**Data Analysis and Optimization for Bike Sharing**

Danni Fu, Junhao Gao, Siying Gong, Ozair Meghani, Zhe Zhao

**Author Contributions:**

D.F., J.G., S.G., O.M., and Z.Z. - Experimental Design/Project Outline;

J.G. -Data Cleaning;

J.G., O.M., and Z.Z. Modeling and Analysis;

O.M., and Z.Z. -Prediction;

D.F., J.G., S.G., O.M., and Z.Z. -Figures and Plots;

D.F., J.G., S.G., O.M., and Z.Z. -Writing

**Abstract:**

Solving multicollinearity during modeling is a significant problem. This report aims to address collinearity among variables while developing three most likely valid models from the 2011 bike sharing in Washington D.C dataset to predict bike rental counts on any given day of 2012. There are three different strategies involved in each of our three models: All subsets variable selection for the Ordinary Least Squares (OLS) based model, m8; penalized modeling through shrinkages on coefficient estimates, the Ridge regression model, ridge.mod; Penalized modeling through shrinkages and Variable selections, Lasso regression model, lasso.mod. To determine the best of these top three valid models for prediction, we tested how they each performed in predicting daily bike rental count on the 2012 dataset based on Mean Squared Error. We concluded that in terms of prediction specifically, the Lasso Regression model was the best, but the issue of bias raises concerns in terms of interpreting the coefficient estimation.

**Introduction:**

Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, bike rental and return has become automatic. It is an economic, convenient and environment-friendly way of public transportation and is gaining popularity in a number of U.S. cities, such as Detroit, New Haven, and New Orleans. These cities have either selected vendors or are planning to launch systems. More people on Bike Share means that more people are saving time, and realizing new opportunities for exercise and health. Currently there are approximately 28 million sharing bikes in the U.S on service and the number is growing. [[1]](#footnote-1)

Because of the significant role bike sharing data has, we are interested in predicting the number of bike rentals daily in Washington D.C in 2012, taking into account different effects such as temperature, wind-speed, humidity, etc. and developing models from the 2011 data. From the perspective of researchers, bike sharing systems can be more accurate in collecting data on travel time, which can be used for sensing the mobility in the city.

To fulfill our goal, we need to find the appropriate model utilizing the methods we learned in class such as the ordinary least squares method, generalized least squares method, Lasso and Ridge regression. Throughout the modeling process, we encountered multiple problems such as a missing value in variable, lags, and collinearity between numbers of independent variables. In this report, we specify each steps we took to overcome these problems and present our finalized, best model in order to have a valid prediction of 2012 daily bike rental counts based on 2011’s data.

**Background**

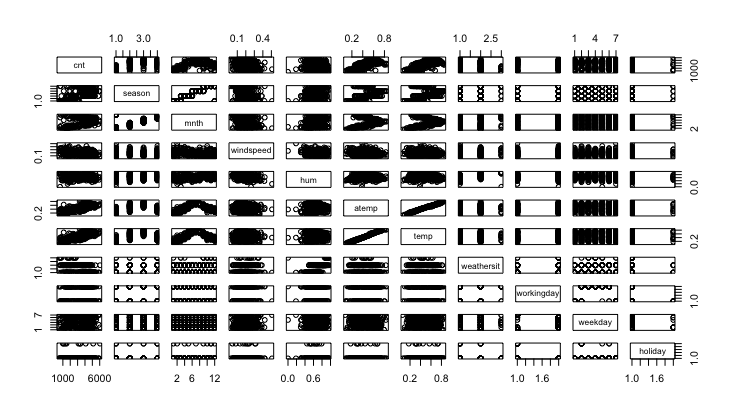
The prediction is that bike-sharing daily rental numbers are correlated to, or to some of, season, holiday, weekday, workingday, weather, temp, atemp (feels-like temperatures), humidity and windspeed. The original data is from Capital Bikeshare system, Washington D.C.,USA. The two-year historical log of year 2011 and 2012 is the core data. Data is collected on a daily basis and corresponding weather and seasonal information is added accordingly. To make sure the accuracy of the data, trips taken by staffs that service and inspect the system and taken from or to their warehouses’ “test” station has been removed. In addition, any trip that is under 60 minutes has been removed since potential users might start falsely and try to re-dock a bike to ensure its security. [[2]](#footnote-2)

We have three main problems for prediction. The first problem is picking up useful variables. With all 9 variables above in the model, the data collection would be costly and the model would be redundant. If we failed to do so, useful variable coefficients would be inflated and uncorrelated variables just fit the noise of the data. The second problem is that year 2012 is a leap year while the year 2011 is not. When predicting, it is important that the prediction data set and training dataset should have same subspaces or the prediction would make no sense. The third problem is detecting errors when developing models. We notice some unusual data in the dataset and we need to determine if they are errors that need to be deleted (and re-predicted) or outliers and leverages that need to be kept.

**Modeling and Analysis**

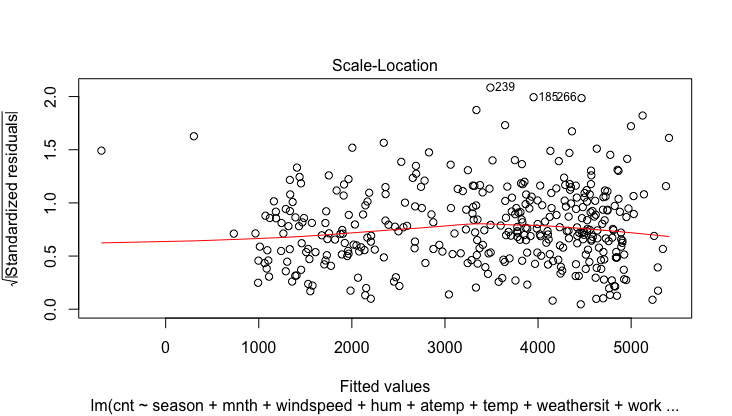
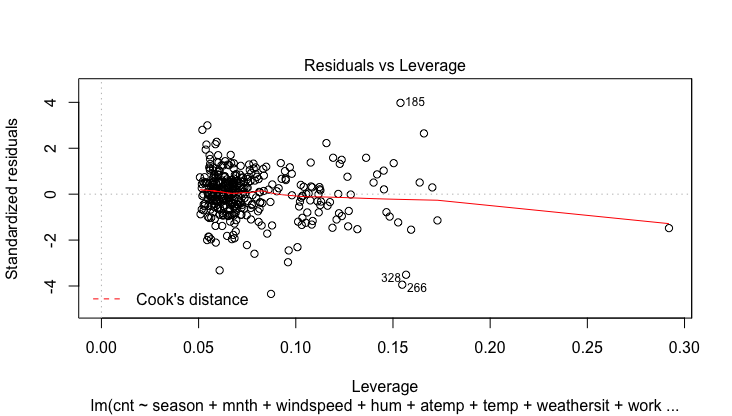
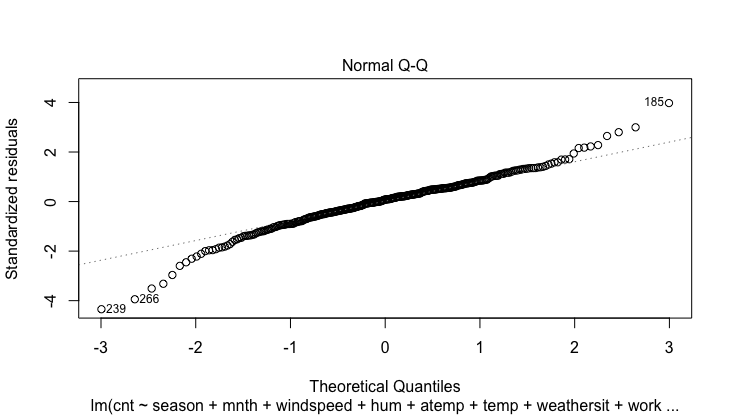
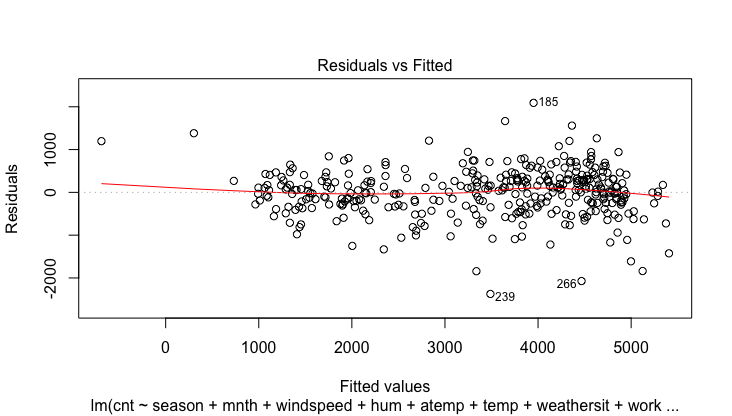
As we were modeling, we found that our main purpose was to eliminate multicollinearity. To achieve this goal, we built an Ordinary Least Squares model with variable selection, ridge regression (shrinkage) and Lasso regression (variable selection and shrinkage).

The dataset that we were provided with included the following variables that we will mention throughout the paper: date (dteday), season, year (yr), month (mnth), whether the day is a holiday or not (holiday), day of the week (weekday), workingday, the type of weather that day (weathersit), normalized temperature in celsius (temp), normalized feeling temperature in celsius (atemp), normalized humidity (hum), normalized windspeed (windspeed), count of casual users (casual), count of registered users (registered), and bike rental count (cnt). We began our analysis by first splitting up our data set into two separate sets based on the year 2011 being the training set for developing our model, and 2012 being the test set for prediction. Our first step with the training data was to take our non-continuous predictor variables (weathersit, weekday, holiday, workingday, season, and month) and convert them into factors so that they could be treated as qualitative variables in our model. Following this, we created a scatterplot matrix (Figure 1) with all of the variables in the data set including bike rental count in order to assess the correlation among variables. We found that there were a few groups of predictor variables that had multicollinearity. For example, temp and atemp had a strong linear relationship based on the plot, while month and temp appeared to have a non linear relationship. Additionally, there was non-constant variance between bike rental count and temp, atemp, and windspeed. Motivated by our matrix plots, we looked at a correlation matrix of the predictor variables in order to detect which variables may be experiencing multicollinearity. The correlation matrix shows us that variables month and season, and temp and atemp are highly correlated. This step provided us with a heads up for later on in the analysis- a better idea of what we might expect during the variable selection process.

As we fit the Ordinary Least Square model with all variables, we proceeded to check the validation of this model by graphing diagnostic plots. We observed a curved pattern in the residuals plots (figure 2), which indicates adding terms; curved pattern in standardized residuals vs. fitted values plot (figure 3) for the OLS model and non-normal distribution in the Normal Q\_Q plots (figure 4), indicating non-constant variance, pointing us in the direction of performing transformations on the variables or adding additional terms to the model. Upon looking at the Residuals vs. Leverage plot (figure 5) we noted 3 major bad leverages, 2 of which we attributed to national holidays (Independence Day and Thanksgiving), and one that we were not able to trace to any specific holiday or major event. Since there are no typos or any other kinds of errors in the data, we decided to keep them.

Figure

Performing the Box-Cox method next helped us determine which variables needed transformations and specifically what kind transformation needed to be done. From the Box-Cox results we determined that the following three variables needed a square root transformation: windspeed, temp, atemp; and hum needed an additional quadratic term. After applying these transformations, we performed several simple linear regressions with bike rental count vs each predictor variable and observed their diagnostic plots to confirm that the appropriate transformations were made. Upon doing so, we noticed an outlier in the bike rental count vs hum leverage plot and discovered that there was an error at data point 69 for hum as its value was listed as 0. Given the Earth’s climate and weather conditions, it is not physically possible for the humidity value to ever reach zero. Thus, we addressed this issue by local polynomial fitting with a

kernel weight and were able to predict an estimate for the humidity on that day based on the rest of the data.

Figure

Figure 4

Figure 5

Figure

After making the above changes and obtaining our most likely valid model, we began to address the issue of multicollinearity by doing variable selection. The method for variable selection that we used is Criterions-based comparing all potential subsets. The criterions of comparison are adjusted R-squared, AIC, AICc and BIC values. The reason we used all-possible-subsets is that it involves a broader search and compares more models than the stepwise method does. The Stepwise method, which drops the variable that has the highest (backward) or adds the lowest (forward) p-value in each step, often increases the coefficient estimates of selected variables. The p-value does not have a proper meaning during the stepwise process. We first came up with every possible subset of the variables at each size for example, all possible subsets with one predictor variable, all possible subsets with 2 predictor variables, and so on and so forth. We chose the best model from each size of variables based on adjusted R-squared, and were then left with 9 models to compare. The two models, m8 (with 8 variables, lowest AIC, AICc and BIC) and m9 (with 9 variables and highest adjusted R-square) were our top two models and upon comparing them we found m8 to be more valid(constant variance) based on residual plots (residual plots: figure 6, Standardized residuals plot: figure 7, Normal QQ plot: figure 8 and leverage plots: figure 9 ) and with every variable being significant at less than 0.001 level (table 1 of summary of m8).

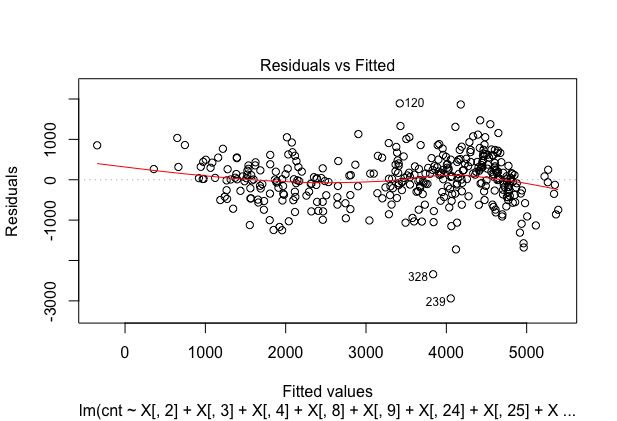


Figure 7

Figure 6

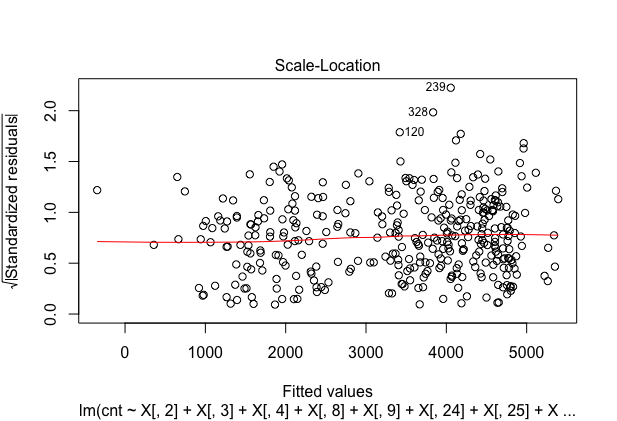
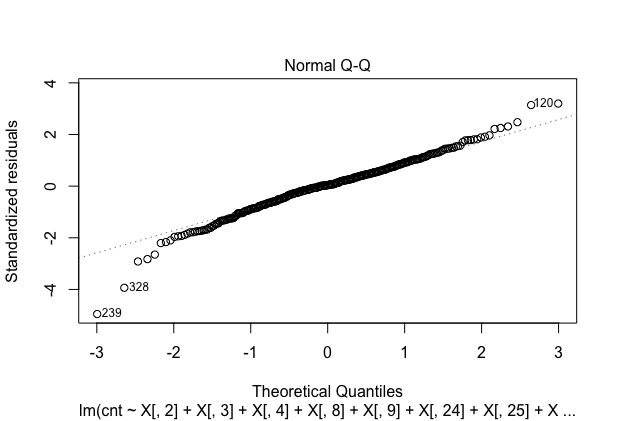
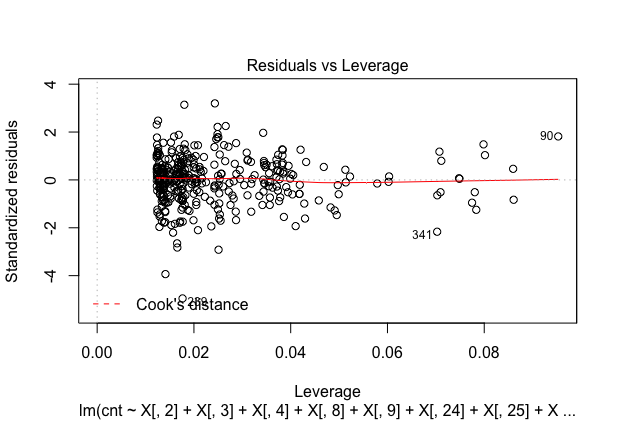
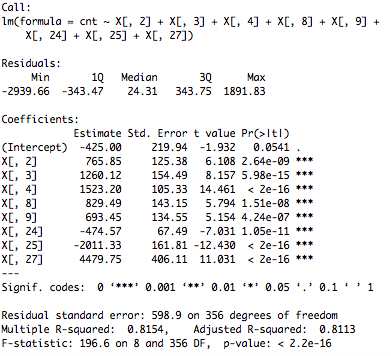


Figure 8

Figure 9

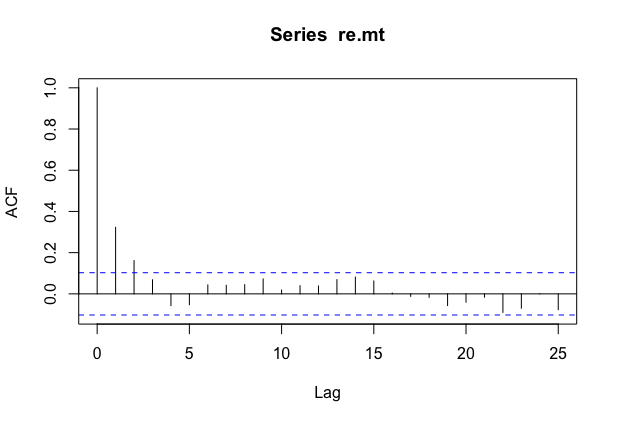
Table 1



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X[, 2] | X[, 3] | X[, 4] | X[, 8] | X[, 9] | X[, 24] | X[, 25] | X[, 27] |
| season2 | season3 | season4 | mnth5 | mnth6 | weather2 | weather3 | sqrt(temp) |

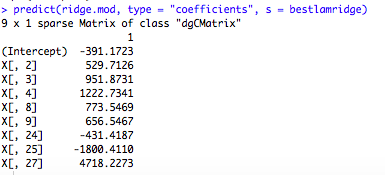
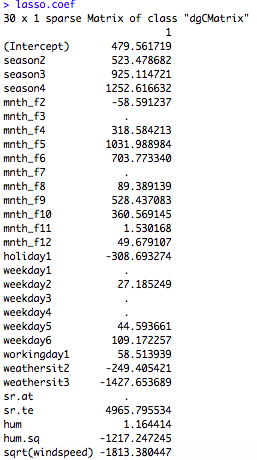
However, when we looked at the diagnostic plots of the model obtained from variable selection, we saw a little curved pattern in the standardized residuals plot. We suspected that this was because of the correlation among residuals. We used the Auto-Correlation Function to detect autocorrelation sequence. The result indicated significant lags between standardized error[i] and standardized error[i + lag] (in figure 10). We solved this problem by fitting into an AR(1) model since it was the only technique that we learned to solve the Auto-Correlation problem The problem was: even though significant lags seemed to be eliminated, a few of the lags seemed to be amplified. For example, lag at standardized error[i + 25] in AR(1) model was larger than the Ordinary Least square model, m8. Furthermore, the diagnostic plots of the LS model transformed from AR(1) model showed worse results. In the end, we considered linear model, m8, to be a better model. In addition, we checked the correlation table again to determine if the variable selection was good enough for eliminating multicollinearity. We suspected one of season 2, season 3, mnth 5 or square root of temp would need to be further adjusted.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | season2 | season3 | season4 | Mnth5 | mnth6 | weather2 | weather3 | sqrt(temp) |
| Inflation factor | 3.0152 | 4.643 | 2.0813 | 1.620 | 1.39 | 1.0396 | 1.0498 | 3.4597 |

The purpose of Ridge was to eliminate multicollinearity furthermore because we found relatively high colinearity in the correlation table of the model matrix of m8. Before we fitted into Ridge and Lasso regression, we checked the inflation factor of the OLS model obtained from Variable selection to assume which variable would be most likely to be eliminated in Ridge and Lasso regression. Our speculation based on the inflation factor (table 2) was that season 3 and the square root of temp would be most likely to shrink. The result of ridge regression (table 4) in model ridge.mod showed shrinkage for season 3’s coefficient estimate, from 1260.1225 to 951.8731. Others had more or less shrinkage except for the square root of temp. The variable’s coefficient estimate increased, from 4479.7546 to 4718.2273. We plotted the Beta Coefficient against lambda in figure 13 to see why there is more inflation with the best lambda chosen. However, the plot we obtained told us that all variables have more or less shrinkage at the best lambda chosen. We could not get information on why ridge inflates the variable. We speculated that maybe the shrinkage of other variables increases the weight of sqrt(temp) in the model.

Since Lasso regression will help us both on the shrinkage of coefficients estimates and the variable selection, we fitted all variables from the 2011 dataset into the Lasso regression and compare with the OLS model we obtained from all-possible subsets selection (lasso output in table 3). Lasso regression chose more variable than the all-possible subsets selection did; however, mnth 11, hum that Lasso regression chose were with very little coefficients (1.53 and 1.16). For the shrinkage part, except season 2 and season 3, Lasso regression model had less shrinkages than the ridge regression model does. However, one thing worth noticing was that the coefficient estimate of square root of temp inflated even more than ridge regression model did. When we plotted the Lasso Regression model with coefficients against log(lambda) in figure (14), we found that one variable, after another variable came in, soon fell to zero as lambda increased. We suspected that variable is sqrt(temp), and the other variable that caused sqrt(temp) to fall is sqrt(atemp) because of more than 90% of collinearity between them. It suggested that before sqrt(atemp) adding to the model, LASSO regression did not shrink much for sqrt(temp). Before the best lambda level, because the lasso added potentially correlated variables with sqrt(temp) would have no effect on the path sqrt(temp) (for example, season 3), we saw clearly that sqrt(temp) never experienced any shrinkage.

Figure 10

In the end, we obtained three models that best fulfill our two purposes. They are the Ordinary Least Square Model m8, the Ridge regression model.ridge.mod, and the Lasso regression model, lasso.mod.

Table

Table

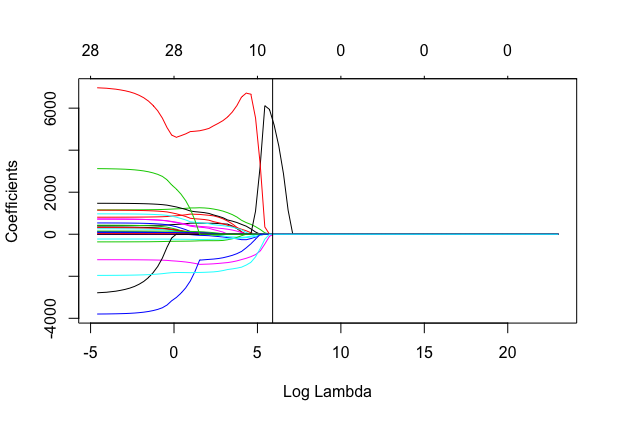
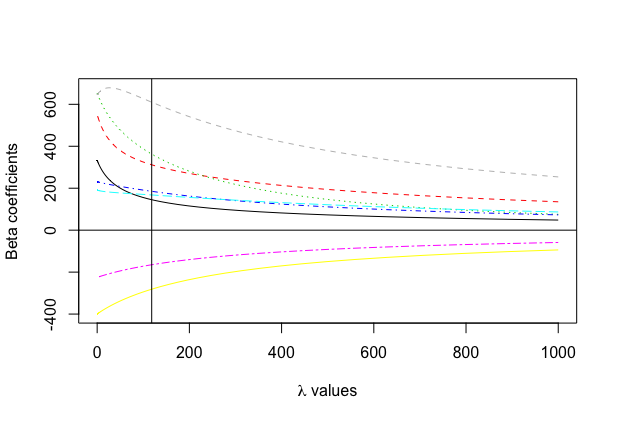


Figure 12

Figure 11

**Prediction:**

One problem was that, since we developed the model in the training set, which had 365 instants in year 2011, when we predicted rental count in 2012, which has one more day of February 29th, we needed to adjust the two datasets so that their rows would corresponded. Prediction would make no sense if their subspaces are not the same. It also does not make sense to predict rental count on Feburary 29th of 2012 due to the issue of interpolation. We then delete the row 60, which is February 29th, and proceed to predict by comparing Mean Squared Error.

In predicting the bike rental count, our variable of interest, we applied our three best models mentioned above to the 2012 testing data set. We are looking for the model that minimizes the mean squared error (MSE) of the predicted vs. actual bike rental count value. The three MSE values we obtained are as follows: the m8 MSE = 6,542,634, the Ridge regression model MSE = 5,564,730, and the Lasso regression model MSE = 5,130,210. Based on these results, it is clear that the Lasso model was the best model for predicting bike rental count on any given day of the year.

However, we need to be careful when using the lasso regression model to predict. Firstly, coefficient estimates are biased because of its bias-variance trade-off nature. The coefficient estimates in lasso may not make any sense, but with biased coefficient estimates, it could get a lower MSE. Secondly, the lasso regression model, which provides more efficient way for variable selection, doesn’t check all available subsets of dataset. Even though it has a better prediction result than the OLS model, m8, does, we are not certain that the lasso regression model is the most-likely true model that could eliminate most of multicollinearity.

**Discussion:**

In this report, we discussed our strategies used in order to develop our three most valid models based on Bike Share data from Washington D.C. in 2011. Upon obtaining these three models we wanted to test how they would perform in predicting bike rental count, our variable of interest, on any given day in the year 2012.

For prediction, our three models suggest significance for variables’ coefficients estimates: season spring (positive), season summer (positive), season fall (postive), month May (positive), month June (positive), weather three (negative), weather two (negative) and temperature (positive). Knowing what factors’ levels have significant influence on the bike rental counts (weather, season, months), could allow the company Capital Bike Share to make decisions on fluctuating the price on certain days or during certain months in order to maximize their profits, as well as to promote increased use of their bikes. Temperature is also an important predictor because it has a significantly positive correlation with the rental count. Investors could consider cities’ climates when entering other markets or expanding their business.

Problems remain in each model. The OLS model, which owns non-biased coefficient estimates and has been chosen from all-possible-subsets, produces the largest Mean Squared Error (MSE) when predicting in the 2012 dataset; the ridge regression model inflates one coefficient estimate that we expected to be shrinked; lasso regression model, which has the least MSE for prediction, chooses more variables than the OLS model does and owns biased coefficient estimates, for which we aren’t able to provide useful information when discussing predictors’ influence on the variable of interests.

To improve our prediction, we would consider the variables registered and casual to see if the variables in our models have a different influence on them, and we could provide more information to companies that are interested in loyal customers. Our best prediction model, lasso.mod, couldn’t provide useful information by its coefficients estimates. We suspect that there are missing variables that we failed to consider. In the future, we would focus on and improve our OLS model, m8, by including other missing predictors.

In the future, we could improve our modeling by firstly, searching through all subsets because R only let us reach subsets of size 9, while theoretically there were many more until 29. Secondly, even though the Lasso regression model finished variable selection effectively, we suspect that it’s not enough and more variables could be dropped. We are curious to see if we could fit our OLS model, m8 into lasso and let the Lasso regression model further improve our variable selection. Ultimately, we feel that our Lasso regression model could serve as a valuable tool in predicting the daily bike rental counts for Bike Share in Washington D.C.

**Bibliography**

Ink, Social. "Bike Share in the US: 2010-2016." National Association of City Transportation Officials. Accessed December 11, 2017. <https://nacto.org/bike-share-statistics-2016/>.

International, Inc. Motivate. "System Data." Capital Bikeshare. Accessed December 09, 2017. https://www.capitalbikeshare.com/system-data.

**Appendix:**

R code:

Matrix Plots (figure 1):

pairs(~cnt+season+mnth\_f+windspeed+hum+atemp+temp+weathersit+workingday+weekday+holiday)

Diagnostic Plots of OLS model with all variables (figure 2 ~ 5):

m.ols.mnf <- lm(cnt~season+mnth\_f+windspeed+hum+atemp+temp+weathersit+workingday+weekday+holiday)

plot(m.ols.mnf)

All subsets Variable Selection:

m.final <- lm(cnt~season+mnth\_f+holiday+weekday+workingday+weathersit+sr.at + sr.te + hum + hum.sq + sqrt(windspeed))

library(leaps)

library(car)

X<-model.matrix(m.final)

b<-regsubsets(as.matrix(X),cnt,data=data2011)

rs1 <- summary(b)

par(mfrow=c(1,1))

m1<-lm(cnt~X[,26])

m2<-lm(cnt~X[,26]+X[,25])

m3<- lm(cnt~X[,4]+X[,25]+X[,27])

m4<- lm(cnt~X[,4]+X[,24]+X[,25]+X[,27])

m5<- lm(cnt~X[,4]+X[,8]+X[,24]+X[,25]+X[,27])

m6<- lm(cnt~X[,2]+X[,3]+X[,4]+X[,24]+X[,25]+X[,27])

m7<- lm(cnt~X[,2]+X[,3]+X[,4]+X[,24]+X[,25]+X[,27]+X[,29])

m8<- lm(cnt~X[,2]+X[,3]+X[,4]+X[,8]+X[,9]+X[,24]+X[,25]+X[,27])

m9<- lm(cnt~X[,2]+X[,3]+X[,4]+X[,8]+X[,9]+X[,25]+X[,27]+X[,28]+X[,29])

n1 <- nrow(data2011)

lm.n <- function(l) {

npar = length(coef(l)) + 1

AIC = extractAIC(l, k = 2)[2]

BIC = extractAIC(l, k = log(n1))[2]

return( c(rs1$adjr2[npar - 2], #adjr2

AIC,

AIC + 2 \* npar \* (npar + 1) / (n1 - npar + 1), # AICc

BIC))

}

matrix(unlist(lapply(list(m1, m2, m3, m4, m5, m6, m7,m8,m9),lm.n)), byrow = TRUE, ncol = 4, dimnames = list(1:9, c("adjr2", "AIC", "AICc", "BIC")))

Summary table of m8 (table 2):

summary(m8)

Diagnostic Plots of m8 (figure 6 ~ 9):

plot(m8)

AR(1) model and ACF plot (figure 10):

re.mt <- rstandard(m8)

acf(re.mt)

library(nlme)

m.gls.f.ad <- gls(cnt~X[,2]+X[,3]+X[,4]+X[,8]+X[,9]+X[,24]+X[,25]+X[,27],

data = data2011, correlation = corAR1(form =~instant), method = "ML")

summary(m.gls.f.ad)

rho.ar <- 0.3338995

X.ar <- model.matrix(m.ols.mnf)

Sigma.ar <- diag(length(instant))

Sigma.ar <- rho.ar^abs(row(Sigma.ar)-col(Sigma.ar))

sm.ar <- chol(Sigma.ar)

smi.ar <- solve(t(sm.ar))

xsta.ar <- smi.ar %\*% X.ar

ystar.ar <- smi.ar %\*% cnt

mltls.ar <- lm(ystar.ar~xsta.ar-1)

red.st.ar <- rstandard(mltls.ar)

Inflation Factors Table (table 1):

vif(m8)

Ridge Regression Table (table 2):

lambda <- 10^seq(10, -2, length = 100)

X.r <- X.ols[,-1]

library(glmnet)

train = data2011

test = data2012

ridge.mod <- glmnet(X.r, train$cnt, alpha = 0, lambda = lambda)

cv.out <- cv.glmnet(X.r, train$cnt, alpha = 0)

bestlamridge <- cv.out$lambda.min

predict(ridge.mod, type = "coefficients", s = bestlamridge)

Lasso Regression Table (table 3):

X.l <- X[,-1]

lasso.mod <- glmnet(X.l, train$cnt, alpha = 1, lambda = lambda)

cv.out.L <- cv.glmnet(X.l, train$cnt, alpha = 1)

bestlamlasso <- cv.out.L$lambda.min

lasso.coef <- predict(lasso.mod, type = 'coefficients', s = bestlamlasso)

Ridge Plot of Coefficients against Lambda (figure 11):

x <- data.frame(X[,2], X[,3], X[,4], X[,8], X[,9], X[,24], X[,25], X[,27])

x <- as.matrix(x)

y <- as.vector(cnt)

ridge.plot(y, x, lambda = 10^seq(3, -2, length = 100))

Lasso Plot of Coefficients against Lambda (figure 12):

plot(lasso.mod, xvar="lambda", label = T)

1. Ink, Social. "Bike Share in the US: 2010-2016." National Association of City Transportation Officials. Accessed December 11, 2017. https://nacto.org/bike-share-statistics-2016/. [↑](#footnote-ref-1)
2. International, Inc. Motivate. "System Data." Capital Bikeshare. Accessed December 09, 2017. https://www.capitalbikeshare.com/system-data. [↑](#footnote-ref-2)