STAT 5302 Final (practice)

Elliot Orenstein

## Introduction

We have been asked to analyze a dataset containing the daily number of bikes shared in London from 2015 to 2016 and to focus on two main objectives. The first is to determine which variables among those provided in the dataset contribute most to increasing the daily number of bikes shared. In addition to the number of bikes shared daily, the dataset contains daily information in the form of seven numerical and categorical variables. The numerical variables are temperature, feels-like temperature (both measured in Celcius), humidity (percentage), and wind speed (measured in km/hr). The categorical variables are weekend (1 for weekend), bank holiday (1 for bank holiday), and season (0 for spring, 1 for summer, 2 for fall, 3 for winter).

Second, the analysis will uncover which model(s) can best help in predicting daily bikes shared for future dates. Given the seven variables provided, there will be many predictive models from which to choose. This analysis will approach this objective systematically, using domain knowledge as well as statistical methods and guidelines to find a smaller subset from which to choose.

## Methods

The analysis used many statistical methods to help in achieving these two main objectives. A key tool used in determining which variables contribute most to the daily number of bikes shared were pairwise scatterplots and the correlation matrix. Scatterplots of the predictors versus the response can indicate if and how the response changes as the predictor varies. Whether a strong relationship exists between a predictor and the response is important in gaining a preliminary understanding of a predictor’s significance.

If a relationship appears to exist, these scatterplots can also indicate whether it appears linear, quadratic, logarithmic, or otherwise. This can be of great help in finding a well-fitting predictive model. The scatterplots between predictors are also important as they can indicate instances of collinearity. Collinearity between predictors will lead to inflated variances, which can interfere with standard errors and therefore with the p-values and thus the significance of predictors in a model. The correlation matrix serves a similar purpose as the correlation matrix in that it helps determine the relationship between the predictors and response and among the predictors themselves. This puts a numeric value to the linear relationship, which can help quantify this in a way that pairwise scatterplots cannot. However, correlations only indicate linear relationships. Therefore, using both is essential to obtaining a full picture on the underlying relationships in the data.

In addition to these tools, which are used throughout the analysis, the following methods were used primarily in finding and choosing a predictive model. An analysis of variance (ANOVA) test was used as the primary method for determining the goodness of fit of the models. ANOVA allowed us to find a subset of models that provide significantly better fits than their counterparts. Because of the marginality principle ANOVA relies on, it is especially useful in determining which interaction effects are significant, which was key to this analysis as potential models contained many interaction effects due to the three categorical variables. After finding a subset of models, summary outputs of the linear models displayed important information such as adjusted R squared as well as estimated values, p-values, and standard errors of the predictors’ coefficients. These were used in interpreting as well as determining the significance and contributions of the predictors in estimating the response as well as for constructing an estimate or prediction of the response for unseen data. Given that we were not provided a test set, we used k-fold cross validation to determine how well the chosen model(s) generalized to new data. The resulting test mean square error (MSE) of a model is a good sign of its predictive accuracy. Following this, diagnostic tools such as histograms of the residuals, qq-plots, standardized residual plots, and Cook’s distance plots were employed to determine whether the assumptions of the errors held under the model. If not, a transformation of the predictors and/or response may be necessary. Given that the response is strictly positive, the Box-Cox method could be used to determine an optimal transformation of the response.

## Results

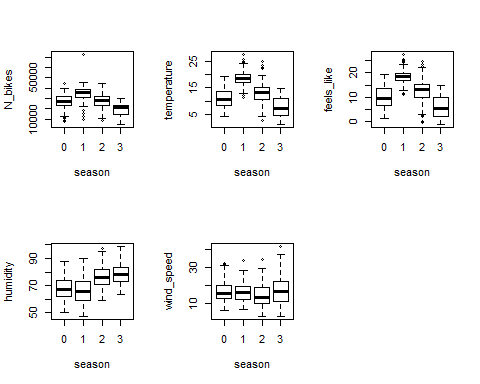
To gain a better understanding of the data, we started by performing exploratory data analysis and data cleansing.

## Loading required package: carData

## N\_bikes temperature feels\_like humidity   
## Min. : 4869 Min. : 1.312 Min. :-1.250 Min. :47.23   
## 1st Qu.:21764 1st Qu.: 8.646 1st Qu.: 6.594 1st Qu.:65.43   
## Median :26927 Median :12.469 Median :12.354 Median :72.20   
## Mean :27074 Mean :12.515 Mean :11.563 Mean :72.20   
## 3rd Qu.:33340 3rd Qu.:16.172 3rd Qu.:16.172 3rd Qu.:79.32   
## Max. :72504 Max. :27.375 Max. :27.438 Max. :98.69   
## wind\_speed holiday weekend season   
## Min. : 2.771 Min. :0.000 Min. :0.000 Min. :0.000   
## 1st Qu.:11.417 1st Qu.:0.000 1st Qu.:0.000 1st Qu.:0.000   
## Median :15.615 Median :0.000 Median :0.000 Median :1.000   
## Mean :16.245 Mean :0.026 Mean :0.292 Mean :1.484   
## 3rd Qu.:20.141 3rd Qu.:0.000 3rd Qu.:1.000 3rd Qu.:2.250   
## Max. :41.389 Max. :1.000 Max. :1.000 Max. :3.000

Because holiday, weekend, and season are categorical variables, we set them as factors with their references as non-holiday, non-weekend, and spring, respectively. Examining the summary output of the dataset, we see that the number of bikes, temperature, humidity, and wind speed all have a fairly wide range. This could be a sign that all or some may benefit from a logarithmic transformation. Feels-like also has a wide range but has negative values, as evidenced by its negative minimum value, so it cannot be log transformed. Also, while season and weekend are fairly well-disbursed, there are only 13 bank holidays. This may cause issues in generating higher-level interactions. Because holiday and weekend cannot both be ones for the same day, they could be combined into a new variable, which we called day type. However, we decided to keep them as separate as it led to a more straightforward interpretation.

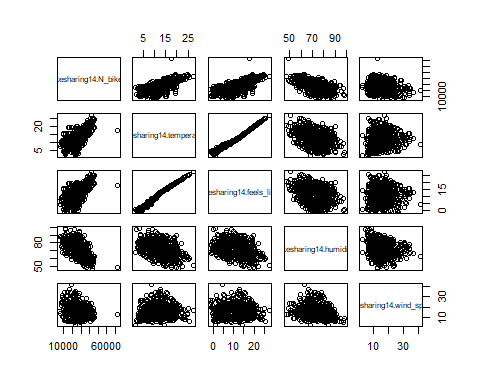
We can see some patterns between categorical and numerical data, but it is difficult to discern from these plots, so we will supplement them with domain knowledge. Below, we have focused in on some of these relationships in the scatterplots.



For example, we can see that temperature ranges are much different depending on the season, which makes sense given our knowledge about weather in London. We see the same pattern between feels-like and season. There is some evidence from the box plot that the number of bikes is influenced by the season and a weaker relationship between season and humidity and nearly no relationship with wind speed.

The relationships between weekend and holiday with the numerical variables provides an initial perspective into the impact these categorical variables have on the number of bikes and examples of potential collinearity. There is no reason to believe these time-based categories would correlate with weather-related variables but we examined the plots to verify this assumption.

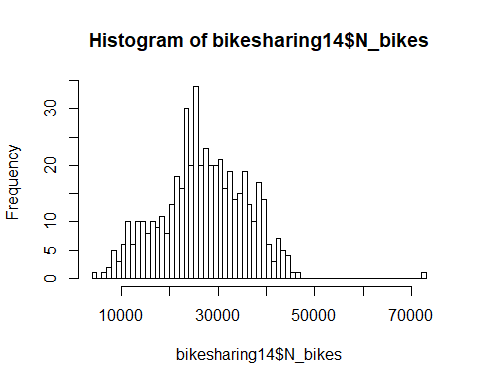
Focusing now on the numerical data, we see that the number of bikes has strong positive relationships with temperature and feels-like, a negative relationship with humidity, and a weaker negative relationship with wind speed. There are also very strong correlations between the predictors, especially temperature and feels like, which have a near one-to-one linear relationship (0.992 estimated correlation).



## bikesharing14.N\_bikes bikesharing14.temperature  
## bikesharing14.N\_bikes 1.0000000 0.67209044  
## bikesharing14.temperature 0.6720904 1.00000000  
## bikesharing14.feels\_like 0.6698234 0.99247085  
## bikesharing14.humidity -0.5808705 -0.38135513  
## bikesharing14.wind\_speed -0.2229315 0.02481386  
## bikesharing14.feels\_like bikesharing14.humidity  
## bikesharing14.N\_bikes 0.6698234 -0.5808705  
## bikesharing14.temperature 0.9924709 -0.3813551  
## bikesharing14.feels\_like 1.0000000 -0.3494702  
## bikesharing14.humidity -0.3494702 1.0000000  
## bikesharing14.wind\_speed -0.0163449 -0.1779635  
## bikesharing14.wind\_speed  
## bikesharing14.N\_bikes -0.22293154  
## bikesharing14.temperature 0.02481386  
## bikesharing14.feels\_like -0.01634490  
## bikesharing14.humidity -0.17796353  
## bikesharing14.wind\_speed 1.00000000

In summary, from this preliminary data exploration, we found that temperature, feels-like, season, and weekend appear to have the strongest relationship with the daily number of bikes shared. Similarities between any two of temperature, feels-like, and season are strong enough to potentially create an issue with collinearity. Therefore, we considered models with only one of these predictors. Lastly, we may need to consider a log transformation of our response or some of our predictors to abide by the assumptions on the errors.

Before running an ANOVA to determine the models with the best goodness-of-fit, we consider the distribution of the data.



The number of bikes shared daily is count data (non-negative, whole numbers over a fixed time), which could indicate it follows a Poisson distribution. With larger means, a Poisson distribution begins to approximate a Normal distribution. The histogram of the number of bikes indicates that the distribution may be normal although it is slightly asymmetical. Because of this, we will assume a (normal) multiple linear regression for now but will consider a poisson linear regression model if necessary. Additionally, there is no indication contextually or in the data that the errors are correlated. For now, our covariance matrix will be .

We then ran an ANOVA on model 1, where model 1 is the largest possible model of temperature, wind speed, weekend, holiday, and humidity.

## Anova Table (Type II tests)  
##   
## Response: N\_bikes  
## Sum Sq Df F value Pr(>F)   
## temperature 8262453475 1 492.8102 < 2.2e-16 \*\*\*  
## wind\_speed 3382571149 1 201.7519 < 2.2e-16 \*\*\*  
## weekend 3162475558 1 188.6244 < 2.2e-16 \*\*\*  
## holiday 997671923 1 59.5057 6.919e-14 \*\*\*  
## humidity 5405885972 1 322.4315 < 2.2e-16 \*\*\*  
## temperature:weekend 86064340 1 5.1333 0.0239089 \*   
## temperature:holiday 1459459 1 0.0870 0.7680886   
## wind\_speed:weekend 268553059 1 16.0177 7.250e-05 \*\*\*  
## wind\_speed:holiday 8532966 1 0.5089 0.4759370   
## weekend:humidity 238719205 1 14.2383 0.0001809 \*\*\*  
## holiday:humidity 44889857 1 2.6774 0.1024243   
## Residuals 8181806194 488   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

At the significance level of , the ANOVA resulted in six models with significant F-tests. These models – which we labeled models 11-16, from the top of the ANOVA table down – are the alternate hypotheses of the first six rows of the table (the remaining two are identical to the sixth). Model 11, for example, includes temperature but not its interactions. Model 16 is the full model, equivalent to model 1.

Performing an ANOVA on model 2, where temperature is replaced with feels-like, resulted in nearly identical results.

## Anova Table (Type II tests)  
##   
## Response: N\_bikes  
## Sum Sq Df F value Pr(>F)   
## feels\_like 8259905665 1 491.5762 < 2.2e-16 \*\*\*  
## wind\_speed 2980229624 1 177.3640 < 2.2e-16 \*\*\*  
## weekend 3145618776 1 187.2069 < 2.2e-16 \*\*\*  
## holiday 1026929674 1 61.1162 3.340e-14 \*\*\*  
## humidity 5948952537 1 354.0432 < 2.2e-16 \*\*\*  
## feels\_like:weekend 69067792 1 4.1105 0.04316 \*   
## feels\_like:holiday 2533668 1 0.1508 0.69795   
## wind\_speed:weekend 282430856 1 16.8085 4.843e-05 \*\*\*  
## wind\_speed:holiday 9653518 1 0.5745 0.44884   
## weekend:humidity 261742892 1 15.5772 9.084e-05 \*\*\*  
## holiday:humidity 39025933 1 2.3226 0.12816   
## Residuals 8199815076 488   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

At the significance level of , the ANOVA resulted in six models with significant F-tests. These models were labeled 21-26, from the top of the ANOVA table down.

Lastly, an ANOVA was performed on model 3, which uses season in place of temperature or feels-like. Because season is a factor, it will have additional interactions to consider compared to models 1 and 2.

## Note: model has aliased coefficients  
## sums of squares computed by model comparison

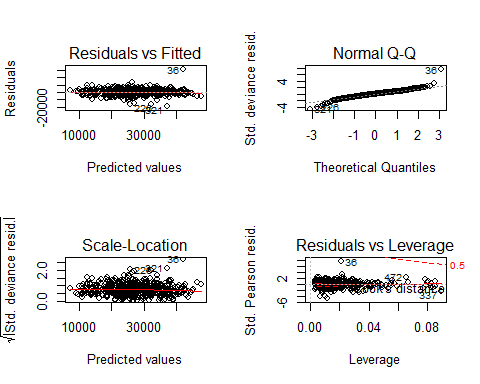
## Anova Table (Type II tests)  
##   
## Response: N\_bikes  
## Sum Sq Df F value Pr(>F)   
## season 5657740629 3 88.4721 < 2.2e-16 \*\*\*  
## wind\_speed 1605278797 1 75.3069 < 2.2e-16 \*\*\*  
## weekend 3223553408 1 151.2235 < 2.2e-16 \*\*\*  
## holiday 662632644 1 31.0855 4.174e-08 \*\*\*  
## humidity 5455398820 1 255.9239 < 2.2e-16 \*\*\*  
## season:weekend 100698532 3 1.5747 0.1946892   
## season:holiday 57419326 2 1.3468 0.2610676   
## wind\_speed:weekend 129125530 1 6.0575 0.0142061 \*   
## wind\_speed:holiday 50515706 1 2.3698 0.1243777   
## weekend:humidity 128505842 1 6.0285 0.0144388 \*   
## holiday:humidity 8350430 1 0.3917 0.5316929   
## season:wind\_speed:weekend 300280231 6 2.3478 0.0303330 \*   
## season:weekend:humidity 552605208 6 4.3206 0.0002954 \*\*\*  
## season:wind\_speed:holiday 28529 1 0.0013 0.9708325   
## season:holiday:humidity 16260166 1 0.7628 0.3829014   
## Residuals 9997432547 469   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

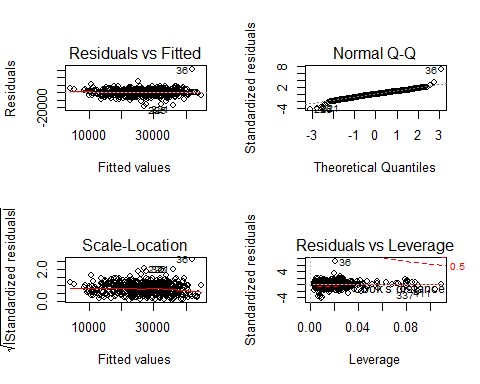
At the significance level of , the ANOVA resulted in seven models with significant F-tests. These models were labeled 31-37, from the top of the ANOVA table down.

Based on each model’s R squared as the criteria for goodness of fit, models 11, 14, 16, 21, 24, and 26 were the top candidates with adjusted R squared values all over 0.78. Adjusted R squared takes into account generalizability by penalizing models with more parameters but with them so close in value, we used an additional criteria, estimated test MSE generated by k-fold cross validation, on all of the models from the three ANOVAs to help aid in model selection.

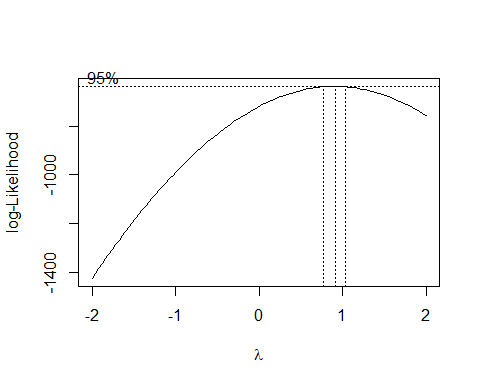
## Adjusted error m11 Adjusted error m12 Adjusted error m13 Adjusted error m14   
## 17200198 17949255 18581286 17072423   
## Adjusted error m15 Adjusted error m16 Adjusted error m21 Adjusted error m22   
## 17794618 17516386 16994442 17705764   
## Adjusted error m23 Adjusted error m24 Adjusted error m25 Adjusted error m26   
## 18531547 17058846 17499437 17149314   
## Adjusted error m31 Adjusted error m32 Adjusted error m33 Adjusted error m34   
## 22734667 23086093 23376695 23408517   
## Adjusted error m35 Adjusted error m36 Adjusted error m37   
## 24027537 22942741 26000852

Model 14 had the lowest estimated test MSE so we chose it as our model. To determine whether the assumptions of multiple linear regression held for model 14, we ran diagnostics, which can be found below on the left. Model 14’s residual plots validate the assumptions of the errors as they are centered around zero, have no trend, and maintain constant variance. The qq-plot deviates slightly, most significantly at the left tail. The Cook’s distance plot does not show any data is influential, so there is no evidence to remove certain points from the dataset. If in fact the errors are not normally distributed, this would affect inference but not the estimates of the coefficients and their standard errors. Applying a log transformation to temperature and wind speed, as seen on the right, results in a better-fitting qq-plot but introduces a slight trend and non-zero mean at the tails. As mentioned previously, Poisson regression was another consideration. However, testing various Poisson models, including one identical to model 14, led to worse estimated test MSE and diagnostics that violated the assumptions on the error.





Lastly, we considered a transformation of the response. Using Box-Cox, we found the optimal transformation to be , or the identity transformation.



As a result, model 14 was the ultimate choice to predict daily bikes shared. The summary output is below.

##   
## Call:  
## lm(formula = N\_bikes ~ temperature + wind\_speed + weekend + holiday +   
## humidity + temperature:weekend + wind\_speed:weekend + humidity:weekend,   
## data = bikesharing14)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18732.9 -2194.3 219.6 2329.8 30540.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 48065.25 2226.23 21.590 < 2e-16 \*\*\*  
## temperature 806.49 45.06 17.900 < 2e-16 \*\*\*  
## wind\_speed -331.53 36.91 -8.983 < 2e-16 \*\*\*  
## weekend1 8414.03 4312.89 1.951 0.051637 .   
## holiday1 -8973.77 1166.10 -7.696 7.79e-14 \*\*\*  
## humidity -329.68 24.55 -13.427 < 2e-16 \*\*\*  
## temperature:weekend1 202.10 88.40 2.286 0.022674 \*   
## wind\_speed:weekend1 -257.76 62.69 -4.112 4.60e-05 \*\*\*  
## weekend1:humidity -169.24 46.89 -3.610 0.000338 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4104 on 491 degrees of freedom  
## Multiple R-squared: 0.7864, Adjusted R-squared: 0.7829   
## F-statistic: 226 on 8 and 491 DF, p-value: < 2.2e-16

As compared to model 16, which is the same as model 14 but also includes interaction effects of holiday, model 14 has a more clear interpretation. None of the interaction effects with holiday are significant in model 16 whereas all but the main effect of weekend are significant at in model 14. It is unsurprising that these interactions would not be significant in model 16 given that so few data points (13 of the 500 samples) are bank holidays. Interpreting model 14 shows that bank holidays result in approximately 8,974 fewer bikes shared that day compared to non-holidays, adjusting for all other predictors. Holiday, followed by weekend and temperature have the most impact on bike sharing numbers. However, weekend is not significant at so we cannot say definitively that it is different from zero. Therefore, we can confidently claim that, under this model, that holiday and temperature have the greatest impact on the number of bikes shared (for a one unit change or compared to its reference category for factors).

## Discussion

The analysis on the daily bike sharing data had two main objectives. First was to find the variables that contributed most to the number of bikes shared. Preliminary data exploration hinted that some variables such as temperature contributed more than others such as wind speed. The second objective was to choose the best model(s) to predict bike sharing in the future. Using ANOVA and k-fold cross-validation, we determined the best model to be model 14. The analysis brought up an interesting question regarding the first objective. Simply looking at the data to determine the variables that contribute the most to the response resulted in a different set of variables than we ultimately ended with in model 14. This was due to the deliberate choice to avoid issues of collinearity. Ultimately, answering the question of which variables contribute the most comes down to the context. Since the analysis concluded in finding a predictive model, the first objective can be achieved in a less open-ended manner by examining it in the context of the second objective.

Another question that arose from this analysis was how to interpret the diagnostic tests of model 14. While the second objective is focused more on prediction than interpretation, the model could be more useful if we could more confidently believe in its inference. This could help more confidently address the first objective, again depending on how it is framed. In summary, gaining access to more data could help the dataset asymptotically approach normality and therefore provide more confidence in its inference.

Future analyses that could further improve the predictive ability of our model could be constructed if we had more predictors. This is not guaranteed to help as it depends on the predictors. Some examples of potentially useful predictors would be whether a promotion or special sale was occurring for the service, weather forecasts, unemployment rates, traffic conditions or road closures, and the supply of bikes available through the service. Although some of these may correlate with existing predictors, they could provide a stronger relationship with the response than those currently present.