Statistical Learning Assignment - Semester 1, 2022

- **INSTRUCTIONS**:

1. The assignment must be typed (not handwritten). You may use either Microsoft Word (or similar) or R markdown in RStudio for the assignment. Note that the final project will require the use of R markdown. **When answering this question, it should be no longer than 10 A4 pages [single sided] with a font size no smaller than 11 point.**
2. The assignment due date is listed on the Wattle (Turn-it-in) site. Upload the assignment through Wattle using Turn-it-in. You should submit your assignment in **two different parts**. **If you are using R markdown:**
   (a) A pdf file [or HTML file] of your assignment (this should include important R code to highlight what you have done).
   (b) A '.Rmd' file [an R markdown file].

   If you are using Microsoft Word (or similar):
   (a) A Word file of your assignment (this should include important R code to highlight what you have done).
   (b) A '.R' file of your R code.
3. In answering the questions, write your answers clearly and succinctly. Use appropriate graphs and tables when you think they help to describe your point or thinking process. Do not just "print" a set of results. Every result should be discussed and have a reason for being presented. **No points will be awarded unless you clearly discuss what you are doing.**
4. No late assignments will be accepted.
5. **The assignment you turn in must be your own work. This includes all computer code, writing, and mathematics. Please see the university resources on Academic Integrity** https://www.anu.edu.au/students/academic-skills/academic-integrity **for more details.** For this assignment, you do not have to cite material from either of the two textbooks or class slides. Any other material should be appropriately cited.
6. **Have fun with the exploration!**

---

1. **(100 points)** We will explore some of the techniques we are considering by examining data on health insurance claims (in total USD over a 24 month period) by women with coronary heart disease for $n = 603$ cases. The data set consists of the following variables:

   - Y = Claims(\$)
   - x_1 = Age($years$)
   - x_2 = Procedures (number)
   - x_3 = Prescribed Drugs (number)
   - x_4 = Emergency Room Visits (number)
   - x_5 = Complications (number)
   - x_6 = Comorbidities (low, medium, high)
   - x_7 = Duration (days)

Consider a multiple regression model to examine the relationship between *Claims* (Y) and their covariate information ($\boldsymbol{x}$). The first 300 cases will be the training data and the second $n - 300$ cases will be the test data. For this assignment set $\alpha = 0.05$.

(a) **(21 points)** Using the training data, conduct an exploratory data analysis. In doing your analysis make sure to identify any unusual points and discuss why they are unusual. For this assignment do not remove any unusual points, only comment on them (if they exist). In addition to visualisations of the raw data, consider the natural log transformation of the response. You may also consider any transformations of the covariates. For the rest of the assignment (unless stated in the question), if you believe the transformations are appropriate (provide justification - this can simply be a discussion), use those transformations.

(b) **(8 points)** Using the training data and the untransformed covariate *Prescribed Drugs*, based on traditional regression approaches (possibly: t-tests, F-tests, etc.), determine if there exists a non-linear (quadratic, cubic, etc.) between the covariate and the response. How flexible should the model be? Make sure to fully outline any tests and conclusions (this may include appropriate mathematics, code, etc.).

(c) **(8 points)** Using your personal training and personal testing data, along with the notion of squared error loss, determine if there exists a non-linear (quadratic, cubic, etc.) relationship between the untransformed covariate *Prescribed Drugs* and the response. How flexible should the model be? Discuss the results compared to 1b.

(d) **(8 points)** Consider all the covariates which we are using in this assignment. Using the training data and traditional regression approaches, determine if any of the variables are statistically significant. Are you able to reduce the model (i.e. not use all the covariates)? Here you do not need to consider any non-linearities or interactions. Make sure to fully outline any tests and conclusions.

(e) **(8 points)** Based on the ordering of the covariates in your final model (and potentially along with the final process for the determination of that model) in the previous question (1d), using the training and testing data, along with the notion of squared error loss, determine which covariates should be included in the model. Discuss the results compared to 1d.

(f) **(8 points)** Consider all the covariates which we are using in this assignment (in addition to results from 1d and 1e). Using the training data and traditional regression approaches, determine if *Cormbidities* has a statistically significant interaction with any of the other covariates. You may have up to six interactions in your model. Make sure to fully outline any tests and conclusions.

(g) **(8 points)** Based on the ordering of the covariates in your final model (and potentially along with the final process for the determination of that model) in the previous question (1f), using the training and testing data, along with the notion of squared error loss, determine which interactions should be included in the model. Discuss the results compared to 1f.

(h) **(10 points)** Consider all the covariates which we are using in this assignment. Based on a multiple regression framework with normal errors ($\epsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2)$), you may now consider any regression modelling that you wish using the training data. You may also consider any type of model selection approach (i.e. traditional or based on squared-error loss for the testing data). Make sure to fully outline any tests and conclusions. Calculate the mean-squared error on the testing data.

(i) **(21 points)** Provide a full discussion of your final model from 1h. This may include, but is not limited to, discussions of the coefficients, visualisations of the fitted model, and model checking.