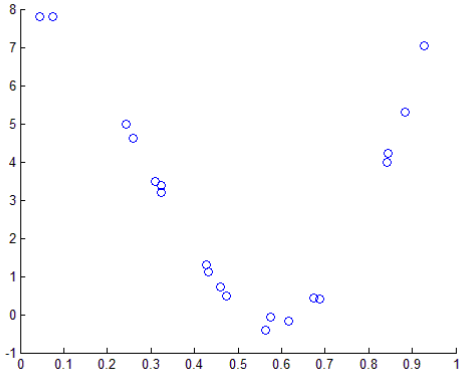
Homework Set 2

Problem 1 (Regression)

In this problem you will implement linear regression in MATLAB. The data for this problem is available on the course website. You will learn functions using the training data and evaluate their generalization on the test data. The picture below was generated by loading the training data into vectors `x` and `y` and plotting the sample points using the command scatter as follows:

```
>> X = load('02HW1_Xtrain');
>> Y = load('02HW1_Ytrain ');
>> scatter(X,Y);
```



Fit the data with a polynomial function from the 1st to the 3rd order: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots$
Finish the following tasks. Remember to submit your MATLAB source code along with your homework.

a) Report the training error obtained using the polynomials. Observe if the residuals follows the least squares assumptions. Perform a residual analysis and elaborate the results.

b) Submit a plot showing the training data and the polynomials you learned.

Problem 2 (Variance of linear regression model)

Suppose that the data are in fact generated by a linear model $Y = Xw + \varepsilon$, in which the ε is an i.i.d. random variable with zero mean. Recall that we can prove that the least-squares estimate $\widehat{w}$ given by the normal equations is an unbiased estimate of the true weights w, assuming that the data are in fact generated by a linear model. The proof goes as follows.

$E[\widehat{w}] = E[(X^T X)^{-1} X^T Y]$	from the normal equations
$= (X^T X)^{-1} X^T E[Y]$	because the input values are fixed here
$= (X^T X)^{-1} X^T E[Xw + \varepsilon]$	by the model assumption above
$= (X^T X)^{-1} X^T (Xw + E[\varepsilon])$	as the true parameters are fixed
$= (X^T X)^{-1} X^T (Xw) + 0$	because errors are zero-mean
$= (X^T X)^{-1} (X^T X) w = w$	by associativity

Using a similar approach, show that the variance of the least-squares estimate $\widehat{w}$ is given by $\mathrm{Var}(\widehat{w}) = (X^T X)^{-1}\sigma^2$, where σ^2 is the variance of ε. [Hint: begin from the following definition of the variance of a random vector Z: $\mathrm{Var}(Z) = E[(Z - E[Z])(Z - E[Z])^T]$.

Problem 3 (Residue analysis)
The dataset "`02HW3_Diabold_Li_data.txt`" contains yields of different maturities at each month from 1970 to 2000. Use the data on 1990.05.31 to perform a detailed analysis and answer the following questions.

a) Generate a scatter plot of the yields vs maturity. Comment on the figures.

b) You want to fit a polynomial regression model to the data. Since you don't know the order you need, you fit six polynomial models, with orders from 1 to 6. Plot the R^2 vs the polynomial order k. Comment on the result.

c) For the 4th-order polynomial model, draw a residual plot (vs maturity). Comment on it.

d) Use Levene test (MATLAB function: `vartestn`) to check if the variance of the first half of the data set is the same as the second half. Does the conclusion confirm what you see from the residual plot? Try to explain.

e) Draw a histogram and a quantile-quantile (Q-Q) plot of the residuals. Comment on it.