MATH 4044 – Statistics for Data Science

Practical Week 4

Note: All tasks in this week's practical are to be performed in SAS Enterprise Guide.

Exercise 1

Data file for this exercise is based on a sample of 103 students who participated in a study on exam anxiety. The data is stored in a SAS data file called `examanxiety.sas7bdat` located in `mydata` library on the SAS OnDemand server. The data statement to access this file is `data=mydata.examanxiety`

Variables in that file are as follows:

Variable	Description
number	Subject ID
revise	Time spent revising
exam	Exam performance (percentage score)
anxiety	Exam anxiety questionnaire score (out of 100)
gender	1=male, 2=female

(a) Use ***Tasks > Multivariate > Correlations…*** to perform correlation analysis by *Gender*. Under 'Options' select both Pearson and Spearman and tick 'Fisher options' to obtain *P*-values and confidence limits. Under 'Results' tick 'Create a scatterplot for each correlation pair'.

Modify the code produced by the task to include a step of creating new formats for *Gender*, replacing '1' with 'Male' and '2' with 'Female'. Under PROC CORR, edit the PLOTS statement to produce the scatterplot matrix only, with histograms on the diagonal. Also add the NOSIMPLE option to omit simple statistics results. Run your version of the program to produce a new set of results.

Report and comment on your results. Would you recommend Spearman's rho over Pearson's correlation coefficient for any pair of variables? Explain briefly.

(b) Use ***Tasks > Regression > Linear Regression…*** to fit a simple linear regression model with *Anxiety* as the dependent variable and *Revise* as the explanatory variable.

Select *Gender* as a 'group analysis by' variable to obtain two models, one for males and one for females. Under 'Statistics' tick 'Confidence limits for parameter estimates'. You may also tick 'Partial correlations' if you wish.

Under 'Plots' choose 'Custom list of plots' and tick the following boxes:

- Residuals by predicted values plot
- Normal quantile plot of residuals
- Scatter plot with regression line.

Modify the code produced by the task to include a step of creating new formats for *Gender*, replacing '1' with 'Male' and '2' with 'Female'. Run your version of the program to produce a new set of results.

Report your results, including:

- Interpretation of slope and intercept
- Goodness of fit as measured by the coefficient of determination
- Inference for the slope
- Inference for overall model fit
- Assumption checking.

Comment on your results. How do the two models compare?