QUESTIONS:

2. Stamey et al. (1989) examined the correlation between level of prostate specific antigen (LPSA) with a number of clinical measures on 97 patients with prostate cancer (data in file prostate.xls). Determine which clinical measure has the highest correlation with LPSA and build a linear model (using standard OLS) between the clinical measure and LPSA. In order to test the goodness of the linear model use coefficient of determination or R^2.

3. Anscombe (1973) has provided four synthetic data sets consisting of two variables x and y (data in file anscombe.xls). Find the best fit linear model for the four data sets using standard OLS. What do you observe? For which of the four data sets do you think that a linear model is appropriate and why?

SOLUTIONS:

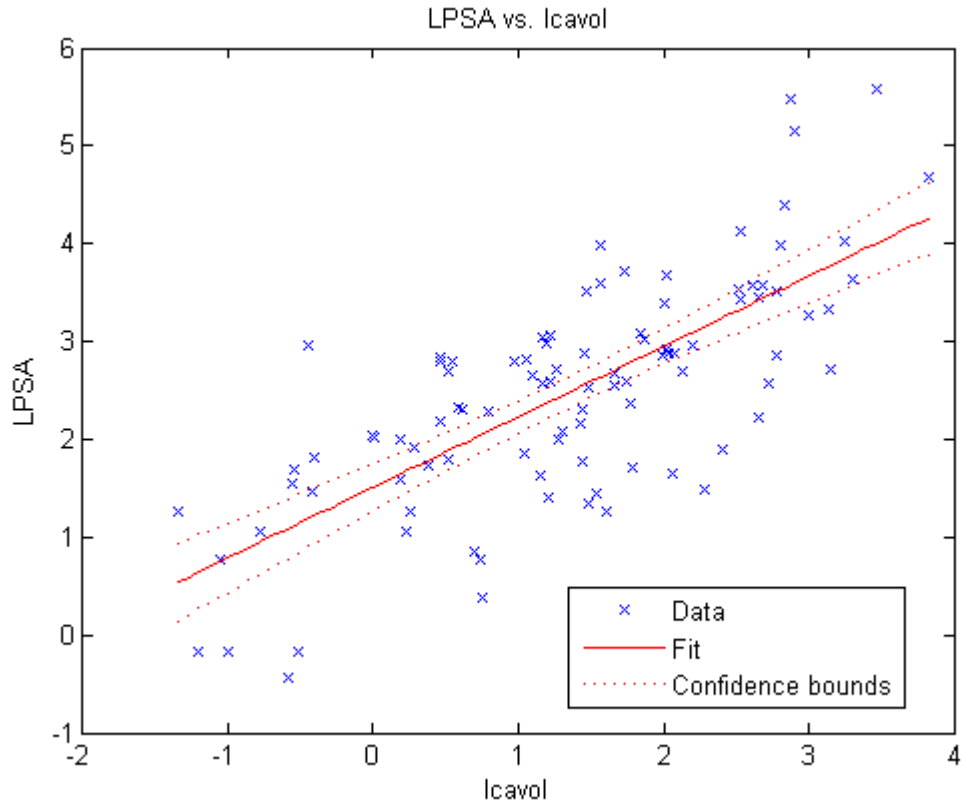

```
clc
clear all
close all
```

Question 2

```
%Reading numerical data into num and column names into txt
[num,txt,raw] = xlsread('prostate.xlsx');
Nvar=length(txt); % No. of variables

%Determining clinical measure has the highest correlation with LPSA
  Correlation= corrcoef(num); % Returns correlation coefficient matrix

%Correlation coefficient b/w each clinical measure(X)and LPSA(Y)-lastcolumn)
  Corr_XY = Correlation(1:Nvar-1,Nvar)

% Determining measure with highest correlation to LPSA
 [ HighestCorr, Index]=max(Corr_XY)
 ClinMeas=cell2mat(txt(Index)) % Clinical Measure with Highest Correlation to LPSA

 % Fitting linear model using standard ols (without weights)
 mdl = fitlm(num(:,Index),num(:,Nvar),'linear','VarNames',{ClinMeas,'LPSA'})

% Plotting the model
figure(1)
plot(mdl)
 CoeffOfDetermination = mdl.Rsquared.Ordinary % R squared Value
```

```
Corr_XY =

    0.7345
    0.4333
```

```
    0.1696
    0.1798
    0.5662
    0.5488
    0.3690
    0.4223


HighestCorr =

    0.7345


Index =

     1


ClinMeas =

lcavol


mdl =


Linear regression model:
    LPSA ~ 1 + lcavol

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     1.5073      0.12194    12.361    1.7223e-21
    lcavol         0.71932     0.068193    10.548    1.1186e-17


Number of observations: 97, Error degrees of freedom: 95
Root Mean Squared Error: 0.787
R-squared: 0.539,  Adjusted R-Squared 0.535
F-statistic vs. constant model: 111, p-value = 1.12e-17

CoeffOfDetermination =

    0.5394
```



LCAVOL has the highest correlation with LPSA determined by the correlation coefficient showing the highest value (closest to one). Linear regression model:

LPSA = 1.5073 + 0.71932*lcavol

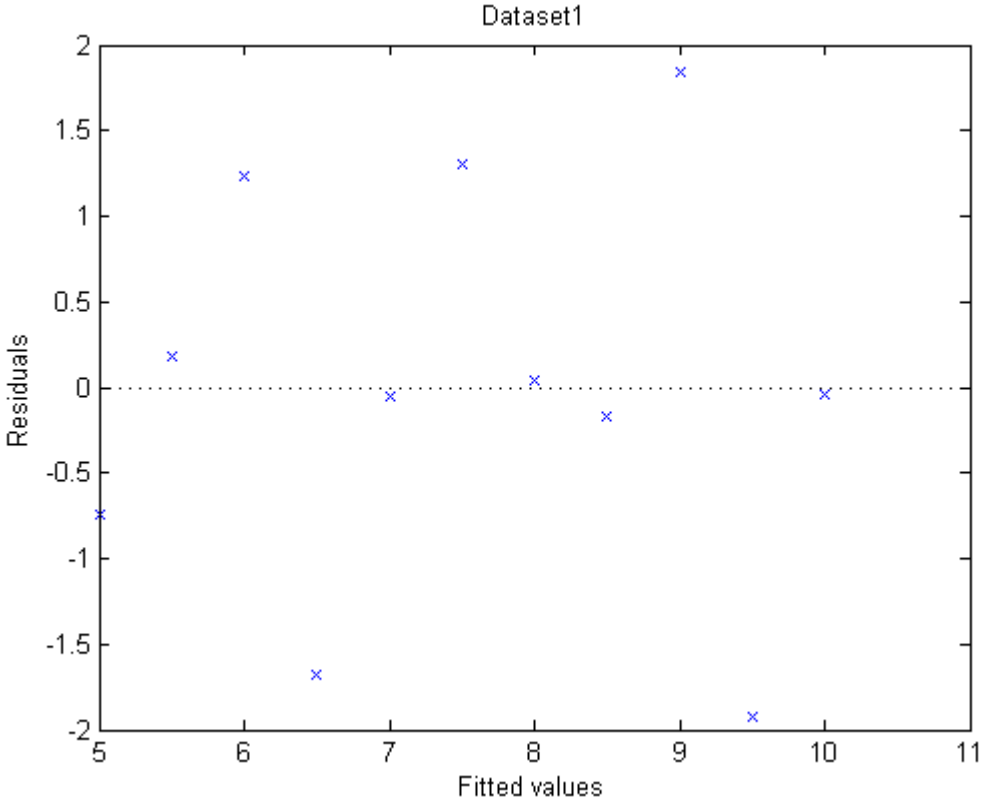
Question 3

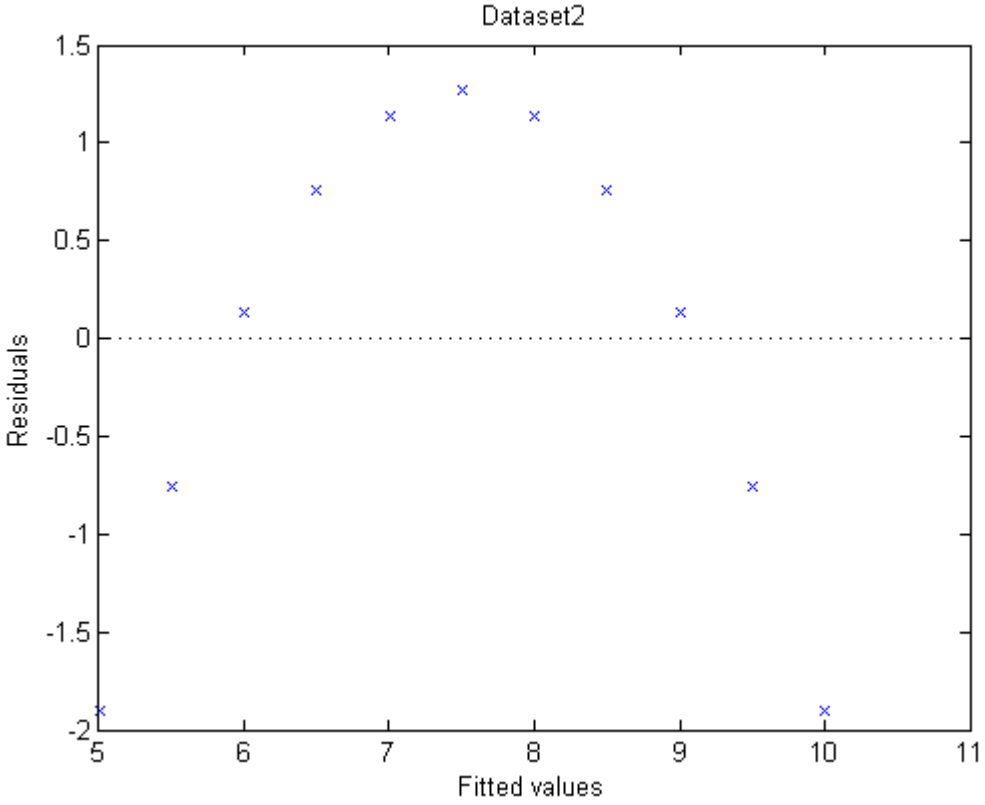
```
 %Reading numerical data into num2
num2 = xlsread('anscombe.xls');

data1=num2(1:11,1:2); % [X Y] for 1st data set

 % Fitting linear model using standard ols (without weights)
 mdl1 = fitlm(data1(:,1),data1(:,2),'linear','VarNames',{'X1','Y1'})
 Rsq1 = mdl1.Rsquared.Ordinary % Goodness of fit
 figure(2) % Plotting the model
 plotResiduals(mdl1,'fitted')
 title('Dataset1')

data2=num2(1:11,3:4); % [X Y] for 2nd data set
 mdl2 = fitlm(data2(:,1),data2(:,2),'linear','VarNames',{'X2','Y2'})
 Rsq2 = mdl2.Rsquared.Ordinary % Goodness of fit
  figure(3) % Plotting the model
  plotResiduals(mdl2,'fitted')
  title('Dataset2')

data3=num2(1:11,5:6); % [X Y] for 3rd data set
 mdl3 = fitlm(data3(:,1),data3(:,2),'linear','VarNames',{'X3','Y3'})
 Rsq3 = mdl3.Rsquared.Ordinary % Goodness of fit
  figure(4) % Plotting the model
  plotResiduals(mdl3,'fitted')
```

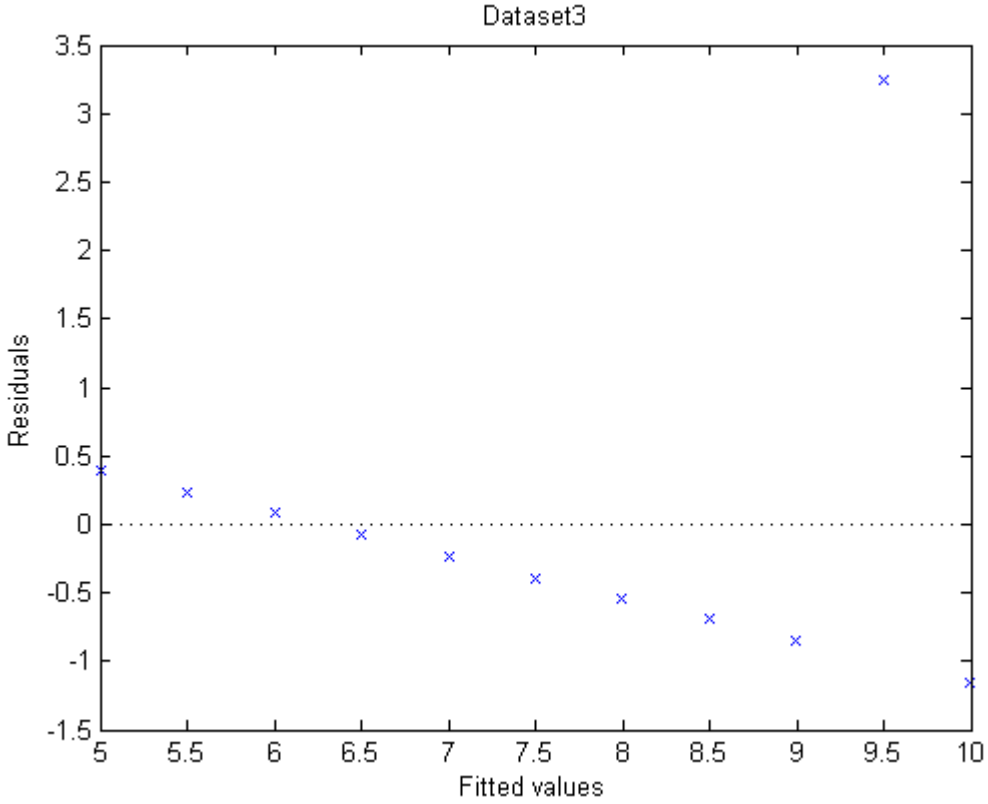
```
 title('Dataset3')

data4=num2(1:11,7:8); % [X Y] for 4th data set
 mdl4 = fitlm(data4(:,1),data4(:,2),'linear','VarNames',{'X4','Y4'})
 Rsq4 = mdl4.Rsquared.Ordinary % Goodness of fit
  figure(5) % Plotting the model
 plotResiduals(mdl4,'fitted')
 title('Dataset4')
```

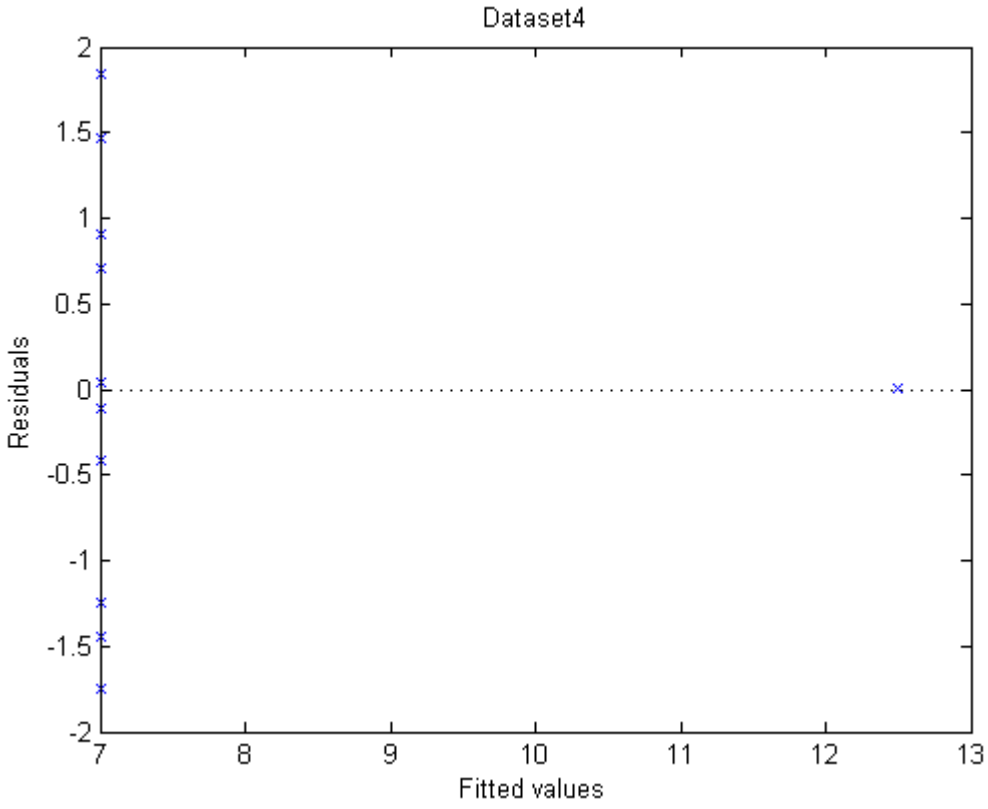
```
mdl1 =


Linear regression model:
    Y1 ~ 1 + X1

Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    _______    ______    _________

    (Intercept)     3.0001      1.1247    2.6673     0.025734
    X1             0.50009     0.11791    4.2415    0.0021696


Number of observations: 11, Error degrees of freedom: 9
Root Mean Squared Error: 1.24
R-squared: 0.667,  Adjusted R-Squared 0.629
F-statistic vs. constant model: 18, p-value = 0.00217

Rsq1 =

    0.6665


mdl2 =


Linear regression model:
    Y2 ~ 1 + X2

Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    _______    ______    _________

    (Intercept)    3.0009       1.1253    2.6668     0.025759
    X2                0.5      0.11796    4.2386    0.0021788


Number of observations: 11, Error degrees of freedom: 9
Root Mean Squared Error: 1.24
R-squared: 0.666,  Adjusted R-Squared 0.629
F-statistic vs. constant model: 18, p-value = 0.00218

Rsq2 =

    0.6662
```

```
mdl3 =


Linear regression model:
    Y3 ~ 1 + X3

Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    _______    ______    _________

    (Intercept)     3.0025      1.1245    2.6701     0.025619
    X3             0.49973     0.11788    4.2394    0.0021763


Number of observations: 11, Error degrees of freedom: 9
Root Mean Squared Error: 1.24
R-squared: 0.666,  Adjusted R-Squared 0.629
F-statistic vs. constant model: 18, p-value = 0.00218

Rsq3 =

    0.6663


mdl4 =


Linear regression model:
    Y4 ~ 1 + X4

Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    _______    ______    _________

    (Intercept)     3.0017      1.1239    2.6708      0.02559
    X4             0.49991     0.11782     4.243    0.0021646


Number of observations: 11, Error degrees of freedom: 9
Root Mean Squared Error: 1.24
R-squared: 0.667,  Adjusted R-Squared 0.63
F-statistic vs. constant model: 18, p-value = 0.00216

Rsq4 =

    0.6667
```


Dataset1
Residuals
Fitted values
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
5
6
7
8
9
10
11


Dataset2
Residuals
Fitted values
1.5
1
0.5
0
-0.5
-1
-1.5
-2
5
6
7
8
9
10
11





Published with MATLAB® R2014a

Linear regression models:

Dataset 1: Y1 = 3.0001+ 0.50001*X1

Dataset2: Y2 = 3.0009 +0.5* X2

Dataset 3: Y3 =3.0025 +0.49973* X3

Dataset 4: Y4 =3.0017 +0.49991* X4

We observe that although the line fit is almost the same for all 4 cases and although R^2 value, mean, variance, correlation coefficient, etc are same for all the datasets, not all of them are good for a linear fit. This is determined by plotting the residual plots.

We can see that data set 1 is appropriate for a linear fit as the residuals are distributed randomly over the fitted values. The rest of the data sets are not distributed randomly but follow some pattern which shows that a non linear fit perhaps is more accurate.