Sachit Patel

DSC 323 Final Report

Seoul Bike Sharing Dataset

1. Introduction

This dataset is about the rental bike system in Seoul and addressing the concern over if there are enough bikes available to those who need them in the city. There is a wide array of potential predictor variables given, such as rainfall, snowfall, holidays, time of day, and temperature. These predictors fall under two groups of contexts: weather conditions and date-time conditions. These are the contexts through which the relationships between data will exist. The objective of this data set is to help predict and estimate a stable supply of rental bikes throughout the year for the city's demand. As such, the rental bike count will be the dependent variable in the models created within this analysis.

2. Exploratory Analysis:

One of the most critical things when given a new dataset is to understand not only what each type of value means, but also get an intuitive understanding of these data points. Various methods were used to explore the data set and get a rough "feel" or understanding of the data. The purpose of the exploratory analysis is to answer as many concerns as wanted regarding "what do I want to know about the data?"

First, histograms were used to explore the distribution of rental bike count. This felt like a natural first step after deciding that rental bike count would be the dependent variable. This led to the discovery of how skewed the distribution of the rental bike count per hour was (see figure 2.1 in the appendix), and attempted transformations of it (more on that later). The distribution of rental bike count per hour on the histogram is abnormal, and as such breaks the givens associated with normal distribution.

Correlation tables and scatterplot matrices were used to visualize the effect that each predictor has on the dependent variable (rental bike count and transformations of rental bike count). These showed me that some values have slight positive correlation (roughly around .5), and some have very small negative correlation (see figure 2.2 in the appendix). The predictors weren't much stronger after transformations, either, so they notably lack a significant linear relationship with rental bike count. One area of interest then is to transform the predictor variables if it's a strong option. Histograms of predictor variables could be analyzed.

The means procedure was also used to see various statistics about each predictor and the dependent variable (see figure 2.3 in the appendix). This is where I first identified that 295 observations of 0 are found in rental bike count (because the log transformation left these values as undefined, and there are no negative values in this dataset). From this means procedure, other predictors can be identified with a non-normal distribution through their high standard deviation values. Several of the weather related predictors (visibility, temperature, humidity) are affected by this.

3. Trying Interaction Variables:

A few interaction variables were considered for this model. The idea behind these interaction terms is that they would add new insights into the model and cover certain conditions more.

One interaction variable attempted was between the hour and functioningDay predictors, and their product. The predictor functioningDay was turned into a dummy variable (0 if the rental bike system was not functioning during that hour, and a 1 if it was). Since it's either a 0 or 1, only the cases in which the dummy variable (numFunc) is 1 is the predictor accounted for. This would strictly account for the rental bike count during business hours. However, when this term is used, it is read as a "biased" predictor within SAS and replaces another predictor (usually hour). This interaction term was not deemed later on during variable selection and was thus removed. Given how we want to predict how many rental bikes are needed anyway, including cases where rental bikes are 0 (they are not available that day) is counterproductive to the objective of analyzing this dataset.

Another interaction variable that was considered was the interaction between hour and numHoliday (the dummy variable of holiday where 0 is equal to "No Holiday" and 1 is equal to "Holiday"). This would allow for a better explanation of the rental bike count during holidays. This I thought would help address clusters of observations that are really high in terms of rental bike count, since outliers and influential points within the data set seemed to crop up within batches. It makes sense as well that there would be more bike activity over a holiday time-frame, where students are not in school and/or there are no business hours. It fits the real-world context of there being a change in the date-time context (around specific holidays) as well as the weather conditions in some cases (snowfall on Christmas, etc.). This interaction variable was a significant predictor (p value < 0.05) and remained within the final model.

4. **Using Transformations:**

Histograms were used in the data exploration stage to analyze the distribution of the dependent variable (rental bike count in this case). Reviewing the case of the histogram for rental bike count (figure 2.1 in the appendix), there are a lot of outliers primarily on the right side of the graph. The main issue is the skewness of the histogram. The skew is heavily skewed to the left, or negatively skewed. The skewness value of the rental bike count histogram is roughly 1.15. Additionally, the p values for all 3 Goodness-of-Fit Tests (Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling) were found to be statistically significant (less than 0.01, less than

0.005 for Cramer-von Mises and Anderson-Darling respectively). Thus, we have sufficient evidence to reject the null hypothesis which states that the distribution of the rental bike count is a normal distribution.

As a result, transformations were attempted to create a more balanced dependent variable. The log transformation was the first to be attempted (see figure 4.1 in the appendix for details). The log transformation has several benefits to it. The log transformation is more positively skewed, or skewed to the right than the original histogram (roughly -0.804). The log histogram is not necessarily much better with its distribution. It is closer to normal distribution than the original, but not by much. The second attempt was through the square root transformation. The square root transformation (see figure 4.2 in the appendix) also provides similar, yet relevant results upon use. The skew of the graph is the closest to normal distribution of the 3, approximately 0.237. Unlike the log histogram, the root histogram is skewed more to the left or positively skewed, like the original histogram.

The choice between the square root and log transformations is not necessarily clear. It's probable that a model could be built off of either transformation. When comparing statistics, which model is better? Although the square root transformation has a skew closer to normal distribution, the log transformation has other variables closer to normal distribution or show a better fit. Such examples include: Kurtosis (approximately 0.52 compared to approximately -0.66), Standard Deviation (approximately 1.12 compared to approximately 12.45), and Variance (approximately 1.35 compared to approximately 155). There can be a model creation and validation process done for both the log and square root versions of the rental bike count variable to see which whole model is a better fit. Although the square root transformation has a skewness value closer to 0 (the skewness value of normal distribution), there are multiple other values in the logarithmic transformation that are closer to normal distribution that overall make it a better fit (Kurtosis, Standard Deviation, and notably Variance). This would be the end of the comparison, but there is one other concern. When performing the log transformation, 295 observations are removed from the data set. This is because the log of 0 is undefined, so these observations are not counted. This reduces the observations the log transformation looks at. As touched on in the exploratory analysis, the observations of 0 coincide when it is not a functioning day within the dataset (when rental bikes are not offered). These observations do not fit with the objective of the analysis (to help estimate and predict the rental bike count per hour based on weather and date-time conditions) and thus obscures the model. As such, the log transformation is a stronger fit for this model than the square root transformation.

5. Comparing Potential Models:

To start, two full models were examined. The same set of predictors (hour of day, temperature, humidity, wind speed, visibility, dewpoint, solar radiation, snowfall, rainfall, season, holiday, functioning day) between the two models. The difference is in the dependent variable. One model uses the regular rental bike count variable, the other uses the log transformation of rental bike count.

There are some differences between the two models. The rental bike count model will be referred to as M1, and the log transformation model as M2. Each of their variance analyses can be found in figure 5.1 in the appendix. M1 generally has lower values than M2. Adjusted r-squared (0.5439 against 0.6003), F-value (871.55 against 1156.4). Model M2 also has much lower error terms across the board (a lower RMSE by around 430 for example), By statistics alone, model M2 is a better fit. However, there’s one concern with model M2. Within SAS, numFunc (dummy variable for the functioningDay predictor) is overlapped with the intercept value and read as a “biased” predictor. This is because numFunc is a “linear combination of other variables” according to SAS. M1 does not have this error regarding numFunc. The model seems to output roughly the same performance without numFunc. As such, this predictor will be removed from model M2.

6. Checks for Multicollinearity:

Multicollinearity is one of the 3 checks done early on with a model (alongside outliers and influential points). To get an understanding of a predictor’s multicollinearity, we can look at the predictor’s VIF (or Variance Inflation Factor) value. Generally, a predictor with a VIF of 10 or higher shows significant multicollinearity that needs to be addressed.

Within these particular models there are 3 predictors with over a 10 in VIF (see figure 6.1 in the appendix). These three predictors are temperature, humidity, and dew point temperature. Of these, dew point temperature has the highest multicollinearity, of over 116 in model M2. As for what predictors dew point temperature seems to be collinear with, temperature and humidity mainly, with around 88 and 20 VIF respectively within model M2.

Both models have similar presenting issues in this case, so the first step action was taken equally for each model. The first step was to remove dew point temperature as a predictor from model M2 and recalculate their VIF values to see if multicollinearity was still present within each model. After removing dew point temperature, no significant multicollinearity was found in model M2. The VIF factors in each model for the temperature and humidity predictors were around 2, which is no longer significant multicollinearity. I believe this is fine to be left in the model, as temperature and humidity are still both significant predictors. Dew point temperature was a significant predictor in the full model (with a p value less than 0.05) with all predictors involved. However, the removal of multicollinearity (in a case where multicollinearity can not be ignored) is more prominent. This type of multicollinearity can't be ignored since dew point temperature is another predictor that is a control variable, it's being used to help predict rental bike count. It is not a predictor that's merely just a combination of others (although it is closely related to temperature).

7. Outlier & Influential Point Analysis:

To start, recall how the log transformation left 295 values unread during the regression within the full model. This is because these 295 observations are not observed as functioning days within the date-time sense, meaning rental bikes were not available during these entries (for each hour). Since these entries aren't relevant to the model, the log transformation can be used efficiently. As thus, these 295 observations were deemed outliers, and removed from the model.

All remaining observations were analyzed for potential outliers. One way to classify outliers is by studentized residuals. An observation can be considered an outlier by studentized residuals if the residual for that observation is either greater than 3, or less than negative 3. In this dataset there were 110 observations that had a studentized residual greater than 3 or less than negative 3. All of these outliers were removed from the dataset. This improved the model overall, with about a 0.06 increase in adjusted r-squared for example.

Influential points can be determined using their dffits value. This uses the formula 2*sqrt(number of predictors+1/n), n being the number of observations. The number of observations (after 295 observations are overlooked because of log transformations and 110 outliers were removed) is 8355. After the calculation, the dffits value for a highly influential point is roughly .07257. An observation with the absolute value of it greater than this amount would be considered a highly influential point. When examining the dataset, there are plenty of influential points within the model. Influential points can't be removed exactly 1 to 1 like outliers. This is because generally removing outliers are a net positive, but not influential points. Removing influential points will change the nature of the model, not just improving it overall. Additionally, the dfbeta value is the change in the specific parameter when that specific observation is cut out of the dataset. Generally we’re looking for when the absolute value of a dfbeta is greater than 2 divided by the square root of the total amount of observations, which comes out to around .02188. This allows us to tell which influential points influence what predictors.

No significant action was taken during the full model stage with influential points. Two observations that had very high dffits and dfbetas values were removed for experimentation, but not much more, seeing they had made minimal positive impacts (nothing negative, at the very least).

8. Residual Plot Analysis and Model Assumptions:

Using normality plots and residual plots, we can see if a model passes the 4 assumptions. The first assumption is that the linear regression model is that the model has a linear relationship between x and y. In the normality plot there is roughly a line, a linear relationship between the CDF of studentized residual (x) and the normal cumulative distribution. Model M2 passes the first assumption. See figure 8.1 in the appendix for a visual. The second assumption is that the model predictors are generally independent from each other. This assumption does not pass for model M2, as there are clear patterns within the residual graphs. Either there is some pattern, or the residual points are clustered for model M2. See figure 8.2 in the appendix for clarification. By this same logic, the third assumption also fails, since constant variance is violated as a result. With both of these assumptions failing, we know that prediction intervals and confidence intervals will be incorrect. This is a fatal flaw in the model. The fourth assumption is that each x-variable has a linear relationship through scatter plots and residuals. Using a scatterplot matrix, it was found that the predictors didn't have a linear relationship with the log transformation of rental bike count. The fourth assumption is also violated.

What's concerning is that solutions have already been tried and utilized. Up to this point, multicollinearity has already been addressed in this model. No predictor has a VIF over 3 with the removal of dew point temperature from the model. Additionally, transformations were already attempted earlier on to fit closer to normal distribution. The log transformation was settled on, and none of the other transformation options pass these assumptions either. This is a loss the model has to take to my knowledge. This is a significant error with the model that doesn't seem to have a way to be addressed.

9. Training and Testing Set Model Validation:

A training and testing split is done with the total amount of observations to estimate a model equation's accuracy in predicting new input data. For this model, a 75% to 25% split was performed, 75% going to the training set, and 25% to the testing set. This left the training set with 6265 observations (do note that this is based on the 8353 count of observations. This count includes the removal of outliers, some influential points, and cases where rental bike count is equal to 0).

There is a strong correlation between lnrbc and yhat (the trained and predicted values, see figure 9.1 in the appendix), suggesting that the model fits decently. Both RMSE and MAE are reflective of this as well, with their values being a decimal (lower than 1, greater than 0, see figure 9.2 in the appendix). Again, because the assumptions are violated, predictions can't be certain. The least squares sum is also off because of the normality assumptions, so the errors here are reflective of that.

10. Variable Selection/Model Selection Processes:

From working with the variables originally, I already had some information on their relevance. Most of the predictors used were relevant (by the p value less than 0.05) from the exploratory phase. The major exception was wind speed, which wasn't relevant to the model even after addressing influential points and outliers.

Model selection methods were used to help determine the final model. Multiple different selection methods were utilized to help determine which potential model would fit best after the training and testing split. Backward, stepwise, and cp model selection methods were used (figure 10.1 in the appendix).

As an aside, other transformations were attempted, such as the square transformation, the cube transformation, and the reciprocal transformation within the exploratory analysis stage. These histograms had more severe errors than the log and square root histograms, and thus were not considered. Quadratic and Cubic models were also tested using the glmselect procedure after the model validation stage. Each version respectively gave higher adjusted r-squared for the model but added a high number of predictors and lowered other statistics about the model (such as the F-value). The quadratic and cubic models could not repair the errors in the model assumptions, either.

11. Examine Final Model:

Model selection methods were used to help determine the final model. Multiple different selection methods were utilized to help determine which potential model would fit best. The final model. Two model selection methods (backwards and CP) ended up with the same set of predictors as their best choice. See figure 11.1 in the appendix for output.

The final model has the following predictors: hour, temperature, humidity, visibility, rainfall, snowfall, season, and holiday (if the day is a holiday or not). The final model has an adjusted r-squared of roughly 0.6465. There are other strong terms, such as the RMSE of 0.6, and the F-Value of around 1900.

The final model equation is the following: lnrbc = 5.6161 + 0.047(hour) + 0.041(temperatureC) – 0.0135(humidity) -557E-7(visibility10m) – 0.3288(rainfallMm) - 0.0867(snowfallCm) + 0.2551(numSeason) -0.2945(numHoliday)-0.0084(hour_numHoliday) + 3623.23549 [this is the error sum of squares].

12. Prediction Intervals:

95% confidence intervals were computed for the modified dataset (8353 observations). As expected, the 95% confidence intervals have a noticeably high range for lnrbc, roughly predicting 1.3 to 1.5 increase or decrease in either direction for lnrbc (for the upper and lower intervals). This is due to the assumptions being violated, thus the predictions having a wide margin of error.

13. Potential Model Improvements:

There are several ways this model can be improved, but they mostly, if not entirely stem from the model assumptions being violated, which result in multiple errors. Errors such as the prediction intervals having a large range for a 95% interval, higher error terms (such as MAE and RMSE). To improve this model, the residual plots would need to be independent of each other, they would require constant variance, and the predictors need a linear relationship with the dependent variable (rental bike count). In addition, the adjusted r-squared significant room for improvement. The final model presented within this analysis is an insufficient one due to the reasons listed above.

Appendix:

Section 2: Exploratory Analysis

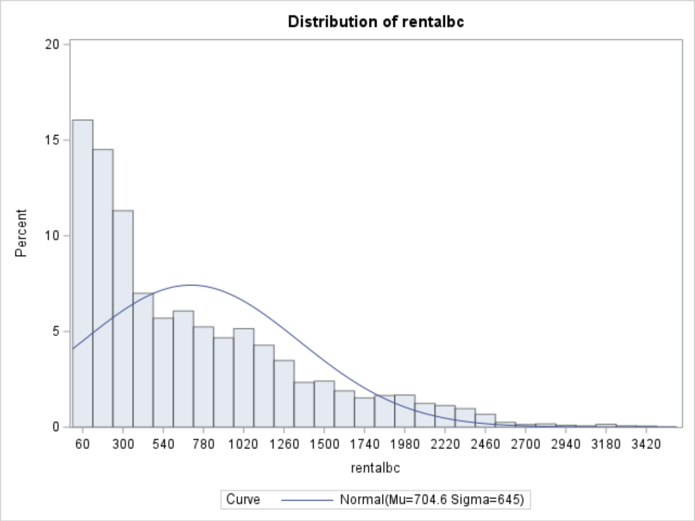
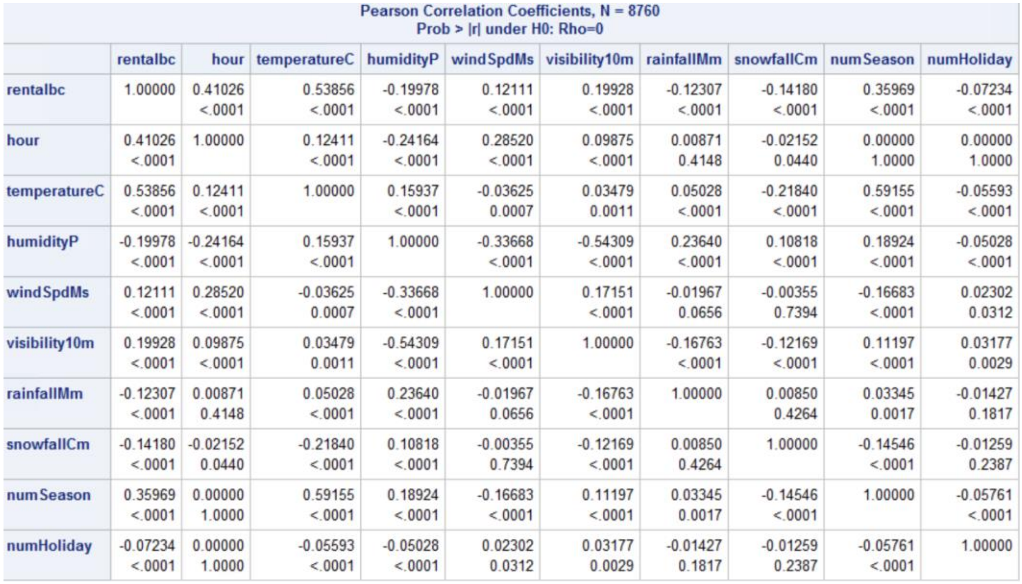


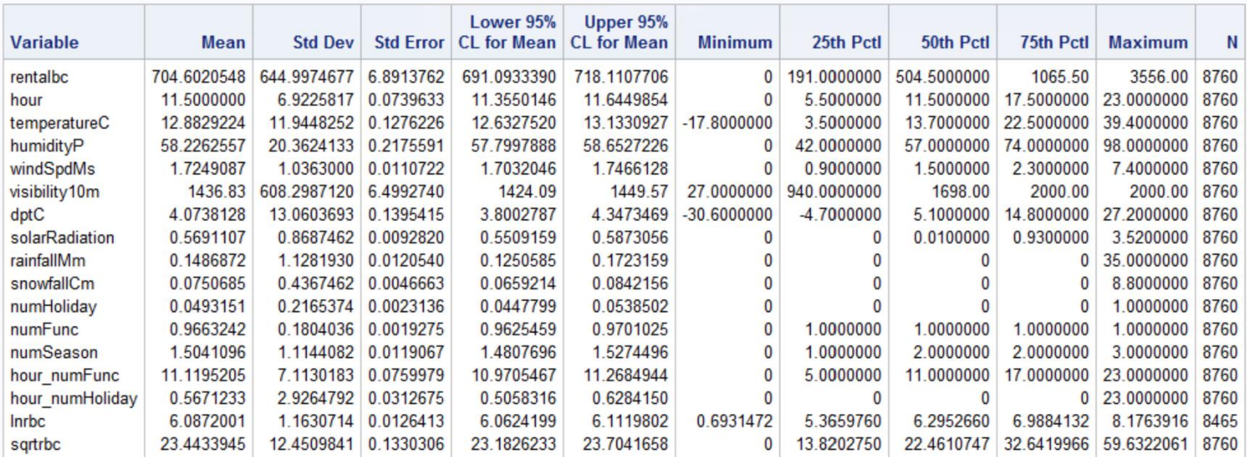
Figure 2.1: A histogram which depicts the distribution of rentalbc, the dependent variable which represents rental bike count.



Pearson Correlation Coefficients, N = 8760
Prob > |r| under H0: Rho=0

	rentalbc	hour	temperatureC	humidityP	windSpdMs	visibility10m	rainfallMm	snowfallCm	numSeason	numHoliday
rentalbc	1.00000	0.41026 <.0001	0.53856 <.0001	-0.19978 <.0001	0.12111 <.0001	0.19928 <.0001	-0.12307 <.0001	-0.14180 <.0001	0.35969 <.0001	-0.07234 <.0001
hour	0.41026 <.0001	1.00000	0.12411 <.0001	-0.24164 <.0001	0.28520 <.0001	0.09875 <.0001	0.00871 0.4148	-0.02152 0.0440	0.00000 1.0000	0.00000 1.0000
temperatureC	0.53856 <.0001	0.12411 <.0001	1.00000	0.15937 <.0001	-0.03625 0.0007	0.03479 0.0011	0.05028 <.0001	-0.21840 <.0001	0.59155 <.0001	-0.05593 <.0001
humidityP	-0.19978 <.0001	-0.24164 <.0001	0.15937 <.0001	1.00000	-0.33668 <.0001	-0.54309 <.0001	0.23640 <.0001	0.10818 <.0001	0.18924 <.0001	-0.05028 <.0001
windSpdMs	0.12111 <.0001	0.28520 <.0001	-0.03625 0.0007	-0.33668 <.0001	1.00000	0.17151 <.0001	-0.01967 0.0656	-0.00355 0.7394	-0.16683 <.0001	0.02302 0.0312
visibility10m	0.19928 <.0001	0.09875 <.0001	0.03479 0.0011	-0.54309 <.0001	0.17151 <.0001	1.00000	-0.16763 <.0001	-0.12169 <.0001	0.11197 <.0001	0.03177 0.0029
rainfallMm	-0.12307 <.0001	0.00871 0.4148	0.05028 <.0001	0.23640 <.0001	-0.01967 0.0656	-0.16763 <.0001	1.00000	0.00850 0.4264	0.03345 0.0017	-0.01427 0.1817
snowfallCm	-0.14180 <.0001	-0.02152 0.0440	-0.21840 <.0001	0.10818 <.0001	-0.00355 0.7394	-0.12169 <.0001	0.00850 0.4264	1.00000	-0.14546 <.0001	-0.01259 0.2387
numSeason	0.35969 <.0001	0.00000 1.0000	0.59155 <.0001	0.18924 <.0001	-0.16683 <.0001	0.11197 <.0001	0.03345 0.0017	-0.14546 <.0001	1.00000	-0.05761 <.0001
numHoliday	-0.07234 <.0001	0.00000 1.0000	-0.05593 <.0001	-0.05028 <.0001	0.02302 0.0312	0.03177 0.0029	-0.01427 0.1817	-0.01259 0.2387	-0.05761 <.0001	1.00000

Figure 2.2: A Pearson Correlation Coefficients table which lists the correlation values for all predictors (that which carry over to the final model at the end of the entire analysis) alongside rentalbc.



Variable	Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean	Minimum	25th Pctl	50th Pctl	75th Pctl	Maximum	N
rentalbc	704.6020548	644.9974677	6.8913762	691.0933390	718.1107706	0	191.0000000	504.5000000	1065.50	3556.00	8760
hour	11.5000000	6.9225817	0.0739633	11.3550146	11.6449854	0	5.5000000	11.5000000	17.5000000	23.0000000	8760
temperatureC	12.8829224	11.9448252	0.1276226	12.6327520	13.1330927	-17.8000000	3.5000000	13.7000000	22.5000000	39.4000000	8760
humidityP	58.2262557	20.3624133	0.2175591	57.7997888	58.6527226	0	42.0000000	57.0000000	74.0000000	98.0000000	8760
windSpdMs	1.7249087	1.0363000	0.0110722	1.7032046	1.7466128	0	0.9000000	1.5000000	2.3000000	7.4000000	8760
visibility10m	1436.83	608.2987120	6.4992740	1424.09	1449.57	27.0000000	940.0000000	1698.00	2000.00	2000.00	8760
dptC	4.0738128	13.0603693	0.1395415	3.8002787	4.3473469	-30.6000000	-4.7000000	5.1000000	14.8000000	27.2000000	8760
solarRadiation	0.5691107	0.8687462	0.0092820	0.5509159	0.5873056	0	0	0.0100000	0.9300000	3.5200000	8760
rainfallMm	0.1486872	1.1281930	0.0120540	0.1250585	0.1723159	0	0	0	0	35.0000000	8760
snowfallCm	0.0750685	0.4367462	0.0046663	0.0659214	0.0842156	0	0	0	0	8.8000000	8760
numHoliday	0.0493151	0.2165374	0.0023136	0.0447799	0.0538502	0	0	0	0	1.0000000	8760
numFunc	0.9663242	0.1804036	0.0019275	0.9625459	0.9701025	0	1.0000000	1.0000000	1.0000000	1.0000000	8760
numSeason	1.5041096	1.1144082	0.0119067	1.4807696	1.5274496	0	1.0000000	2.0000000	2.0000000	3.0000000	8760
hour_numFunc	11.1195205	7.1130183	0.0759979	10.9705467	11.2684944	0	5.0000000	11.0000000	17.0000000	23.0000000	8760
hour_numHoliday	0.5671233	2.9264792	0.0312675	0.5058316	0.6284150	0	0	0	0	23.0000000	8760
lnrbc	6.0872001	1.1630714	0.0126413	6.0624199	6.1119802	0.6931472	5.3659760	6.2952660	6.9884132	8.1763916	8465
sqrtrbc	23.4433945	12.4509841	0.1330306	23.1826233	23.7041658	0	13.8202750	22.4610747	32.6419966	59.6322061	8760

Figure 2.3: The MEANS procedure output for the Seoul Bike Sharing dataset.

References:

Section 4: Transformations

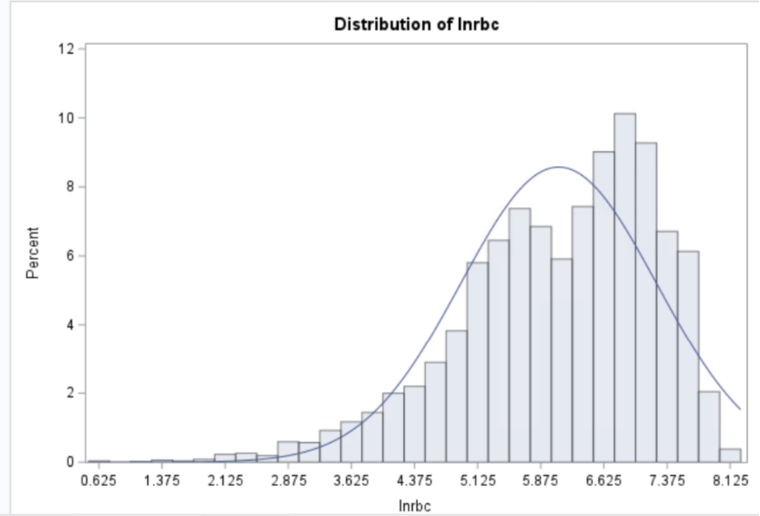


Figure 4.1: Histogram displaying the distribution of lnrbc, a transformed dependent variable (lnrbc = log(rentalbc)).

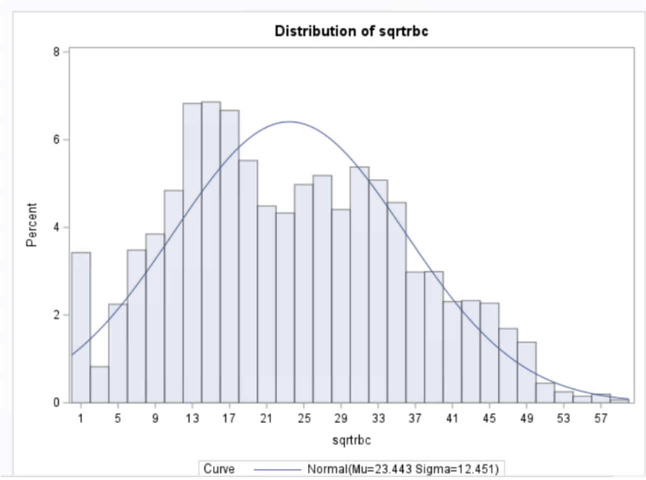
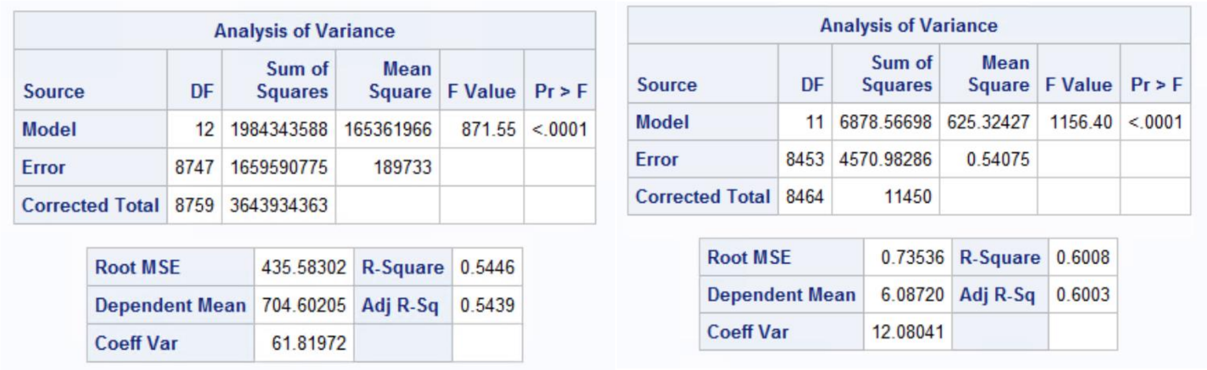


Figure 4.2: Histogram depicting the distribution of sqrtrbc, a variable that is a transformation of rentalbc (sqrtrbc = sqrt(rentalbc)).

Section 5: Potential Models



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	1984343588	165361966	871.55	<.0001
Error	8747	1659590775	189733		
Corrected Total	8759	3643934363			

Root MSE	435.58302	R-Square	0.5446
Dependent Mean	704.60205	Adj R-Sq	0.5439
Coeff Var	61.81972		

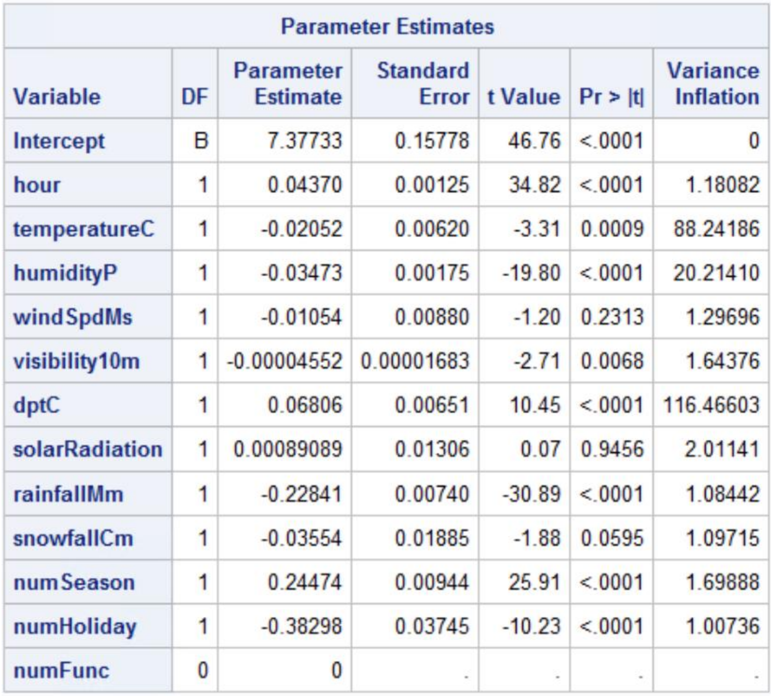
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	6878.56698	625.32427	1156.40	<.0001
Error	8453	4570.98286	0.54075		
Corrected Total	8464	11450			

Root MSE	0.73536	R-Square	0.6008
Dependent Mean	6.08720	Adj R-Sq	0.6003
Coeff Var	12.08041		

Figure 5.1: Variance analysis comparison between two models: model M1 (left), and model M2 (right).

Section 6: Multicollinearity



Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	B	7.37733	0.15778	46.76	<.0001	0
hour	1	0.04370	0.00125	34.82	<.0001	1.18082
temperatureC	1	-0.02052	0.00620	-3.31	0.0009	88.24186
humidityP	1	-0.03473	0.00175	-19.80	<.0001	20.21410
windSpdMs	1	-0.01054	0.00880	-1.20	0.2313	1.29696
visibility10m	1	-0.00004552	0.00001683	-2.71	0.0068	1.64376
dptC	1	0.06806	0.00651	10.45	<.0001	116.46603
solarRadiation	1	0.00089089	0.01306	0.07	0.9456	2.01141
rainfallMm	1	-0.22841	0.00740	-30.89	<.0001	1.08442
snowfallCm	1	-0.03554	0.01885	-1.88	0.0595	1.09715
numSeason	1	0.24474	0.00944	25.91	<.0001	1.69888
numHoliday	1	-0.38298	0.03745	-10.23	<.0001	1.00736
numFunc	0	0	.	.	.	.

Figure 6.1: Variance Inflation Factor check for model M2. High cases of multicollinearity are detected, particularly between temperature, humidity, and dew point temperature.

Section 8: Residual Plot Analysis and Model Assumptions

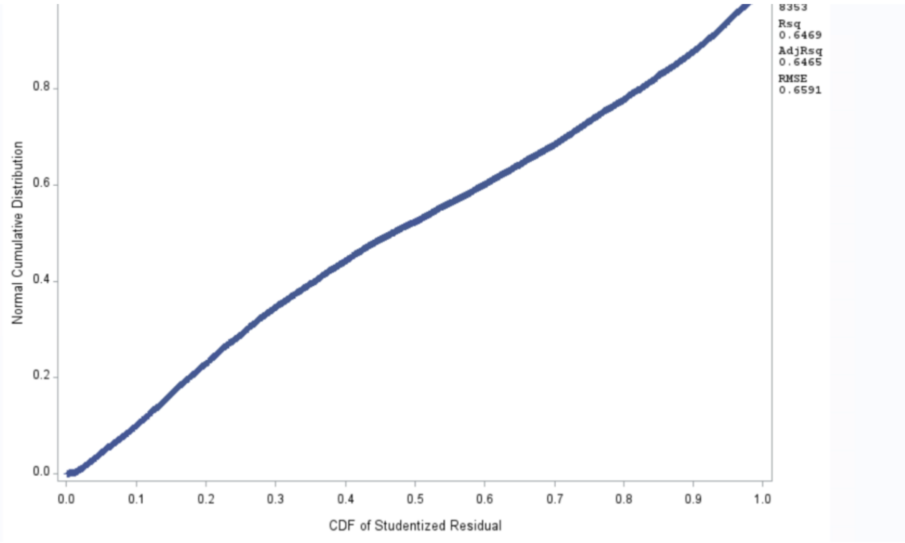


Figure 8.1: CDF of Studentized Residual Plot for model M2.

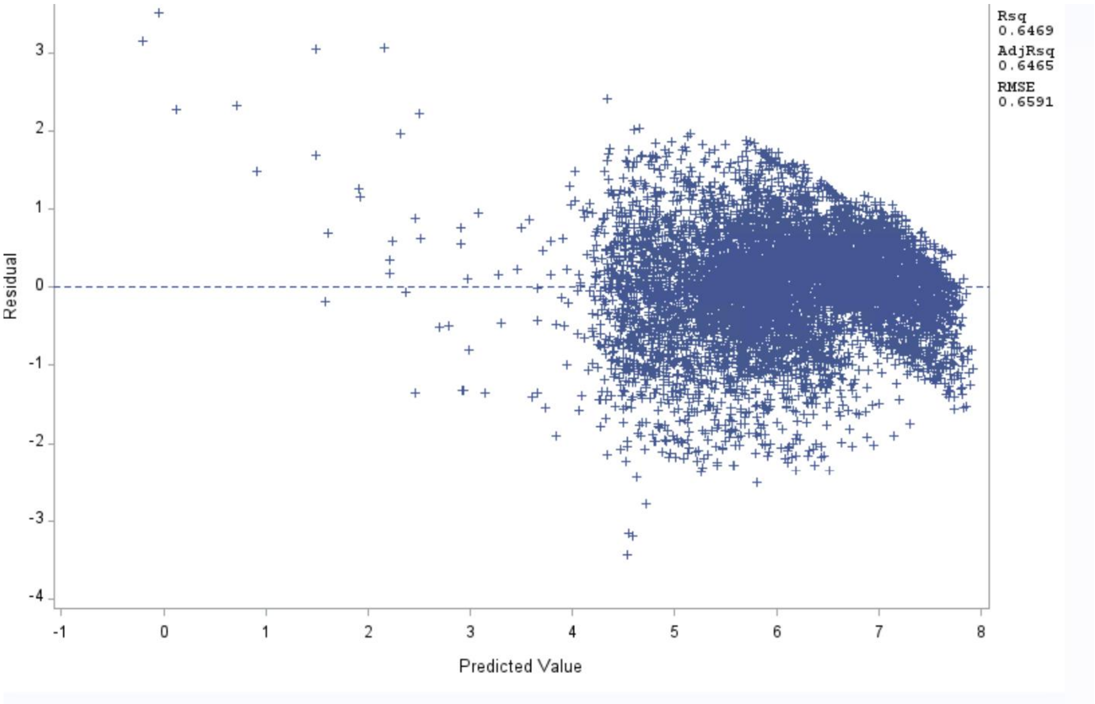
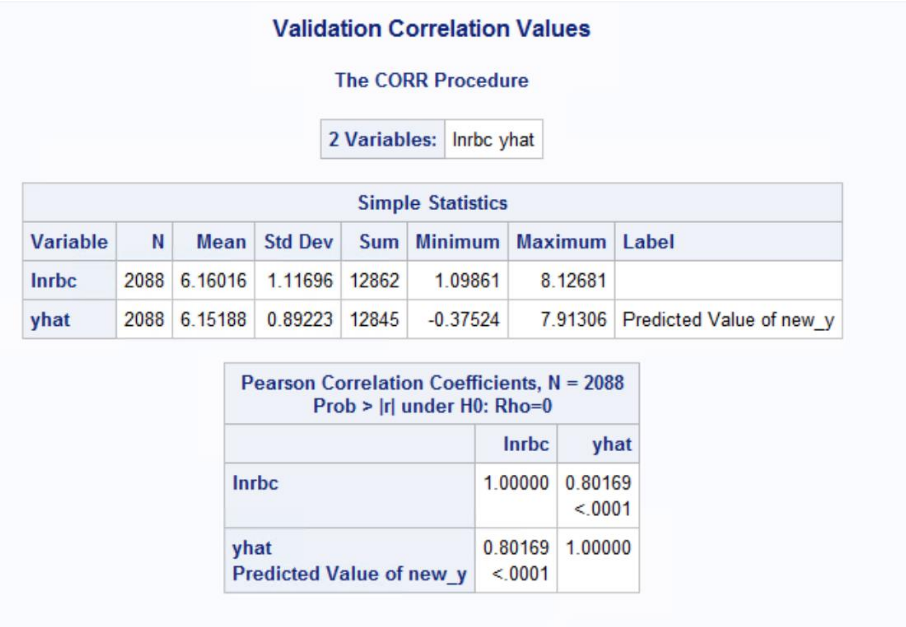


Figure 8.2: Predicted Value vs Residual Value plot for model M2. Multiple assumptions are violated (constant variance and independence) due to the patterned cluster of residuals.

Section 9: Training and Testing Set Model Validation



Validation Correlation Values

The CORR Procedure

2 Variables:	lnrbc yhat

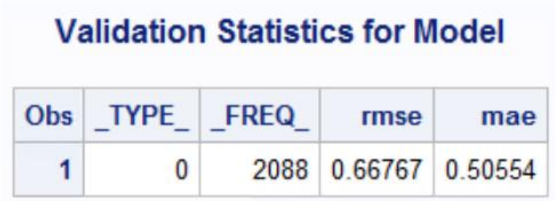
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
lnrbc	2088	6.16016	1.11696	12862	1.09861	8.12681	
yhat	2088	6.15188	0.89223	12845	-0.37524	7.91306	Predicted Value of new_y

Pearson Correlation Coefficients, N = 2088
Prob > |r| under H0: Rho=0

	lnrbc	yhat
lnrbc	1.00000	0.80169 <.0001
yhat Predicted Value of new_y	0.80169 <.0001	1.00000

Figure 9.1: The correlation procedure for the dataset outm1.stats, which is the testing set taken from model validation. The correlation is shown between lnrbc and yhat (the predicted value for new_y) which predicts lnrbc for that data entry.

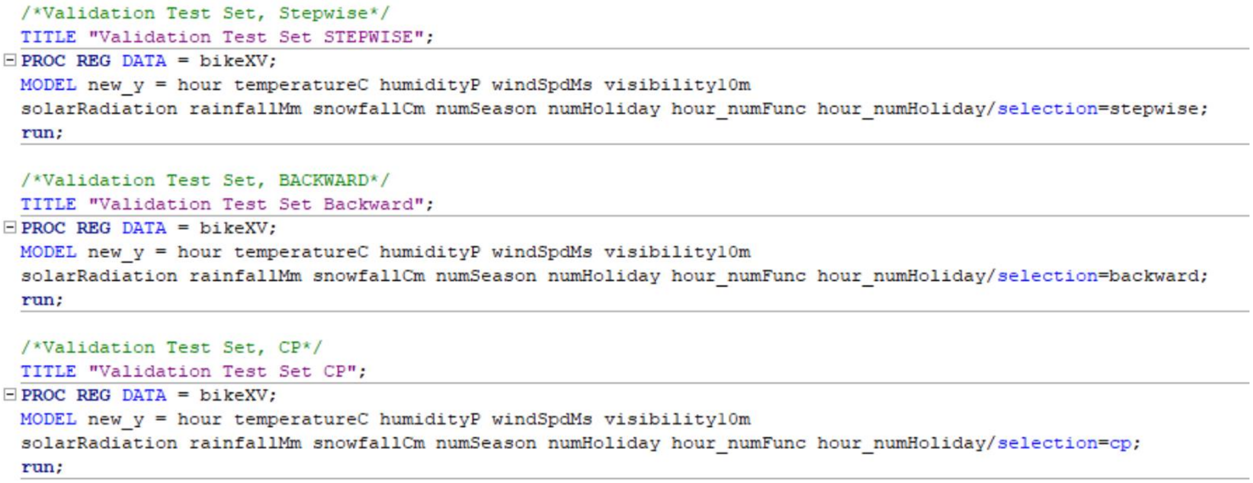


Validation Statistics for Model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	2088	0.66767	0.50554

Figure 9.2: Root mean square error and mean absolute error for the testing set in model M2.

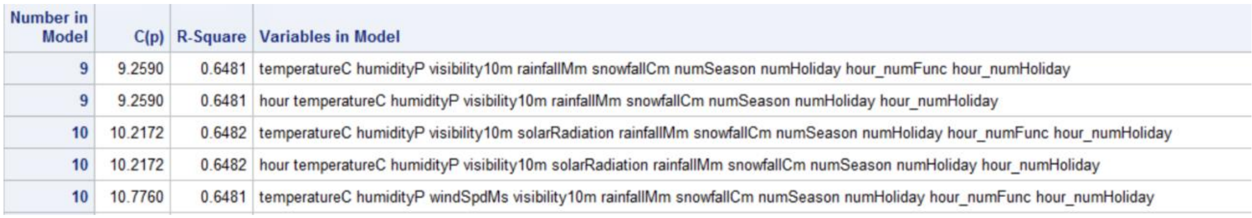
Section 10: Variable Selection/Model Selection Processes



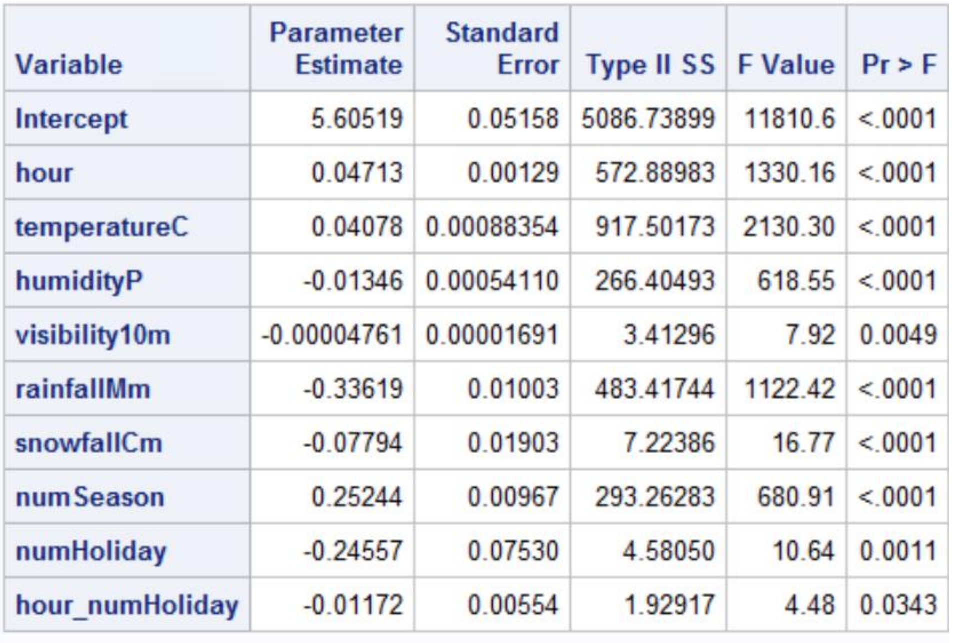
```
/*Validation Test Set, Stepwise*/
TITLE "Validation Test Set STEPWISE";
PROC REG DATA = bikeXV;
MODEL new_y = hour temperatureC humidityP windSpdMs visibility10m
solarRadiation rainfallMm snowfallCm numSeason numHoliday hour_numFunc hour_numHoliday/selection=stepwise;
run;

/*Validation Test Set, BACKWARD*/
TITLE "Validation Test Set Backward";
PROC REG DATA = bikeXV;
MODEL new_y = hour temperatureC humidityP windSpdMs visibility10m
solarRadiation rainfallMm snowfallCm numSeason numHoliday hour_numFunc hour_numHoliday/selection=backward;
run;

/*Validation Test Set, CP*/
TITLE "Validation Test Set CP";
PROC REG DATA = bikeXV;
MODEL new_y = hour temperatureC humidityP windSpdMs visibility10m
solarRadiation rainfallMm snowfallCm numSeason numHoliday hour_numFunc hour_numHoliday/selection=cp;
run;
```

Figure 10.1: Model selection processes utilized for the new_y model involving the training set.

Section 11: Examine Final Model



Number in Model	C(p)	R-Square	Variables in Model
9	9.2590	0.6481	temperatureC humidityP visibility10m rainfallMm snowfallCm numSeason numHoliday hour_numFunc hour_numHoliday
9	9.2590	0.6481	hour temperatureC humidityP visibility10m rainfallMm snowfallCm numSeason numHoliday hour_numHoliday
10	10.2172	0.6482	temperatureC humidityP visibility10m solarRadiation rainfallMm snowfallCm numSeason numHoliday hour_numFunc hour_numHoliday
10	10.2172	0.6482	hour temperatureC humidityP visibility10m solarRadiation rainfallMm snowfallCm numSeason numHoliday hour_numHoliday
10	10.7760	0.6481	temperatureC humidityP windSpdMs visibility10m rainfallMm snowfallCm numSeason numHoliday hour_numFunc hour_numHoliday



Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	5.60519	0.05158	5086.73899	11810.6	<.0001
hour	0.04713	0.00129	572.88983	1330.16	<.0001
temperatureC	0.04078	0.00088354	917.50173	2130.30	<.0001
humidityP	-0.01346	0.00054110	266.40493	618.55	<.0001
visibility10m	-0.00004761	0.00001691	3.41296	7.92	0.0049
rainfallMm	-0.33619	0.01003	483.41744	1122.42	<.0001
snowfallCm	-0.07794	0.01903	7.22386	16.77	<.0001
numSeason	0.25244	0.00967	293.26283	680.91	<.0001
numHoliday	-0.24557	0.07530	4.58050	10.64	0.0011
hour_numHoliday	-0.01172	0.00554	1.92917	4.48	0.0343

Figure 11.1: Model selection outputs for code above (figure 10.1) regarding the CP and backwards selection methods respectively. The backwards selection and the 2nd CP selection option have the same amount and set of predictor. This model was chosen as the final model.