Spring 2021   **Ali Berra**

University of Central Florida
College of Business

QMB 6911
Capstone Project in Business Analytics

Solutions: Problem Set #6

Probability Density Function of Truck ror

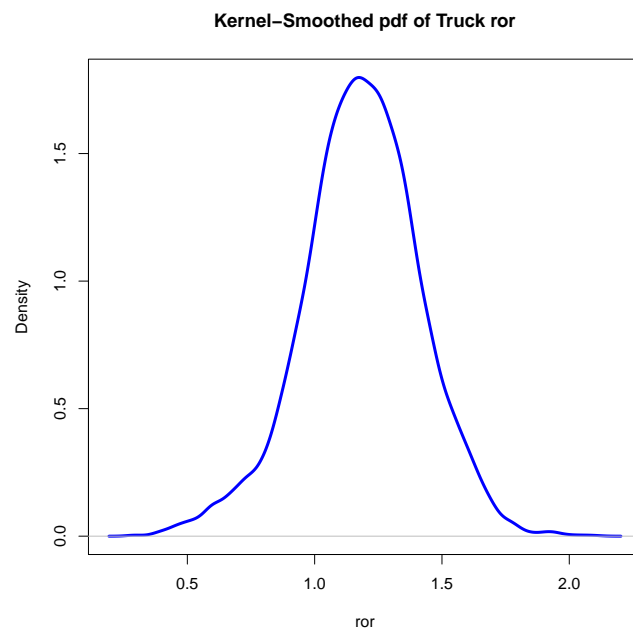Figure 1 shows the kernel-smoothed probability density function of Truck ror.

Figure 1: Probability Density Function of Truck rors

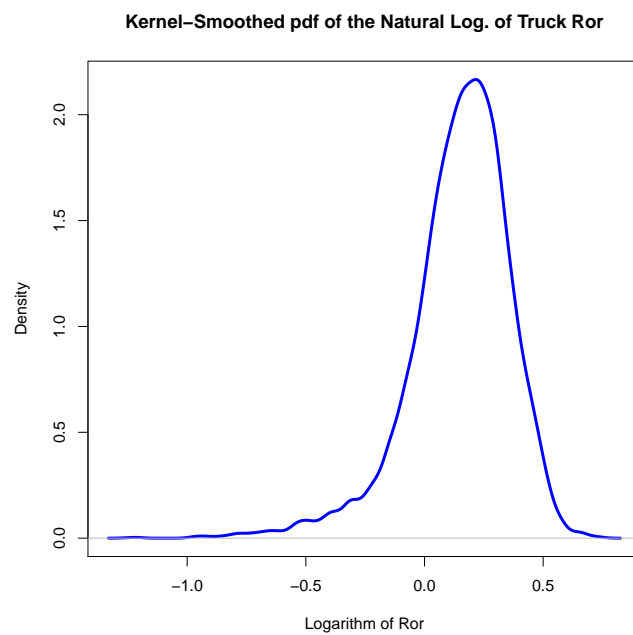As a comparison, Figure 2 shows the kernel-smoothed probability density function of the natural logarithm of ror.

Kernel–Smoothed pdf of the Natural Log. of Truck Ror

Figure 2: Probability Density Function of the Logarithm of Truck ror

Normality of the Original and Transformed Variables

Figure 8 shows a pair of Q-Q plots, comparing quantiles of the empirical distribution against the quantiles of the normal distribution. In the left panel, Figure 3a shows this comparison for the original level of the Truck ror, without transformation. In the right panel, Figure 3b shows this comparison for the logarithmic transformation of Truck ror, without transformation. Consistent with the pair of distributions estimated above, each plot shows a divergence from a normal distribution, suggesting that an optimal transformation might lie somewhere in the middle. The Box-Cox transformation allows for this possibility.
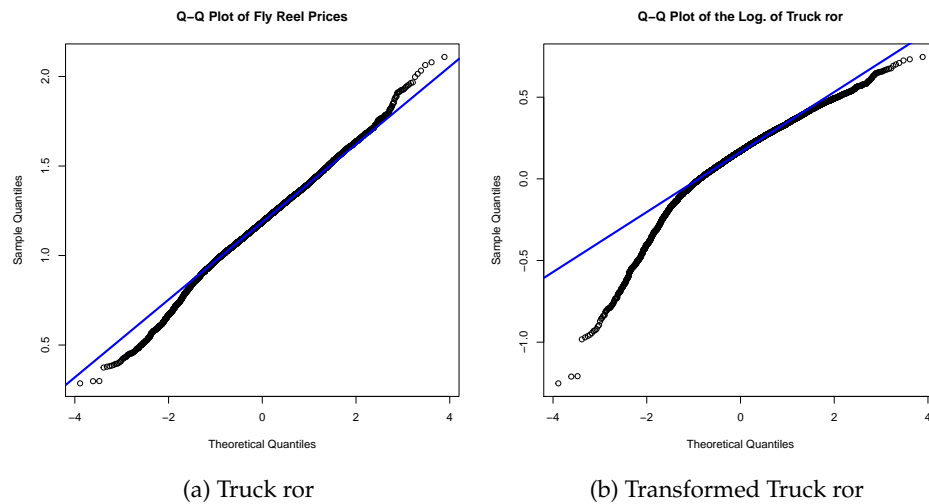
(a) Truck ror   (b) Transformed Truck ror

Figure 3: Q-QPlots of the Log. and Levels of Ror

0.1 Box-Cox Transformation of Truck ror

Under the Box–Cox transformation of P_n, Ror n is calculated as follows,

$$\Lambda(P_n) \equiv \begin{cases} \frac{P_n^\lambda - 1}{\lambda} & \text{if } \lambda > 0 \\ \log P_n & \text{if } \lambda = 0. \end{cases}$$

The following code block defines a function that performs a Box-Cox transformation.

```
# Box-Cox transformation.
Lambda_Price <- function(price, lambda) {
  if (lambda == 0) {
    return(log(price))
  } else {
    return((price^lambda - 1)/lambda)
  }
}
```

0.1.1 Log-likelihood Function

Under the Box-Cox transformation, the fly reel prices can be decomposed into a location parameter μ^0 and an error U, so

$$\Lambda(P_n) = \mu^0(\lambda) + U_n,$$

where the U_ns are independent, mean-zero, constant-variance $\sigma^2(\lambda)$, Gaussian (normal) errors. In the above equation, for clarity, the dependence of μ^0 and $\sigma^2(\lambda)$ on λ is made explicit.

The next code block defines a likelihood function for the normal distribution of the errors as a function of the parameter λ.

```
log_like_uni <- function(ror, lambda) {

  # Calculate maximum likelighood estimates of the
      parameters.
  n <- length(ror)
  lambda_ror <- Lambda_ror(price, lambda)
  mu_0_lambda <- mean(lambda_ror)
  sigma_2_lambda <- sum((lambda_ror - mu_0_lambda)^2)/n

  # Calculate the log-likelihood from the sum of the
      logarithms
  # of the density of the normal distribution.
  like <- - n/2*log(2*pi*sigma_2_lambda)
  like <- like - 1/2/sigma_2_lambda*sum((lambda_ror -
      mu_0_lambda)^2)
  like <- like + (lambda - 1)*sum(log(ror))
  return(like)
}
```

As a first approximation, One can calculate the value of the log-likelihood function on a grid of values to find an optimal value of λ. The plot of this likelihood function is shown in Figure 4. The red points represent the values of the log-likelihood at the optimum $\lambda = 0.43$ and at $\lambda = 0$ and $\lambda = 1$.

```
# Calculate values of the log-likelihood function.
lambda_grid <- seq(-1, 2.5, by = 0.001)
like_grid <- 0*lambda_grid
for (lambda_num in 1:length(lambda_grid)) {
  like_grid[lambda_num] <- log_like_uni(ror = Truck[,
      'ror'],
                                lambda = lambda_grid[lambda_num])
}
```
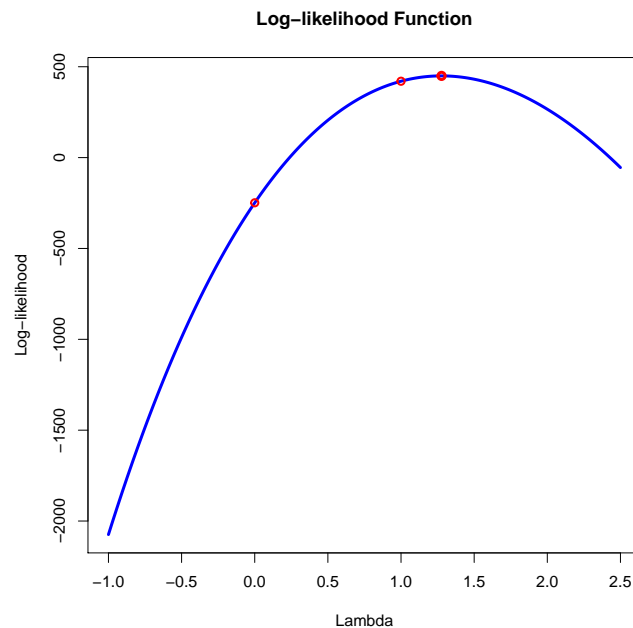

Figure 4: Log-likelihood Function for Box-Cox Transformation

0.1.2   Testing for an Appropriate Transformation

Now we consider the statistical properties of these estimates by calculating a likelihood ratio statistic.

```
> # Calculate likelihood ratio statistics.
> LR_stat_0 <- - 2*(like_mu_0 - like_MLE)
> print(LR_stat_0)
[1] 11.76066
> LR_stat_1 <- - 2*(like_mu_1 - like_MLE)
> print(LR_stat_1)
[1] 2.937618
>
>
> # Compare to quantile of chi-squared distribution with 1
   degree of freedom.
> LR_cv_5 <- qchisq(p = 0.95, df = 1)
> print(LR_cv_5)
[1] 3.841459
>
> # Calculate p-values for these tests.
> p_value_0 <- 1 - pchisq(q = LR_stat_0, df = 1)
> print(p_value_0)
[1] 0.0006049
> p_value_1 <- 1 - pchisq(q = LR_stat_1, df = 1)
> print(p_value_1)
[1] 0.08653825
>
```

Statistically, this is evidence to reject them both. This suggests using the transformation at the MLE. However, one may want to investigate further to find out whether it is worth transforming the data. There exists a trade-of between interpretability and the accuracy of the statistical specification.

0.2 R Packages for the Box-Cox Transformation

Using the MASS Package

As an illustration, we calculated the likelihood ourselves. However, there exist other packages to output the estimation results for an optimal Box-Cox transformation.

One option is to use the function from the MASS package. This is an R package that accompanies a well-know statistics textbook and has a great reputation. In the MASS package, the notation is the same as for a linear model.

```
# In the MASS package, the notation is the same as for a
    linear model.
# i.e., summary(lm(ror ~ 1, data = Truck))
bc_grid_MASS <- MASS::boxcox(ror ~ 1,
                    data = Truck,
                    lambda = lambda_grid)
```
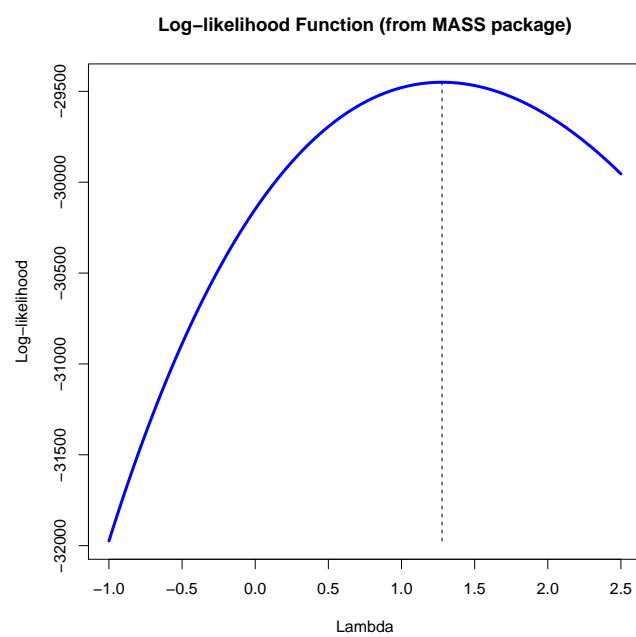
The output is plotted in Figure 5.

Figure 5: Log-likelihood Function for Box-Cox Transformation (`MASS` package)

Using the `car` Package

The `car` package is another well-known option. With this function, the optimization produces a figure automatically from the code below.

```r
bc_grid_car <- car::boxCox(object = lm(data = Truck,
                                      formula = ror ~ 1),
                           lambda = lambda_grid)
```
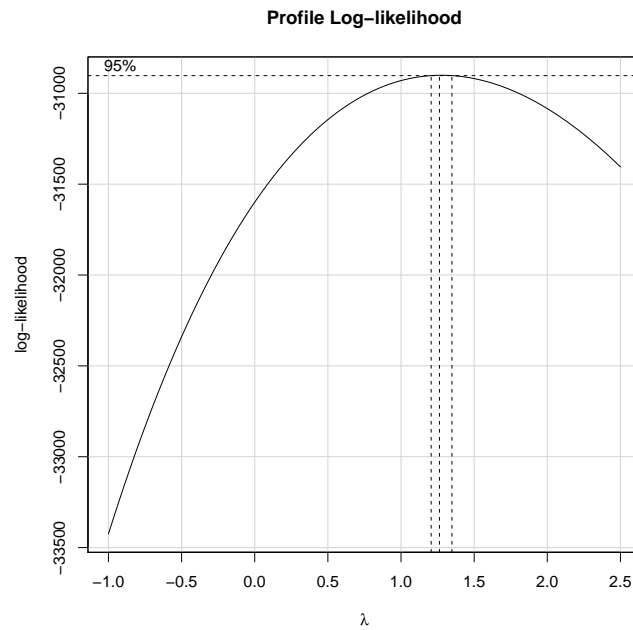
The output is plotted in Figure 6.

Figure 6: Log-likelihood Function for Box-Cox Transformation (`car` package)

Using the `EnvStats` Package

The `EnvStats` package is another option but it is one designed for environmental statistics. That is, it is not a generic package designed for the population of statisticians at large. For that reason, it is missing some of the features that a statistician would expect. The notation and interpretation, however, are similar, except that the straight call to `boxcox` simply does the calculation, unless you specify otherwise.

```
> # Find optimal value of lambda.
> bc_grid_ES_opt <- EnvStats::boxcox(x = Truck[, 'ror'],
+                               lambda = range(lambda_grid),
+                               optimize = TRUE,
+                               objective.name =
   "Log-Likelihood")
>
> bc_grid_ES_opt$lambda
[1] 0.6665808
>
```
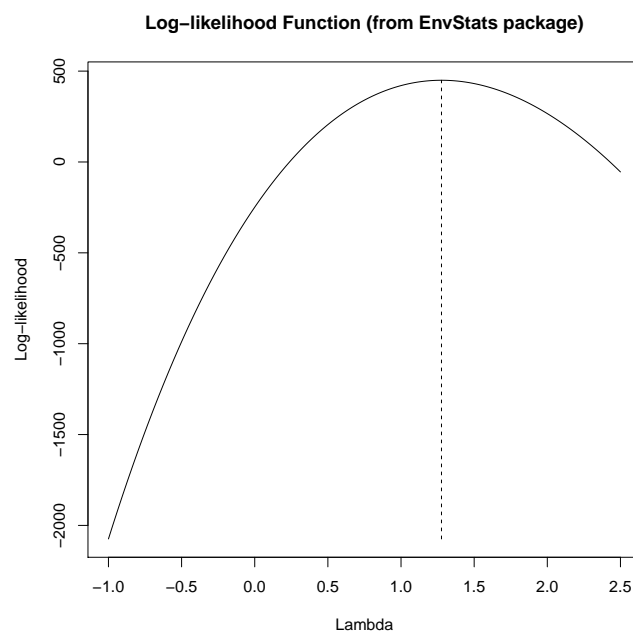
The output is plotted in Figure 7.

Log−likelihood Function (from EnvStats package)

Figure 7: Log-likelihood Function for Box-Cox Transformation (`EnvStats` package)

Normality of the Transformed Variable

Now compare the quantiles of the distribution of the transformed variable with the original. We already plotted normal QQ plot for fly reel prices when considering the log transformation. Now we can generate a new dependent variable with the results from the estimates above.

```
# Generate new dependent variable with results from
   estimates above.
flyreels[, 'Trans_Price'] <- Lambda_Price(price =
   flyreels[, 'Price'],
                             lambda = lambda_hat)
```

Figure **??** shows this comparison and the panel on the right, Figure 8b, shows that the quantiles of the distribution of the transformed variable nearly overlap with those of the normal distribution. From a purely statistical perspective, this provides evidence that the prices are best modeled with the transformation at the optimal $\lambda = 0.43$. From a practical point of view, however, it is still an open question whether this added complexity is warranted when other variables are added to the model.
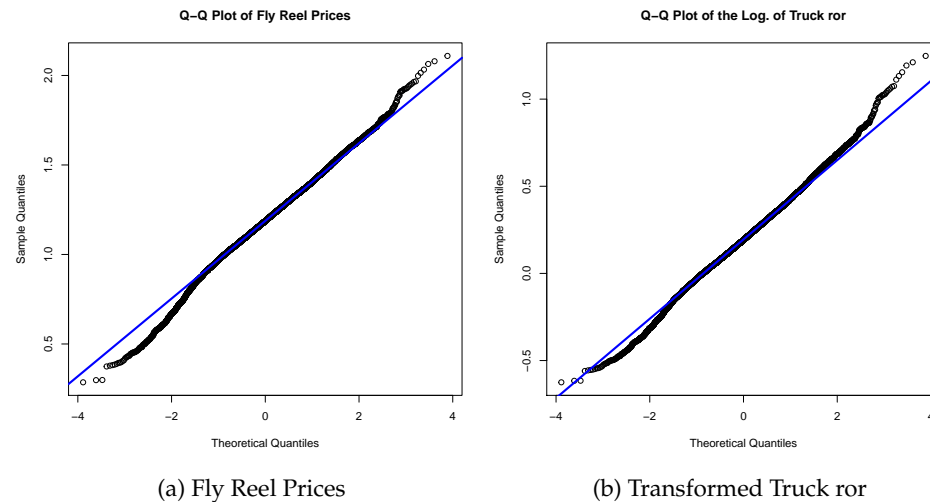

(a) Fly Reel Prices (b) Transformed Truck ror

Figure 8: Q-QPlots of the Transformed Truck ror