Chalmers-University of Gothenburg Department of Mathematical Sciences

Probability, Statistics and Risk, MVE240

Project 1

In the project description we sketch the analysis of the problem we expect you to do. (Obviously you are welcome to do more.) To pass the project a short report should be written and handed in to the project supervisor. In addition the group should present their results in class. The presentation should take about 15 minutes. Please include a short introduction which will facilitate for other students to understand the results of the project. (Do not assume that the audience knows the subject.)

1 Introduction - Regression

When dealing with two or more variables, the functional relation between the variables is often of interest. In almost any activity where experiments have been performed, linear regression might be an extremely useful tool for analysis.
In this project you will use examples from engineering, to illustrate the following aspects of regression:

(1) Linear regression

(2) Non-linear regression

(3) Multiple non-linear regression

The Matlab program `regress` can be used.

2 Linear regression - Component Testing

In a tensile test in a laboratory, test specimens are exposed to load. For a given load t (kN), one measures the elongation x (10^{-2} mm) of the item. Some measurements have been saved on the file `testdata`. Load the data in Matlab, check the variables, and plot data:

```
load testdata
plot(t,x,'*')
```

Judging from the plot, there seems to be a (linear) relation: when the load increases, the elongation increases. However, there is also randomness occuring. For a given load (7.9 kN), four items were tested. Obviously, there is a variability in the elongation. We want to model this with tools from probability. In other situations (and maybe in this too), measurement errors is a source of randomness.

2.1 A Probabilistic Model

A naive model would be to write

$$x \approx \alpha + \beta t. \quad (1)$$

However, in order to handle the randomness in the measurements, we introduce the variables ϵ_i, $i = 1, \ldots, 9$. Assume that $\epsilon_i \in \mathrm{N}(0, \sigma^2)$; then we rewrite Eq. (1) as

$$x_i = \alpha + \beta t_i + \epsilon_i. \tag{2}$$

Thus, we have 9 pairs of values $(t_1, x_1), \ldots, (t_9, x_9)$ where $t_1, \ldots, t_9$ are given quantities and $x_1, \ldots, x_9$ are observations of independent random variables $X_1, \ldots, X_9$. The expectations represents a straight line and the (unknown) variance σ^2 models the deviation from the line. The variable t is often called the independent variable, or covariate, while x is called the dependent variable.

2.2 Estimates of Parameters

To estimate the parameters α and β in the model given in Eq. (2) one can use the Maximum Likelihood method. Derive the formulas for the ML-estimators of α and β. in the case of Gaussian ϵ_i.
In Statistics Toolbox in Matlab, there is a routine `regress` which performs least-squares estimation. (For Gaussian ϵ_i LS and ML methods give the same estimators.) The parameters will be obtained in a vector `b`. The purpose of the part `ones(9,1)` in the call is to handle the so-called intercept (the constant term). Here is the call:

```
b = regress(x,[ones(9,1) t])
```

The parameter β is given by `b(2)` while α is found as `b(1)`. From the help text to the routine (`help regress`), you find that confidence intervals for the estimated parameters may be given as output if requested.
Plot the regression line in the same figure as the original data:

```
hold on
xest = b(1) + b(2)*t;
plot(t,xest,'r.-')
```

2.3 Check of Residuals

In the model it is assumed that the residuals ϵ_i belonged to a Gaussian distribution; let therefore examine the residuals. Create a vector `res` of residuals:

```
res = xest - x;
```

Plot first the residuals in a probability paper.

```
wnormplot(res)
```

Of course, to draw any conclusions from only nine observations is almost nonsense. Assuming normality, we can estimate the variance σ^2 of ϵ_i (and check if zero mean):

```
wnormfit(res)
```

3 Non-linear regression - Fatigue of Metals

In some physical phenomena the strength of material, i.e. the capacity of carrying loads, is decreasing with time. The material is "aging" or "degrading". Typical examples of such phenomena are fatigue of materials, corrosion or failure of cables insulation. The speed of degradation may depend on the applied load.

In laboratory experiments one often subjects a specimen of a material to a constant amplitude load, e.g. $L(t) = \frac{s}{2}\sin(\omega t)$ where s and ω are constants, and counts the number of cycles (periods) until it breaks. The number of load cycles $N(s)$ as well as the amplitudes s are recorded. The data set is called SN data.
In practice, one often uses a simple model for $N(s)$,

$$N(s) = K\epsilon^{-1}s^{-\beta}, \quad s > 0, \tag{3}$$

where K is a (material dependent) random variable, usually lognormally distributed, i.e. $\ln K \in \mathrm{N}(0, \sigma_K^2)$, and ϵ, β are fixed constants. Clearly, this model is non-linear in s.
However by taking the logarithm[1] of Eq. (3), we obtain

$$\ln N(s) = \ln K - \ln\epsilon - \beta \ln s. \tag{4}$$

Convince yourself that

$$\ln N(s) \in \mathrm{N}(-\ln\epsilon - \beta \ln s, \sigma_K^2),$$

i.e regression model.
Load the SN data which are stored in the file `SN.mat` and estimate parameters in the regression, check assumptions by plotting the residuals.

4 Multiple non-linear regression - Mean Oxygenation Rate for a Stream

The mean rate of oxygenation from the atmospheric reaeration process for a stream depends on the mean velocity of stream flow and average depth of the stream bed. Data from 12 experiments have been recorded; they are found in the file `oxyrate.mat`. Clear the workspace, load data into Matlab and check variables.
Suppose we know from experts that a relation of the type

$$X = \alpha V^{\beta_1} H^{\beta_2}$$

can be used to estimate the mean oxygenation rate. How can we estimate the parameters α, β_1, and β_2 from data? (Hint: Again, take the logarithm and obtain a linear relation then use Matlab function `regress` to estimate the parameters.)
Visualise the regression by plotting the regression plane together with measurements (use `plot3`). Analyse residuals.

[1] Recall that $\ln XY = \ln X + \ln Y$