"Data Story to explain factors impacting count of daily bike rentals in Washington DC"

Introduction

Background - Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return has been automated. Through these systems, user can easily rent a bike from one position and return back at another position. *Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.*

Dataset Description – The dataset contains the daily count of rental bikes between years 2011 and 2012 in ***Capital bikeshare system*** with the corresponding weather and seasonal information. Dataset includes weather attributes like temperature, humidity, windspeed as well as seasonal attributes like season, month, weekday or working day etc. *Number of Observation*: 731, *Number of attributes*: 16, *Response variable* – Count of Daily bike rentals

Variable Names - "instant" "dteday" "season" "yr" "mnth" "holiday" "weekday" "workingday" "weathersit" "temp" "atemp" "hum" "windspeed" "casual" "registered" "**cnt**"

Goal – we will try to determine the factors that influence the total number of bike rentals on any particular day. Or in other words, what are the factors that influence number of bike rentals on a day?

Exploratory Data Analysis

First, we checked the data for missing values and found that there is no missing data in our dataset.

Few interesting observations while conducting EDA:

1. Our dependent variable (count of users) increases with increase in temp and reduces with increase in humidity and windspeed. When we cross-checked this, we found that more people rent bikes on "good" (clear/few clouds) weather days rather than "bad" (heavy rain/fog/thunderstorm) weather days

2. Variables "temp" and "atemp" show high positive correlation (0.99) and hence both can't be used for the analysis due to assumption of singularity. It means that actual temperature and feels like temperature change together in same direction

3. Count of "registered" users and "total" users is also highly correlated (0.97) which means that registered users constitute a major component of total users on any day. When we cross-checked, we found that registered users indeed form 70-90% of total users on any given day.

4. Bike users are spread almost evenly across days of the week

5. Interestingly, when we plotted count of users on monthly bases across both years (2011 and 2012), we see a bell-shaped curve. Count of users peaked in sept increasing gradually through Jan-Sept and then declined again in Nov and December. Same is confirmed when

we look at season and temperature graphs, bike rentals increase with rise in temperature in summer/fall and then see a decline in winter and on low temperature days.

6. Our Y variable (Count of daily rentals) is almost normally distributed

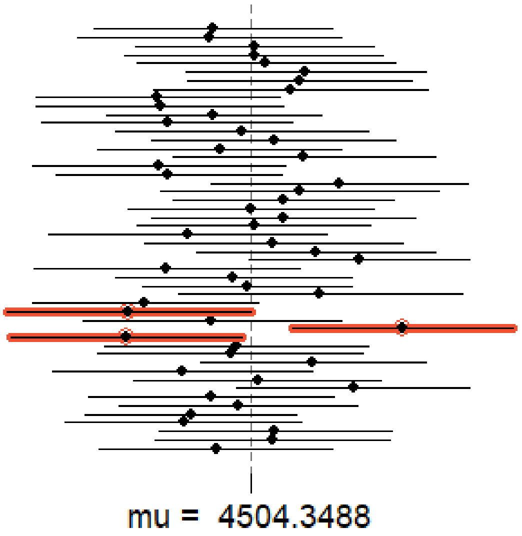
7. More registered users are renting bikes on working days and on the contrary more casual users are renting bikes on non-working days

8. Average bike users are higher on a working day as compared to non-working day. Also, higher casual users are seen on weekends as compared to weekdays. This means that mostly registered users use bikes to commute to work on working days.

Inference

Research question: Estimate the average number of daily bike rentals.

We used loop function to take 50 random samples of size 60 from the available daily data of 731 days. We used these samples to create sampling distribution for the dataset.



95% Confidence Interval

Using these 50 sample distributions, we created 50 CI for mean daily bike rentals. Upon plotting, we can see that 47/50 CI contain the actual population mean. Hence, we can infer with 95% Confident that real average daily bike rentals lie between 4071 & 4905.

Data Manipulation & Preparation

Once we completed our EDA and had a fair understanding of the data, we went ahead to manipulate the data to make it ready for modelling. We:

- changed categorical variables to dummy variables using one-hot encoding
- dropped columns – instant (simply count of rows), dteday (date of row), casual (number of casual rentals on a day), registered (number of registered rentals on a day), atemp (feels like temperature)
- transformed months to quarters for easy of modelling
- bring together all the columns and our final data is ready for modelling

First, we checked assumptions for linear regression namely

- normality
- linearity
- additivity
- homogeneity/homoscedasticity

Results – Our data is nearly normally distributed and passes through other assumptions of linear regression.

Preparing Data for modelling

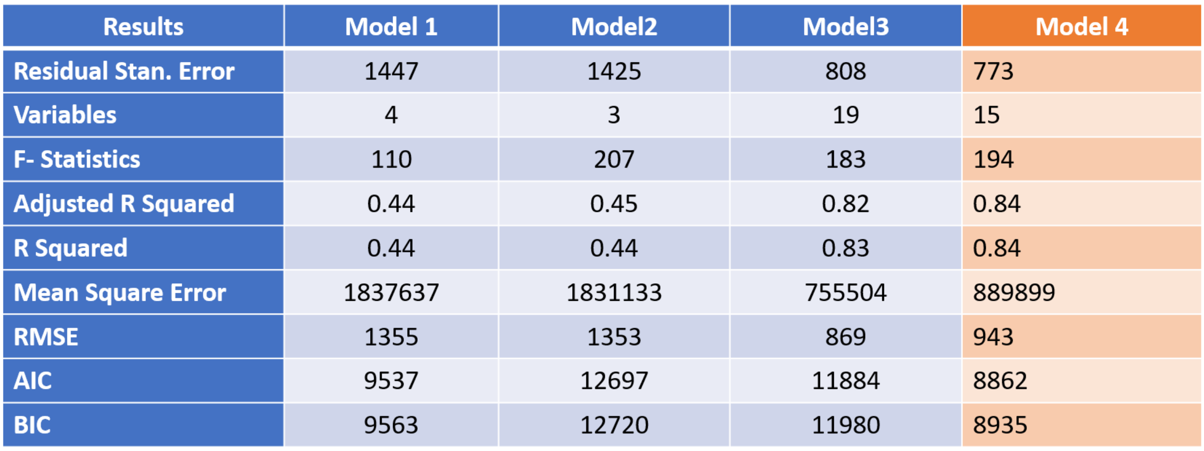
We split the dataset into train and test datasets in 75:25 proportion using random sampling. We used train dataset to fit the regression model and test dataset to predict y variable using the model fitted on train dataset.

Linear Regression Modelling – The "Best" Model

Process - We used lm() function in R to fit ***4 linear regression models*** for y variable using different sets of independent variables in each model.

Decision Parameters - We used adjusted R squared as our bases for model selection and improvement along with other scores like RMSE, MSE, AIC, BIC. Model with higher value of adjusted R squared and lower values of AIC, BIC, MSE & RMSE is a better fit.

Here, we can see that ***model 4 is the best fit*** amongst the 4 models that we created. It has highest adjusted R2 value at 0.84 which means the model explains 84% of the variation in Y variable (count of daily bike rentals) using selected independent variables.



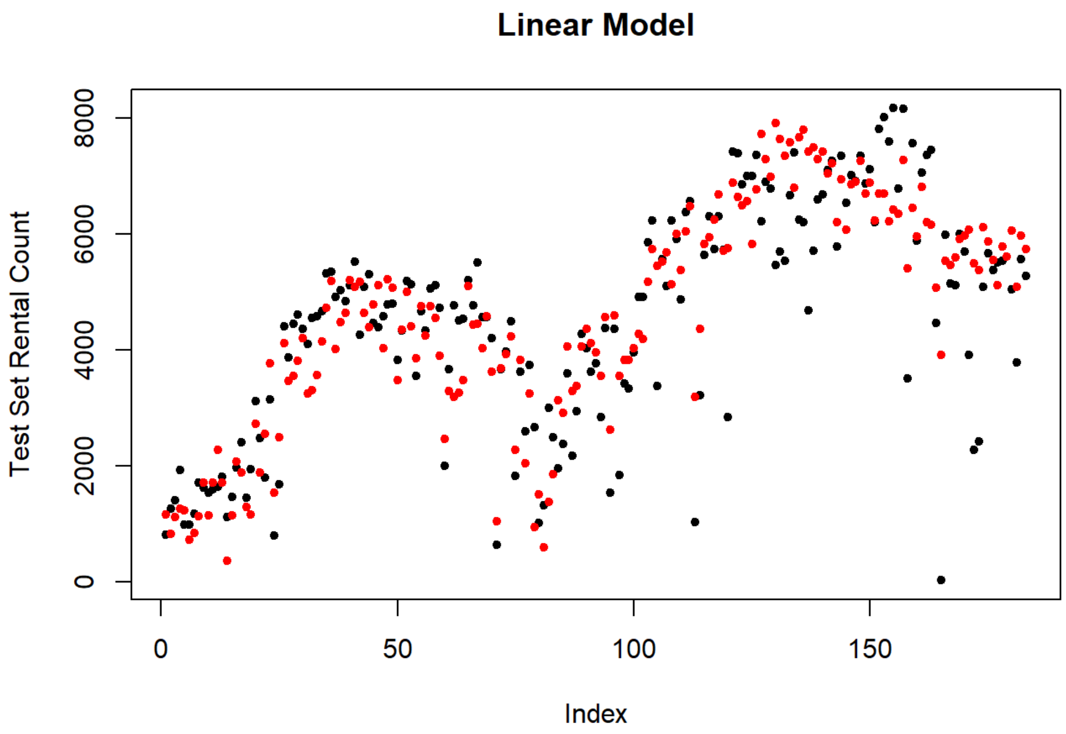
Results	Model 1	Model2	Model3	Model 4
Residual Stan. Error	1447	1425	808	773
Variables	4	3	19	15
F- Statistics	110	207	183	194
Adjusted R Squared	0.44	0.45	0.82	0.84
R Squared	0.44	0.44	0.83	0.84
Mean Square Error	1837637	1831133	755504	889899
RMSE	1355	1353	869	943
AIC	9537	12697	11884	8862
BIC	9563	12720	11980	8935

Interpretation

We built multiple linear regressions by putting all variables against response variable and removed insignificant predictor variables from earlier models. The best fit model achieved 0.84 adjusted R-squared, which indicates a good fit. Also, p value for all predictor variables are below 0.05 which indicate statistical significance. Also, checking the residual plot and QQ plot we can see that residuals have no pattern and are normally distributed, which means the model fit the data well. All of this assumes that our assumptions of linear regression are fulfilled and model results will be irrelevant if these assumptions were not met.

Prediction

When we plot the prediction of best fit model (model 4) on test dataset against the true values, the graph shows that the spread of the response variable is similar to multilinear model. Still, we cannot depend on this because we worked on a small data and this dataset does not contain other variables such as daily hours and bike stations, which can help more in accuracy.



Conclusion

"An ideal day for highest bike rentals would be a warm working day in summer/fall with low humidity and slow wind"

We can conclude that count of bike rentals on a day is dependent of a number of factors – both seasonal and weather related.

● **How well can you predict your response variable?**
Our model is a good fit and can predict response variable with good accuracy as we can see above in the graph of actual test data against the prediction

● **What are the caveats to your analysis?**
We did not incorporate company's growth plans in our analysis. We could see that bike rentals increased from 2011 to 2012 over same time period. So using our model to predict for future won't be right as it does not incorporate company's expansion plans.

● **Does this data set lack information that you would have liked to use?**
We did not have hour wise data which could have validated our finding that working people who are registered users use bike to commute to office. Bike station wise data visibility could also be helpful to improve the model