STAT202 Assignment 3: Multiple linear regression II

Due on 14th August 1 pm

This assignment deals with the analysis of the `aquatic_toxicity.xlsx` dataset, which is used to predict **LC50** (concentration that causes death in 50% of test *Daphnia magna* over a test duration of 48 hours). There are 8 molecular descriptors that act as predictors: **tpsa**, **saacc**, **h_050**, **mlogp**, **rdchi**, **gats1p**, **nn** and **c_040**. Details can be found in https://journals.sagepub.com/doi/pdf/10.1177/026119291404200106

1. As always, create a new folder for your work and also a new RStudio project in that folder. Download the `aquatic_toxicity.xlsx` files in your folder.

2. Use the function `read_excel()`, from the `readxl` package, to read the file and name the object `toxic`.

3. You will use random selection of 500 rows of the dataset, which depends on your student ID. Name your sample `my_toxic`.

4. Estimate the correlations between all variables[1] and, based on the correlations, choose three variables to predict `lc50`. Using `ggpairs` (from the `GGally` package) create a scatterplot matrix with your three predictors and `lc50`. Explain in 50 words the relationships you observe in those plots.

5. Fit the model with your chosen predictors (call it `m1`) and write down the coefficients, the adjusted-r2 and residual standard error.

6. Now use the `leaps` package to fit all regression subsets. If working in your laptop, you'll have to install this package. Plot the results of regression subsets, and explain which predictors are contained in the best model for each number of predictors.

```
library(leaps)
all_models <- regsubsets(model with all predictors, data = my_toxic)
plot(all_models, scale = "adjr2")
```

7. Now fit the best model from the previous step (hint: `round(summary(all_models)$adjr2, 2)` will help you), call it `m2`, and compare its adjusted-r2 and residual standard error with `m1`. Discuss in 50 words the similarities and differences between the results of the 2 models.

8. Create diagnostic plots for the residuals of model `m2`. Check the model for assumptions for the residuals. Explain in no more than 70 words if there is anything unusual or wrong.

[1] If you are feeling adventurous or prefer something prettier than `cor()` give the `corrplot` package a try. You'll need to install it and can see a tutorial https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

9. Obtain the predicted LC50 (95% **prediction** and 95% **confidence** intervals, the distinction is presented in Lecture 10 slides) for chemicals with the following characteristics:

```
tpsa   saacc h_050 mlogp rdchi gats1p    nn c_040
69.97  97.43     0  3.12  3.72   1.26     0     2
 3.24   3.12     0  9.15  5.49   1.56     1     0
```

You will need to create a `tibble` and use the `predict()` function for this. Have a look at slides in lecture 10 to get an idea of how this should be done.

10. Let's have a look again at `all_models` (from question 6). Plot it again using `scale = "r2"` and `scale = "bic"`. Compare the 3 plots (adjr2, r2, bic) and explain in 50 words how/why they differ.

11. We will use the formula $(\mathbf{X}`\,\mathbf{X})^{-1}\,\mathbf{X}`\mathbf{y}$ to obtain the vector of regression coefficients (this was covered in class). Use R's matrix algebra capabilities to reproduce the coefficients estimated in `m2`.

12. Put together your answers, figures and code in a Word document and upload it to Learn.