Building a Linear Regression Model to Predict Total Daily Rental Bike Demands

School of Global Public Health, New York University, New York, NY, USA

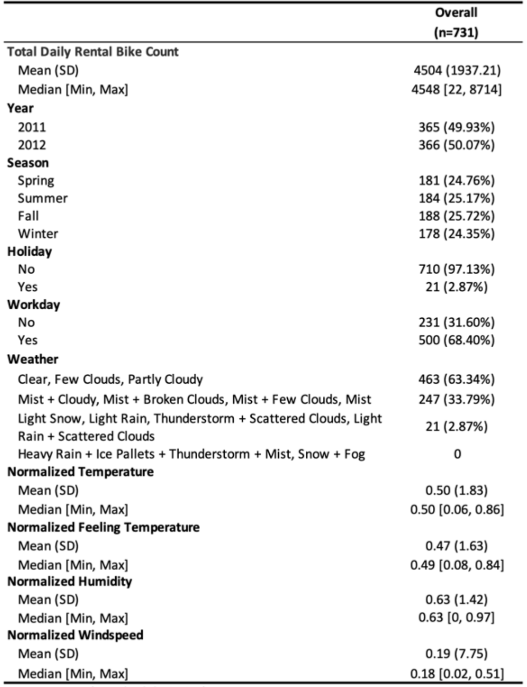
Spring 2024, GPH-GU 2353 Regression I: Linear Regression and Modeling

Helen Liang, Ivy Zhao, Xiaotong Zhao

Introduction

This project aims to create a linear regression model to predict the bike-sharing daily demand using the Bike Sharing Dataset. The dataset (Source: UCI Machine Learning Repository) includes daily rental bike counts from the Capital bike-share system in Washington DC from 2011 to 2012. This dataset contains 731 observations and 16 variables. We excluded time variables (e.g., date, month, etc.) and focused on using 9 of those 16 variables for our study (Table 1). The main outcome of this study, daily rental bike counts, is a continuous variable and is the count of total rental bikes including casual and registered. Other continuous variables include normalized temperature, normalized feeling temperature, normalized humidity, and normalized wind speed. Categorical variables include year (2011 and 2012), season (winter, spring, summer, and fall), holiday (no or yes), workday (no or yes), and weather conditions. Weather conditions were ordinal and categorized as clear/few clouds/partly cloudy, mist/cloudy/broken clouds, and light snow/rain.

Table 1. Summary of the Bike Sharing Dataset (n=731), 2011-2012. Mean (SD), median (min, max) are reported for continuous variables. Frequencies (%) are reported for categorical variables.



	Overall (n=731)
Total Daily Rental Bike Count	
Mean (SD)	4504 (1937.21)
Median [Min, Max]	4548 [22, 8714]
Year	
2011	365 (49.93%)
2012	366 (50.07%)
Season	
Spring	181 (24.76%)
Summer	184 (25.17%)
Fall	188 (25.72%)
Winter	178 (24.35%)
Holiday	
No	710 (97.13%)
Yes	21 (2.87%)
Workday	
No	231 (31.60%)
Yes	500 (68.40%)
Weather	
Clear, Few Clouds, Partly Cloudy	463 (63.34%)
Mist + Cloudy, Mist + Broken Clouds, Mist + Few Clouds, Mist	247 (33.79%)
Light Snow, Light Rain, Thunderstorm + Scattered Clouds, Light Rain + Scattered Clouds	21 (2.87%)
Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog	0
Normalized Temperature	
Mean (SD)	0.50 (1.83)
Median [Min, Max]	0.50 [0.06, 0.86]
Normalized Feeling Temperature	
Mean (SD)	0.47 (1.63)
Median [Min, Max]	0.49 [0.08, 0.84]
Normalized Humidity	
Mean (SD)	0.63 (1.42)
Median [Min, Max]	0.63 [0, 0.97]
Normalized Windspeed	
Mean (SD)	0.19 (7.75)
Median [Min, Max]	0.18 [0.02, 0.51]

Acryonym: SD (standard deviation)

Methods

We conducted univariable analysis to observe the distribution of main outcomes of interest, daily rental bike counts and 8 predictors (Figure 1) in the Bike Sharing dataset. Density plots were used for continuous variables and bar charts are used to for categorical variables for visualization. Then we conducted bivariable analysis to help understanding the relationships between variables and identify patterns (Figures 2 and 3). Scatter plots are used to visualize the results of bivariable analysis.

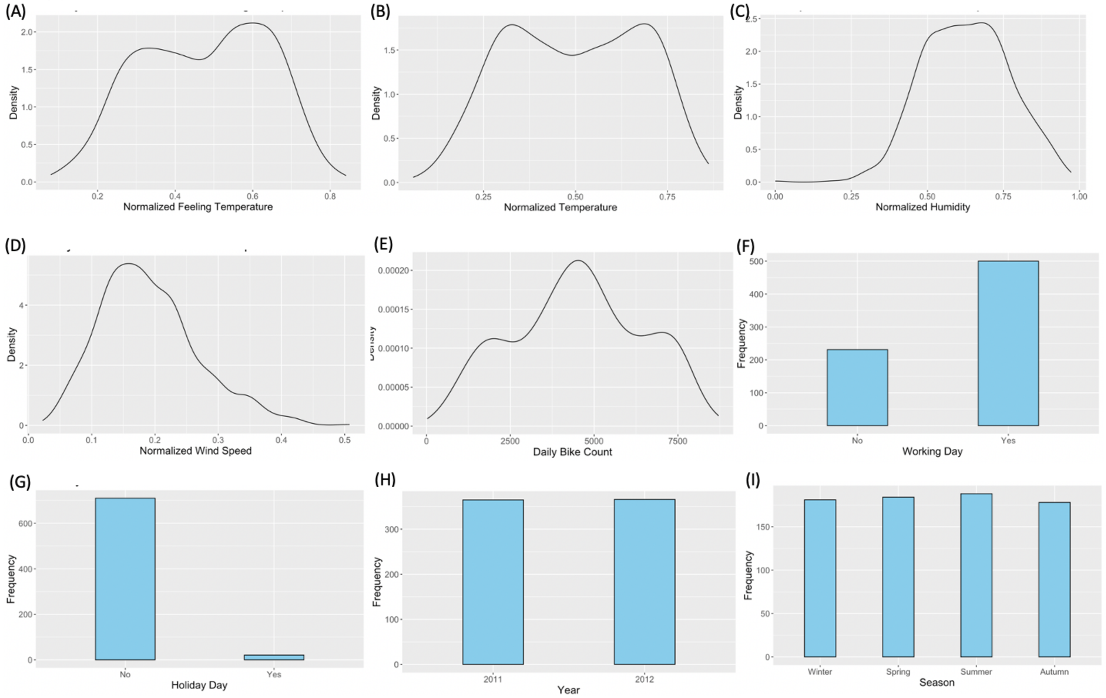
We first used all 9 predictors to build a linear regression model to predict daily rental bike counts. The Adjusted R-squared values were used for model comparison. We checked assumption of linear regression model using diagnostic plots including standardized residuals against the fitted values, Quantile-Quantile plots, histogram of standardized residuals, Scale-Location plot, and Residuals vs. Leverage plots (Figure 4). We calculated leverage, Cook's distance and checked outlier using Bonferroni Criterion, which was visualized using half-norm plots. Influential outliers were removed for model improvement. We calculated variance inflation factors (VIF) to test the existence of multicollinearity issue (Tables 2 and 3) and predictor was removed to solve multicollinearity issue.

For further model improvement, Akaike Information Criterion (AIC)-based model selection was conducted to select the best model (Figure 5). We checked for significant interaction terms using ANOVA out of 28 potential interaction terms. We utilized the Box-Cox method to transform the model and found the λ that maximized the log-likelihood. Finally, we used the plot of predicted daily bike count values using our final model vs. actual values (Figure 7), to visualize the performance of our final model.

Results

We first conducted univariable analysis (Figure 1). The main outcome, bike counts, follows a normal distribution centered around the median. Both temperature and feeling temperature show a bimodal distribution, indicating that two specific temperatures are most common. The normalized windspeed density plot shows a right skew, with a peak at approximately 0.1, highlighting that lower windspeeds are more frequent. Conversely, the normalized humidity plot, peaking once and skewed to the left, suggests that lower humidity levels are rarer than higher ones within the dataset. For categorical variables, there are very few observations that are holidays compared to non-holidays. Working days are about twice the number and non-working days. For year and season, observations are distributed evenly in each category.

Figure 1. Univariable Analysis. A-E are density plots for continuous variables and F-I are bar charts for categorical plots. A, Normalized Feeling Temperature; B, Normalized Temperature; C, Normalized Humidity; D, Normalized Wind Speed; E, Daily bike count; F, working day; G, Holiday; H, Year; I, Season.



We then conducted bivariable analysis on each predictor (Figure 2 and Figure 3). For continuous variables (Figure 2), we found that both temperature and feeling temperature are positively correlated with daily counts, and wind speed has an inverse relationship. Humidity exhibits a non-linear relationship, with daily counts peaking at moderate humidity levels and potentially decreasing at very low or high levels. For categorical variables (Figure 3), scatterplots revealed distinct clusters for daily counts by year, indicating variability within each year. Seasonal distribution shows that spring and summer have the highest rental bike demand. Although there's a wide spectrum of counts for holidays and working days, higher counts tend to occur on working days.

Figure 2. Bivariable plots for continuous variables. A) Normalized Temperature. B) Normalized Feeling Temperature. C) Normalized Humidity. D) Normalized Wind Speed.

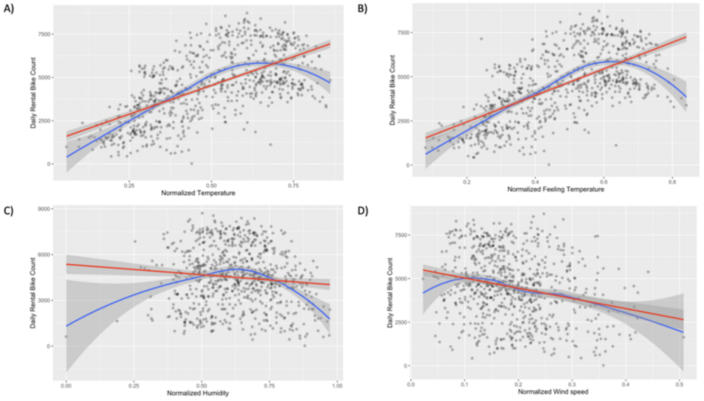
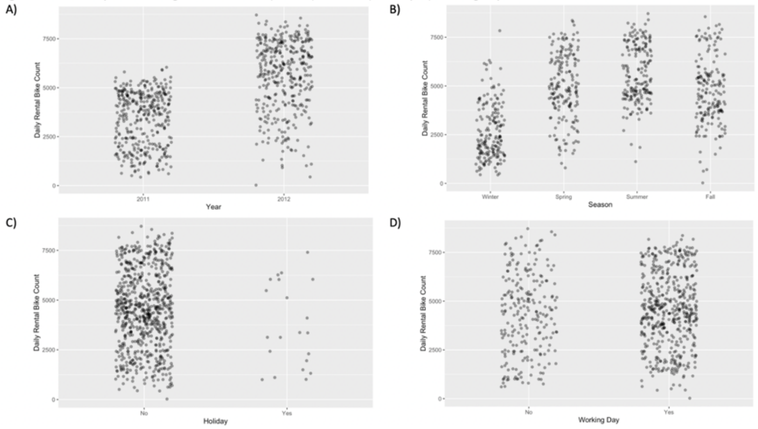


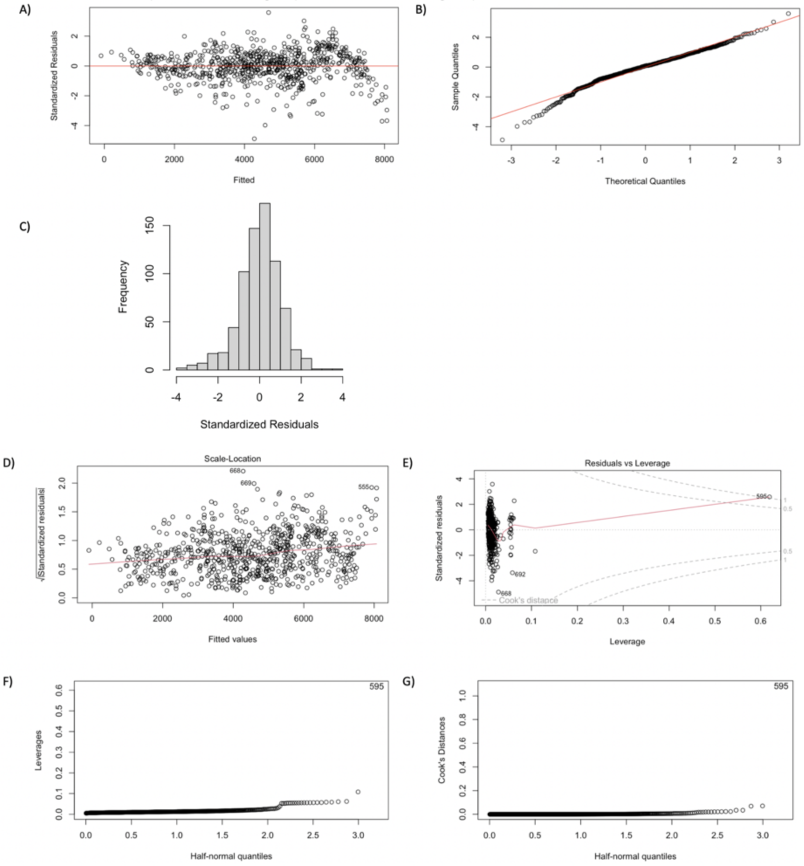
Figure 3. Bivariable plots for categorical variables. A) Year. B) Season. C) Holiday. D) Working Day.



Our initial full model utilizes linear regression to predict the response variable "daily rental bike counts" based on all 9 predictors: "year", "season", "holiday", "workday", "weather", "normalized temperature", "normalized feeling temperature", "normalized humidity", and "normalized windspeed". Our initial model was:

```
Daily bike counts = 1730.19 + 2034.07 (year) + 409.40 (season) – 621.06 (holiday) +
124.12 (workday) – 580.37 (weather) + 2326.31 (temperature) + 3312.12 (normalized
feeling temperature) – 1202.79 (humidity) – 2582.41 (windspeed)
```

Figure 4. Plots for the initial model. A) Residual vs. Fitted Plot. B) Quantile-Quantile Plot. C) Histogram. D) Fitted vs. Square-Root of Standardized Residuals. E) Residuals vs. Leverages. F) Half-Norm Plot for Leverages. G) Half-Norm Plot for Cook's Distances.



This model has an adjusted $R^2 = 0.7914$. In this model, the year being 2012, season, working day, increase in normalized temperature and normalized feeling temperature would increase daily rental bike counts, whereas holiday, weather, increase in normalized humidity and windspeed would decrease daily rental bike counts. the overall model is highly significant, as indicated by the F-statistic with $p < 0.01$ and predictors are statistically significant ($p < 0.01$), except working day ($p = 0.081$) and normalized feeling temperature ($p = 0.449$).

To check for assumptions of linear regression model, we plotted diagnostic plots for our initial full model (Figure 4). The standardized residuals against the fitted values (Figure 4A) illustrated that the residuals are mostly randomly scattered around the horizontal line at zero and that there are no clear patterns. However, the spread of the residuals did not seem to be completely uniform across all levels of the fitted values, suggesting potential heteroscedasticity in the model. The Quantile-Quantile plot (Figure 4B) and the histogram of standardized residuals (Figure 4C) showed that the distribution of residuals is generally normal. However, although the residuals largely follow the reference line in the Q-Q plot, there are deviations at the tail ends of the distribution, indicating the presence of outliers. The Scale-Location Plot also showed deviations from the horizontal line suggesting potential heteroscedasticity and data points in the top right deviating notably from the main cluster, indicating potential outliers (Figure 4D).

To improve our model, we decided to identify and remove potential outliners or influential points. We calculated leverages, Cook's distance and drew half-norm plots. The 595th observation stood out in both half-norm plots for leverages (Figure 4F) and Cook's distance (Figure 4G). We also applied Bonferroni Criterion to check for outliers and the 595th observation also stood out, which aligned with the Residuals vs. Leverage plot (Figure 4E). Based on the results, we identified that the 595th observation is an influential outlier and removed it from the initial model to improve model performance.

We then tested whether if multicollinearity exists in our model by calculating the variance inflation factors (VIF) for coefficients in our initial model (Table 2). Both predictors "normalized temperature" and "normalized feeling temperature" contained VIFs of more than 60, indicating multicollinearity, with the latter having a higher value. Therefore, we decided to remove the predictor "normalized feeling temperature" to test for improvement(s) in model performance.

Table 2. Variance Inflation Factors (VIF) for the initial model. This table shows the VIFs calculated for each predictor in the initial model after removal of leverages/outliers/influential points.

Variable	Year	Season	Holiday	Workday	Weather	Normalized Temperature	Normalized Feeling Temperature	Normalized Humidity	Normalized Windspeed
VIF	1.02	1.19	1.07	1.08	1.72	63.20	64.22	1.87	1.20

After removing both the influential outlier (595th observation) and predictor "normalized feeling temperature", our new model (Model 2) is:

```
Daily bike counts = 1701.99 + 2060.78 (year) + 429.52 (season) – 637.14 (holiday) +
135.76 (workingday) – 567.47 (weather) + 5160.16 (temperature) – 1044.52 (humidity) –
2284.42 (windspeed)
```

We calculated VIFs for coefficients of the Model 2 and all VIFs were close to one, indicating that multicollinearity issue was solved (Table 3). The adjusted R^2 remained almost the same (from 0.7999 to 0.7988) for the new model while "normalized temperature" became statistically significant.

Table 3. Variance Inflation Factors (VIF) of the new model after removing "Normalized Feeling Temperature". This table shows the VIFs calculated for each predictor in the new model after removing "Normalized Feeling Temperature" from the model.

Variable	Year	Season	Holiday	Workday	Weather	Normalized Temperature	Normalized Humidity	Normalized Windspeed
VIF	1.02	1.19	1.07	1.08	1.71	1.20	1.86	1.17

Our next step is to use the Akaike Information Criterion (AIC)-based model selection to select the best model (Figure 5), to see if

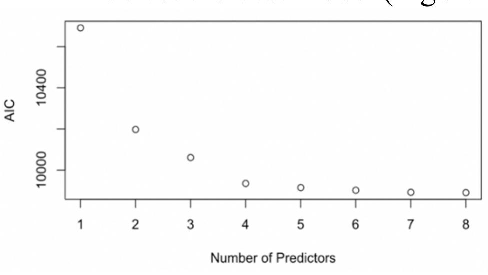


Figure 5. AIC vs. number of predictors plot. The x axis shows number of the predictors. And the y-axis shows the AIC log-likelihood values for model selection for the linear regression model with 8 predictors.

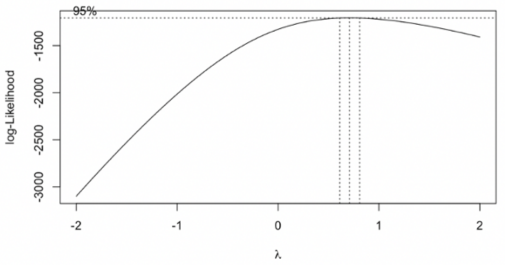


Figure 6.Log-likelihood vs. Power transformation. This plot shows the log-likelihood vs power transformation of the linear regression model. Dotted lines are the optimal power transformations with 95% confidence interval

dropping any other predictor(s) will further improve our model. The results shows that the new model containing all 8 predictors has the lowest AIC value and thus, we decided to keep the Model 2.

Next, we checked for interaction terms using ANOVA as our first step and found that there are 12 out of 28 significant interaction terms, all with $p < 0.05$. The addition of these 12 significant interaction terms improved the performance of our Model 3 since the adjusted R^2 increased to 0.8252, compared to an adjusted $R^2 = 0.7988$ for the previous model, Model 2. To further improve our model performance, we utilized the Box-Cox method to transform Model 3 and found that the λ that maximized the likelihood was 0.7071 (Figure 6). This transformation was carried out and the adjusted R^2 of the model has increased to 0.8388. The Transformed Model 3 has the best model performance since it has the highest adjusted R^2 compared to all other models and thus, it was determined as our final model.

Our final model consists of 8 predictors and 12 interaction terms, as following:

(Daily bike counts$^{0.7}$-1)/0.7 = -26.816 + 102.404 (year) + 71.918 (season) - 70.071 (holiday) + 10.215 (workingday) + 64.941 (weather) + 588.851 (normalized temperature) + 62.654 (normalized humidity) + 33.738 (windspeed) + 2.352 (year*season) + 13.147 (year*workingday) - 11.910 (year*weather) + 41.648 (year*temperature) - 106.206 (season*temperature) - 4.571 (season*humidity) + 61.627 (holiday*temp) - 55.069 (workingday*temperature) + 99.090 (workingday*windspeed) - 79.309 (weather*humidilty) - 202.108 (weather*windspeed) + 75.596 (temperature*windspeed)

Our final model achieved an adjusted $R^2 = 0.8388$, which can explain 83.88% of variance on daily bike counts using 8 predictors and 12 interaction terms. In the plot of predicted daily bike count values using final model vs. actual values (Figure 7), the performance of our final model was promising.

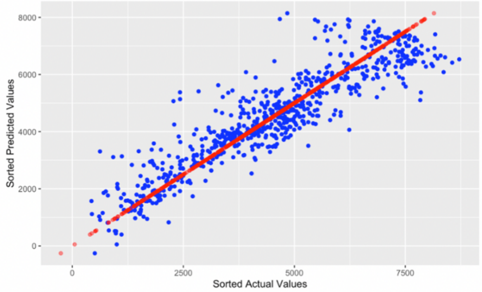


Figure 7. Predicted values vs. actual values for the final linear regression model. This plot shows the predicted daily bike counts using the final linear regression model, versus actual bike counts values.

Summary

Our final model containing 8 predictors and 12 interaction terms has an adjusted $R^2 = 0.8328$ compared to our initial model containing only 9 predictors including "normalized feeling temperature", which only has an adjusted $R^2 = 0.7914$. This indicates that the variance in bike counts that can be explained by the predictors in the model increased from 79.14% to 83.28%, meaning that we can more accurately predict the total daily rental bike count using our final model compared to using our initial model.

There are limitations in our model, as the diagnostic plots show that moderate heteroscedasticity may exist in our model. For potential analysis or study design improvements, machine learning techniques such as K-Nearest Neighbors (KNN), random forests, and time series analysis can be implemented to reduce overfitting/underfitting and increase flexibility in model prediction.

Contributions

Xiaotong found and proposed the dataset. Ivy wrote the Introduction and the first half of Methods. Xiaotong wrote the second half of Methods and the first half of Results. Helen wrote the second half of Results and Summary. Ivy complied Tables 1, 2, and 3 and illustrated Figure 1. Xiaotong illustrated Figures 2 and 3. Helen illustrated Figures 4, 5, and 6. Helen and Xiaotong illustrated Figure 7.

We all worked very hard on this project.