STAT 5531 Final — Fall 2010

Thanh Doan

1. (20%) Suppose that a study involves two characteristic measurements.

1. Evaluate and test the hypothesis with α = 0.05.

Comparing with critical value

We see and thus we reject at the level α = 0.05 and conclude that the population mean vector is not equal to.

1. Determine the lengths and directions for the axes of the 95% confidence ellipse.

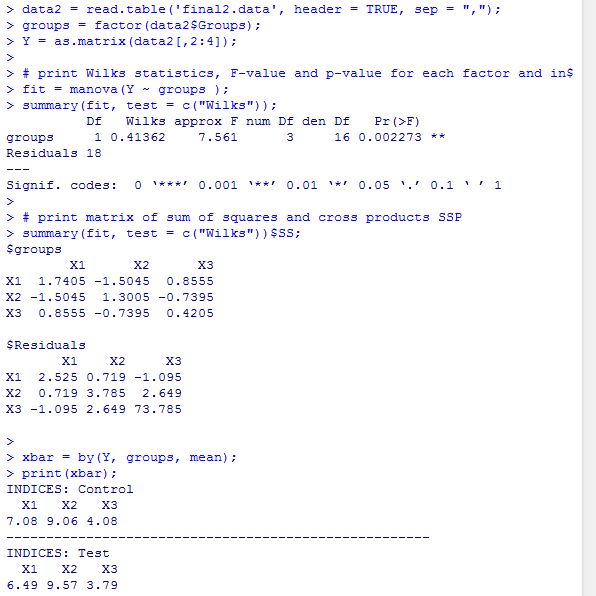


The center is at and the half-lengths of the major/minor axes are:

The axes lie along

2. Suppose that a clinical study involves 3 characteristic measurements in two groups.

1. Perform the analysis of one-way MANOVA at α = 0.05.

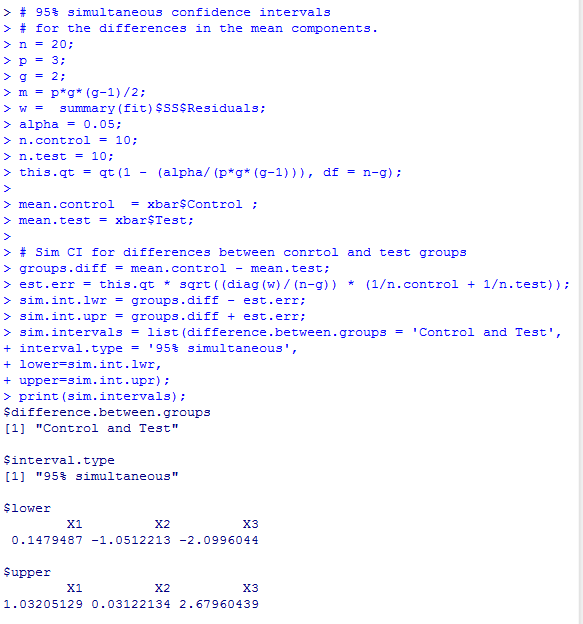


From the Wilks statistics above, it appears

there are differences between the control group and the test group

(b) Construct 95% simultaneous confidence intervals for the differences in the mean

components.



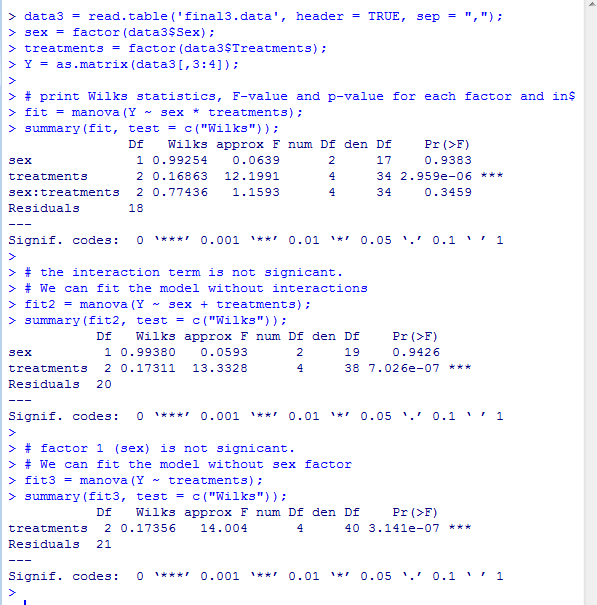
From 95% simultaneous confidence interval above it appears that X1 is significantly has difference between control group and test group.

3. (20%) A scientist carried out a study for comparing the loss in weights of male and

female mice under three different treatments. Four mice of each sex were randomly

assigned to each of three treatments and weight losses were measured at the end of the first and second weeks.

1. Perform the analysis of two-way MANOVA with alpha = 0.05.



Both interaction and factor 1 (sex) is not significant

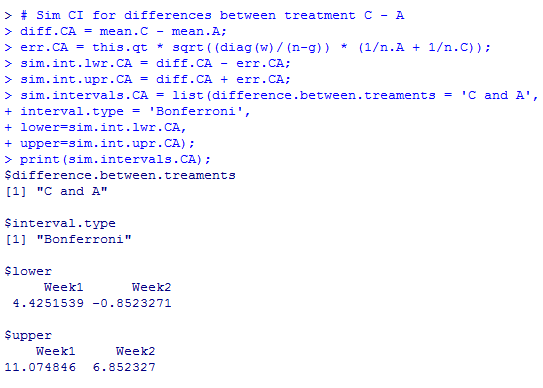
(b) Construct Bonferroni simultaneous 95% confidence intervals for differences of the

components of the significant factor(s).

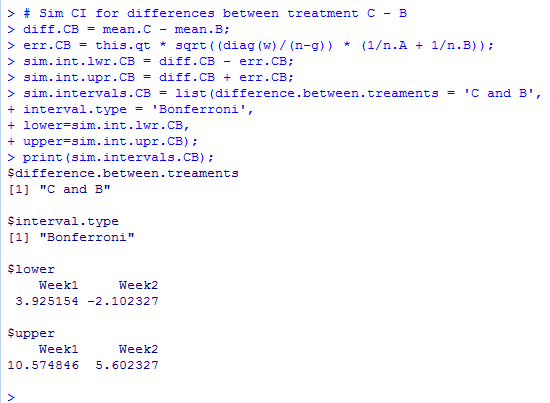
*First, Bonferroni simultaneous 95% confidence intervals for differences between treatment A and treatment B.*



*Next, Bonferroni simultaneous confidence intervals for differences between treatment C and treatment A.*

**

*Finally, below are the bonferroni simultaneous confidence intervals for differences between treatment C and treatment B.*

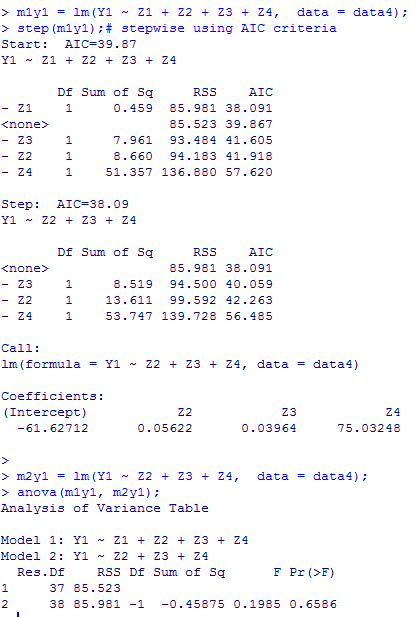


4. (20%) In a regression study, there are three dependent variables and four independent variables.

(a) Perform a regression analysis using each of the response variables Y1, Y2 and Y3.

1. Suggest and fit appropriate linear regression models

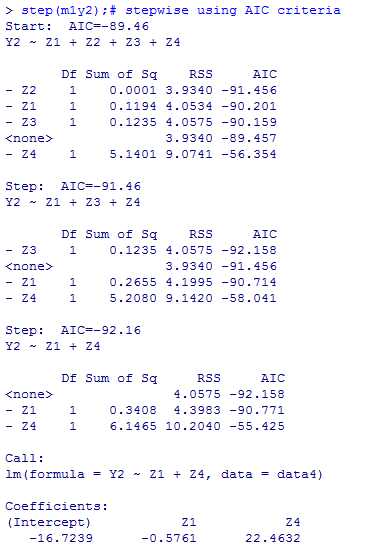
Use stepwise to find appropriate linear model for Y1



After stepwise model building process and manual comparison...

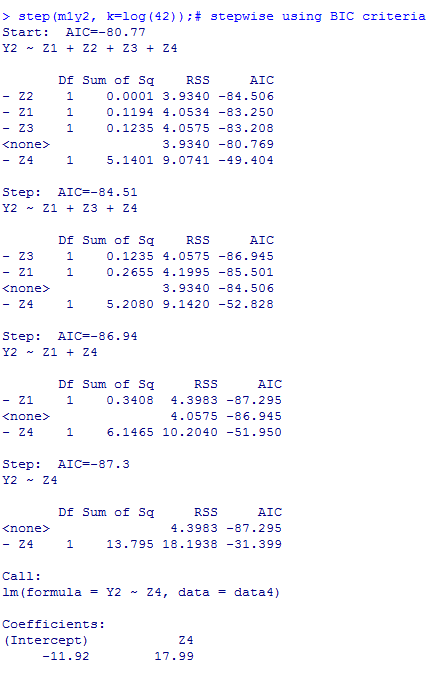
I suggest model 2 for Y1: **m2y1 = lm(Y1 ~ Z2 + Z3 + Z4, data = data4)**

Use stepwise (AIC criterion) to find appropriate linear model for Y2



**For Y2, stepwise procedure (using AIC criterion) suggest keeping Z1 and Z4 in the model**

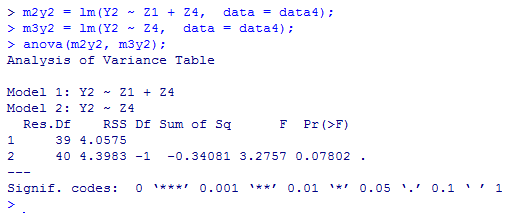
Use stepwise (BIC criterion) to find appropriate linear model for Y2



**For Y2, stepwise procedure (using BIC criterion) suggest keeping only Z4 in the model**

After compare model 2 and model 3 manually using anova...

I suggest model 3 for Y2: **m3y2 = lm(Y2 ~ Z4, data = data4)**



After compare model 2 and model 3 manually using anova...

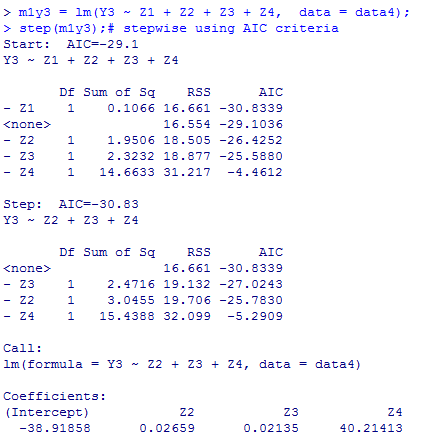
I suggest model 3 for Y2: **m3y2 = lm(Y2 ~ Z4, data = data4)**

**Note:**

**My suggestion above is based on simple analysis. If I analyze this problem in the regression models course or working on a real-world problem then I would spend more time and might suggest different models.**

* 1. **I might standardize the data because Z2 and Z3 have different scales/units and different variation comparing to Z1 and Z4**
  2. **I might explore different types of regression models such as considering transformation, or adding quadratic terms or interaction to the model exploration process.**

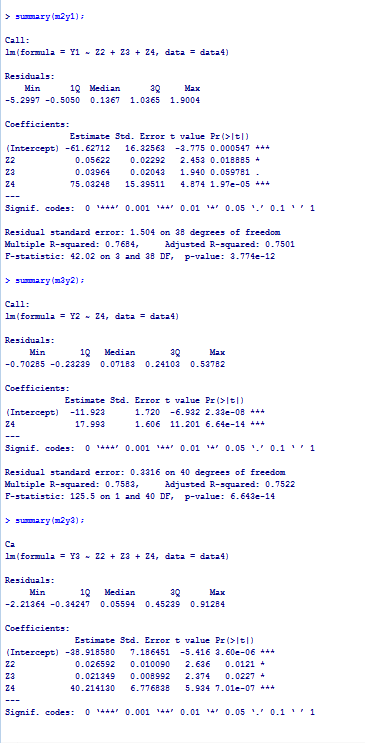
Use stepwise (AIC criterion) to find appropriate linear model for Y3



After stepwise procedure (using both AIC and BIC criterions) and manual comparison...

I suggest model 2 for Y3: **m2y3 = lm(Y3 ~ Z2 + Z3 + Z4, data = data4)**

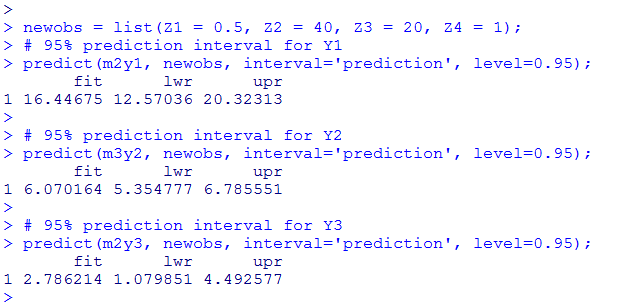
In summary, here are the fitted linear regression models for Y1, Y2 and Y3



Please see attached R code for residuals analysis of these models

ii. Construct a 95% prediction interval for each individual response Yi at z1 =

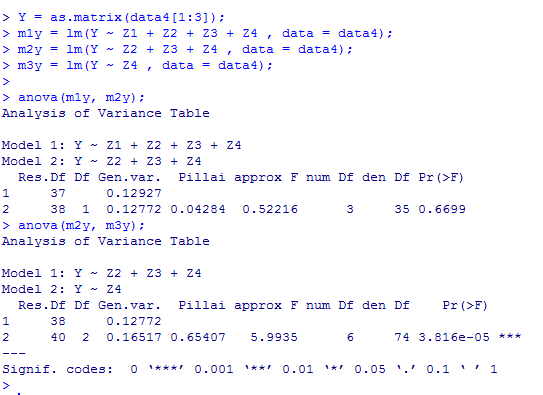
0.5, z2 = 40, z3 = 20, z4 = 1



(b) Perform a multivariate multiple regression analysis using all three response vari-

ables, Y1, Y2 and Y3, and the four independent variables, Z1, Z2, Z3 and Z4.

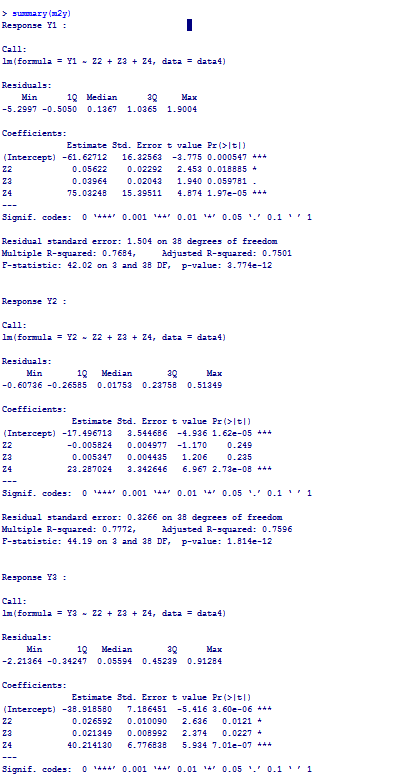
1. Suggest and fit appropriate linear regression models.



After exploration and manual comparison...

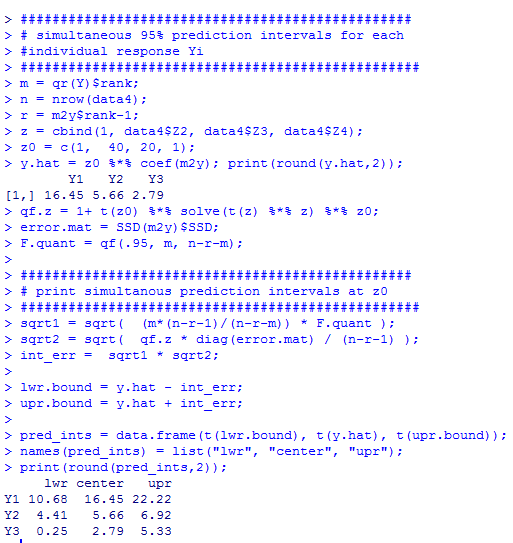
I suggest model 2 for Y = (Y1, Y2, Y3): **m2y = lm(Y ~ Z2 + Z3 + Z4, data = data4);**

Here is the fitted linear regression model for Y = (Y1, Y2, Y3)



ii. Construct simultaneous 95% prediction intervals for each individual re-

sponse Yi at z1 = 0.5, z2 = 40, z3 = 20, z4 = 1



iii. Compare and comment the simultaneous prediction interval with the predic-

tion interval in part a (ii).

The ***simultaneous*** 95% prediction intervals for y1 and y3 in part B are wider/longer than ***individual*** prediction intervals for y1 and y3 in part A

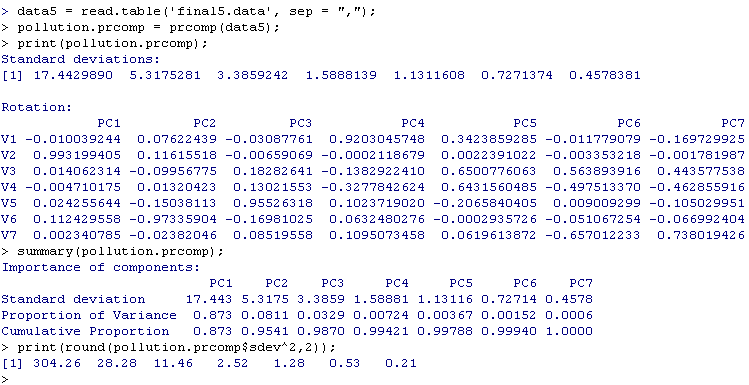
For y2, it is not easy to compare because in part A the response Y2 was fitted with only Z4 as predictor variable while in part B the response is fitted with different predictor variables.

5. The data set contains measurements on seven air-pollution variables recorded

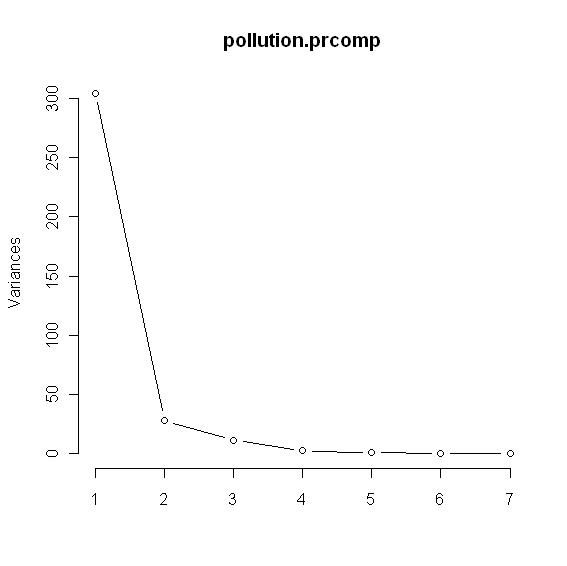
at certain time in the Houston area on different days.

(a) Principal Components Analysis from the sample covariance matrix S

i. Construct the sample principal components



ii. Determine the proportion of the total sample variance explained by the first few principal components. Interpret these components.



**Findings & Interpretation:**

* The first principal component

explains 87.3% of the total sample variance

* The first principal two components,

collectively, explain 95.41% of the total sample variance

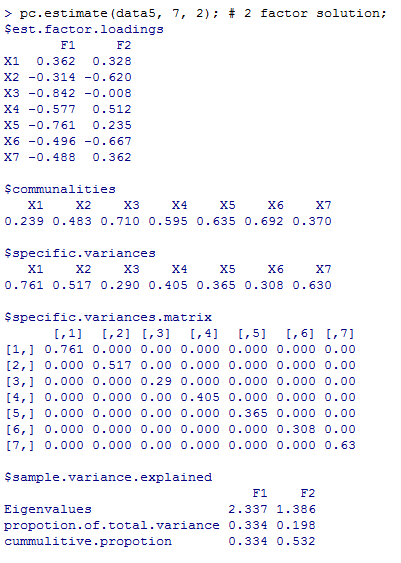
* The above findings and the *scree* plot suggest that sample variance is summarized very well by two principal components and a reduction in the data from 42 observations on 7 air-pollution variables to 42 observations on 2 principal components is reasonable.
* From the component coefficients, the first principal component appears to be essentially a weighted sum between X2 and X6. The contributed determination from other variables appears to be negligible.
* From the component coefficients, the second principal component appears to be a weighted difference between X2 and (a weighted sum X5+X6). The contributed determination from other variables appears to be negligible.

**Note:**

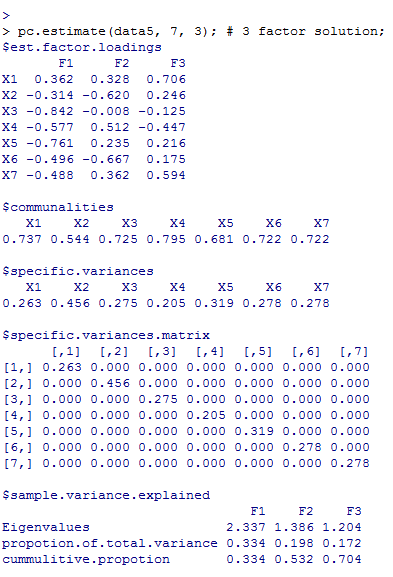
* This problem is investigated under both matlab and R computing environments to compare the result and the capability of each package. The results from both packages are identical.
* The attached matlab code contains more plots and further exploratory work than this report.

1. Factor Analysis from the sample correlation matrix R
2. Obtain the principal component solution to a factor model

**Two-factor solution**

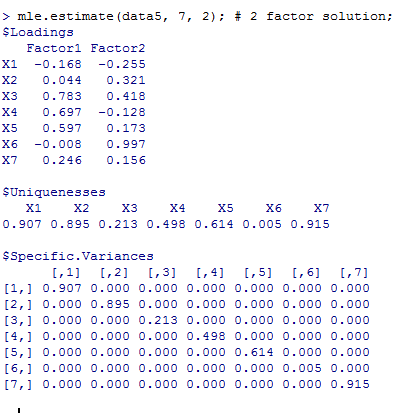
****

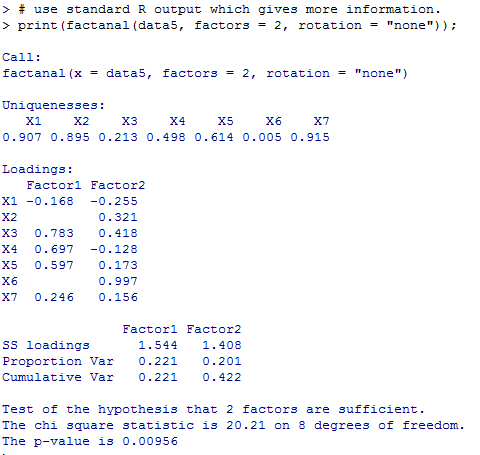
**Three-factor solution**

****

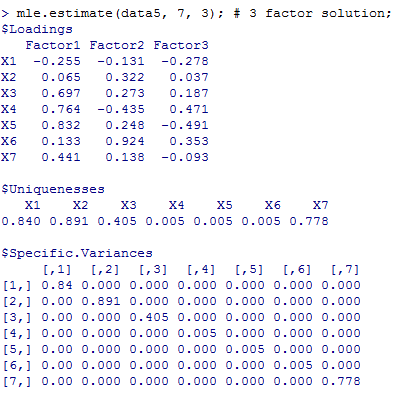
1. Find the maximum likelihood estimates of L (loadings) and (specific variances)

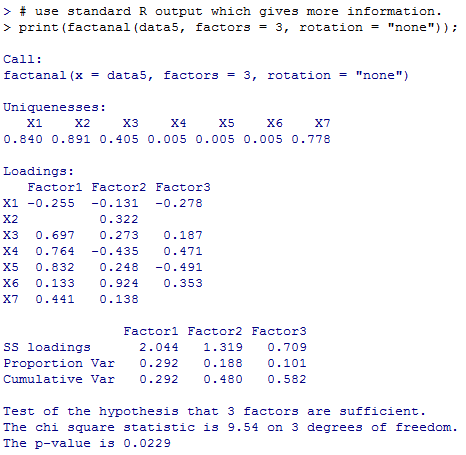
**Two-factor solution**





**Three-factor solution**





**References**

1. Johnson, A. R. and Wichern, D. W. “Applied Multivariate Statistical Analysis”
2. Li, Y. “lectures slides of Applied Multivariate Statistical Analysis”
3. Hewson, P. J. “Multivariate Statistics with R”
4. <http://www.statmethods.net/advstats/factor.html>