Bike Sharing Data Analysis

Khushi Khatri

4/11/2020

1. Introduction

In the last decade, bike-sharing systems have become increasingly prevalent in urban cities. These programs automate the process of short-term bike rentals, thereby providing greater mobility and affordability while promoting sustainable transport. The success of bike-sharing systems can be attributed to modern advances in information technology as well as to data-driven decision making. In this report, we will analyze bike rental data provided by Capital Bikshare in Washington D.C., and utilize statistical procedures such as linear modeling to find the best regression model for predicting the total number of bike rentals in a particular hour from environmental conditions and other predictors. We will also interpret this final model and determine the validity of the inference drawn from it.

2. Data Description

The dataset of interest has been collected by Capital Bikshare System, Washington D.C. with the intent of predicting the number of bike rentals in a given hour using the seasonal and environmental conditions for that hour. This dataset contains 17379 observations collected over two years (2011 and 2012) with each observation representing one particular hour. For the purpose of validation of the final model in this analysis, we reserve a random sample of 1000 observations for the test set, leaving 16379 observations to train the competing models on.

3. Analysis

i) Exploratory Data Analysis

We begin our analysis by exploring the contents of the dataset. There are 17 columns, and the response variable (Total Count of Bike Rentals in an Hour), is the sum of two other variables, which represent the number of bike rentals by unregistered (casual) users and registered users for that hour respectively. We remove these latter two columns from our analysis as they are perfectly collinear with the response and will hinder the existence of unique Least Squares solutions. Similarly, we remove the factor that indicates a working day as it has a linear dependency with and can be discerned from the factors that indicate holidays and weekdays. We also remove the column that only gives the unique observation number, a column representing the date, which is in a format (yyyy-mm-dd) that is not compatible with the regression methods used in this report, and a factor indicating the year, which takes values 0 and 1 for 2011 and 2012 respectively. The latter two variables won't be beneficial for future prediction purposes as they will not be repeated (the same date will not occur again in the future).

We move on to examining the distribution of the response variable, which is heavily positively skewed as can be seen in Figure 1. Some hours have a significantly higher count of bike rentals than others. These are unconditional outliers that could influence regression results, but we will test measures of leverage later to determine if this is true. Figure 1 also shows the result of using a power transformation of 1/3 (cube root)

to make the distribution of the response more symmetric and to pull the right tail in. This transformation makes the model less interpretable but since prediction is our main goal we apply it anyway.

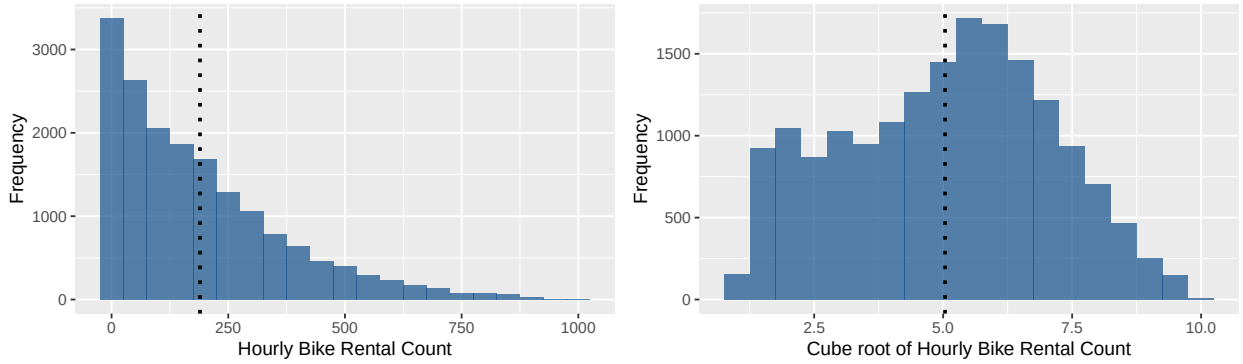

Figure 1: We plot the distribution of the response variable, Total Bike Rentals in an Hour, before (left) and after (right) the cube root transformation and notice how the distribution becomes significantly more symmetric about the mean, which is indicated by the dotted line in each graph.

Upon examining the marginal distributions of the four quantitative predictors, we notice that the distributions of Normalized Temperature and Normalized Feeling Temperature are fairly symmetric. The distributions of Normalized Humidity and Normalized Wind speed are left- and right-skewed respectively. We apply a power transformation to each of these variables, taking the square of Normalized Humidity and the square root of Normalized Wind speed, which is indicated in Figure 2. Not only do these transformations help towards reducing skewness, but they also help make the relationship between these predictors and the response more linear. Furthermore, these specific transformations were applied in favor of others because the final model is to be used for prediction: the chosen transformations resulted in the lowest Cross-Validation Mean Squared Error (discussed later) upon running the analysis with a range of different transformations.

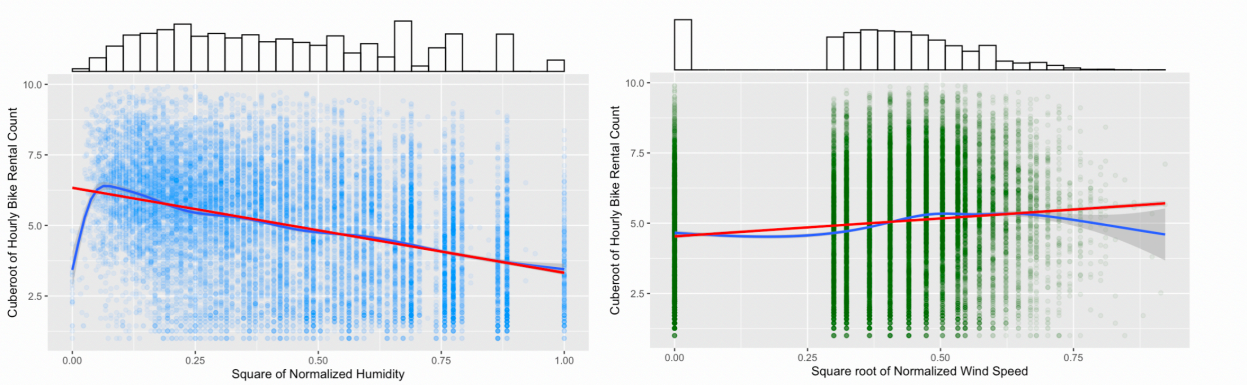

Figure 2: We provide scatterplots to visualize the post-transformation relationships between the response and each of two predictors (Normalized Humidity on the left and Normalized Wind Speed on the right) as well as the predictors' marginal distributions, which are pictured in the histograms above. The red line in the graphs is fit by least squares regression while the blue line is a smooth line fit by generalized additive modeling for comparison purposes.

In Figure 3, we examine bivariate relationships between the transformed response and the quantitative explanatory variables (incorporating transformations for Normalized Humidity and Normalized Wind Speed) through pairwise correlations as opposed to scatterplots because the very large number of observations makes it difficult to interpret relationships using the latter. The heatmap of correlations reveals a nearly perfect positive correlation between Normalized Temperature and Normalized Feeling Temperature, which is expected. This creates concerns about collinearity, which we will keep in mind while performing model selection. We notice that the Pearson Correlation Coefficient between the response and the Square root

of Normalized Wind Speed has a smaller magnitude relative to that of the response with the other three quantitative explanatory variables, which are relatively similar (~0.4) in magnitude.

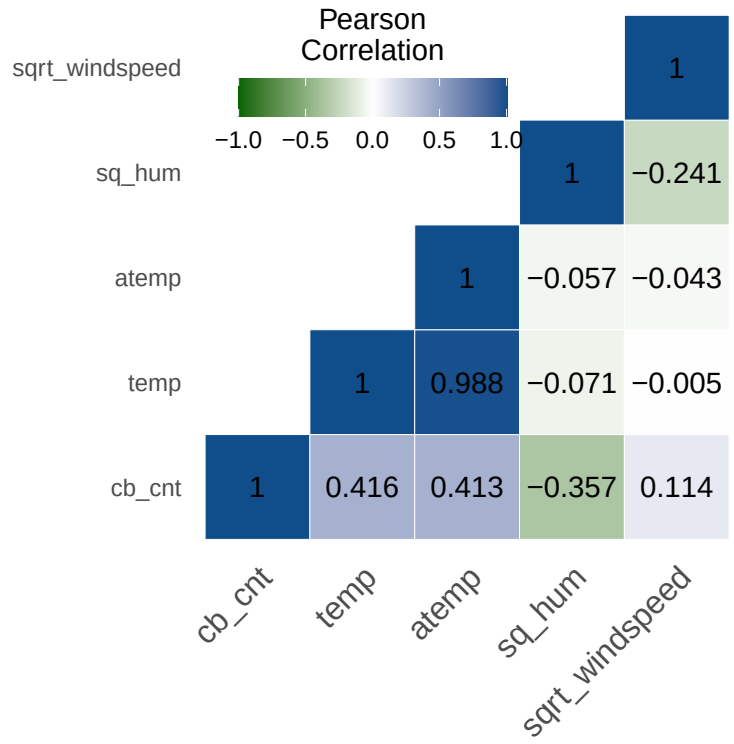

Figure 3: We create a heatmap of the pairwise correlations between the Cube Root of Hourly Bike Rental Counts and the four quantitative explanatory variables.

Upon analyzing the distributions of the six categorical explanatory variables, we notice that while most factors such as those indicating season, weekday, month and hour have a mostly uniform distribution of recorded observations across their categories, the variable indicating holiday that is coded 1 if the observation is recorded on a holiday and 0 otherwise, has a very unequal distribution between its two categories, as can be seen in Figure 4. Another highly unequal distribution of observations across categories is present in the factor that differentiates between 4 categories of different weather types. As is indicated in Figure 4, only 3 of the 17379 observations have a value of 4 for this factor, which is reasonable as this 4th category represents extreme weather conditions such as heavy rain, thunderstorm, etc.

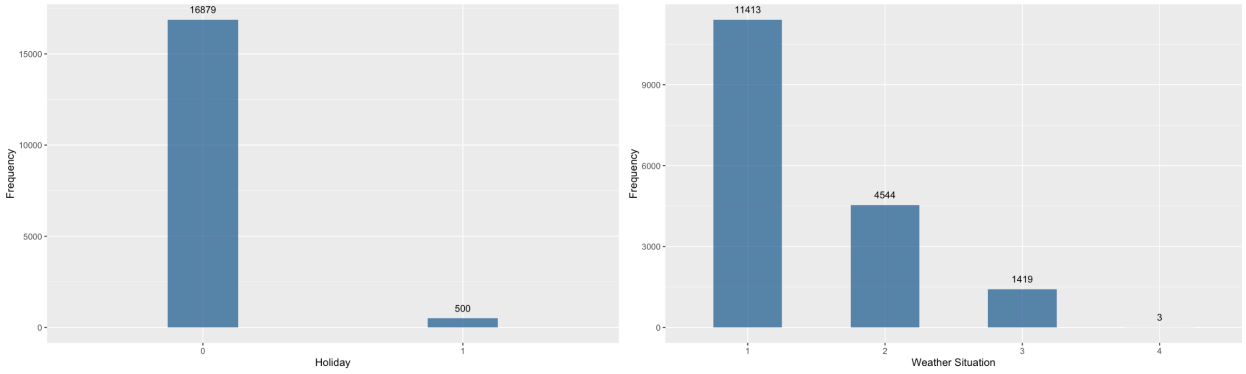

Figure 4: We create bar charts of the counts of recorded observations across the categories of the factors that indicate Holidays (left) and Weather Situation (right).

An analysis of side-by-side box plots revealed no notable interactions between pairs of factors. We also

examined co-plots for the regression of the Cube root of Hourly Bike Rental Count on the 4 quantitative explanatory variables across the categories of various factors to check for interactions. The most notable observation was the slightly different slope of the fitted line for the regression of the response on the quantitative explanatory variables for the 4th category of the factor Weather Situation. However, we will refrain from treating this as evidence of an interaction between the factor Weather Situation and the quantitative explanatory variables for two reasons. The first reason is that the slopes of the fitted lines for the other 3 categories of this factor look almost identical when the response is regressed on any of the 4 quantitative predictors, and the second being that it is highly likely that this different slope is a consequence of the fact that the number of recorded observations for which this factor has a value of 4 is extremely small (as discussed earlier), which is evident in Figure 4.

Lastly, the (pre-transformation) distribution of the response is highly right-skewed, so we wanted to ensure that the results of our analysis aren't too dependent on a few extreme observations. We looked for regression outliers (measured by studentized residuals) as well as high-leverage observations (measured by hat values) and influential observations (measured by Cook's distance and COVRATIO). Upon running the analysis with and without the five observations with the highest Cook's distance and COVRATIO values, we noticed that the results did not change much, so we refrain from deleting these observations from the dataset.

ii) Model Selection

To find the best regression model for predicting the total number of bike rentals in a particular hour using our explanatory variables, we will attempt two methods of regularization: (1) Model Shrinkage and (2) Variable Selection for Models (fit by the Least Squares approach).

Since our main goal is prediction, we will focus on finding the model that has the smallest Mean Squared Error of Prediction. For each of the following methods, we will estimate predictive accuracy (or out of sample fit) by using a measure of in-sample fit.

(1) Model Shrinkage
The following methods find models that fit the data well (making the Residual Sum of Squares small) while constraining the size of the coefficient estimates by utilizing a penalty term that shrinks these estimates towards zero.

Ridge Regression: The vector of ridge regression coefficients β is that which minimizes the following equation:

$$||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Ridge Regression utilizes an l_2 norm, which only shrinks the coefficients to exactly zero when the tuning parameter $\lambda = \infty$. Therefore, unless $\lambda = \infty$, Ridge Regression always returns a final model with all p explanatory variables.

LASSO: The vector of LASSO coefficients β is that which minimizes the following equation:

$$||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

LASSO, unlike Ridge Regression, utilizes an l_1 norm, which shrinks certain coefficients to exactly zero when the tuning parameter λ is large enough (not necessarily infinity).

The value of the tuning parameter λ was selected to optimize predictive accuracy, which, in this case, is evaluated by minimizing the cross-validation error using 10 fold cross-validation. Figure 5 displays the error paths (Cross-Validation Mean Squared Error) as a function of the tuning parameter λ (on a log scale), which varies across a range of values for both Ridge Regression (featured on left) and LASSO (featured on right).

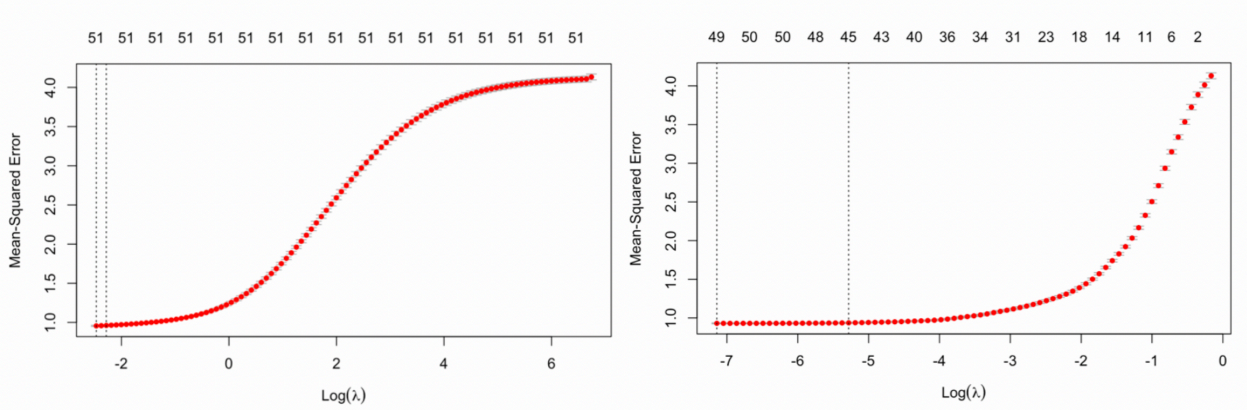

Figure 5: We plot Mean Squared Cross-Validation error paths as a function of the tuning parameter λ (on a log scale) for Ridge Regression (left) and LASSO (right). The first vertical dotted line indicates the λ value that minimizes this error while the second vertical dotted line indicates the largest λ value that lies within one standard error of the λ that minimizes the error.

(2) Variable Selection Models fit by the Least Squares Approach

The method of least squares chooses a coefficient vector β that minimizes the following measure:

$$||\mathbf{y} - \mathbf{X}\beta||^2$$

While performing variable selection, we used the "regsubsets" function from the "leaps" package to find the models of each size that minimize the Residual Sum of Squares using forward and backward selection with the 10 explanatory variables. Ideally, we would use all subsets selection but it proved to be problematic with categorical variables and too computationally expensive for a dataset of this size. We used three criteria: BIC, Mallows C_p and Adjusted R^2 to pick between models of different sizes that were chosen by forward and backward selection. We did this to have a range of models to choose from, but the best model resulting from each of these step-wise methods and criteria included all four quantitative explanatory variables and either most or all of the categories of each factor.

The omission of only a few categories and not others indicated that each factor contributed to explaining the variation in the response. For brevity, we show the comparison of the minimum Mean Squared Error from 10-fold cross-validation of the full Least Squares model (with the transformations that were applied in the Exploratory Data Analysis section of this report) with that of Lasso and Ridge Regression to tangibly compare their predictive accuracy. These results are summarized in Table 1 below.

Method	Optimal λ value	Minimum CV MSE Error
Ridge Regression	0.0848	0.9559
LASSO	0.000791	0.9325
Least Squares Regression	NA	0.9299

Table 1: Comparison of Minimum Mean Squared Error from 10-fold cross-validation to evaluate predictive performance across Ridge Regression, LASSO, and Least Squares Regression.

It can be seen that the full model fit by Least Squares to the training data has the smallest Cross-Validation Mean Squared Error, and hence is best for prediction purposes. It is important to note that the optimal λ value for both LASSO and Ridge Regression is very small. This indicates that according to both methods, a very small amount of shrinkage is needed from the Least Squares coefficients to get the optimal fit. The optimal LASSO and Ridge Regression coefficients hence only have a slight bias and a slightly smaller variance as compared to Least Squares coefficients. This also explains why the Least Squares fit does only marginally better than the Model shrinkage methods. The optimal λ for LASSO is much smaller than that of Ridge Regression, and while there is a reasonably small difference of Mean Squared Error between the latter and the Least Squares method, the difference between LASSO and Least Squares is almost negligible

as is evident from Table 1. This slightly better predictive performance of Least Squares could be attributed to the fact that when λ is set to its optimal value for LASSO, this method shrinks the coefficients of only 1 category in each of 2 factors (the 4th category of month and the 2nd category of weekday), as opposed to all the categories of a factor, to zero. LASSO, in using the l_1 norm, implicitly assumes that a certain number of population coefficients have a true value of zero, so it understandable that it does not perform as well in this scenario if none of the true coefficients are equal to zero.

iii) Model Diagnostics and Interpretation

We have now determined that the best regression model for predicting total number of bikes rented in a particular hour is fit using the Least Squares approach and has the following equation:

$$(\text{Hourly Bike Rental Count})^{1/3} \sim \text{Normalized Temperature} + \text{Normalized Feeling Temperature} + (\text{Normalized Humidity})^2 + (\text{Normalized Wind Speed})^{1/2} + \text{Season} + \text{Month} + \text{Hour} + \text{Weekday} + \text{Holiday} + \text{Weather Situation}$$

Although the predictive accuracy of this model is more desirable than that fit by model shrinkage methods, it is imperative to analyze model diagnostics to determine the validity of inference drawn from it. We begin by examining diagnostic plots for the model fit by least squares, which is shown in Figure 6 below. The red line in the Residuals against Fitted Values plot (top left) of Figure 6 reveals that the residuals are mostly centered on a mostly horizontal line at zero for all fitted values. The assumption of linearity is therefore met. However, this plot also reveals that the residuals display a "floor" effect wherein they appear to be centered around a mostly straight line but the non-constant residual variance becomes apparent in a somewhat cone shape. This indicates that we must not interpret the standard errors rigorously. The Normal Q-Q plot (top right) of Figure 6 reveals that the assumption of normality is somewhat violated as the residuals have a slight left skew: there is a greater proportion of large negative residuals than is expected from a normal distribution. The Leverage plot (bottom right) of Figure 6 reveals that there are two very high leverage points but their studentized residuals are not too large, because of which Cook's Distance values for all observations are reasonable, as was discussed earlier in the Exploratory Data Analysis section of this report. This highlights the ability of data samples as large as this one to absorb unusual data while ensuring that it does not influence the final results, except in extreme cases.

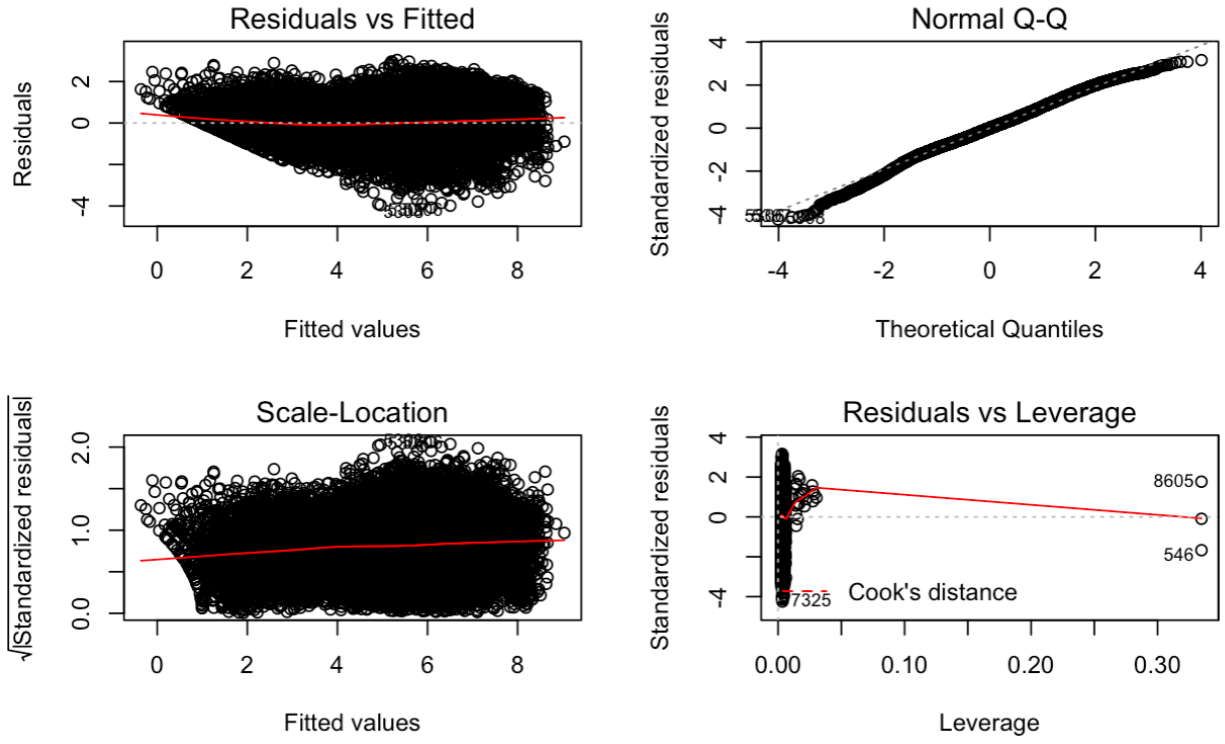

Figure 6: We plot the Residuals against Fitted Values (top left), Standardized Residuals against the Theoretical Quantiles of the Normal Distribution (top right), Square root of Standardized Residuals against Fitted Values (bottom left) and Standardized Residuals against Leverage (or hat) values (bottom right) to test the assumptions of linear regression.

When the final model is fit to training data, we see that the R^2 value is 0.776, indicating that this model explains about 77% of the total variation of the response variable. We also notice that the estimated coefficients for all categories of three factors- indicating season, hour and holiday - are statistically significant. The estimated coefficients for the quantitative explanatory variables- Normalized Temperature (1.69), Normalized Feeling Temperature (1.39), the Square of Normalized Humidity (-0.81) and the Square root of Normalized Wind Speed (-0.18)- are highly statistically significant (according to the t-test with 16327 degrees of freedom and $\alpha = 0.01$). In the Exploratory Data Analysis section, we noted that the Square root of Normalized Wind Speed had a smaller correlation relative to the other three quantitative explanatory variables with the response. This fact manifests in this predictor having a significantly smaller estimated coefficient than the other three quantitative predictors.

For this ordinary least squares regression, the estimated coefficient of 1.69 for a quantitative variable such as Normalized Temperature can be interpreted as the average change in the cube root of the total number of bike rentals in an hour associated with a one-unit increase in Normalized Temperature, holding other explanatory variables constant. On the other hand, the estimated coefficient of 0.45 for a dummy regressor such as season2 can be interpreted as the average difference in the cube root of the total number of bike rentals in an hour between the category represented by season2 (Summer) and the baseline category (Spring), holding other explanatory variables constant.

Not all the categories of Weather Situation, Weekday and Month have statistically significant estimated coefficients. Particularly months 4, 9 and 10 and weekdays 1, 2 and 3 have standard errors almost as large as their estimated coefficients. This could hint at the possibility that there is no difference between these categories and the baseline category of their respective factors. The 4th category of Weather Situation has an estimated coefficient with a very high standard error that is several times the coefficient itself. This once again could be attributed to the fact that only 3 out of the 17379 recorded observations have this category of weather.

Note: It is crucial to keep in mind that this inference should not be treated too rigorously since we have seen that after model selection not all the assumptions of linear models, that these interpretations depend on, were satisfied. There is uncertainty about the validity of these theoretical inferences.

4. Discussion

The analysis above reveals that the full model fit by the Least Squares approach is the best regression model for predicting total hourly bike rental count from the given explanatory variables. Although we were concerned about the high correlation between Normalized Temperature and Normalized Feeling Temperature, both variables were present in all the best models chosen by the various variable selection methods and criteria used. Moreover, a vast majority of the estimated coefficients in the full least squares model are statistically significant despite the collinearity between these two variables.

It is imperative to acknowledge the role of the Bias-Variance tradeoff in determining the best model for prediction. The models fit by Ridge Regression and LASSO intentionally trade off an increase in bias in an attempt to reduce variance and make the Mean Squared Error small. To investigate the better performance of Least Squares over LASSO further, I performed a Group LASSO using functions from the "gglasso" package. I grouped together the categories of each factor, and the Group LASSO did not shrink the coefficients of any category in any factor to exactly zero. This result is consistent with the fact that the full model fit by Least Squares has a lower Minimum Cross-Validation Mean Squared Error than that fit by LASSO.

Although the estimated coefficients in this chosen Least Squares model may be used to approximate partial effects for summarization purposes, it is not reasonable to view any of the fitted parameters as causal effects. The data that we fit this model to is observational data, not experimental data. There is no explicit guarantee

of randomness or control, which is why we are skeptical about our results to begin with. Moreover, the final model chosen was the one that optimized predictive accuracy as opposed to interpretability. Treating a parameter as a causal effect involves making a bold statement about the impact of an intervention, and is appropriate in scenarios where we define a treatment and an effect while controlling for confounding factors and other explanatory variables. Since none of these measures are taken in the procedure followed to obtain the data or the final model, no fitted parameter must be assumed to be a causal effect.

There are other weaknesses to the approaches used in this report that must be noted. The step-wise selection methods we used while doing variable selection are sub-optimal and do not consider all possible models with all subsets of regressors. It was too computationally inefficient to use best subset selection and cross-validation to choose among competing models and so those methods were not explored further even though they could provide further insight. The variable selection methods were also problematic with categorical variables. Another, although imperfect, approach would be to do variable selection with just the quantitative variables and to manually determine what categorical variables to include one at a time based on their explanatory capacities or their contribution to the Adjusted R^2. Other weaknesses include the removal of the variable representing year from the analysis, which was excluded on the basis that it would not be an appropriate regressor for future prediction purposes as the same years (2011 and 2012) will not occur again. Investigating the slightly higher mean and the significantly higher variance of hourly bike rental counts in 2012 as compared to 2011, and the potential reasons behind it, could help explain the variation in the response further and provide additional insights as to how the number of total bikes rented in an hour grows over time.

This model thus provides certain guidance for future studies of bike-share use. Apart from working towards resolving the aforementioned weaknesses, we could look deeper into the possibility of an interaction between the factor indicating weather situation and each of the quantitative explanatory variables (as discussed earlier) because of the different regression slope of the 4th category. Exploring the possibility that the 2 categories (1 from each of 2 factors), whose coefficients the LASSO shrunk to exactly zero, are no different from their respective baseline categories could also be productive. Finally, we could also utilize the bootstrap (residual or non-parametric) to do hypothesis testing on the obtained coefficient estimates and create confidence intervals to quantify their stability, although this could be computationally expensive given the large sample size.

5. Conclusion

Overall, the Least Squares model with the 10 explanatory variables and the aforementioned transformations, which has no bias but high variance, has the lowest Prediction Mean Squared Error. We used this final model, together with those resulting from the model shrinkage methods, to make predictions on the test data that we held out initially. We obtained Mean Squared Prediction Errors of 1.072, 1.063 and 1.058 for Ridge Regression, LASSO and Least Squares Regression respectively. This validates our results about the relative performance of the three models but hints at a possibility of slight overfitting, which would require further exploration to justify. Therefore, from this analysis, we conclude that the best regression model for predicting the total number of bikes rented in a particular hour using environmental conditions and other available explanatory variables is the chosen Least Squares Model. However, it is of utmost necessity to be mindful of the uncertainty regarding the validity of interpretations and conclusions drawn from this final model before performing further inference with it.