MIDTERM STAT-COMP 2021

Enter your name here:

This exam is open book/notes but strictly individual.

Please refer any questions that yo may have to the instructor.

How to report your results and scripts

- I have uploaded in D2L the exam in three formats: Rmd, fillable pdf, and word.
- **If you choose to work with RStudio**, you can provide your exam in one file (compiled to either pdf or html) that displays your script, results, and answers. You can use the Rmd as as a template, simply add your code and answers and compile (be sure you use echo=TRUE, eval=TRUE).
- **If you choose not to use RStudio**, you can use either the word or pdf exam files to enter your output, answers, and scripts (word will be preferred because you can easily change the size of the boxes used for reporting).
- **For those using the word or pdf files**, report in the output box just the requested output (e.g., an estimate, SEs, etc.), copy on the script box your code.

Submission

- You should submit your exam file in D2L by 4:20pm.
- Exams uploaded after 4:25 will recieve a 5 point penalty.
- The submission folders won't be avilable after 4:30pm.

Questions

The exam has two questions (50 points each), each question has several sub-items. There is a third bouns question that can give you up to 15 extra point. Your final score will be the minimum of the sum of the points you obtain in questions 1 through 3 and 100.

Data

Throughout the exam you will use the following data set (I'll upload the data set in Github, you should change the link and read it from there)

```
fname='~/Dropbox/STAT_COMP/2021/PROSTATE_CANCER.csv'
DATA=read.csv(fname,header=TRUE)
dim(DATA)
```

```
## [1] 97  2
```

```
head(DATA)
```

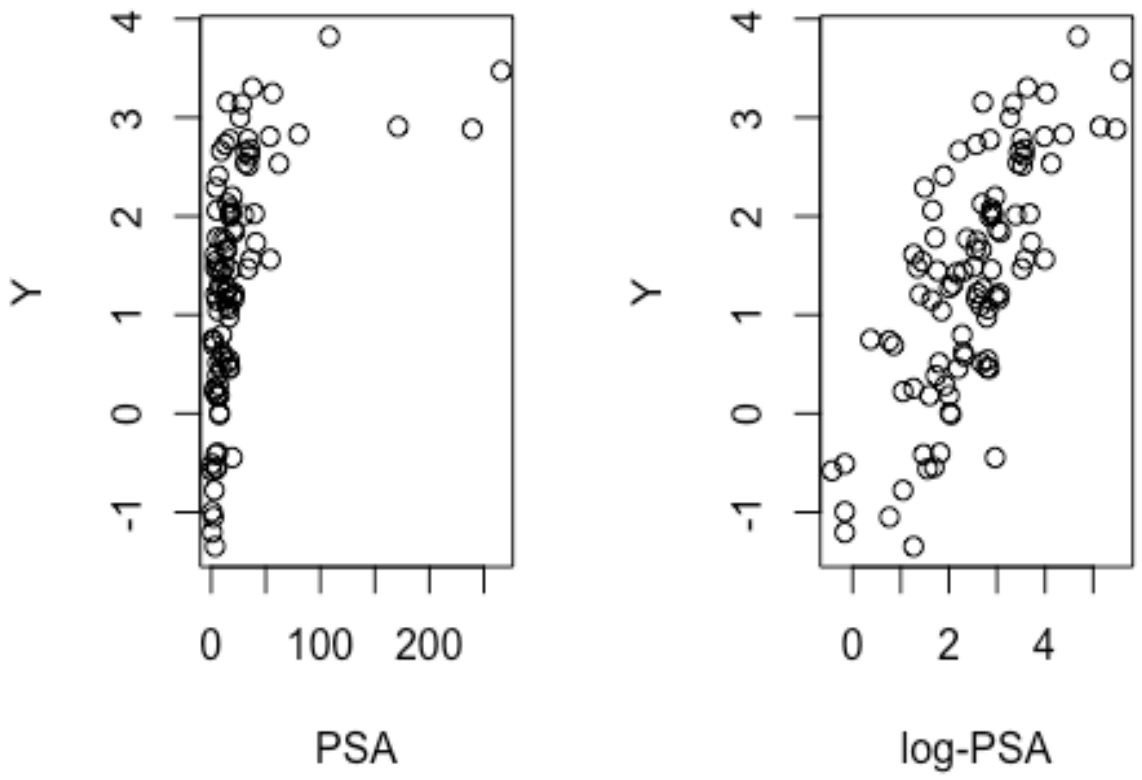
```
##            Y  PSA
## 1 -0.5798185 0.65
## 2 -0.9942523 0.85
## 3 -0.5108256 0.85
```

The data set contains information on 97 prostate cancer patients.

- Y: is the logarithm of the volume of the primary tumor
- PSA: Is the prostate specific antigen, a marker for prostate cancer.

The goal is to study how PSA co-varies with the logarithm of the volume of the cancer tumor.

Here are two plots with the response in the vertical axis and PSA (left) and the log-PSA (right) in the horizontal axis.



Question 1 (50 points)

1.1) Fit each of the following models using Y as the response

(i) A linear model with PSA as predictor
(ii) A cubic spline for PSA with 5 degree of freedom
(iii) A linear model log(PSA) as predictor

You can to use `lm()` and the functions of the spline library to answer this question.

Copy your output here

Copy your script here

1.2) For each of the models report: adjusted R-squared, AIC, and BIC.

Hint: For adjusted R-squared, you can use sumary(fm)$adj.r

Copy your output here

Copy your script here

1.3) What model do you recommend and why?

Provide your answer/explanation here

Question 2 (50 points)

2.1) Write a function that for a model of the form y=Xb+e evaluates the $RSS = (y - Xb)'(y - Xb)$. Your function should take as inputs y, X, and b, and return the residual sum of squares.

Copy your code here

2.2) Use the function you developed in 2.1 to fit the third model of Question 1 (the one using log(PSA) as the predictor) via ordinary least squares using `optim()`.

Hints:

- Do not worry about centering log-PSA in your incidence matrix (I tested and it converges wihtout centering)
- Initialize the intercept to the mean of Y and 0 for the coefficient on log-PSA.

Report the estimates you obtained here

Copy your script here

2.3) Compute the SE, z-statistic, and p-values using the results provided by optim(), report a table like the one produced by `summary(fm)` derived completly from the results returned by optim().

Notes:

- To derive pvalues, assume that estimates follow normal distributions, and use the standard approch we used in maximum likelihood estimation to approximate the SEs and pvalues.
- You should expect differences in the 1st or 2nd decimal place for estimates and SEs, and functions thereof.

Copy your output here

Copy your script here

Question 3 (up to 15 bouns points)

Use Bootstrap to approximate the SEs for the coefficients of model *iii* of Question 1.1 (the one using log(PSA) as the predictor). Report below your results, and your scripts

Report your Boostrap estimate of the SE here

Copy your script here