**Math312 HW08 Spring 2020**

**Multiple Regression Model Building – 01 (without 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 at most be* ***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 hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 108 patients was available for analysis. From each patient record, the following information was extracted from the pre-operation evaluation:

|  |  |  |
| --- | --- | --- |
| Feature | Type | Label (meaning) |
| Y | Num | Survival Time |
| X1 | Num | Blood Clotting Score |
| X2 | Num | Prognostic Index |
| X3 | Num | Enzyme Function Test Score |
| X4 | Num | Liver Function Test Score |
| X5 | Num | Age (in years) |
| X6 | Char | Gender (0:Male; 1:Female) |
| X7 | Char | History Of Alcohol Use: 1: none; 2: moderate; 3: severe |

1. (10 pts) Exploratory Data Analysis:
   1. Generate a scatter matrix for all the numerical variables (6 × 6) with marginal histograms on the diagonal. Explain the information about possible associations among variables and their distributional behaviors.
   2. Generating boxplots for all the numerical variables, make them in a (2 × 3) display. Comment on their distributional information again.
   3. Generating visuals for all the categorical variables, e.g. frequency tables or pie charts. Fit the visuals in reasonable size.
   4. Generate side by side boxplots of survival time across different gender groups and different alcohol usage groups. Comment on whether you suspect significant mean difference across gender and alcohol use groups.
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.a. Project the result of regression model if to be implemented.

1. (5 pts) Frequency Analysis on categorical features

Evaluate the possible association between alcohol consumption and gender.

1. Regression Analysis
   1. (13 pts) Fit the FULL regression model for regressing survival time (Y) on Blood Clotting Score (X1), Prognostic Index (X2), Enzyme Function Test Score (X3), Liver Function Test Score (X4) and Age (X5).
      1. Testing on regression effect.
      2. Perform model diagnostics. Report any serious violations.
      3. (if no violation) Report the LSE fit for the full model and adjusted R2.
   2. (13 pts) Fit the FULL regression model but with nature-logarithm-transformed response (i.e. log(Y)). Repeat tasks (i.-iii.) in a. Comment on how this transformation help with the analysis.
   3. (20 pts) Perform model selection via (1) Adjusted R2, (2) AIC, (3) BIC, (4) SBC, (5) Cp, (6) Forward, (7) Backward, (8) Stepwise, (9) LASSO and (10) Elasticnet. Prepare a table that lists the top three models from (1)-(5) and best model for each of the (6)-(10).
   4. (13 pts) Select your final model(s) and conduct model diagnostics accordingly.
   5. (6 pts) Report and explain your final model(s). Provide suggestion for possible future improvement if you find the result is barely satisfactory.
2. (10 pts) Find the standardized coefficients for your final model and compare on the plausible impacts for regressors on the response (survival time in logarithm scale) by their coefficients. (note that this tasks is assigned for completeness, overpowering is risky and use with caution in reality)

Please refer to the annotated MR example-01 for the understanding of the flow and SAS codes.