**Non-linear Models**

We observe predominantly non-linear trends of humidity, temp, windspeed variables with respect to the outcome variables (square root transformed registered and casual counts). This suggests accounting for non-linear terms of these numeric variables may be appropriate.

|  |  |
| --- | --- |
| Registered Rentals | Casual rentals |
|  |  |

We first explore the prediction for registered rentals. To baseline the results of more sophisticated non-linear models, we train a linear regression model on the training data set and evaluate the Mean Square Error (MSE) on a test data set using predictors season, holiday, workingday, weather, temp, humidity, windspeed, hr\_reg, day, year.

We achieve a test MSE of 15.31429 from a relatively simple linear model. Examining the associated summary statistics, we observe a significant yearly and hourly trend, weather, seasonal, and working-day effect, and linear effect with respect to temp and humidity on the conditional mean. Notably, in a model that includes all variables, the effects of windspeed and holiday are insignificant at the 5% level.

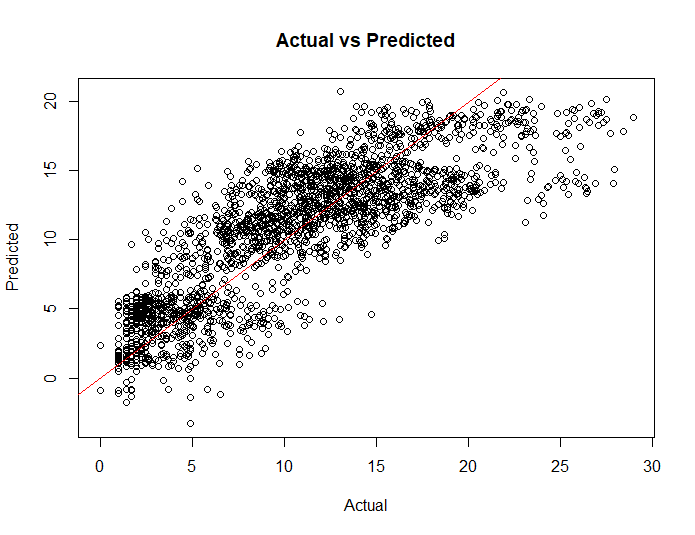
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **term** | **estimate** | **std.error** | **statistic** | **p.value** |
| (Intercept) | 0.591338 | 0.283079 | 2.088951 | 0.036741 |
| seasonSummer | 0.904341 | 0.155088 | 5.831154 | 5.70E-09 |
| seasonFall | 0.341996 | 0.197973 | 1.72749 | 0.084115 |
| seasonWinter | 2.509025 | 0.129602 | 19.35947 | 8.75E-82 |
| holidayHoliday | 0.493762 | 0.301106 | 1.639831 | 0.101076 |
| workingdayWorkday | 1.390502 | 0.158426 | 8.776995 | 2.00E-18 |
| weatherMist | 0.141193 | 0.102823 | 1.373166 | 0.169736 |
| weatherLight Snow and Rain | -1.69821 | 0.173193 | -9.8053 | 1.40E-22 |
| weatherHeavy Rain | -0.84648 | 3.941241 | -0.21477 | 0.829948 |
| temp | 0.204349 | 0.009295 | 21.98482 | 2.67E-104 |
| humidity | -0.04062 | 0.002745 | -14.7969 | 6.02E-49 |
| windspeed | -0.00096 | 0.00559 | -0.17256 | 0.863004 |
| hr\_reg | 1.847702 | 0.024063 | 76.7866 | 0 |
| dayMonday | -0.43278 | 0.161975 | -2.67189 | 0.007557 |
| dayTuesday | -0.28664 | 0.160661 | -1.78413 | 0.074437 |
| dayWednesday | -0.1244 | 0.159165 | -0.78155 | 0.4345 |
| dayThursday | -0.06361 | 0.158972 | -0.40013 | 0.689074 |
| dayFriday | NA | NA | NA | NA |
| daySaturday | 0.484683 | 0.156497 | 3.097083 | 0.001961 |
| year2012 | 2.604206 | 0.085136 | 30.58886 | 2.85E-195 |

We then include squared terms of humidity, temp, and windspeed to account for the non-linear effect of those variables with respect to registered rentals. The test MSE decreased slightly to 15.1957. Including cubic terms of humidity, temp, and windspeed further decreased MSE to 15.05779, while adding polynomial terms of these variables greater than 3 degrees increases the MSE. Getting rid of nonsignificant predictors of windspeed and holiday do not improve test MSE either.

Now we consider interaction effects. We expect that the effect of temp and humidity on registered may depend on season. In addition, there might be an interaction effect between temp and humidity. After calculating the test MSE adding those interaction terms, the model that includes cubic temp and season interaction effect further reduces MSE to 15.04829.

Lastly, we fit a GAM to predict counts using smoothing spline functions of temp, humidity, and windspeed. The GAM model significantly reduced test MSE to 12.47339. Using local regression for the three terms with a span of 0.7 gives the same test MSE.

Below is the test set actual vs predicted plot from GAM. This model is predicting well in the range of 0-20 square root of registered rentals, but underpredicts when the square root of registered rentals is above 20.



Similarly, we arrive at GAM with smoothing splines for temp, humidity, and windspeed as the best prediction model with a test MSE as 3.647885 for casual rentals using the above process. We observe again that the model underpredicts when the square root of registered rentals is above 14.

