Q1: 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. **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 to remove a variable from a dataset:  
#----------------------------------------------------**

**#Example: Remove the 2nd Column (Size) from the dataframe HospitalCost1 and write to a new dataframe HospitalCost**

**HospitalCost=HospitalCost1[-2]**

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

**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**

**#Modify this line to include the additional variable**

**RegressedCost1<- lm(Cost ~ predictor1+predictor2+predictor3,data=HospitalCost1)**

**summary(RegressedCost1)**

**# INSERT R CODE TO REMOVE THE VARIABLE**

**#PERFORM REGRESSION ON THE NEW DATASET**

**RegressedCost<- lm(Cost ~ predictor1+predictor2,data=HospitalCost)**

**summary(RegressedCost1)**

**#predicting new values**

**newdata=data.frame(Size=aaa,Visibility=bb)**

**predict(RegressedCost,newdata)**

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

**4240.131-3.762\*aaa-29.896\*bb**

**Include your R code here:  
---------------------------------**

1. **Build a multiple regression model (the algebraic form of the model)**RegressedCost1<- lm(Cost ~ Size+Visibility,data=HospitalCost)

summary(RegressedCost1)

**cost =4240.131 -3.762\*size + -29.896\*visibility + e**

1. **Explain the results of the tests (t-test, F-test) and goodness of model (Adj R-sq)  
   a. t-test results**p-value size 0.0350  
   p-value visibility 0.0374  
   both are significant for a cut off of 0.05 **b. F-test results**p value = 0.01556  
   p value is below the cutoff point and hence the model is statistically significant and a good fit **c. Goodness of the model (variance explained) – Adj. Rsq**Adjusted R-squared: 0.6086  
   60.86% of the predicted cost can be determined by size and visibility
2. **What is the predicted value of Cost?**

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

2099.671

Q2: **Simple Regression Example:**

To explore the relationship between muscle mass and age in women, a nutritionist randomly selected four women from each 10 year age group, beginning with age 40 and ending with age 79. The results are shown.

Age: 71, 64, 43, 67, 56, 73, 68, 56, 76, 65, 45, 58, 45, 53, 49, 78

Muscle: 82, 91, 100, 68, 87, 73, 78, 80, 65, 84, 116, 76, 97, 100, 105, 77

Assume the simple linear regression model.

1. Obtain the estimate regression function.
2. Test at the 5 % level whether muscle mass decreases with age. Estimate the p–value of the test.

**Note:**

# Here is the R code to create the dataframe containing the data. So no csv file is needed.

Age = c(71, 64, 43, 67, 56, 73, 68, 56, 76, 65, 45, 58, 45, 53, 49, 78)

Muscle=c(82, 91, 100, 68, 87, 73, 78, 80, 65, 84, 116, 76, 97, 100, 105, 77)

**MuscleAge=data.frame(Age,Muscle)**

View(MuscleAge)

*rmage <- lm(Muscle ~ Age, data=MuscleAge)*

*summary(rmage)*

*MuscleMass = 148.0507 -1.0236\*Age*

*H0 = muscle mass does not change with age*

*H1 = muscle mass decreases with age*

*Since the coefficient of age is negative (-1.0236) the muscle mass decreases with age. The p-value is 8.27e-05 thus is statistically significant for 5% level and hence we reject the null hypothesis and conclude that muscle mass decreases with age.*