Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Predicting Demand for Rental Bikes: Exploratory Data Analysis

1

About This Lesson

2

Predicting Demand for Rental Bikes

Bike sharing systems are of great interest due to their important role in traffic management.
Dataset: Historical data for years 2011-2012 for the bike sharing system in Washington D.C.
Data Source: UCI Machine Learning Repository

***Acknowledgement**: This example was prepared with support from students in the Masters of Analytics program, including Naman Arora, Puneeth Banisetti, Mani Chandana Chalasani, Joseph (Mike) Tritchler and Kevin West*

Georgia Tech

3

Response & Predicting Variables

The response variable is:

Y (*Cnt*): Total bikes rented by both casual & registered users together

The qualitative predicting variables are:

Season: Season which the observation is made (1 = Winter, 2 = Spring, 3 = Summer, 4 = Fall)
Yr: Year on which the observation is made
Mnth: Month on which the observation is made
Hr: Day on which the observation is made (0 through 23)
Holiday: Indictor of a public holiday or not (1 = public holiday, 0 = not a public holiday)
Weekday: Day of week (0 through 6)
Weathersit: Weather condition (1 = Clear, Few clouds, Partly cloudy, Partly cloudy, 2 = Mist & Cloudy, Mist & Broken clouds, Mist & Few clouds, Mist, 3 = Snow, Rain, Thunderstorm & Scattered clouds, Ice Pallets & Fog)

The quantitative predicting variables are:

Temp: Normalized temperature in Celsius
Atemp: Normalized feeling temperature in Celsius
Hum: Normalized humidity
Windspeed: Normalized wind speed

Georgia Tech

4

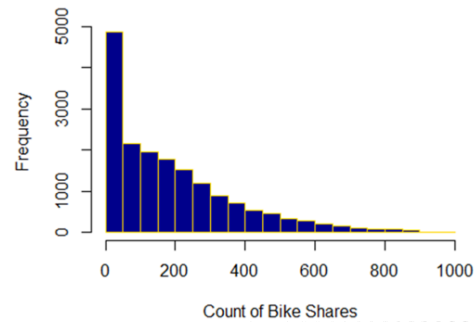
Exploratory Data Analysis in R

Read data using read.csv

```
data<-read.csv("Bikes.csv")
dim(data)[1] # how many observations?
[1] 17379
```

Test initial intuitions/assumptions on the behavior of the data

```
hist(data$cnt,
     main="",
     xlab="Count of Bike Shares",
     border="gold",
     col="darkblue")
```



The frequency of zero bike shares is high, which skews the demand data.

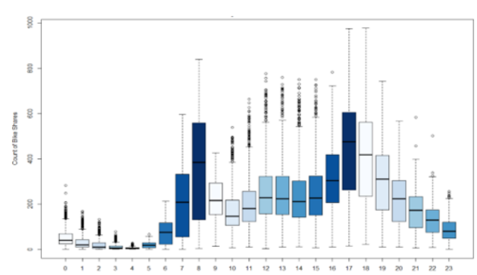
Georgia Tech

5

Exploratory Data Analysis in R (cont'd)

Evaluate intuitions/assumptions on the behavior of the data and understand patterns

```
boxplot(cnt~hr,
     main="",
     xlab="Hour",
     ylab="Count of Bike Shares",
     col=blues9,
     data=data)
```

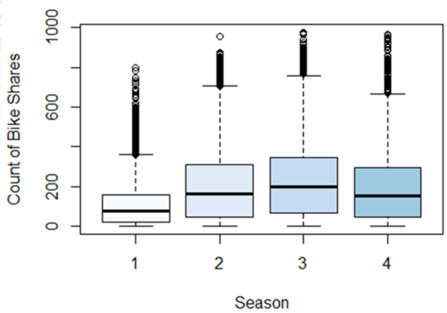


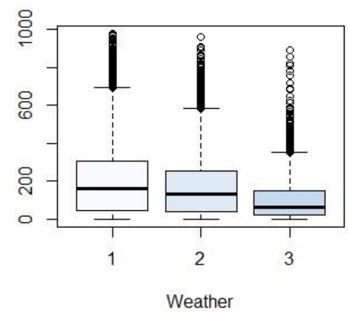
The number of bike shares between hour 0 and hour 6 is low. The majority activity as expected is focused between hour 7 and hour 23, peaking at hour 8 and hour 17.

Georgia Tech

6

Exploratory Data Analysis in R (cont'd)





The number of bikes rented during winter are the lowest.

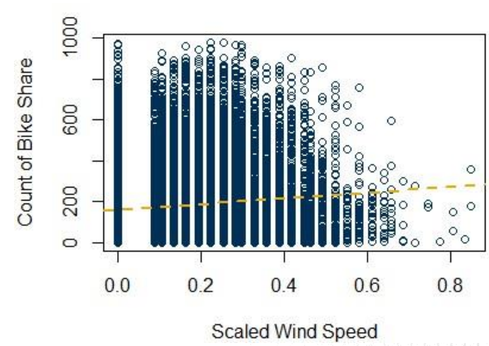
The number of bikes decreases as the weather becomes unfavorable.

7

Exploratory Data Analysis in R (cont'd)

plot(data$windspeed,
data$cnt,
xlab='Scaled Wind Speed',
ylab='Count of Bike Share',
main='', col="darkblue")

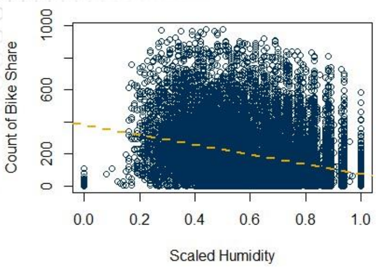
abline(lm(cnt~windspeed, data=data),
col=buzzgold,
lty=2, lwd=2)



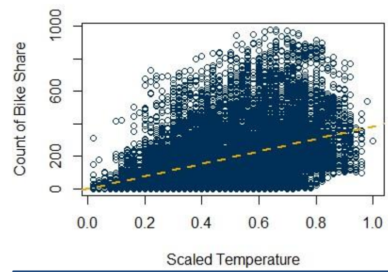
The count of rental bikes seems to decrease as windspeed increases.

8

Exploratory Data Analysis in R (cont’d)



The count of rental bikes seems to decrease as humidity increases although the demand varies within similar ranges at varying humidity levels.



The count of rental bikes seems to increase as temperature increases however with much wider variability at larger temperature levels.

Georgia Tech

9

Preparing the Data

```
## Divide data into train and test data
# Set seed for reproducibility
set.seed(9)
# Test Train split
sample_size = floor(0.8*nrow(data))
picked = sample(seq_len(nrow(data)),size = sample_size)

# Remove irrelevant columns from training data
train = data[picked,]
train <- train[-c(1,2,9,15,16)]

## Converting the numerical cateogrical variables to predictors
train$season = as.factor(train$season)
train$yr = as.factor(train$yr)
train$mnth = as.factor(train$mnth)
train$hr = as.factor(train$hr)
train$holiday = as.factor(train$holiday)
train$weekday = as.factor(train$weekday)
train$weathersit = as.factor(train$weathersit)
```

Georgia Tech

10

Fitting the Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-79.4201	7.3917	-10.744	< 2e-16 ***
season2	41.7616	5.3578	7.794	6.93e-15 ***
season3	33.3129	6.3740	5.226	1.75e-07 ***
season4	67.2826	5.4338	12.382	< 2e-16 ***
yr1	86.2941	1.7468	49.401	< 2e-16 ***
mnth2	0.7558	4.3763	0.173	0.862893
mnth3	12.2441	4.9101	2.494	0.012655 *
mnth4	3.5236	7.2709	0.485	0.627950
mnth5	20.2297	7.7696	2.604	0.009233 **
mnth6	2.6876	7.9759	0.337	0.736150
mnth7	-10.7018	8.9548	-1.195	0.232069
mnth8	11.4522	8.7278	1.312	0.189491
mnth9	31.6884	7.7488	4.089	4.35e-05 ***
mnth10	18.1808	7.1948	2.527	0.011517 *
mnth11	-9.8396	6.9461	-1.417	0.156635
mnth12	-9.4086	5.5098	-1.708	0.087728 .
hr1	-21.5064	5.9532	-3.613	0.000304 ***

 # Applying multiple linear regression model
model1 = lm(cnt ~ .,data=train)
summary(model1)

In the full output there are 51 predictor rows in addition to the intercept.

11

Statistical Significance

 # Applying multiple linear regression model
model1 = lm(cnt ~ .,data=train)
summary(model1)

 # Find insignificant values
which(summary(model1)$coeff[,4]>0.05)

mnth2	mnth4	mnth6	mnth7	mnth8	mnth11	mnth12
6	8	10	11	12	15	16

Statistically insignificant variables at 0.05 significance level:
- Month-2, month-4, month-6, month-7, month-8, month-11, month-12 are not statistically different from month-1 (baseline)

12

Summary

Georgia Tech

13