# Problem Set 3

## Statistics 509 – Winter 2022

## Due by Wednesday, February 2 in class

**Instructions.** You may work in teams, but you must turn in your own work/code/results. Also for the problems requiring use of the R-package, you need to include a copy of your R-code. This provides us a way to give partial credit in case the answers are not totally correct.

**1.** Exercise 2 on page 81 in Ruppert/Matteson, but as modified below. The data set RecentFord.csv is in Data directory under Files in Canvas.
**(a)** Create a normal Q-Q plot of the returns. Do the returns look normally distributed? If not, how do they differ from being normally distributed?

**(b)** Test for normality using the Shapiro–Wilk test. What is the p-value? Can you reject the null hypothesis of a normal distribution at 0.01?

**(c)** Create several t-plots of the returns using a number of choices of the degrees of freedom parameter (df). What value of df gives a plot that is as linear as possible? Make sure that your QQ-plots and discussion justify your final answer.

**(d)** Based on results from **(c)**, what can you say about asymmetries between the left and right tails?

**2.** Suppose $X_1, X_2, \ldots, X_{100}$ are iid Generalized Error Distribution GED$(0, 1, 1.5)$ and you use a kernel density estimate with a rectangular kernel.

**(a)** Analytically derive an expression for the expected value of the kernel density estimate $\hat{f}_b(x)$ in terms of the cdf of GED$(0, 1, 1.5)$.
*Hint:* You can use that the width, $w$, of rectangular kernel with bandwidth $b$ is approximately $w = b \cdot 3.464$ , and also note that rectangular kernel with bandwidth parameter $b$ and width $w$ satisfies that

$$K_b(x) = \begin{cases} \frac{1}{w} & -\frac{w}{2} \le x \le \frac{w}{2} \\ 0 & \text{otherwise} \end{cases}$$

**(b)** Based on **(a)**, derive an expression for the bias in terms of cdf and pdf of the appropriate Generalized Error Distribution, and then plot the bias of the kernel density estimate as a function of $x$ for $b = .2, .4, .6$.

**(c)** Based on **(a)**, derive an expression for the standard deviation of $\hat{f}_b(x)$ in terms of cdf and pdf of the appropriate Generalized Error Distribution, and then plot the standard deviation of the kernel density estimate as a function of $x$ for $b = .2, .4, .6$.

**(d)** Based on **(b)** and **(c)**, plot the mean-squared error of $\hat{f}_b(x)$ as a function of $x$ for $b = .2, .4, .6$. From these plots, which $b$ is preferred? Explain your answer.

**(e)** If the sample size was increased from $n = 100$ to $n = 500$, could that possibly change your bandwidth selection, and if so what are the possible changes?

*Hint:* Recall that $\text{Bias}(x) = E(\hat{f}_b(x)) - f(x)$. Also as given in Lecture 3, there are R-functions for GED distribution in package fGarch.