STAT 462 – Applied Regression Analysis
Fall 2017, Homework 5

Applied part (100 points)

For this last homework your report can be up to 8 pages (instead of the usual 5 pages).

Consider again the dataset contained in "BODY_FAT.TXT".

In the following analyses, use the body fat percentage you re-computed in Homework1.
In addition, please remove the observation with erroneous measurement as well as the influential observation that you identified in Homework4.

Remember that the final aim of this analysis is to produce a satisfactory regression model for the percentage of body fat based on all or on a subsample of the available predictor variables. A regression model should be satisfactory in three aspects:

- variability explanation;
- diagnostics on the model assumptions;
- parsimony and interpretability.

A. (15 Points) Select a model for predicting the body fat percentage, using backward elimination with alpha-to-remove $\alpha_R = 0.15$ (on the p-value of the t-test for single terms). Begin with the full model that considers all 13 predictors in the dataset (12 numerical predictors and the categorical predictor Over45; remember you DON'T want to include Density as predictor, since you use it to compute the body fat percentage). Report which predictor you remove at each step, and the final model that you obtain.
[Hint: use function `update` *to update the model removing predictors.]*

B. (15 Points) Select a model for predicting the body fat percentage, using forward selection with alpha-to-enter $\alpha_R = 0.15$ (on the p-value of the t-test for single terms). Begin with the empty model that considers only the intercept and try adding each of the 13 predictors in the dataset. Report which predictor you add at each step, and the final model that you obtain. Is the model different from the one selected at point A?
[Hint: use function `update` *to update the model adding predictors.]*

C. (30 Points) Select the best model of each size according to the RSS, and compute the RSS for each size. Then use the RSS to compute adjusted R-squared ${R_{adj}}^2$, Mallow's C_p, AIC and BIC. Plot these values versus the size, comment on the model that each of them choose and how similar/different they are.
[Hint: use function `regsubset` *from the package* `leaps` *to select the best model of each size. Use the argument* `nvmax=13` *in order to compute the best model up to size 13. Use the function summary on its output to compute the* RSS*]*

D. (40 Points) Evaluate the model selected in points A, B and C using diagnostic plots and checking for multicollinearity. Refine the models as needed, and select a satisfactory

model for the body fat percentage. Interpret your final model, and produce a prediction interval for the average men (a men with average values for each predictor).
If you encounter more than one model that you believe to be satisfactory, and if you see an interest in terms of interpretation and prediction in presenting more than one final model, you are allowed and encouraged to do so (but do not present more than two final models).
[Hint: use function `vif` *from the package* `car` *to compute the* VIF*. Use function* `predict` *to produce prediction interval]*