Practice Data Analysis Project 2

Yutong Li

11.10.2015

Introduction

Bike sharing systems, as variants of traditional bicycle rentals, allow users to rent bikes from self-served bike stations and return them to any other stations within the service area. Such an efficient system not only provides people a flexible and inexpensive way of transportation, but also has the potential of reducing traffic congestions and air pollution. In this paper, we predict the daily level of bicycle rentals from environmental and seasonal variables. In particular, we hypothesise that having more registered users renting bikes on a given day predicts higher total bike rentals, that warmer temperature is associated with more bike rentals, and that higher percentage of humidity is associated with less rentals. We also hypothesise that the relationship between temperature and the number of bikes rented is the same in both 2011 and 2012, and that the relationship between humidity and daily rentals is different under different weather conditions.

Exploratory Data Analysis

The dataset contains data points taken on 731 days in 2011 and 2012. The response variable of interest is the count of total bike rentals each day, and there are 13 predictor variables: date, season, whether the year is 2011 or 2012, month, whether the data point is taken on a holiday, on which day of the week is the data point taken, weather conditions, normalised temperature, normalised feeling temperature, normalised humidity, normalised wind speed and number of daily bike rentals by registered users. It is worth to note that season, year, month, and weekday should all have approximately uniform distribution because the data points are taken on continuous dates. All continuous variables are summarised in Table 1.

	Count	Temperature	Feeling Temperature	Humidity	windspeed	Resigtered
Mean	4504	0.50	0.47	0.63	0.19	3656
Median	4548	0.50	0.49	0.63	0.18	3662
Variance	3752788	0.03	0.03	0.02	0.01	2434400
Range	[22, 8714]	[0.06, 0.86]	[0.08, 0.84]	[0.00, 0.97]	[0.02, 0.51]	[20,6946]

Table 1: Summary Statistics

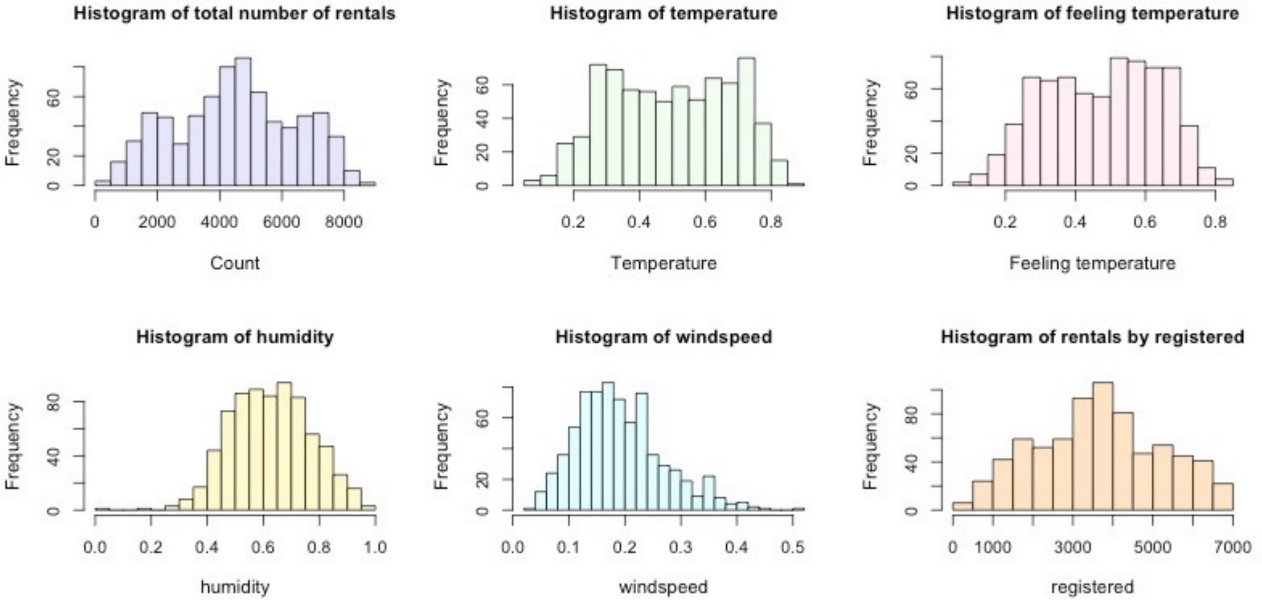

Figure 1: Histograms for total number of rentals, temperature, feeling temperature, humidity, windspeed and rentals by registered users.

Figure 1 illustrates the histograms of total count of daily rentals, temperature, feeling temperature, humidity, windspeed and number of rentals by registered users each day. The count of total bike rentals, the response variable, has range from 22 to 8714. Its distribution is symmetric with a mode around 4500 with no apparent outliers. The distribution of rentals made by registered users is also symmetric, and is roughly unimodal with a mode around 4000. The distribution of temperature is roughly symmetric and bimodal, with modes at 0.25 and 0.7. Feeling temperature has a similar bimodal distribution with modes at 0.25 and 0.5, but the distribution is less symmetric. Both of the temperature variables have no apparent outliers. The distribution of humidity is skewed left with a mode around 0.7, while the distribution of windspeed is skewed right with a mode around 0.15. Both of them have some outliers - there are a few days with relative humidity less than 0.1, and there are a few days with normalised wind speed greater than 0.45.

Table 2: Frequency tables for holiday, working day and weather

Non Holiday	Holiday	Non Working Days	Working Days	Clear/ partly	Mist	Light rain/ snow	Heavy rain /snow
710	21	231	500	463	247	21	0

Table 2 shows the frequency tables of the categorical variables - whether the day is a holiday, whether the day is a working day, and weather conditions. Only minority of days are holidays, and about 2/3 of the days are working days. There are only a few days with light rain or snow, and there is no observation of heavy rain or snow during the years.

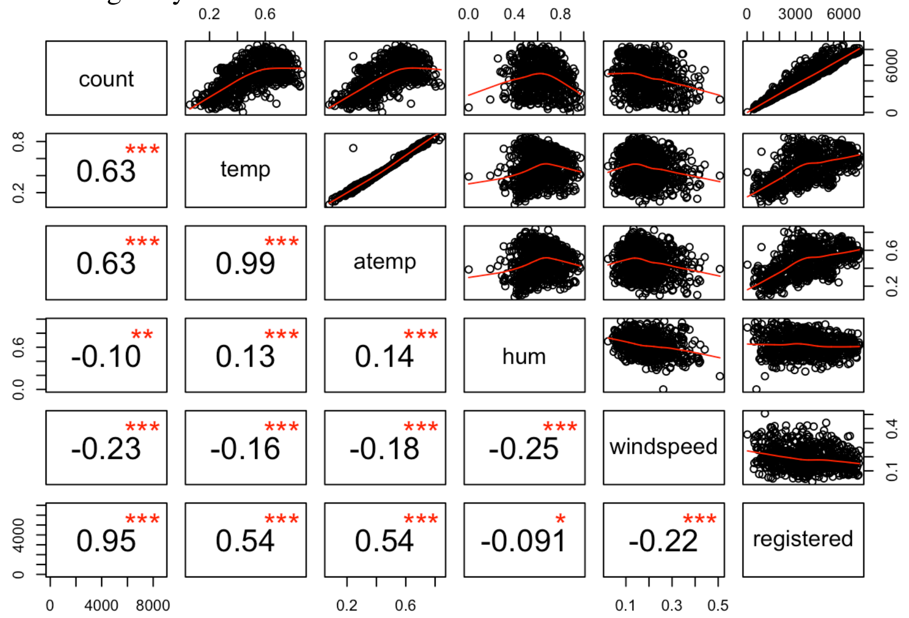

Figure 2: Scatter plots for continuous predictor variables versus count

Scatterplots of continuous variables are shown in Figure 2. The relationship between number of bike rentals by registered users and total number of bike rentals is linear, positive and strong, meaning that having more registered users renting bikes on a given day is associated with higher total bike rentals. Besides, on days with higher temperature or feeling temperature, the number of total rentals tend to be greater. The correlation between temperature and feeling temperature is strong, suggesting that these two variables are linear combinations of each other. There is also a moderate correlation between temperature and the number of rentals by registered users. The relationship between windspeed and total count of rentals is negative, indicating that on days with stronger wind, there tend to be less rentals. There is no apparent linear relationship between humidity and number of total rentals.

In Figure 3, box plots are used to show the bivariate relationship between categorical variables and total count of bike rentals each day. In general, daily level of bicycle rentals in 2012 is more than that in 2011, and there are more rentals on non-holidays. Whether the day is a working day makes little difference, and numbers of total rentals do not differ much on different days of the week. Total daily rentals in the second and third seasons of a year tend to be larger than that in the first and forth season. The relationship between month and count has similar shape. In fact, month and season are highly correlated with correlation about 0.8. Finally, daily level of bicycle rentals on clear and partly cloudy days is higher than that on mist days, and is even higher than that on lightly rainy or snow days. We cannot say anything about the daily rentals on days with heavy rain or snow because no observation was made under such weather condition.

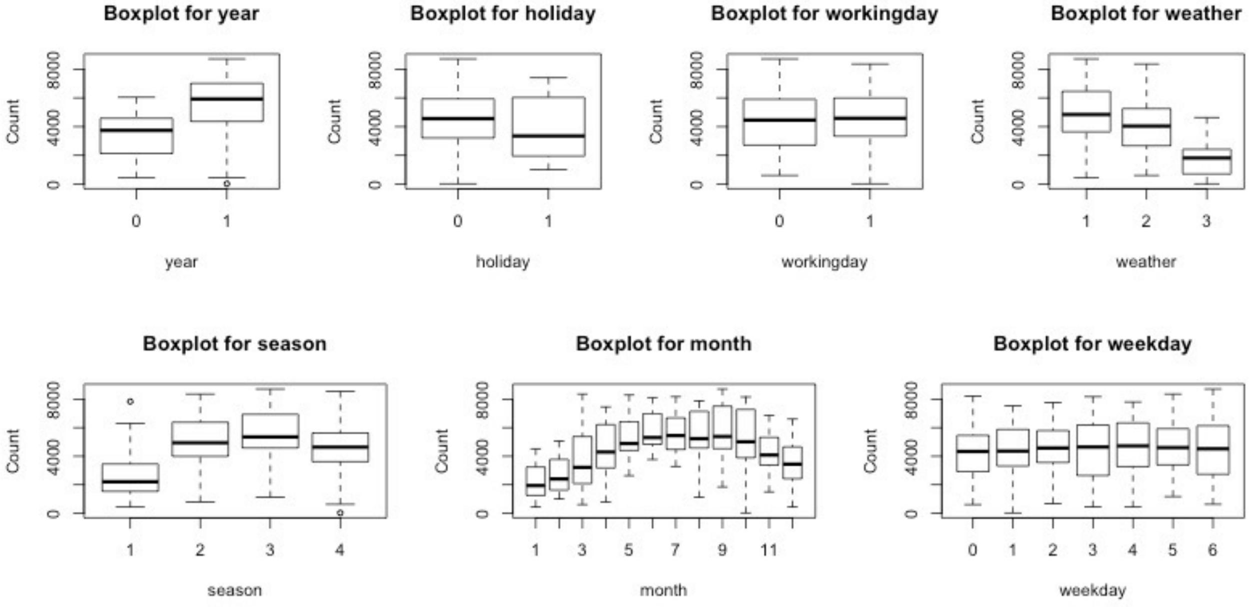

Figure 3: Box plots for categorical variables

Initial Modeling

To build the initial model, we first include year, weather conditions, temperature, relative humidity and number of bike rentals by registered users each day to answer the client's questions. We also include the interaction between temperature and year, and the interaction between humidity and weather for client's interest. Though our EDA suggests that feeling temperature is associated with daily level of bike rentals, we cannot include the variable in our model because it is highly correlated with temperature. Similarly, we can only choose one variable from month and season because they present similar information due to correlation. We choose to include season rather than month because daily rentals in months of each season do not present significant difference. It seems reasonable to include holiday and windspeed because they are both associated with total number of rentals. However, we left out the variable indicating whether the day is a working day and the variable indicating days of the week because daily level of bike rentals does not differ much as these variables differ. We do not include date in our model because the important pieces of information, year and season, are already included.

Exploring different ways to combine or split categories suggests that we should use two categories for the variable season: one includes season 1 (the reference group), and the other includes all other seasons. If we split the categories instead, the standard error of estimated coefficients will be increased, suggesting that the slopes are imprecisely estimated. For the similar reason, it is better to use two categories for weather: one includes clear or partly cloudy days (the reference group), and the other includes days with other weather conditions. For the rest of the categories, we will use days in year 2011 and days that are non-holidays as the reference groups.

Finally, we denormalise temperature before we include it in the model for the sake of easier interpretation.

Diagnostics and Model Selection

We then check the model assumptions by the diagnostic plots (Figure 4). The residuals in plot against fitted values shows a random pattern. The line 0 roughly goes through the centre of the data. However, the variance around count values of 5000 is slightly larger than the variance at other values. Thus, the assumption of constant variance is violated. Plots of residuals against numerical predictor variables have similar problems with constant variance (plots are omitted). The distributions of residuals against discrete predictors are mostly homogeneous except that residuals of data with light rain or snow weather is in general below 0. One possible explanation is that we only have 21 days with light rain or snow out of 731 days. The normal probability plot suggests that the distribution of residuals is tailed rather than approximately normal.

To improve the model, we first consider a BoxCox transformation on the response variable. Yet λ given by the BoxCox procedure is approximately 1, indicating that no transformation should be performed on the response variable. We also explore other possible transformations on the predictor variables based on the observations of our EDA, but there is no significant improvement either. Clearly, we need to consider other adjustments of our model.

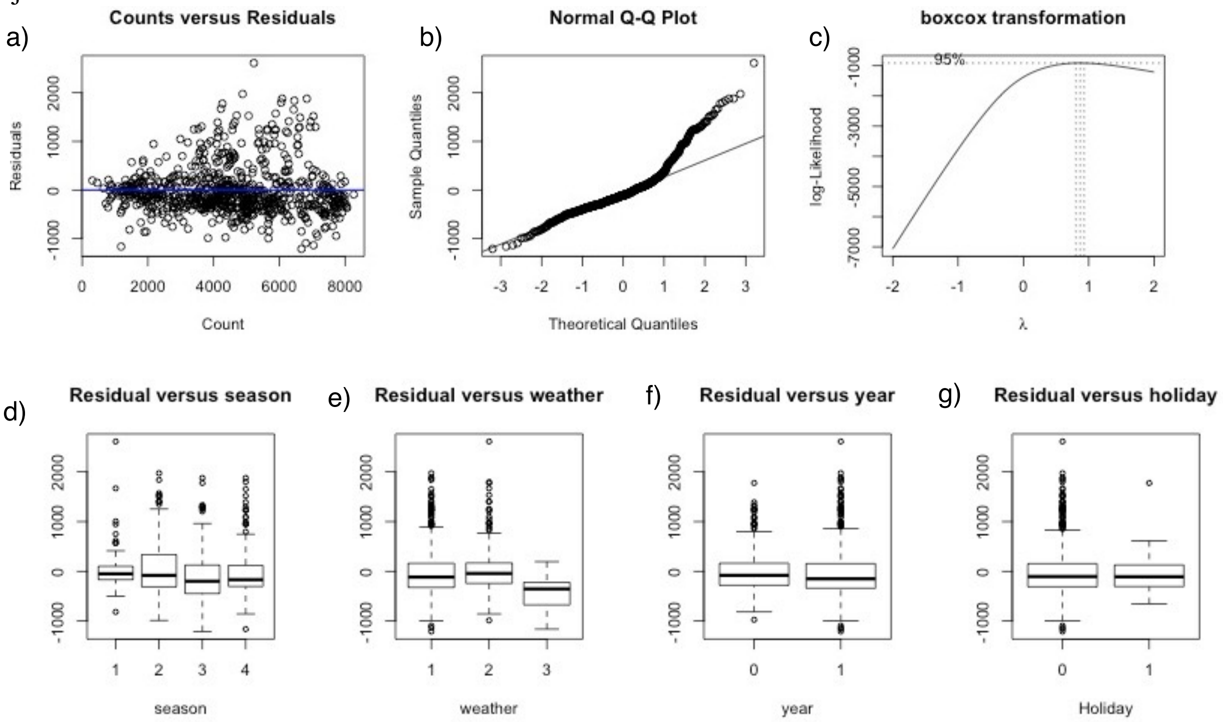

Figure 4: a) Residual against fitted valued b) Normal Probability Plot c) BoxCox procedure (c)-(g) Residual plots against categorical variables

We first turn our attention to determining if we can remove any influential outliers. In the data set, there is a point with 0 humidity, which means that no water vapour is present in the air. It is safe to remove this data because while humidity can approach 0 in the real world, this is unlikely to happen in Washington, D.C. To determine other outliers, Figure 5 is created to show the leverages, studendized residuals, and the Cook's distance statistics for each point. The clustering of extreme residuals appear to be around 500, and there is a less extreme clustering of residuals around 200. There is no extreme outliers greater than chi-square 0.1 in the Cook's distance plot. Hence, we only remove the outlier with humidity 0 from our model.

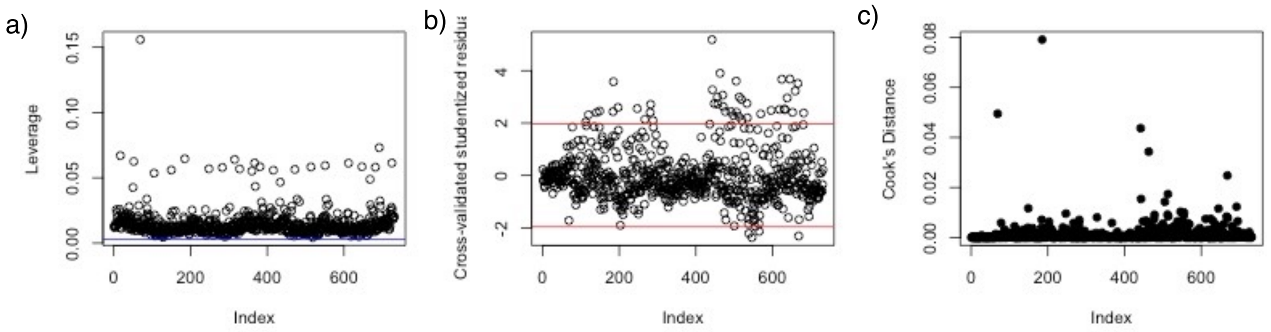

Figure 5: (a) Leverage (b) Studendized Residual with 95% t-distribution sampling intervals (c) Cooks Distance with $\chi^2(0.1)$

We then consider to exclude some interactions. Figure 6(a) shows the residual plot against temperature, with different colors representing different years. Since the colors heavily overlap and there is no apparent pattern to the colors, we do not have strong evidence of an interaction between temperature and year. Hence, we may try to turn off the interaction during model selection. Figure 6(b), on the other hand, shows the residual plot against humidity, with different colors representing different weather conditions. There is a pattern to the colors: relative humidity is higher on days with light rain or snow, and is lower on clear days. Therefore, we do have initial evidence of an interaction, and thus we will always include the interaction between humidity and weather in our model. We also try to include interaction between temperature and windspeed because based on our background knowledge, the relationship between total rental counts and temperature should differ when windspeed differ. This interaction should be independent of the interaction between temperature and year because intuitively, windspeed should not interact with year. A residual plot (omitted) against windspeed with different colors representing difference years will prove the independence.

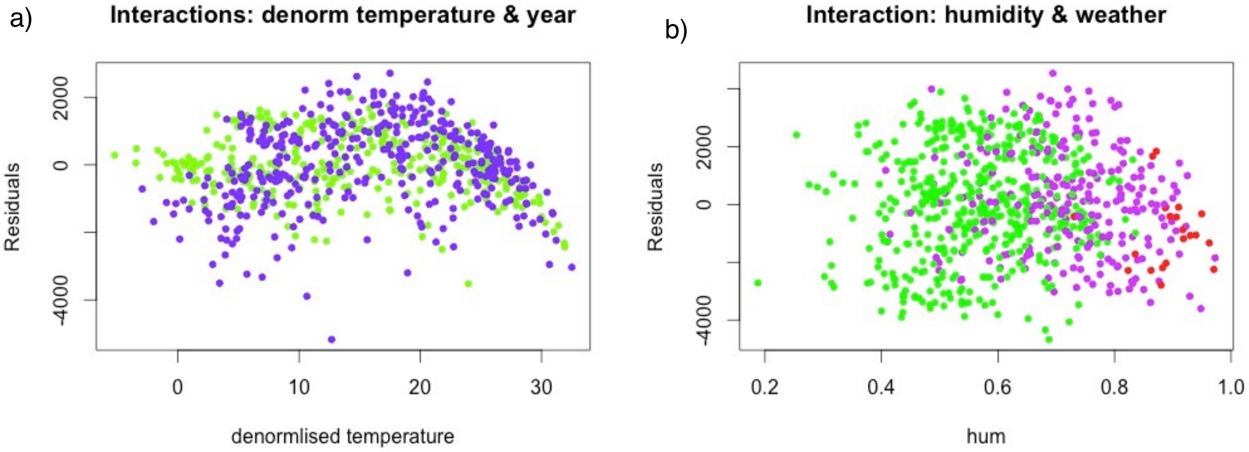

Figure 6: a) Residuals agains temperature, with the color varies with year.
b) Residuals against humidity, with color varies with weather

Finally, we may consider to add weekday or working day into our model because the residual plots suggests that the two variables actually matter. When including weekday in the model, we use Sunday (weekday == 0) as the reference group, and all other days of the week in another group.

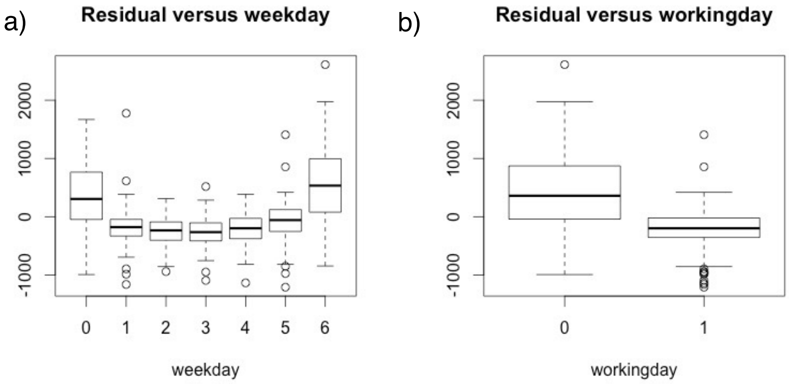

Figure 7: a) Residuals agains weekday (b) Residuals against working day

Our examination of data and the diagnostic plots suggest 4 possible modifications to our baseline model: add variable indicating whether the day is a working day, add variable indicating days of the week, turn off interaction between temperature and year, and turn on interaction between temperature and windspeed. We use leave-one-out cross-validation to select one model from 16 possible models. In particular, we run the cross-validation on half of the randomly selected data (about n/2 data points), and fit the model using (n/2 -1) data points, and choose the model that best predicts the missing data. Comparing to the baseline model, the new model includes both working day and weekday as predictor variables. Everything else stays the same.

The diagnostics is checked again for the final model (Figure 11). The residual plot against fitted values shows a random pattern of residuals. The line 0 roughly goes through the centre of data, and the variance is reasonably constant. The issue of constant variance in plots of residuals against numerical predictor variables are also greatly improved (plots are omitted). The distributions of residuals conditional on discrete predictors are mostly homogeneous except that residuals for holidays are in general below zero and is smaller than residuals for non-holidays. One possible explanation is that we have only a few data entries representing holidays (21 out of 731). While all other assumptions of multiple linear regression model are roughly met, the distribution of residuals is still far from Gaussian. The Gaussian noise assumption is violated.

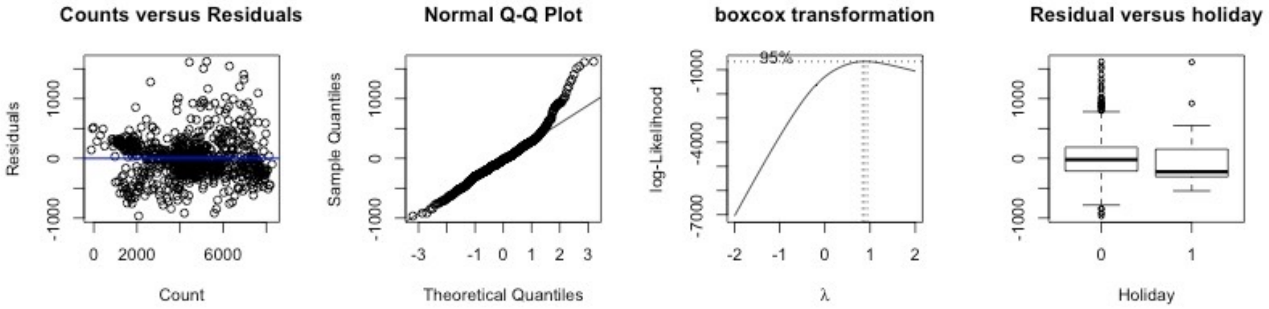

Figure 7: Main diagnostic plots

Final Model Inference and results

To predict the daily level of total bike rentals, the final model includes 10 predictor variables: temperature, relative humidity, windspeed, number of rentals by registered users, year (either 2011 or 2012), whether the day is a holiday, weather condition, season, whether the day is a working day and on which day of the week is the data taken. Besides, the interaction between temperature and year and the interaction between humidity and weather conditions are also included in the model. Estimates for the final model is contained in Table 3.

Table 3: Estimates and confidence interval for coefficients for the final model

	Estimated β	St. Error	95$ CI	p-value
Intercept	862.00	124.00	[617.00,1110.00]	9.82E-12
Temperature	25.30	2.67	[20.10, 30.60]	3.45E-20
Humidity	-381.00	168.00	[-710.00, -51.10]	2.36E-02
Windspeed	-578.00	195.00	[-961.00, -194.00]	3.23E-03
Rentals by registered users	1.12	0.02	[1.08, 1.16]	1.70E-259
Year	-78.60	64.10	[-204.00, 47.20]	2.2E-01
Holiday	-329.00	88.90	[-504.00, -155.00]	2.31E-04
Weather	99.40	176.0	[-246.00, 445.00]	5.72E-01
Season	124.00	47.20	[31.20, 217.00]	8.89E-03
Workingday	-992.00	42.90	[-1080.00, -908.00]	7.62E-89
Weekday	108.00	51.30	[7.67, 209.00]	3.50E-02
Temperature:year	10.60	3.21	[4.26, 16.90]	1.04E-03
Humidity:weather	-202.00	260	[-713.00, 308.00]	4.37E-01

As shown in the table, the estimated intercept is 862, which means that when all predictor variables are 0, the daily level of bike rentals is 862. To interpret the slope of temperature, we hold all other variables to be fixed. Then if a day is 1°C warmer than the other, on average, total bike rentals on that day will be 25 more than the total bike rentals on the other day. Similarly, if all other variables are fixed, total bike rentals on a day with relative humidity 1 percent higher will be 3.81 less; total bike rentals on a day with normalised windspeed one percent larger will be 5.78 less on average; on a day with 1 more rental made by registered users, daily level of bike rentals will be 1.12 higher; Days in 2012 have on average have 78.6 less bike rentals each day than days in 2011; Bike rentals on holidays are on average 329 less than bike rentals on non-holidays; days with clear or cloudy weather on average have 99.4 less bike rentals, and days in season 1 has on average 124 less bike rentals; Daily bike rentals on working days are 922.00 less on average than bike rentals on non-working days; Sundays, on average, have 108 more daily rentals than other days of the week. The interpretations of interactions are as follows: 1) Given two days which are otherwise equal, and both days are in year 2012, the day with a higher temperature is predicted to have 10.6 more total rentals on average for each degree Celsius; 2) Given two days which are otherwise equal, and both days do not have clear or cloudy weather, the day with higher humidity is predicted to have 202 less total rentals on average for each percentage difference of relative humidity.

The root mean square error of our final model is 368.4. Comparing to the root mean square of the initial model (517.4), this is a great improvement. We can say that our model predicts pretty well. However, we are not confident about whether the estimated coefficients are statistically significant because we cannot rely much on the confidence intervals and p-values without the Gaussian assumption. In order to answer the client's question about interactions more precisely, we generated Figure 7 to look for graphical evidence of interactions. Figure 7(a) shows that the daily level of bike rentals in 2012 is higher than the daily level of bike rentals in 2011. The two subgroup lines are fairly parallel, meaning that while the intercepts are different, the positive relationship between temperature and total daily rentals is similar in 2011 and 2012. There is no strong evidence of interaction between temperature and year. Figure 7(b) suggests an interaction between relative humidity and weather conditions. While higher relative humidity is associated with less total daily bike rentals in general, the rate of decreasing rentals is much slower when the days are clear and partly cloudy.

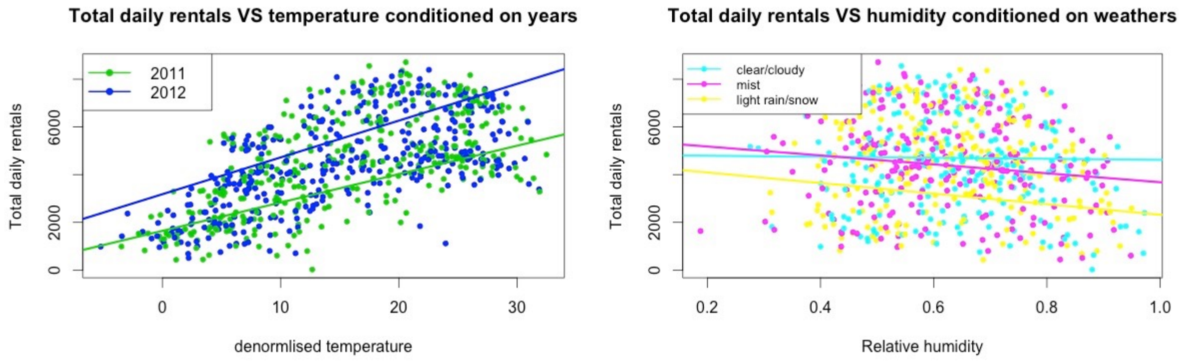

Figure 7: Rainbow plot looking for Graphical evidence of Interactions.

Discussion

As a conclusion, we found evidence that daily level of bike rentals is associated with temperature, relative humidity, windspeed, number of rentals by registered users, year, whether the day is a holiday, weather conditions, season, whether the day is a working day and on which day of the week is the data taken. In particular, the model suggests that more registered users renting bikes on a given day predict higher total bike rentals, and that higher relative humidity predict less total bike rentals. The relationship between temperature and the number of bikes rented is the same in both 2011 and 2012, and the relationship between humidity and rentals is different under different weather conditions. All hypothesises hold. Violating the Gaussian noise assumption has brought some disadvantages to our model, but overall the model gives good predictions with root mean square error about 368. One thing that does not make sense is that the daily level of total rentals in 2012 is indicated to be less than that in 2011 by our model. However, Figure 7(a) shows the opposite relationship. This is not contradictory because our model suggests an extremely broad confidence interval including 0 for the variable year, meaning that we do not actually have any predictive implications of the variable. Another strange thing is that the daily level of total bike rentals on holidays is on average less than that on non-holidays. But the daily level of total bike rentals on working days is also less than that on non-working days. One possible explanation is that we only have 21 holidays out of 731 days, and the incompletion of our data may affect the correctness of our model. Hence my suggestion for future researchers is to get data set with more complete information so that we can be more confidence about the model.