Probability and Statistical Inference

Week 4 Exercise

1. Load/open your regression dataset (regression.sav)

   This dataset comprises a sample of 4,059 young people (aged 16) selected from 65 difference secondary schools from six inner London Education Authorities.This is a sub-sample from a much larger study undertaken by Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., et al. (1993) A multilevel analysis of school examination results, Oxford Review of Education, 19, pp. 425-433. The dataset has been specifically prepared to accompany Rasbash, J. et al. (2005) A User's Guide to MLwiN 2.0 (Bristol, Centre for Multilevel Modelling).

Variable	Description
School	A unique numeric identifier for each school
Student	A unique numeric identifier for each student
Normexam	Student's exam score at age 16, normalised to have approximately a standard Normal distribution and a mean of 0 and standard deviation of 1. (Note that the normalisation was carried out on a larger sample, so the mean in this sub-sample is not exactly equal to 0 and the variance is not exactly equal to 1).
Cons	A column of 1's. This is used in the multilevel modelling package MLwiN to represent the intercept in a statistical model.
Standlrt	Student's score at age 11 on the London Reading Test (LRT), standardised using Z-scores.
Girl	1 = girl, 0 = boy
Schgend	School's gender (1 = mixed school, 2 = boys' school, 3 = girls' school)
Avslrt	Average LRT score in school
Schav	Average LRT score in school, coded into 3 categories (1 = bottom 25%, 2 = middle 50%, 3 = top 25%)
Vrband	Student's score in test of verbal reasoning at age 11, coded into 3 categories (1 = top 25%, 2 = middle 50%, 3 = bottom 25%)

2. To read in the file we need to do the following:

```
library(foreign)
#Read in the file
regression <- read.spss("regression.sav", use.value.labels=TRUE, max.value.labels=Inf,
to.data.frame=TRUE)

#Setting the column names to be that used in the dataset
colnames(regression) <- tolower(colnames(regression))
```

3. We are interested in the following variables:
   Normexam – students score age 16
   Standlrt – students score at age 11 on the LRT
   Review these for normality
4. Investigate whether there is a difference between Normexam for students of different gender (grouping by Girl)
5. Investigate whether there is a difference between Standlrt for students of different gender (grouping by Girl)
6. Using Survey.dat
   - Investigate the normality of the following:
     i. perceived stress (tpstress)
     ii. positive affect (tposaff)
     iii. negative affect (tnegaff)
     iv. life satisfaction (tlifesat)
     v. self-esteem (tslfest)
7. Investigate whether there is a difference in the following for different gender values:
   - Positive Affect
   - Negative Affect
   - Life Satisfaction