STAT/Q SCI 403: Introduction to Resampling Method
Spring 2018
Homework 05

Instructions:

- You have to submit all your answers in a single PDF file generated by either LaTeX or *Rmarkdown*.
- You may use the LaTeXtemplate `HW_template.tex` to submit your answer.
- For questions using R, you have to attach your code in the PDF file. If the question ask you to plot something, you need to attach the plot in the PDF as well.
- If the question asks you to show a figure, the clarity of the figure will also be graded.
- The total score of this homework is 8 points.
- Questions with ♠ will be difficult questions.

Questions:

1. In this question, we will study the statistical consistency of logistic regression. For simplicity, we consider a univariate covariate X versus a univariate response Y such that $Y = \{0, 1\}$. Recall that the logistic regression models

$$P(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

   We use Monte Carlo Simulation to study the distribution of the fitted parameters. We assume that $\beta_0 = 1, \beta_1 = 2$.

   (a) **(1 pt)** First generate $n = 500$ data points such that $X_1, \cdots, X_n \sim N(0, 1)$ and $Y_1, \cdots, Y_n$ from

$$P(Y_i = 1|X_i) = \frac{e^{1+2\cdot X_i}}{1 + e^{1+2\cdot X_i}}.$$

   Namely, given X_i, the value of Y_i is from a Bernoulli distribution with parameter being $\frac{e^{1+2\cdot X_i}}{1+e^{1+2\cdot X_i}}$. Fit the logistic regression, what are the fitted parameter $\widehat{\beta}_0$ and $\widehat{\beta}_1$?

   (b) **(1 pt)** Using Monte Carlo simulations to repeat the above procedure $N = 2,000$ times. This gives us $2,000$ realizations of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Use two histograms to show the distribution of them. Moreover, attach a vertical line to each histogram to denote the true value of β_0 and β_1.

(c) **(1 pt)** Judging from the previous two histograms, do $\widehat{\beta}_0$ and $\widehat{\beta}_1$ follow roughly a Normal distribution? Why or why not?

(d) **(1 pt)** Now increase the sample size to $n = 2000$, repeat the procedure in question (b) and plot the histograms. Does the histogram concentrate more around the true parameter values?

(e) **(1 pt; ♠)** Now we focus on the slope parameter β_1. Use Monte Carlo simulations and a plot to illustrate the convergence of $\widehat{\beta}_1$ toward β_1.
Hint 1: (Method 1) Use Monte Carlo simulations to compute the variance $\mathsf{Var}(\widehat{\beta}_1)$ or the MSE at different sample sizes. And then use a line graph to show that the variance/MSE converges when sample size increases.
Hint 2: (Method 2) Use Monte Carlo simulations to obtain distributions of $\widehat{\beta}_1$ at different sample sizes and then (i) overlap histograms, or (ii) use boxplots to show the concentration of distributions.
You may earn an extra point if you do not use the methods mentioned in the above two hints.

2. In the Logistic Regression, we often encounter a function like

$$f(x) = \frac{e^x}{1 + e^x}.$$

An interesting fact is that: this function $f(x)$ is actually a CDF.

(a) **(1 pt)** If a random variable X has CDF $F(x) = f(x) = \frac{e^x}{1+e^x}$, what is its PDF $p(x)$? What are the mean and median of this random variable?

(b) **(1 pt; ♠♠)** Write down a procedure to generate X. Namely, find a procedure to generate random points from the CDF $F(x) = \frac{e^x}{1+e^x}$.

(c) **(1 pt; ♠♠)** Use Monte Carlo Simulation to generate at least $n = 10,000$ points from the CDF $F(x) = \frac{e^x}{1+e^x}$. Compare the histogram to the density curve $p(x)$ of the CDF $F(x)$.

Hint: The Cauchy $(0, 1)$ distribution has a PDF

$$p_{\mathsf{Cauchy}}(x) = \frac{1}{\pi}\frac{1}{1 + x^2}$$

and the double exponential distribution has a PDF

$$p_{\mathsf{dExp},\lambda}(x) = \frac{\lambda}{2}e^{-\lambda|x|},$$

where λ is the rate parameter similar to the usual exponential distribution. These two distributions are supported on $(-\infty, \infty)$. You can use `rcauchy()` to generate points from the Cauchy distribution; you need to come up with your own way to generate points from the double exponential distribution if you want to use it.

Bonus points: You will earn a bonus point if you use a double exponential distribution to solve Q2-(c).