Regression Analysis

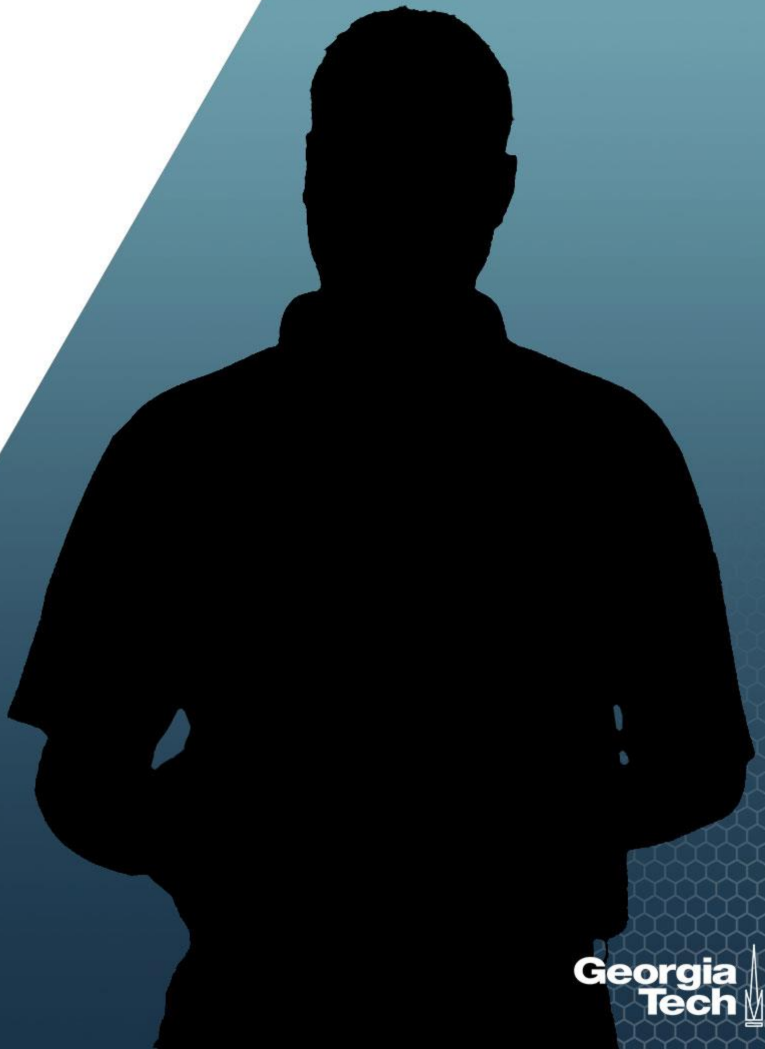
Multiple Linear Regression

Nicoleta Serban, Ph.D.

Professor

School of Industrial and Systems Engineering

Predicting Demand for Rental Bikes: Exploratory Data Analysis



About This Lesson

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*

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

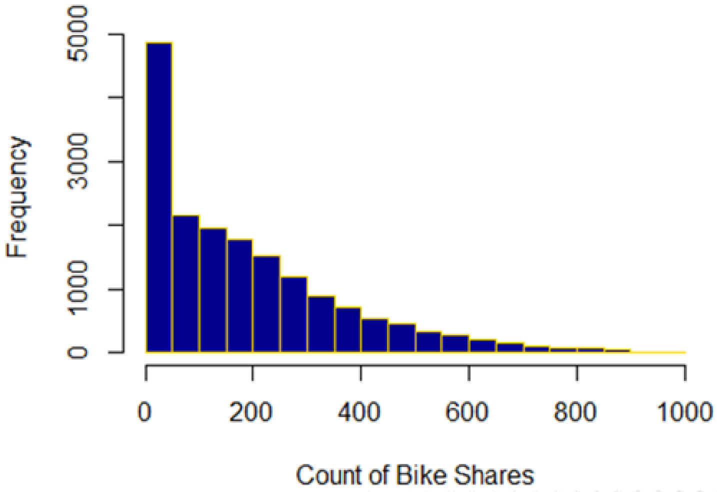
Atemp: Normalized feeling temperature in Celsius

Hum: Normalized humidity

Windspeed: Normalized wind speed

Exploratory Data Analysis in R

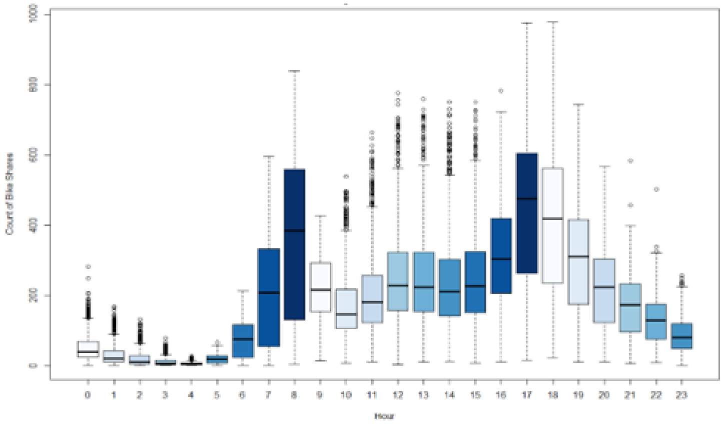
```
## Read data using read.csv
data<-read.csv(“Bikes.csv")
dim(data)[1] # how many observations?
[1] 17379
## Test initial intuitions/assumptions on the
behavior of the data
hist(data$cnt,
        main="",
        xlab="Count of Bike Shares",
        border="gold",
        col="darkblue")
```



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

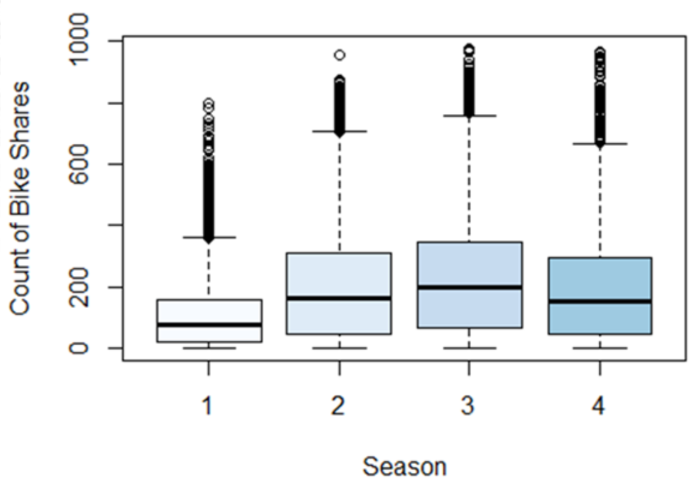
Exploratory Data Analysis in R (cont’d)

```
## 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)
```

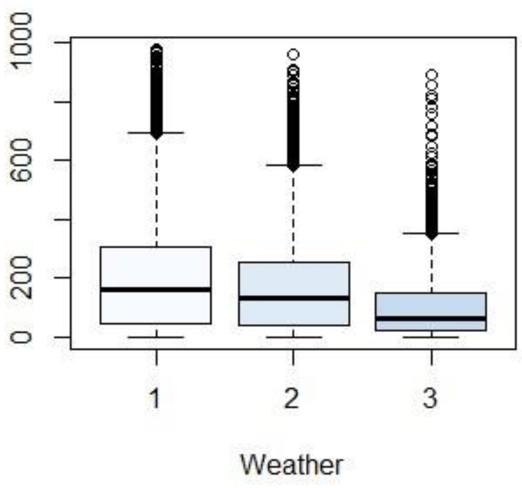


The number of bike shares between hour 0 and hour 6 is low. The majority activity as expected is focused between hour 7 and hour 23, peaking at hour 8 and hour 17.

Exploratory Data Analysis in R (cont’d)



The number of bikes rented during winter are the lowest.

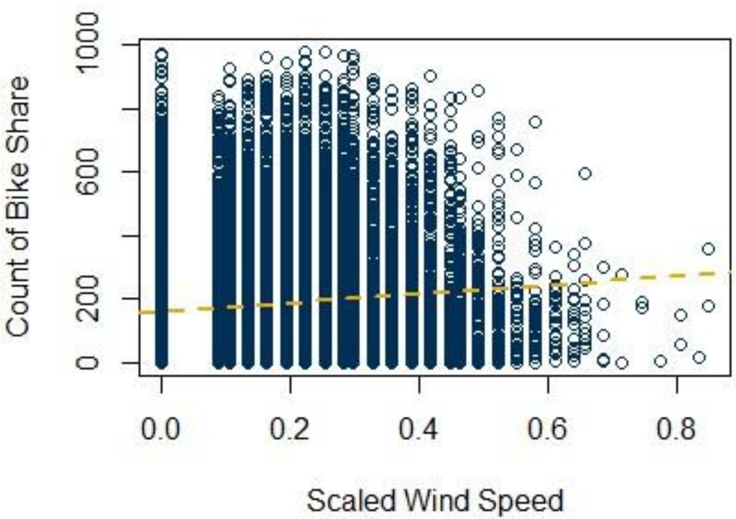


The number of bikes decreases as the weather becomes unfavorable.

Exploratory Data Analysis in R (cont’d)

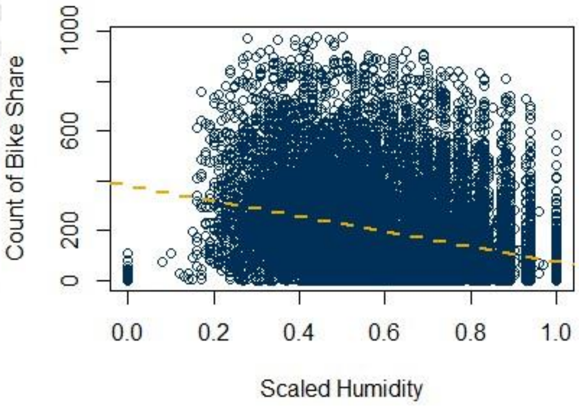
```
plot(data$windspeed,
        data$cnt,
        xlab='Scaled Wind Speed',
        ylab='Count of Bike Share',
        main=‘’,  col="darkblue")

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

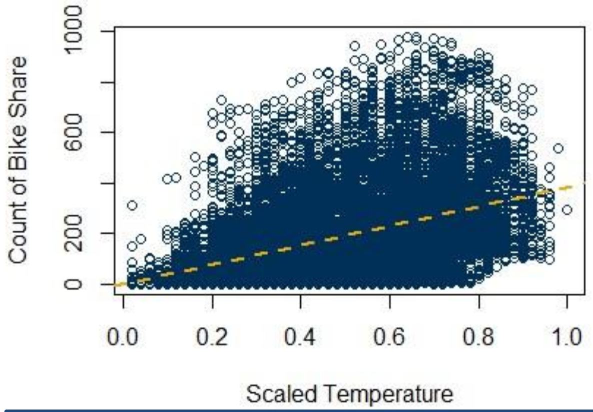


The count of rental bikes seems to decrease as windspeed increases.

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.

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)
```

Fitting the Regression Model

```
              Estimate   Std. Error    t value       Pr(>|t|)
(Intercept)   -79.4356       7.4390    -10.678   < 2e-16 ***
season2        34.9268       5.4110      6.455   1.12e-10 ***
season3        27.0055       6.4438      4.191   2.80e-05 ***
season4        65.3435       5.4690     11.948   < 2e-16 ***
yr1            85.3415       1.7487     48.804   < 2e-16 ***
mnth2           4.1666       4.3853      0.950       0.342060
mnth3          16.4733       4.9267      3.344    0.000829***
mnth4          12.5834       7.3038      1.723     0.084936 .
mnth5          26.4616       7.8357      3.377   0.000735 ***
mnth6          11.5056       8.0535      1.429       0.153131
mnth7          -7.8872       9.0547     -0.871       0.383736
⋮
---
```

```
# 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.

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)

Summary