An Analysis of the Washington D.C./Arlington Bike Sharing System

Nicol Lo

Introduction

Bike sharing, a system that allow people to borrow a bike from point A and return it at point B, is becoming increasingly popular in major cities due to its low cost and flexibility. Since biking encourages a healthy lifestyle and is more environmentally friendly than other form of transportation, it is important to understand how and what factors affect bike share usage. We are interested in understanding bike share usage of registered users, the people who keeps the business model running. In particular, we would like to explore (1) the relationship between number of registered and casual users, (2) if the relationship between number of registered users and weather is dependent on whether it's a holiday, 3) whether "feel like" temperature is more important than actual temperature, wind speed and humidity to predict number of registered users, and 4) if the relationship between number of registered users and particular time of day depends on whether it's a weekend.

Exploratory Data Analysis

Bike sharing is unique in that we the duration of travel, departure and arrival locations are recorded. Here we have a sample of 910 hours over a randomly chosen day between 2011 and 2012 for the Washington D.C/Arlington, VA/MD area. Specifically, we have data on the number of registered bike users, the date, year, month, day and hour of observation, whether or not it was on a holiday or on a workday, the type of weather, actual temperature, "feels like" temperature, humidity, wind speed and number of casual bike users during the hour of observation.

The general characteristics of our sample are summarized in Table 1,2 and Figure 1. The number of registered users ranges from 0 to 775, with a right skewed distribution and a mean of 150. The distribution of number of casual users is even more right skewed, with majority of its data ranging between 0 and 50. Both normalized actual and "feels like" temperatures are approximately normally distributed around 0.5, although comparatively speaking "feels like" temperature have a slightly lower mean and standard deviation. Normalized humidity is also normally distributed with a mean of 0.629. Wind speed has a slight right skew with two primary modes of 0 and between 0.1 and 0.2. Generally speaking all the date-related variables (year, month, day and hour) are distributed evenly between its categories. 68.7% of our observations are recorded during workdays, while only 3.5% of them are recorded during holidays. Weather is divided into three categories, in order of decreasing pleasantness and also decreasing frequency.

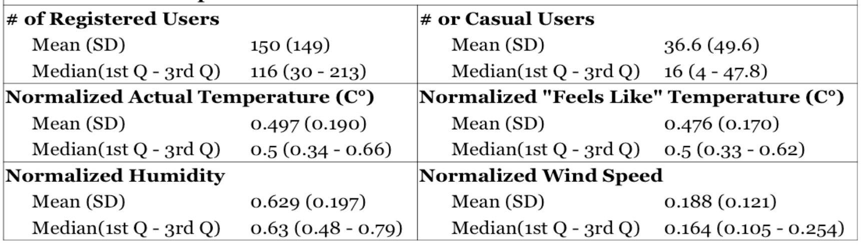


Table 1. General Population Statistics

# of Registered Users		**# or Casual Users**	
Mean (SD)	150 (149)	Mean (SD)	36.6 (49.6)
Median(1st Q - 3rd Q)	116 (30 - 213)	Median(1st Q - 3rd Q)	16 (4 - 47.8)
Normalized Actual Temperature (C°)		**Normalized "Feels Like" Temperature (C°)**	
Mean (SD)	0.497 (0.190)	Mean (SD)	0.476 (0.170)
Median(1st Q - 3rd Q)	0.5 (0.34 - 0.66)	Median(1st Q - 3rd Q)	0.5 (0.33 - 0.62)
Normalized Humidity		**Normalized Wind Speed**	
Mean (SD)	0.629 (0.197)	Mean (SD)	0.188 (0.121)
Median(1st Q - 3rd Q)	0.63 (0.48 - 0.79)	Median(1st Q - 3rd Q)	0.164 (0.105 - 0.254)

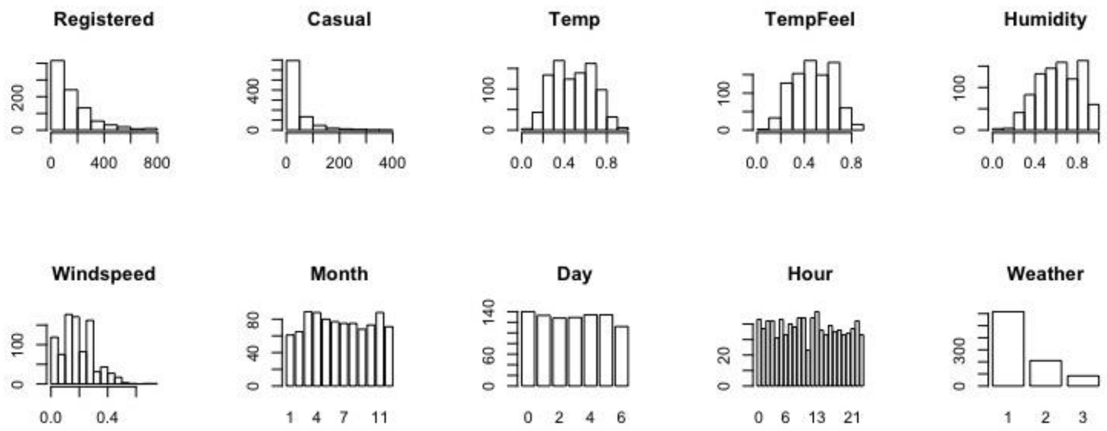


Figure 1. Distribution of Registered Bikers, Casual Bikers, Actual Temperature, "Feels like" Temperature, Humidity, Wind Speed, Month, Day, Hour and Weather

Table 2. General Population Statistics

Year		Workday?		Holiday?	
2011	447 (49.1%)	Yes (1)	626 (68.7%)	Yes (1)	32 (3.5%)
2012	465 (50.9%)	No (0)	284 (31.2%)	No (0)	878 (96.5%)

The bivariate relationships between registered bikers, casual bikers, and the continuous weather-related variables are shown in Figure 2 on the bottom left corner. Although wind speed does not seem to be a good predictor, the number of registered bikers is significantly positively correlated with actual and "feels like" temperature (r = 0.307 and 0.305), and negative correlated with humidity (r = -0.29). Compared to that of registered bikers, number of casual bikers shows a similar but even stronger relationship with the weather-related variables, suggesting that weather impacts casual ridership more than registered ridership. Unsurprisingly actual and "feels like" temperature are highly correlated, but at the moment we cannot tell whether one is a better predictor of registered bikes than another.

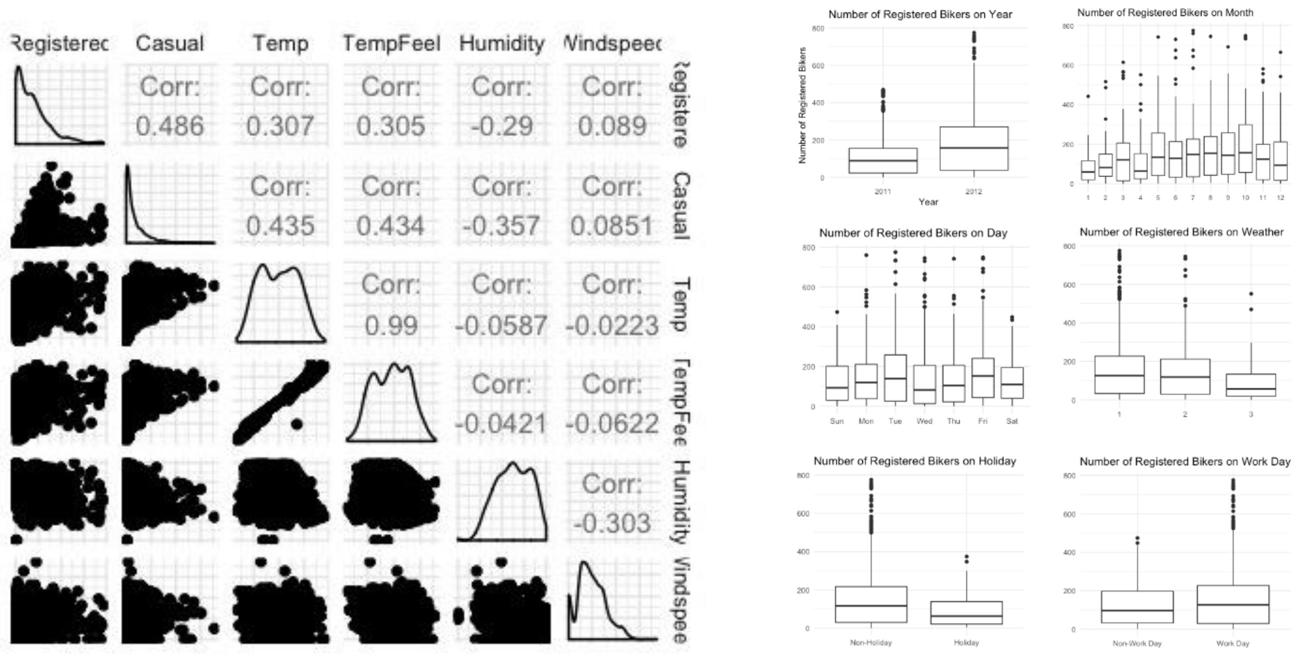


Figure 2. Left - Bivariate Relationships between Number of Registered and Casual bikers, Actual and "Feels like" Temperature, Humidity and Wind Speed. Right – The Number of Registered Bikers on Year, Month, Day of Week, Weather, Holiday and Work

Table 3 and the right of Figure 2 shows the conditional distribution between number of registered bikers on Year, Month, Day of Week, Weather, Holiday and Work Day. On average 187.57 registered riders are on the road every hour during 2012, substantially higher than 2011 which only averages 110.26 riders per hour. Very little people ride bikes during winter, but the number increases steadily after January (81.9) and peaks during October (206) before dropping sharply again in November. However, the standard deviations of each month are very close to the averages, it suggest that other factors may have contributed to the difference in ridership than which month it is. Generally speaking more people ride bikes during work days and non-holidays than during holidays or days off. Interestingly, it seems that around 20 more people ride during Tuesday and Friday than other weekdays (174.23 and 177.93 compared to 141.33 – 151.6). People are a little less likely to bike during fair weather (weather 2), but are considerably much less likely to bike during mediocre weather (weather 3). Every single category shown here has very high standard deviations for its mean, possibly implicating that there are interactions between the terms.

Table 3 - Summary of Conditional Distributions of Number of Registered Bikers: Mean (SD)

Month		**Year**		**Weather**	
Jan	81.9 (80.39)	2011	110.26 (104.47)	1 (good)	158.25 (152.35)
Feb	109.23 (106.43)	2012	187.57 (173.58)	2	149.7 (150.28)
Mar	142.53 (145.8)			3 (poor)	85.82 (95.92)
Apr	112.84 (118.73)	**Day**			
May	168.04 (146.63)	Sun	122.25 (107.63)	**Work Day**	
Jun	161.32 (162.22)	Mon	151.6 (146.32)	Yes	162.03 (163.26)
Jul	174.05 (174.07)	Tue	174.23 (174.97)	No	122.19 (105.96)
Aug	173.83 (148.64)	Wed	150.35 (177.36)		
Sep	176.16 (158.88)	Thu	141.33 (145.45)	**Holiday**	
Oct	206 (185.7)	Fri	177.93 (160.5)	Yes	100.38 (104.93)
Nov	146.86 (144.95)	Sat	128.35 (104.23)	No	151.39 (149.97)
Dec	134.03 (142.66)				

Exploring the relationship between registered and casual bikers a little more closely, we can see that there seem to be two distinct groups, which turn out to be determined by whether or not it's a workday. Generally speaking workdays have more registered bikers than non-workdays, while casual bikers are more active on non-workdays. Additionally, while the instances with less than 500 registered bikers seems to be mildly related to the presence of casual bikers on workdays, instances with over 500 registered bikers does not seem to be influenced by them. On non-workdays, there is a much stronger relationship between the two types of bikers.

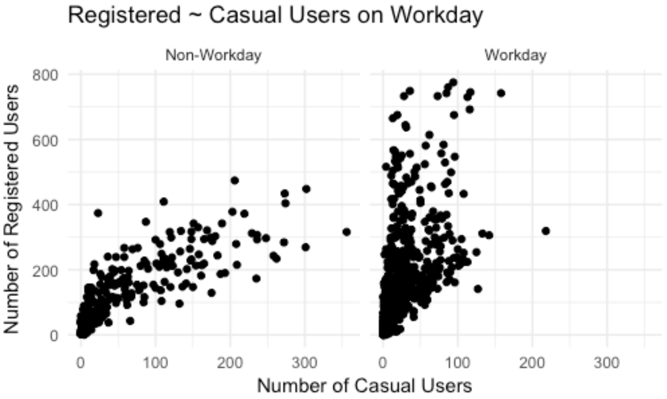


Figure 3. The Relationship Between the Number of Registered and Casual bikers Conditioned on Workday.

Modeling the Data

To explore if the relationship between registered users and time of day is conditional on whether it's a weekend, we created another variable called "weekend", which is distinct from the opposite of work day as work day excludes weekdays that are just so holidays.The conditional means and standard errors of the number of registered bikers on hour of day and weekend are shown on the left of Figure 4. During weekdays, the number of registered bikers peak around morning and evening rush hours from around 7 to 9am and around 4 to 7 pm.   During weekends, the number of registered bikes slowly increases from 7 am to around noon, and slowly decreases until 10pm.

In the previous section we have demonstrated how the variables workday and holiday are also very useful factors sat predicting registered bikers ridership. However it is not hard to see why weekend, workday and holiday are highly correlated, and the low t-values from all three t-test for each pair of these variables (0.30 - 1.25) confirms that these variables share too much information for one to be statistically different from another. We found that conditioning the number of registered bikers on hour of day and workdays yield more stable results than the results from conditioning on weekends (right of Figure 4), since it appears to have a clearer trend and has lower standard deviations within each category.   Based on our findings, we tentatively choose workday to include into our model.

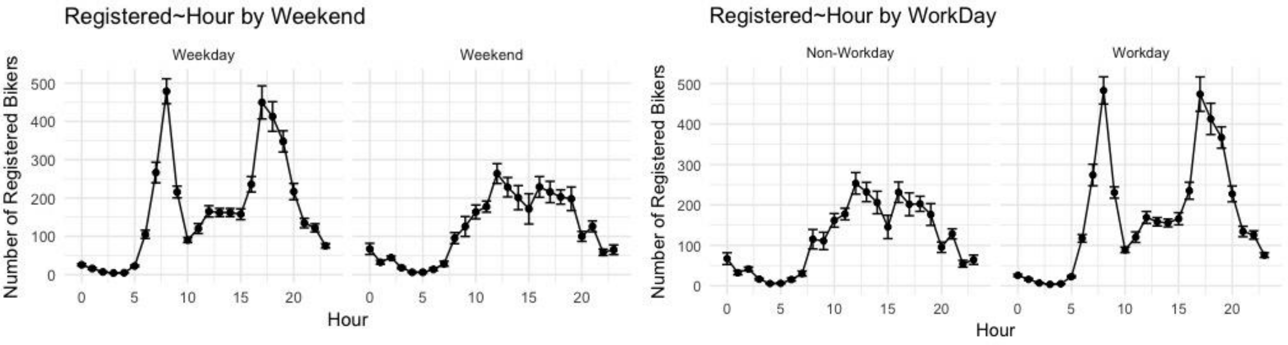


Figure 4. The Conditional Mean and Standard Error of Registered Bikers on Hour of Day and Weekend (Left) or Workday

We also note that from looking at the summary conditional distribution results, there does not seem to be a great enough distinction between weather 1 and 2 to separate them into two categories.The student t-test returns a p-value of 0.45, which means we should collapse weather 1 and 2 into one category because the two groups are not statistically different from each other.   Table 4 shows the initial attempts to model the interaction between hour and workday and weather. Considering the number of hours there are in a day, we will need to combine the hours into fewer categories to improve stability and generalizability. The decisions are primarily based on size of coefficients and standard error, p-values and interpretability. We successful collapsed 48 interaction terms between hour and work day into 14, increasing our F-statistic from 59.09 to 127 with an adjusted R-squared of 0.6431. We choose 11pm to 5 am as our reference point based on low standard error and the number of observations.   As expected, collapsing weather 1 and 2 improved the significance of our variables.

One of the things we are interested in knowing is if the relationship between number of registered users and weather is dependent on whether it's a holiday. In Figure 5, we can see that there are much more registered bikers during non-holidays than there are during holiday for all three types of weather, and for non-holidays we have more riders in better weather (left of each graph is weather 1 or 2, right is weather 3. However it must be noted that due to the small number of holiday and weather 3 observations we have in general, and there are two observations for in holiday-weather 3. This means that we cannot conclude anything from our analysis.

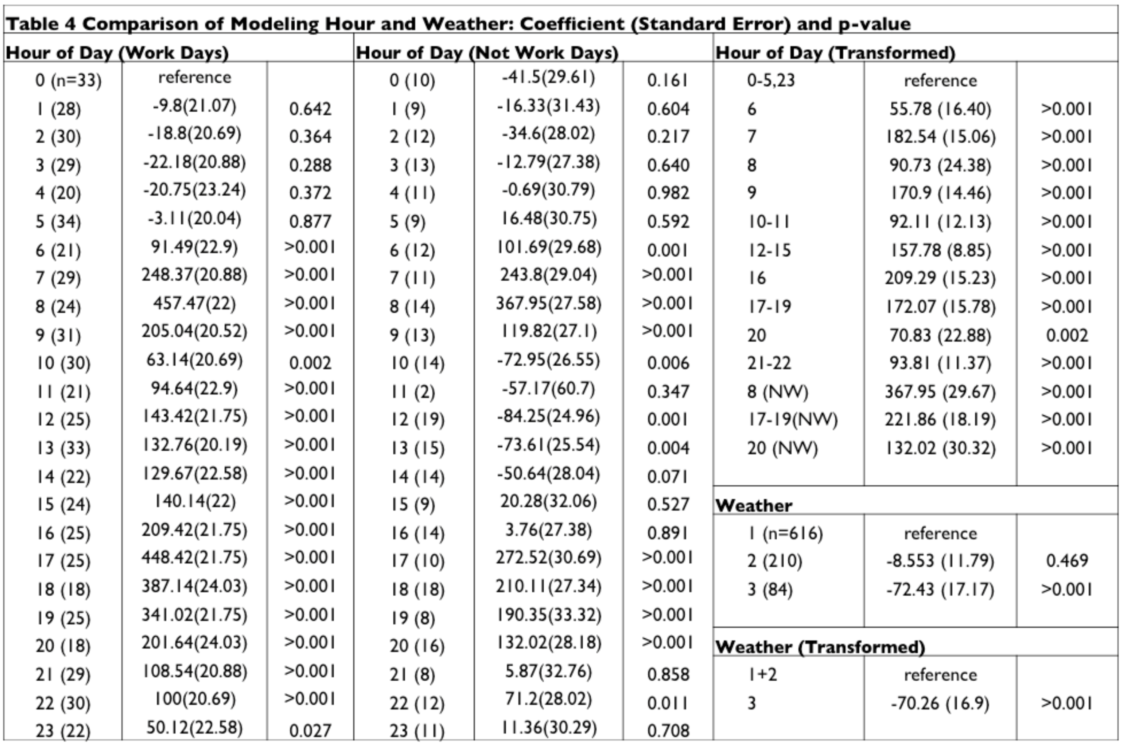


Table 4 Comparison of Modeling Hour and Weather: Coefficient (Standard Error) and p-value

Hour of Day (Work Days)			Hour of Day (Not Work Days)			Hour of Day (Transformed)		
0 (n=33)	reference		0 (10)	-41.5(29.61)	0.161	0-5,23	reference	
1 (28)	-9.8(21.07)	0.642	1 (9)	-16.33(31.43)	0.604	6	55.78 (16.40)	>0.001
2 (30)	-18.8(20.69)	0.364	2 (12)	-34.6(28.02)	0.217	7	182.54 (15.06)	>0.001
3 (29)	-22.18(20.88)	0.288	3 (13)	-12.79(27.38)	0.640	8	90.73 (24.38)	>0.001
4 (20)	-20.75(23.24)	0.372	4 (11)	-0.69(30.79)	0.982	9	170.9 (14.46)	>0.001
5 (34)	-3.11(20.04)	0.877	5 (9)	16.48(30.75)	0.592	10-11	92.11 (12.13)	>0.001
6 (21)	91.49(22.9)	>0.001	6 (12)	101.69(29.68)	0.001	12-15	157.78 (8.85)	>0.001
7 (29)	248.37(20.88)	>0.001	7 (11)	243.8(29.04)	>0.001	16	209.29 (15.23)	>0.001
8 (24)	457.47(22)	>0.001	8 (14)	367.95(27.58)	>0.001	17-19	172.07 (15.78)	>0.001
9 (31)	205.04(20.52)	>0.001	9 (13)	119.82(27.1)	>0.001	20	70.83 (22.88)	0.002
10 (30)	63.14(20.69)	0.002	10 (14)	-72.95(26.55)	0.006	21-22	93.81 (11.37)	>0.001
11 (21)	94.64(22.9)	>0.001	11 (2)	-57.17(60.7)	0.347	8 (NW)	367.95 (29.67)	>0.001
12 (25)	143.42(21.75)	>0.001	12 (19)	-84.25(24.96)	0.001	17-19(NW)	221.86 (18.19)	>0.001
13 (33)	132.76(20.19)	>0.001	13 (15)	-73.61(25.54)	0.004	20 (NW)	132.02 (30.32)	>0.001
14 (22)	129.67(22.58)	>0.001	14 (14)	-50.64(28.04)	0.071			
15 (24)	140.14(22)	>0.001	15 (9)	20.28(32.06)	0.527	**Weather**		
16 (25)	209.42(21.75)	>0.001	16 (14)	3.76(27.38)	0.891	1 (n=616)	reference	
17 (25)	448.42(21.75)	>0.001	17 (10)	272.52(30.69)	>0.001	2 (210)	-8.553 (11.79)	0.469
18 (18)	387.14(24.03)	>0.001	18 (18)	210.11(27.34)	>0.001	3 (84)	-72.43 (17.17)	>0.001
19 (25)	341.02(21.75)	>0.001	19 (8)	190.35(33.32)	>0.001			
20 (18)	201.64(24.03)	>0.001	20 (16)	132.02(28.18)	>0.001	**Weather (Transformed)**		
21 (29)	108.54(20.88)	>0.001	21 (8)	5.87(32.76)	0.858	1+2	reference	
22 (30)	100(20.69)	>0.001	22 (12)	71.2(28.02)	0.011	3	-70.26 (16.9)	>0.001
23 (22)	50.12(22.58)	0.027	23 (11)	11.36(30.29)	0.708			

We would also like to explore whether “feel like” temperature is more important than actual temperature, wind speed and humidity to predict number of registered bikers. We already know that wind speed is not a good predictor of registered bikers. Testing out models that includes different combinations of one of the two temperature variables, humidity and their interactions, we find out that best model is with “feel like” temperature and its interaction between humidity. The same model but with actual temperature instead is marginally worse at predicting number of registered users (“feels like” - F-statistic 101.9 and adjusted R-squared 0.1816; actual - 101.1 and 0.1806). Since we can only choose one of the two temperatures to include in our final model, we will choose “feels like” temperature.

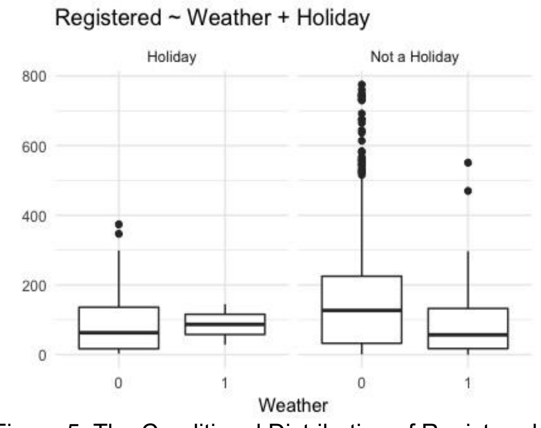
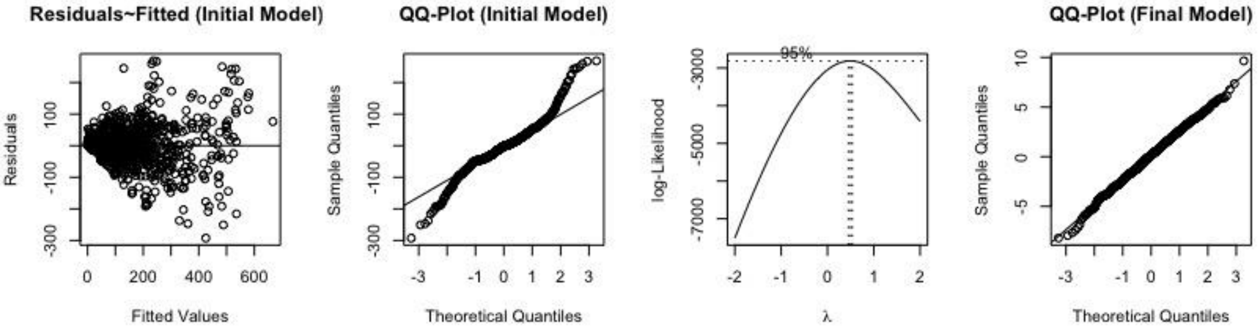


Figure 5. The Conditional Distribution of Registered Bikers on Type of Weather and Holiday

In our initial model, we choose to predict number of registered bikers using year, number of casual bikers, “feels like’ temperature, the hour the observation is recorded (divided into categories of 6am, 7am, 8am, 9am, 10 to 11am, 12 to 3pm, 4pm, 5 to 7pm, 8pm, 9 to 10 pm, and 11pm to 5am), the interaction term between “feels like” temperature and humidity, the interaction terms between workday and number of casual bikers, and the interaction terms between workday and 8am, workday and 12 to 3pm, and workday and 5 to 7 pm. We will first assess our model in the next section.

Diagnostics

Figure 5. Three from left - Diagnostics plots for Normality & Response Transformation. Right – Diagnostics Post-Transformation



We can go straight to check our normality assumptions using residual diagnostic plots shown in Figure 5. While the expectation of the errors is approximately 0, our graphs show non-constant variance that increases as fitted values increase, which indicates that a transformation of our response variable is necessary. Using the box-cox power transformation (lambda = 0.505, we use 0.5 for simplicity). Our post-transformation QQ-plot confirms that no further transformation is necessary

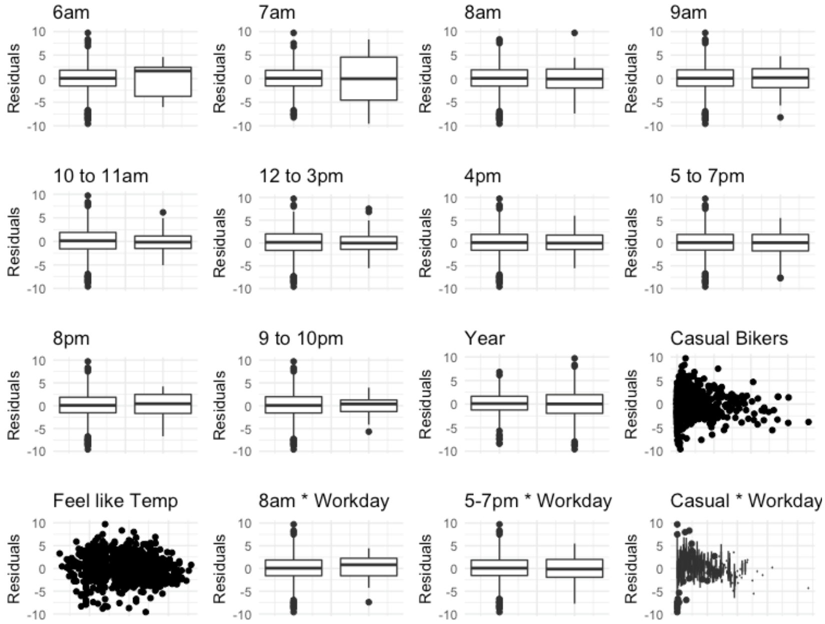


Figure 6 Diagnostic plots for our initial model

The residual diagnostic plots for all the variables and interactions in our initial model are shown in Figure 6. The right of each box plot for every hour variable shows the distribution of residuals of observations collected during that hour, and the left is of observations collected during any other hour. The left and right box on the plot for year shows the residuals for 2011 and 2012 respectively. The right box on the plot for workday interaction shows the residuals for workdays. We can see that the residuals for both casual bikers and the interaction between casual bikers and workday have non-constant variance that decreases as number of casual bikers increases. This indicates that we need transform the variables to a lower power – we will first try to square root them for simplicity. The residuals for both 6am and 7am seems a bit have more variation than other hours suggesting that there might be confounding variables, so we decide to explore the interaction between 6 am and workday and 7am and workday in our final model.

Model Inference & Results

The regression results of our transformed initial model and our final model after final adjusting are presented in Table 5. Since our transformed initial model is already good enough to explain over 95% of the variation in the data (adjusted R-squared = 0.9558), our decision to remove variables is based comparing the model and with the same model with the most insignificant variable removed until we maximize the F-statistic without compromising the strength of our model. The order of removal is also documented in the table. Overall we eliminated 6 am, 7 am, the interaction between "feels like" temperature and humidity and the interaction between 8 pm and workday, with a final F-statistic of 1825 but and adjusted R-squared of 0.9698.

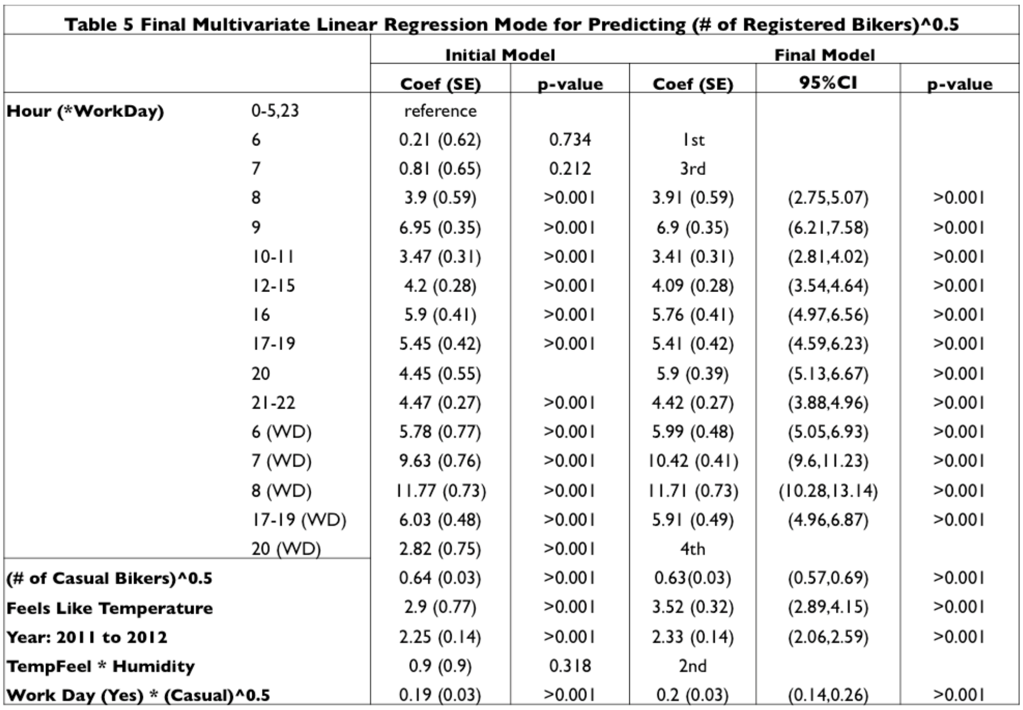


Table 5 Final Multivariate Linear Regression Mode for Predicting (# of Registered Bikers)^0.5

		Initial Model		Final Model		
		Coef (SE)	p-value	Coef (SE)	95%CI	p-value
Hour (*WorkDay)	0-5,23	reference				
	6	0.21 (0.62)	0.734	1st		
	7	0.81 (0.65)	0.212	3rd		
	8	3.9 (0.59)	>0.001	3.91 (0.59)	(2.75,5.07)	>0.001
	9	6.95 (0.35)	>0.001	6.9 (0.35)	(6.21,7.58)	>0.001
	10-11	3.47 (0.31)	>0.001	3.41 (0.31)	(2.81,4.02)	>0.001
	12-15	4.2 (0.28)	>0.001	4.09 (0.28)	(3.54,4.64)	>0.001
	16	5.9 (0.41)	>0.001	5.76 (0.41)	(4.97,6.56)	>0.001
	17-19	5.45 (0.42)	>0.001	5.41 (0.42)	(4.59,6.23)	>0.001
	20	4.45 (0.55)		5.9 (0.39)	(5.13,6.67)	>0.001
	21-22	4.47 (0.27)	>0.001	4.42 (0.27)	(3.88,4.96)	>0.001
	6 (WD)	5.78 (0.77)	>0.001	5.99 (0.48)	(5.05,6.93)	>0.001
	7 (WD)	9.63 (0.76)	>0.001	10.42 (0.41)	(9.6,11.23)	>0.001
	8 (WD)	11.77 (0.73)	>0.001	11.71 (0.73)	(10.28,13.14)	>0.001
	17-19 (WD)	6.03 (0.48)	>0.001	5.91 (0.49)	(4.96,6.87)	>0.001
	20 (WD)	2.82 (0.75)	>0.001	4th		
(# of Casual Bikers)^0.5		0.64 (0.03)	>0.001	0.63(0.03)	(0.57,0.69)	>0.001
Feels Like Temperature		2.9 (0.77)	>0.001	3.52 (0.32)	(2.89,4.15)	>0.001
Year: 2011 to 2012		2.25 (0.14)	>0.001	2.33 (0.14)	(2.06,2.59)	>0.001
TempFeel * Humidity		0.9 (0.9)	0.318	2nd		
Work Day (Yes) * (Casual)^0.5		0.19 (0.03)	>0.001	0.2 (0.03)	(0.14,0.26)	>0.001

The squared root transformation of our response variables tells us that the effect of these factors decreases with increased registered bikers. While all variables in the model significantly contribute to our model, morning and evening hours and their interactions with work day remain to be most important predictors of the variation between registered bikers in Washington D.C./Arlington area. Specifically, compared to non-workdays, 7 am and 8 am on workdays are associated with 10.42 and 11.71 increase in squared root of registered bikers respectively, and 5 to 7pm are associated with a slightly 5.91 increase in squared root of registered bikers.

With regards to the research hypotheses we have several findings. We concluded that in contrast to our initial prediction, an increase in casual bikers are associated with a general increase in registered bikers, but the effect is reversed during the day of work days. Due to the small sample size we have of holiday observations, it is not certain whether the relationship between number of registered users and weather is dependent on whether it's a holiday, but considering the strong correlation between work day and holiday and that ridership during non-work days are more weather-dependent, it is likely registered ridership is similarly more weather-dependent during holidays. "Feels like" temperature is a much better predictor of registered bikers than wind speed and humidity, but performs only marginally better than actual temperature in our sample. So although we choose "feels like" over actual temperature to include in our model, we cannot say whether one is inherently better than the other without more data.

As expected, the relationship between registered users and hour of day does greatly depend on whether it's a weekend. While more people ride bike in general and during mornings and evenings during weekdays, comparatively more registered bikers bike during the afternoon on weekends. However, since workday and weekday status share a lot of information and the workday/non-workday division is more stable, we did not incorporate weekend into our final model.

Discussion

Surprisingly, the number of registered bikers can be predicted with very few variables – hour and year of observation, whether it's a workday, number of casual bikers and "feels like" temperature. We see that there are significantly more registered riders in 2012 than 2012, but this is most likely due to the increase is popularity of bike shares in the area and is not likely a factor that affects ridership for bike share system in other areas as well, so it might be necessary to modify the model when analyzing bike share data from other cities. In our sample of 910 hours, we have too little data of type 3 and 4 weather to conclude any relationship between registered bikers and type of weather, and similar too little data collected on holidays to conclude the association between registered bikers and holidays. Therefore although we already have a very stable model with our current sample, a bigger sample may be helpful to discover more factors than influences bike share ridership.