5.19 R Lab

5.19.1 Earnings Data

Run the following R code to find a symmetrizing transformation for 1998 earnings data from the Current Population Survey. The code looks at the untransformed data and the square-root and log-transformed data. The transformed data are compared by normal plots, boxplots, and kernel density estimates.

```
library("Ecdat")
?CPSch3
data(CPSch3)
dimnames(CPSch3)[[2]]

male.earnings = CPSch3[CPSch3[ ,3] == "male", 2]
sqrt.male.earnings = sqrt(male.earnings)
log.male.earnings = log(male.earnings)

par(mfrow = c(2, 2))
qqnorm(male.earnings ,datax = TRUE, main = "untransformed")
qqnorm(sqrt.male.earnings, datax = TRUE,
   main = "square-root transformed")
qqnorm(log.male.earnings, datax = TRUE, main = "log-transformed")

par(mfrow = c(2, 2))
boxplot(male.earnings, main = "untransformed")
boxplot(sqrt.male.earnings, main = "square-root transformed")
boxplot(log.male.earnings, main = "log-transformed")

par(mfrow = c(2,2))
plot(density(male.earnings), main = "untransformed")
plot(density(sqrt.male.earnings), main = "square-root transformed")
plot(density(log.male.earnings), main = "log-transformed")
```

Problem 1 *Which of the three transformation provides the most symmetric distribution? Try other powers beside the square root. Which power do you think is best for symmetrization? You may include plots with your work if you find it helpful to do that.*

Next, you will estimate the Box–Cox transformation parameter by maximum likelihood. The model is that the data are $N(\mu, \sigma^2)$-distributed after being transformed by some λ. The unknown parameters are λ, μ, and σ.

Run the following R code to plot the profile likelihood for λ on the grid `seq(-2, 2, 1/10)` (this is the default and can be changed). The command `boxcox` takes an R formula as input. The left-hand side of the formula is the variable to be transformed. The right-hand side is a linear model (see Chap. 9). In this application, the model has only an intercept, which is indicated by

"1." "MASS" is an acronym for "Modern Applied Statistics with S-PLUS," a highly-regarded textbook whose fourth edition also covers R. The MASS library accompanies this book.

```
library("MASS")
par(mfrow = c(1, 1))
boxcox(male.earnings ~ 1)
```

The default grid of λ values is large, but you can zoom in on the high-likelihood region with the following:

```
boxcox(male.earnings ~ 1, lambda = seq(0.3, 0.45, 1 / 100))
```

To find the MLE, run this R code:

```
bc = boxcox(male.earnings ~ 1, lambda = seq(0.3, 0.45, by = 1 / 100),
   interp = FALSE)
ind = (bc$y == max(bc$y))
ind2 = (bc$y > max(bc$y) - qchisq(0.95, df = 1) / 2)
bc$x[ind]
bc$x[ind2]
```

Problem 2 *(a) What are* ind *and* ind2 *and what purposes do they serve?*
(b) What is the effect of interp *on the output from* boxcox*?*
(c) What is the MLE of λ?
(d) What is a 95 % confidence interval for λ?
(e) Modify the code to find a 99 % confidence interval for λ.

Rather than trying to transform the variable male.earnings to a Gaussian distribution, we could fit a skewed Gaussian or skewed t-distribution. R code that fits a skewed t is listed below:

```
library("fGarch")
fit = sstdFit(male.earnings, hessian = TRUE)
```

Problem 3 *What are the estimates of the degrees-of-freedom parameter and of ξ?*

Problem 4 *Produce a plot of a kernel density estimate of the pdf of* male.earnings*. Overlay a plot of the skewed t-density with MLEs of the parameters. Make sure that the two curves are clearly labeled, say with a legend, so that it is obvious which curve is which. Include your plot with your work. Compare the parametric and nonparametric estimates of the pdf. Do they seem similar? Based on the plots, do you believe that the skewed t-model provides an adequate fit to* male.earnings*?*

Problem 5 *Fit a skewed GED model to* `male.earnings` *and repeat Problem 4 using the skewed GED model in place of the skewed t. Which parametric model fits the variable* `male.earnings` *best, skewed t or skewed GED?*

5.19.2 DAX Returns

This section uses log returns on the DAX index in the data set `EuStockMarkets`. Your first task is to fit the standardized t-distribution (std) to the log returns. This is accomplished with the following R code.

Here `loglik_std` is an R function that is defined in the code. This function returns minus the log-likelihood for the std model. The std density function is computed with the function `dstd` in the `fGarch` package. Minus the log-likelihood, which is called the objective function, is minimized by the function `optim`. The `L-BFGS-B` method is used because it allows us to place lower and upper bounds on the parameters. Doing this avoids the errors that would be produced if, for example, a variance parameter were negative. When `optim` is called, `start` is a vector of starting values. Use R's help to learn more about `optim`. In this example, `optim` returns an object `fit_std`. The component `fig_std$par` contains the MLEs and the component `fig_std$value` contains the minimum value of the objective function.

```
data(Garch, package = "Ecdat")
library("fGarch")
data(EuStockMarkets)
Y = diff(log(EuStockMarkets[ ,1]))  #  DAX

#####  std  #####
loglik_std = function(x) {
   f = -sum(dstd(Y, x[1], x[2], x[3], log = TRUE))
   f}
start = c(mean(Y), sd(Y), 4)
fit_std = optim(start, loglik_std, method = "L-BFGS-B",
   lower = c(-0.1, 0.001, 2.1),
   upper = c(0.1, 1, 20), hessian = TRUE)
cat("MLE =", round(fit_std$par, digits = 5))
minus_logL_std = fit_std$value  # minus the log-likelihood
AIC_std = 2 * minus_logL_std + 2 * length(fit_std$par)
```

Problem 6 *What are the MLEs of the mean, standard deviation, and the degrees-of-freedom parameter? What is the value of AIC?*

Problem 7 *Modify the code so that the MLEs for the skewed t-distribution are found. Include your modified code with your work. What are the MLEs? Which distribution is selected by AIC, the t or the skewed t-distribution?*

Problem 8 *Compute and plot the TKDE of the density of the log returns using the methodology in Sects. 4.8 and 5.17. The transformation that you use should be* $g(y) = \Phi^{-1}\{F(y)\}$, *where* F *is the* t*-distribution with parameters estimated in Problem 6. Include your code and the plot with your work.*

Problem 9 *Plot the KDE, TKDE, and parametric estimator of the log-return density, all on the same graph. Zoom in on the right tail, specifically the region* $0.035 < y < 0.06$. *Compare the three densities for smoothness. Are the TKDE and parametric estimates similar? Include the plot with your work.*

Problem 10 *Fit the F-S skewed* t*-distribution to the returns on the FTSE index in* `EuStockMarkets`. *Find the MLE, the standard errors of the MLE, and AIC.*

5.19.3 McDonald's Returns

This section continues the analysis of McDonald's stock returns begun in Sect. 2.4.4 and continued in Sect. 4.10.2. Run the code below.

```
1  data = read.csv('MCD_PriceDaily.csv')
2  adjPrice = data[ ,7]
3  LogRet = diff(log(adjPrice))
4  library(MASS)
5  library(fGarch)
6  fit.T = fitdistr(LogRet, "t")
7  params.T = fit.T$estimate
8  mean.T = params.T[1]
9  sd.T = params.T[2] * sqrt(params.T[3] / (params.T[3] - 2))
10 nu.T = params.T[3]
11 x = seq(-0.04, 0.04, by = 0.0001)
12 hist(LogRet, 80, freq = FALSE)
13 lines(x, dstd(x, mean = mean.T, sd = sd.T, nu = nu.T),
14    lwd = 2, lty = 2, col = 'red')
```

Problem 11 *Referring to lines by number, describe in detail what the code does. Examine the plot and comment on the goodness of fit.*

Problem 12 *Is the mean significantly different than 0?*

Problem 13 *Discuss differences between the histogram and the parametric fit. Do you think that the parametric fit is adequate or should a nonparametric estimate be used instead?*

Problem 14 *How heavy is the tail of the parametric fit? Does it appear that the fitted t-distribution has a finite kurtosis? How confident are you that the kurtosis is finite?*

5.20 Exercises

1. Load the CRSPday data set in the Ecdat package and get the variable names with the commands

   ```
   library(Ecdat)
   data(CRSPday)
   dimnames(CRSPday)[[2]]
   ```

   Plot the IBM returns with the commands

   ```
   r = CRSPday[ ,5]
   plot(r)
   ```

   Learn the mode and class of the IBM returns with

   ```
   mode(r)
   class(r)
   ```

   You will see that the class of the variable r is "ts," which means "time series." Data of class ts are plotted differently than data not of this class. To appreciate this fact, use the following commands to convert the IBM returns to class numeric before plotting them:

   ```
   r2 = as.numeric(r)
   class(r2)
   plot(r2)
   ```

   The variable r2 contains the same data as the variable r, but r2 has class numeric.
   Find the covariance matrix, correlation matrix, and means of GE, IBM, and Mobil with the commands

   ```
   cov(CRSPday[ ,4:6])
   cor(CRSPday[ ,4:6])
   apply(CRSPday[ ,4:6], 2, mean)
   ```

   Use your R output to answer the following questions:
   (a) What is the mean of the Mobil returns?
   (b) What is the variance of the GE returns?
   (c) What is the covariance between the GE and Mobil returns?
   (d) What is the correlation between the GE and Mobil returns?

2. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$, where μ is *known*. Show that the MLE of σ^2 is
$$n^{-1} \sum_{i=1}^{n} (Y_i - \mu)^2.$$
3. Show that $f^*(y|\xi)$ given by Eq. (5.15) integrates to $(\xi + \xi^{-1})/2$.
4. Let X be a random variable with mean μ and standard deviation σ.
   (a) Show that the kurtosis of X is equal to 1 plus the variance of $\{(X - \mu)/\sigma\}^2$.
   (b) Show that the kurtosis of any random variable is at least 1.
   (c) Show that a random variable X has a kurtosis equal to 1 if and only if $P(X = a) = P(X = b) = 1/2$ for some $a \neq b$.
5. (a) What is the kurtosis of a normal mixture distribution that is 95 % $N(0, 1)$ and 5 % $N(0, 10)$?
   (b) Find a formula for the kurtosis of a normal mixture distribution that is $100p\%$ $N(0, 1)$ and $100(1 - p)\%$ $N(0, \sigma^2)$, where p and σ are parameters. Your formula should give the kurtosis as a function of p and σ.
   (c) Show that the kurtosis of the normal mixtures in part (b) can be made arbitrarily large by choosing p and σ appropriately. Find values of p and σ so that the kurtosis is 10,000 or larger.
   (d) Let $M > 0$ be arbitrarily large. Show that for any $p_0 < 1$, no matter how close to 1, there is a $p > p_0$ and a σ, such that the normal mixture with these values of p and σ has a kurtosis at least M. This shows that there is a normal mixture arbitrarily close to a normal distribution but with a kurtosis above any arbitrarily large value of M.
6. Fit the F-S skewed t-distribution to the gas flow data. The data set is in the file `GasFlowData.csv`, which can be found on the book's website.
7. Suppose that $X_1, \ldots, X_n$ are i.i.d. exponential(θ). Show that the MLE of θ is $\overline{X}$.
8. For any univariate parameter θ and estimator $\widehat{\theta}$, we define the bias to be $\text{Bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta$ and the MSE (mean square error) to be $\text{MSE}(\widehat{\theta}) = E(\widehat{\theta} - \theta)^2$. Show that
$$\text{MSE}(\widehat{\theta}) = \{\text{Bias}(\widehat{\theta})\}^2 + \text{Var}(\widehat{\theta}).$$
9. Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} Normal(\mu, \sigma^2)$, with $0 < \sigma^2 < \infty$, and define $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$. What is Bias($\hat{\mu}$)? What is MSE($\hat{\mu}$)? What if the distribution of the X_i is not Normal, but Student's t distribution with the same mean μ and variance σ^2, and tail index (ν, df) of 5?
10. Assume that you have a sample from a t-distribution and the sample kurtosis is 9. Based on this information alone, what would you use as an estimate of ν, the tail-index parameter?
11. The number of small businesses in a certain region defaulting on loans was observed for each month over a 4-year period. In the `R` program below,

the variable y is the number of defaults in a month and x is the value for that month of an economic variable thought to affect the default rate. The function dpois computes the Poisson density.

```
start =c(1,1)
loglik = function(theta) {-sum(log(dpois(y,
   lambda = exp(theta[1] + theta[2] * x))))}
mle = optim(start, loglik, hessian = TRUE)
invFishInfo = solve(mle$hessian)
options(digits = 4)
mle$par
mle$value
mle$convergence
sqrt(diag(invFishInfo))
```

The output is

```
> mle$par
[1] 1.0773 0.4529
> mle$value
[1] 602.4
> mle$convergence
[1] 0
> sqrt(diag(invFishInfo))
[1] 0.08742 0.03912
```

(a) Describe the statistical model being used here.
(b) What are the parameter estimates?
(c) Find 95 % confidence intervals for the parameters in the model. Use a normal approximation.

12. In this problem you will fit a t-distribution by maximum likelihood to the daily log returns for BMW. The data are in the data set bmw that is part of the evir package. Run the following code:

```
library(evir)
library(fGarch)
data(bmw)
start_bmw = c(mean(bmw), sd(bmw), 4)
loglik_bmw = function(theta)
{
-sum(dstd(bmw, mean = theta[1], sd = theta[2],
   nu = theta[3], log = TRUE))
}
mle_bmw = optim(start_bmw, loglik_bmw, hessian = TRUE)
CovMLE_bmw = solve(mle_bmw$hessian)
```

Note: The R code defines a function loglik_bmw that is minus the log-likelihood. See Chap. 10 of *An Introduction to R* for more information about functions in R. Also, see page 59 of this manual for more about maximum likelihood estimation in R. optim minimizes this objective function and returns the MLE (which is `mle_bmw$par`) and other information, including the Hessian of the objective function evaluated at the MLE (because `hessian=TRUE`—the default is not to return the Hessian).

(a) What does the function `dstd` do, and what package is it in?
(b) What does the function `solve` do?
(c) What is the estimate of ν, the degrees-of-freedom parameter?
(d) What is the standard error of ν?

13. In this problem, you will fit a t-distribution to daily log returns of Siemens. You will estimate the degrees-of-freedom parameter graphically and then by maximum likelihood. Run the following code, which produces a 3 × 2 matrix of probability plots. If you wish, add reference lines as done in Sect. 4.10.1.

```
library(evir)
data(siemens)
n = length(siemens)
par(mfrow = c(3, 2))
qqplot(siemens, qt(((1 : n) - 0.5) / n, 2),
   ylab = "t(2) quantiles",
   xlab = "data quantiles")
qqplot(siemens,qt(((1:n)-.5)/n,3),ylab="t(3) quantiles",
   xlab="data quantiles")
qqplot(siemens,qt(((1:n)-.5)/n,4),ylab="t(4) quantiles",
   xlab="data quantiles")
qqplot(siemens,qt(((1:n)-.5)/n,5),ylab="t(5) quantiles",
   xlab="data quantiles")
qqplot(siemens,qt(((1:n)-.5)/n,8),ylab="t(8) quantiles",
   xlab="data quantiles")
qqplot(siemens,qt(((1:n)-.5)/n,12),ylab="t(12) quantiles",
   xlab="data quantiles")
```

`R` has excellent graphics capabilities—see Chap. 12 of *An Introduction to R* for more about `R` graphics and, in particular, pages 67 and 72 for more information about `par` and `mfrow`, respectively.

(a) Do the returns have lighter or heavier tails than a t-distribution with 2 degrees of freedom?
(b) Based on the QQ plots, what seems like a reasonable estimate of ν?
(c) What is the MLE of ν for the Siemens log returns?

References

Arellano-Valle, R. B., and Azzalini, A. (2013) The centred parameterization and related quantities of the skew-t distribution. *Journal of Multivariate Analysis*, 113, 73–90.

Azzalini, A. (2014) *The Skew-Normal and Related Families (Institute of Mathematical Statistics Monographs, Book 3)*, Cambridge University Press.