STAT 6337 (Advanced Statistical Methods)

Project # 1 (Due on October 25, 2018)

- For starting seed value, use the last 4-5 digits of your ID.
- This is an individual project rather than a group project. You may discuss it with others but you must write your own code and answers. If the submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Your project report must be typed. It must consist of answers to the specific questions asked, appropriate justifications for the statistical methods used, annotated `SAS` code, and at most 6 pages of relevant parts of SAS output with relevant numbers highlighted. You may not get any credit if you just dump code and output. Be sure to justify the choice of any statistical method you use and mention and verify any inherent assumptions.
- You are welcome to ask me or TA questions. However, first try to find the answer on your own. Don't be afraid to google! It is a necessary skill for developing expertise in any programming language.
- Some useful links:
  - `https://stats.idre.ucla.edu/sas/library/`
  - `https://stats.idre.ucla.edu/sas/library/sas-libraryoverview-of-the-sas-language/`
  - `https://stats.idre.ucla.edu/sas/library/sas-libraryoverview-of-sas-procedures/`
  - `http://www.ssc.wisc.edu/sscc/pubs/4-8.htm`
  - `https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=allprodsproc&docsetTarget=procedures.htm&locale=en`

---

1. A prostate cancer dataset is available on the website. It consists of data on 97 men with advanced prostate cancer. Following is a description of the variables:

ID	1 to 97
PSA level	Serum prostate-specific antigen level (mg/ml)
Cancer Volume	Estimate of prostate cancer volume (cc)
Weight	prostate weight (gm)
Age	Age of patient (years)
Benign prostatic hyperplasia	Amount of benign prostatic hyperplasia (cm^2)
Seminal vesicle invasion	Presence (1) or absence (0) of seminal vesicle invasion
Capsular penetration	Degree of capsular penetration (cm)
Gleason score	Pathologically determined grade of disease (6, 7 or 8)

(a) (25 points) Make scatterplots of PSA level with other variables. Based on these, choose one numerical variable that you think may be used effectively to predict PSA level. Highlight any potential outliers on the scatterplot of this variable with PSA level. Fit a simple linear regression model and carry out regression diagnostics. The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If an assumption is not met, attempt to remedy the situation. Comment on the fit the final model using appropriate tests and statistics. Use the final model to provide interval estimates of the mean PSA level for three patients whose predictor variable values are at first, second, and third quartiles.

(b) (25 points) Choose one more variable to add to the above model and justify your choice. Fit a model with the two chosen variables, and repeat the above analysis.

2. "Uric Acid and Cardiovascular Risk Factors": The cardio dataset contains data on 998 individuals on the following variables (Source: Heritier et al., Robust Methods in Biostatistics, 2009):

#	Variable	Description
1	uric	Uric acid level
2	dia	Diastolic blood pressure
3	hdl	High-density lipoprotein cholesterol
4	choles	Total cholesterol
5	trig	Triglycerides level in body fat
6	alco	Alcohol intake (ml per day)

(a) (35 points) Make scatterplots of Uric acid level with other variables and calculate the corresponding correlation coefficients. Based on these, choose one numerical variable that you think may be used effectively to predict Uric acid. Highlight any potential outliers on the scatterplot of this variable with Uric acid. Fit a simple linear regression model and carry out regression diagnostics (including Brown Forsythe Test using SAS). The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If an assumption is not met, attempt to remedy the situation. Comment on the fit of the final model using appropriate tests and statistics. Use the final model to provide interval estimates of the mean uric acid level for patients whose predictor variable values are at first, second, and third quartiles. Construct a 95% confidence ellipse for (β_0, β_1) and plot it. Use the plot to test $H_0 : (\beta_0, \beta_1) = (100, 100)$ at 5% level of significance.

(b) (15 points) Remove five outliers (justify your choices) and redo the above analysis. Compare the two sets of results.