Project 2
Computational Numerical Statistics
DM, FCT-UNL

Deadline: 13/12/21

The writing of the project report, either in LaTeX or R Markdown, should detail all the theoretical aspects of the methods involved in this project and provide all the necessary graphical displays essential for the understanding of the particular cases.

Bootstrap & Jackknife

In the resolution of the following problems **do not** use any of the `R` bootstrap and jackknife built-in functions.

Fix your R random seed to 1234 in all simulations.

1. Let $X \sim F$ such that $E(X) = \mu$ with μ unknown. Further, let $X_1, \ldots, X_n \underset{iid}{\sim} X$ and $T = g(X_1, \ldots, X_n)$ be an estimator of μ. Show that when $T = \overline{X}$ then: (i) the jackknife estimator of μ, say T_{jack}, coincides with T; and (ii) $V(T_{jack}) = \dfrac{n-1}{n} \sum_{i=1}^{n} (T_i^* - T_{jack})^2$ simplifies to $\dfrac{S^2}{n} = V(\overline{X})$.

2. Consider the observed sample referring to the survival times of some electrical component pertaining to a car assembly factory.

1766	884	2420	695	1825	1014	2183	2586	627	965
1577	2195	1354	1325	1552	1299	71	3725	1354	159

   (a) Let μ refer to the mean survival time of that component. Use the **non-parametric** bootstrap ($B = 10000$ samples) to test the hypotheses

$$H_0: \ \mu \leq 1020 \qquad \text{vs} \qquad H_1: \mu > 1020,$$

   at the 10% significance level. Would it be possible to perform this test using an exact test? If so, do it and compare the results.

   (b) Compute the 90% bootstrap **pivotal** and **percentile** confidence intervals for μ. Plot the histogram of the B bootstrap estimates of μ. Which CI do you think is more adequate?

   (c) Research the literature for the bootstrap **bias corrected and accelerated** (BCa) confidence interval Thoroughly present and discuss the BCa CI. Compute the 90% bootstrap BCa CI for μ.

   This car assembly factory needs that all these components are replaced after 1100 hours of service. Let

$$T = \text{ number of survival hours of a component.}$$

One is interested in estimating the proportion of components that live more than 1100 hours, i.e., one wishes to estimate $p = P(T > 1100)$. It is known that

$X =$ number of components that live more than 1100 hours in n inspected components, where $p = P(T > 1100)$ is the probability of a success, has distribution $\sim Bin(n, p)$

(d) Show that $\mathcal{P} = \dfrac{X}{n}$ is an unbiased and consistent estimator of p. Estimate p and $\mathtt{SE}(\mathcal{P})$.

(e) Describe and discuss in detail the **non-parametric bootstrap** and **jackknife** techniques. Use both approaches ($B = 10000$ samples in the case of the bootstrap) to estimate the variance, standard error and bias of $\mathcal{P}$. Compare the results. Check whether there is need to correct the original estimate of p for bias and if such report the corrected estimate of p.

(f) The **jackknife-after-bootstrap** technique provides a way of measuring the uncertainty associated with the bootstrap estimate $\widehat{\mathtt{SE}}(T)$ (T some estimator of interest). Research the literature for this technique and apply it in order to estimate the standard error of the bootstrap estimate of $\mathtt{SE}(P)$ obtained in (d). (if you are unable to program the method use some `R` built-in function to complete the exercise)

3. Consider the following data

```
x   34.00 28.00 31.00 28.00 30.0 27.0 32.0  25.0  34.0 34.00  29.0 26.00 24 33.00
y   23.44  7.95 17.04  9.57 16.9  9.3 16.2   3.2    24 19.02  11.2  7.32  3 18.63
```

(a) Graphically inspect that there is a linear trend in the data. Comment on the linear trend. Fit a linear regression model to your data in `R` presenting and commenting in detail all the summary results referring to the fitted model (returned by the `R` function `summary()`). In addition,

- plot the data *versus* the fitted line; and
- use the `R` built-in function `confint()` and report a 90% CI for the slope parameter.

(b) Carefully check for the linear model's underlying assumptions - use both visual inspection and adequate statistical tests (at the 5% level) to check the assumptions. Does the fitted model validate all the underlying assumptions?

(c) Use the **bootstrap of the pairs** (with $B = 10000$) to

- estimate the bias and standard error of the slope parameter estimator; check if there's need to correct the original estimate and if so report the corrected estimate;
- construct a **pivotal** 90% CI for the slope parameter; compare it with the CI obtained in (a).

(d) The **wild bootstrap** (there are several variants) is a bootstrap technique that has been shown to be more effective in the case of error heteroskedasticity than bootstrapping the pairs. Provided a detailed discussion of this method. Redo (c) using the wild bootstrap (with a variant of your choice). (if you are unable to program the method use the `R` built-in function `wild.boot()` to complete the exercise)

Optimization

4. Let $X \sim Pareto(1, \alpha)$, which has p.d.f.

$$f(x;\alpha) = \frac{\alpha}{x^{\alpha+1}}, \qquad \alpha > 0, \qquad x \geq 1.$$

Let

```
1.977866 1.836622 1.097168 1.232889 1.229526 2.438342 1.551389 1.300618 1.068584
1.183466 2.179033 1.535904 1.323500 1.458713 1.013755 3.602314 1.087067 1.014013
1.613929 2.792161 1.197081 1.021430 1.111531 1.131036 1.064926
```

be an observed sample from X.

(a) Derive the likelihood, log-likelihood and score functions (simplify the expressions as much as possible). Derive both the **maximum likelihood estimator** (MLE) and **method of moments estimator** (MME) of α and use them to estimate α. Why are ML estimators so attractive?

(b) Noting that the Pareto distribution belongs to the exponential family, derive the **Fisher information** $I_n(\alpha)$. Use the Fisher information to estimate the variance of the MLE.

Assume herein that it was not possible to derive the MLE of α.

(c) Display graphically (side-by-side) the likelihood, log-likelihood and score functions in order to locate the ML estimate of α. Indicate an interval that contains the ML estimate.

(d) Use the `R` function `maxLik()` from library `maxLik` to approximate the ML estimate of α. Feed `maxLik()` with the initial estimate of α given by the method of moments.

(e) Describe and discuss in detail the algorithms of **bissection**, **Newton-Raphson** and **secant** that enable, in particular, the approximation of the ML estimate of α. Implement those in `R` and use them and the sample above to estimate α - report all the iterations together with the error. Justify your choice of the initial estimates for each method and discuss the results.

Use the absolute convergence criterion as a stop rule with $\varepsilon = 0.000001$.

Note: the **secant** method is a variation of the method of Newton-Raphson. It considers the update equations

$$\theta^{(t+1)} = \theta^{(t)} - \mathcal{S}(\theta^{(t)}) \frac{\theta^{(t)} - \theta^{(t-1)}}{\mathcal{S}(\theta^{(t)}) - \mathcal{S}(\theta^{(t-1)})}$$

where $\mathcal{S}$ is the score function whose zero we want to approximate. In particular, this method needs two initial estimates, which need to be carefully addressed in order for the method to converge.

(f) Describe and discuss in detail the **Fisher scoring** method. Show, analytically, that the methods of **Newton-Raphson** and **Fisher scoring** coincide in this particular case. Implement it in `R` and use it and the sample above to estimate α - report all the iterations together with the error. Justify your choice of the initial estimates.