Daily Assignment # 5 – Scatter Plots, regression and scatter plor of data and fitted values in R

|  |  |  |  |
| --- | --- | --- | --- |
| **Countries** |  | Production Capacity (thousand barrels/day) | Marginal Cost |
| Venezuela |  | 4400 | 20 |
| Iraq |  | 3700 | 16 |
| Kuwait |  | 3300 | 13 |
| Iran |  | 4600 | 10 |
| Saudi Arabia |  | 12000 | 9 |
| Nigeria |  | 2700 | 7 |
| UAE |  | 3000 | 5 |

DA\_5 solutions

Use DA5.R for R commands

Daily Assignment # 5

#-------------------------------------------

#install needed R packages

#you will learn this in Sections

#you only need to install them once then only call them using library()

#-------------------------------------------

#for reading escell data file install the package below

#install.packages("readxl")

#for OLS linear regression model install below

#install.packages("lmsupport")

#for summary stats install below

#install.packages("psych")

#for scatter plots etc install ggplot2

#install.packages("ggplot2")

#install haven

#install.packages("haven")

#install.packages("data.table")

#install.packages("dplyr")

#install.packages("foreign")

#call them in

# Loading packages

**library(dplyr)**

**library(haven)**

**library(readr)**

**library(knitr)**

**library(haven)**

**library(dplyr)**

**library(readxl)**

**library(psych)**

**library(ggplot2)**

**library(stats4)**

**library(lmSupport)**

**library(magrittr)**

**library(qwraps2)**

**library(stargazer)**

1. Change into your working directory and Read data into R

#-------------------------------------------

#set your working directory

#-------------------------------------------

**setwd("/Users/sofiavillas-boas/Dropbox/EEP118\_Spring2021/Daily Assignments/5-DA-Lecture5")**

#-------------------------------------------

#1. Read in data and see the top rows to see column names etc

#read in DA5 data set

**my\_data <- read\_excel("opec.xlsx")**

**head(my\_data)**

1. Summary stats of data

**describe(my\_data,skew=FALSE)**

1. Create a scatter plot of Capacity against Marginal cost in R using the seven OPEC countries’ data. Use appropriate units, and make sure to properly label the axes, put a title, etc.

**#scatter plot**

**scatter\_data <- ggplot(data = my\_data) + geom\_point(aes(x=capacity, y=margCost, color = "data")) +**

**xlab("Production Capacity (thousand barrels/day)") + ylab("Marginal Costs (Dollars)") +**

**ggtitle("Marginal Costs and Capacity (thousand barrels/day)")**

**scatter\_data**

1. Can you figure out how to label the points, Label the points that are Venezuela and the Saudi Arabia on the graph. Print out the graph and turn it in with a couple of sentences of comment.
2. Run regression of Marginal Costs on capacity.

#-----------------------------------------------------------------------------------------------

#run regressions

#-----------------------------------------------------------------------------------------------

**reg1<-lm(margCost ~ capacity,my\_data)**

**summary(reg1)**

1. Created a variable equal to the fitted values from the linear regression model

**my\_data$margCosts\_hat <-reg1$fitted.values**

1. Plot a scatter plot of the data, marginal costs on Y axis, and capacity on the X axis and in the same plot, in blue, add the fitted values on the same Y axis. Add title also

**scatter\_data\_fittedVals <- ggplot(data = my\_data) + geom\_point(aes(x=capacity, y=margCost, color = "data")) +**

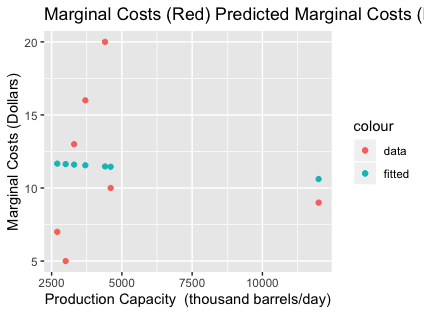
**geom\_point(aes(x=capacity, y=margCosts\_hat, color = "fitted")) +**

**xlab("Production Capacity (thousand barrels/day)") + ylab("Marginal Costs (Dollars)") +**

**ggtitle("Marginal Costs (Red) Predicted Marginal Costs (Blue) and Capacity")**

**scatter\_data\_fittedVals**

See R file for solutions. And see solutions doc for how the graphs should look



**REPLICATION of Lecture 5**

Replicate Lecture 5 using Lecture5.R and Lecture5.csv dataset.

#-------------------------------------------

#set your working directory

#-------------------------------------------

#setwd("/Users/sberto/Desktop/")

setwd("/Users/sofiavillas-boas/Dropbox/EEP118\_Spring2020/Lectures/Lecture5")

#-------------------------------------------

#1. Read in data and see the top rows to see column names etc

#-------------------------------------------

my\_data <- read.csv("Lecture5.csv")

head(my\_data)

#summarize data

describe(my\_data)

#lecture 4 only used year=87

my\_data2 <- my\_data[my\_data$year ==87,]

head(my\_data2)

#regression

regLecture4 <- lm(crmrte ~ polpc,my\_data2)

#show output

summary(regLecture4)

#predicted crime rate

my\_data2$crmrte\_hat <- regLecture4$fitted.values

#--------------------------------------------------------------------------------------

#in section you will learn how to plot scatter plots etc and also in lecture 5 in a

#jupyter notebook with me

#--------------------------------------------------------------------------------------

#-------------------------------------------

#for Lecture 5 scatter plot Y and Yhat

#-------------------------------------------

fig1<-plot(my\_data2$crmrte,my\_data2$crmrte\_hat, main="Scatter of Crime Rate and Predicted Crime Rate",

xlab="Crime Rate", ylab="Predicted Crime Rate")

#--------------------------------------------------------------------------------------

# vary sample size and show how standard errors of beta hats change

#--------------------------------------------------------------------------------------

#load sample N=100

#read in a Stata dataset

sample100 <- read\_dta("sample100.dta")

#load sample N=400

#read in a Stata dataset

sample400 <- read\_dta("sample400.dta")

#use sample N=630

#read in a Stata dataset

sample630<-my\_data

#-----------------------------------------------------------------------------------------------

#run regressions and compare the standard errors of the beta hats as the sample size N increases

#-----------------------------------------------------------------------------------------------

reg100<-lm(crmrte~polpc,sample100)

summary(reg100)

reg400<-lm(crmrte~polpc,sample400)

summary(reg400)

#--------------------------------------------------------------------------------

#in Jupyter notebook you compare summary(regLecture4) using N=90 with N=630 below

#regression

regLectureN630 <- lm(crmrte ~ polpc,my\_data)

#show output

summary(regLectureN630)

#generate fitted values

my\_data$crmrte\_hat<-regLectureN630$fitted.values

#make combined scatter plot of crime rate data and fitted values of crime rate given regression estimates

scatter\_data\_fittedVals <- ggplot(data = my\_data) + geom\_point(aes(x=polpc, y=crmrte, color = "data")) +

geom\_point(aes(x=polpc, y=crmrte\_hat, color = "fitted")) +

xlab("Police Per Capita") + ylab("Crime Rate ") +

ggtitle("Crime Rate (Red) and Predicted Crime Rate (Blue) and Police Per Capita")

scatter\_data\_fittedVals