UCD School of Mathematics and Statistics

STAT40840: Data programming with SAS
Dr Laura Kirwan

Review

End of Semester Exam

- There will be a two hour lab exam on May 18th from 9am – 11am.
- The exam will be open book
- Build on the sas code that we've developed over the module.
- Use the SAS help documentation
- Include annotations so that I know what you are doing (make it easy for me to give you marks)
- In some questions, I ask you to comment on the output – this means that I want you to interpret the output.

End of Semester Exam

- The exam is worth 70% of your grade.
- There will be two questions, each worth 50 marks.
  1. Data management and manipulation
  2. Data analysis
- You will work with datasets that I provide in the Assessment folder on blackboard.
- You will be required to submit two script files and two output files on blackboard

1. Data Management and Manipulation

- Structure of a SAS dataset
- Applying formats and labels
- Subsetting datasets
- Sorting datasets
- Reading in SAS dataset
- Reading in external dataset
- Using one SAS dataset to create another
- Merging datasets
- Creating new variables
- Creating output

1. Data Management and Manipulation

Structure of a SAS dataset

- Permanent vs temporary dataset
  - In a permanent library or the work folder
- SAS dataset contains variables (columns) and observations (rows)
- Descriptor portion (proc contents)
  - Names, attributes, labels, etc
- Data portion (proc print)
  - Data values
- Missing Values
  - Numeric (full stop), Character (blank)

Applying formats and labels

- Formats can be applied using statements within the DATA step, statements within a PROC or by applying PROC Format
- Formats used to change how values are displayed in a report
- Range of character, numeric, date and user-defined formats
- Labels, titles and footnotes can all be added to enhance reports

Subsetting datasets

- the WHERE statement to print only a subset of the data, namely the subset meeting the condition specified in the WHERE statement
- Specifying conditions (<, >, ne, contains, etc)

Sorting datasets

- Using the SORT procedure
- based on the values of one or more variables, specified in the BY statement
- by default, SAS sorts the values of the variables appearing in the BY statement in ascending order. If you want them sorted in descending order, you need to use the BY statement's DESCENDING option.
- If you want to group any procedure using a BY statement, then the data must first be sorted accordingly

1. Data Management and Manipulation

Reading in Data

- SAS dataset from a permanent library
- Instream data using cards or datalines
- Importing external data (e.g. from a csv file) using infile statement
- Creating a new SAS dataset from an existing SAS dataset using the SET statement

Merging datasets

- Concatenate datasets using the SET statement (like and unlike structured)
- Merge datasets using the MERGE and BY statements
  - one-to-one
  - one-to-many
  - Non-matches

1. Data Management and Manipulation

Creating new variables

- Use assignment statements in the DATA step
- Numeric operators
- In-built SAS functions
- Conditional processing (IF / THEN / ELSE)

1. Data Management and Manipulation

Creating output

– you should be able to use the PRINT procedure to :

– apply a title or footnote to a printed page of SAS output
– use the VAR statement to print a subset of the variables in a SAS data set
– use the NOOBS option to suppress the printing of the observation number
– use the LABEL option to print variable labels
– use the SPLIT= option to split labels used as variable headings
– use the ID statement to emphasize key variable(s)
– use the FORMAT statement to print a variable in a previously specified format
– use the WHERE statement to print only a subset of the data, namely the subset meeting the condition specified in the WHERE statement
– use the SUM statement to specify the sum of certain variables
– use a BY statement to print observations in groups based on the values of the different BY groups

2. Data Analysis

- Numeric and graphical descriptive Statistics
- Correlations
- Regression
- Linear model
- Generalised linear model
- Linear mixed model
- Generalised linear mixed model

2. Data Analysis

Numeric and graphical descriptive Statistics

- Use the UNIVARIATE procedure to produce summary statistics
- Use the normal option to test for normality and the QQPLOT statement to produce a qqplot
- Use the ODS graphics statement to produce graphical output
- Use the ODS select statement to select the printing of particular output tables
- Use the ODS output statement to save output tables as datasets

2. Data Analysis

Correlations

- Use the CORR procedure to produce a correlation matrix
- Produce an output scatterplot matrix
- Use the CORR procedure to estimate partial correlations

2. Data Analysis

Regression

- Use the REG procedure to fit simple linear and multiple regressions
- Produce output with diagnostic plots
- Create an output dataset with residuals and predicted values

2. Data Analysis

Linear model

- Use the GLM procedure to fit a linear model
- Include categorical explanatory variables using the class statement
- Create an output dataset with residuals and predicted values

Generalised linear model

- Use the GENMOD procedure to fit a linear model where the data is not necessarily normally distributed
- Use the link= option to specify how the mean of a population depends on the linear predictor through the nonlinear link function
- Use the dist= option to specify the response probability distribution function

Linear mixed model

- Use the MIXED procedure to fit a linear mixed model
- Use the random statement to include a g-side random effect
- Use the repeated statement to include an r-side random effect
- Use the type= option to specify the structure of the variance-covariance matrix

2. Data Analysis

- Generalised linear mixed model

- Use the GLIMMIX procedure to fit a generalised linear mixed model
- Use the link= option to specify how the mean of a population depends on the linear predictor through the nonlinear link function
- Use the dist= option to specify the response probability distribution function
- Use the random statement to specify the g-side and r-side (using the _residual_ keyword) random effects