model is given below.

$$\text{x} \sim \text{Fréchet}(\alpha, \beta)$$

$$\alpha \sim \text{Gamma}(\alpha_0, \beta_0)$$

$$\beta \sim \text{Gamma}(\alpha_1, \beta_1)$$

Unfortunately, the Fréchet Likelihood has no corresponding conjugate prior distribution, so no simple conjugate relationships can be taken advantage of. Therefore, Gamma priors are chosen to match the right-skew and positive support of the Fréchet likelihood. Computation is carried out through an MCMC algorithm, updating α and β using a Metropolis-Hastings approach. This process seeks to explore the space around each parameter by proposing new values and accepting or rejecting them based on calculation of the Metropolis-Hastings ratio.

Specifically, this analysis made n=10,000 iterations; each time randomly selecting either α or β to update. Proposal values were drawn from a Uniform(0,1) distribution and multiplied by a constant c. The majority of this analysis was carried out through a simulation study, with various values of α and β. Correspondingly, the constant c took on different values[1] depending on the current parameter values. The values of c were determined by diagnostic checking of acceptance rates and trace plots.

2.4 Simulation Study

The three methods of finding estimators were compared through a simulation study. The simulation tested various combinations of values for α and β, with $\alpha \in (3, 4, 5)$ and $\beta \in (0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2)$, by calculating the average bias and MSE for each estimator. This was done under two configurations of sample size, n=100 and n=500. For each of these runs through the simulation, data from the Fréchet Distribution was generated through the Probability Integral Transformation, given in equation (11). This transformation allows for draws from a distribution that is difficult to sample from by evaluating its inverse CDF at points drawn from a random Uniform(0,1) Distribution.

Let $U_1, ..., U_n \sim Unif(0, 1)$, with $n = 10000$

$$Y_i = F^{-1}(U_i|\alpha, \beta) = \beta\Big(\frac{-1}{\ln(U_i)}\Big)^{\frac{1}{\alpha}} \tag{11}$$

3 Results

For brevity, the simulation results are presented with values of β varying over α fixed at 3. The results conclusively show that the MLE and Bayes estimator performed markedly better than the method of moments estimator in terms of both bias and mean squared error. However, there is almost no differentiation between the MLE and Bayes estimator at any point, in either metric. Figures 2 and 3 both show that the Bayes and MLE are unbiased and have low MSE across values of β and that trends are more consistent as the number of samples increases.

[1] $c_\alpha \in (0.2, 0.3, 0.4)$, $c_\beta \in (0.0075, 0.015, 0.025, 0.03, 0.0375, 0.045, 0.0525, 0.06)$

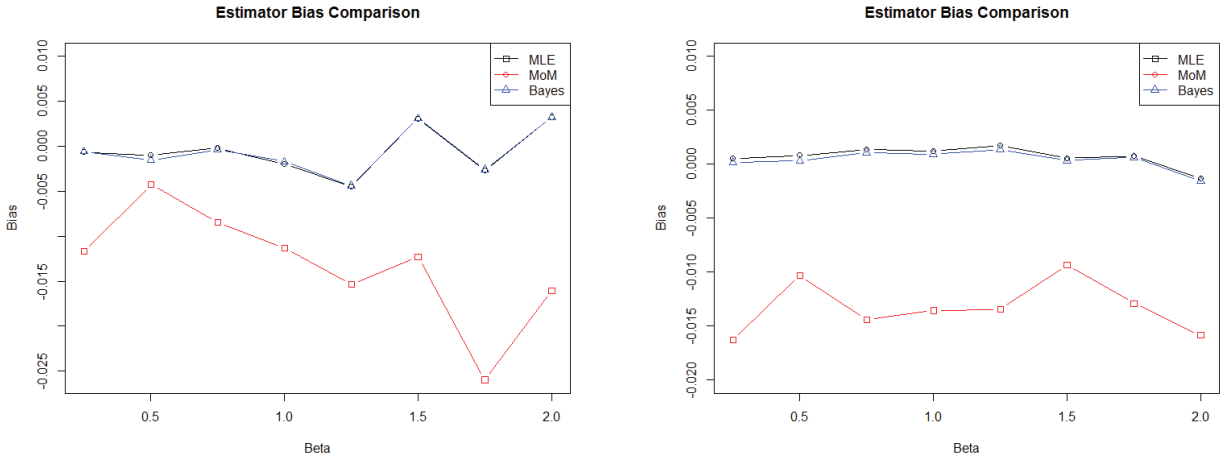


Fig. 2: Comparisons of bias in the three estimators, displayed as the average bias in samples of size 100 on the left and 500 on the right

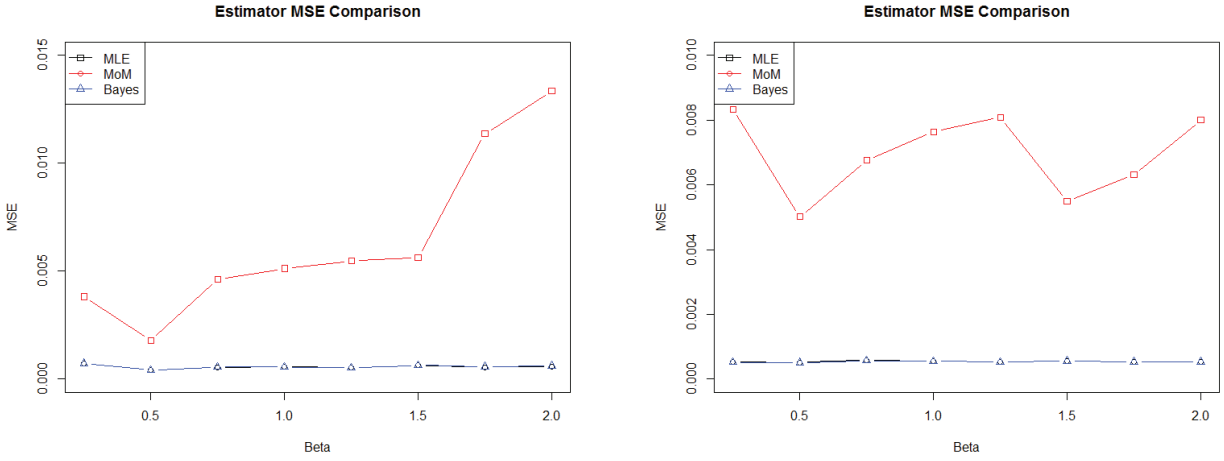


Fig. 3: Comparisons of mean squared error in the three estimators, displayed as the average MSE in samples of size 100 on the left and 500 on the right

3.1 Wind Speed Data

Data of extreme wind speeds was used to test the parameter estimation techniques. This data is located in *Extreme Value Theory in Engineering*[2]. This data is Fréchet distributed, and contains 50 observations. However, the density plots characterized by the parameter estimates and displayed in Figure 4 show exceptional fits with the data histogram.

[2] Castillo (1988)

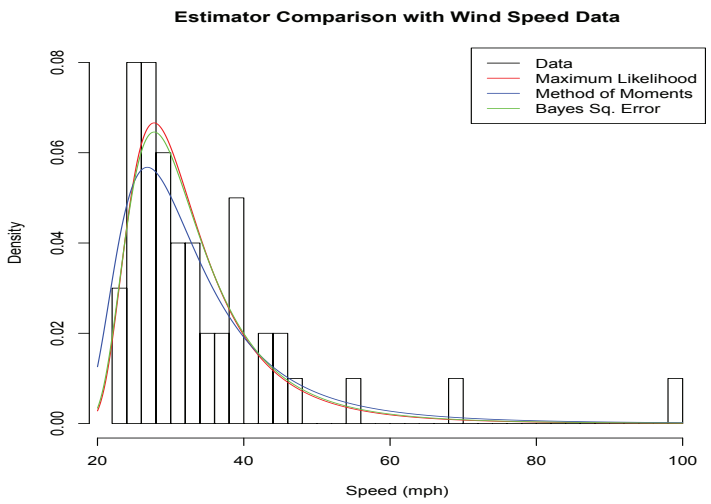


Fig. 4: The MLE and Bayes estimator offer a slightly better fit to the data than the method of moments estimator.

3.2 Model Misspecification

The performance of the three estimation techniques was also tested under a misspecified model. Here, data was generated from a Weibull Distribution with shape: $\lambda = 2$ and Scale: $\kappa = 2$. The Weibull Distribution is an appropriate distribution to misspecify data from as it is the inverse Fréchet Distribution and the type III GEV model. In order to obtain

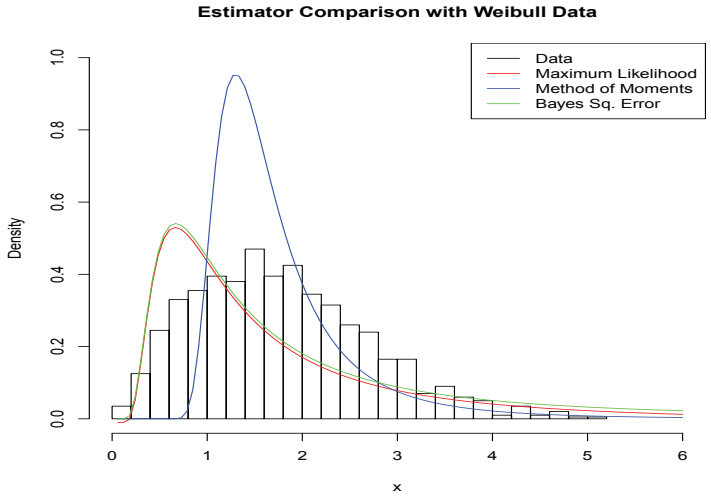


Fig. 5: The Fréchet parameter estimation techniques are not robust when asked to represent data drawn from a Weibull distribution.

results, parameter estimators for α and β under each method were computed as outlined previously. These parameter values were then used in Figure 5 to draw comparative density functions. This plot shows the method of moments result again differing from that obtained through the maximum likelihood and Bayes estimators. However, none of the methods offer overwhelmingly accurate fits.

4 Conclusion

This simulation study showed that the maximum likelihood estimator and Bayes estimator perform almost identically, and both clearly outperform the method of moments estimator with low bias and mean square error. The change in sample size produced little effect on both bias and MSE except to increase the clarity of the trend.

The estimators performed well with actual wind speed data, but even better results could possibly be attained with a larger data set. However, under model misspecification, the estimators struggled to fit Weibull data. Overall, both comparisons aligned with the comprehensive results of the analysis by continuing to show that the maximum likelihood and Bayes estimators are equally superior to the method of moments estimator.

Part II

Other Projects

Bayesian Analysis of PGA Tour Hole Difficulty

Stephen Merrill

Brigham Young University

1 Introduction

1.1 Motivation

Golf is a sport that requires both skill and strategy in order to have success. I enjoy the sport and was interested in assessing the difficulty of individual golf holes and the factors that influence that difficulty. I had PGA Tour data from research with a professor and felt I could use it to evaluate golf hole difficulty and the variables that determine difficulty. I wanted to develop a flexible model that could be used on any golf hole on the PGA Tour. This model could be used in many applications related to the PGA Tour and golf in general. For example, golf course design, which is a lucrative business that requires creativity to come up with unique and challenging golf holes. I feel that golf course designers could use the results of this model to find holes that are difficult in a certain aspect and draw inspiration for their design. Another natural application would be for PGA Tour players to use the model when planning their strategy on a hole. They could find valuable insights on how to lower their scores by applying these results to individual holes on different courses.

1.2 Data Exploration

This analysis was done using data from the 2012 Northern Trust Open held at Riviera Country Club in Pacific Palisades, California. Ideally I would have liked to use one of the four major tournaments, which most of the top ranked professional players participate in, but none of that data was available. Instead, I chose this tournament because there is still a relatively high level of professional competition and the course has no water hazards, which I thought would be difficult to properly account for. The data comes from ShotLink, a data collection service used by the PGA Tour to gather data on every shot. Specifically, there is data on over 31,000 shots at the Northern Trust Open.

2 Methods

The model uses the response of score relative to par on a hole, which usually will take on one of 5 values seen below in Table 1. This number is the difference between the expected number of strokes to complete a hole, called par, and the actual number taken by the golfer. I combined all scores less than -1 into the Birdie category and all scores greater than +2 into the Double Bogey category due to sparsity of data for these extreme scores. There is a natural ordering to golf scoring, and as a result I chose to treat the response as ordinal data and model it accordingly.

In order to answer the research question of which variables affect course difficulty, I selected four covariates that cover a golfer's entire skill set and fit β coefficients for each one. There is a natural correlation between the covariates since one shot affects the next, and this is accounted for in the model, which is formally defined below. These variables of interest are given as:

- β_1: Driving distance $>$ 300 yards. This is a binary variable for whether the drive is hit over 300 yards or not. 300 yards is considered an average driving distance for a PGA player.
- β_2: Distance off from center in 5 yard increments. Off from center refers to the distance away from the center of the fairway, measured from the location where the drive lands.
- β_3: Distance to scramble in 25 yard increments. Scramble distance is the distance from the hole, accumulated over shots where the golfer is not yet on the green, but should be. For example, on a par four hole any shots after the second that do not originate on the green will contribute to scramble distance for that hole.
- β_4: Distance of first putt in 5 foot increments. This is the distance from the location of the first putt attempt to the hole.

Name	Eagle	Birdie	Par	Bogey	Double Bogey
Score relative to par	-2	-1	0	+1	+2

Table 1: There is a natural ordinal structure to golf scores

I determined that effects were best seen by incrementing the distance measures, which makes sense because there is often a threshold of how far off a golf shot is from ideal placement that needs to be passed before there would be an impact on the score. For example, hitting a drive a few yards off from center will not matter until it is far enough off that the threshold of the fairway is passed and the ball ends up in the rough. Figure 1 shows relationships among these variables from the data and leads us to believe that a regression framework is appropriate for this analysis.

2.1 Latent Variable Bayesian Hierarchical Probit Model

The ordinal data is modeled with a latent variable Z, which uses a probit link to model the underlying distribution of the response on a standardized scale. I decided to use a latent variable bayesian probit model because it allowed us to fit regression coefficients for the covariates and I could account for covariance, knowing the covariates were not independent of one another. The model is used for ordinal response variables so it worked well with score relative to par, which is ordinal. I extended the model to be hierarchical so I could evaluate

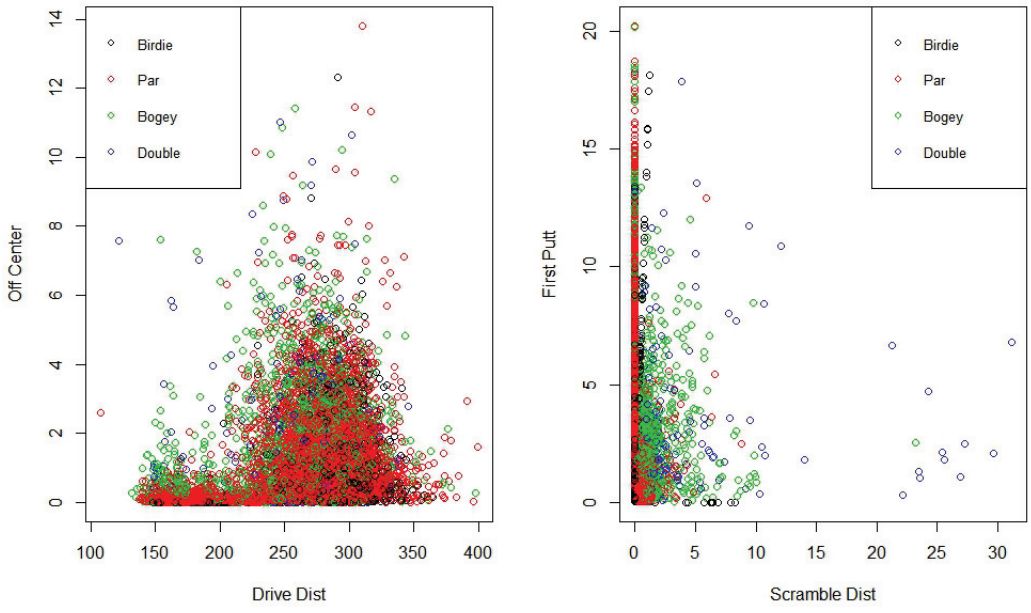


Fig. 1: An exploratory glance shows that the data has linear relationships that are intuitively expected.

multiple holes at once while getting separate coefficient estimates for each hole. Another nice feature of a bayesian model is that I could get posterior predictive distributions for each hole, allowing me to rank the holes on predicted mean score relative to par on the hole.

See equations (1) - (5) below for the formal definition of the model. There are three γ cutpoints, which are fixed at 0, 1 and 2 as those are natural points to cut. Values for the prior on $\boldsymbol{\mu}$ are set as $\mathbf{m} = \mathbf{0}$ and $\mathbf{V} = \mathbf{I}$, a 4x4 identity matrix. Values for the prior on $\boldsymbol{\Sigma}$ are set as w $= 1 + n_{holes}$ and $\mathbf{I}$ = a 4x4 identity matrix. The model is implemented by deriving complete conditionals (see Appendix) and using a Gibbs Sampling technique to obtain draws from the posterior distributions. The results here are from a run of 5,000 draws with a burn of 1,000. Figure 2 addresses the diagnostics of this model by showing that the trace plots from the Gibbs Sampler converge well.

$$\mathbf{Z}_{ik} \sim \mathcal{N}(\mathbf{X}_k' \boldsymbol{\beta}_k, 1) \tag{1}$$

$$\mathbf{Y}_{ik} = j, \quad \text{if} \quad \gamma_{j-1} \leq \mathbf{Z}_{ik} \leq \gamma_j \tag{2}$$

$$j = \{-1, 0, 1, 2\}$$

$$k = 1, \ldots, n_{holes}$$

$$i = 1, \ldots, n_{obs}$$

$$\boldsymbol{\beta}_k \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3}$$

$$\boldsymbol{\mu} \sim MVN(\mathbf{m}, \mathbf{V}) \tag{4}$$

$$\boldsymbol{\Sigma} \sim InverseWishart(w, \mathbf{I}) \tag{5}$$

3 Results

3.1 β Interpretation

Examining the β posterior distributions answers the research question of the effect of different factors on individual hole difficulty. As this is a probit model, the effect sizes are

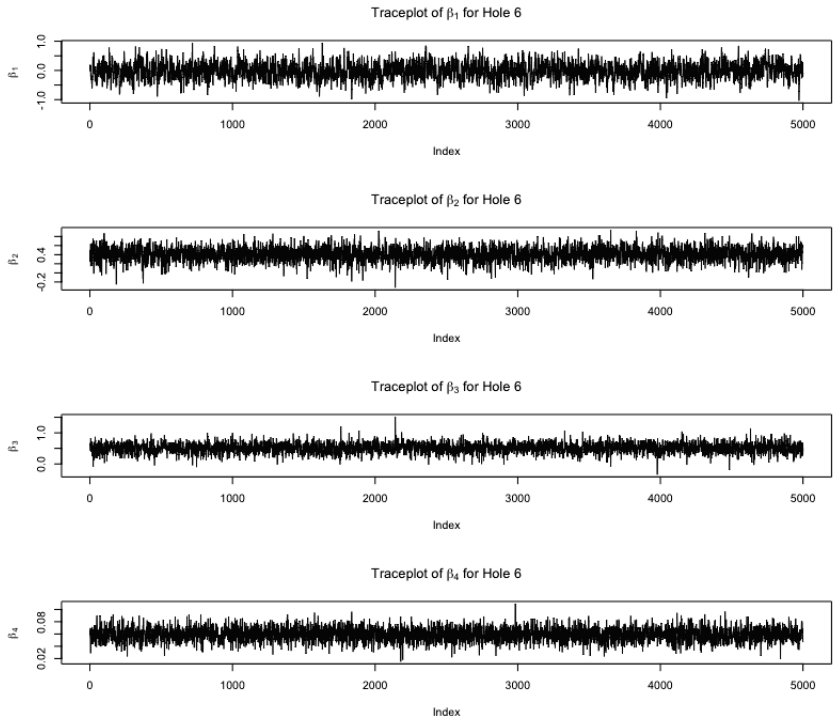


Fig. 2: Trace plots for the $\boldsymbol{\beta}$ on Hole 6

interpreted as the change in Z-score across the scale of the ordinal response. A few interesting observations from these results are highlighted below.

The first plot in Figure 3 shows that driving distance over 300 yards is much more beneficial on par 5 holes than par 4 holes, with almost no effect on par 3 holes. This result can be used to recommend that golfers drive as far as possible on par 5 holes, but use discretion on par 4 holes because a long drive could lead to trouble. Par 3 holes rarely if ever have drives over 300 yards because of how short they are, so precision placement of shorter drives is much more important. This can be seen especially in the effect of the off center distance. As par decreases, it is more and more vital to stay in the center of the hole. This makes sense because being off center on a par 4 or 5 hole allows for more opportunities to make corrections than on an unforgiving par 3. Scramble distance indicated a similar trend with par 3 holes being affected more. This result indicates that running into trouble and scrambling to make it to the green is worse on a par 3 hole. This was interesting because I believed scramble distance would be independent of par since scrambling is dependent on missing the green, irrespective of par. Finally, there is no clustering by par for first putt

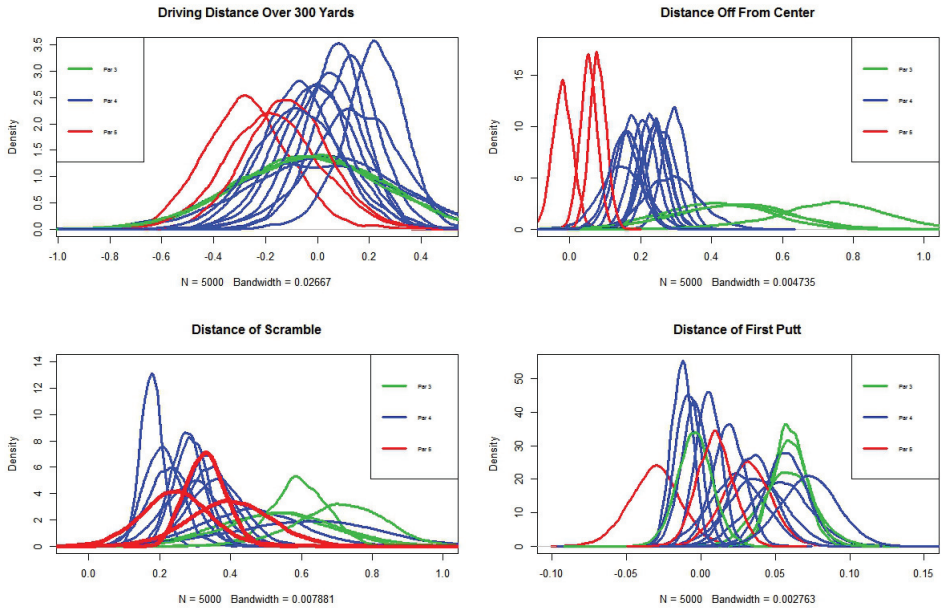


Fig. 3: Posterior distributions for the $\boldsymbol{\beta}$ effects, with holes distinguished by par amounts (Par 3 is green, Par 4 is blue, Par 5 is red)

distance, nor is there a clear effect at all. Clearly putting is different for each hole, which is understandable because every hole has a different green with different challenges.

3.2 Posterior Predictive Distribution

Figure 4 shows a cumulative distribution of difficulty for three holes, obtained from the density of the means of the draws of each hole's Z-distributions. This distribution shows the difficulty of the hole in that it will produce scores with the same relative distribution. To answer the research question of direct hole difficulty evaluation, the cursory glance at hole difficulty shown by this plot can be formalized by extension to a posterior predictive distribution. This distribution reflects the probability of a hole's next score being a birdie, par, bogey, or worse and is seen in Table 2.

I found a cumulative distribution of difficulty of each hole and then found the proportion of the distribution between each of the fixed cut points to calculate the posterior predictive distribution on each hole. Table 2 shows this calculated posterior predictive distribution. This distribution reflects the discrete multinomial data with π_j probabilities of each hole's next score of being in category j. As would be expected, most of the holes have par as

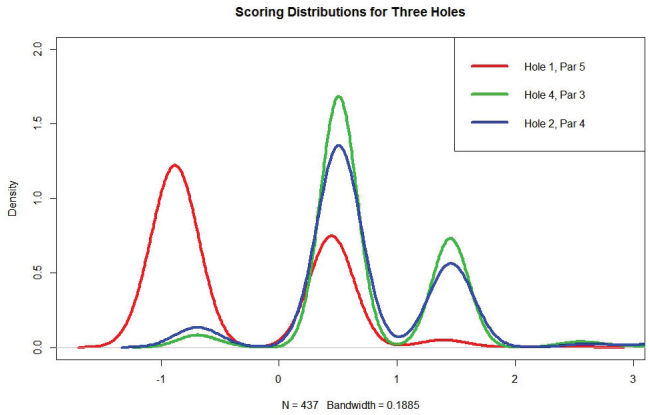


Fig. 4: Hole 1 scores much lower than Holes 2 and 4. This scoring comparison between Par 3, 4 and 5 holes is fairly common.

the score with the greatest probability of occurring. Some of the easier holes such as holes 1 and 17 have birdie as more probable than a par and some of the more difficulty holes such as 2 and 12 have a higher probability of bogey or worse.

I then took the posterior predictive distribution and calculated the mean score relative to par. I then ranked the holes from most difficult to least difficult based on this predicted mean score relative to par. To compare with what was observed in the data I also looked at the observed mean score to par for each hole and ranked them accordingly so I could see whether the results made sense. These predicted and observed ranks and scores can be see in

Hole	Birdie	Par	Bogey	Double
1	0.585	0.289	0.108	0.018
2	0.271	0.360	0.249	0.120
3	0.338	0.364	0.230	0.068
4	0.272	0.356	0.267	0.106
5	0.300	0.357	0.235	0.107
6	0.372	0.358	0.208	0.062
7	0.356	0.360	0.214	0.070
8	0.301	0.356	0.237	0.106
9	0.282	0.351	0.247	0.120
10	0.329	0.366	0.227	0.078
11	0.403	0.356	0.186	0.056
12	0.264	0.348	0.272	0.116
13	0.322	0.351	0.239	0.088
14	0.280	0.353	0.271	0.097
15	0.286	0.352	0.254	0.109
16	0.398	0.367	0.182	0.052
17	0.400	0.371	0.183	0.045
18	0.275	0.359	0.256	0.110

Table 2: The posterior predictive distributions for each hole give the probability of the next score.

Fig. 5: Hole 1, (Par 5) the least difficult

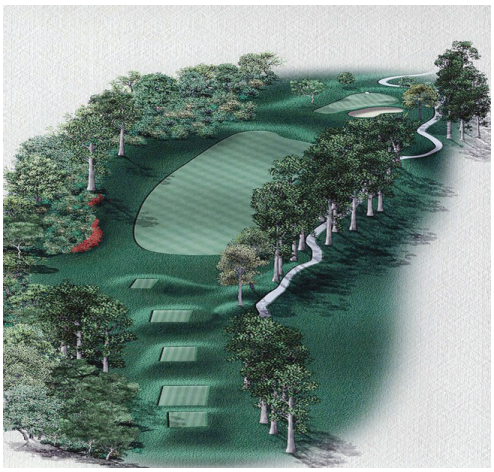


Fig. 6: Hole 12, (Par 4) the most difficult

Table 3. In Table 3 I can see that the ranks aren't exactly the same but the ranks of each hole are in the same general vicinity for the predicted and observed. The predicted mean scores were less extreme, positive or negative, than the observed mean scores. This makes sense because I simulated 10,000 scores for the posterior predictive on each hole and there were only 437 observations per hole so I would expect more extreme values in the observations.

Figures 5 and 6 present visualizations of holes 1 and 12, which were ranked the least and most difficult, respectively. Hole 1 is a short par 5 that does not have any imposing

Rank	Pred. Hole	Pred. Mean Score to Par	Obs. Hole	Obs. Mean Score to Par
1	12	0.240	15	0.311
2	2	0.218	4	0.293
3	4	0.207	12	0.293
4	9	0.205	2	0.259
5	18	0.200	18	0.256
6	15	0.186	9	0.222
7	14	0.184	14	0.215
8	5	0.149	8	0.211
9	8	0.148	13	0.188
10	13	0.093	5	0.162
11	10	0.053	7	0.076
12	3	0.028	16	0.073
13	7	-0.002	3	0.030
14	6	-0.040	10	-0.016
15	11	-0.105	6	-0.048
16	16	-0.111	17	-0.128
17	17	-0.126	11	-0.160
18	1	-0.442	1	-0.588

Table 3: Predicted mean score to par and rank by difficulty of each hole alongside the observed mean score and rank by difficulty of each hole

hazards to worry about on the drive. Players do not have to be as cautious on the drive so they can go for length rather than accuracy. This then allows them to attempt to hit the green in 2 shots rather than the standard 3 shots for a par 5, which leads to more eagles and birdies. Hole 12 is a long par 4 with a thick tree line on the right side of the hole and a narrow green surrounded by thick rough (long grass) and a deep sand bunker. The β for driving accuracy on hole 12 is high suggesting that hitting far off from the center of the fairway raises your score significantly. This could be explained by the trees; if you get behind them it is next to impossible to hit the green on your second shot. Another factor is the narrowness of the green, which makes the angle from which you approach determine your margin for error.

4 Conclusion

I feel that this model allows us to rank PGA Tour holes based on difficulty and evaluate how the covariates influence that difficulty. From the analysis I see that on average driving accuracy, measured by distance off from center, was much more influential on score than was hitting a drive over 300 yards. This is dependent, however, on the individual characteristics of the hole, such as its par. I also see the importance of scrambling distance on every hole that I analyzed, with the effect appearing to be influenced by par value for some unknown reason. The effect of first putt length is dependent on the hole and the characteristics of the green such as slope, speed and size. In the future I would like to extend this model to only look at putting, which is a unique part of golf. I could use this model to evaluate the effect of covariates such as slope, speed and distance on the number of putts taken on a green.

5 Appendix

5.1 Complete Conditionals

$$\boldsymbol{\beta}_k|* \sim MVN(\Sigma^*(\Sigma^{-1}\mu + X_k'Z_k), (\Sigma^{-1} + X_k'X_k)^{-1})$$

$$\boldsymbol{\mu}|* \sim MVN(V^*(V^{-1}\mu + \Sigma^{-1}\bar{\boldsymbol{\beta}}_k), (V^{-1} + \Sigma^{-1})^{-1})$$

$$\boldsymbol{\Sigma}|* \sim InverseWishart(w + n_{scores}, (I + \sum_{k=1}^{n_{holes}} (\beta_k - \mu)(\beta_k - \mu)')^{-1})$$

$$\mathbf{Z}_{ik}|*, Y_{ik} = j \sim \mathcal{TN}(\mathbf{X}_k'\boldsymbol{\beta}_k, 1, \gamma_{j-1}, \gamma_j)$$

5.2 Code

```
length<-5000
burn<-1000
h<-18
obs<-nrow(clean.dat)/h
clean.dat$score.to.par[clean.dat$score.to.par>2]<-2
clean.dat$score.to.par[clean.dat$score.to.par<(-1)]<-(-1)
y<-matrix(nrow = obs, ncol = h)
for(k in 1:h){
  y[,k]<-clean.dat$score.to.par[clean.dat$hole==k]
}

X<-array(0,dim = c(obs,4,h))
for (k in 1:h){
  X[,,k]<-as.matrix(clean.dat[clean.dat$hole==k,c("long.drive",
        "drive.off.full","dist.scramble","first.putt2")])
}

beta<-matrix(1,ncol=ncol(X),nrow = h)
```

```
beta.save<-array(dim = c(h,ncol(X),(burn+length)))
beta.save[,,1]<-beta

mu<-matrix(0,ncol=4,nrow=(burn+length))
mu[1,]<-rep(1,4)

m<-rep(0,4)
V<-10*diag(4)

w<-h+1
I<-diag(4)
Sigma<-array(dim = c(ncol(X),ncol(X),(burn+length)))
Sigma[,,1]<-I

z.o<-matrix(0,nrow=obs,ncol = h)
Z<-array(dim = c(c(obs,h,(burn+length))))
Z[,,1]<-z.o

gamma<-c(0,1,2)

birdie<-list()
even<-list()
bogey<-list()
dubbogey<-list()

for(k in 1:h){
   birdie[[k]]<-which(y[,k]==-1)
   even[[k]]<-which(y[,k]==0)
   bogey[[k]]<-which(y[,k]==1)
   dubbogey[[k]]<-which(y[,k]==2)
```

```
}

for(d in 2:(length+burn)){
  # update beta (k = 1,...,n_h)
  for (k in 1:h){
    sigstar <- solve( solve(Sigma[,,(d-1)]) + t(X[,,k])
        %*%X[,,k] )
    mustar <- sigstar %*% (solve(Sigma[,,(d-1)])
        %*%mu[(d-1),] + t(X[,,k])%*%Z[,k,(d-1)])
    beta.save[k,,d] <- mvrnorm(1,mustar,sigstar)
  }

  # update mu
  Vstar <- solve( solve(V) + solve(Sigma[,,(d-1)]) )
  mstar <- Vstar %*% (solve(V)%*%m +
        solve(Sigma[,,(d-1)])%*%colMeans(beta.save[,,d]))
  mu[d,] <- mvrnorm(1,mstar,Vstar)

  # update Sigma
  wstar<-w+h
  S.mu<-matrix(0,nrow = 4,ncol = 4)
  for(k in 1:h){
    S.mu<-S.mu+(beta.save[k,,d]-mu[d,])%*%
        t((beta.save[k,,d]-mu[d,]))
  }

  Istar<-solve(I+S.mu)
  Sigma[,,d] <- riwish(wstar,Istar)

  # update Z
```

```
  for(k in 1:h){
    for (i in 1:obs){
      Z[i,k,d]<-rtnorm(1,t(X[i,,k])%*%
        beta.save[k,,d],1,gamma[3],Inf)
      if(y[i,k]==1)
      {
        Z[i,k,d]<-rtnorm(1,t(X[i,,k])%*%
        beta.save[k,,d],1,gamma[2],gamma[3])
      }
      if(y[i,k]==0)
      {
        Z[i,k,d]<-rtnorm(1,t(X[i,,k])
        %*%beta.save[k,,d],1,gamma[1],gamma[2])
      }
      if(y[i,k]==-1)
      {
        Z[i,k,d]<-rtnorm(1,t(X[i,,k])%*%
        beta.save[k,,d],1,-Inf,gamma[1])
      }
    }
  }
}
```

Predicting Soil Water Content from Crop Water Stress Index

Stephen Merrill

Brigham Young University

1 Introduction

Irrigation assists farmers in distributing water in an efficient manner. As water is limited and therefore expensive, the efficient use of water is vital to the success of farmers. In addition to using water for a productive crop yield, producers can also sell surplus water at an increased price. The goal of this analysis is to predict soil hydration from the appearance of the plant. A low cost system to measure how much water a plant needs will aid the farmers in maximizing profits.

The hydration of the soil is not easily measured; however, the soil hydration can be estimated by visually measuring the approximate hydration of the plant. Soil hydration is referred to as soil water content, or SWC. CWSI is the crop water stress index which measures how dehydrated the plant appears to be. The purpose of this analysis is to predict SWC from CWSI.

1.1 Data Exploration

In order to accurately predict SWC, we need to understand the relationship between CWSI and SWC. Figure 1 reveals that the relationship appears to be non-linear. Non-linearity restricts the type of analysis that can be used for prediction. In addition, we are not assuming the plant samples are independent. It is logical that if the soil is dehydrated in a location, the soil near the original sample will have a similar level of hydration. Dependence will also limit the type of analysis that can be conducted. Analysis that can compensate for both dependence and non-linearity will be discussed in future sections. The data comes from land in Southern Colorado and contains information on 44 samples. CWSI can only take on values from 0 to 1, where 0 indicates virtually completely dehydrated and 1 indicates full

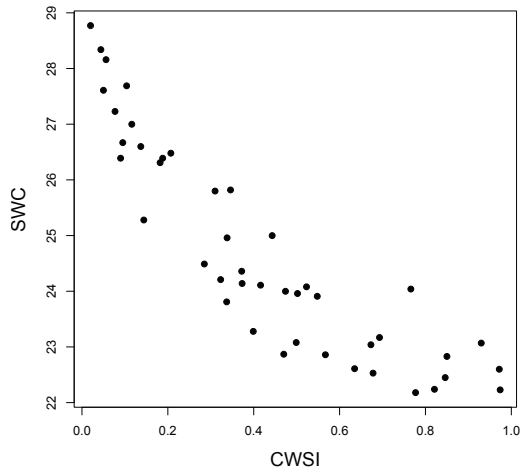


Fig. 1: Note the non-linear relationship between CWSI and SWC. The goal of the analysis is to predict the soil hydration from that appearance of the plant's hydration.

hydration. The mean CWSI value for the 44 observations is 0.42. The mean SWC value is 24.7. Table 1 contains other summary statistics.

2 Methods

Gaussian process regression is a viable method for analyzing data that is non-linear with correlated observations. A Gaussian process is a type of Stochastic process. A Stochastic process is a collection of random variables over a specified interval, where only a finite collection of random variables are observed. In this case, we observe 44 observation of SWC which correspond to CWSI. A Gaussian process is a stochastic process where any finite collection of random variables follow a multivariate normal (MVN) distribution. When we

	X	CWSI	SWC
Min.	1.00	0.02	22.18
1st Qu.	11.75	0.17	23.06
Median	22.50	0.39	24.12
Mean	22.50	0.42	24.70
3rd Qu.	33.25	0.64	26.39
Max.	44.00	0.97	28.77

Table 1: Summary Statistics for SWC & CWSI

observe SWC at various CWSI, we need to find a function such that the residuals are small and the function is smooth. If we let the function at each CWSI be random, then we can model the function with a distribution. This can be notated as follows:

$$\boldsymbol{Y} \mid \boldsymbol{W} = \begin{bmatrix} y(x_1) \\ \vdots \\ y(x_N) \end{bmatrix} \sim N(\boldsymbol{W}, \tau^2 \boldsymbol{I}_N) \tag{1}$$

Where $y(x_i)$ denotes the SWC for the i^{th} plant or CWSI.

x_i denotes the i^{th} CWSI measurement.

$\boldsymbol{Y} \mid \boldsymbol{W}$ represents the distribution of the of the SWC given the function.

$\boldsymbol{W}$ is the collection of random variables that describes the non-linear relationship of CWSI with SWC.

τ^2 is the variance in the distribution of SWC given W.

σ^2 is the variance in the distribution of W.

μ is the constant mean of the Gaussian Process.

$\boldsymbol{R}$ is the variance of $\boldsymbol{W}$.

$$\boldsymbol{W} = \begin{bmatrix} w(x_1) \\ \vdots \\ w(x_N) \end{bmatrix} \sim N(\mu \mathbf{1}_N, \sigma^2 \boldsymbol{R}) \tag{2}$$

Where $w(x_i)$ denotes the function for the i^{th} CWSI. x_i denotes the i^{th} CWSI measurement. $w(x)$, the Gaussian Process, also needs to identify a covariance structure. We use a Mattern function, which satisfies the requirements of being positive definite and being strongly correlated over small distances. To this end we use a Mattern function with parameters α and ν which govern decay, or change in correlation, and smoothness, respectably.

By definition, we can obtain the joint distribution of the SWC and the function. After we have the joint distribution, we can integrate over $\mathbf{W}$ to obtain the marginal distribution of $\mathbf{Y}$. The distribution of $\mathbf{Y}$ can be expressed as:

$$\boldsymbol{Y} \sim N(\mu \mathbf{1}_N, \sigma^2 \boldsymbol{R} + \tau^2 \boldsymbol{I}_N) \tag{3}$$

Since the goal of the analysis is to predict SWC based on CWSI, we need to use our model to predict. First, we decide the CWSI values where we are intending to predict. Then, by the properties of the MVN, we can obtain the conditional distribution of $\boldsymbol{Y}^* \mid \boldsymbol{Y}$. By the process described above, we know the mean and variance of the conditional distribution. In other words, we can find the mean and variance of the predicted values, given the observed SWC values. Additionally, we can obtain prediction intervals by taking the corresponding quantiles of the conditional distribution using the empirical rule.

Gaussian process regression (GPR) does not carry the same assumptions as linear regression. GPR is not restricted to model only linear relationships, nor do observations need to be independent. Thus, dependence will be present and the variance will be relative to the amount of data. Therefore, the only assumption that is made with GPR is that the residuals are normally distributed.

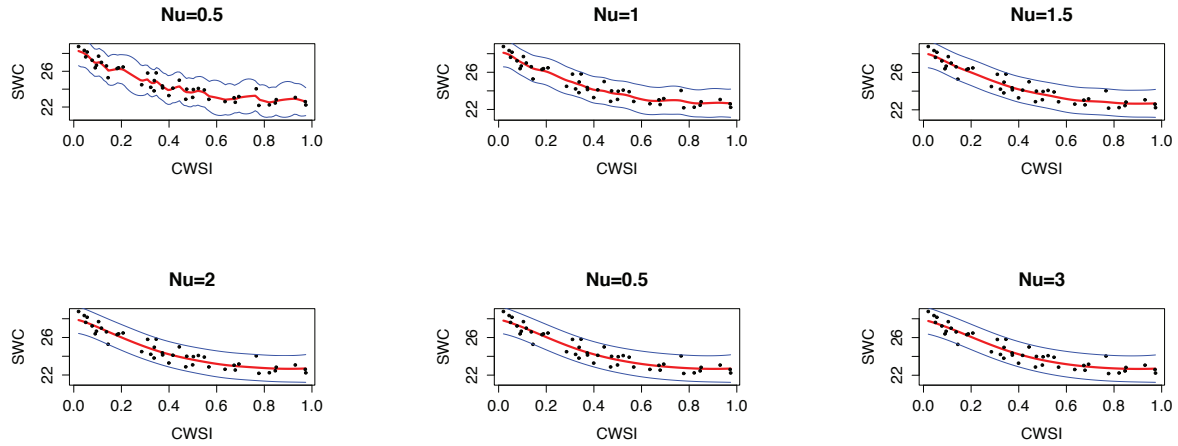
2.1 Model Justification

Gaussian Process Regression allows us to model a non-linear relationship between SWC and CWSI. Additionally, this method allows for us to model data where various CWSI are dependent on each other.

GPR assumes that residual values are normally distributed. However, this assumption is not easily verified since we modeled all observed SWC values as one draw from the multivariate normal, and one draw will yield a vector of length 44. Therefore, it is not possible to verify that a single draw comes from the MVN. We will move forward assuming that our data comes from a MVN distribution with the knowledge that the distribution is robust to misclassification. We will be aware that a violation on the normality assumption will increase the width of the prediction intervals and invalidate our use of the t distribution for constructing prediction intervals.

GPR also uses a decay and smoothness parameter. The decay parameter, α, is chosen using cross-validation. However, the smoothing parameter, ν, is not easily chosen. A small ν value is very sensitive to the data and the fitted line will be very jagged. Whereas a large ν value will may be too robust to the information in the data and ignore important trends.

The figures below show predicted SWC for various ν values. We chose a ν of 1.5 as it seemed to be robust to smaller trends, while still capturing the global pattern of the data.



2.2 Performance Evaluation

This study is solely concerned with prediction of SWC, rather than inference, so we evaluate our model based on prediction performance. Since there are only n=44 observations we used a Leave-one-out Cross Validation (LOOCV) method to simulate predictions and obtain results (see Table 2 and Figure 2).

Of particular note in Table 2 are bias and coverage. Although the bias interval ranges fairly widely, the overall prediction nature of our model is remarkably unbiased. This can be seen by comparing average distances between actual points and the predicted fit line in Figure 2 below, and is due to a property of the Multivariate Normal Distribution, which we assumed fit our data. The coverage statistic also supports the validity of our model in that it is so close to .95, the theoretical value with our assumed error rate. This means that the

Table 2: Prediction Diagnostics

	Estimate	Lower	Upper
Bias	-0.001713	-1.385079	1.381653
RMSE	0.697747	-0.789005	1.263421
Coverage	0.931818	0.437794	1.425842
Interval Width	2.830984	2.624461	3.037506

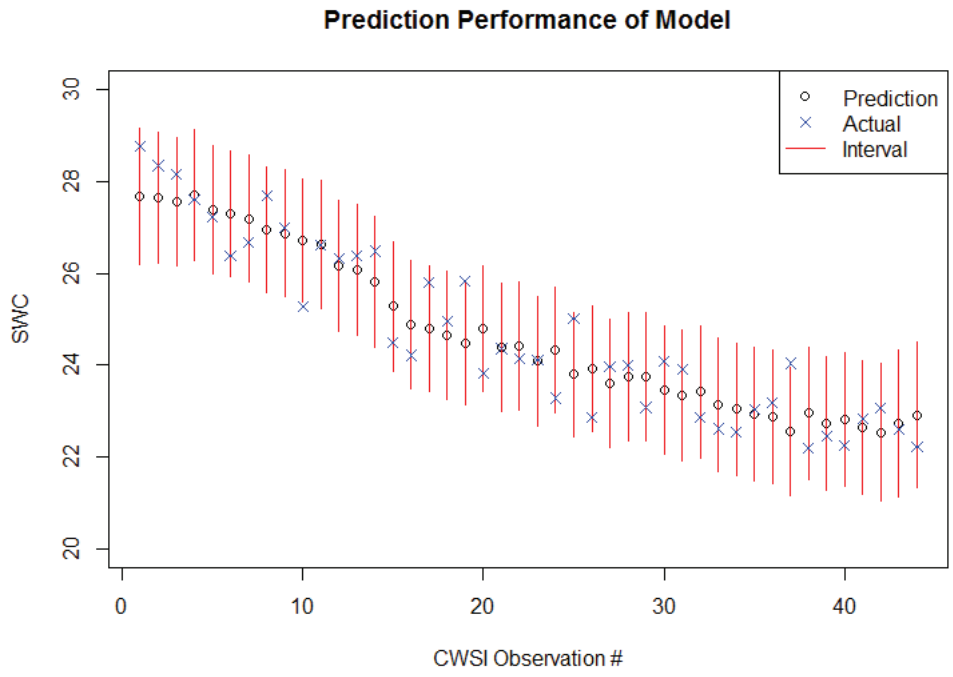


Fig. 2: 93.18% of the prediction intervals generated through the cross-validation method contained the true SWC response value.

correct number of actual points were contained within the intervals of error for our predicted SWC values. The coverage can be seen graphically in Figure 2. All of the intervals in the table are quite wide and sometimes uninterpretable. That is a function of the low number of trials our sample size restricts us to and is a weakness of our results.

3 Results

Our results include estimates of the covariance parameters and a confidence interval for μ, presented below in Table 3.

Table 3: Parameter Estimates

σ^2	τ^2	α	ν	μ	μ CI
2.392	0.440	4.972	1.5	24.982	(24.970, 24.994)

These parameters are defined as: σ^2 and τ^2 are variances in Multivariate Normal Distributions that make up the model. α is the decay parameter, which governs the correlation of the regression line and ν is the smoothness parameter, which governs the smoothness of the regression line. ν cannot be estimated and is assigned based on visual inspection of different values. Both are parameters in the Mattern covariance function. μ is the mean of the regression line, or the mean of the predicted values of SWC.

The model addresses the motivation for the study by making accurate prediction of SWC and CWSI. The predictive accuracy was validated through a simulation. The final results, the effect of varying levels of CWSI on SWC, can be seen in Figure 3. There is a curved negative trend, fit by the red Gaussian Process regression line. In the future, farmers can use Figure 3 to predict values of SWC from observed CWSI levels.

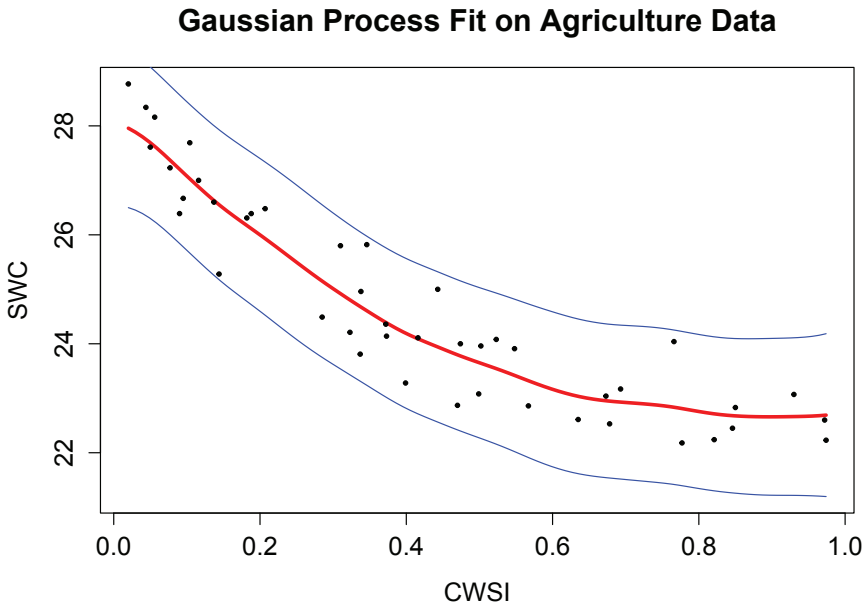


Fig. 3: The red fitted line is nearly unbiased on average and the blue confidence intervals appropriately contain the data.

4 Conclusion

In order for farmers to most efficiently use their irrigation water, it is of interest to be able to predict the SWC level of the soil from the CWSI level. Our proposed Gaussian Process model meets this goal by making unbiased predictions on the non-linear relationship. Analysis of the prediction results shows a very effective approach to modeling the desired relationship.

One shortcoming of the study is the low amount of sample data that creates wide uncertainty intervals in our results. In the future, if more data were gathered our results would be able to support the model more strongly. We could also consider a model with more covariates, such as location of the land plots, type of crops being grown, etc. However, in order for Gaussian Process Regression to continue to be accurate we will need to ensure there are no problems with high dimensional data.

Credit Card Case Study

Stephen Merrill

Brigham Young University

1 Introduction

Credit card companies are interested in predicting outstanding balances of potential customers. This is due to the potential profit to be made from interest on outstanding balances and the potential loss from customers that declare bankruptcy. A company has collected data about their current customers and hope this information can be used to predict potential customer behavior. An accurate prediction method will allow the company to identify potential customers that are more likely to maintain a moderate monthly balance, who have the greatest potential for profit.

The dataset used in this analysis includes information on the outstanding credit card balance and several quantitative and qualitative characteristics of each cardholder. The main problem is that there are 90 cardholders out of the 400 in the dataset that have zero outstanding balance. This disrupts the assumption that the data are normally distributed. Also it negatively affects the linear relationship between balance and other quantitative variables.

This analysis is meant to produce a predictive model to be used for potential customers. The most important characteristics to predict balance are identified and a model is derived. Assumptions that are made for the model are checked to see whether or not they are reasonable. This model is then tested for prediction accuracy. All of these steps are meant to provide the most predictive model possible using available methods.

2 Methods

2.1 Multiple Linear Regression (MLR) Model

$$Y = \beta_0 + X_I\beta_I + X_R\beta_R + X_S\beta_S + X_A\beta_A + \epsilon$$

- Y : Credit Balance
- β_0 : Intercept - balance for a non-student subject, aged zero, with no income or rating.
- β_I : Effect for Income - for every \$1000 change in income, on average balance will change by β_I dollars.
- β_R : Effect for Rating - for every one unit change in rating, on average balance will change by β_R dollars.
- β_S : Effect for Student status - a status change from student to non-student or vice versa will change balance by β_S dollars.
- β_A : Effect for Age - for every one year change in age, on average balance will change by β_A dollars.
- ϵ : The random error in the model.

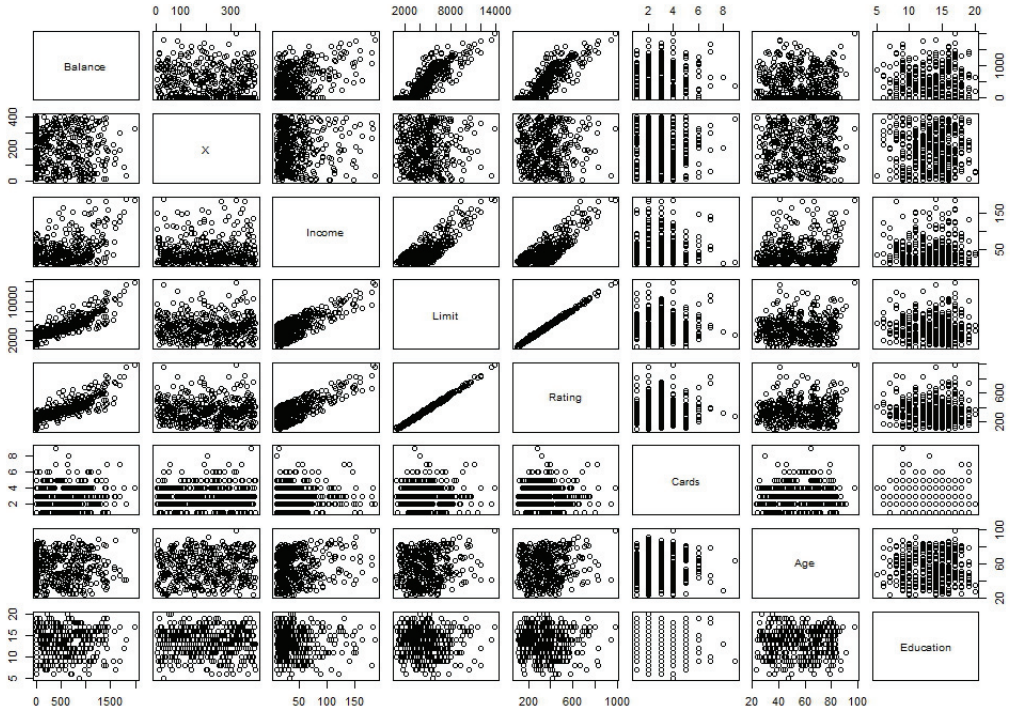


Fig. 1: All non-factor variables. Note the observations with zero balance and the colinearity between Limit and Rating

A linear regression model allows for both inference and prediction to be made. For this problem, interest lies in making predictions of future customers' credit card balances. The model accomplishes this goal by determining the effect sizes (the β values) of the significant variables: income, rating, student status and age. Once determined, prediction for balance can be made by gathering data on the subjects and making calculations according to the model.

2.2 MLR Assumptions

In order for multiple linear regression to be a valid model, the following assumptions about the data must be met.

Linearity Each variable must have a linear relationship with the response. If this is not the case, the entire model is invalid since it would be fitting a line to non-linear data. Transformations of the data are often used to solve this problem.

Independence The data must be independent. If this assumption is violated, measures of variability will typically be too small. This is a difficult assumption to verify. Usually prior knowledge of the data is required.

Normality The errors must be normally distributed. Otherwise, confidence and prediction intervals that depend on t distributions are incorrect.

Equal Variance The errors also must have equal variances. Without this, measures of variability will once again be invalid.

2.3 Model Selection

Adjusting for Zero Balance Prior to finalizing the model selection, data analysis revealed serious violations of the normality assumption. This problem was a consequence of the 90 data points with zero balance. Since no transformation created normality, those 90 data points were removed. After doing so, all the assumptions were met.

There are consequences of removing those data points. It is probable that there is some kind of relationship between the subjects with zero balance, and by removing them, the

data can no longer be considered a random sample. This weakens the model's ability to make predictions since it is now built on the assumption that the subject has a non-zero balance. Therefore the credit card companies will not be able to accurately identify customers that will have zero balance using this model.

Best Subset Selection The raw sample data contained 10 different explanatory X variables. In order to determine which variables were most significant and find the model that yielded the best prediction of balance, the best subset selection method was used. This method looks at all possible combinations of variables in the model and selects the "best" one. However, there are many different criteria that define which model is best.

A k-fold cross validation algorithm was selected as the criteria in selecting the best subset of variables to include in the model. This algorithm selects k subsets (called training sets) and determines the Mean Square Error (MSE) for each possible number of variables included in the model by making a prediction and comparing to the data not included in the training set. That data are known as the testing set. Once the number of variables to include is determined, best subset selection is done on the entire data set and the best model of the specified size is selected. Here, best is quantified using Residual Sums of Squares, the default criteria for best subset selection.

This algorithm, and use of cross validation error as the acceptance criteria, is appropriate for the data since the question of interest lies in making prediction and cross validation depends upon the subset prediction method previously outlined.

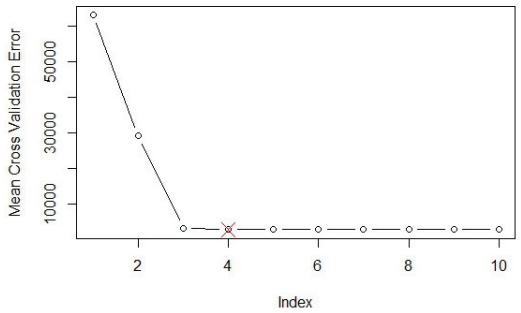


Fig. 2: Choosing four variables in the model minimizes error.

Interaction A possible interaction between income and student status was considered. However, including this interaction in the model did not produce a significant effect, so it was not included in the final model.

2.4 Model Assumptions Verification

Linearity This Added-Variable plot matrix shows a clear linear relationship between each of the four explanatory variables and the response.

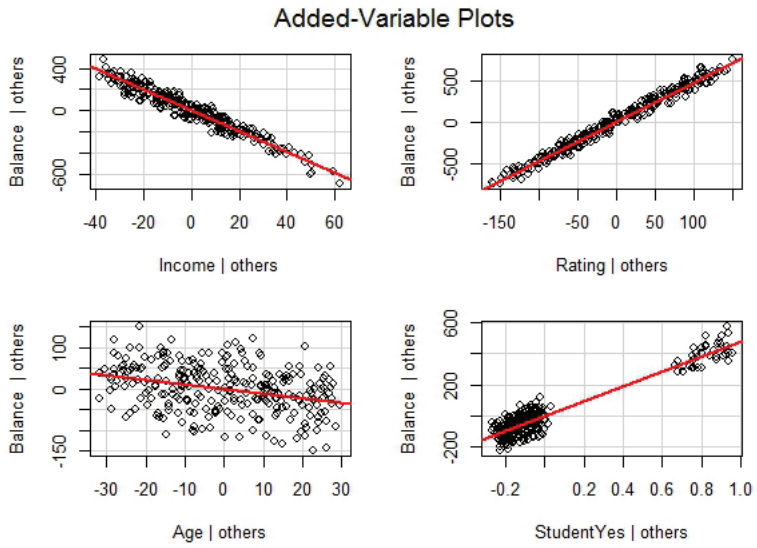


Fig. 3: Added variable plots indicate that the linearity assumption is satisfied.

Independence Independence is difficult to determine. Because there is no prior knowledge that would suggest a violation of this assumption, it is assumed to be met.

Normality This histogram and quantile plot of the residuals show that the errors are normally distributed.

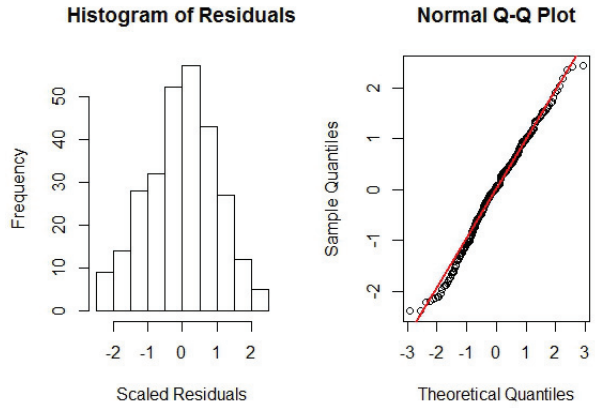


Fig. 4: The histogram of residuals and Q-Q plot show no major concerns.

Equal Variance This plot of the residuals offers no evidence of an unequal variance.

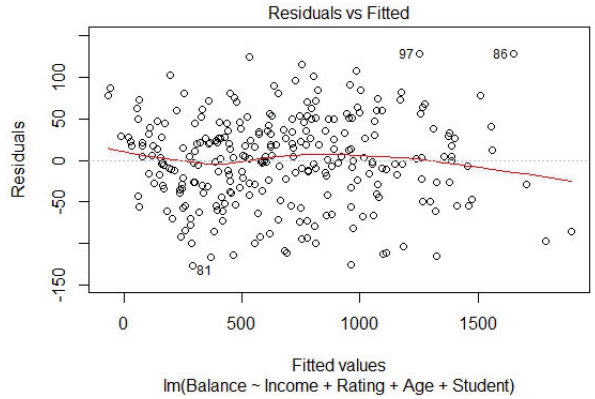


Fig. 5: The residual plot shows no signs of heteroscedasticity.

2.5 Model Fit

The model fits the data very well. Every β has a highly significant t value, which suggests that each individual effect is significant, and the entire model has a significant F-statistic, which means there is a significant effect from at least one β.

3 Results

In order to assess the ability of the model to predict accurately, another cross validation algorithm was used. The data were randomly subset into a training and testing set, with 90% of the data in the training set and 10% in the testing set. The model was fit to the training set, and then used with the data in the testing set to calculate prediction intervals. For this subset of the data, the prediction intervals contained all 31 points in the testing set.

The model was able to predict the balance of potential customers. In order to use the regression model to make a prediction the subject's income, credit rating, age and student status are needed. The table below gives the estimates of the effect size of each variable in the model along with 95% confidence intervals.

Income, age and rating can all be interpreted in the following way. The estimate for income effect size is about -9.72, this means balance would be expected to decrease by $9.72 on average as income is increased by $1000. The student variable is interpreted a little differently. The estimate of student status effect size is roughly 479.51, this means balance would be expected to be $479.51 more on average if a customer is a student.

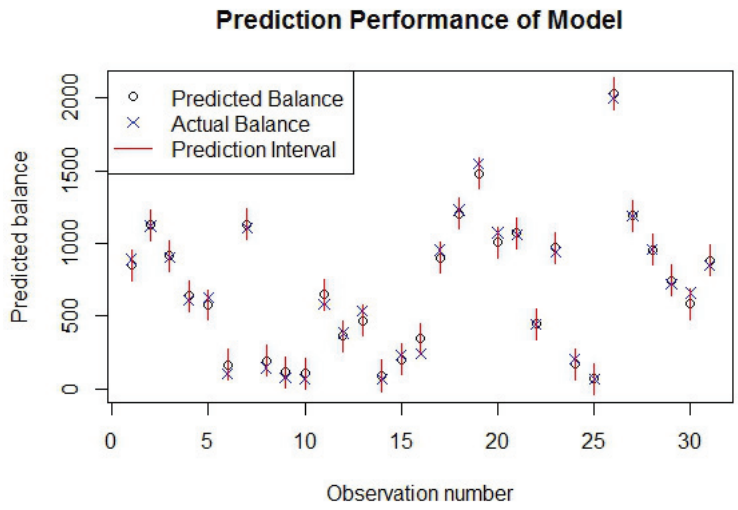


Fig. 6: Intervals contained all 31 points in the testing set

	Estimate 95% CI
Intercept	-776.86 (-808.9, -744.81)
Income	-9.72 (-10.02, -9.42)
Credit Rating	4.77 (4.69, 4.86)
Age	-1.08 (-1.46, -0.7)
Student Status	479.51 (460.64, 498.39)

Table 1: Estimates and 95% Confidence Intervals for Variables

4 Conclusion

This analysis was done to provide banks the most predictive model for potential customer balance. This was accomplished by finding the best number of variables to use in the model and then carefully selecting that number of variables that were most predictive of balance. Underlying assumptions for the model were then verified. The model was then used to predict the values of a small subset of the data and accuracy of the model was assessed.

There were problems that had to be addressed. Over 20% of the customers in the dataset had zero balance. This complicated the assumptions of normality and linearity. It was decided to exclude these customers because no transformation or combination of transformations could be found to deal with these zero balance customers. For future analysis a logistic regression equation could be used to deal with customers that potentially could have zero balance. A logistic regression would predict whether a potential customer would have zero balance or not. Then those potential customers predicted to have a non-zero balance would be run through the model described in this paper to predict their balance. This two-step analysis would allow for the use of all of the data and to take into account potential customers that could have zero balance.

Gene Expression Analysis

Stephen Merrill

Brigham Young University

1 Introduction

Genes are subunits of DNA that are responsible for carrying out instructions for various functions within the body. Some genes are responsible for regulating cell growth and differentiation. When these genes are altered, cancer can develop, since cancer is a disease of unregulated cell growth. If these altered genes could be identified, we would be able to understand cancer better.

A data set was collected having information on the cancer status of 102 patients along with the gene expressions of 12625 of their genes Cancer status was assigned by doctors to a variable called "Malignant" on a scale from 0 to 1 that indicated the level of cancer. If Malignant was close to 1, a patient is said to have very invasive tumors. If an association between a set of genes' expressions and the cancer status of the patients could be found, we would hopefully be able to identify some of the genes that were altered that resulted in cancer.

One of the issues with using the data set we have is the high dimensionality of the data: there are far more variables than there are observations. This would make typical linear regression methods inappropriate due to overfitting. Through some exploratory data analysis, we identified a couple of the highest positive and negative correlated variables with cancer status.

In the pairs plot in Figure 1 we see a couple more issues. The response variable "Malignant" does not have a linear relationship with the explanatory variables: values for malignant seem to cluster around 0 and 1. Performing a transformation that spreads out these clusters, such as a logit transformation, would help improve our analysis. Another issue we see is colinearity with the explanatory variables. This will have to be considered in

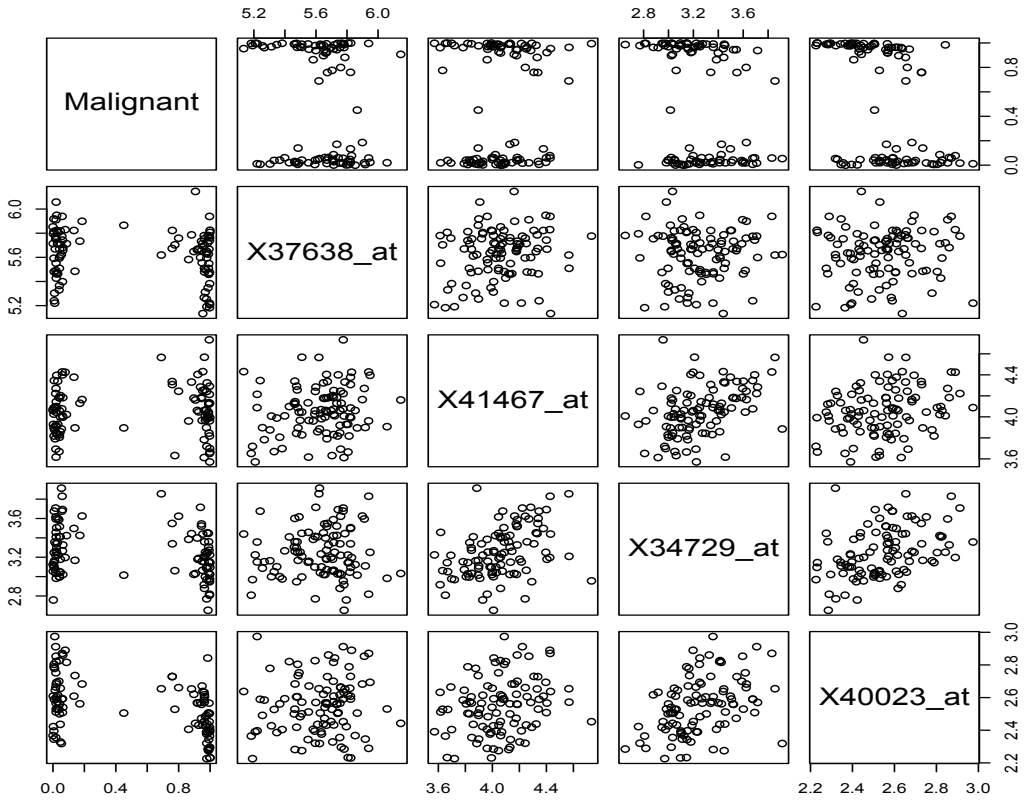


Fig. 1: Pairs Plot with a Few Selected Genes and Malignant Variable

our analysis, but because of the high dimensionality of the data the colinearity is difficult to fully address.

2 Methods

2.1 Multiple Linear Regression (MLR) Model using LASSO

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_9\beta_9 + \epsilon$$

– Y : Logit transformation of the malignancy rating. This transformation was necessary to change the support of the data from [0,1] to all real numbers, which allows for our assumptions to be met.

- β_0 : Intercept - malignancy rating for a subject with a zero expression level for each gene.
- β_i : Effect for the ith gene. For every one unit increase in expression level, the odds ratio of malignancy rating will increase by $e^{estimate}\%$. (See the Results section for further explanation) The genes are defined as follows. $i=$
  1. X32330_at
  2. X32782_r_at
  3. X33369_at
  4. X33920_at
  5. X35330_at
  6. X35347_at
  7. X36989_at
  8. X37578_at
  9. X41125_r_at
- ϵ : The random error in the model. The error follows a Normal distribution with mean 0 and variance σ^2I.

A linear regression model allows for both inference and prediction to be made. For this problem, interest lies in making inference on the effect gene expression levels have on malignancy ratings of cancerous tumors. The model accomplishes this goal by determining the effect sizes (the β values) of the genes that were determined to be significant. These effect sizes can then be interpreted in the context of the problem. In order to solve the high dimensionality problem, we used a shrinkage method called LASSO. This method shrunk the majority of the β's in the full data set down to zero, allowing for the formulation of a concise model.

2.2 Model Assumptions

Due to the issues with high dimensional data, some of the standard regression assumptions will be difficult to verify. However, in order for multiple linear regression to be a valid model, the following assumptions about the data must still be met.

Linearity Each variable must have a linear relationship with the response. If this is not the case, the entire model is invalid since it would be fitting a line to non-linear data. Transformations of the data are often used to solve this problem.

Independence The data must be independent. If this assumption is violated, measures of variability will typically be too small. This is a difficult assumption to verify. Usually prior knowledge of the data is required.

Normality The errors must be normally distributed. Otherwise, confidence and prediction intervals that depend on t distributions are incorrect.

Equal Variance The errors also must have equal variances. Without this, measures of variability will once again be invalid.

2.3 Model Selection

Model selection was done using the LASSO method. Similar to typical linear regression, the LASSO method chooses estimates of the effects of the explanatory variables (the β's) such that the residuals around the fit are minimized, but uses an additional constraint:

$$\sum_{j=1}^{p} |\beta_j| < \lambda$$

In this formula, p is the number of explanatory variables and λ is a selected shrinkage factor. The shrinkage factor will force the β_j's to be smaller and further, constrain some of the β_j's to be zero. Using this method actually forces many of the effect sizes of the 12625 genes on Malignant to be zero, thus simplifying our results.

After obtaining estimates from the LASSO method, we created 95% confidence intervals for the nonzero estimates through a bootstrapping technique. This was done by drawing a sample size equal to the number of nonzero estimates from the original 102 observations. This was done with replacement. With the new set of observations we recalculated our estimates again using the LASSO method and stored those results. After repeating this process a large number of times, we used our stored estimates in order to calculate standard errors and 95% confidence intervals for the estimates of effects of genes on malignancy scores. We then used these confidence intervals to determine which genes had a statistically significant effect on cancer and selected those variables to form our final model.

2.4 Model Assumptions Verification

Linearity The scatterplot matrix (Figure 2) shows a vague linear relationship between each of the nine explanatory variables and the response. No other kind of relationship is prominent, so we considered this assumption to be met.

Independence Independence is difficult to determine. Because there is no prior knowledge that would suggest a violation of this assumption, it is assumed to be met.

Normality This histogram of the residuals (Figure 3) show that the errors are normally distributed.

Equal Variance This plot of the residuals (Figure 4) offers no evidence of an unequal variance.

2.5 Model Fit

Standard methods of evaluating model fit such as R^2, t values, and F statistics cannot be used in this context of LASSO-reduced high dimensionality. Instead, we consider the cross-validation MSE associated with different values of the shrinkage factor λ. We choose λ such that it minimizes this MSE. In our analysis, we used the mean of the λ values from 100

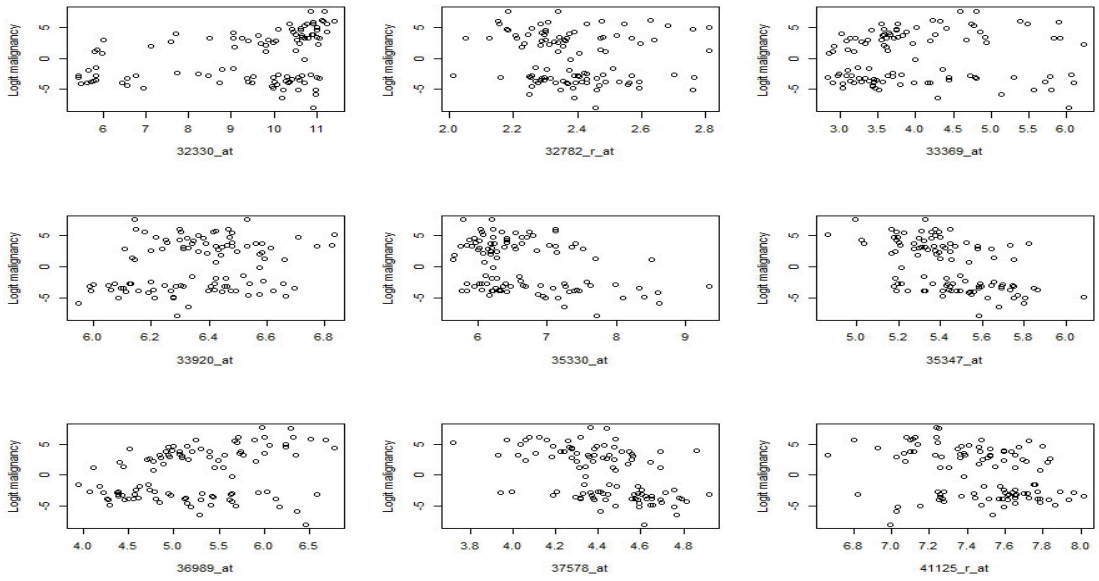


Fig. 2: Scatterplot matrix (Figure 2) validates linearity assumption

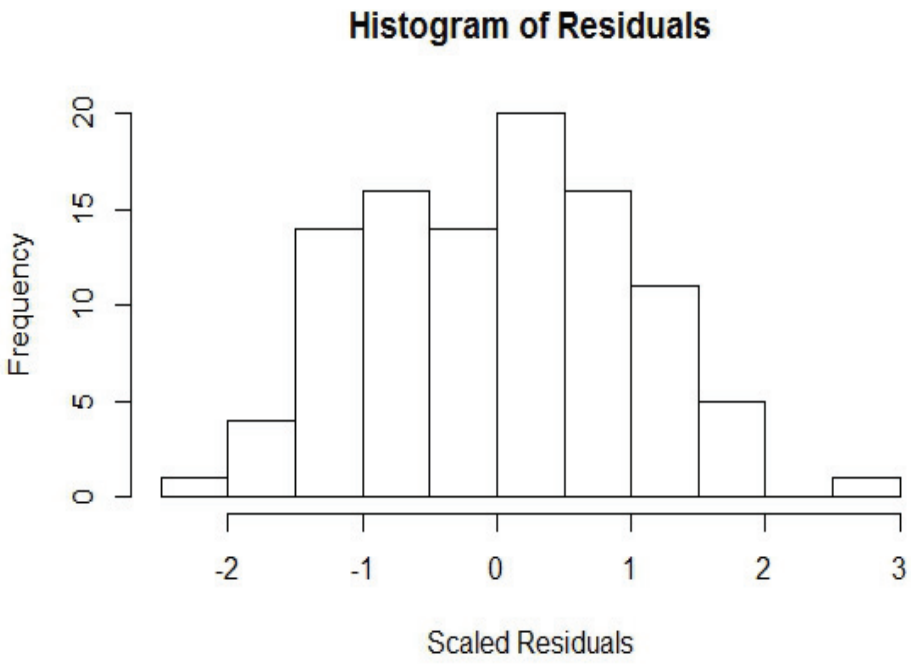


Fig. 3: Histogram of residuals validates normality assumption

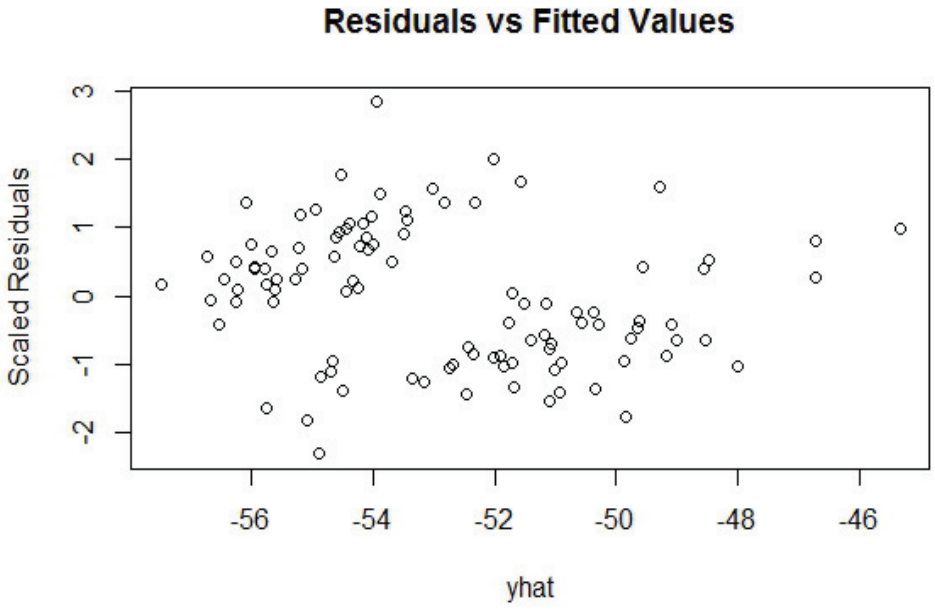


Fig. 4: Plot of residuals vs fitted values validates equal variance assumption

different cross-validated training sets so as not to have as much variability in λ and make the results reproducible. This method of minimizing MSE suggested that the LASSO method was executed well and that the resulting model fit the data well. Figure 5 shows a comparison of MSE for different values of λ.

3 Results

After using the method as described above, we received the following results in Table 1. Only the genes that had a statistically significant on cancer are shown.

In order to interpret the estimates for the coefficients, it's important to remember that a logit transformation was used on the Malignant variable. That is $logit(Malignant) = log(\frac{Malignant}{1-Malignant})$. The fraction of $\frac{Malignant}{1-Malignant}$ is often referred to as the odds ratio, and in this case, an odds ratio greater than one indicates a higher likelihood of cancer, while an odds ratio less than one indicates a lower likelihood of cancer. Having this in mind, a back

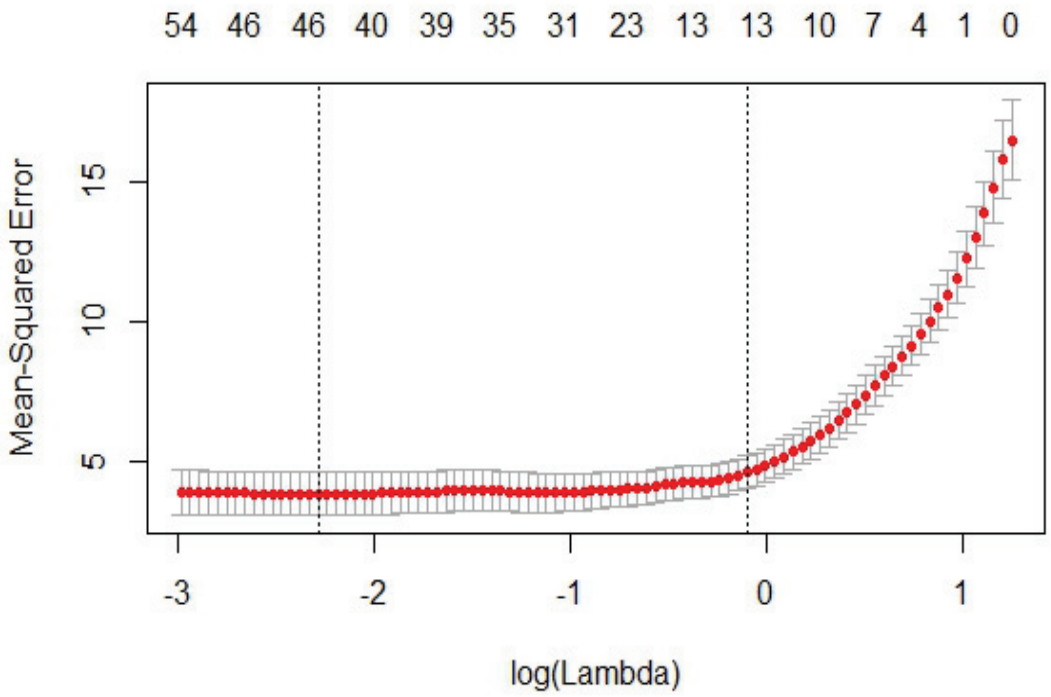


Fig. 5: λ was chosen to minimize MSE

Genes	Estimates	Standard.Error	Lower.Bound	Upper.Bound
1 x32330_at	0.365	0.176	0.020	0.709
2 x32782_r_at	-2.144	1.088	-4.276	-0.011
3 x33369_at	-1.030	0.390	-1.795	-0.266
4 x33920_at	2.400	0.636	1.153	3.646
5 x35330_at	-0.812	0.177	-1.159	-0.465
6 x35347_at	-4.712	0.177	-5.059	-4.365
7 x36989_at	1.459	0.497	0.484	2.434
8 x37578_at	-4.021	0.314	-4.636	-3.406
9 x41125_r_at	-2.836	1.407	-5.593	-0.078

Table 1: Coefficient Estimates of Genes that had a significant effect on Cancer

transformation would yield the following interpretation: as the gene expression for a given gene increases by one unit, we would expect the odds ratio of Malignant (given that the other gene expressions are held constant) to go up or down by $e^{estimate}\%$. By this interpretation the positive estimates correspond to genes that have an increasing relationship with cancer, while the negative estimates correspond to go genes that have a decreasing relationship with cancer.

The lower and upper bounds in the table represent the ends of the 95% confidence interval for our estimates. For example, with gene x32330_at, we are 95% confident that the estimate for that gene is in between .020 and .709. Or if we were to back transform the estimate through exponentiation, we could say we are 95% confident that a one unit increase for the gene expression of x32330_at would correspond to an increase of the odds ratio Malignant anywhere from from 22.1% to 101.4%.

Again, the purpose of this study was to identify the genes that would possibly be correlated with cancer. Because we used the LASSO method to analyze this high dimensional data, we were able to appropriately constrain many of the estimates to be zero so that we were left with smaller subset of variables that might be correlated with cancer. By creating confidence intervals through a bootstrap method, we were able to determine which of those genes had a statistical significant effect on the likelihood of cancer. This analysis meets the goal of our study as the 9 genes in the Table 1 represent those genes that we found are associated with cancer.

4 Conclusion

Again, by using the LASSO method to find estimates, and by using a bootstrapping technique to find confidence for those intervals, we were able to identify genes that had a significant effect on cancer. Those genes were X32330_at, X32782_r_at, X33369_at, X33920_at, X35330_at, X35347_at, X36989_at, X37578_at, and X41125_r_at.

However with this said, we realize that the high dimensionality of our data set (i.e. we have way more variables than observations) might hurt the validity of our results. If we were to do repeat this study with a different data set and the same method, we might get entirely different results, even though the method we used is valid. This is because of the phenomenon that many statisticians refer to as the curse of dimensionality. If we had more observations in our data, we would have had a more accurate answer.

It's also important to note that while the LASSO method has some advantages over other methods we could have used, it also has some disadvantages. One advantage of the LASSO method is that reduces many of the effects of the explanatory variables to zero, making the results much simpler. We thought this was important because eliminating the unneeded variables would make identifying genes that were associated with cancer much easier. Another method we could have used was principle component regression which would have addressed issues of collinearity a bit better. Principle component regression also usually has higher predictive power than the LASSO method. However, this wouldn't constrain any of the estimates to go down to zero, leaving us with a huge subset of possible variables that might be related to cancer, making identification more complicated. But, it still might be a good idea to do another analysis using principle component regression and see how our results compare.

Evaluation of Ankle Taping Methodology

Stephen Merrill

Brigham Young University

1 Introduction

The use of athletic tape to protect vulnerable muscles is an established physical therapy technique. This study attempts to quantify the effects of three different taping methods: tape casting, air casting, and tape bracing. These techniques will be investigated in a mixed model setting, using data from measurements of ankle inversion, an adverse condition present in rolled ankles, collected before and after exercise was performed. The data was gathered by treating 16 subjects with each technique as well as a control of no taping. For each of these four treatments, five measurements were taken before and after one hour of exercise was performed. There are therefore 640 observations in the data set, with the response being a measurement of ankle inversion with larger negative values indicating more severe inversion.

The goal of this study is to answer the five following questions:

- Do taping methods effect ankle inversion?
- Is one taping method superior prior to exercise?
- Does exercise diminish the effect of taping?
- Does the effect of different taping methods differ with exercise?
- Which taping method is most effective?

2 Methods

To accomplish the goal of this study, I considered the data in a mixed model setting and fit several different models with slightly different structures. Of these, I chose to use a random coefficients model, based on a BIC comparison. This model allows for analysis to be done on fixed effects, comprising the intercept, which is the treatment effect, and the slope, which changes due to exercise. The model can be defined succinctly as follows:

$$y_{ij} = \beta_{0i} + \beta_{1i}X + U_{0j} + U_{1j}X + \epsilon_{ij} \tag{1}$$

Where β is defined as the slope and intercept of the fixed effects and U defines the random effects, which are the subjects. This model accounts for the correlation between slope and intercept and can also be used to predict for individual subjects, if desired.

With a model defined, I approached each of the five motivating questions. Each was answered with consideration to an appropriate F or Likelihood Ratio Test. The details are given below.

3 Results

Do taping methods effect ankle inversion? Examination of Figure 1 shows that there is an apparent difference between the first column (control) and the others (taping), but exercise effect and differences in taping methods are not obvious. An F-test of the null hypothesis, that each treatment is equal, $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ yielded a p-value $< .0001$ and the null can therefore be rejected in favor of the conclusion that there is a treatment effect and taping is an effective therapy technique.

Is one taping method superior prior to exercise? Figure 2 makes it clear that before exercise, taping methods clearly perform better than the control, but determining which method is most effective will require more analysis. I preformed multiple F-tests to compare pairs of

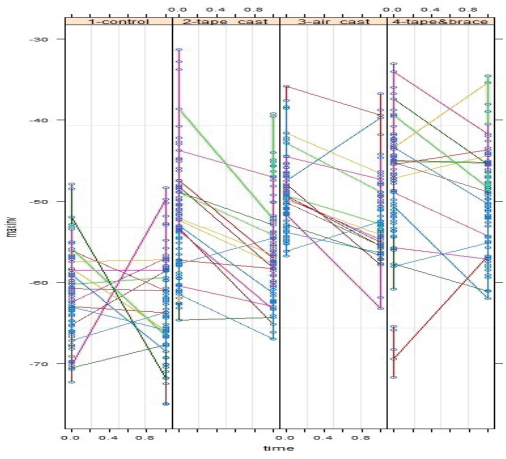


Fig. 1: Comparison of each treatment across exercise

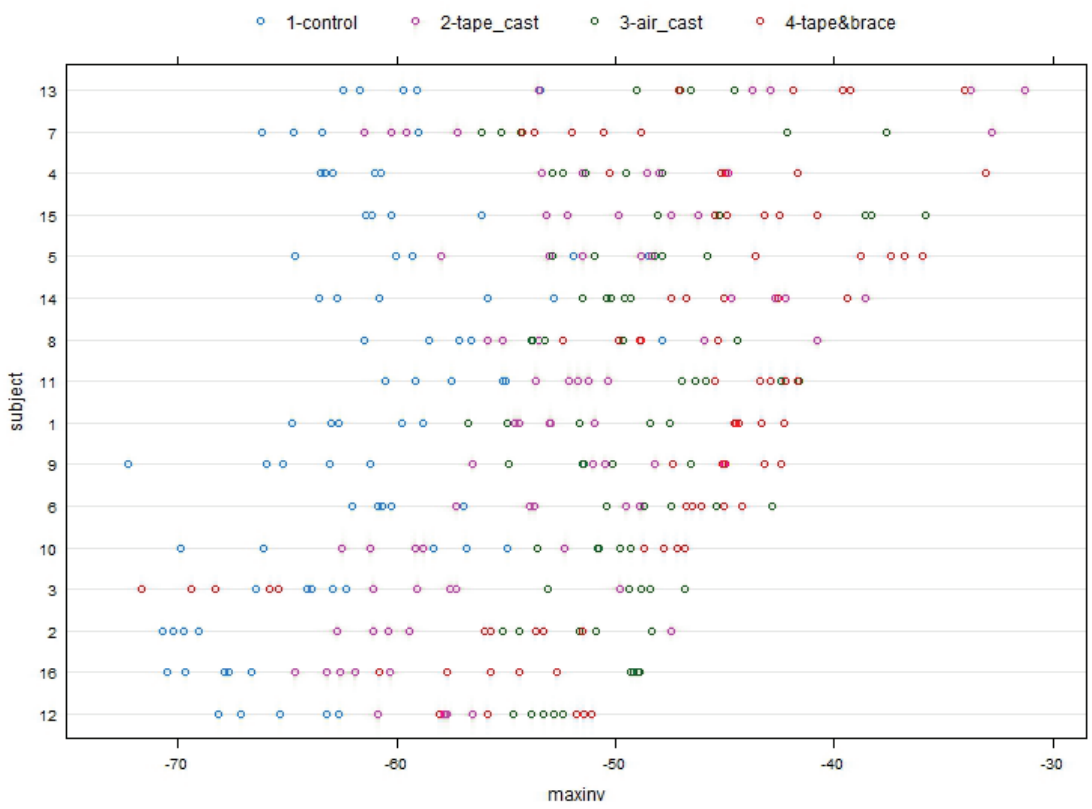


Fig. 2: Comparison before exercise

treatment intercepts and found significant differences in all treatments except for air casting and tape bracing. These two methods are both superior to the others prior to exercise, but there is not enough evidence to conclude which of them is superior overall.

Does exercise diminish the effect of taping? Figure 3 suggests that regardless of treatment, inversion seems to decrease after exercise. This impression was validated by a Likelihood Ratio Test between the full model and a reduced model without the fixed effect for the slope included. The result of this test was a p-value $< .0001$ and I therefore concluded that the full model that included the exercise effect was necessary.

Does the effect of different taping methods differ with exercise? I tested this interaction between treatments and exercise with multiple F-tests in the same manner as I did with testing the treatments prior to exercise. However, I found that there was only a significant

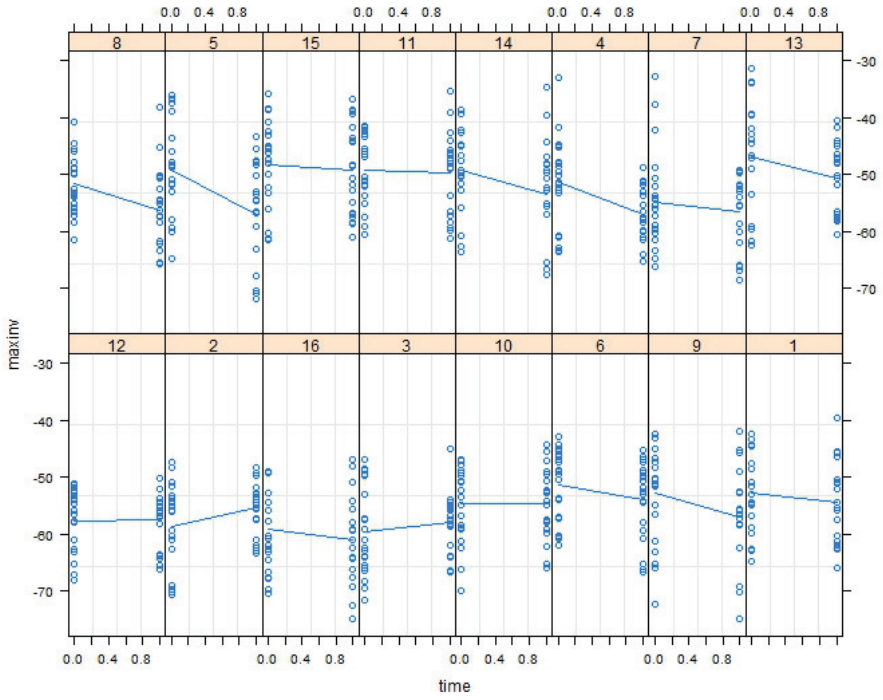


Fig. 3: All observations before and after exercise, by subject

difference in slope between control and tape casting and control and air casting. There was not enough evidence to suggest that any of the other slopes were unequal.

Which taping method is most effective? Two methods, air casting and tape bracing are superior to the others prior to exercise. However, there was not enough evidence to prove that one was better than the other, nor was there evidence that one performed better after exercise than the other. Therefore, I conclude that air casting and tape bracing are the two best methods, but further testing is required to determine which is the best method overall.

4 Conclusion

Figure 4 summarizes the results by plotting each treatment effect, as fit by the model, and the effect due to exercise. It is apparent by the overlapping confidence regions that although the treatments appear to be dissimilar there is too much variability to conclude that one treatment is the best at decreasing inversion overall. The best this model can do is conclude that air casting and tape bracing are superior to both tape casting and no taping.

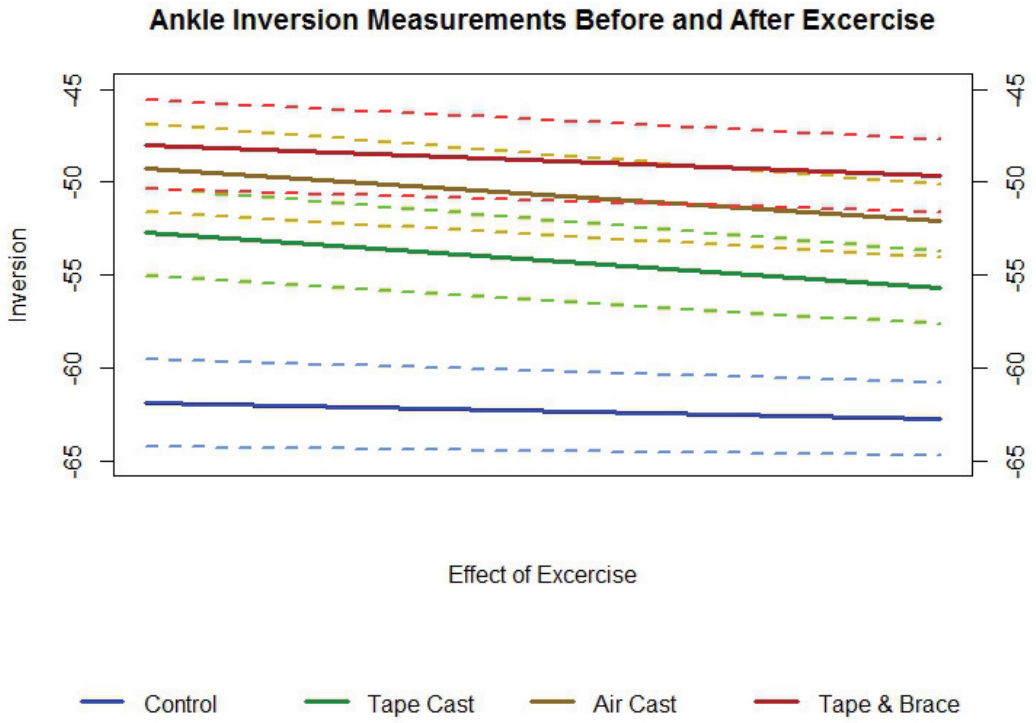


Fig. 4: Model-fitted effects for each treatment, with 95% confidence intervals

This model could be improved with a Bayesian analysis. Several tests were nearly significant, which means that this Frequentist approach limits the power of my results. If I was able to compare Bayesian posterior distributions, however, I would be able to directly report the probability of one treatment being superior to another.

Mixed Model Simulation Study

Stephen Merrill

Brigham Young University

1 Introduction

In a repeated measures setting, there are multiple ways to test treatment effects in a model. In this paper we investigate the use of the Likelihood Ratio Test (LRT) and F-test in determining if the treatments are different from one another in terms of slope across timed measurements. To compare these two tests, we will implement a simulation study with varying amounts of subjects, time measurements, and models. The goal of this study is to determine which test is better to use when testing treatment effects, and if there are certain situations where it is better to use one over the other.

2 Methods

To accomplish the goal of this paper, we simulated repeated measures data under an AR1 structure, assuming a $\rho = 0.8$. A repeated measures data set contains different response measures for a subject at each time measurement. For example, a person's weight could be measured once a week, and a certain amount of subjects could be assigned different diets so that a diet effect could be estimated from the data. For each different run of the simulation we considered all pairwise combinations of the following variables in a three treatment study scenario.

- Subjects (30,60,90)
- Time measurements (3,4,5)
- Treatment slopes
  - Slopes equal (set each at one)
  - Slopes slightly different (0.95, 1, 1.05)
  - Slopes different (0.9, 1, 1.1)

To help visualize the simulated data, Table 1 represents the first nine observations of a simulated data set that contains 30 subjects with three time measurements at equal slopes.

Table 1: Simulated Data Example

subject	trt	time	response
1	1	1	8.32
1	1	2	9.14
1	1	3	12.02
2	1	1	10.71
2	1	2	10.89
2	1	3	12.89
3	1	1	9.66
3	1	2	11.88
3	1	3	11.75

We also wanted to understand which significance test performs best in each of the different models, so we estimated the fixed effects using an AR1 model and a random coefficients model. Although the data was generated under an AR1 structure, the other models can estimate the fixed effects accurately. For the purpose of this study, we will assume that the random effects and variance components estimated are correct, though we understand that there is ample discussion related to the choice of model that best estimates the random effects and variance components.

For each of the different simulations, we will generate 1000 data sets, then perform a LRT and F-test for each of the simulated data sets to assess if the treatments are significantly different from one another. The LRT will consist of building a model with only one slope assuming the treatments are equal and another "full" model that contains a treatment slope effect for each treatment, and then comparing the likelihoods of the two different models. The F-test will consist of testing the null hypothesis that all treatments are equal using a contrast matrix, yielding a p-value that will determine if at least one of the treatments is different. The denominator degrees of freedom used for the F-test will be the result of Equation 1, where q is the number of subjects, t is the number of treatments, and r is the

number of variance components estimated by the model. This *df* is more conservative, but we feel it is the most intuitive definition of denominator *df*.

$$df_{denom} = q - t - r \tag{1}$$

After the simulations are run, the two different tests will be compared to understand which test is better in the different circumstances. The best test will be determined by a combination of Type I error and power, subject to an α level $= 0.05$, shown in the simulations.

3 Results

The goal of this simulation study and analysis was to compare the performance of the LRT and the F-test in different situations. The performance of both tests was assessed using tables 2, 3 and 4, which contain the percentage of the 1000 runs that resulted in a significant test statistic. Table 2 gives us the Type I error rate and Tables 3 and 4 give us the power when the treatment slopes are slightly different and clearly different.

The Type I error rate of the LRT and F-test for both models did not vary too much as number of subjects or time measurements increased. For both models we saw that the LRT was closer to the desired Type I error rate, 0.05, than the F-test for almost all the pairwise combinations of subjects and time measurements. There was only two occasions, which are bolded, where the F-test Type I error rate was better or equally as good as the LRT Type I error rate. The power of the LRT and F-test, for both models and both slope differences, increased as number of subjects and time measurements increased. The power was higher when comparing datasets with a greater difference between treatment slopes, which we would expect. For both models and slope differences, the power for the LRT is higher than the power for the F test for every number of subjects and time measurements.

4 Conclusion

We ran this simulation study in order to determine whether the Likelihood Ratio Test or the F-test performed better in finding significant differences between treatments in a repeated measures setting. Based on the results of our simulation study, we would recommend

Table 2: Equal Treatments

(a) AR1 Model

		3 Times	4 Times	5 Times
30 Subjects	LRT	.047	.039	.054
	F	.028	.029	.039
60 Subjects	LRT	.057	.051	.057
	F	.041	.033	.040
90 Subjects	LRT	.043	.044	.047
	F	.026	.027	.035

(b) Random Coefficients Model

		3 Times	4 Times	5 Times
30 Subjects	LRT	.047	.055	.041
	F	.033	.042	.032
60 Subjects	LRT	.049	.061	.059
	F	.035	**.044**	**.041**
90 Subjects	LRT	.052	.045	.056
	F	.034	.030	.039

Table 3: Slightly Different Treatments

(a) AR1 Model

		3 Times	4 Times	5 Times
30 Subjects	LRT	.099	.121	.146
	F	.071	.089	.115
60 Subjects	LRT	.111	.171	.247
	F	.082	.126	.181
90 Subjects	LRT	.167	.245	.335
	F	.134	.200	.282

(b) Random Coefficients Model

		3 Times	4 Times	5 Times
30 Subjects	LRT	.084	.115	.123
	F	.067	.095	.097
60 Subjects	LRT	.110	.185	.221
	F	.094	.160	.185
90 Subjects	LRT	.152	.228	.327
	F	.122	.195	.267

Table 4: Different Treatment

(a) AR1 Model

		3 Times	4 Times	5 Times
30 Subjects	LRT	.198	.305	.427
	F	.161	.232	.365
60 Subjects	LRT	.371	.571	.719
	F	.306	.498	.672
90 Subjects	LRT	.499	.760	.879
	F	.435	.700	.848

(b) Random Coefficients Model

		3 Times	4 Times	5 Times
30 Subjects	LRT	.196	.319	.401
	F	.167	.265	.348
60 Subjects	LRT	.342	.546	.681
	F	.298	.483	.635
90 Subjects	LRT	.544	.707	.875
	F	.475	.665	.844

the use of the Likelihood Ratio Test over the F test when dealing with repeated measures data. As was stated in the results, the Likelihood Ratio Test consistently outperformed the F test in obtaining the desired Type I error rate and in power.

We only included two different models, a random coefficient model and another with an AR1 covariance structure, in our simulation study so our analysis could not include all possible simulations. For further investigation into the effectiveness of the LRT and the F-

test we could consider different covariance structures such as compound symmetry, Toeplitz or unstructured. We could also look to change the covariance structure when generating the data. We used an AR1 covariance structure, assuming a $\rho = 0.8$, but we could use any of the aforementioned covariance structures to simulate the data. If different F-tests needed to be compared using different denominator *df* then a similar simulation could also be performed. Overall, we were pleased with the results of this study and the insight it provided into the most effective way to test treatment effects in a repeated measures setting.

Hypergeometric Distribution

Stephen Merrill

Brigham Young University

1 Introduction

The Hypergeometric Distribution is a discrete probability distribution that is used to describe the probability of a number of successes in a sequence of draws. The draws are done without replacement from a finite population with a specified number of success states. Each draw is either a success or failure, or a Bernoulli random variable. The distribution has three parameters: the population size, the number of success states, and the number of draws.

History The Hypergeometric Distribution is named for the Hypergeometric Function, which appears in the CDF and moment generating function of the distribution. The first recorded use of the distribution was in 1657, in a solution to a problem in Huygens's *De Ratiociniis in Ludo Aleae*. James Bernoulli and de Moivre both provided solutions. However, the first use of the phrase "Hypergeometric Distribution" did not appear until 1936 in H. T. Gonin's article "The use of factorial moments in the treatment of the Hypergeometric Distribution and in tests for regression."

2 Methods

2.1 Distribution Description

PMF

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{1}$$

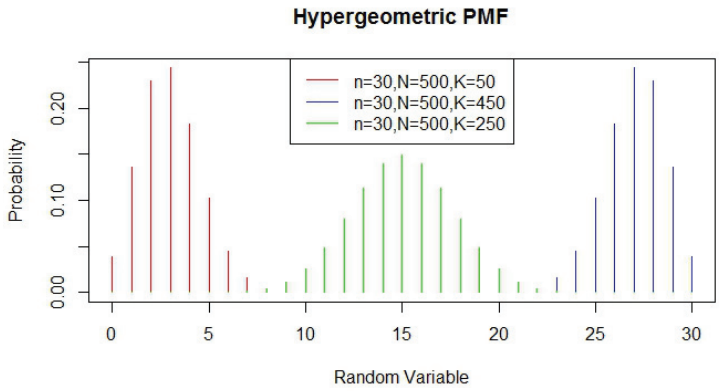
Support

$$k \in \{\max(0,\, n + K - N),\, \ldots,\, \min(n,\, K)\} \text{ – number of successes}$$

Parameters

- $N \in \{0, 1, 2, \ldots\}$ – population size
- $K \in \{0, 1, 2, \ldots, N\}$ – number of success states in the population
- $n \in \{0, 1, 2, \ldots, N\}$ – number of draws

PMF Graph



CDF

$$1 - \frac{\binom{n}{k+1}\binom{N-n}{K-k-1}}{\binom{N}{K}} \, {}_3F_2\left[\begin{matrix} 1,\ k+1-K,\ k+1-n \\ k+2,\ N+k+2-K-n \end{matrix}; 1\right] \tag{2}$$

Mean

$$n\frac{K}{N}$$

Variance

$$n\frac{K}{N}\frac{(N-K)}{N}\frac{N-n}{N-1}$$

MGF

$$\frac{\binom{N-K}{n} \, {}_2F_1(-n, -K; N-K-n+1; e^t)}{\binom{N}{n}} \tag{3}$$

Hypergeometric Function

The ${}_pF_q$ notation seen in (2) and (3) is defined as the *Hypergeometric Function*, with the general form:

$${}_pF_q(a_1, ..., a_p; b_1, ..., b_q; z) = \sum_{k=0}^{\infty} \prod_{i=1}^{p} \frac{\Gamma(k+a_i)}{\Gamma(a_i)} \prod_{j=1}^{q} \frac{\Gamma(b_j)}{\Gamma(k+b_j)} \frac{z^k}{k!} \tag{4}$$

3 Results

3.1 Special Cases and Relationships

Let $X \sim$ Hypergeometric(K, N, n) and $p = \frac{K}{N}$.

Bernoulli

If $n = 1$, then X has a Bernoulli Distribution with parameter p.

Binomial

Let $Y \sim$ Binomial(n, p)

If N and K are large compared to n, then X and Y have similar distributions and $P(X \leq k) \approx P(Y \leq k)$.

Normal

If n is large, N and K are large compared to n, and p is not close to 0 or 1, then:

$$P(X \leq k) \approx \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right)$$

Where Φ is the standard normal distribution function

4 Conclusion

4.1 Common Uses

The Hypergeometric Distribution is commonly seen in instructional problems involving balls and urns, but more useful applications also exist.

Texas Hold'em Probability calculations in Poker are often made with the Hypergeometric Distribution. By identifying N, the number of cards available to draw, K, the number of ways to achieve a success, n, the number of cards to draw, and k, the number of successes needed, a poker player can calculate their probability of success using the Hypergeometric PMF.

Subpopulation Representation A Hypergeometric test, which uses the Hypergeometric Distribution to calculate the test statistic, can identify if subpopulations are over or under

represented in a sample. This is useful for identifying if different demographic groups are being properly represented in a sampling study.

4.2 Published Journal Example

In a study published as "Group Testing, the Pooled Hypergeometric Distribution, and Estimating the Number of Defectives in Small Populations" (Theobald & Davie) the Hypergeometric Distribution is used to propose a new distribution called the Pooled Hypergeometric Distribution. This distribution is characterized by grouped samples from equal-sized groups and is used to effectively reduce the cost of identifying defective individuals in populations containing small proportions of detectives. It is used especially to identify the prevalence of rare diseases that are difficult to detect.

4.3 Homework Problem

In Poker, players make the best hand they can combining the two cards in their hand with the five cards that are eventually turned up on the table. Assume a player has two hearts in his hand and there are three cards showing on the table, two of which are also hearts. What is the probability that one of the next two cards to be shown is also a heart, completing his flush?

Solution Define the parameter values and compute the probability using the Hypergeometric PMF.

- $N = 47$ – Number of total unseen cards
- $K = 9$ – Number of unseen hearts
- $n = 2$ – Number of cards to be flipped
- $k = 1$ – Number of hearts needed to achieve success

$$P(X = 1) = \frac{\binom{9}{1}\binom{47-9}{2-1}}{\binom{47}{2}} = .316$$

Car Prices Case Study

Stephen Merrill

Brigham Young University

Abstract. The price at which a used car will sell is a matter of debate between the salesman and the customer. Often, used car dealers receive a variety of cars into their lots, and then they must determine the price at which they will offer to sell the vehicle. Since this may be a complicated process, we use regression from data on used car sells in order to find out which factors best aid in predicting the price for which a used car will sell. As there are non-linearities in the data and with some of the explanatory variables, a generalized additive is used with a natural spline on one of the variables. Best subset selection with cross-validation is used in order to determine which covariates should be used. After cross-validating the model, we find that the model fits the data fairly well and has good predictive accuracy. The selling price is determined mostly by the manufacturing year, weight, number of miles, horsepower, manufacture, automatic air conditioning, and powered windows. Therefore, both internal and external features are important in the future selling price of any given used car.

1 Introduction

Used car dealerships seek to swindle their customers by purchasing a used car at a low price and then reselling the car at a much higher price. If the dealerships know the value a car will sell for, they will hold an advantage over their customer. Therefore, in order to aid these in-famous car salesmen, our purpose is to create a model that will predict the future selling price of a used car given characteristics of the car.

Although we have a data set of 1436 observations and 21 descriptive explanatory variables about the characteristics of the cars, this data set contains several problems. Most noteably is the clear non-linear relationship between Miles and Price. In order to build a model, this non-linearity needs to be addressed in some manner. There are also apparent outliers in the data. One, in the cc data, was a misprint and was easily corrected. However, others in the Horsepower data were not as easily rectified and will be considered in the results. Finally, some data cleanup was necessary. The cylinder data was found to be uninformative and thus removed. Additionally, we were considered about sparse observations in categories in the Color, Doors, and Air Conditioning data. However, none of the data combination

techniques we explored seemed to provide extra accurateness or effectiveness, so we left this data as is. A brief glance at the data can be seen in Figure 1.

The goal of this analysis is to predict the future used car selling price, given characteristics about the used car.

2 Methods

2.1 Generalized Additive Model (GAM)

Since there are variables with a nonlinear relationship with respect to the selling price, we need to account for them using some type of linear regresssion. We choose to use natural splines to account for these variables, because we can still use parametric functions to interpret them as well as they predict well on the tails of the distributions.

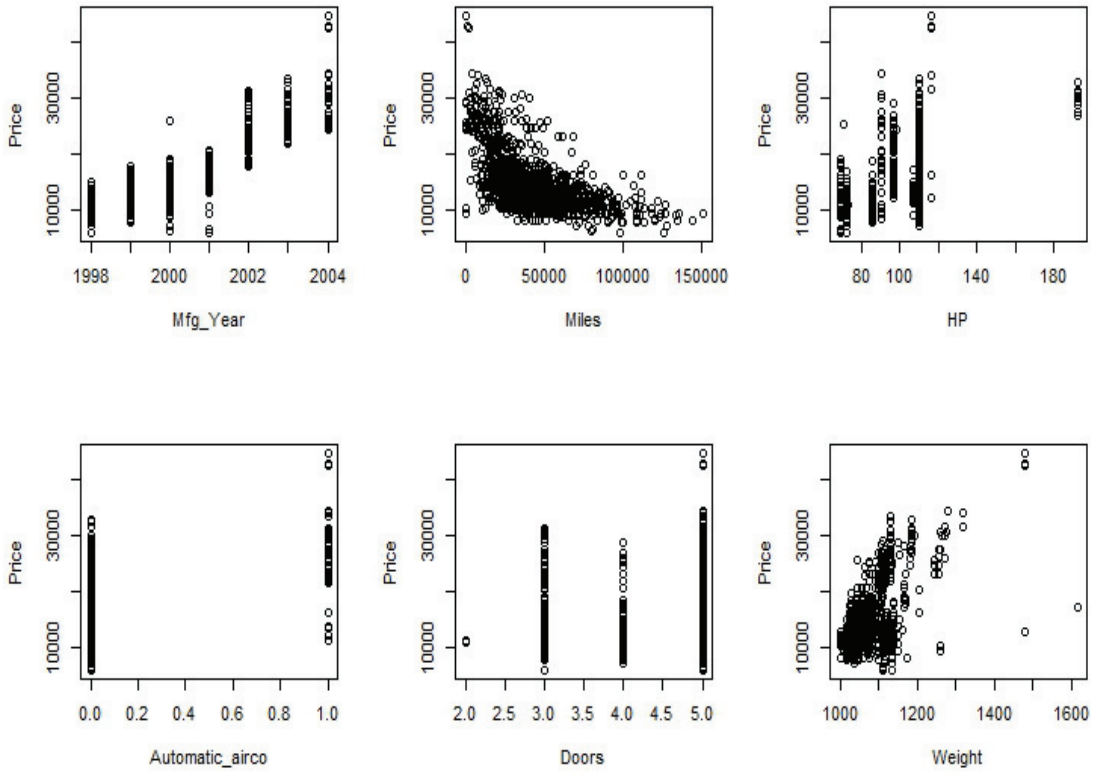


Fig. 1: Scatterplot matrix of selected covariates and price

Then, we use a Generalized Additive Model (GAM), because it is an extension of the standard linear model but additionally allows for the use of non-linear functions, such as we have with this data set.

Our GAM model is defined by functions for each variable:

$$y_i = \beta_0 + \sum_{p=1}^{P} f_p(x_p) + \epsilon_i \tag{1}$$

$\epsilon_i \sim N(0, \sigma^2 I)$

f_1 is a Natural spline for the Miles variable with three degrees of freedom and two knots. $f_2 \dots f_{P=6}$ are each linear functions with respect to the explanatory variables of Manufacturing Year, Weight, HP, Manufacture Guarantee, Automatic Air-conditioning, and Powered Windows. The generalized additive model includes an overall intercept and an overall error term as well as a coefficient measuring the slope for each of the explanatory variables.

The β_0 represents the price when all of the predictor variables are equal to zero. Each of the β_i in the linear functions of the Generalized Additive Model are representations of slope, meaning they are the amount that the Price will increase by for each one unit increase in each explanatory variable, respectively.

Our model allows for both inference and prediction to be made. For this problem, interest lies in making predictions of used car sales prices. The model accomplishes this goal by determining the effect sizes (the β values) of the significant variables outlined above, as well as the relationship between Price and Miles. Once determined, prediction for balance can be made by gathering data on the cars and making calculations according to the model.

2.2 Model Assumptions

Since we are using a Generalized Additive Model, the model assumptions are the same as those for a linear model.

Linearity Each variable must have a linear relationship with the response. If this is not the case, the entire model is invalid since it would be fitting a line to non-linear data. However, we know that Miles is non-linear with Price, and we've fit a natural spline accordingly. Therefore, this assumption only needs to hold for the linear functions in the GAM.

Independence The data must be independent. If this assumption is violated, measures of variability will typically be too small. This is a difficult assumption to verify. Usually prior knowledge of the data is required.

Normality The errors must be normally distributed. Otherwise, confidence and prediction intervals that depend on t distributions are incorrect.

Equal Variance The errors also must have equal variances. Without this, measures of variability will once again be invalid.

2.3 Model Justification and Performance

We used two Cross-Validation techniques in order to build the model. For each technique, we randomly generated 100 different test and training sets from the original data set in order to have consistent results.

We first used Cross-Validation to determine the optimal number of degrees of freedom to use in the natural spline for Miles. This was determined to be three. Results can be seen in Figure 2.

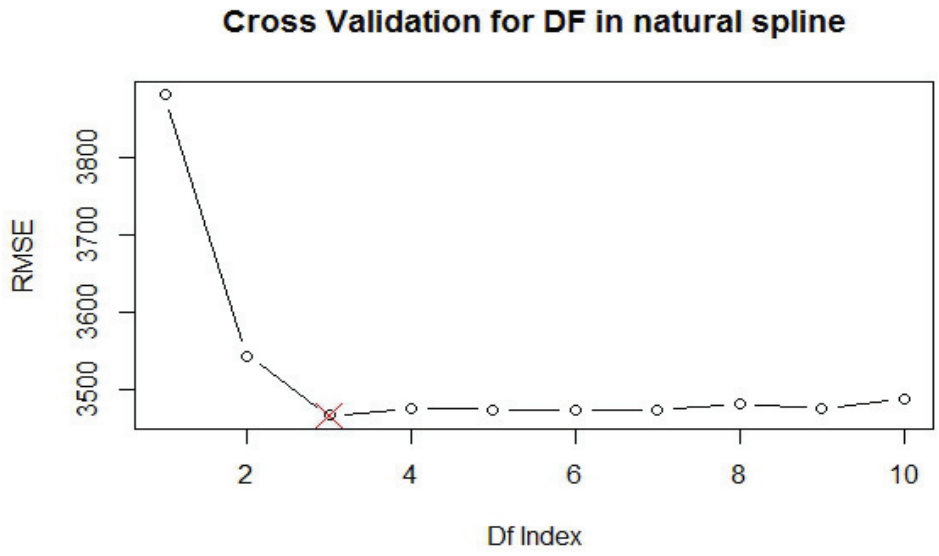


Fig. 2: Three degrees of freedom in the natural spline minimized RMSE

We then used best subset selection in order to select the functions for our GAM model. This process selected nine variables - the natural spline, which counted as three due to its degrees of freedom, and the other six variables, which we considered linear functions in order to add them to the GAM. We therefore fit the GAM with seven functions. Results can be seen in Figure 3.

2.4 Model Assumptions Verification

Linearity The scatterplot matrix (Figure 4) shows a vague linear relationship between each of the nine explanatory variables and the response. No other kind of relationship is prominent, so we considered this assumption to be met. As previously discussed, Miles is not considered.

Independence Independence is difficult to determine. Because there is no prior knowledge that would suggest a violation of this assumption, it is assumed to be met.

Normality This histogram of the residuals (Figure 5) show that the errors are skewed and not precisely normally distributed. This means that our interval estimates that depend on t distributions will be inaccurate.This also implies that the standard errors will be inflated.

Equal Variance This plot of the residuals (Figure 6) offers no evidence of an unequal variance.

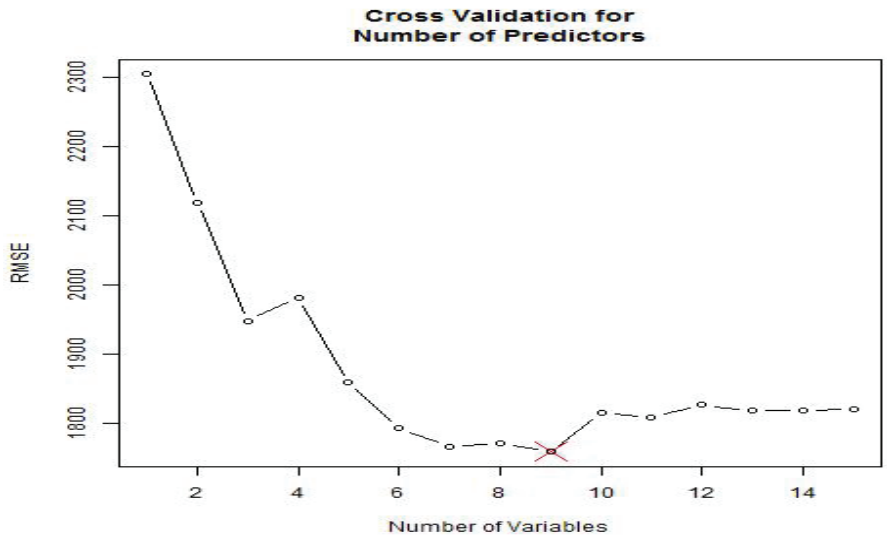


Fig. 3: Best subset indicated nine variables minimized RMSE

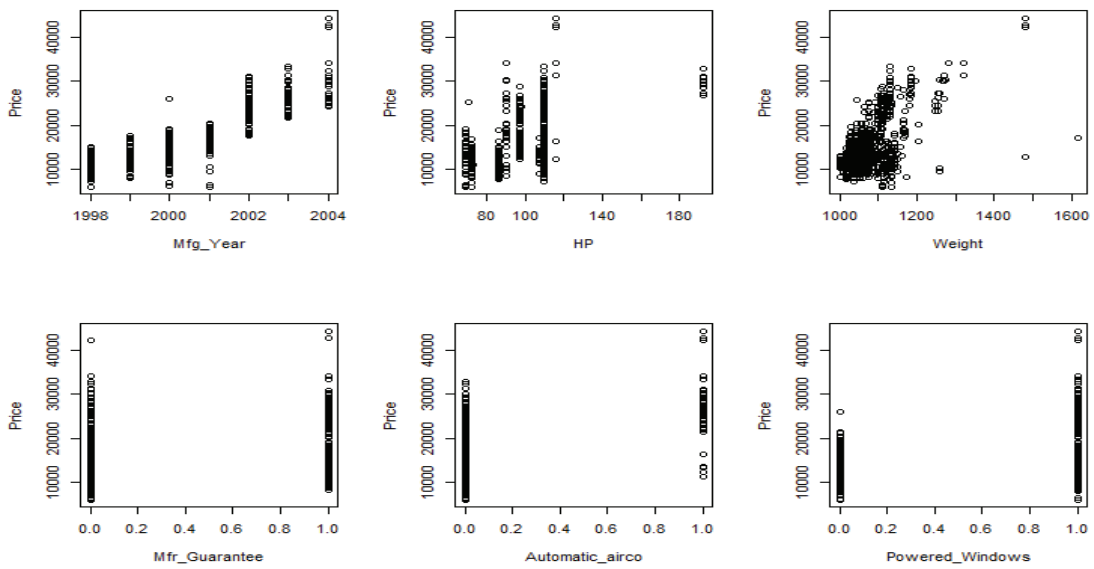


Fig. 4: Scatterplot matrix (Figure 2) validates linearity assumption

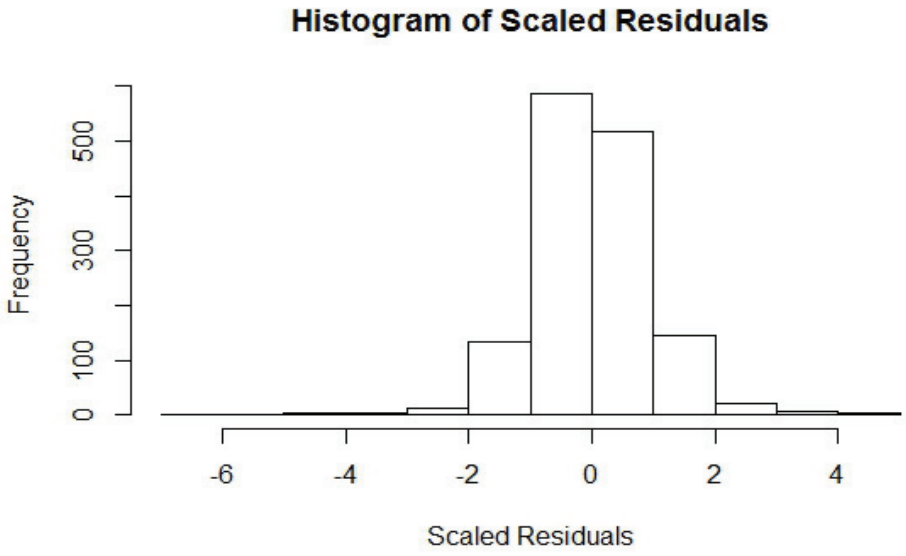


Fig. 5: Histogram of residuals shows concern for normality assumption

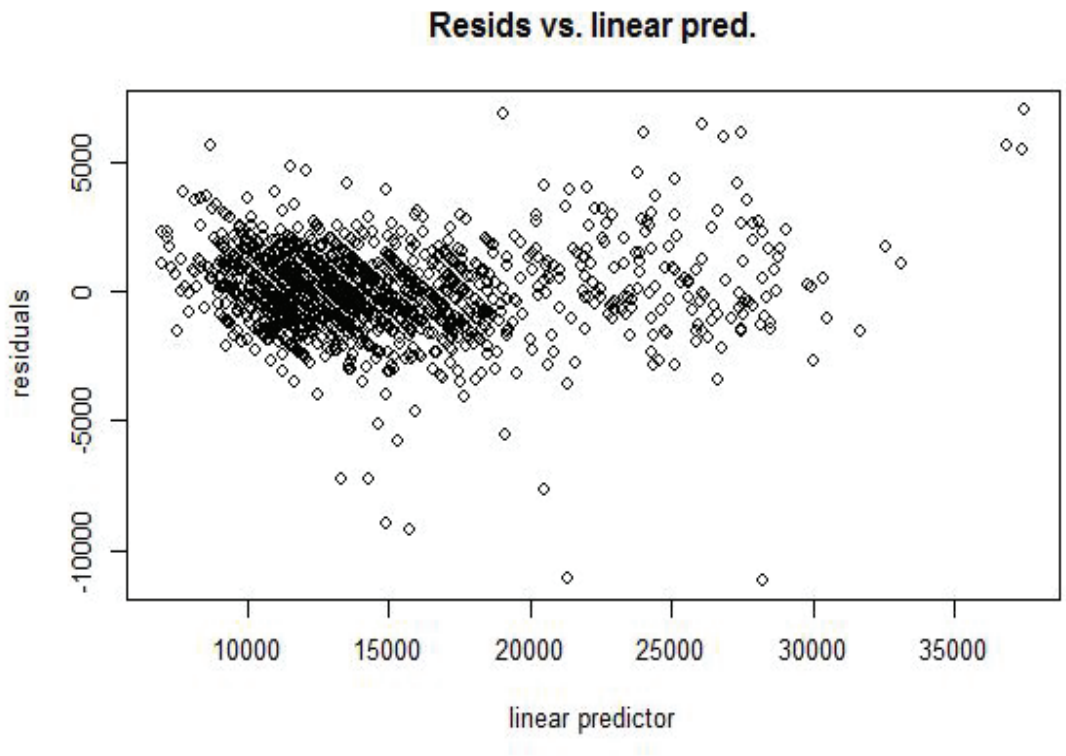


Fig. 6: Plot of residuals vs fitted values validates equal variance assumption

2.5 Model Fit and Prediction

The model fits the data as determined by the minimization of RMSE during Cross-Validation techniques, and R^2, which had a value of .8959. This means that 89% of the variability in Price is explained by the variability in all of the covariates.

The goal of this study was to predict the price of used cars. Using Cross-Validated testing and training sets, we simulated prediction for many different test sets and recorded the results in Table 1 and Figure 7. Of particular interest is Confidence Interval coverage, which should be 95% but may have been effected by the lack of normality in the residuals.

Notice that although some of these results seem quite large, the Price variable was also large and quite variable, and this may have some impact upon the results. Some of these intervals are more tight than others which may reflect the lack of meeting of all the assumptions.

Table 1: Prediction Diagnostics

	Estimate	Lower	Upper
Bias	-28.773	-112.147	54.602
MSE	2,341,443.000	2,227,397.000	2,455,489.000
RMSE	1,530.177	1,492.743	1,567.612
Coverage	0.967	0.966	0.969
Interval Width	6,045.905	6,044.992	6,046.818

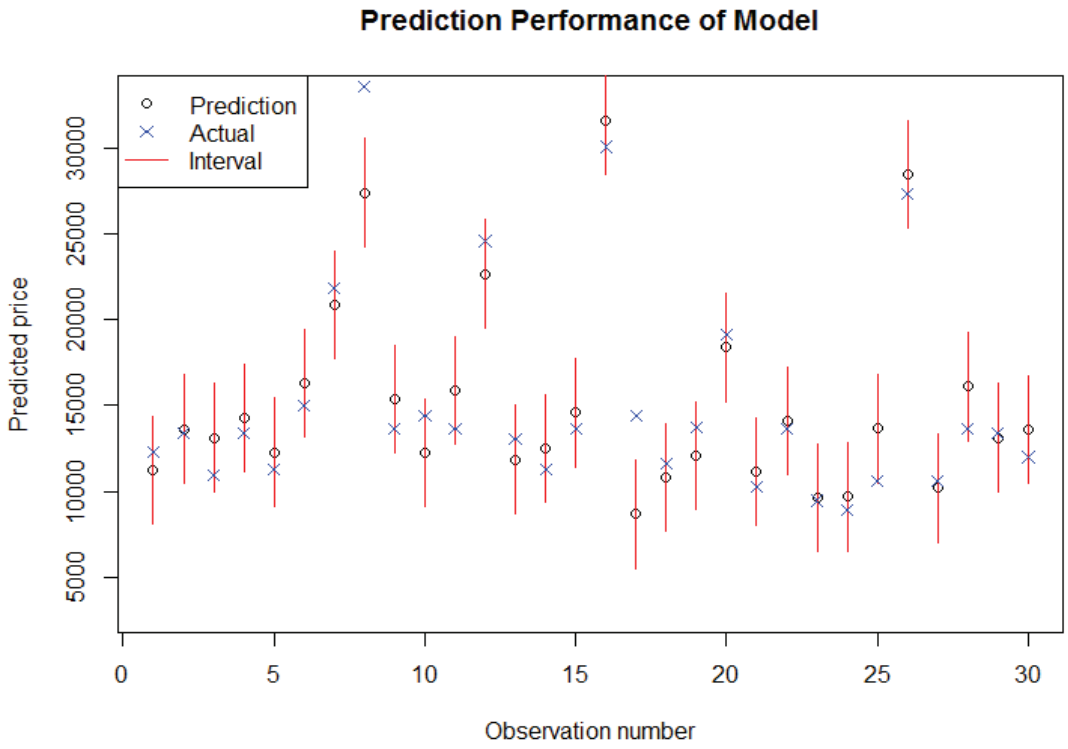


Fig. 7: Assessment of prediction with n=30 observations in the test set. Note that Confidence Interval coverage is 28 out of 30, 93.3%, but due to variability in these simulations, this number is not as informative as the coverage calculated in Table 1.

While our estimates are not exactly what we would expect, they are close, and from this we determine that our model predicts well on any given test set.

3 Results

This report adequately answers the questions posed in the case study, because we have used the data from the used car dealership in order to create a model that predicts the future sale price of any used car. The model is based upon natural splines and linear functions added together, yet this model seems to work will with a variety of combinations of test sets generated from the real data set.

Since we used a natural spline for the Miles variable as it appeared to have a nonlinear relationship with the Price response variable from our scatterplot as shown earlier, it is common to interpret a graph representing the relationship with Price and Miles and the effect of the natural spline. Overall, the effect of miles on price is decreasing. As shown in the graph of Price vs. Miles with the natural spline overlaid, it appears that the as the number of miles increases, the price decreases. In fact, this is what we would expect to happen, because usually cars who have been driven more miles are likely to be either well used or have more problems or both. In such cases, customers are less likely to buy the car and so we see a decrease in price. Notice also from the plot of Miles and Price with the natural spline, that the Price drops quickly soon after Miles increase from zero, but around 50000 miles, the price decreases at a slower rate than before. This seems to indicate that

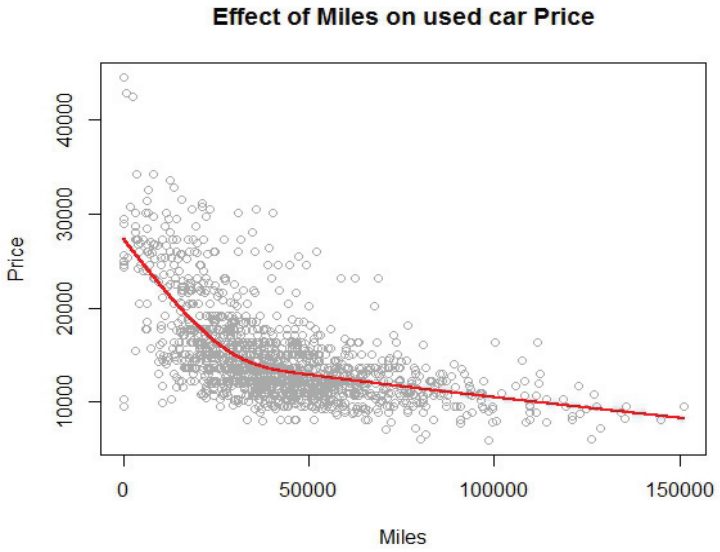


Fig. 8: The natural spline shows a decreasing price effect.

when a car has been driven a few miles, the amount of the drop in price is significant, but after reaching a large threshold of Miles, the decrease is less significant.

The other explanatory variables included in the model: Manufacturing Year, Weight, HP, Manufacture Guarantee, Automatic Air-conditioning, and Powered Windows are linear relationships with Price. Therefore, we can interpret the overall intercept from the Generalized Additive Model as well as each one of the β coefficients.

The parameter estimates are shown in Table 2. We found the estimates for the coefficients as well as 95% confidence intervals for each of the coefficients with the t-distribution and standard errors from the GAM model. These confidence intervals show our uncertainty. If the process of generating intervals was repeated many times, on average 95 percent of the intervals would contain the true value for the intercept and each of the listed slopes.

However, it must be noted that these estimates are quite wide, because the Normality assumption of the model was not completely met. There were a few outliers in the data set in terms of Horsepower and Weight that are likely affecting the Normality condition of the data set. Further investigation should be done with the dealership to determine the cause of these outliers.

The intercept is quite large, because it includes the effect of the year variable where the year is measured in the 2000s and so the intercept has to scale up for that. This means that if a car had zero of the explanatory variables, then the price would be $-\$3,728,966.000$. The interpretation of the intercept is not quite reasonable in the context of this problem.

Table 2: Parameter Estimates

	bhat	lower	upper
(Intercept)	-3,728,966.000	-3,893,027.000	-3,564,906.000
Mfg_Year	1,861.402	1,779.086	1,943.718
HP	26.094	19.833	32.356
Weight	19.463	17.482	21.443
Mfr_Guarantee	415.363	239.692	591.033
Automatic_airco	3,317.640	2,885.639	3,749.640
Powered_Windows	629.884	449.155	810.614

Instead, we focus on the predictions that this model is able to make, because that is the purpose of the problem.

We will interpret one of the β coefficients for an example with the others following similar interpretations.

For example, as the Manufacturing Year increases by one, the price will increase on average by $1,861.402$. Also, we are 95% confident that the true slope parameter for the Manufacturing Year lies between the interval of $(1,779.086,\ 1,943.718)$.

The other variables and their respective estimates for their contribution to the slope with their confidence intervals on the slope found from the standard errors of the slope coefficients on a t-distribution have similar interpretations as previously interpreted for Manufacturing Year.

Simply stated, the main points of the results are that we found an overall additive model for predicting the used car sale price. In this model, we included a nonlinear relationship of Miles and found that as the Miles increase the Price decreases. The other slope values for the variables of Manufacturing Year, Weight, HP, Manufacture Guarantee, Automatic Air-conditioning, and Powered Windows showed positive values, because the overall model included a very large negative intercept. The model also includes some random error to account for the uncertainty in the data and statistical method as well as to allow for the possibility of prediction by not creating a perfect fit of the data.

We found parameters which model the linear relationship of these variables - their slope coefficients and overall intercept with associated standard errors and confidence intervals. We also found a method to account for the non-linearity in the Miles variable and included this in the overall additive linear model. Finally, we found that despite some minor violations in the assumptions of the general additive linear model, the model performs well in predictions of used car selling price.

4 Conclusion

In summary, the goals of the analysis were met, because we created a model that is useful for predicting the selling price of used cars given some information about them, such as their Manufacturing Year, Weight, Miles, Horsepower, Manufacture , Automatic Air

conditioning, and Powered Windows. It appears that those variables are the most significant of all the possible variables and characteristics about the used cars that are available in the data set about the used cars. It would be interesting to see if the model and associated explanatory variables would change given more current data. The newest car in this data set was made in 2004, which was 12 years ago. Perhaps people would be more interested in a Board Computer with the recent increases in technology or a USB connection with the spread of smartphones. Or it may be that given another more current data set that these variables would still be the most important in determining selling price.

There are a few relevant shortcomings with this model and approach used. First, we were surprised that after running our model with Cross-validation approaches and Best-subset selection and even accounting for non-linearity in some variables by using a natural spline or using some variables as factors instead of continuous that the Normality conditions of the model were still not completely met. We believe this was due to the presence of some influential observations in the Horsepower and Weight variables.

However, in future work, we could consider adding interactions between the variables into the model. This would need to be done carefully with best subset selection in an effort to not over- the model with linear dependencies. We found that as we would add more variables to the model as factors, that some of them would be linearly dependent with each other and have to be combined. It is difficult work for a statistician to combine all of these seemingly important variables and it would be more helpful to know which interactions should be investigated instead of needing to do it all by hand. Although we did not consider interactions in this case, they could be considered with a similar process of function selection or also with two-dimensional or splines.

For further research, investigation could be made into including more variables which exhibited an almost non-linear pattern into the model. Better methods for function selection could be developed so that the number of degrees of freedom for a natural spline could be tested in conjugation with the function selection or different functions of the same variable could be used in the best subset selection to know which truly is the best function to use for each variable. We could also investigate how well natural splines do in comparison to other

methods, such as wavelets, smoothers, or local regression. These methods all seem to have different advantages and it would be interesting to learn more about their similarities.