Cork Institute of Technology
Institiúid Teicneolaíochta Chorcaí

Semester 2 Examinations 2019/20

Module Title:	Data Analysis with R
Module Code:	STAT9006
Internal Examiner:	Dr Seán Lacey
Instructions:	Write a Report
Duration:	2 hours 30 minutes
Sitting:	Assessment 04, April 2020

Assessment 04: Dataset, on Canvas, lists estimates of body fat percentage determined by underwater weighing and various body circumference measurements for 252 men. A variety of popular health books suggest that the readers assess their health, at least in part, by estimating their percentage of body fat. In Bailey (1994), for instance, the reader can estimate body fat from tables using their age and various skin-fold measurements obtained by using a caliper. Other texts give predictive equations for body fat using body circumference measurements (e.g., chest circumference) and/or skin-fold measurements - e.g., Behnke and Wilmore (1974), pp. 66-67; Wilmore (1976), p. 247; or Katch and McArdle (1977), pp. 120-132.

Requirements: In your assessment, you are tasked with compiling a report investigating whether or not body fat (%) is dependent on subject's age (years), chest circumference (cm), density (g/cm^3), knee circumference (cm) and weight (lbs), based on the data in *Assessment 04: Dataset,* on Canvas. Your report should use *R* to compile the appropriate tables/graphs/tests/etc. You are expected to **explain all concepts and procedures** used in the statistical inference on the data.

Note:

- Report and interpret your findings using 0.05 level of significance;
- Outliers are present in the dataset. Do not remove them;
- The initial multiple regression model will contain five explanatory variables. The final model will contain three explanatory variables.

Breakdown of marks (200 marks)
Your report should have the following sections:

- 15 marks for a clear introduction to the assessment with reference to the:
  - **Type of data** presented, **type of study**, **sample size**.
- 15 marks for presentation of descriptive statistics (sample size, mean, standard deviation, median, 1st quartile, 3rd quartile, min, max) and determination of normality on all measurements (mean, median, skewness, verdict, *p*-value). Note: **no explanation required on statistics outputted for this section**.
- 40 marks for the **scatterplots**:
  - 20 marks for presentation of five plots describing the relationship between the response variable and each of the predictors (explanatory variables);
  - 20 marks for the appropriate **correlation coefficients** (with brief explanations) for each scatterplot.

- 20 marks for presentation and interpretation of the **correlation matrix** (predictors only):
  - 10 marks for clearly interpreting the correlation matrix;
  - 10 marks for a general comment on **multi-collinearity**.
- 50 marks for using **model selection** to develop a regression model:
  - 40 marks for presentation and explanation of the full model (adjusted R^2, regression coefficients, standardised regression coefficients, p-values);
  - 10 marks for applying and interpreting the Variance Inflation Factor (VIF) to determine whether or not multi-collinearity is present.
- 25 marks for **presentation of the final model** [adjusted R^2, regression coefficients (with 95% confidence interval), standardised regression coefficients, p-values]. Note: **no explanation required on statistics outputted for this section**.
- 25 marks for determining whether or not the conditions placed on the **residuals** are violated for the final model:
  - 10 marks for presenting the relevant plots;
  - 15 marks for correct conclusions (with corresponding p-values).
- 10 marks for **two conclusion comments** on the model presented. Highlight whether or not there are any concerns with the model.

Upon completion of your report, upload your report **AND** R script as an assignment submission on Canvas