# LINEAR 7 - Data set: BIKE SHARING

## INTRODUZIONE

Il data set è composto da 731 osservazioni raccolte con cadenza giornaliera riguardanti il numero di biciclette affittate a Washington. Le variabili raccolte sono le seguenti:

1. DATE: giorno della rilevazione
2. SEASON, YEAR, MONTH, WEEKDAY: stagione, anno, mese e giorno della settimana della rilevazione
3. HOLIDAY: dummy per indicare se il giorno è festivo oppure no
4. WORKINGDAY: dummy per indicare se il giorno considerato è lavorativo o meno
5. WEATHERSIT: condizioni climatiche con 4 modalità (cielo chiaro, nebbia o nuvoloso, pioggia, temporale)
6. TEMP: temperatura media
7. ATEMP: temperatura media percepita
8. HUM: valore dell'unidità normalizzato
9. WINDSPEED: velocità massima del vento normalizzata

Analisi proposte:

1. Statistiche descrittive
2. Regressione lineare

```r
#-- R CODE

library(pander)
library(car)
library(olsrr)
library(systemfit)
library(het.test)
panderOptions('knitr.auto.asis', FALSE)

#-- White test function
white.test <- function(lmod,data=d){
  u2 <- lmod$residuals^2
  y <- fitted(lmod)
  Ru2 <- summary(lm(u2 ~ y + I(y^2)))$r.squared
  LM <- nrow(data)*Ru2
  p.value <- 1-pchisq(LM, 2)
  data.frame("Test statistic"=LM,"P value"=p.value)
}

#-- funzione per ottenere osservazioni outlier univariate
FIND_EXTREME_OBSERVARION <- function(x,sd_factor=2){
  which(x>mean(x)+sd_factor*sd(x) | x<mean(x)-sd_factor*sd(x))
}

#-- import dei dati
ABSOLUTE_PATH <- "C:\\Users\\sbarberis\\Dropbox\\MODELLI STATISTICI"
d <- read.csv(paste0(ABSOLUTE_PATH,"\\F. Esercizi(22) copia\\4.tutto(4)\\1.tutto\\Day.csv"),sep=",")

#-- vettore di variabili numeriche presenti nei dati
```

```r
VAR_NUMERIC <- c("temp","atemp","hum","windspeed","cnt")

#-- print delle prime 6 righe del dataset
pander(head(d),big.mark=",")
```

Table 1: Table continues below

| instant | dteday | season | yr | mnth | holiday | weekday | workingday |
|---|---|---|---|---|---|---|---|
| 1 | 2011-01-01 | 1 | 0 | 1 | 0 | 6 | 0 |
| 2 | 2011-01-02 | 1 | 0 | 1 | 0 | 0 | 0 |
| 3 | 2011-01-03 | 1 | 0 | 1 | 0 | 1 | 1 |
| 4 | 2011-01-04 | 1 | 0 | 1 | 0 | 2 | 1 |
| 5 | 2011-01-05 | 1 | 0 | 1 | 0 | 3 | 1 |
| 6 | 2011-01-06 | 1 | 0 | 1 | 0 | 4 | 1 |

| weathersit | temp | atemp | hum | windspeed | casual | registered | cnt |
|---|---|---|---|---|---|---|---|
| 2 | 0.3442 | 0.3636 | 0.8058 | 0.1604 | 331 | 654 | 985 |
| 2 | 0.3635 | 0.3537 | 0.6961 | 0.2485 | 131 | 670 | 801 |
| 1 | 0.1964 | 0.1894 | 0.4373 | 0.2483 | 120 | 1,229 | 1,349 |
| 1 | 0.2 | 0.2121 | 0.5904 | 0.1603 | 108 | 1,454 | 1,562 |
| 1 | 0.227 | 0.2293 | 0.437 | 0.1869 | 82 | 1,518 | 1,600 |
| 1 | 0.2043 | 0.2332 | 0.5183 | 0.08957 | 88 | 1,518 | 1,606 |

## STATISTICHE DESCRITTIVE

Si propongono la matrice di correlazione tra le variabili e alcune descrittive di base. Si analizza l'influenza delle variabili climatiche sull'affitto di biciclette a Washington nel periodo 2011-2012.

```r
#-- R CODE
pander(summary(d[,VAR_NUMERIC]),big.mark=",") #-- statistiche descrittive
```

Table 3: Table continues below

| temp | atemp | hum | windspeed |
|---|---|---|---|
| Min. :0.05913 | Min. :0.07907 | Min. :0.0000 | Min. :0.02239 |
| 1st Qu.:0.33708 | 1st Qu.:0.33784 | 1st Qu.:0.5200 | 1st Qu.:0.13495 |
| Median :0.49833 | Median :0.48673 | Median :0.6267 | Median :0.18097 |
| Mean :0.49538 | Mean :0.47435 | Mean :0.6279 | Mean :0.19049 |
| 3rd Qu.:0.65542 | 3rd Qu.:0.60860 | 3rd Qu.:0.7302 | 3rd Qu.:0.23321 |
| Max. :0.86167 | Max. :0.84090 | Max. :0.9725 | Max. :0.50746 |

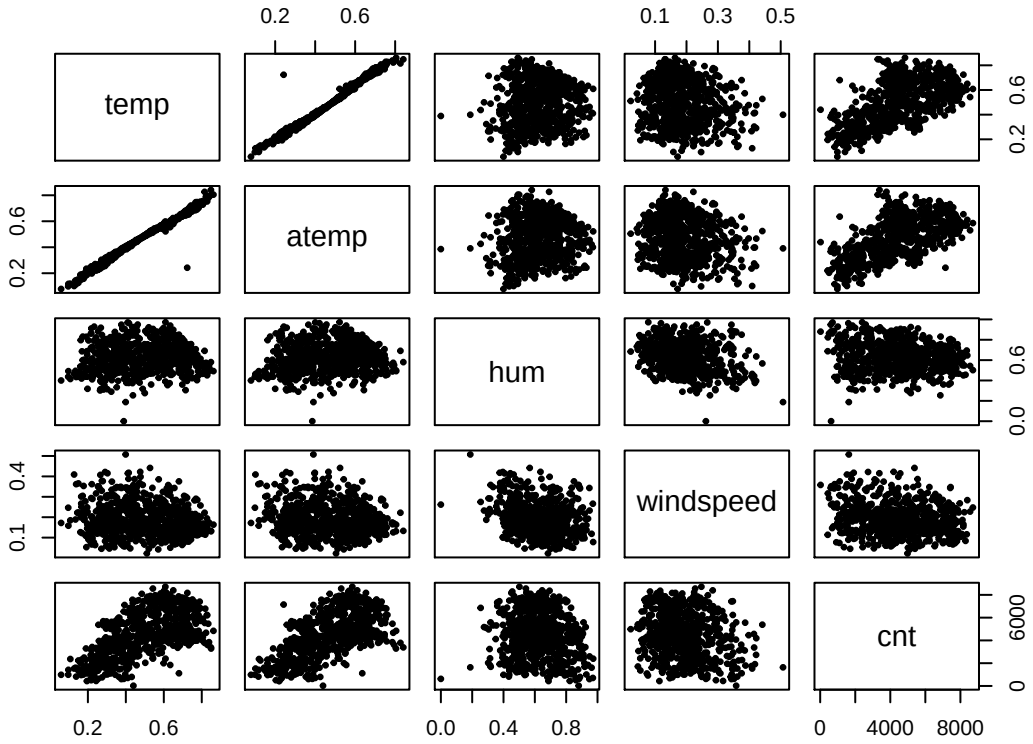
| cnt |
|---|
| Min. : 22 |
| 1st Qu.:3152 |
| Median :4548 |

| cnt |
|---|
| Mean :4504 |
| 3rd Qu.:5956 |
| Max. :8714 |

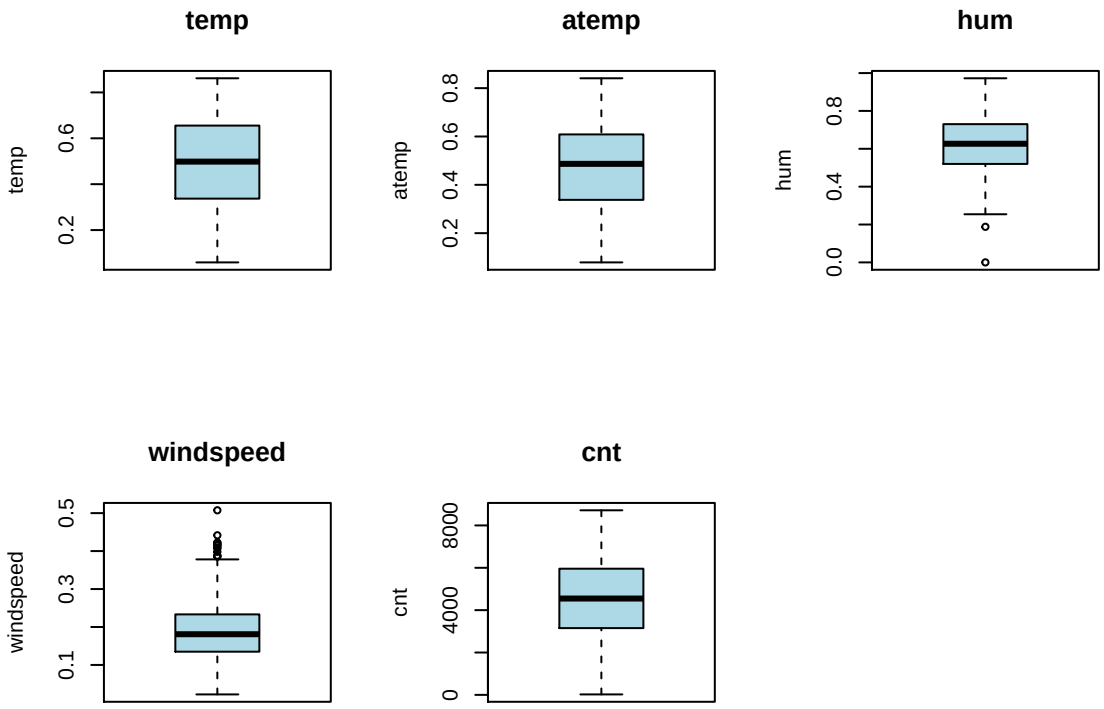
```
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

| | temp | atemp | hum | windspeed | cnt |
|---|---|---|---|---|---|
| **temp** | 1 | 0.9917 | 0.127 | -0.1579 | 0.6275 |
| **atemp** | 0.9917 | 1 | 0.14 | -0.1836 | 0.6311 |
| **hum** | 0.127 | 0.14 | 1 | -0.2485 | -0.1007 |
| **windspeed** | -0.1579 | -0.1836 | -0.2485 | 1 | -0.2345 |
| **cnt** | 0.6275 | 0.6311 | -0.1007 | -0.2345 | 1 |

```
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```



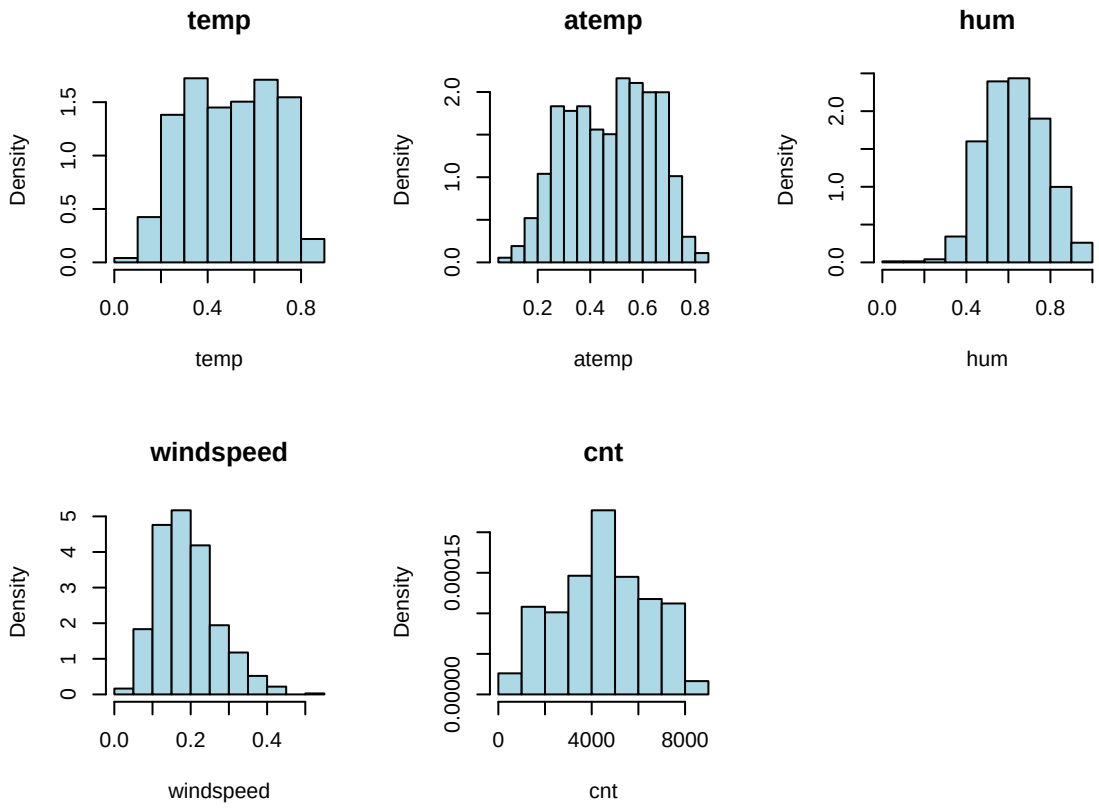

```
par(mfrow=c(2,3))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
par(mfrow=c(2,3))
```



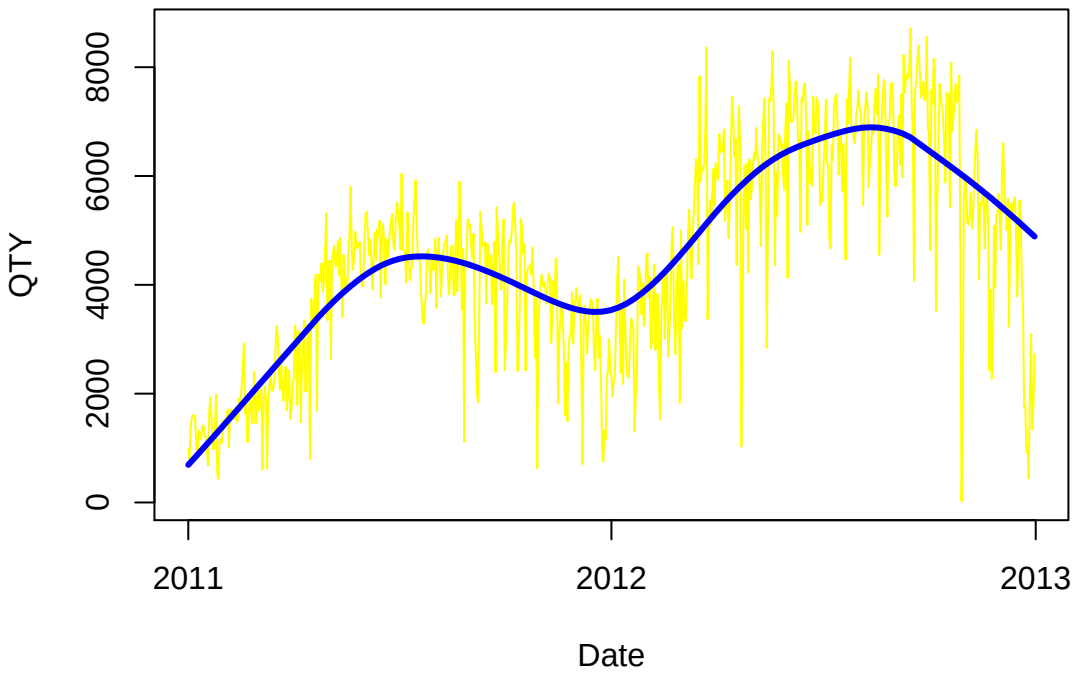

```
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}

d$dteday <- as.Date(paste(d$dteday))

aggr <- aggregate(d$cnt,list(day=d$dteday),sum)
aggr$day <- as.Date(aggr$day)
par(mfrow=c(1,1))
```




```
plot(aggr[,1],aggr[,2],type="l",xlab="Date",ylab="QTY",col="yellow")
lines(aggr[,1],lowess(aggr[,2],f = .3)$y,col="blue",lwd=3)
```




La serie storica denota dei picchi nei mesi estivi e quindi stagionalità e un trend crescente. Sarebbe da verificare anche la omoschedastictà degli errori ma analizziamo solo la multicollinerità e gli outlier. Prima di impostare il modello lineare si analizza la correlazione far le variabili esplicative.

Le variabili esplicative inizialmente utilizzate sono "temp", "atemp", "hum", "windspeed". Si nota la fortissima collinearità fra "temp" e "atemp", come è naturale essendo una la temperatura registrata e l'altra la temperatura percepita. Entrambe sono correlate positivamente con "cnt": la gente usa più la bicicletta tanto più fa caldo.

A questo punto si propone un primo modello lineare

## REGRESSIONE

```r
#-- R CODE
mod1 <- lm(cnt~temp + atemp + hum + windspeed,d)
pander(summary(mod1),big.mark=",")
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | 3,860 | 355.4 | 10.86 | 1.427e-25 |
| **temp** | 2,112 | 2,282 | 0.9253 | 0.3551 |
| **atemp** | 5,139 | 2,577 | 1.994 | 0.0465 |
| **hum** | -3,149 | 384 | -8.201 | 1.082e-15 |
| **windspeed** | -4,529 | 721.1 | -6.28 | 5.815e-10 |

Table 7: Fitting linear model: cnt ~ temp + atemp + hum + windspeed

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 731 | 1422 | 0.4638 | 0.4609 |

```
pander(anova(mod1),big.mark=",")
```

Table 8: Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **temp** | 1 | 1.079e+09 | 1.079e+09 | 533.2 | 7.049e-89 |
| **atemp** | 1 | 12,774,317 | 12,774,317 | 6.314 | 0.0122 |
| **hum** | 1 | 99,420,281 | 99,420,281 | 49.14 | 5.457e-12 |
| **windspeed** | 1 | 79,801,308 | 79,801,308 | 39.44 | 5.815e-10 |
| **Residuals** | 726 | 1.469e+09 | 2,023,211 | NA | NA |

```
pander(white.test(mod1),big.mark=",")
```

| Test.statistic | P.value |
|---|---|
| 30.55 | 2.32e-07 |

```
pander(dwtest(mod1),big.mark=",")
```

Table 10: Durbin-Watson test: `mod1`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 0.4098 | 9.24e-104 * * * | true autocorrelation is greater than 0 |

Dall'F value si verifica che è respinta l'ipotesi nulla che il modello nel suo complesso non spieghi per nulla la variabile esplicativa affitto delle biciclette. Il valore dell'$R^2$ che si colloca su valori medi ci dice che la capacità di spiegazione non è elevatissima ma neanche scarsa. I parametri significativi per cui l'ipotesi nulla viene respinta sono "hum" e "windspeed" e in parte "atemp". Considerando il segno le biciclette sono affittate tanto più quanto più si percepisce un' alta temperatura e quanto meno è umido e c'e vento.