In fitting a multiple linear regression model (MLR) the identification of the relevant explanatory variables to include in the model is required. The identification must be performed against the principle of a parsimonious model. A parsimonious model is a model that accomplishes the desired level of explanation or prediction with as few predictor variables as possible. In the course, thus far, we identified significant explanatory variables by testing the associated coefficients of the explanatory variables for significant contribution to the model, those coefficients significantly different from 0. If a large number of explanatory variables are available the identification task is complex and nontrivial. The goal is to determine the best subset of variables to be used in the model. This subset can be identified by evaluating the performance of all models obtained from all possible subsets of the explanatory variables. The number of possible subsets become very large as the number of explanatory variables increase, rendering the evaluations of all subset models unfeasible. Over the time different techniques and approaches were proposed to identify the most significant explanatory variables. These techniques include amongst others forward, backward and stepwise selection methods. These approaches are often used by statistics practitioners. More recent techniques are shrinkage techniques, also referred to as regulation or penalise regression methods. These techniques include methods such as the LASSO, Ridge penalty, Elastic net and others. The goal of this assignment is to explore the LASSO techniques for selecting relevant variables for a model.

The seminal paper by the author Robert Tibshirani on the LASSO was published in January 1996. This publication opened up an ongoing area of research in the fields of Statistics and Machine learning. Please consider this paper for this assignment, specifically the formulation from section 1 and the application of the method on prostate cancer data.

Regression Shrinkage and Selection via the Lasso: Robert Tibshirani. Journal of the Royal Statistical Society. Series B (Methodological), 1996, Vol. 58, No. 1 (1996), pp. 267-288 , https://www.jstor.org/stable/2346178

For the assignment, consider also other resources you might identify. You might also find the following of value:

SAS/STAT 15.3® User's Guide The GLMSELECT Procedure, SAS® Documentation, 31 January 2023 (Supplied)

https://video.sas.com/detail/video/3646879895001/lasso-selection-with-proc-glmselect

Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences: Daniel M. McNeish Multivariate Behavioral Research, 2015, 50:5, 471-484, https://doi.org/10.1080/00273171.2015.1036965

Assignment

1. Consider the LASSO method to fit a regression model.
   (a) Describe the "LASSO" method by defining the optimisation required to fit the regression model. $(2\ marks)$
   (b) The optimization in (a) can be obtained using a goal function including a regularization parameter. Describe this function and clearly identify the "OLS" term, the penalize term and the regularization parameter. $(4\ marks)$
2. It is known that OLS coefficient estimates are unbiased.
   (a) Comment if the estimates obtained using LASSO is unbiased or not. Motivate your answer. $(2\ marks)$
   (b) Using the bias-variance trade off, motivate why LASSO estimators are "better" estimators $(2\ marks)$
3. Consider the MLR model
$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_{20} x_{20} + \beta_{21} x_{21} + \beta_{40} x_{40} + \varepsilon. \tag{1}$$
   Generate, using SAS software, 500 data observations from this model with the variables $x_1,\ \ldots, x_{20}$ significant (relevant) explanatory variables and the variables $x_{21},\ \ldots, x_{40}$ non significant (no association with y) variables. Give a short description on the following:
   (a) How the variables $x_1,\ \ldots, x_{20}$ was generated. Give an algebraic formulation of the generation process followed. Attache the code as an appendix. $(2\ marks)$
   [Hint: Consider the data simulation programmes discussed during the course.]

(b) How the variables $x_{21}, \ldots, x_{40}$ was generated. Give an algebraic formulation of the generation process followed. Attache the code as an appendix. $(2\ marks)$

(c) How the variables $y_1, \ldots, y_{40}$ was generated. Give an algebraic formulation of the generation process followed. Attache the code as an appendix. $(4\ marks)$

4. Use, Proc GLMSELECT within SAS software package implementing the LASSO method and fit a regression model to the simulated data using all variables $x_1, \ldots x_{40}$. Attache the code used as part of the code appendix.

   (a) Present a graph indicating which variables are included in the model as a function of the regularization parameter. $(4\ marks)$

   (b) Comment on the effect of the regularization parameter in shrinking the coefficients. $(2\ marks)$

   (c) Give the final model selected, clearly indicating which significant and non significant variables are included in the final model. $(2\ marks)$

Instructions:

- You can work in groups of at most 5 to complete this assignment.
- Your assignment must be submitted on Gradescope (the link is on clickUP) before the due date. When you submit, you have to add/list group members to ensure that they are "linked" to the assignment. Only one person per group needs to submit the assignment.
- UP Rules and Regulations regarding this assignment are valid. If any unethical allegations (including but not limited to, adding a classmate's name to an assignment if they have not contributed suitably come to light, the matter will be referred to the UP Legal Office and all group members would be scrutinised and considered complicit.
- Submit your assignment in PDF format.
- Your submission must be free from spelling/grammar errors and be presented formally and professionally.
- Your submission may be at most **four** A4 pages (not including the appendix). Use a sans serif font with usual A4 margins with font size 11. **Start each question on a new page**. Do not include a cover page or a table of contents!

Submission deadline: 25 May 2025 at 23h50