STAT 341: Assignment 1 - Spring 2020

Name

72 Marks, Due: Friday, June 26 at 10:00am

NOTES

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark/LEARN. This means that your responses for different questions should be in separate .pdf files. Your .pdf solution files must have been generated by R Markdown unless otherwise specified. Additionally:

- For mathematical questions: your solutions must be produced by LaTeX (from within R Markdown). Handwritten and scanned/photographed solutions will not be accepted and you will receive zero points.
- For computational questions: R code should always be included in your solution (via code chunks in R Markdown). If code is required and you provide none, you will receivee zero points.
- For interpretation question: plain text (within R Markdown) is fine.

Organization and comprehensibility is part of a full solution. Consequently, points will be deducted for solutions that are not organized and incomprehensible.

- You will submit your solutions in the form of one pdf file per question through LEARN For example, for Q1 you should submit one pdf file containing the solution to the first question only. Failing to follow the formatting instructions may result in your whole paper or individual questions receiving a grade of 0%.

Question 1 - 36 Marks

- The relationship between age and survival for female passengers on Titanic

For this question you will need the Titanic data which can be found on in the `carData` package. Here we will focus on the female passengers and the relationship between age and survival.

```
library(carData)
data(TitanicSurvival)
Titanic = na.omit(TitanicSurvival)
Titanic = Titanic[Titanic$sex == "female", ]
```

```r
Titanic$survived1 = as.numeric(Titanic$survived == "yes")
head(Titanic)
```

```
##                                 survived    sex age passengerClass survived1
## Allen, Miss. Elisabeth Walton        yes female  29            1st         1
## Allison, Miss. Helen Loraine          no female   2            1st         0
## Allison, Mrs. Hudson J C (Bessi       no female  25            1st         0
## Andrews, Miss. Kornelia Theodos      yes female  63            1st         1
## Appleton, Mrs. Edward Dale (Cha      yes female  53            1st         1
## Astor, Mrs. John Jacob (Madelei      yes female  18            1st         1
```

- Remember that each plot should be clearly labelled.

a) **[3 Marks]** Form two histograms (side-by-side) of ages by
   - using **equal bin widths** with bin width equal to 8 years and
   - using **varying bin widths** with 10 bins (use zero as the smallest value).

b) **[5 Marks]** We can model the relationship between survival and age non-parametrically by calculating the proportion of survivors for a given age range. Construct two plots (side-by-side) of age (`x`-axis) versus survival (`y`-axis).
   - Then using the two age partitions (**equal bin widths** and **varying bin widths**) from part a), add points using the mid-point of the age interval and the proportion of survivors within each age group.
   - In addition, generate a table where each age partition has the following columns: age range, total number of units in that range, number of survivors, and the proportion of survivors. It might be helpful to write a function that takes in the age groups and then creates a plot and outputs the table.
   - Why is the age partition with **varying bin widths** preferred over the range partition with **equal bin widths**?

c) Another approach is to model the proportion using a parametric model. i.e. Use a function to model the proportions that is bounded to the range $[0, 1]$. One example is the logistic function but here we will consider using the **probit function** which is the cumulative distribution function of the standard normal distribution denoted by Φ

$$\Phi(z) = \int_{-\infty}^{z} \phi(u)du = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

i) **[2 Marks]** Using the R-function `pnorm` and plot the probit function over the range $[-6, 6]$

ii) **[2 Marks]** Recreate the plot from b) using **varying bin widths** and overlay the probit function with the argument given by

$$\Phi\left(\widehat{y}\right) \qquad \text{where} \qquad \widehat{y} = -1/2 + 0.03 \times \text{Age}$$

d) **[4 Marks]** Differentiate the following log-likelihood of N bernoulli trials with varying probability of success with respect to α and β.

$$l(\boldsymbol{\theta}) = l(\alpha, \beta) = \sum_{i=1}^{N} \left[y_i \log \frac{p_i}{1 - p_i} + \log(1 - p_i) \right]$$

where

$$p_i = \Phi(\widehat{y}) = \Phi(\alpha + \beta \left[x_i - \overline{x} \right])$$

In particular, show that

$$\frac{\partial l}{\partial(\alpha,\beta)} = \sum_{i=1}^{N} \frac{y_i - p_i}{p_i(1-p_i)} \times \phi(\widehat{y}) \times \begin{bmatrix} 1 \\ x_i - \overline{x} \end{bmatrix}$$

where $\phi(z)$ is the probabilty density function for the standard normal density. i.e.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{z^2/2}$$

Hint: Use the chain rule.

e) Fit the model given by the equation from part d) via gradient descent. **Note**, under maximum likelihood we would maximize the function given in d), so we want to minimize the negative of the log-likelihood function.

- i) **[2 Marks]** Modify the `createLeastSquaresRho` function to calculate the objective or negative of the log-likelihood function, $-l(\boldsymbol{\theta})$. Call this new function `createObjBinary`.
- ii) **[2 Marks]** Modify the `createLeastSquaresGradient` function to calculate the gradient of the objective function or the gradient of the negative of the log-likelihood. Call this new function `createBinaryLogisticGradient`
- iii) **[4 Marks]** Using the functions `gradientDescent`, `gridLineSearch`, and `testConvergence` from notes and the functions you created in part e), perform gradient descent until convergence with `theta=c(0,0)` and `lambdaStepsize = 0.0001, lambdaMax = 0.01`
- iv) **[4 Marks]** What are the values of α and β that correspond to the age having no effect on survival? Use these parameter values as a starting value for gradient descent. Is there any improvement?

f) Now, we will assess the fitted probit curve.

- i) **[2 Marks]** Recreate the plot from b) with **varying bin widths** and overlay the fitted probit function.
- ii) **[2 Marks]** Compare the fitted values from the probit curve to the points added to the plot.
- iii) **[2 Marks]** What is implicit assumption behind the parametric model and non-parametric model?
- iv) **[2 Marks]** At what age would the probit curve report as a 50-50 chance of survival.

Question 2 - 10 Marks

Suppose were going to have a test that covers Section 2.1, 2.2 and 2.3. Construct a one page **handwritten** study sheet. The page should be one sided and the allowable size is 8.5×11 inches.

- **Note:** Output from word, markdown or etc is not allowed.

Rubric

Criteria	Descriptor	Marks
Content	Coverage, Key Concepts, Examples, Terminology	/7
Format	Creativity, Clarity and Organization	/3

Question 3 - 10 Marks

In your own words summarize the subsections **3.0-Samples**, **3.1-All_Possible_Samples** and **3.1.1-Consistency_and_Sample_Size**.

- You are limited to 1 to 2 pages.
- You are recommended to use a combination of formulas, full sentences an example.

Rubric

Criteria	Descriptor	Marks
Format	Organization	/3
Writing	Clarity & Grammar	/2
Content	Coverage, Depth, Relevant Terminology used and Example	/5

Question 4 - 16 Marks

Demonstrate the gradient descent using some non-trivial objective function.

- The objective function;
  - should be a function of more than one variable that cannot be decomposed into a sum of two single-variable functions (i.e. $f(x, y) \neq g(x) + h(y)$),
  - it can be either an implicitly defined attribute of interest such as robust regression or an explicit multivariate function such as the Himmelblau's function.
- Your demonstration should
  - clearly provide the objective function, ρ, and gradient, g, in LaTeX and R code,
  - include a variety of line search methods,
  - different starting values,
  - some sort of graphic to help understand problem or the solution, and
  - a summary of what was demonstrated and a conclusion.
- You answer should be limited to 1 to 3 pages.
  - Note that any functions used in the notes or function glossary can loaded using `echo=FALSE` but any other code chunks should have `echo=TRUE`. e.g. the code chuck loading `gradientDescent` can use `echo=FALSE` but chunks that call `gradientDescent` should have `echo=TRUE`.

Rubric

Criteria	Descriptor	Marks
Function/Gradient	Difficulty, Description, Written in LaTeX & R	/4
Format	Clarity and Organization	/4
Results	different line search methods and starting values, Graphic	/4
Summary	Justification and Relevant Terminology used	/4