Stat 481 Project   Due Time: 11:59 PM on Wednesday, April 2.

Instructions:

- Project must be typed in a word or pdf file for credit. Use COMPLETE SENTENCES and justify your findings.
- **Include only necessary figures and tables, and comment each of them. DO NOT include R or SAS code in the main section of the report. Attach your code in a separate Appendix section at the end of the project .**
- R Markdown or a set of code plus a sheet with brief answers will not be accepted for grading.
- No late projects will be accepted. Should you have any questions, only brief responses will be provided, and longer questions may not be answered.
- Gradescope Entry Code: 6KGPBG

Dataset Location:
Use the dataset provided on Blackboard: `Diabetes.txt` or `Diabetes.xlsx`

Diabetes Data:
In this data set, there are 10 baseline numerical predictor variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of 442 diabetes patients, as well as the response of interest, one target variable, a quantitative measure of disease progression one year after baseline. This dataset was originally used in "Least Angle Regression" by Efron et al. (2004) in Annals of Statistics. It was also posted on the Blackboard.

	AGE	SEX	BMI	BP	··· Serum Measurements ···						Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

Table 1. Diabetes study. 442 diabetes patients were measured on 10 baseline variables.

Project Goal:
Our goal is to perform a regression analysis to identify the predictor variables that significantly contribute to understanding the variations in diabetes progression.

The Diabetes data set contains the following variables:

Variable Name	Description
AGE	age of the patient
SEX	gender of the patient
BMI	Body mass index
BP	Average blood pressure
S1	Serum total cholesterol level
S2	Low-density lipoproteins (LDL)
S3	High-density lipoproteins (HDL)
S4	Total cholesterol / HDL ratio
S5	Serum triglyceride level
S6	Blood sugar level
Target	Disease progression indicator (response)

Key Items of Statistical Report:

- Title Page. Include project title, your name, and a brief summary of the project.
- Part 1. Data Introduction and Description. Provide descriptive statistics such as five-number summary, mean, variance/sd, histogram, or boxplots.
- Part 2. Multiple Linear Regression Analysis.
  - Provide an introduction on each model/test/method you use in the regression analysis. Write down appropriate statistical models at each step of your analysis. State your problems with null and alternative hypotheses when necessary, based its ANOVA table and summary statistics, conclude accordingly.
  - Check for pairwise correlation and multicollinearity (VIF values) of all variables. This test only needs to be done once at the beginning of the analysis. If consider removing one variables because of severe multicolinearity issue, delete only one variable.
  - Model Diagnosis: linearity, independence, normality and equal variance of residuals
    - Provide any applicable plots (in a reasonable size) and appropriate tests (Shapiro-Wilks test, Breusch–Pagan test, etc.), and interpret the results.
    - If any of the model assumptions are violated, suggest ways to "fix" your data, for example the Box-Cox transformation. To simplify things, if you need to do a transformation, do *one* transformation for response, and proceed with analysis.
    - Be sure to re-check all the model assumptions after any transformations and address each of the assumptions in your report.
- Part 3. Variable Selection. Based on the model in the end of Part 2, build the "best" model possible by using either backward selection with the criteria for inclusion as having a significance of 0.10 or lower. Compare them with reasonable tests (ANOVA tables) or model criterion. Make sure to check all the model assumptions on the final model (L.I.N.E.).

- Part 4. Conclusion. Draw conclusions/interpret your regression model. Include a statement about R^2s before and after creating the "best" model possible. Include statements about each variable kept in the final model. *Make sure these conclusions can be understood by a general reader.*
- Appendix. **Attach R or SAS code in this section only.**

Grading:
Your project will be graded on the following items:

- Include a detailed description about data and Data Summary with graphical display
- Introduction of Regression Model, necessary hypotheses and analysis outcome interpretation, and model assumptions Check
- Drawing Conclusions based on the "best" model possible. Includes final model statement, interpretation of parameter values, how R^2 changes before and after creating the "best" model.
- Code Provided in Appendix.

Some useful `R` functions :

- read.table() or read.csv() to import data
- lm() and plot(lm()) to fit linear regression model
- cor() for correlation coefficient, shapiro.test(x) for normality check
- boxcox() in MASS package
- vif() and ncvTest() in car package
- Chapter 11 (R Introduction
  `https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf`

Some useful `SAS` procedures :

- DATA or PROC IMPORT
- PROC REG
  SAS PROC REG - model options: `https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_reg_sect013.htm`
- PROC CORR
- PROC UNIVARIATE
- PROC TRANSREG
- PROC MODEL
- SAS Procedure Help
  `https://documentation.sas.com/?docsetId=proc&docsetTarget=titlepage.htm&docsetVersion=9.4&locale=en`