Department of Statistics
STATS 330 Statistical Modelling
Assignment 4 (2020; Semester 1)

Total: 50 marks **Due: 23:59 NZDT, June 7, 2020**

Notes:

(i) Write your assignment using R Markdown. Knit your report to either a Word or PDF document.
(ii) Create a section for each question. Include all relevant code and output in the final document.
(iii) 7 presentation marks will be used; these will be deducted if offences occur. Please think of your markers—keep your code and plots neat and remember to check your spelling.
(iv) Please remember to upload your Word or PDF document to Canvas by the due date.
(v) Please remember to upload your R Markdown file to Canvas before the deadline, too. If the markers identify an error in your work, being able to run the code you have written can help determine what you did wrong
(vi) **Coversheet**: please make sure you do **one** of the following else your assignment will not be marked:
   (i) Sign the Cover Sheet and combine with your assignment document (pdf or Word) into a single file before submission, OR
   (ii) Type or write for the following at the beginning of your assignment: Your name (as it appears in Canvas), your UPI, and the following statement: "I have read the declaration on the cover sheet and confirm my agreement with it."
(vii) **Remember to comment on all your output using your own words!**

1. [25 marks] **Womens' BMI** In the previous assignment a quadratic curve was fitted to the BMI of female Europeans. Of interest is θ, the age in which BMI is a maximum. From the scatter plot, we saw it was about 60 years old. We want to obtain an approximate 95% confidence interval for θ.

   (a) Read in the data into a data frame. It is a good idea to sort by `age`. Fit a linear model with a quadratic in `age`. Giving the details, obtain the point estimate $\widehat{\theta}$. [Hint: if you can't do any calculus then look up the formula for the roots of a quadratic ($ax^2 + bx + c = 0$) and use the fact that the minimum/maximum lies midway between the roots.] Produce a scatterplot of the data and overlay the fitted linear model and $\widehat{\theta}$. Comment. [5 marks]

   (b) Use parametric bootstrapping to obtain an approximate 95% confidence interval for θ. Use $\widehat{\sigma}$ from your fit in (a) and show some details. Note: make your answer reproducible. [6 marks]

   (c) Use nonparametric bootstrapping to obtain an approximate 95% confidence interval for θ. Then compare your answer to the parametric bootstrap. [6 marks]

   (d) Install the `msm` package and apply the delta method. Then compare your answer to the previous answers. [5 marks]

   (e) The data are a subset of a prospective observational study conducted in the mid-1990s. Having 10,500+ participants and about 28% females, the study can be considered an approximate random sample of the New Zealand working population at the time. Altogether there were four major ethnic groups: "Europeans", "Maori", "Polynesian" and "Others". Comment on the generalizability of the results to here and now. [3 marks]

2. [5 marks] **Dredging** Using the model called `fm1` in the `LifeCycleSavings` online help file as the 'global' model object, use `dredge()` to obtain a model selection table of models.

   Notes:

   - Your table must only contain models that include the variable `ddpi`. That is, models without `ddpi` must be excluded.
   - Use BIC to measure how good models are, not AIC.

3. [20 marks] **Using smooth.spline()** Consider the following R code to generate data coming from a quadratic trend and smoothing it.

   (a) Substitute the last 3 digits of your student ID number into `set.seed()` below and run the code and obtain a plot. Comment. [3 marks]

```r
# Generate the 'original' data set.
set.seed(123)  # Substitute the last 3 digits of your student ID number!!
n <- 100
X <- scale(3 * (1:n)/n, scale = FALSE)
myfun <- function(x)
    2 - x + 3*x*x
Y <- myfun(X) + rnorm(n)
plot(X, Y, col = "blue")
fit <- smooth.spline(X, Y, df = 3 , all = TRUE)
lines(fit, lty = 1, col = "darkgreen", lwd = 2)
```

   (b) Add smooth curves corresponding to `df = 2` and `df = 20` to your plot. Comment. [2 marks]

   (c) For a wide range of values of `df` from 2 to n plot the mean residual sum of squares $n^{-1} \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ versus `df` (or some more suitable function of `df`). You should smooth `Y` versus `X`. Comment. [5 marks]

   (d) Add to your (c) plot the mean residual sum of squares corresponding from new (test) data generated from the model. Your plot should look a bit like the figure on Slide 32 in Handout 15. Comment. [5 marks]

   (e) Let `smooth.spline()` determine the 'best' smoothing parameter—use the default which is GCV, but set `all = TRUE`. What value of `df` does that correspond to? Plot the scatter plot with the smoother going through it. Comment. [3 marks]

   (f) Comment on this whole question—why is relevant to statistical modelling? [2 marks]