Chalmers-University of Gothenburg Department of Mathematical Sciences

Probability, Statistics and Risk, MVE240

Project 2

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 - Has the intensity of hurricanes changed?

Discussions on climate change are intense. Broadly speaking, one aspect is whether certain events, for instance storms or hurricanes, have been more common recently; another if such events (possibly in addition) have become more severe and violent. A difficult problem related to the later issue is the relation to climate change: are changes to occur statistically speaking, as a part of the inherent variability of nature, or are these an effect from human activity.
In this project you will investigate possible trend change by means of regression model for the trends.
Focusing on hurricanes in the Atlantic, the season 2005 resulted in many broken records: 27 named storms, 14 hurricanes, 7 major hurricanes, 4 hurricanes of Category 5. Moreover, the hurricane Katrina implied very large and severe losses. Researchers agree that a change in intensity has occurred in recent years. When investigating a possible trend over time for the yearly number of hurricanes, analyses are often based on regression models where the number of hurricanes is supposed to be a random variable dependent on covariates like North Atlantic oscillation index. Some researcher found that only weak linear trends can be ascribed to hurricane activity and that multidecal variability is more characteristic of the region.
The region of the Atlantic basin will be studied. Different type of data can be found on the Tropical Prediction Center homepage of Unisys Weather,
http://weather.unisys.com/hurricane/atlantic/.
For example The National Oceanic and Atmospheric Administration (NOAA) uses an Accumulated Cyclone Energy (ACE) index as a measure of total seasonal activity. This refers to the collective intensity and duration of named storms and hurricanes in the North Atlantic during a given season. The ACE index is defined as the sum of the squares of the maximum sustained surface wind speed measured every six hours for all named systems while they are at least tropical storm strength.
The data is called `acedata.dat`. It contains 2 columns; the year is given in the first column while the second column contains ACE index. Load the data and employ the regression model to check for trends in the data.

2 Linear regression

Denote by X_i the values of ACE index year i. Since ACE is an average of large number of components thus, by Central Limit Theorem, X_i could be approximately normally distributed. Hence we start with assumption that $X_i \in N(m_i, \sigma^2)$.

We are interested if there is some systematical variability of parameters m_i, σ_i^2. Obviously the simplest model would be that the parameters are constant and that weather is independent from year to year. Let check it first.
Load the data, plot it to visually check the hypothesis that X_i has the same cdf. Use also normal probability paper to control the normality assumption.
The linear regression is often used to model the variability of m_i. In its simplest form one assumes that

$$m_i = \alpha + \beta \cdot i.$$

and that $\sigma_i^2 = \sigma^2$, i.e. variance is constant. Alternatively one can write the model as follows

$$X_i = \alpha + \beta \cdot i + \epsilon_i, \tag{1}$$

where ϵ_i are iid $N(0,\, \sigma^2)$ distributed random variables. The problem is to estimate parameters α, β, and σ^2.

2.1 Estimates of Parameters

To estimate the parameters α and β in the model given in Eq. (??) one can use the Maximum Likelihood method. In Statistics Toolbox in Matlab, there is a routine `regress` which performs least-squares estimation. (For Gaussian errors ϵ_i LS and ML methods are equivalent.) Type `help regress`. The following Matlab script can be used to find `b`:

```
>> load Acedata.dat -ascii
>> t=Acedata(:,1); x=Acedata(:,2);
>> b = regress(x,[ones(length(t),1) t]);
```

The purpose of the part `ones(length(t),1)` in the call is to handle the so-called intercept (the constant term α). 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.

```
>> [b bint] = regress(x,[ones(length(t),1) t]);
```

Use the confidence interval to test hypothesis that there is no trend in the data, i.e. if $\beta = 0$ is in the confidence interval. Conclusions!
In order to be able to relay on the test one need also to check if the assumptions behind fitted model are reasonably fulfilled. Plot the regression line in the same figure as the original data:

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

Conclusions.

2.1.1 Check of Residuals

In the model it is assumed that the errors ϵ_i, called also residuals, are independent and 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.

```
>> figure(2)
>> wnormplot(res)
```

Assuming normality, we can estimate the variance σ^2 of ϵ_i (and check if zero mean):

```
>> figure(2)
>> par=wnormfit(res)
```

Conclusions. A useful test is to perform a sort of "parametric bootstrap" to check whether simulated data from the model resemble the measured data.

```
>> figure(1)
>> plot(t,xest+wnormrnd(par(1),par(2),length(xest),1),'ro')
```

Conclusions.

3 Quadratic regression

Since there is some indications that both in 50th and recent years the simple regression somewhat underestimates the average ACE index we propose to use quadratic regression to model m_i, viz.

$$m_i = \alpha + \beta_1 \cdot i + \beta_2 \cdot i^2.$$

It is also assumed that $\sigma_i^2 = \sigma^2$, i.e. variance is constant. Alternatively one can write the model as follows

$$X_i = \alpha + \beta_1 \cdot i + +\beta_2 \cdot i^2 + \epsilon_i,$$

where ϵ_i are iid $N(0, \sigma^2)$ distributed random variables.
Again the parameters α, β_1 and β_2 can be estimated by means of the program (`regress`). The call is

```
>>[b bint] = regress(x,[ones(length(t),1) t t.^2]);
```

- Does the model explains significantly better the variability in the data?
- Are the trends significant?
- Check again the model by plotting residuals. Is the variance of residual constant?[1]

If the variance seams to be lager in the region where m_i are high then one may consider logarithmic transformation of the variable x in order to "stabilize" the variance. Try the following log normal model

$$\ln(X_i) = \alpha + \beta_1 \cdot i + +\beta_2 \cdot i^2 + \epsilon_i,$$

where ϵ_i are iid $N(0, \sigma^2)$ distributed random variables.
Again one can use ((`regress`)) to estimate parameters parameters α, β_1, β_2 and σ^2

```
[b bint] = regress(log(x),[ones(length(t),1) t t.^2]);
```

Analyse the results and draw final conclusions.

[1] You can check it by splitting data in 3 or 4 groups, estimate the variance of residual in each group.