Assignment 4: Variable selection

Luis A. Apiolaza
Due on 12 noon 15 September

This time we will work with 2 datasets:

- The first one contains biometric data for the !Kung San people[1] (`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.csv`) 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. As in the previous assignment, we will read the data file directly from a web server.

1. Read the datasets from `http://stats.apiolaza.net/data/` and subset only the values that correspond to your student code number to a dataset called `my_kungsan`. For example,

```
kungsan <- read_csv('http://stats.apiolaza.net/data/kungsan_full.csv')
set.seed(your student id number)
```

2. Take a sample of 540 observations call it `my_kungsan` and create a squared version of `weight (weight^2)` in that dataset. The new variable should be called `weight2`.

3. Create a scatterplot matrix with `weight`, `weight2`, `sex`, and `height` using the `ggpairs()` function from the `GGally` package. Explain in no more than 100 words the relationships you observe in that plot (this time also including relationships between predictors). It is a good idea to list `height` last in the list of variables to plot, as it then will be on the y-axis for reading the scatterplot matrix.

4. Fit the following models to predict height: `m1` uses weight only, `m2` uses `weight` & `weight2`, and `m3` uses `weight`, `weight2` and `sex`[2]. Check the variance inflation for `m2` and `m3` (and explain what these factors mean); you will need the `vif()` function from the `car` package for this.

5. Create diagnostic plots for the residuals of `m1` and `m3`, checking the model 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.

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

[2] Nerdy note: R provides many ways of fitting a second-degree polynomial. Besides what the question suggests, it would be possible to use: `height ~ poly(weight,2)` and `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 2? 20 words maximum.

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) for a male and a female who weighs 50 kg using model `m4`. The average weight for the population is 36 kg.

9. Take a sample of 4800 wines and fit a 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.

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

11. 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.

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