MA575 Lab6

Shuqiang Chen, Hengchang Hu, Haikuo Lu, Chi Chen, Shreya Gupta, Azar Ghahari

10.24.2018

```
# 1. Choose the response variable (Y) and one covariate (X).
#Please put some though for your response and covariate variable selection.
# import dataset day.csv
BikeSharingInDay <- read.csv(file = "C:/Users/hugo1/Documents/ma575/Proj/day.csv")

# Sneak peak at the data
head(BikeSharingInDay)
```

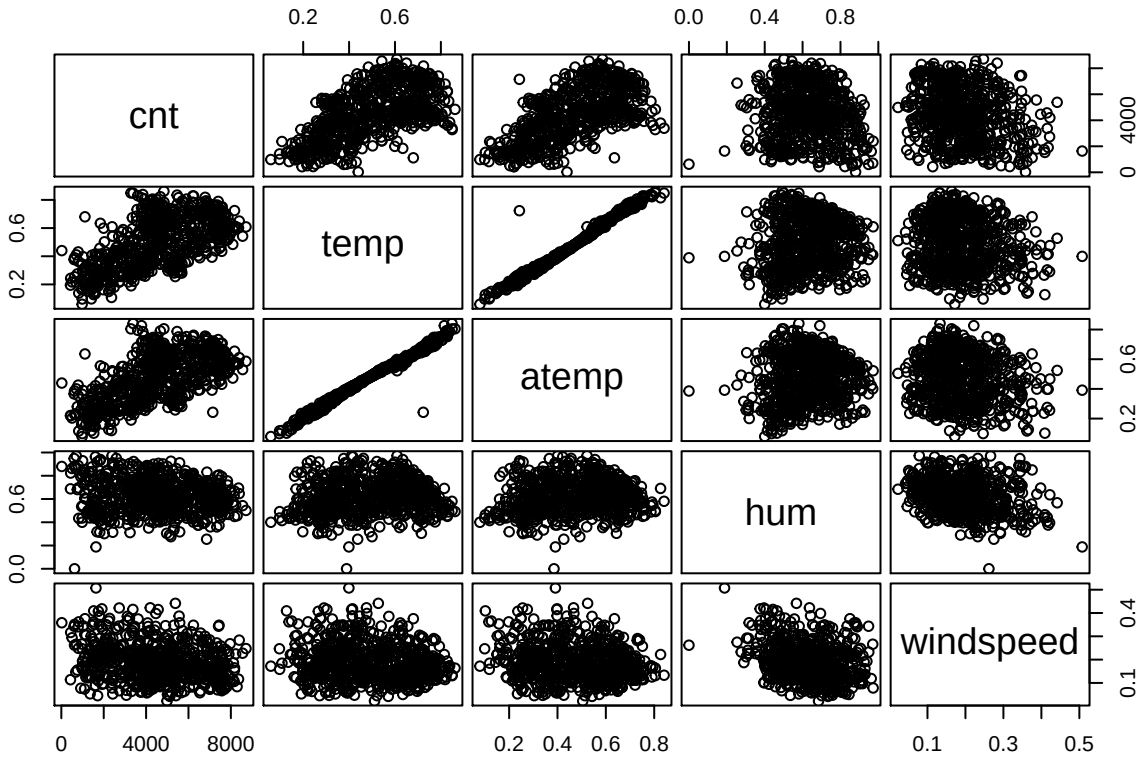
```
##   instant     dteday season yr mnth holiday weekday workingday weathersit
## 1       1 2011-01-01      1  0    1       0       6          0          2
## 2       2 2011-01-02      1  0    1       0       0          0          2
## 3       3 2011-01-03      1  0    1       0       1          1          1
## 4       4 2011-01-04      1  0    1       0       2          1          1
## 5       5 2011-01-05      1  0    1       0       3          1          1
## 6       6 2011-01-06      1  0    1       0       4          1          1
##       temp    atemp      hum windspeed casual registered  cnt
## 1 0.344167 0.363625 0.805833 0.1604460    331        654  985
## 2 0.363478 0.353739 0.696087 0.2485390    131        670  801
## 3 0.196364 0.189405 0.437273 0.2483090    120       1229 1349
## 4 0.200000 0.212122 0.590435 0.1602960    108       1454 1562
## 5 0.226957 0.229270 0.436957 0.1869000     82       1518 1600
## 6 0.204348 0.233209 0.518261 0.0895652     88       1518 1606
```

```
# choose counts of total rental bikes as response variable (Y)
# hypothesis: temp(X1),atemp(X2),hum(X3),windspeed(X4),holiday(X5),weathersit(X6),year(X7) have impact
cnt <- BikeSharingInDay$cnt

# choose covariate (X1~X7)
temp <- BikeSharingInDay$temp
atemp <- BikeSharingInDay$atemp
hum <- BikeSharingInDay$hum
windspeed <- BikeSharingInDay$windspeed
holiday <- BikeSharingInDay$holiday
weathersit <- BikeSharingInDay$weathersit
year <- BikeSharingInDay$yr
```

We choose these four covariates (temp,atemp,hum,windspeed) because they are all numeric variables and easy to interpret. And we choose other three covariates (holiday,weathersit,year) because we thought it may affect the using of bikes.

```
# 2. Plot Y VS. X1-x4 (i.e. a scatterplot) from the data.
# Plot scatter matrix
pairs(~cnt+temp+atemp+hum+windspeed,gap=0.4)
```




```
m.mls <- lm(cnt ~ temp + atemp + hum + windspeed + holiday + weathersit + year)
summary(m.mls)
```

```
##
## Call:
## lm(formula = cnt ~ temp + atemp + hum + windspeed + holiday +
##     weathersit + year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3531.3  -601.1    42.2   681.6  2644.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2334.18     251.13   9.295  < 2e-16 ***
## temp         2258.50    1572.65   1.436  0.15141
## atemp        4229.96    1778.84   2.378  0.01767 *
## hum          -760.59     346.82  -2.193  0.02862 *
## windspeed   -3405.28     506.84  -6.719 3.72e-11 ***
## holiday      -703.01     217.19  -3.237  0.00126 **
## weathersit   -591.90      87.42  -6.771 2.65e-11 ***
## year         2031.59      73.13  27.782  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 979 on 723 degrees of freedom
```

```
## Multiple R-squared:  0.7471, Adjusted R-squared:  0.7446
## F-statistic: 305.1 on 7 and 723 DF,  p-value: < 2.2e-16
```

```
# correlation matrix
X <- cbind(cnt, temp, atemp, hum, windspeed, holiday, weathersit, year)
c <- cor(X)
round(c,3)
```

```
##               cnt   temp  atemp    hum windspeed holiday weathersit   year
## cnt         1.000  0.627  0.631 -0.101    -0.235  -0.068     -0.297  0.567
## temp        0.627  1.000  0.992  0.127    -0.158  -0.029     -0.121  0.048
## atemp       0.631  0.992  1.000  0.140    -0.184  -0.033     -0.122  0.046
## hum        -0.101  0.127  0.140  1.000    -0.248  -0.016      0.591 -0.111
## windspeed  -0.235 -0.158 -0.184 -0.248     1.000   0.006      0.040 -0.012
## holiday    -0.068 -0.029 -0.033 -0.016     0.006   1.000     -0.035  0.008
## weathersit -0.297 -0.121 -0.122  0.591     0.040  -0.035      1.000 -0.049
## year        0.567  0.048  0.046 -0.111    -0.012   0.008     -0.049  1.000
```

We delete 'atemp' variable because it is not that signficant and this variable is highly correlated with 'temp' variable.

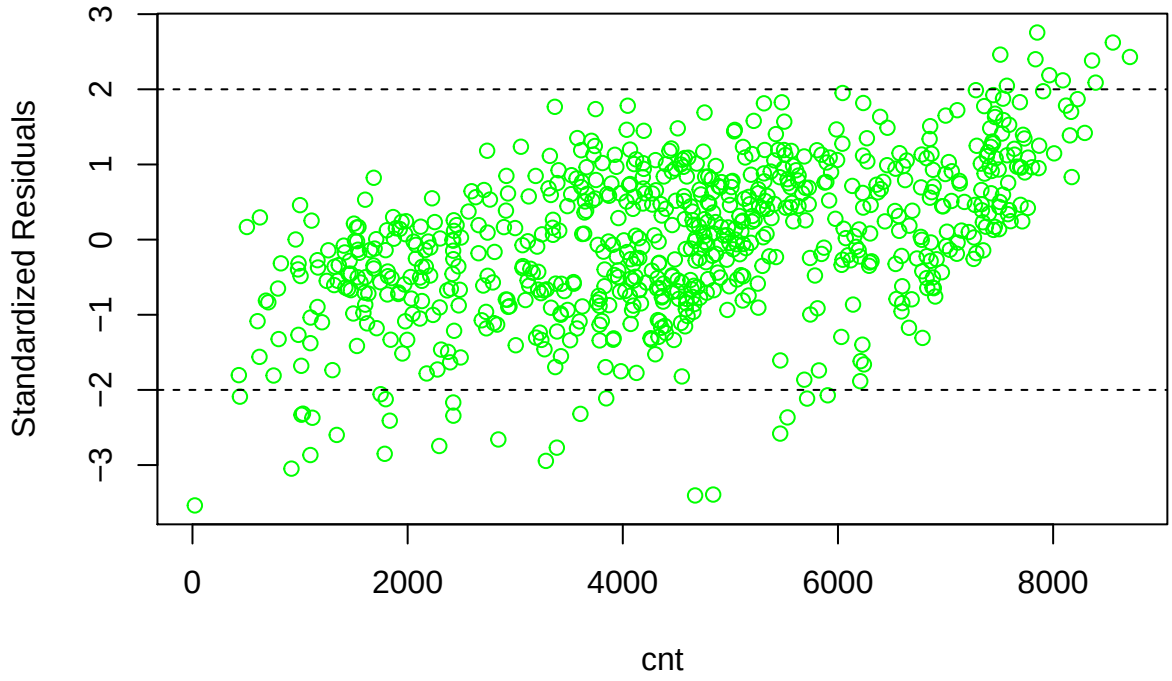
```
# 3. Perform MLR using R on your response (Y) and covariates (X1,X2, ..Xr) .

m.mls <- lm(cnt ~ temp + hum + windspeed + holiday + weathersit + year)
summary(m.mls)
```

```
##
## Call:
## lm(formula = cnt ~ temp + hum + windspeed + holiday + weathersit +
##     year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3435.2  -629.2    27.1   694.3  2699.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2517.01     239.84  10.495  < 2e-16 ***
## temp         5965.62     207.68  28.725  < 2e-16 ***
## hum          -684.22     346.44  -1.975 0.048647 *
## windspeed   -3616.58     500.60  -7.225 1.28e-12 ***
## holiday      -721.11     217.75  -3.312 0.000973 ***
## weathersit   -606.31      87.49  -6.930 9.30e-12 ***
## year         2031.85      73.36  27.696  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 982.1 on 724 degrees of freedom
## Multiple R-squared:  0.7451, Adjusted R-squared:  0.743
## F-statistic: 352.7 on 6 and 724 DF,  p-value: < 2.2e-16
```

4.In the output of this MLR, we can see that except 'hum' is significant under 0.05 significant level other covariates are all significant under 0.001 significant level.

```
# 5.Standard Residuals
StanResMLS <- rstandard(m.mls)
par(mfrow=c(1,1))
plot(cnt,StanResMLS,xlab="cnt", ylab="Standardized Residuals", col="green")
abline(h=2,lty=2)
abline(h=-2,lty=2)
```



From the plot, we think it is a good model.