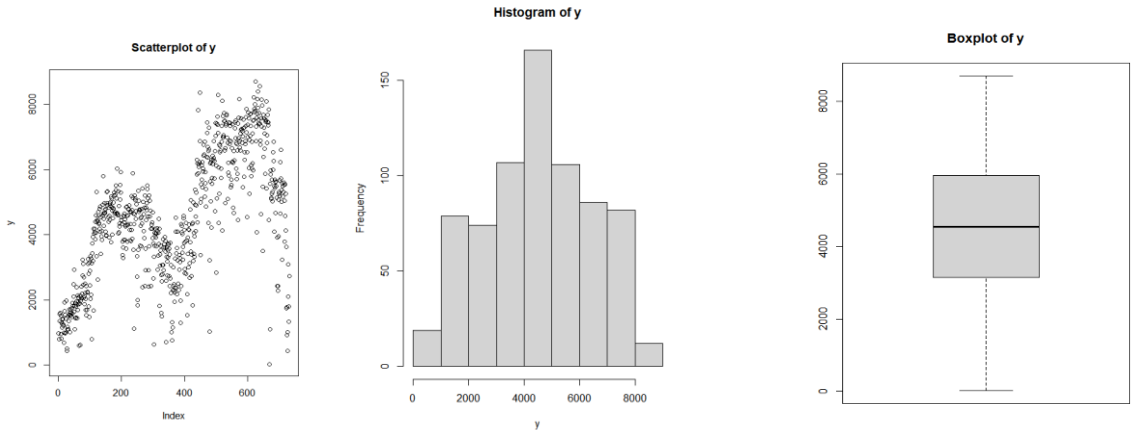
Building a Linear Regression Model on Bike Rental Dataset

This personal project proposes building a Supervised Linear Regression Model to predict the total daily count of bike rental based on potential regressors. It looks at regressor-variables such as season, work-day, weather, temperature, humidity, and windspeed. Model is built using R programming language in RStudio and dataset is obtained from :
https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset#

Q1.Exploring Suitability for a linear regression model.

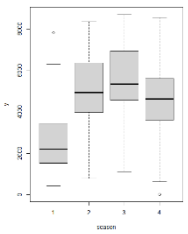
```
summary(y)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   22    3152    4548    4504    5956    8714
```



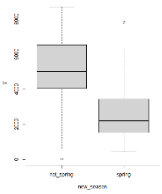
- Suitable.
- Quantitative Response variable, y. Shown by scatterplot.
- Mean and Median almost similar, almost normal, centered around mean
- Histogram resembles bell shaped, almost symmetrical normal distribution
- N=731 >30, sampling distribution will resemble normal
- No outliers in boxplot, mean and standard deviation can be effectively used, as no strong outliers present to skew distribution. Also, tails are pretty symmetrical, and median is on middle of box. Possibly almost normal.

Q2. Testing association between regressors (Model Building Part A)

Response and season (Categorical)

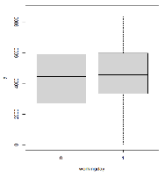


- season seem to have an association with response.
- Season 2,3,4 (Summer, fall, winter) have significantly higher median response
- compared to Season 1(Spring)
- Highest average bike rentals occurs in 3(Fall), followed by 2(summer), then 4(winter) and 1(Spring).



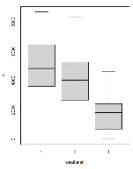
- Transforming season to categorical variable with 2 categories ("not_spring" vs "spring") shows strong association, as there is a large change in median response when comparing both catergories.
- Suitable prospective regressor due to strong association between season and response.
-

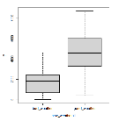
Response and workingday (Categorical)



- Working day and bike rental does not seem to have strong association with response
- Median for 0(weekend/holiday) and 1(otherwise) are approximately the same
- Shows that explanatory variable workingday might not affect response
- Not-suitable prospective regressor as weak association between workingday and response.

Response and weathersit (Categorical)

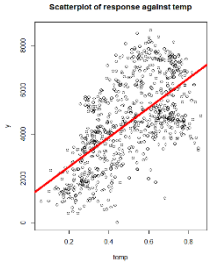




- Weathersit seems to have a very strong association with response
- Median response for weathersit= 1 and 2 higher than 3
- Shows that explanatory variable weathersit might have a strong effect on response

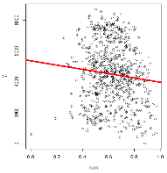
- Transforming weathersit to categorical variable with 2 categories ("bad_weather" vs "good_weather") shows strong association, as there is a large change in median response when comparing both catergories.
- Suitable prospective regressor due to strong association between weathersit and response.

Response and temp (Quantitative)

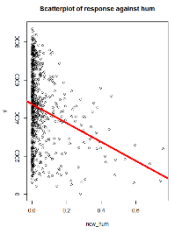


Correlation= 62.7%
Positive, Strong correlation
Possibly linear
Temp have almost strong association to response.
Suitable prospective regressor due to almost strong association between temp and response.

Response and hum (Quantitative)



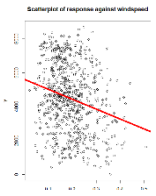
Correlation= -10.1%
Negative, weak correlation
Possibly not linear from scatterplot
Weak to no association with response



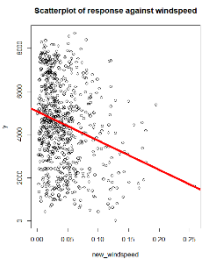
After transformation of hum^10(higher order),
Increased Correlation= -24.6%
Negative, weak correlation
Possibly Slightly linear from scatterplot
Weak association with response

- Might be suitable prospective regressor due to slightly weak association between hum and response. Will not be first choice but will add to model if needed.

Response and windspeed (Quantitative)



Correlation= -23.5%
Negative, weak correlation
Very Possibly not linear from scatterplot
Weak to no association with response.



After transformation of hum^2(higher order),
Increased Correlation= -24.0%
Negative, weak correlation
Possibly Slightly linear from scatterplot
Weak association with response

- Might be suitable prospective regressor due to slightly weak association between windspeed and response. Will not be first choice but will add to model if needed.

Q3. Initial Proposed Model M_1 using proposed regressors (Model Building Part B)

3) Proposed Regressors: ① temp (Quantitative), almost Strong association with response
② weathersit (Categorical) } both strong association with response
⑤ season (Categorical)
* detailed explanation for why association is strong is explained earlier in Qn 2.

Y = response = cnt

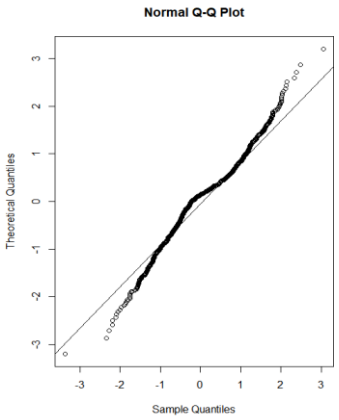
$$Y = \beta_0 + \beta_1(X_1) + \beta_2 \cdot I(X_2 = \text{good\_weather}) + \beta_3 \cdot I(X_3 = \text{Spring}) + \epsilon$$

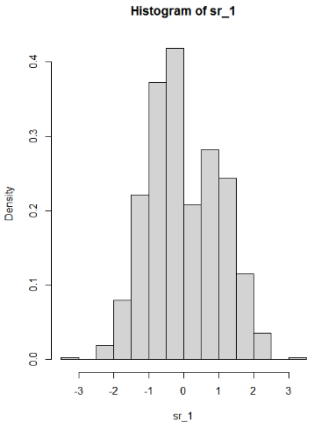
Fitted line (M1):

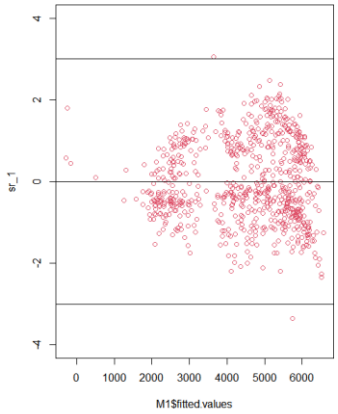
$$\hat{Y} = 108.4 + 4509.5(X_1) + 2573.6\, I(X_2 = \text{good\_weather}) - 1363.7\, I(X_3 = \text{Spring})$$

Q4. Analysis of M_1 for adequacy via residual plot

M1:








```
> which(sr_1>3 | sr_1<(-3))
239 442
239 442
```

```
> cook_m1= cooks.distance(M1)
> which(cook_m1>1)
named integer(0)
```

Model is not adequate.

Normality Violated, SR>3,SR<-3 present,
QQ-plot have lighter left and right tails than line, however histogram slight resembles normal

Equal Variance Violated, Funnel shape present in SR against y-hat scatterplot

Linearity assumption not violated, rectified and explained in Qn.2 earlier.

Outliers: 239 and 442. Influential point: None.

(For future models, will transform response(y) to rectify equal variance assumption and add/drop explanatory variables & try adding interaction variables for normality assumption. No further action for outliers that are non-influential points.)

Q5. Analysis on significance of regressors in M_1

```
> summary(M1)

Call:
lm(formula = y ~ temp + c_new_weathersit + c_new_season, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4633.5  -984.8  -313.3  1155.6  4199.0

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                  108.4      348.8   0.311    0.756
temp                        4509.5      358.0  12.595  < 2e-16 ***
c_new_weathersitgood_weather 2573.6      307.8   8.362 3.14e-16 ***
c_new_seasonspring          -1363.7      151.5  -9.000  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1384 on 727 degrees of freedom
Multiple R-squared:  0.4919,    Adjusted R-squared:  0.4898
F-statistic: 234.6 on 3 and 727 DF,  p-value: < 2.2e-16
```

T-test (regressor Significance):

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$ (Significant)

All regressors are highly significant.

Regressor, Beta-1, temp, p-value < 2e-16, p-value= ~0, very small, reject H0, accept H1, significant

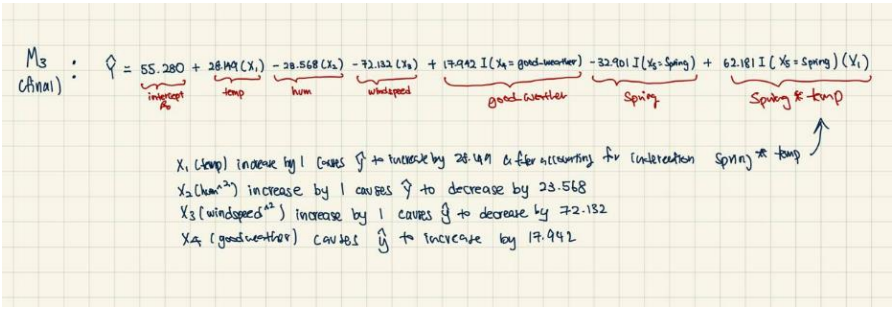
Regressor, Beta-2, weathersit(good_weather), p-value = 3.14e-16, p-value= ~0, very small, reject H0, accept H1, significant

Regressor, Beta-3, season(spring), p-value < 2e-16, p-value= ~0, very small, reject H0, accept H1, significant

In the event that a regressor is non-significant, I will propose to add an interaction variable and if the interaction variable is significant, keep the regressor. Otherwise, drop it from model.

Q6. Improved model M_3 and interpretation

1. Transformed response sqrt(y), funnel shaped mainly gone, equal variance assumption not violated
2. Added interaction variable temp*weathersit(not significant, dropped interaction) and temp*season (Significant, kept interaction).
3. Added hum^10 and windspeed^2, dropped hum^10 and added hum^2 because it let to better normality. Windspeed^2 has better normality than windspeed.
4. Normality improved, QQplot left tail on line while right tail still slightly lighter, Histogram almost symmetrical, SR[-3.5,3], almost normal.
5. Added interaction variables: hum* weathersit , hum*season , windspeed^2*weathersit , windspeed^2*season, but violated equal variance and normality.
6. Conducted anova test, and dropped all insignificant interaction variables added in point 5 above. Kept hum* weathersit . But decided to drop it later because it didn't benefit normality much and added SR>3.
7. Chose point 4 as best model out of others (M3).
   - T-test for all regressors individually have p-value almost 0, hence reject H0 and accept H1, Significant
   - F-test p-value < 2.2e-16, reject H0 and accept H1, indicating that at least one regressor significant, hence model significant.
   - Adjusted R^2 used for multiple regressor model co-efficient of determination, A.R^2= 0.6046 , hence model has goodness of fit which has 60.5% indicating the proportion of response that can be predicted from regressors.
   - Equal variance assumption satisfied, no funnel shape
   - Almost normality, symmetrical histogram, SR[-3.5,3], qqplot left tail on line, right tail slightly lighter.
   - No influential outliers reported



Appendix For final model (M3):