Jeffrey Chou

STAT 4210

December 1, 2016

Final Project

Utilizing Weather to Predict Daily Bike Rentals in Washington DC’s Capital Bikeshare System

**ABSTRACT**

Washington DC’s Capital Bikeshare provides a wealth of data regarding daily bike rentals that can be supplmented with daily weather metrics. In a previous study (STAT 3130 Project), a model was built to predict daily rental counts based on three weather variables, but the adjusted R2 was somewhat low at 46%. More variables, including categorical, were added to the model in this study and boosted the adjusted R2 to 57%. It is difficult to get an R2 value above 65% due to other factors besides weather affecting rental counts, but suggestions for further analysis are provided.

**INTRODUCTION**

Like most other large cities, Washington, DC has a bike rental system called Capital Bikeshare that allows anyone to rent and return a bicycle at any of its 350+ bike docks throughout the area. In addition, the system collects data on every single rental, such as duration, origin/return location, and user type (e.g. casual vs registered user). The data collected is uploaded to their website and open to the general public.

Data used for this study comes from two differences sources. The first source is the “Bike Sharing Dataset” on UCI’s Machine Learning Repository website (http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset), which contains bike rental and basic weather data of every day from Jan 1, 2011 to Dec 31, 2012. For this study, we used the R package ‘weatherData’ to extract more weather variables from wunderground.com. Both datasets were then merged.

In the previous study, only three quantiative variables were used in the prediction model. This study will analyze if five additional quantitative variables and three additional categorical variables can improve the adjusted R2 value.

The insights gained from the study has the potential to provide cost-reduction benefits to Capital Bikeshare. Bicycles left outside tend to rust faster due to rain and humidity exposure. If a useful model can be built to predict bike rental counts, the model can then be used to estimate how many bikes should be stored indoors during periods of suboptimal weather.

**DATA OVERVIEW**

**Table 1**: Description of Variables

| **Name** | **Type** | **Description/Notes** | **Source** |
| --- | --- | --- | --- |
| **Date** | Identifier | Date | UCI ML |
| **tempmax** | Quantitative, Continuous | Daily max temperature in (°F) | Wunderground.com |
| **tempmin** | Quantitative, Continuous | Daily min temperature (°F) | Wunderground.com |
| **humidity** | Quantitative, Percentage | Average Humidity (%) | Wunderground.com |
| **pressure** | Quantitative, Continuous | Average Barometric air pressure (inHg) | Wunderground.com |
| **visibility** | Quantitative, Integer | Average Visibility (3-10 integer miles) | Wunderground.com |
| **wind** | Quantitative, Continuous | Average wind speed (mph) | Wunderground.com |
| **precip** | Quantitative, Continuous | Total precipitation (in) | Wunderground.com |
| **CloudCover** | Quantitative, Integer | Average cloud cover (integer scale from 0-8, where 0=no clouds and 8=complete cloud cover) | Wunderground.com |
| **Workday** | Categorical, Nominal | 0 is workday, otherwise 1 is weekend or holiday | UCI ML |
| **Ds\_spring**  **Ds\_summer**  **Ds\_fall** | Categorical, Ordinal | Dummy variable for ‘seasons’, where base level=winter. | UCI ML |
| **Dwc\_1**  **Dwc\_2** | Categorical, Ordinal | Dummy variable for ‘weather category’, where base level=3(rain/snow). Dwc\_1=sunny and Dwc\_2=cloudy. | UCI ML |
| **Rentals (y)** | Quantitative, Continuous | **Dependent Variable of Interest**: Number of Daily Bike Rentals | UCI ML |

Table 1 describes all variables used for this study. Tables 2A-C provide distribution of the three categorical variables in the dataset. There are no missing values, and there is no appearance of any unusal distributions.

**Table 2A**: Frequency Table of Seasons

| **Season** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1\_Winter** | 181 | 24.76 | 181 | 24.76 |
| **2\_Spring** | 184 | 25.17 | 365 | 49.93 |
| **3\_Summer** | 188 | 25.72 | 553 | 75.65 |
| **4\_Fall** | 178 | 24.35 | 731 | 100.00 |

**Table 2B**: Frequency Table of Workday

| **Workday** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0\_No** | 231 | 31.60 | 231 | 31.60 |
| **1\_Yes** | 500 | 68.40 | 731 | 100.00 |

**Table 2C**: Frequency Table of Weather

| **Weather** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1\_Sunny** | 463 | 63.34 | 463 | 63.34 |
| **2\_Cloudy** | 247 | 33.79 | 710 | 97.13 |
| **3\_Rain/Snow** | 21 | 2.87 | 731 | 100.00 |

Table 3 provides summary statistics for all quantiative variables and confirms that there are no missing values. Of particular note is that visibility is heavily left-skewed and precipitation is heavily right skewed, otherwise all variables are normally distributed.

**Table 3**: Summary Statistics of All Quantitative Variables

| **Variable** | **N Miss** | **Minimum** | **Lower Quartile** | **Median** | **Mean** | **Upper Quartile** | **Maximum** | **Std Dev** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| tempmax tempmin humidity pressure visibility wind precip CloudCover | 0 0 0 0 0 0 0 0 | 28.000 17.000 22.000 29.110 3.000 1.000 0.000 0.000 | 55.000 38.000 54.000 29.900 9.000 6.000 0.000 4.000 | 71.000 52.000 63.000 30.020 10.000 8.000 0.000 5.000 | 69.274 52.250 63.064 30.025 9.278 8.056 0.109 5.103 | 84.000 68.000 71.000 30.160 10.000 10.000 0.030 7.000 | 105.000 84.000 97.000 30.640 10.000 24.000 3.850 8.000 | 17.592 16.184 12.657 0.215 1.355 3.295 0.338 2.167 |

**Figure 1**: Scatter Plot Matrix of All Quantitative Variables

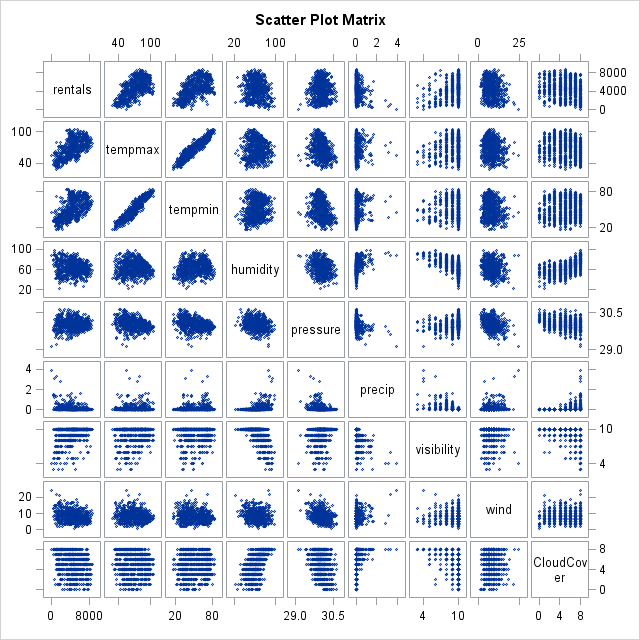


Figure 1 provides visualizations of all quantitative variables versus each other. Tempmin and tempmax appear to have the strongest linear relationships out of all the variables, but they probably have multicollinearity issues since there is also a strong correlation seen in the tempmin and tempmax matrices. In addition, it appears that visibility and cloudcover are both discrete variables, but they will be treated as a quantitative variable here.

**DATA COLLECTION AND PRE-PROCESSING**

There were not enough variables provided in the original dataset provided by UCI Machine Learning, so additional weather variables were collected using the ‘weatherData’ package in R (https://cran.r-project.org/web/packages/weatherData/weatherData.pdf). The package scapes data from wunderground.com for any given location and time range. We programmed the package the collect selected weather variables in DCA (Ronald Reagan National Airport in Washington, DC) from January 1, 2011 to December 31, 2012.

Next, the dataset collected was imported into SAS from R and then merged with the original UCI dataset. A few unnecesary variables were dropped and most of the remaining variables were renamed to shorter titles. Two categorical variables, weathercategory and season were coded into dummy variables.

Weathercat consisted of three categories describing the weather, i.e. sunny, cloudly, or rain/snow. It was recoded as follows, where rain/snow is the base level:

Season consisted of the categories winter, spring, summer, and fall. It was recoded as follows, where the base category was winter.

The last step was to recode the value of ‘T’ in precipitation, which stands for trace precipitation (not enough measurable rain, usually <0.001 inches). ‘T’ was simply converted to 0.00 inches. For some odd reason, precipitation was originally coded as a categorical variable in SAS, but a manipulation was done to change the format to a numerical variable.

**MODEL I: FIRST-ORDER MODEL**

We will start the analysis by proposing a complete first-order linear model to determine if there are some insignificant variables that should be dropped from consideration in the next models. The proposed model is:

Where

* Y = daily bike rental count
* X1 = tempmax
* X2 = tempmin
* X3 = humidity
* X4 = pressure
* X5 = precip
* X6 = visibility
* X7 = wind
* X8 = cloudcover

A global F-Test (alpha=0.05) will first be conducted to determine if at least one coefficient βk that is a statistically significant predictor of daily bike rentals. Thus, the null hypothesis is that all βk values are equal to zero, and the alternate hypothesis is that at least one of the βk values is not equal to zero. Written mathematically:

* HO:
* HA:  at least one

**Table 4**: Global F-Test

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 8 | 1423646052 | 177955756 | 97.64 | <.0001 |
| **Error** | 722 | 1315889340 | 1822561 |  |  |
| **Corrected Total** | 730 | 2739535392 |  |  |  |

From table 4, the results of the global F-test indicate an F-value of 97.64, corresponding to a p-value of <0.0001. Since the p-value is less than alpha, there is evidence to reject the null hypothesis. Thus we conclude that at least one coefficient is significant with regards to predicting bike rentals. The next step is to run the complete regression model in SAS.

**Table 5:** Parameter Estimates

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | 1 | -26963 | 8755.04277 | -3.08 | 0.0022 | 0 |
| **tempmax** | 1 | 44.64599 | 9.85494 | 4.53 | <.0001 | 12.03490 |
| **tempmin** | 1 | 29.30931 | 10.65246 | 2.75 | 0.0061 | 11.90054 |
| **pressure** | 1 | 872.02156 | 287.64635 | 3.03 | 0.0025 | 1.53587 |
| **precip** | 1 | -852.73103 | 182.95181 | -4.66 | <.0001 | 1.53166 |
| **visibility** | 1 | 163.61990 | 47.50468 | 3.44 | 0.0006 | 1.65944 |
| **wind** | 1 | -49.16426 | 17.31238 | -2.84 | 0.0046 | 1.30280 |
| **CloudCover** | 1 | -72.33298 | 28.49896 | -2.54 | 0.0114 | 1.52683 |

After running the complete model, all variables except humidity were found to be significant in predicting rental counts. Humidity was dropped from the model and the remaining variables were analyzed again in SAS, all of which were found to be significant according to table 5. Thus, the proposed fitted first-order model is:

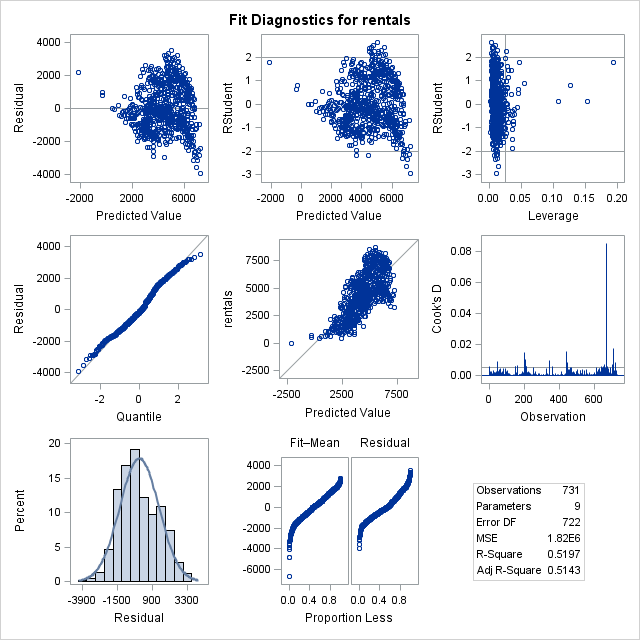
There are some multicollinearity issues with tempmax and tempmin since their VIF’s are above 10. It is possible to z-score the variables in order to reduce the problem. However, it is not completely necessary here since the VIF’s are not severely high, and using the coding procedure will not increase R2 or adjusted R2 values.

**Table 6:** Model Performance Metrics

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 1350.23541 | **R-Square** | 0.5189 |
| **Dependent Mean** | 4504.34884 | **Adj R-Sq** | 0.5142 |
| **Coeff Var** | 29.97626 |  |  |

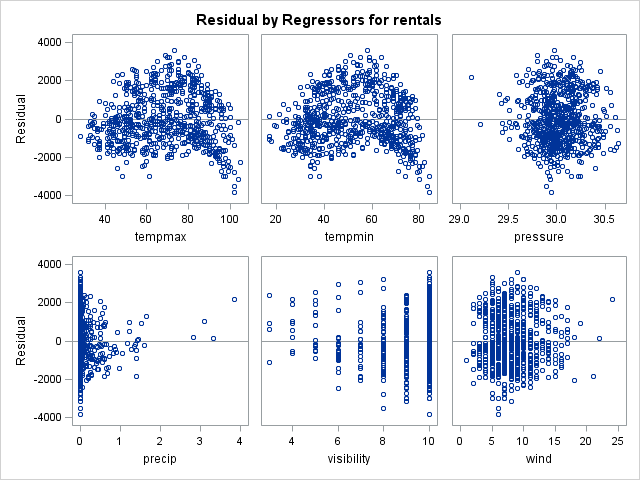
From table 6, the adjusted R2 value is 0.5142. This indicates that 51.42% of the variation in rental counts are explained by the model. Thus, the model is only slightly useful and will need additional variables added to improve performance. Before proceeding, we will conduct residual analysis on this model.

**Figure 2**: Residual Analysis



Judging from figure 2, the assumption of equal variances is met since there are no unusual trends or patterns in the residuals vs predicted values plot. The other assumption of normality is met since there is mostly a linear trend along y=x in the normal probability plot and close to normal distribution in the histogram of residuals. There are also some moderate problems with outliers as seen in the leverage plot. The observed datapoints beyond a leverage of 0.05 have undue influence on their own predicted values. In addition, there is at least one datapoint in the Cook’s D plot indicating a y value that has extremely strong influence on the estimated beta coefficients.

**Figure 3**: Residuals by Each Independent Variables



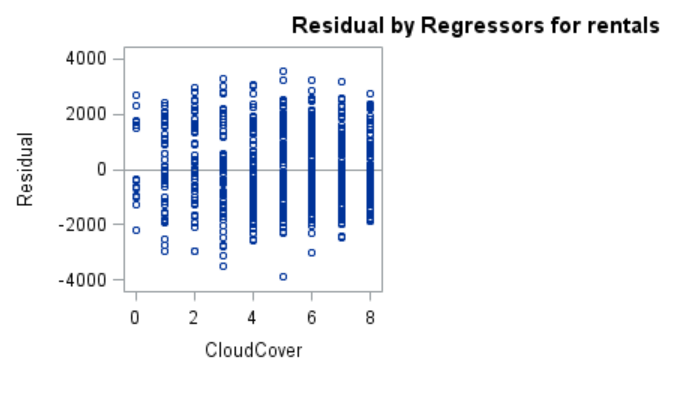


Figure 3 plots the residuals for rentals by each independent variable used in the model. Tempmax and tempmin appear to have constant variance, except for the lower right corner of both graphs appearing to have a “tail” appearance. However, it is not severe enough to be a cause for concern. Precipitation also has a slight issue where there is very large variance towards 0, and less variance towards larger precipitation values. It is interesting to note that it still had a significant p-value in the model, which may have been caused by the large sample size. All other graphs appear to follow equal variances.

**MODEL II: FIRST-ORDER MODEL WITH CATEGORICAL PREDICTORS**

For this model, we will now be using the dummy categorical variables which were coded and discussed in the “data collection and pre-processing” section, in order to further increase R2 value. A stepwise regression was utilized to screen for predictors which are useful in building the model. The process was run in SAS with an SLE of 0.15 and SLS 0.10. Results are summarized in the tables below.

**Table 6**: Summary of Stepwise Selection

| **Summary of Stepwise Selection** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Variable Entered** | **Variable Removed** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | tempmax |  | 1 | 0.4089 | 0.4089 | 287.534 | 504.36 | <.0001 |
| **2** | precip |  | 2 | 0.0720 | 0.4810 | 165.905 | 101.02 | <.0001 |
| **3** | ds\_fall |  | 3 | 0.0461 | 0.5271 | 88.7127 | 70.93 | <.0001 |
| **4** | pressure |  | 4 | 0.0119 | 0.5390 | 70.3619 | 18.67 | <.0001 |
| **5** | ds\_spring |  | 5 | 0.0115 | 0.5505 | 52.5568 | 18.61 | <.0001 |
| **6** | humidity |  | 6 | 0.0099 | 0.5604 | 37.6221 | 16.25 | <.0001 |
| **7** | ds\_summer |  | 7 | 0.0090 | 0.5693 | 24.2081 | 15.08 | 0.0001 |
| **8** | wind |  | 8 | 0.0042 | 0.5735 | 19.0751 | 7.03 | 0.0082 |
| **9** | tempmin |  | 9 | 0.0031 | 0.5766 | 15.7085 | 5.32 | 0.0213 |
| **10** | workday |  | 10 | 0.0020 | 0.5786 | 14.3181 | 3.37 | 0.0666 |
| **11** | visibility |  | 11 | 0.0015 | 0.5801 | 13.6868 | 2.63 | 0.1056 |
| **12** |  | visibility | 10 | 0.0015 | 0.5786 | 14.3181 | 2.63 | 0.1056 |

**Table 7**: Summary of Stepwise Selection

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 1266.26072 | **R-Square** | 0.5786 |
| **Dependent Mean** | 4504.34884 | **Adj R-Sq** | 0.5727 |
| **Coeff Var** | 28.11196 |  |  |

From tables 6 and 7, it appears the final model as an R2 of 0.579, adjusted R2 value of 0.573, and a C(p) value of 14.31. Adding the categorical variables has increase the R2 value slightly, and the C(p) value is somewhat close to 12 (p+1=11+1=12), indicating low bias and small mean square error. The model selected included 6 categorical variables and two of the three categorical variables.

**Table 8**: Parameter Estimates

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | 1 | -24233 | 8255.91012 | -2.94 | 0.0034 | 0 |
| **workday** | 1 | 186.50945 | 101.52565 | 1.84 | 0.0666 | 1.01571 |
| **tempmax** | 1 | 45.91071 | 8.94254 | 5.13 | <.0001 | 11.26753 |
| **tempmin** | 1 | 25.20805 | 10.99299 | 2.29 | 0.0221 | 14.41025 |
| **humidity** | 1 | -26.24067 | 4.80783 | -5.46 | <.0001 | 1.68601 |
| **pressure** | 1 | 849.32029 | 268.33777 | 3.17 | 0.0016 | 1.51975 |
| **precip** | 1 | -975.93993 | 160.60774 | -6.08 | <.0001 | 1.34213 |
| **wind** | 1 | -43.96485 | 16.81795 | -2.61 | 0.0091 | 1.39792 |
| **ds\_spring** | 1 | 932.20333 | 174.79094 | 5.33 | <.0001 | 2.62349 |
| **ds\_summer** | 1 | 568.98532 | 237.02788 | 2.40 | 0.0166 | 4.89320 |
| **ds\_fall** | 1 | 1417.55092 | 149.43885 | 9.49 | <.0001 | 1.87546 |

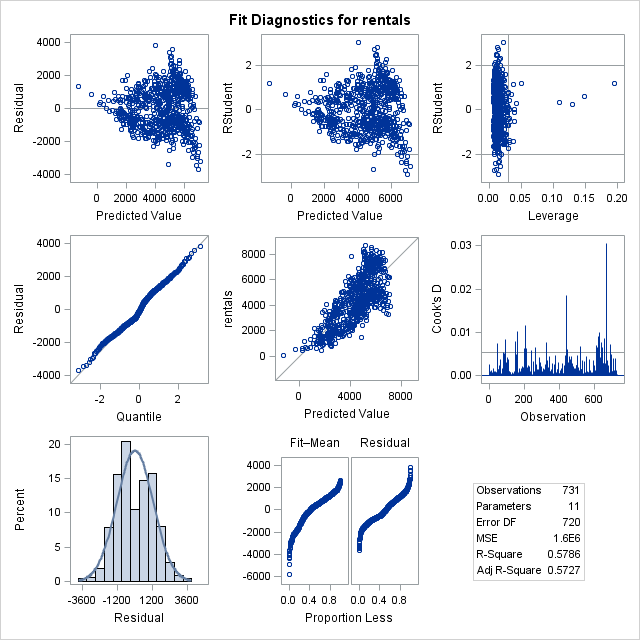
The proposed fitted model equation from the stepwise regression is:

Where

* = daily bike rental count
* X1 = workday
* X2 = tempmax
* X3 = tempmin
* X4 = humidity
* X5 = pressure
* X6 = precip
* X7 = wind
* X8 = ds\_spring
* X9 = ds\_summer
* X10 = ds\_fall

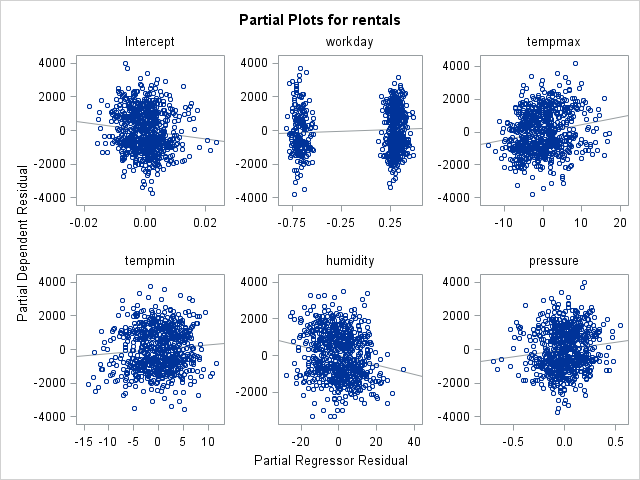
There still is the multicollinearity issue with tempmax and tempmin, but they have been previously addressed and the variables will be left as is.

**Figure 4**: Residual Analysis



Now we will analyze the residual plots for the model in figure 4. The assumption of equal variances is met since there are no major unusual trends or patterns in the residuals vs predicted values plot. The other assumption of normality is met since there is mostly a linear trend along y=x in the normal probability plot. There is a minor bimodal distribution in the residual histogram, but not enough to be a cause for concern. There are also some moderate problems with outliers as seen in the leverage plot. There are multiple observed datapoints beyond a leverage of 0.05 that have undue influence on their own predicted values. In addition, there are several datapoints in the Cook’s D plot indicating a y value that has strong influence on the estimated beta coefficients.

**Figure 5**: Partial Residual Plots Vs Predictors



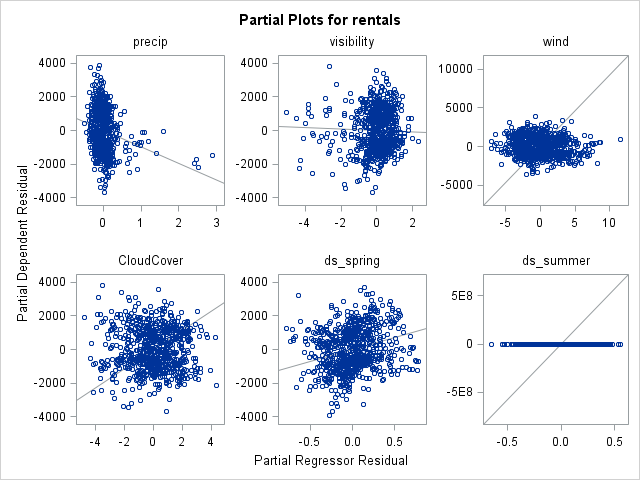


Figure 5 plots the partial residuals versus each predictor (ignore categorical predictors workday, ds\_spring, and ds\_summer). The plots measure the influence of each xk on the dependent variable after the effects of the other independent variables have been removed or accounted for. Since there does not appear to be any unusual patterns or trends, none of the variables have lack of fit problems.

**MODEL III: ALL-POSSIBLE-REGRESSION SELECTION METHOD**

Since there were no unusal trends seen in the residuals and partial plots for each independent variable (for models I and II), we will not be conducting any variable transformations such as x2, 1/x, etc. In the previous study (from STAT 3130 project), a two-way ANOVA was utilized to determine if there was an interaction between workday and weathercategory. No significant interaction was found, so there will not be justification for attempting interaction terms for such variables. In addition, it will be time-consuming to try interactions for every singly possible interaction term.

Instead, the last model for this analysis will be using an all-possible regression procedure to see if there will be any differences from the stepwise procedure. We will primarily focus on adjusted R2 and Cp  criterion to select the best model.

**Table 9**: Summary of All-Possible-Regression Selection Procedure

| **Number in Model** | **R-Square** | **Adjusted R-Square** | **C(p)** | **Variables in Model** |
| --- | --- | --- | --- | --- |
| **1** | 0.4089 | 0.4081 | 287.5342 | tempmax |
| **1** | 0.3558 | 0.3549 | 378.8027 | tempmin |
| **1** | 0.1201 | 0.1188 | 783.3721 | ds\_summer |
| **2** | 0.4810 | 0.4795 | 165.9046 | tempmax precip |
| **2** | 0.4605 | 0.4590 | 201.0135 | tempmax visibility |
| **2** | 0.4559 | 0.4544 | 208.8437 | tempmax pressure |
| **3** | 0.5271 | 0.5251 | 88.7127 | tempmax precip ds\_fall |
| **3** | 0.5065 | 0.5045 | 124.0538 | tempmax visibility ds\_fall |
| **3** | 0.5001 | 0.4981 | 134.9995 | tempmax pressure precip |
| **4** | 0.5390 | 0.5364 | 70.3619 | tempmax pressure precip ds\_fall |
| **4** | 0.5386 | 0.5361 | 70.8997 | tempmax humidity precip ds\_fall |
| **4** | 0.5371 | 0.5346 | 73.4734 | tempmax precip visibility ds\_fall |
| **5** | 0.5505 | 0.5474 | 52.5568 | tempmax pressure precip ds\_spring ds\_fall |
| **5** | 0.5500 | 0.5469 | 53.4274 | tempmax humidity precip ds\_spring ds\_fall |
| **5** | 0.5473 | 0.5441 | 58.1035 | tempmax humidity pressure precip ds\_fall |
| **6** | 0.5605 | 0.5569 | 37.3261 | tempmax humidity precip wind ds\_spring ds\_fall |
| **6** | 0.5604 | 0.5567 | 37.6221 | tempmax humidity pressure precip ds\_spring ds\_fall |
| **6** | 0.5596 | 0.5560 | 38.8613 | tempmax humidity precip ds\_spring ds\_summer ds\_fall |
| **7** | 0.5693 | 0.5652 | 24.2081 | tempmax humidity pressure precip ds\_spring ds\_summer ds\_fall |
| **7** | 0.5689 | 0.5648 | 24.8880 | tempmax humidity precip wind ds\_spring ds\_summer ds\_fall |
| **7** | 0.5686 | 0.5645 | 25.4074 | tempmax tempmin humidity pressure precip ds\_spring ds\_fall |
| **8** | 0.5735 | 0.5688 | 19.0751 | tempmax humidity pressure precip wind ds\_spring ds\_summer ds\_fall |
| **8** | 0.5734 | 0.5686 | 19.2808 | tempmax tempmin humidity pressure precip wind ds\_spring ds\_fall |
| **8** | 0.5723 | 0.5676 | 21.0455 | tempmax tempmin humidity pressure precip ds\_spring ds\_summer ds\_fall |
| **9** | 0.5766 | 0.5713 | 15.7085 | tempmax tempmin humidity pressure precip wind ds\_spring ds\_summer ds\_fall |
| **9** | 0.5755 | 0.5702 | 17.6006 | workday tempmax humidity pressure precip wind ds\_spring ds\_summer ds\_fall |
| **9** | 0.5754 | 0.5701 | 17.7971 | tempmax tempmin humidity pressure precip visibility wind ds\_spring ds\_fall |
| **10** | 0.5786 | 0.5727 | 14.3181 | workday tempmax tempmin humidity pressure precip wind ds\_spring ds\_summer ds\_fall |
| **10** | 0.5781 | 0.5722 | 15.2227 | tempmax tempmin humidity pressure precip visibility wind ds\_spring ds\_summer ds\_fall |
| **10** | 0.5774 | 0.5715 | 16.4357 | workday tempmax tempmin humidity pressure precip visibility wind ds\_spring ds\_fall |
| **11** | 0.5801 | 0.5737 | 13.6868 | workday tempmax tempmin humidity pressure precip visibility wind ds\_spring ds\_summer ds\_fall |
| **11** | 0.5797 | 0.5732 | 14.4776 | tempmax tempmin humidity pressure precip wind ds\_spring ds\_summer ds\_fall dwc\_1 dwc\_2 |
| **11** | 0.5789 | 0.5725 | 15.7184 | workday tempmax tempmin humidity pressure precip wind CloudCover ds\_spring ds\_summer ds\_fall |
| **12** | 0.5818 | 0.5748 | 12.7927 | workday tempmax tempmin humidity pressure precip wind ds\_spring ds\_summer ds\_fall dwc\_1 dwc\_2 |
| **12** | 0.5805 | 0.5735 | 15.0852 | workday tempmax tempmin humidity pressure precip visibility wind CloudCover ds\_spring ds\_summer ds\_fall |
| **12** | 0.5803 | 0.5733 | 15.3609 | workday tempmax tempmin humidity pressure precip visibility wind ds\_spring ds\_summer ds\_fall dwc\_2 |
| **13** | 0.5825 | 0.5749 | 13.6546 | workday tempmax tempmin humidity pressure precip visibility wind ds\_spring ds\_summer ds\_fall dwc\_1 dwc\_2 |
| **13** | 0.5821 | 0.5746 | 14.2421 | workday tempmax tempmin humidity pressure precip wind CloudCover ds\_spring ds\_summer ds\_fall dwc\_1 dwc\_2 |
| **13** | 0.5811 | 0.5735 | 16.0114 | workday tempmax tempmin humidity pressure precip visibility wind CloudCover ds\_spring ds\_summer ds\_fall dwc\_2 |
| **14** | 0.5829 | 0.5747 | 15.0000 | workday tempmax tempmin humidity pressure precip visibility wind CloudCover ds\_spring ds\_summer ds\_fall dwc\_1 dwc\_2 |

The best model was selected primary on looking for a high adjusted R2 value and low Cp value (or closest to p+1), which was the highlighted first model with 12 predictors. It had the second adjusted R2 value of 0.5748 and the lowest highest Cp value of 12.79. Interestingly, it is has the same subset of variables used in model II, with just the dummy weather categories (sunny, cloudy, etc…) added.

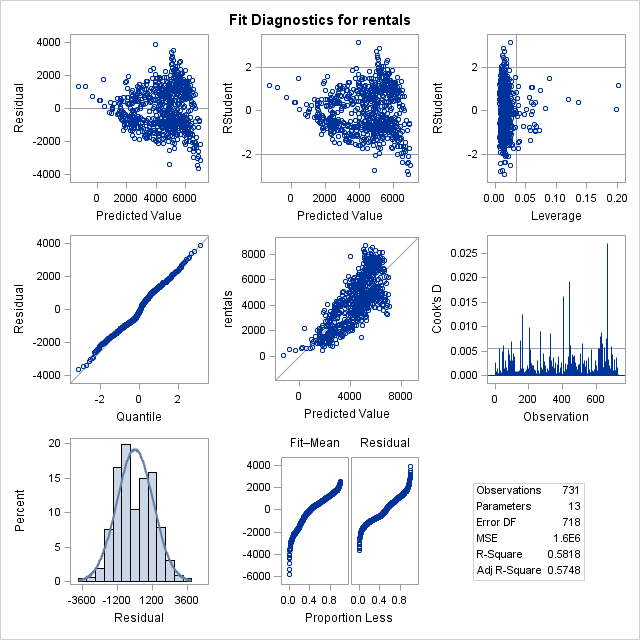
**Table 10**: Estimates for Selected Model from All-Possible-Regression Selection Procedure

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | 1 | -25223 | 8258.34458 | -3.05 | 0.0023 | 0 |
| **workday** | 1 | 194.75693 | 101.44161 | 1.92 | 0.0553 | 1.01900 |
| **tempmax** | 1 | 42.13465 | 9.41404 | 4.48 | <.0001 | 12.54820 |
| **tempmin** | 1 | 28.21290 | 11.27005 | 2.50 | 0.0125 | 15.21998 |
| **humidity** | 1 | -23.42417 | 5.40201 | -4.34 | <.0001 | 2.13892 |
| **pressure** | 1 | 850.99238 | 268.33381 | 3.17 | 0.0016 | 1.52715 |
| **precip** | 1 | -807.80353 | 175.98630 | -4.59 | <.0001 | 1.61936 |
| **wind** | 1 | -39.93466 | 16.86449 | -2.37 | 0.0181 | 1.41256 |
| **ds\_spring** | 1 | 944.05197 | 174.43756 | 5.41 | <.0001 | 2.62569 |
| **ds\_summer** | 1 | 578.89202 | 236.81137 | 2.44 | 0.0147 | 4.90819 |
| **ds\_fall** | 1 | 1439.10285 | 149.65385 | 9.62 | <.0001 | 1.89008 |
| **dwc\_1** | 1 | 846.69791 | 370.05600 | 2.29 | 0.0224 | 14.56831 |
| **dwc\_2** | 1 | 779.13746 | 334.23034 | 2.33 | 0.0200 | 11.44966 |

The proposed fitted model equation from the stepwise regression is:

Where

* = predicted daily bike rental count
* X1 = workday
* X2 = tempmax
* X3 = tempmin
* X4 = humidity
* X5 = pressure
* X6 = precip
* X7 = wind
* X8 = ds\_spring
* X9 = ds\_summer
* X10 = ds\_fall
* X11 = dwc\_1
* X12 = dwc\_2



The residual plots for model III are nearly identical to model II. The assumption of equal variances is met since there are no major unusual trends or patterns in the residuals vs predicted values plot. The other assumption of normality is met since there is mostly a linear trend along y=x in the normal probability plot. There is a minor bimodal distribution in the residual histogram, but not enough to be a cause for concern. There are also some moderate problems with outliers as seen in the leverage plot (slightly more than in model III). There are multiple observed datapoints beyond a leverage of 0.05 that have undue influence on their own predicted values. In addition, there are several datapoints in the Cook’s D plot indicating a y value that has strong influence on the estimated beta coefficients.

No residual versus individual predictors are provided since they are also nearly identical to model II.

**CONCLUSION**

The best models for predicting daily bike rental counts are in Models II and III, with both having an adjusted R2 value of approximately 0.57. Model III might be slightly better since its Cp value of 12.79 is lower than Model II (14.31) by a small margin. The difference in R2 values is almost neglible.

Regrettably, we were unable to find fit a model with an adjusted R2 beyond 0.65, but at least the R2 value of 0.57 from this analysis is a slight improvement over the previous study’s R2 value of 0.46. It seems as if addiding additional variables, especially categorical ones, did help with improving the model.

In reality, it is unlikely for there to be an R2 value beyond 0.65, since it is somewhat unrealistic to completely predict daily bike rental counts based on weather data alone. There may be other variables at play such as the economy. In addition, Capital BikeShare debuted in September 2010, and the dataset only covers daily counts from January 2011 to December 2012. It may have taken time for all DC area residents to fully take advantage of the system, biasing the results. A future recommendation is to pull more recent data directly from their website and not just rely on the dataset provided by UCI Irvine.

There are a multitude of suggestions for future analysis. One would be to figure out a faster way to analyze interactions, since it was not possible to try every single possible interaction variable. Another possible issue to look at was the weather type categories of sunny, cloudy, and rain/snow. It is likely these are poorly-defined and do not do a good job of distinguishing true differences in rental between all types of weather such as sunny, rain, thunderstorm, snow, fog, etc. Finally, the largest issue here is that probably is better to use a poisson regression since we are modeling daily counts here. Hopefully, there issues can be addressed in a future study and further improve the R2 value.

**APPENDIX I: R Code (Extract data using ‘weatherData’ package)**

#Install 'weatherData' package

install.packages('weatherData')

#Run 'weatherData' package

require('weatherData')

#Extract all weather data for DCA

DCAall <- getWeatherForDate("DCA", start\_date="2011-01-01",

                      end\_date = "2012-12-31",

                      opt\_detailed = FALSE,

                      opt\_all\_columns = TRUE)

#Extract weather data for DCA in 2011

#!Info: DCA is Ronald Reagan Washington National Airport

#!Info: 'custom\_columns' are selected variables

DCA11 <- getWeatherForDate("DCA", start\_date="2011-01-01",

                           end\_date = "2011-12-31",

                           opt\_custom\_columns=TRUE,

                           custom\_columns=c(2,4,9,12,15,18,20,21,22))

#Extract weather data for DCA in 2012

#!Note: It wasn't possible to do >1 year at a time. Reason unknown.

DCA12 <- getWeatherForDate("DCA", start\_date="2012-01-01",

                           end\_date = "2012-12-31",

                           opt\_custom\_columns=TRUE,

                           custom\_columns=c(2,4,9,12,15,18,20,21,22))

#Combine DCA11 and DCA12 datasets

DCA1112 <- rbind(DCA11,DCA12)

#Export 'DCA1112' dataset to CSV File

write.csv(DCA1112, file = "file:///C:/Users/Jeffrey/Documents/School/Fall 2016/STAT 4210 Applied Regression/Final Project/DCA1112.csv")

#List all weather events

table(DCA1112$Events)

**APPENDIX II: SAS Code**

libname project "C:\Users\Jeffrey\Documents\School\Fall 2016\STAT 4210 Applied Regression\Final Project";

\*DATA PREPARATION--------------------------------------------------;

\*Rename "date" variable in original 'bike' dataset;

**data** project.bike2;

set project.bike;

rename dteday=Date;

**run**;

\*Merge 'bike2' and 'dca1112' datasets;

**data** project.dcbike;

merge project.bike2 project.dca1112;

by Date;

**run**;

\*Print list of variables in 'dcbike' by "position" in dataset;

**proc** **contents** data=project.dcbike position;

**run**;

\*Make subset of 'dcbike' with selected variables;

**data** project.dcbike2 (keep=Date season workingday weathersit cnt

Min\_TemperatureF Max\_TemperatureF Mean\_Humidity

Mean\_Sea\_Level\_PressureIn Mean\_VisibilityMiles

Mean\_Wind\_SpeedMPH PrecipitationIn CloudCover Events);

set project.dcbike;

**run**;

\*rename variables;

**data** project.dcbike3;

set project.dcbike2 (rename=(cnt=rentals

workingday=workday

weathersit=weathercat

Max\_TemperatureF=tempmax

Min\_TemperatureF=tempmin

Mean\_Humidity=humidity

Mean\_Sea\_Level\_PressureIn=pressure

Mean\_VisibilityMiles=visibility

Mean\_Wind\_SpeedMPH=wind

PrecipitationIn=precipitation));

**run**;

\*Dummy variable coding for 'season', where base=winter;

**data** project.dcbike4;

set project.dcbike3;

if season=**2** then ds\_spring=**1**;

else ds\_spring=**0**;

if season=**3** then ds\_summer=**1**;

else ds\_summer=**0**;

if season=**4** then ds\_fall=**1**;

else ds\_fall=**0**;

**run**;

\*Dummy variable coding for 'weathercat', where base=3;

**data** project.dcbike5;

set project.dcbike4;

if weathercat=**1** then dwc\_1=**1**;

else dwc\_1=**0**;

if weathercat=**2** then dwc\_2=**1**;

else dwc\_2=**0**;

**run**;

\*Print list of variables in 'dcbike' by "position" in dataset;

**proc** **contents** data=project.dcbike5 position;

**run**;

\*Drop 'season' and 'events';

**data** project.dcbike6 (drop = season events);

set project.dcbike5;

**run**;

\*Recode 'T' (trace) precipitation to 0.00 inches;

**data** project.dcbike7;

set project.dcbike6;

if precipitation='T' then precipitation=**0**;

**run**;

\*Print list of variables in 'dcbike' by "position" in dataset

!For some odd reason, precipitation is a character variable and should be changed to numerical;

**proc** **contents** data=project.dcbike7 position;

**run**;

\*Change 'precipitation' from categorical to numerical variable.

!Modified variable is now called 'precip;

**data** project.dcbike8;

set project.dcbike7;

precip=input(precipitation,best32.);

**run**;

\*Check to make sure that 'precip' is a numerical variable;

**proc** **contents** data=project.dcbike8 position;

**run**;

\*DATA OVERVIEW-----------------------------------;

\*Frequency tables for all categorical variables;

ods rtf;

**proc** **freq** data=project.bike2;

table season;

table workday;

table weather;

**run**;

ods rtf close;

\*Create histogram/boxplot/PPlot to visualize distributions of each quantitative variable;

**proc** **univariate** data=project.dcbike8 plots;

var tempmax;

**run**;

**proc** **univariate** data=project.dcbike8 plots;

var tempmin;

**run**;

**proc** **univariate** data=project.dcbike8 plots;

var humidity;

**run**;

**proc** **univariate** data=project.dcbike8 plots;

var pressure;

**run**;

**proc** **univariate** data=project.dcbike8 plots;

var visibility;

**run**;

**proc** **univariate** data=project.dcbike8 plots;

var wind;

**run**;

**proc** **univariate** data=project.dcbike8 plots;

var precip;

**run**;

\*Summary statistics for all quantitative variables;

ods rtf;

**proc** **means** data=project.dcbike8 nmiss min q1 median mean q3 max stddev maxdec=**3**;

var tempmax;

var tempmin;

var humidity;

var pressure;

var visibility;

var wind;

var precip;

var cloudcover;

**run**;

**quit**;

ods rtf close;

\*Correlation matrix of all quantitative variables;

ods rtf;

**proc** **corr** data=project.dcbike8 plots(maxpoints=none)=matrix(nvar=**9**);

var rentals tempmax tempmin humidity pressure precip

visibility wind CloudCover;

**run**;

**quit**;

ods rtf close;

\*MODEL 1--------------------------------------------------;

\*Stepwise regression, first-order model;

ods rtf;

**proc** **reg** data=project.dcbike8;

model rentals = tempmax tempmin humidity pressure precip

visibility wind CloudCover / vif;

**run**;

model rentals = tempmax tempmin pressure precip

visibility wind CloudCover / vif influence;

**run**;

**quit**;

ods rtf close;

\*MODEL 2--------------------------------------------------;

\*Stepwise regression with categorical variables;

ods rtf;

**proc** **reg** data=project.dcbike8;

model rentals = workday tempmax tempmin humidity pressure precip

visibility wind CloudCover ds\_spring ds\_summer

ds\_fall dwc\_1 dwc\_2 / selection=stepwise sle=**0.15** sls=**0.10** partial vif;

**run**;

**quit**;

ods rtf close;

\*MODEL 3--------------------------------------------------;

\*All-possible model selection;

ods rtf;

**proc** **reg** data=project.dcbike8;

model rentals = workday tempmax tempmin humidity pressure precip

visibility wind CloudCover ds\_spring ds\_summer

ds\_fall dwc\_1 dwc\_2 / selection=rsquare adjrsq cp best=**3** vif;

**run**;

**quit**;

ods rtf close;

\*Best model from all-possible-regression procedure;

ods rtf;

**proc** **reg** data=project.dcbike8;

model rentals = workday tempmax tempmin humidity pressure precip wind

ds\_spring ds\_summer ds\_fall dwc\_1 dwc\_2/ vif;

**run**;

**quit**;

ods rtf close;