P8130 Biostatistical Methods I – Fall 2022

Final Project

Due: December 16, 2022 at 5:00pm

Guidelines for Project Submission

This group project must be submitted through Courseworks before the deadline. Email submissions WILL NOT be accepted and will receive a score of 0 for all group members!

All graphs, output, and interpretations must be included in ONE PDF file, otherwise it will not be graded. In a separate attachment, you must also submit your R/Rmd code used in your project.

General Writing Instructions

Your project should not exceed 5 double-spaced pages using 11 or 12-point font, EXCLUDING figures and tables, references, appendix, that can be placed at the end of the main text. Be selective in your output and visual displays!

Your brief report should be structured as a publishable research article containing the following sections:

- Abstract (condenses a brief introduction, brief description of methods, and main results into a one-paragraph summary)
- Introduction (brief context and background of the problem)
- Methods (data description and statistical methods)
- Results
- Conclusions/Discussion

Your findings should be written as for an informed (but non-statistical) audience (NO FORMULAS!). Each figure and table should be of publishable quality and well notated, I.e., labeled and/or captioned.

Grading Instructions

The rubric attached will be used to evaluate the project. This is a group project and collaborations within your group are essential and provide great practice for your career.

Academic dishonesty or lack of contribution to the team effort

will be penalized and reflected in individual grades.

Body fat is a useful predictor in many medical situations, though it is not always straightforward to measure. (In fact, body fat is usually estimated using an underwater weighing technique.) This data set contains some variables that are much easier to measure (i.e., with a scale or measuring tape). Your goal is to build a model of some simple measurements that doctors can use to predict a patients' body fat.

You must choose one of the variables in BLUE below as your outcome. The remaining BLUE variables may NOT be used as a predictor in your model. (These are the "difficult to measure" variables that you wish to build a model to estimate.)

(This data set has measurements of 252 men.)

Variable	**Description**
id	Case number: 1-252
bodyfat_brozek	Percent body fat using Brozek's equation:
bodyfat_siri	Percent body fat using Siri's equation:
density	Body density (gm/cm^3)
age	Age (years)
weight	Weight (lbs)
height	Height (inches)
neck	Neck circumference (cm)
chest	Chest circumference (cm)
abdomen	Abdomen circumference (cm) – measured "at the umbilicus and level with the iliac crest"
hip	Hip circumference (cm)
thigh	Thigh circumference (cm)
knee	Knee circumference (cm)
ankle	Ankle circumference (cm)
bicep	Extended biceps circumference (cm)
forearm	Forearm circumference (cm)
wrist	Wrist circumference (cm) – measured "distal to the styloid processes"

As a team, write up your findings in the report. In this report you should describe your final model and interpret its parameters in an accurate and useful manner. It is expected that you would first examine the marginal distributions and pairwise relationships between variables (e.g., to check to see whether any nonlinearities are immediately obvious), that you would explore several candidate models for predicting body fat, and explain why you selected your model. Also, you should check for violations of regression model assumptions, influential observations, multicollinearity, etc.

Your report will be evaluated based on how well you communicate insights about how these body measurements related to body fat, so it would be helpful to be clear about your motivation for carrying

out certain analyses as well as to be clear about interpretations of fitted model parameters. Your report should include a table summarizing parameter estimates associated with your final fitted model, characterizing predictor variables in a way that a reader can clearly understand.

Below you'll find some aspects to be addressed in your report. These are just a few suggestions, but feel free to add your own input/creativity to the analysis:

- Data exploration: descriptive statistics and visualization. You might want to, for instance:
    - Include a descriptive table with summary statistics for all variables;
    - Explore the distribution of the outcome and consider potential transformations (if necessary);
    - See if there are any unusual observations and consider them as potential outliers/influential points.
- In your multiple linear regression model, be watchful for variables that are highly correlated and be selective in the variables you will include in your analysis.
- Consider selective interactions between variables.
- DO NOT IGNORE MODEL DIAGNOSTICS.