Scientific Methods for Health Sciences: Applied Inference (HS851): Fall 2014

http://www.socr.umich.edu/people/dinov/2014/Fall/HS851

Homework 5 Solutions

Problem 1: Using the following survival data

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
T_i	2	5	8	12	15	21	25	29	30	34

Compute by hand the survival time estimates $S(6)$, $S(12)$, and the hazard estimate $h(12)$.

S(t) = P(T_i>t)
S(6) = (1/10)*(2*0+8*1)=8/10 = 0.8
S(12) = (1/10) * (0*4+6*1) = 6/10 = 0.6
h(12): p(12≤t<13 | t>12)) = (1/10)/(7/10) = 0.143

Problem 2: Use SOCR and/or R to generate and interpret the Kaplan-Meyer survival curve for the data below (Example 5 data in the SOCR Survival Analysis Applet).

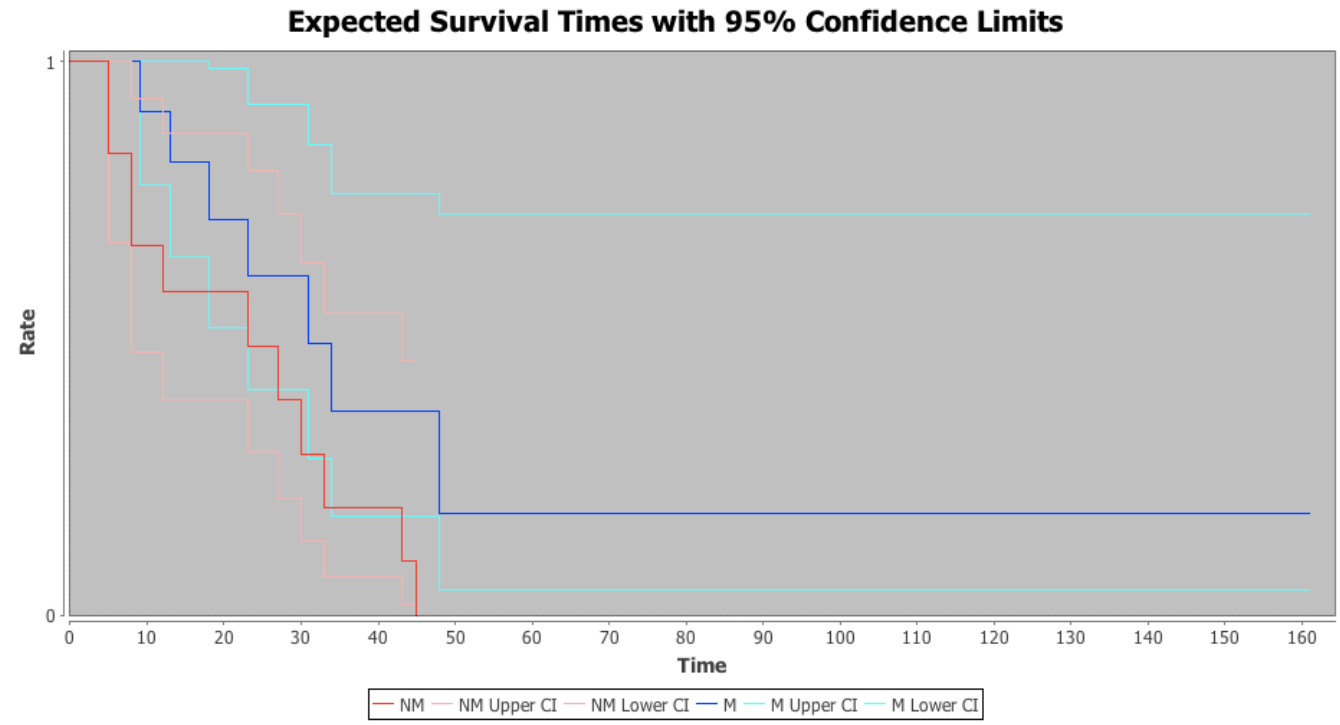
time	censor	group
9	1	M
13	1	M
13	0	M
18	1	M
23	1	M
28	0	M
31	1	M
34	1	M
45	0	M
48	1	M
161	0	M
5	1	NM
5	1	NM
8	1	NM
8	1	NM
12	1	NM
16	0	NM
23	1	NM
27	1	NM
30	1	NM
33	1	NM
43	1	NM
45	1	NM

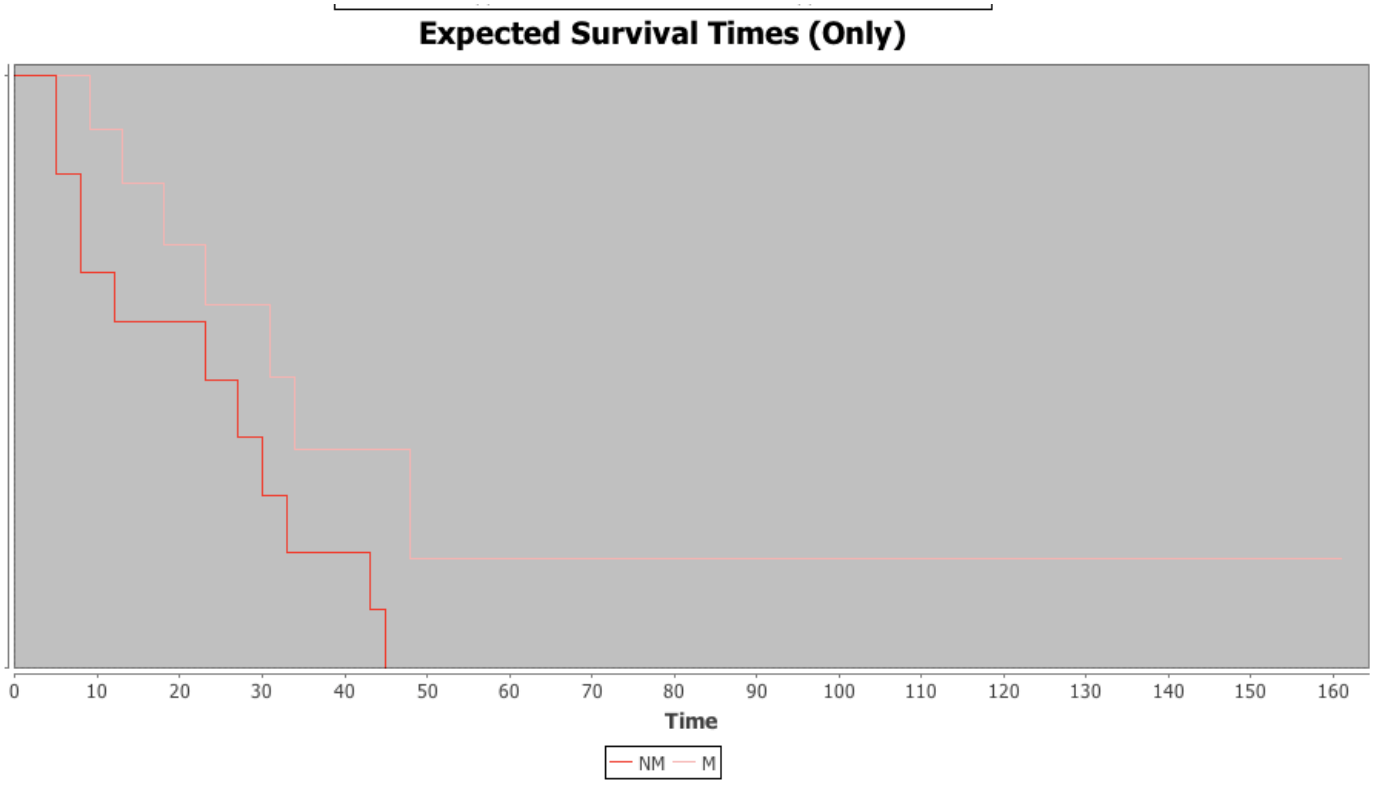
SOCR results:

Samepl Size = 23
Number of Censored Cases= 5
Number of Groups Cases= 2
Groups = NM M

Survival Times (Censored Cases Marked +) =
9.0 13.0 13.0+ 18.0 23.0 28.0+ 31.0 34.0 45.0+ 48.0
161.0+ 5.0 5.0 8.0 8.0 12.0 16.0+ 23.0 27.0 30.0
33.0 43.0 45.0

Time	No. At Risk	Rate	SE of Rate	Upper CI	Lower CI
Group = NM					
5.0	12	.833	.108	1.000	.674
8.0	10	.667	.136	.933	.477
12.0	8	.583	.142	.871	.390
23.0	6	.486	.148	.802	.294
27.0	5	.389	.147	.724	.209
30.0	4	.292	.139	.638	.133
33.0	3	.194	.122	.545	.069
43.0	2	.097	.092	.460	.021
45.0	1	.000	□	□	.000
Group = M					
9.0	11	.909	.087	1.000	.777
13.0	10	.818	.116	1.000	.648
18.0	8	.716	.140	.987	.519
23.0	7	.614	.153	.924	.408
31.0	5	.491	.164	.851	.283
34.0	4	.368	.163	.762	.178
48.0	2	.184	.153	.726	.047



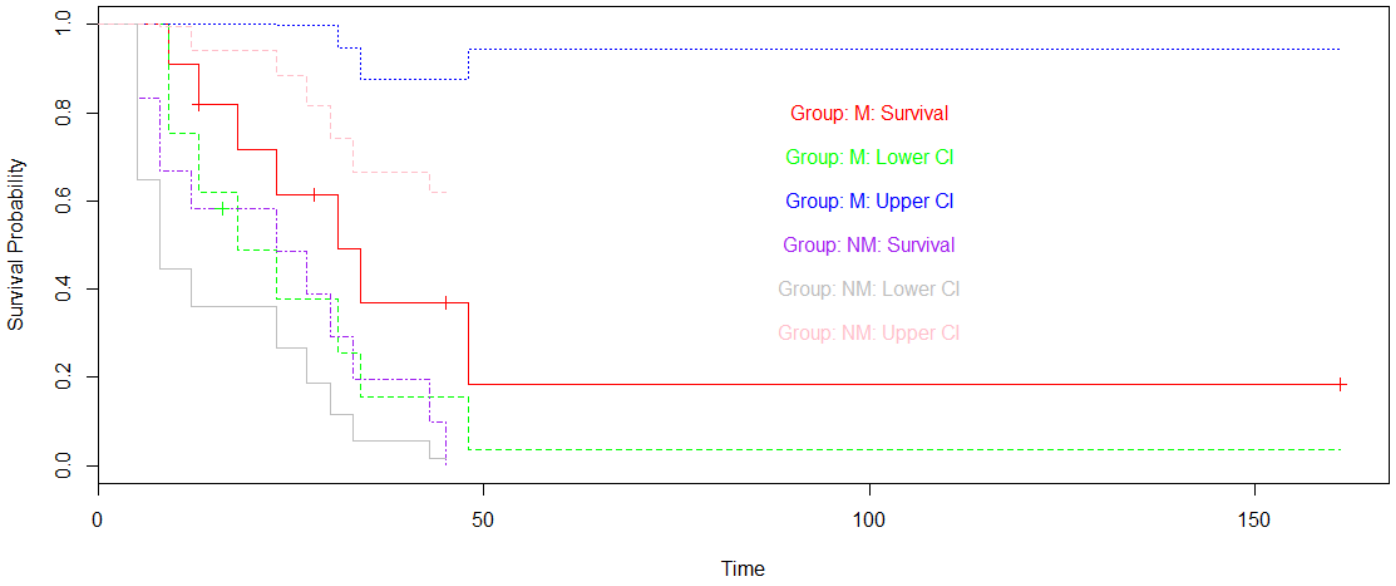


The results suggest that the survival times of the two groups are not significantly different since the CIs have significant overlap. However, on average, group NM dies a bit more quickly. Also, in group M, we see that once you survive to time 50, the hazard is very low.

In R:

```
require('survival')
dataset <- read.csv("C:\\Users\\ Desktop\\data.csv", header=TRUE)
surv<-survfit(Surv(time, censor)~group, data=dataset)
summary (surv)
source('http://www.stat.ucla.edu/~david/teac/surv/conf-bands.R')
my.cb <- conf.bands(surv, type= 'hall', 100, 600)
plot(surv, lty=1:4, col=c('red', 'green', 'blue', 'purple', 'gray', 'pink'), xlab="Time", ylab="Survival
Probability", mark.time=TRUE, conf.int=TRUE)
text(100,  .8, "Group: M: Survival", col='red')
text(100,  .7, "Group: M: Lower CI", col='green')
text(100,  .6, "Group: M: Upper CI", col='blue')
text(100,  .5, "Group: NM: Survival", col='purple')
text(100,  .4, "Group: NM: Lower CI", col='gray')
text(100,  .3, "Group: NM: Upper CI", col='pink')

# If we have additional meta-data about the cases (e.g., Gender, Age), then we can fit
# linear survival models, which can be used to "predict" survival. Hence, we can
# compare alternative survival models:
# survfit1 <- coxph(Surv(time, status)~age+sex, data=dataset, subset=(etype==1), method="breslow")
# survfit2 <- coxph(Surv(time, status)~ sex, data=dataset, subset=(etype==1), method="breslow")
```

In the plot below, the dashed line is for group M and the solid line for group NM.



Problem 3: The multivariate mmreg.csv data includes 600 observations and 8 psychological, academic and demographic (gender) variables. Use some dimensionality reduction methods to interrogate the data and report your findings.

Principal components analysis

```
dataset<-read.csv('http://www.ats.ucla.edu/stat/data/mmreg.csv')
pca.model<-princomp(dataset)
summary(pca.model)
```

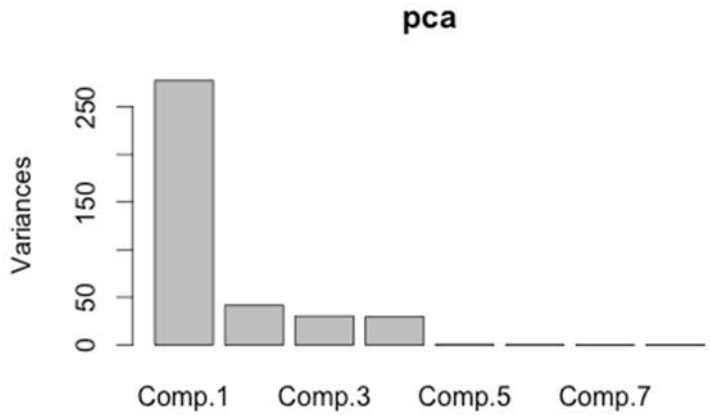
Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	16.6603919	6.4582621	5.49497449	5.43624558	0.734981761
Proportion of Variance	0.7300605	0.1097033	0.07941816	0.07772963	0.001420828
Cumulative Proportion	0.7300605	0.8397639	0.91918202	0.99691165	0.998332478

	Comp.6	Comp.7	Comp.8
Standard deviation	0.5931064249	0.4338453026	0.3065846745
Proportion of Variance	0.0009252385	0.0004950605	0.0002472233
Cumulative Proportion	0.9992577163	0.9997527767	1.0000000000

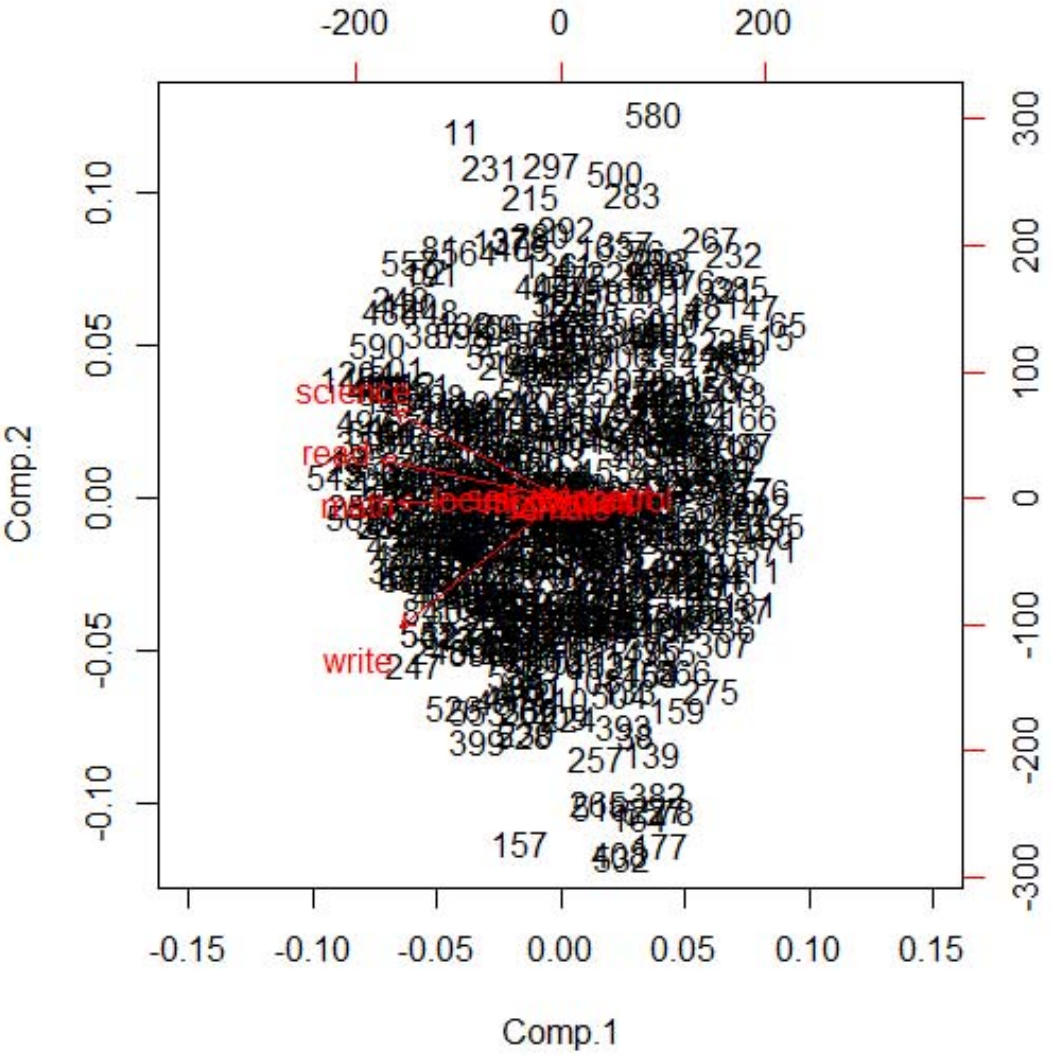
```
plot(pca.model)
```

Scree plot:



This plot shows that the first principal component accounts for the vast majority of the variance in the data, and the first four account for almost all the variance.

Biplot of the first two PCs:
```
biplot(pca.model, xlim=c(-0.15,0.15))
```



This biplot shows that science, reading, math and writing are strongly and positively correlated with each other and share a sign in the first PC. However, along the second PC, writing and science are negatively associated. Science and reading retain a weaker positive association, and math contributes little to differentiation in PC2.

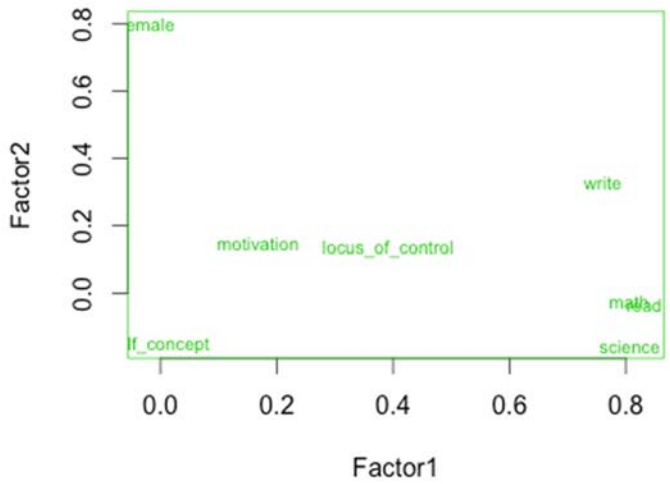
Factor analysis in R

```
# dataset can be a raw data matrix or a covariance matrix.
fit <- factanal(dataset, 3, rotation="varimax")
print(fit, digits=2, cutoff=.3, sort=TRUE)
```

Call:
factanal(x = dat, factors = 3, rotation = "varimax")

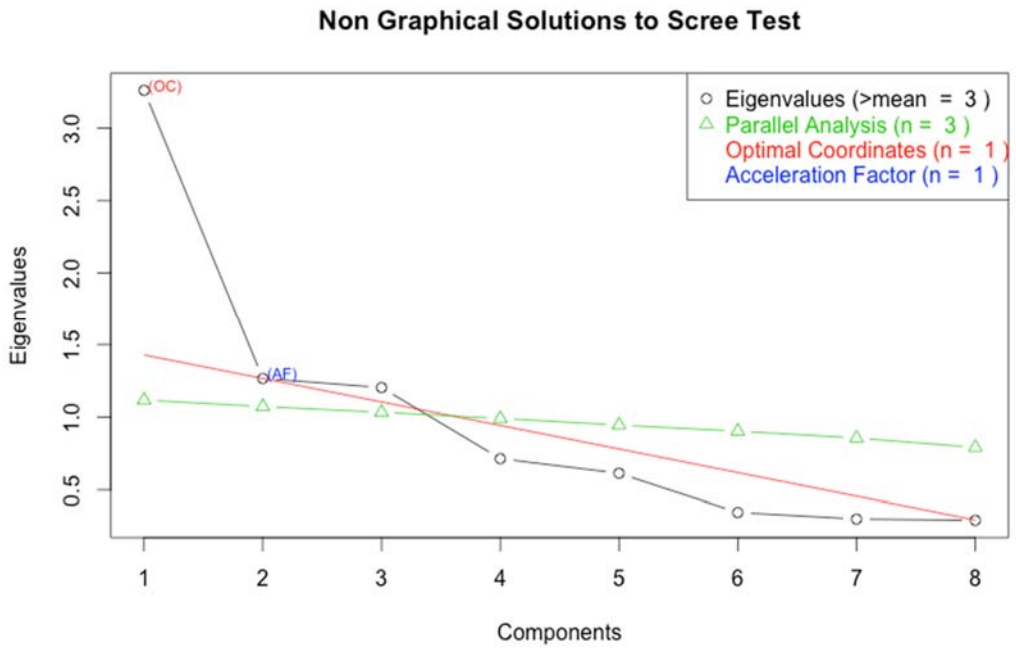
```
Uniquenesses:
locus_of_control    self_concept      motivation            read           write
            0.73            0.65            0.65            0.29            0.30
            math         science          female
            0.34            0.32            0.36

Loadings:
                 Factor1 Factor2 Factor3
read              0.83
write             0.76    0.33
math              0.81
science           0.81
female            0.80
self_concept              0.57
motivation                0.55
locus_of_control  0.39            0.31

                Factor1 Factor2 Factor3
SS loadings        2.75    0.83    0.77
Proportion Var     0.34    0.10    0.10
Cumulative Var     0.34    0.45    0.54

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 11.15 on 7 degrees of freedom.
The p-value is 0.132
```

```
# plot factor 1 by factor 2
load <- fit$loadings[,1:2]
plot(load,type="n") # set up plot
text(load,labels=names(dataset),cex=.7) # add variable names
```



Similarly to the results seen in the PCA biplot, the first axis is strongly driven by a positive correlation among writing, math, reading and science. These variables are negatively correlated with being female and self concept. On the second axis, we see a positive association between female and writing and negative relationships between female and science and female and self-concept.

```
library(nFactors)
ev <- eigen(cor(dataset)) # get eigenvalues
> ap <- parallel(subject=nrow(dataset),var=ncol(dataset), rep=100,cent=.05)
> nS <- nScree(x=ev$values, aparallel=ap$eigen$qevpea)
> plotnScree(nS)
```



Similar to the scree plot from PCA, we see the first factor has a much larger eigenvalue than the following ones and accounts for more of the variance in the data.

Problem 4: Use the Gazi University Student Evaluation Data Set to compute two test reliability measures (use only the responses in Q1-Q28). Intepret the results and discuss your findings in the study-specific context. For extra credit you can think about interpreting the impact of the course descriptive meta-data (Repeat, Attendance, Difficulty) and their potential impact on student responses.

```
require(psy)
dataset <- read.csv("C:\\Users\\Desktop\\gazi.csv", header=TRUE)
dataset <- dataset [,-c(1:5)]
c.alpha<-cronbach(dataset)
c.alpha
        $sample.size
```

```
[1] 5820
$number.of.items
[1] 28
$alpha
[1] 0.9887934
```

A Chronbach's alpha value of 0.99 is very high, and suggests high internal consistency.

```
library("psych" )
t.dataset <-as.data.frame(t(dataset [c(1:20),])) # using first 20 students (raters)
c.kappa<-cohen.kappa(t.dataset)
kappam.fleiss(t.dataset)
```

When the response variable is continuous, the intra-class correlation coefficient may be useful for instrument reliability. Either only subjects/topics can be considered as random effects ("oneway" model, default) or both subjects and raters are considered as randomly chosen ("twoway" model). When differences in raters' mean ratings are of interest, inter-rater "agreement" instead of "consistency" (default) type should be specified.

```
library(irr)
icc(dataset, model="twoway", type="agreement")
    Single Score Intra-class Correlation (ICC)
        Model: twoway
        Type : agreement

        Subjects = 28
        Raters = 20
        ICC(A,1) = -0.00189

        F-Test, H0: r0 = 0 ; H1: r0 > 0
        F(27,5.19) = 0.34 , p = 0.971

       95%-Confidence Interval for ICC Population Values: -0.003 < ICC < 0
```