Linear Regression Assignment

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(pastecs)
library(knitr)
```

Linear Regression Assignment

General instructions

You must submit the following.

Stata

- Stata Log File
- Stata Do File
- Report in Word doc form with tables/figures cut and pasted from Stata output file

R

- RMarkdown file
- Knit report in HTML or PDF

Part 1 - Statistical Model Building

The following analysis uses the Can-Path Dataset available in Canvas.

Your boss tells you to conducted an analysis examining the association between different social and behavioural factors on Body Mass Index using the Can-Path dataset. Your goal as a epidemiologist is to understand factors associated with BMI that can either be intervened on (exposures of interest) or that may be effect modifiers in the association between an exposure of interest and BMI.

It is crucial when discussing obesity and weight status that we avoid adding to stigma as researchers and epidemiologists. There are a number of health consequences to weight stigma that we must attempt to avoid. At the same time, we need to work with people to try and improve health. Some resources here:

1. Puhl, R M., Wharton, C M. Weight Bias: A Primer for the Fitness Industry. Health & Fitness Journal. 2007; 11(3), p 7-11. https://doi.org/10.1249/01.FIT.0000269060.03465.ab
2. Phelan SM, Burgess DJ, Yeazel MW, Hellerstedt WL, Griffin JM, van Ryn M. Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. Obes Rev. 2015;16(4):319-326. https://doi.org/10.1111/obr.12266.
3. Puhl RM, Himmelstein MS, Pearl RL. Weight stigma as a psychosocial contributor to obesity. Am Psychol. 2020;75(2):274-289. https://doi.org/10.1037/amp0000538

Your task in this analysis is to determine which variables should be included in a regression model with BMI as the outcome.

1. Identify the primary outcome of interest and the main predictor(s) of interest
2. Draw a causal diagram for your full causal model
3. Assess potential confounding variables on the relationship between the outcome and exposure of interest using two approaches we learned in class:

- A DAG/theory based approach
  - Pick confounder variables from DAG first, list 4-5 confounders (or more if you want) in DAG. Do not pick Collider.
  - Identify at least one variable that MUST stay in your model no matter what, because it is important from common sense or literature. Give 1-2 sentence explanation for retaining such variable(s).
- A data driven/model fit approach

4. Conduct cross-tabulations of the outcome variable with categorical predictors.
   - Are you concerned about any of the cell sizes?
5. Identify and fit the best model(s) evaluating the association between your exposure and outcome of interest. Describe why you included/excluded predictors and how you compared between models to arrive at your final model.
   - Insert your model output below.
   - Interpret each of the coefficients and odds ratios.
6. Evaluate the model and if the model meets the assumptions of linear regression. Use regression diagnostics.
   - For the rest of the variables you picked, use them to drop some (or all) using model selection, either using AIC, BIC, or likelihood ratio test. Remember, the likelihood ratio test can be only used for comparing nested models.
   - Present 1 or two final models with appropriate conclusion.

Part 2 - Format and Effort

General Formatting

- A combination sentences/paragraphs with some bullet points is appropriate.
- Include a list of references where appropriate. For this assignment, you do not need to worry about providing references to the scales/items within the dataset.

Overall

- Assignments will be evaluated based on the overall effort and thoroughness of the assignment, attention to details, and overall presentation of results.