Predictive Modelling with Linear Regression

# Assignment

The goal of this project is for you to perform data analysis, predictive modelling, and diagnostics on a dataset of your choosing. The statistical method used must be linear regression. You will create a report in Word, R Markdown, or similar software.

The report should contain the following key elements:

## 1. Gathering Data

First you must find a dataset with a quantitative response (numerical, not categorical), and several predictors (at least 3). The hardest part of this project may be finding the right dataset for your report, so make sure you spend a good deal of time finding one. Here are some resources for you:

<https://lib.conestogac.on.ca/statistics>

<https://vincentarelbundock.github.io/Rdatasets/datasets.html>

<https://github.com/rfordatascience/tidytuesday>

These are just a few examples. It is a good idea to search for your own and find one that interests you.

After finding a dataset, write several paragraphs describing the data (how many observations, the predictors/response) and explaining what you hope to predict using it.

## 2. Initial Modelling

Start by considering what variables you believe make sense to include as predictors. Think about possible non-linearities and interactions between variables. Write a paragraph describing which predictors you believe are the most important for predicting your response and your reason for including them in your model. You should also discuss any variables you don’t think will be useful and why.

After this, run your regression in statistical software and report the coefficients in a well formatted table.

If your data is giving you errors when you try to run the regression, you may need to “clean” the data.

If you use R I suggest this resource for cleaning data: <https://r4ds.had.co.nz/wrangle-intro.html>

## 3. Diagnostics

Perform some of the diagnostic tests you’ve learned in this course on your model. Do the major assumptions of a linear regression model hold? Include the plots and the results from any tests you used, as well as a brief paragraph describing your findings. Do not worry if the assumptions seem violated, the focus here is an honest appraisal of the model you chose in part 2. Discuss the changes you might make to your model to improve the results.

## 4. Model Selection

Now consider improving the model, for example, are there any non-linearities you detected in the step 3? Afterwards, perform variable selection using one of the techniques you learned in the course.

Choose one model (the best performing) and include a screenshot of the summary statistics for it. Briefly describe the reasons why you chose this model.

## 5. Prediction and Summary

Use your final model to perform predictions with statistical software. Choose a prediction that makes sense (don’t choose something that is an extrapolation). Describe the results and summarize your model.

# Submission Instructions

* This assignment is to be completed by each student individually.
* Include your dataset and any code/commands you used to do this project
* Your assignment must be submitted to eConestoga by the date/time stated in your Instructional Plan.
* Include a cover page with the following information:
  + Your full name
  + Student number
  + Course number