4.9 Bibliographic Notes

Exploratory data analysis was popularized by Tukey (1977). Hoaglin, Mosteller, and Tukey (1983,1985) are collections of early articles on exploratory data analysis, data transformations, and robust estimation. Kleiber and Zeileis (2008) is an introduction to econometric modeling with R and covers exploratory data analysis as well as material in latter chapters of this book including regression and time series analysis. The R package AER accompanies Kleiber and Zeileis's book.

The central limit theorem for sample quantiles is stated precisely and proved in textbooks on asymptotic theory such as Serfling (1980); Lehmann (1999), and van der Vaart (1998).

Silverman (1986) is an early book on nonparametric density estimation and is still well worth reading. Scott (1992) covers both univariate and multivariate density estimation. Wand and Jones (1995) has an excellent treatment of kernel density estimation as well as nonparametric regression, which we cover in Chap. 21. Wand and Jones cover more recent developments such as transformation kernel density estimation. An alternative to the TKDE is variable-bandwidth KDE; see Sect. 2.10 of Wand and Jones (1995) as well as Abramson (1982) and Jones (1990).

Atkinson (1985) and Carroll and Ruppert (1988) are good sources of information about data transformations.

Wand, Marron, and Ruppert (1991) is an introduction to the TKDE and discusses methods for automatic selection of the transformation to minimize the expected squared error of the estimator. Applications of TKDE to losses can be found in Bolance, Guillén, and Nielsen (2003).

4.10 R Lab

4.10.1 European Stock Indices

This lab uses four European stock indices in R's EuStockMarkets database. Run the following code to access the database, learn its mode and class, and plot the four time series. The plot() function will produce a plot tailored to the class of the object on which it is acting. Here four time series plots are produced because the class of EuStockMarkets is mts, multivariate time series.

```
data(EuStockMarkets)
mode(EuStockMarkets)
class(EuStockMarkets)
plot(EuStockMarkets)
```

If you right-click on the plot, a menu for printing or saving will open. There are alternative methods for printing graphs. For example,

```
pdf("EuStocks.pdf", width = 6, height = 5)
plot(EuStockMarkets)
graphics.off()
```

will send a pdf file to the working directory and the `width` and `height` parameters allow one to control the size and aspect ratio of the plot.

Problem 1 *Write a brief description of the time series plots of the four indices. Do the series look stationary? Do the fluctuations in the series seem to be of constant size? If not, describe how the volatility fluctuates.*

Next, run the following `R` code to compute and plot the log returns on the indices.

```
logR = diff(log(EuStockMarkets))
plot(logR)
```

Problem 2 *Write a brief description of the time series plots of the four series of log returns. Do the series look stationary? Do the fluctuations in the series seem to be of constant size? If not, describe how the volatility fluctuates.*

In `R`, data can be stored as a data frame, which does not assume that the data are in time order and would be appropriate, for example, with cross-sectional data. To appreciate how `plot()` works on a data frame rather than on a multivariate time series, run the following code. You will be plotting the same data as before, but they will be plotted in a different way.

```
plot(as.data.frame(logR))
```

Run the code that follows to create normal plots of the four indices and to test each for normality using the Shapiro–Wilk test. You should understand what each line of code does.

```
par(mfrow=c(2, 2))
for(i in colnames(logR))
{
  qqnorm(logR[ ,i], datax = T, main = i)
  qqline(logR[ ,i], datax = T)
  print(shapiro.test(logR[ ,i]))
}
```

Problem 3 *Briefly describe the shape of each of the four normal plots and state whether the marginal distribution of each series is skewed or symmetric and whether its tails appear normal. If the tails do not appear normal, do they appear heavier or lighter than normal? What conclusions can be made from the Shapiro–Wilk tests? Include the plots with your work.*

The next set of R code creates t-plots with 1, 4, 6, 10, 20, and 30 degrees of freedom and all four indices. However, for the remainder of this lab, only the DAX index will be analyzed. Notice how the reference line is created by the `abline()` function, which adds lines to a plot, and the `lm()` function, which fits a line to the quantiles. The `lm()` function is discussed in Chap. 9.

```
 1 n=dim(logR)[1]
 2 q_grid = (1:n) / (n + 1)
 3 df_grid = c(1, 4, 6, 10, 20, 30)
 4 index.names = dimnames(logR)[[2]]
 5 for(i in 1:4)
 6 {
 7   # dev.new()
 8   par(mfrow = c(3, 2))
 9   for(df in df_grid)
10   {
11     qqplot(logR[,i], qt(q_grid,df),
12        main = paste(index.names[i], ", df = ", df) )
13     abline(lm(qt(c(0.25, 0.75), df = df) ~
14        quantile(logR[,i], c(0.25, 0.75))))
15   }
16 }
```

If you are running R from Rstudio, then line 7 should be left as it is. If you are working directly in R, then remove the "#" in this line to open a new window for each plot.

Problem 4 *What does the code* `q.grid = (1:n) / (n + 1)` *do? What does* `qt(q.grid, df = df[j])` *do? What does* `paste` *do?*

Problem 5 *For the DAX index, state which choice of the degrees of freedom parameter gives the best-fitting t-distribution and explain why.*

Run the next set of code to create a kernel density estimate and two parametric density estimates, t with `df` degrees of freedom and normal, for the DAX index. Here `df` equals 5, but you should vary `df` so that the t density agrees as closely as possible with the kernel density estimate.

At lines 5–6, a robust estimator of the standard deviation of the t-distribution is calculated using the `mad()` function. The default value of the argument `constant` is 1.4826, which is calibrated to the normal distribution since $1/\Phi^{-1}(3/4) = 1.4826$. To calibrate to the t-distribution, the normal quantile is replaced by the corresponding t-quantile and multiplied by `df/(df - 2)` to convert from the scale parameter to the standard deviation.

```
1 library("fGarch")
2 x=seq(-0.1, 0.1,by = 0.001)
```

```
par(mfrow = c(1, 1))
df = 5
mad_t = mad(logR[ , 1],
   constant = sqrt(df / (df - 2)) / qt(0.75, df))
plot(density(logR[ , 1]), lwd = 2, ylim = c(0, 60))
lines(x, dstd(x, mean = mean(logR[,1]), sd = mad_t, nu = df),
   lty = 5, lwd = 2, col = "red")
lines(x, dnorm(x, mean = mean(logR[ ,1]), sd = sd(logR[ ,1])),
   lty = 3, lwd = 4, col = "blue")
legend("topleft", c("KDE", paste("t: df = ",df), "normal"),
   lwd = c(2, 2, 4), lty = c(1, 5, 3),
   col = c("black", "red", "blue"))
```

To examine the left and right tails, plot the density estimate two more times, once zooming in on the left tail and then zooming in on the right tail. You can do this by using the `xlim` parameter of the `plot()` function and changing `ylim` appropriately. You can also use the `adjust` parameter in `density()` to smooth the tail estimate more than is done with the default value of `adjust`.

Problem 6 *Do either of the parametric models provide a reasonably good fit to the first index? Explain.*

Problem 7 *Which bandwidth selector is used as the default by* `density`*? What is the default kernel?*

Problem 8 *For the CAC index, state which choice of the degrees of freedom parameter gives the best-fitting t-distribution and explain why.*

4.10.2 McDonald's Prices and Returns

This section analyzes daily stock prices and returns of the McDonald's Corporation (MCD) over the period Jan-4-10 to Sep-5-14. The data set is in the file `MCD_PriceDail.csv`. Run the following commands to load the data and plot the adjusted closing prices:

```
data = read.csv('MCD_PriceDaily.csv')
head(data)
adjPrice = data[ , 7]
plot(adjPrice, type = "l", lwd = 2)
```

Problem 9 *Does the price series appear stationary? Explain your answer.*

Problem 10 *Transform the prices into log returns and call that series* `LogRet`*. Create a time series plot of* `LogRet` *and discuss whether or not this series appears stationary.*

The following code produces a histogram of the McDonald's log returns. The histogram will have 80 evenly spaced bins, and the argument `freq = FALSE` specifies the density scale.

```
hist(LogRet, 80, freq = FALSE)
```

Also, make a QQ plot of `LogRet`.

Problem 11 *Discuss any features you see in the histogram and QQ plot, and, specifically, address the following questions: Do the log returns appear to be normally distributed? If not, in what ways do they appear non-normal? Are the log returns symmetrically distributed? If not, how are they skewed? Do the log returns seems heavy tailed compared to a normal distribution? How do the left and right tails compare; is one tail heavier than the other?*

4.11 Exercises

1. This problem uses the data set `ford.csv` on the book's web site. The data were taken from the `ford.s` data set in `R`'s `fEcofin` package. This package is no longer on CRAN. This data set contains 2000 daily Ford returns from January 2, 1984, to December 31, 1991.
   (a) Find the sample mean, sample median, and standard deviation of the Ford returns.
   (b) Create a normal plot of the Ford returns. Do the returns look normally distributed? If not, how do they differ from being normally distributed?
   (c) 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?
   (d) Create several t-plots of the Ford 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? The returns include the return on Black Monday, October 19, 1987. Discuss whether or not to ignore that return when looking for the best choices of df.
   (e) Find the standard error of the sample median using formula (4.3) with the sample median as the estimate of $F^{-1}(0.5)$ and a KDE to estimate f. Is the standard error of the sample median larger or smaller than the standard error of the sample mean?
2. Column seven of the data set `RecentFord.csv` on the book's web site contains Ford daily closing prices, adjusted for splits and dividends, for the years 2009–2013. Repeat Problem 1 using these more recent returns. One of returns is approximately -0.175. For part (d), use that return in place of Black Monday. (Black Monday, of course, is not in this data set.) On what date did this return occur? Search the Internet for news about Ford that day. Why did the Ford price drop so precipitously that day?