Math312 **HW09** **Spring 2020**

Multiple Regression Model Building – 02 (with multicollinearity)

Note that

(1) This HW will be graded at the standard of regular exam.

*(2) You NEED to submit your work in a **nonmanual format** for this HW assignment (e.g. MS/WORD and convert into a PDF file).*

*(3) All the analyses are expected to be **done in SAS** and all the SAS **output** included **needs explanation**.*

*(4) **Graphs** (except for scatter matrix and diagnostics panel) can be at most **a-third-page high**, **NUMBERED** and **CAPTIONED.** Tables are in reasonable size. Remove all the unnecessary/unexplained tables or figures. Failure to do so will result in deduction of points. It helps to think about you will be charged by the number of pages if report gets accepted for publication.*

*(5) **Place your SAS code in the appendix**. (for your later reference)*

A study of investigating how peak rate of water flow *Q* (cfs) associates with attributes of watersheds from six storm episodes was conducted. The data consists of the following features is provided in HW09.sas

Feature	Type	Label (meaning)
Q	Num	Peak Rate of Flow
X_1	Num	Area of watershed
X_2	Num	Area impervious to water
X_3	Num	Average slope of watershed
X_4	Num	Longest stream flow in watershed
X_5	Num	Surface absorbency index, 0 = complete absorbency, 100 = no absorbency
X_6	Num	Estimated soil storage capacity
X_7	Num	Infiltration rate of water into soil (inches/hour)
X_8	Num	Rainfall
X_9	Num	Time period during which rainfall exceeded 0.25 inch/hr.

1. (10 pts) Exploratory Data Analysis:
   Generate a scatter matrix for all the numerical variables (10 × 10) with marginal histograms on the diagonal. Explain the information about possible associations among variables and their distributional behaviors.
2. (10 pts) Correlation Analysis on numerical features
   Evaluate all the pairwise correlations among all the numerical features. (assuming bivariate Normalities for all the pairs of features) Report the significantly correlated pairs, make your comments connected to what you discovered in 1. Project the result of regression model if to be implemented.
3. Regression Analysis
   a. (15 pts) Fit the FULL regression model for regressing Peak Rate of Flow (Q) on X_1- X_9.
      i. Testing on regression effect.
      ii. Perform model diagnostics. Report any serious violations.
      iii. (if no violation) Report the LSE fit for the full model and adjusted R^2.
   b. (15 pts) Since there is an evidence about heteroscedasticity, fit the FULL regression model but with nature-logarithm-transformed response (i.e. log(Q), variable in the data). Repeat tasks (i.-iii.) in a. Comment on how this transformation help with alleviating heteroscedasticity.
   c. (5 pts) Evaluate and report whether there is multicollinearity among regressors.
   d. (10 pts) Assume that you found multicollinearity in c, does centering the regressors help to resolve multicollinearity?
   e. (5 pts) Assume that multicollinearity remains after centering, use the uncentered data and generate ridge trace/plot. Confirm that the variables which can be possible removed are X_6 and X_9.
   f. (10 pts) Perform model selection via (1) GROUPLASSO and (2) Elasticnet. Report the models that proposed by those two methods.
   g. (5 pts) If model selection is not attempted, perform ridge regression with ridge parameter $\lambda = 0.4$. Report the estimated ridge regression model.

h. (5 pts) If model selection is not attempted, perform principal component regression with five principal components. Report the estimated principal component regression model.
i. (10 pts) Choose your final model and perform model diagnostics accordingly. Address on how you deal with multicollinearity towards the final model.

Please refer to the annotated MR example-02 for the understanding of the flow and SAS codes.

In terms of flow, regression modeling under issue of multicollinearity could be proceeded as follows:

(1) Fit raw data with full model, perform model diagnostics to see if there is(are) any serious violation(s). If no, proceed to step (2), Otherwise, try transformations to alleviate the violation before proceeding to (2).
(2) Detection of multicollinearity by VIF, condition index. If no concern on multicollinearity, proceed model selection as HW08 (no repetition here); otherwise proceed with (3).
(3) Try following possible attempts to alleviate multicollinearity.
    (3-a) Try centering regressors to rule out the possibility that collinearity happens between any regressors and 1's (intercept).
    (3-b) Try ridge trace (or by domain expertise) to see whether there is possible deletion of regressors whose coefficients are nearly zero.
    (3-c) Based on domain knowledge, combining regressors that are possibly involved in multicollinearity.
    If multicollinearity gets alleviated (substantially at least), one might choose to proceed with model selection as did in HW8, or use directly (too few variables left) as the final model.
(4) If multicollinearity remains and cannot be ignored (e.g. some of VIFs are still >> 10), proceed
    (4-a) GroupLASSO or ElasticNet, if model selection is still intended.
    (4-a) Ridge Rgression or Principal Component regression, if model selection is still intended.