Consider the performance data for ten hospitals in the U.K.

* The variable y is the **cost per hospital case**
* The variable X1 is the **size of the hospital** in number of beds
* The **visibility score (X2)** is an assessment of how the hospital administrator compares his/her hospital with others in terms of cost – the more accurate the view of the administrator, the higher the score (Rationale is that higher the visibility lower the costs for the hospital)
* The **Admissions** (X3) is the average number of patients admitted to the wards on a certain day

1. **Build a multiple regression model**
2. **Evaluate the model based on the four assumptions discussed in class**
3. **Predict the cost per case given visibility=20 and size=410.**

|  |  |  |  |
| --- | --- | --- | --- |
| **y** | **x1** | **x2** | **x3** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Cost** | **Size** | **Visibility** | **Admissions** |
| 2750 | 225 | 6 | 219 |
| 2400 | 200 | 37 | 198 |
| 2920 | 300 | 14 | 292 |
| 1800 | 350 | 33 | 342 |
| 3520 | 200 | 11 | 195 |
| 2270 | 250 | 21 | 248 |
| 3100 | 175 | 21 | 167 |
| 1980 | 400 | 22 | 399 |
| 2680 | 350 | 20 | 350 |
| 2720 | 275 | 16 | 269 |

**R-Code:**

**install.packages("psych")**

**install.packages("nloptr")**

**library(psych)**

**library(car)**

**HospitalCost1 <- read.csv("HospitalCost.csv", stringsAsFactors=FALSE, header=TRUE)**

**# --------Check 1 for linearity and multicollinearity-----------------**

**#Examine the pairs of variables to see the nature of association**

**pairs(HospitalCost1)**

**# Build linear model with all variables**

**RegressedCost1<- lm(Cost ~ Size+Visibility+Admissions, data=HospitalCost1)**

**summary(RegressedCost1)**

**# ---------Remove the redundant variable -----------------------------**

**#Remove the 4th Column (Admissions)**

**HospitalCost=HospitalCost1[-4]**

**pairs(HospitalCost)**

**# Build linear model without Admissions**

**RegressedCost<- lm(Cost ~ Size+Visibility, data=HospitalCost)**

**summary(RegressedCost)**

**#Compare the R-sq of the two models and explain**

**#Quick and Dirty Rule - Select the variable with the highest p-value for possible elimination**

**# ------Inspect residues -------------------------------------------------------------**

**res = resid(RegressedCost) #get the residues**

**res #display the residuals**

**plot(RegressedCost, which=1) #check the plot for constant variance**

**summary(lm(abs(res)~fitted(RegressedCost))) #dirty test for a linear relationship**

**# -------------------------------------------------------------------**

**plot(RegressedCost,which=2) #check normal plot**

**shapiro.test(res) #do a test for normality**

**# -------------------------------------------------------------------**

**durbinWatsonTest(RegressedCost)**

**# -------------------------------------------------------------------**

**#predicting new values**

**newdata=data.frame(Size=410,Visibility=20)**

**predict(RegressedCost,newdata)**

**#Or you can determine the cost directly using the regression equation**

**4240.131-3.762\*410-29.896\*20**

1. **Build a multiple regression model**

**The regression model is:  
Cost = 4240.131 -3.762 \* Size - 29.896 \* Visibility**

**Note: We got this model from the stepwise regression output shown below. But often you will need to do a pairwise check for multicollinearity (i.e. correlation between predictors) or a significance check for variables (i.e. the p-value of the t-test) before you arrive at your final chosen model with predictors you think are meaningful.**

> step(RegressedCost1, direction="backward")

Start: AIC=120.05

Cost ~ Size + Visibility + Admissions

Df Sum of Sq RSS AIC

- Admissions 1 0 734622 118.05

- Size 1 1019 735641 118.06

<none> 734622 120.05

- Visibility 1 668658 1403280 **124.52**

Step: AIC=118.05

Cost ~ Size + Visibility

Df Sum of Sq RSS AIC

<none> 734622 118.05

- Visibility 1 689540 1424162 122.67

- Size 1 713631 1448253 **122.83**

Call:

lm(formula = Cost ~ Size + Visibility, data = HospitalCost1)

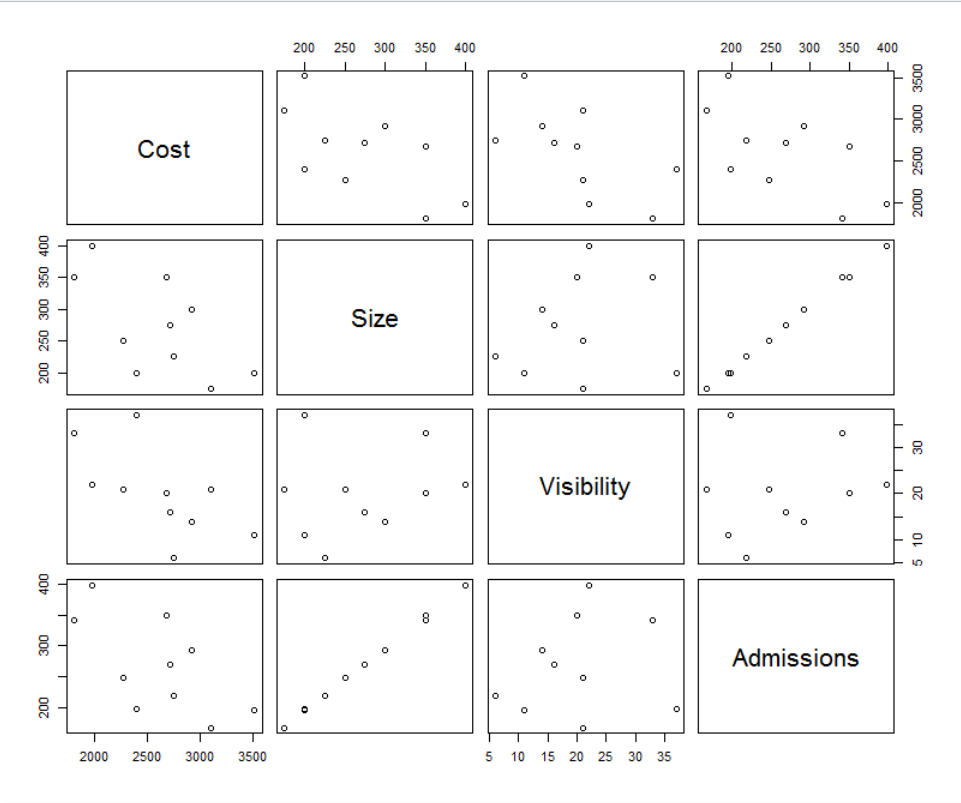
Coefficients:

(Intercept) Size Visibility

4240.131 -3.762 -29.896

1. **Evaluate the model based on the four assumptions discussed in class**

**Assumption 1: Linearity, Multicollinearity, Significance**

****

**Note: The scatterplot shows that Size and Admissions have a perfectly linear correlation. You can test this using the cor(Size, Admission) function to make sure.**

# Build linear model with all variables

**RegressedCost1<- lm(Cost ~ Size+Visibility+Admissions, data=HospitalCost1)**

**summary(RegressedCost1)**

Residuals:

Min 1Q Median 3Q Max

-464.24 -127.15 5.69 206.65 361.24

---------------------------------------------------------

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4239.58752 586.68027 7.226 0.000356 \*\*\*

Size -3.69948 40.54644 -0.091 0.930271

Visibility -29.89209 12.79118 -2.337 0.058087 .

Admissions -0.06214 40.07044 -0.002 0.998813

------------------------------------------------------------

Residual standard error: 349.9 on 6 degrees of freedom

Multiple R-squared: 0.6956, Adjusted R-squared: 0.5434

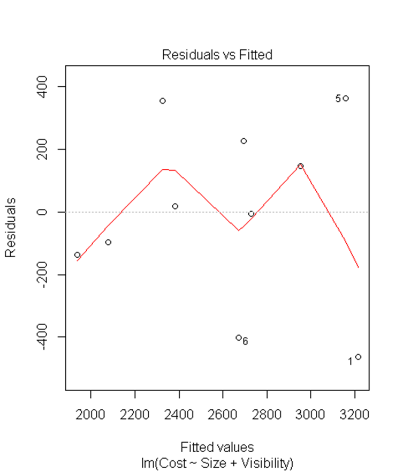
F-statistic: 4.571 on 3 and 6 DF, p-value: 0.05416

**Here we confirm our observations from the scatterplot. Both Size and Admissions have extremely high p-value much above our alpha of 0.05 and are therefore candidates for removal.**

**But which one to remove?**

1. Remove the variable with the highest p-value (this is a thumbrule) or one with the lower t-value (the higher the t-value the better the explanatory power of the variable) and perform a regression again.

**Assumption 2: Variances of errors are constant for predicted values**

****

**There doesn’t seem to be a pattern in this plot of residuals against fitted values. To check for a linear correlation we can do a regression also as shown in the code.**

**summary(lm(abs(res)~fitted(RegressedCost)))**

**Call: lm(formula = abs(res) ~ fitted(RegressedCost))**

**Residuals:**

**Min 1Q Median 3Q Max**

**-235.58 -110.25 14.06 108.11 187.33**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) -269.6970 310.3658 -0.869 0.410**

**fitted(RegressedCost) 0.1879 0.1173 1.602 0.148**

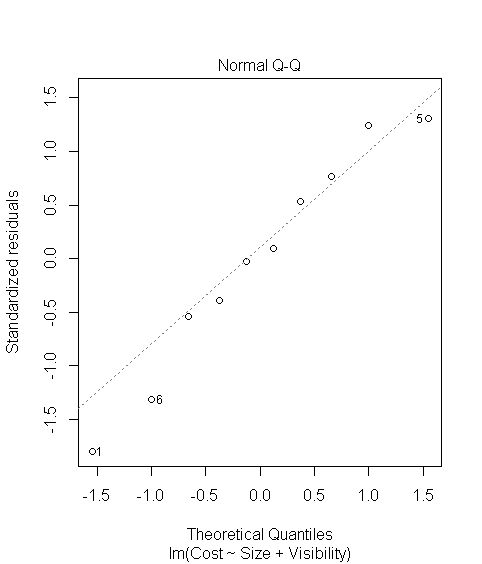
**Multiple R-squared: 0.2429, Adjusted R-squared: 0.1482**

**F-statistic: 2.566 on 1 and 8 DF, p-value: 0.1478**

**There doesn’t seem to be a correlation as the model statistics confirm.**

**Assumption 3:**

* **The error terms are normally distributed.**
  + **Check by making a histogram or normal quantile plot of the residuals. Use shapiro.test(residuals(regressedModel))**

****

shapiro.test(res)

------------------------------------------------------

data: res

W = 0.9407, p-value = 0.5608

The null hypothesis is that the residuals have a normal distribution.

------------------------------------------------------

Here the p-value is much higher than our level of significance (alpha) of 0.05, so do not reject the null hypothesis, which means there is a high likelihood that residues are normally distributed (as is also shown by the plot).

**Assumption 4:** Observations are independent (not from the same source and also not serially dependent). Observations are independent (residuals are not correlated – also called autocorrelation and can be tested with the **Durbin-Watson test)**

> durbinWatsonTest(RegressedCost)

lag Autocorrelation D-W Statistic p-value

1 -0.4628048 2.63217 0.152

Alternative hypothesis: rho != 0

Answer: Negative correlation is suggest by the D-W value greater than 2.

Null hypothesis that residuals are not correlated (correlation=0) cannot be rejected due to high p-value of 0.152.

1. **Predict the cost per case given visibility=20 and size=410.**

> #predicting new values

> newdata=data.frame(Size=410,Visibility=20)

> predict(RegressedCost,newdata)

2099.671

**Answer:** Cost is 2099.67 for a case admitted to the hospital.