Linear Regression Model

Overview

A data contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

Dataset: https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

We aim to propose a linear regression model for the response variable: Count of total rental bike daily

Exploratory Data Analysis

1. Response variable: Count of total daily rental bikes
   The response variable is quantitative.
   Summary statistics and figures
   Minimum: 22
   1st Quartile: 3152
   Median: 4548
   Mean: 4504
   3rd Quartile: 5956
   Maximum: 8714
   Standard Deviation: 1937.211

   Plots

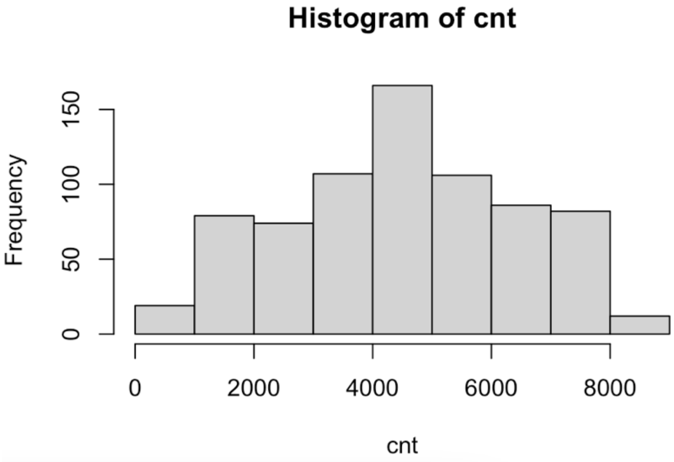


Figure 1: Histogram of response variable

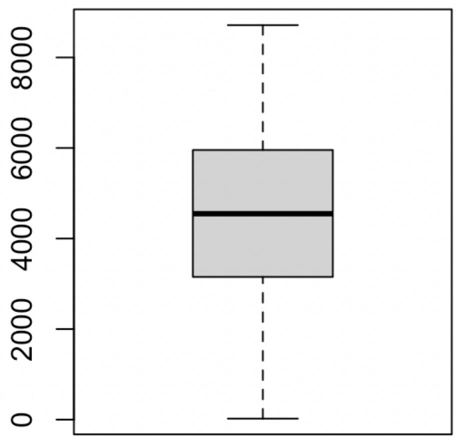


Figure 2: Boxplot of response variable

   As the response variable is quantitative, and the population distribution is symmetric (in this case approximately bell curved as seen from the histogram), and there are no outliers (as seen from the box plot), **it is possible to fit a linear regression model for this response.**

2. Explanatory variable: **temp**
   Correlation(cnt, temp) = 0.627494
   Association between cnt and temp is **positive** and **moderately strong.**
   From scatter plot: temp might be linear and might have a constant variance
   From histogram: multimodal

   Explanatory variable: **hum**
   Correlation(cnt, hum) = -0.1006586
   Association between cnt and hum is **negative** and **weak.**
   From scatter plot: hum might be linear and might have a constant variance
   From histogram: slightly left skewed

   Explanatory variable: **windspeed**
   Correlation(cnt, windspeed) = -0.234545
   Association between cnt and windspeed is **negative** and **weak.**

From scatter plot: windspeed might be linear and might have a constant variance
From histogram: right skewed

Explanatory variable: **season**
From the boxplots, the IQR of all the categories are approximately equal, and thus the spread of data in all categories is approximately equal, however median differs from category to category.

Explanatory variable: **weathersit**
From the boxplots, the IQR of categories 1 and 2 are approximately equal, and the IQR of category 3 is slightly less, thus the spread of data in all categories aren't same, and the median differs from category to category.

Explanatory variable: **workingday**
From the boxplots, the IQR of category 0 is slight more than the IQR of category 1, thus the spread of data for both categories is different, however median of both categories is approximately same.

[For box plots, scatter plots and histograms mentioned above, refer and run the R code attached]

Model

3. Proposed Regressors for the starting model M1 are **temp**, **hum**, **windspeed**, **season**, **workingday**, **weathersit**

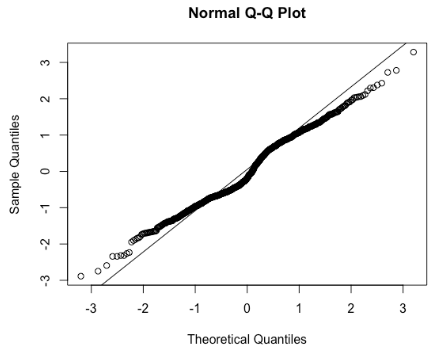
```
> M1 = lm(cnt ~ temp + hum + windspeed + sea + wday + weatsit, data = data)
> summary(M1)

Call:
lm(formula = cnt ~ temp + hum + windspeed + sea + wday + weatsit,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3714.8  -918.5  -243.6  1059.9  4214.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3024.4      349.3   8.659  < 2e-16 ***
temp          6159.1      481.1  12.802  < 2e-16 ***
hum          -2608.2      461.1  -5.656 2.23e-08 ***
windspeed    -3306.0      674.5  -4.901 1.18e-06 ***
sea2           932.3      178.6   5.220 2.34e-07 ***
sea3           483.2      235.6   2.051   0.0406 *
sea4          1499.6      152.4   9.841  < 2e-16 ***
wday1          155.0      103.7   1.496   0.1352
weatsit2      -232.3      127.7  -1.818   0.0694 .
weatsit3     -1929.7      326.8  -5.905 5.43e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1296 on 721 degrees of freedom
Multiple R-squared:  0.5577,	Adjusted R-squared:  0.5522
F-statistic:   101 on 9 and 721 DF,  p-value: < 2.2e-16
```

Figure 3: summary of Model 1

The fitted regression line for M1 is

$$\hat{Y} = 3024.4 + 6159.1 \times temp - 2608.2 \times hum - 3306 \times windspeed + 932.3 \times I(season = 2) + 483.2 \times I(season = 3) + 1499.6 \times I(season = 4) + 155 \times I(workingday = 1) - 232.3 \times I(weathersit = 2) - 1929.7 \times I(weathersit = 3)$$

4.



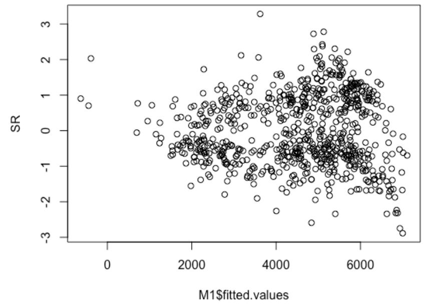
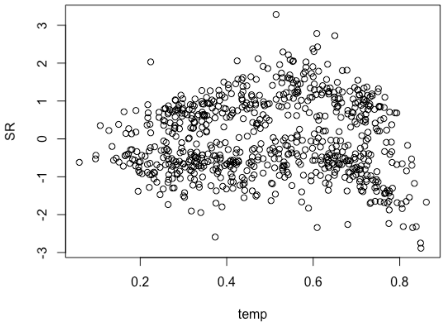
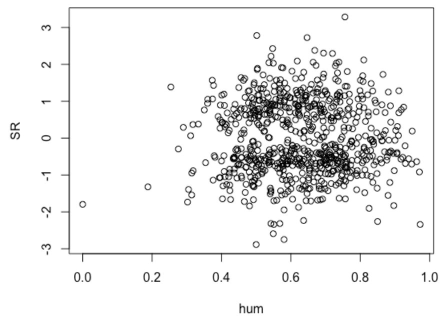


Figure 4: histogram of SR. Figure 5: QQ plot of SR Figure 6: Plot of SR vs fitted values

From the histogram, we can see that the distribution of r_i's isn't completely bell-curved/normal.
The QQ plot is also heavy-tailed and hence, it's safe to say that the normality assumption is violated.

As the scatter plot resembles a funnel shape, we can safely conclude that the constant variance assumption is also violated.





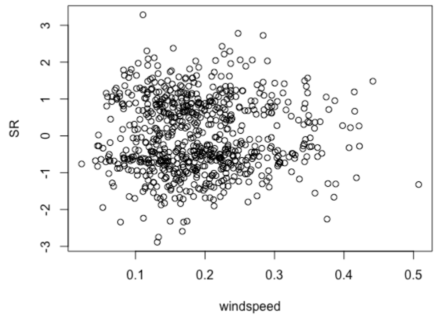


Figure 7: Plot of SR vs temp. Figure 7: Plot of SR vs hum Figure 7: Plot of SR vs windspeed

From the scatterplots, verifying the linearity assumption is difficult. The linearity assumption might or might not be true for hum and windspeed, but is clearly violated for temp.
Thus, model M1 is not adequate, as it violates the assumptions and the adjusted R-squared: 0.5522 is quite low.
There is 1 outlier(442) and no influential point.

```
> which(SR>3 | SR< -3)
442
442
> C = cooks.distance(M1)
> which(C>1)
named integer(0)
```

Figure 8: Find outliers and influential points

5. As P values for quantitative variables temp, hum, and windspeed < 0.001, these regressors are clearly significant[Refer to figure in 3.].
   P value of workingday is greater than 0.001, 0.01, 0.05 and 0.1. So it is clearly insignificant.
   Determining the significance of categorical variables season and weathersit is more complex as some indicator variables are more significant that others and so we use anova() to determine the significance of these 2 variables[Refer and run R code for more details]. Clearly both season and weathersit are significant, as P value < 0.001.
   Proposal: Hence, we might want to drop workingday. However, in case workingday is related to other variables, then we might want to add it along with the interaction term.

6.
#As the scatterplot of SR and fitted values showed a funnel shape, we log the response variable and remove workingday(refer to 5.). We also remove the outlier from the data.
M2 = lm(log(cnt) ~ temp + hum + windspeed + sea + weatsit, data = data)
summary(M2)

#We try to explore the significance of interaction terms, by taking all pairs possible. We also include the workingdays(to test its relation with other variables, refer to 5.).

```
M3 = lm(
  log(cnt) ~ temp + hum + windspeed + sea + wday + weatsit + temp * hum + temp *
    windspeed + temp * wday + temp * weatsit + temp * sea + hum * windspeed + hum *
    wday + hum * weatsit + hum * sea + windspeed * wday + windspeed * weatsit + windspeed *
    sea + wday * weatsit + wday * sea + weatsit * sea,
  data = data
)
summary(M3)
```

#Final Model

#We remove insignificant interaction terms(use intuition and trial and error to see change in adjusted R^2 value and observe changes in the corresponding plots) and also remove the outliers of the new model.

```
M4 = lm(log(cnt) ~ temp + hum + windspeed + sea + wday + weatsit + temp * sea + hum * weatsit
  + windspeed * weatsit + wday * weatsit,
  data = data
)
summary(M4)
```

[Refer and run the R code for further details.]

The fitted regression line is

$$
\begin{aligned}
\widehat{log(cnt)} = 7.14 + 3.45(temp) - 0.57(hum) - \\
& 0.48(windspeed) + 0.87(I(season=2)) + 2.51(I(season=3)) + \\
& 1(I(season=4)) + 0.08(I(workingday=1)) + 0.63(I(weathersit=2)) + \\
& 0.16(I(weathersit=3)) - 1.84(temp \times I(season=2)) - 4.34(temp \times I(season=3)) - \\
& 1.7(temp \times I(season=4)) - 0.68(hum \times I(weathersit=2)) + 0.66(hum \times I(weathersit=3)) - \\
& 0.86(windspeed \times I(weathersit=2)) - 4.75(windspeed \times I(weathersit=3)) - \\
& 0.09(I(workingday=1) \times I(weathersit=2)) - 0.65(I(workingday=1) \times I(weathersit=3))
\end{aligned}
$$

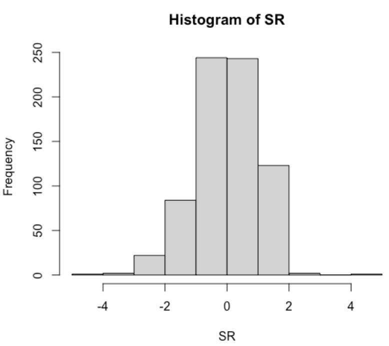


Figure 9: histogram of SR-M4

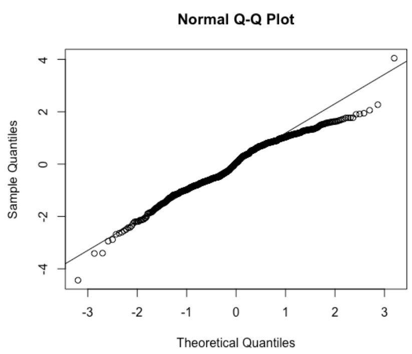


Figure 10: QQ plot of SR-M4.

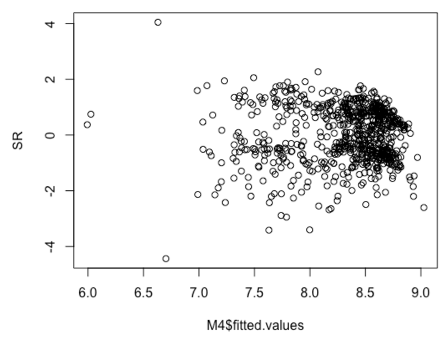


Figure 11: SR vs fitted values-M4

As the histogram and QQ Plot are much closer to a normal distribution, and the adjusted R^2 value increases from 0.5522 to 0.6921, and the scatter plot is approximately randomly distributed about 0 and mostly lies in the interval [-3,3], we can say that even though the model is quite complex, it somewhat depicts the relationship between the explanatory and the response variables in the best possible way. [Refer and run R code for scatter plots between the regressors and the SR].

To interpret the effect of each variable on the response variable, we need to consider different cases for the categorical variables which will give different values of the intercepts and might also change the slope(as interactive terms are involved). There can be 24 such permutations of different categories of season, weathersit and workingday(4*3*2). For all these 24 cases, we get different slopes and intercepts and a change in 1 unit of the explanatory variable will lead to the log of the response variable changing by a value equal to the slope of the corresponding explanatory variable, for that particular case.

In general, an increase in temp would result in an increase log(cnt) and hence cnt, and an increase in hum or windspeed will result in a decrease in log(cnt) and hence cnt. This might not be always true(as we also need to consider the interaction terms). The exact increase or decrease can be easily found out by finding the slope of the explanatory variables after putting in the values of the categorical regressors and obtaining the equation for that particular combination of categorical variables.

Appendix

R code

```
rm(list = ls())
setwd("~/Downloads")
data = read.csv('day.csv')
#data = data[-c(442), ] #Removing outliers for Model 1
data = data[-c(2, 27, 65, 66, 239, 249, 250, 328, 668), ] #Removing outliers for the final model
attach(data)

#Declaring categorical variables
sea = factor(season)
wday = factor(workingday)
weatsit = factor(weathersit)

#Summarizing response variable to show if its suitable to fit a linear
#regression model or not
summary(cnt)
sd(cnt)
hist(cnt)
boxplot(cnt)

#Checking association between response and explanatory variables
cor(temp, cnt)
plot(temp, cnt)
hist(temp)

cor(hum, cnt)
plot(hum, cnt)
hist(hum)

cor(windspeed, cnt)
plot(windspeed, cnt)
hist(windspeed)

table(sea)
barplot(table(sea))
boxplot(cnt ~ season)

table(weatsit)
barplot(table(weatsit))
boxplot(cnt ~ weatsit)

table(wday)
barplot(table(wday))
```

```
boxplot(cnt ~ wday)

#Propose an initial model
M1 = lm(cnt ~ temp + hum + windspeed + sea + wday + weatsit, data = data)
summary(M1)

#Checking assumptions using residual plots to determine the
#adequacy of the model
SR = rstandard(M1)
hist(SR)
qqnorm(SR)
qqline(SR)
plot(M1$fitted.values, SR)

plot(temp, SR)
plot(hum, SR)
plot(windspeed, SR)

#determining outliers and influential points
which(SR > 3 | SR < -3)
C = cooks.distance(M1)
which(C > 1)

#testing the significance of categorical variables sea and weatsit
anova(lm(cnt ~ temp + hum + windspeed + wday + weatsit + sea))
anova(lm(cnt ~ temp + hum + windspeed + sea + wday + weatsit))

#As the scatterplot of SR and fitted values showed a funnel shape,
#we log the response variable and remove workingday(refer to 5.).
#We also remove the outlier from the data.
M2 = lm(log(cnt) ~ temp + hum + windspeed + sea + weatsit, data = data)
summary(M2)

#We try to explore the significance of interaction terms, by taking
#all pairs possible. We also include the workingdays(to test its
#relation with other variables, refer to 5.).
M3 = lm(
  log(cnt) ~ temp + hum + windspeed + sea + wday + weatsit + temp * hum + temp *
    windspeed + temp * wday + temp * weatsit + temp * sea + hum * windspeed + hum *
    wday + hum * weatsit + hum * sea + windspeed * wday + windspeed * weatsit + windspeed *
    sea + wday * weatsit + wday * sea + weatsit * sea,
  data = data
)
summary(M3)

#Final Model
#We remove insignificant interaction terms(use intuition and trial
#and error to see change in adjusted R^2 value and observe changes
#in the corresponding plots) and also remove the outliers of the new model.
M4 = lm(
  log(cnt) ~ temp + hum + windspeed + sea + wday + weatsit + temp * sea + hum * weatsit
  + windspeed * weatsit + wday * weatsit,
  data = data
)
summary(M4)
```

```r
#Determining outliers and influential points and observing changes on
#removing them(Check line 5)
SR = rstandard(M4)
which(SR > 3 | SR < -3)

C = cooks.distance(M4)
which(C > 1)

#Checking assumptions using residual plots to determine the
#adequacy of the model
SR = rstandard(M4)
hist(SR)
qqnorm(SR)
qqline(SR)
plot(M4$fitted.values, SR)

plot(temp, SR)
plot(hum, SR)
plot(windspeed, SR)
```