Exam Statistical Models | Department of Mathematics

December 20, 2006 | *vrije* Universiteit

Question 1: ANOVA

A high school student is interested in studying physics and can choose between the universities VU (Vrije Universiteit), TU (Delft University of Technology) and RL (Rijksuniversiteit Leiden). The student wants to base his choice (of course he shouldn't) upon the expected salary he will earn twenty years from now. For completely different (probably sound) reasons somebody else collected income data of samples of former physics students that graduated in the eighties from the three universities. The high school student wants to know whether there is a difference between the expected monthly income in 2006 of students that graduated in physics, depending on the university and wants to apply Analysis of Variance to this end.

a) Formulate an appropriate mathematical ANOVA model.

b) Describe the test the student has to perform. In particular state the hypothesis, test statistic and explain how to conclude.

The data frame `dat` contains for each alumnus his/her former university as well as his/her monthly income in 2006. The R command

```
> dat.aov<-aov(Income University,data=dat)
> summary(dat.aov)
```

yields the following output:

```
> summary(dat.aov)
            Df   Sum Sq  Mean Sq F value  Pr(>F)
University   2  2369129  1184565    ??   0.02327 *
Residuals   ?? 16795749   294662
Total       59 19164878
```

c) Fill in the values that should be given instead of the question marks. If needed, you may use the approximations $19164878 \approx 20000000, 1184565 \approx 1200000$ and $294662 \approx 300000$.

d) Formulate the conclusion when the test is conducted at level of significance $\alpha = 0.05$. In particular, can you advise the student on his choice based on the information available to you at this point?

Question 2: Nonlinear Regression

For the data $\{(x_i, y_i) : 1 \leq i \leq n\}$ that are visualized in Figure 1 two natural models seem to be possible. These are the non-linear regression models with regression functions

$$f(x, \theta) = \theta_1 \exp(\theta_2 x) + \theta_3 \exp(\theta_4 x), \ x \in \mathbb{R} \text{ and } \theta = (\theta_1, \theta_2, \theta_3, \theta_4)^T \in \mathbb{R}^4$$

and

$$f(x, \theta) = \theta_1 \exp(\theta_2 x), \ x \in \mathbb{R} \text{ and } \theta = (\theta_1, \theta_2)^T \in \mathbb{R}^2$$

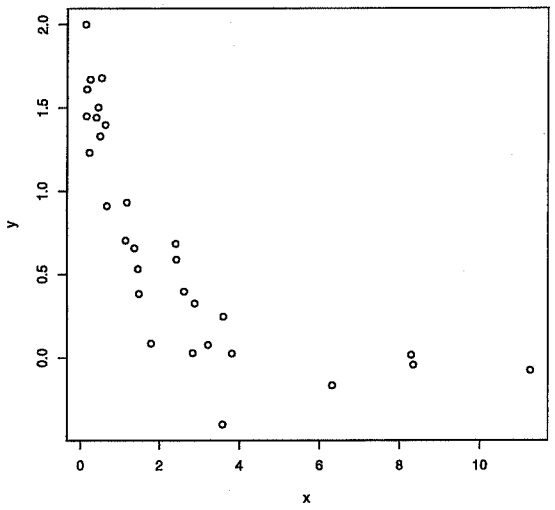


Figure 1: Scatter plot of the data

a) Give the general form of a non-linear regression model.

b) Assume the errors to be normally distributed and describe a test that can be used to choose between the two regression functions given before.

It is decided that the second regression model adequately describes the data. In order to fit the model to the data (estimate the parameters θ_1 and θ_2), the R function `nls` is used. For this function, starting values for the iterations are needed.

c) Based on the scatter plot in Figure 1, give sensible starting values for estimating θ_1 and θ_2 and explain why you chose them.

Question 3: Gereralized Linear Models

a) Describe the logistic regression model and give 2 reasons why this is a generalized linear model.

b) Give an example of a situation where a logistic regression model can be used.

c) In many applications, an explanatory variable is categorical. Suppose we have such an explanatory variable X, having two levels: A and B. Explain how to correctly use this variable in a logistic regression model.

Question 4: Time Series

Look at the time series plotted below in Figure 2, displaying monthly observations for 6 consecutive years.

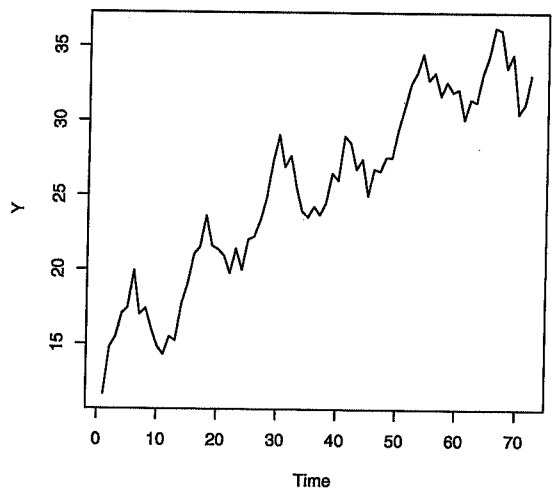
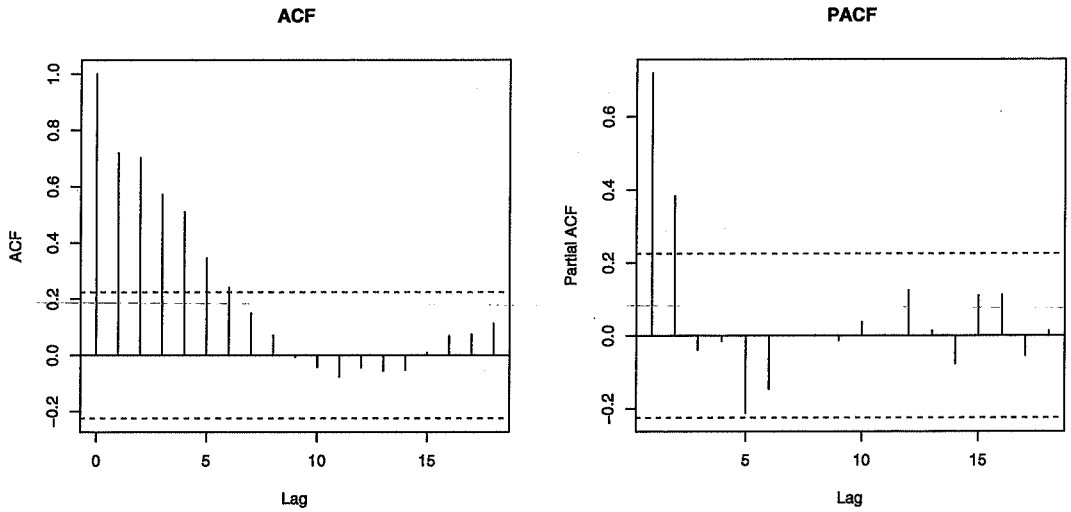


Figure 2: Time plot of the data

a) Comment on what pattern you observe. Define stationarity. Do you think the time series is stationary? Why or why not?

b) Formally describe one method to detrend and deseasonalize a time series.

The following plots show the ACF and PACF of a stationary time series:



c) Which type of time series model would you fit and why? Describe the model you have chosen formally.

Consider an MA process of order one.

d) Derive the autocorrelation function for this model and explain how it can be estimated from the data.

Good luck with this exam!

Grading:

1	2	3	4
a: 2	a: 2	a: 3	a: 3
b: 3	b: 2	b: 1	b: 2
c: 1	c: 1	c: 1	c: 2
d: 2			d: 2

Grade of written exam:
1+(number of points)/3
Final grade (provided the exam result is ≥ 5.5):
mean(weakly exercises)/3 + 2(grade of written exam)/3