STAT 6337 (Advanced Statistical Methods)

Project # 3 (Due in class on Nov 29, 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. See additional instructions at the end.
- 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.

1. (25 points) Consider the prostate cancer dataset from the previous project. 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)

Use these data to build a regression model for predicting PSA level. Treat all the other variables (except ID, of course) as potential predictors. You may use automatic selection procedures discuss in the class to come up with a small subset of models for closer examination. Conduct relevant diagnostics on the selected models (including checking for collinearity and influential observations). If the diagnostics reveal any issue, attempt to remedy it. Finally, propose one model that appears the best in your view. Justify all your answers.

2. (25 points) Consider the `gpa` data stored in the `gpa.csv` file available on eLearning. The data consist of GPA at the end of freshman year (`gpa`) and ACT test score (`act`) for randomly selected 120 students from a new freshman class. Make a scatterplot of `gpa` against `act` and comment on the strength of linear relationship between the two variables. Let ρ denote the population correlation between `gpa` and `act`. Provide a point estimate of ρ, histogram of (nonparametric) bootstrap distribution of the point estimate, bootstrap estimates of bias and standard error of the point estimate, and 95% confidence intervals computed using normal approximation, basic bootstrap, and percentile bootstrap methods. Interpret the results. Be sure to compare the various confidence intervals.

Useful links:

- http://blogs.sas.com/content/iml/2016/08/10/bootstrap-confidence-interval-sas.html
- http://blogs.sas.com/content/iml/2014/01/29/sample-with-replacement-in-sas.html

---

Arrange the contents of the report in the following order (Keep the three components separate; points will be deduced if you mix them!)

1. complete and typed answers or comments about each step, e.g., ... plot/test indicate ..... assumption may be violated. So, ... is attempted, which shows ....
2. at most 10 pages of relevant parts of SAS output with relevant numbers highlighted and arranged in the order of the steps described in above typed comments and labeled.
3. your SAS code including brief typed comments of main steps. Make sure your code is well-annotated otherwise full credit will not be given.