Assignment 4: Variable selection

Luis A. Apiolaza
Due on 1 pm 21st August

This time we will work with 2 datasets:

- The first one contains biometric data for the !Kung San people[*] (kungsan_full.csv): **height** (height in cm, our response variable), **weight** (weight in kg), **age** (in years) and **sex** (female and males).
- The second one (white_wines.xls) refers to Portuguese wines, where the response variable is a **quality** score (from 1 to 10), for which we have eleven chemical composition predictors: fixed acidity (**fix_acid**), volatile acidity (**vol_acid**), citric acid (**cit_acid**), residual sugar (**res_sugar**), chlorides (**chlorides**), free sulphur dioxide (**free_sulphur**), total sulphur dioxide (**total_sulphur**), density (**density**), acidity (**pH**), sulphates (**sulphates**), and alcohol (**alcohol**).

Remember to create a folder for this assignment and an RStudio project in that folder. We will read the first data file directly from a web server (like you did in assignment 1).

1. Read the `kungsan_full.csv` dataset from `http://stats.apiolaza.net/data/kungsan_full.csv` and call it `kungsan.`

2. Create a new variable called `weight2, which is the` squared version of `weight` `(weight^2).` Convert the variable `sex` to a factor, using the `factor()` function (this ensures the variable is treated as categorical), keep only the individuals who are 12 years or over, and take a sample of 400 observations based on your student ID. Assign this sample to the name `my_kungsan`.

3. Create a scatterplot matrix with `height, weight`, `weight2` and `sex`. **Remember that height is the response variable**. Explain in no more than 100 words the relationships you observe in that plot (this time also including relationships between predictors).

4. Fit the following models to predict height: `m1` uses weight only, `m2` uses `weight` & `weight2`, and `m3` uses `weight`, `weight2` and `sex`[†]. Check the variance inflation for `m2` and `m3` (and explain what these factors mean); you will need the `check_collinearity()` function for this.

5. Create diagnostic plots for the residuals of `m1` and `m3`, checking the models for residual assumptions. Explain in 75 words how you think this model meets the assumptions, referring to the names of the specific plots you are basing your answer on.

[*] A description is available here https://en.wikipedia.org/wiki/%C7%83Kung_people

[†] Nerdy note: R provides other ways of fitting a second-degree polynomial. It is possible to use: `height ~ weight + I(weight^2)` in the formula to get `height ~ weight + weight2` without needing to create a new `weight2` variable.

6. Now center the `weight` predictor (call it `weight_c`) and create a squared version of `weight_c` (call it `weight_c2`). Create a scatterplot matrix with `weight_c`, `weight_c2`, `sex`, and `height.` Does it look different from the scatterplot matrix in question 3? Explain in no more than 50 words.

7. Fit `m4` with `weight_c`, `weight_c2` and `sex` as predictors. Check again the variance inflation factors for the predictors.

8. Predict height (and give a **confidence** interval and a **prediction** interval) for a 50 kg male and a 50 kg female **using model `m4`**. Assume that the average weight for the population is 36 kg.

9. Read the `white_wines.xlsx` file using the read_excel() function.

10. Take a sample of 4500 wines and fit model `w1` to predict wine quality using **all** available predictors. Present a summary of the model and point out adj-R2, residual standard error and variance inflation factors (VIF) for all slopes.

11. Choose the best set of predictors for wine quality using `regsubsets()` relying on BIC, fit that model as `w2` and present a summary of the model and point out adj-R2, residual standard error and VIF for all slopes.

12. Fit model `w3` by removing from `w2` the predictor with the highest VIF. Present the new VIF and adjusted-R2. In your opinion, is this a better or a worse model? Justify your answer in no more than 50 words.

13. Include all the answers and code used to obtain them in a Word document and submit it through Learn. Put the code you used to answer the questions next to the answers. Remember to use your own words when answering the questions.