OLLSCOIL NA hEIREANN, CORCAIGH
THE NATIONAL UNIVERSITY OF IRELAND, CORK

COLAISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

ST4060 - ST6015 - ST6040

Continuous assessment 1 - 2021-22

Eric Wolsztynski
eric.w@ucc.ie

Question 1

No code is required for this question.

Consider an i.i.d. sample $\{X_1, \dots, X_N\}$ and a non-parametric estimate $\hat{f}$ of its probability density function f defined for any $u \in \mathbb{R}$ and some real constant $h > 0$ by

$$\hat{f}(u) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{X_i - u}{h}\right)$$

(a) What is the standard deviation of function $K(u)$? (No derivations are required to answer this question.)

(b) What is the standard deviation of function $K_h(u) = K(u/h)/h$? (No derivations are required to answer this question.)

(c) Can $K(u) = \exp(-\frac{u^2}{2})$ be used to compute this estimate? Why?

(d) In order to ensure a finite-sample estimate $\hat{f}$ of f with as small a bias as possible, and using the unbiased sample variance estimate $\hat{\sigma}^2$ of $Var(X)$, indicate which of the following values of h should be used and why:

$$h_1 = 1.06\ \hat{\sigma}\ N^{-\frac{1}{5}}$$
$$h_2 = 2.34\ \hat{\sigma}\ N^{-\frac{1}{5}}$$

Note: no marks awarded if no explanation is provided; a one-sentence explanation is all that is required here.

Question 2

Run `set.seed(4060)` before running the analysis below, and any time you run your code.

Implement a Monte Carlo simulation of $M = 1,000$ random samples, each comprising of $N = 100$ observations $\{Y_i\}_{i=1}^N$ defined as

$$Y_i = \theta^* X_i + \varepsilon_i$$

where:

- $\theta^* = 8$,
- $\{X_i\}_{i=1}^N$ is a unique sample of N realisations of $X \sim \mathcal{U}(1, 2)$ that you generate only once and use in all Monte Carlo iterations (i.e. where X is a unique sample of realisations of a continuous random variable with a Uniform distribution over $[1, 2]$, common to all Monte Carlo samples),
- $\varepsilon \sim t(2)$ are i.i.d. realisations of Student's t-distribution with 2 degrees of freedom.

Compute and store the M estimates for the following three estimators of θ^*:

- the ordinary least squares estimator $\hat{\theta}^{LS}$;
- the estimator defined by $\hat{\theta}^{med} = median(Y/X)$, where *median* denotes the sample median;
- the estimator defined by $\hat{\theta}^{mean} = mean(Y/X)$, where *mean* denotes the sample mean;

(a) Quote the Monte Carlo estimate of the expected value of each of these three estimators.

(b) Quote the Monte Carlo estimate of the standard error of each of these three estimators.

(c) Comment on your results in (a) and (b). Which estimators would you prefer, and why?

(d) Provide a single figure showing boxplots for the Monte Carlo distributions of these three samples of estimates.

Question 3

Create the following dataset in your `R` session:

```
dat = data.frame(wt=mtcars$wt, mpg=mtcars$mpg)
```

Run `set.seed(6040)` before running the analysis below, and any time you run your code.

Implement bootstrapping of the linear regression of dependent variable `mpg` (car consumption, in miles per gallon) onto predictor (i.e. independent variable) `wt` (car weight, in 1,000 lbs). Store the bootstrapped estimates of intercept and slope coefficients, as well as the p-values corresponding to the bootstrapped slope estimates.

(a) Quote the bootstrap estimate of the expected value of the slope coefficient (i.e. of the effect of weight `wt` on consumption `mpg`).

(b) Quote a *naive*, bootstrap 95% confidence interval for the p-value associated with this effect.

(c) Quote an appropriate bootstrap 95% confidence interval for the slope parameter, i.e. the effect of car weight on car consumption.

(d) Compare the bootstrap confidence interval in (c) with the traditional 95% confidence interval for the effect of weight on consumption obtained under the Normal assumption from linear regression of the original data (i.e. without bootstrapping). Coment as you see appropriate.