Spring 2020 Quantitative Analysis
Chung-Ming Kuan

# Midterm: Computer Exam

This is an one hour, computer-based exam. Any material stored in your computer can be used in the exam, while **access to the Internet is strictly prohibited**. Please write your codes on the R Answer Sheet provided by TA, and upload it on NTU COOL Assignments at the end of the exam.

**Needed Package**: `ISLR`

1. (**10 points**) Please load the data set "`mtcars`" from `R`. The data was extracted from the 1974 Motor Trend US magazine, and comprises `mpg` (fuel consumption) and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). Suppose we have the following model:

$$\texttt{mpg}_i = \beta_0 + \beta_1 \texttt{wt}_i + \beta_2 \texttt{hp}_i + \beta_3 \texttt{drat}_i + \beta_4 \texttt{gear}_i + u_i.$$

   (a) Please fit the data to the above model and obtain $\hat{\beta}_1$, the OLS estimator of $\beta_1$.

   (b) Use $B = 1000$ bootstrap samples to compute the "Paired Bootstrap" and "Residual Bootstrap" estimator of $\text{SD}(\hat{\beta}_1)$ without the function **boot()**. And please **set.seed(b)**, $b = 1, ..., 1000$, for each time you do bootstrap.

   (c) Please test the hypothesis $H_0 : \beta_1 + \beta_2 = -3$ and obtain its p-value.

   (d) Please test the hypothesis $H_0 : \beta_1 = \beta_3 = \beta_4 = 0$ that is robust to heteroskedasticity and obtain its p-value.

2. (**10 points**) Please load the data set "`Wage`" from `R`. This data set contains `wage` (workers raw wage) and 11 variables for a group of 3000 male workers in the Mid-Atlantic region.

   (a) How many white men in our dataset are older than or equal to 50 years old with wage smaller than 100?

   (b) Please construct a variable "`Divorced`", which equals 1 when the worker is divorced and 0 otherwise. Then show the divorced rate (`sum(Divorced)/length(Divorced)`).

   (c) Please fit a probit model with `logwage` and `age` as the independent variables and `Divorced` as the dependent variable. What is the estimated coefficient for `logwage`?

   (d) Following (c), instead of fitting a probit model, please fit a logit model. Then, please use `10-fold CV` and `set.seed(1)` to determine which model is better with showing the testing MSE.

3. (**10 points**) Monte Carlo Simulation :

- Sample sizes $N$: 5, 500
- Number of replications: 2000
- Data generating process (DGP): $y_i \sim U(-1, 1)$
- The statistics:

$$M_N = \frac{1}{\hat{\sigma}_N \sqrt{N}} \sum_{i=1}^{N} \phi(y_i), \quad \text{where } \hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^{N} \left( \phi(y_i) - \frac{1}{N} \sum_{i=1}^{N} \phi(y_i) \right)^2,$$

- Moment function: $\phi(y_i) = y_i^3$
- For the total 2 different ways of constructing $M_N$ ($N = 5, 500$), please plot their corresponding histogram using 2000 replications. Note that each graph must be properly labeled with suitable title, then combine these 2 graphs on a single $1 \times 2$ plot.
- Please compute the empirical frequencies of the events: $|M_N| > 1.644854$. Record them under their corresponding graphs.
- Please add the Gaussian kernel density estimate (KDE) of $M_N$ (using blue line) as well as the probability density function (PDF) of $N(0, 1)$ (using red line) for each simulation graphs.