Spring 2021 **Firstname M. Lastname**

University of Central Florida
College of Business

QMB 6911
Capstone Project in Business Analytics

Solutions: Problem Set #6

Probability Density Function of Tractor Prices

Figure 1 shows the kernel-smoothed probability density function of tractor prices.

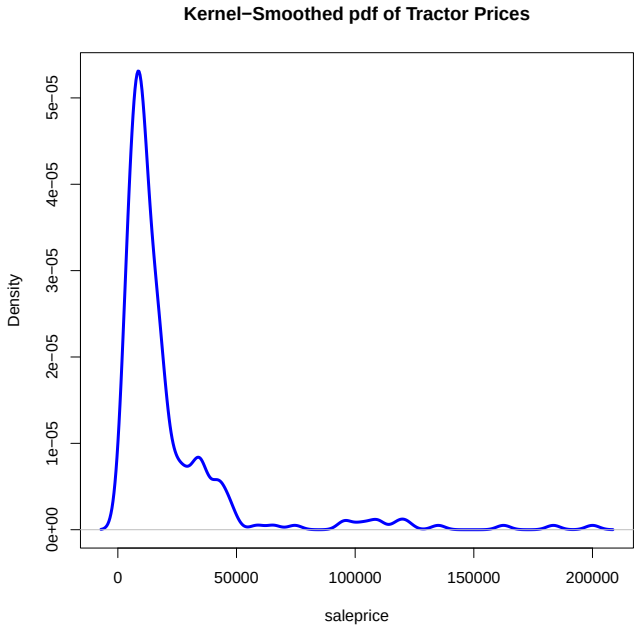


Figure 1: Probability Density Function of Tractor Prices

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

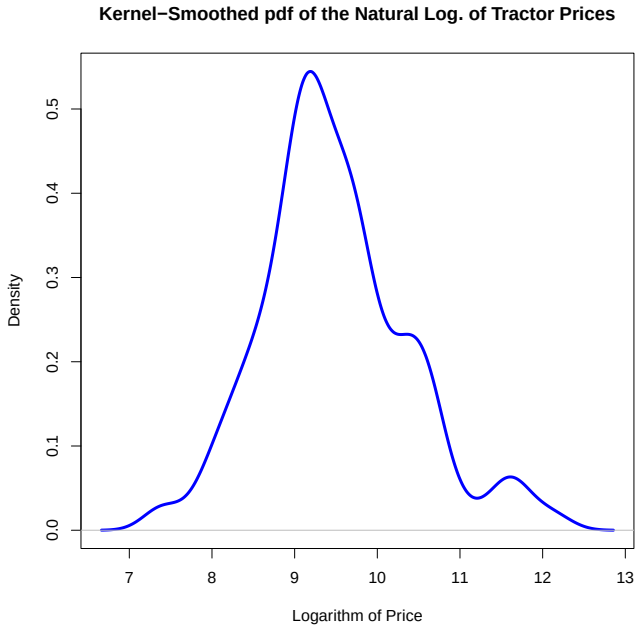
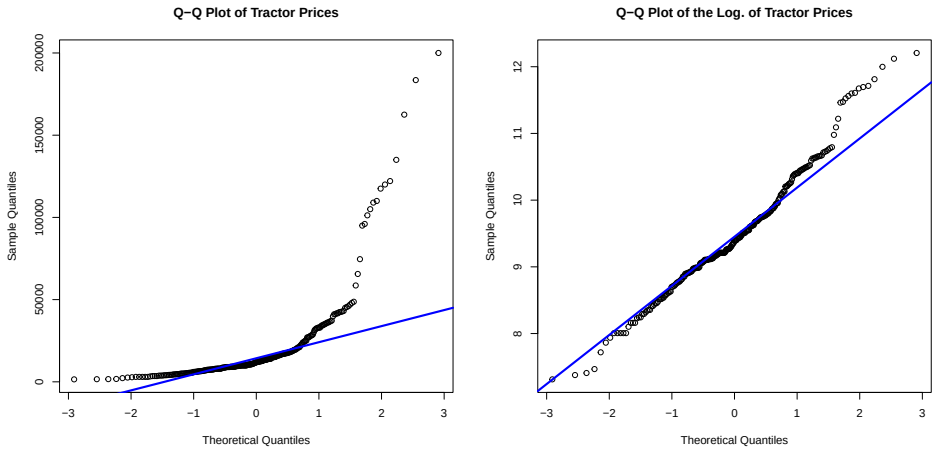


Figure 2: Probability Density Function of the Logarithm of Tractor Prices

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 8a shows this comparison for the original level of the tractor prices, without transformation. In the right panel, Figure 3b shows this comparison for the logarithmic transformation of tractor prices, 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 else. The Box-Cox transformation allows for this possibility.



(a) Tractor Prices (b) Transformed Tractor Prices

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

0.1 Box-Cox Transformation of Tractor Prices

Under the Box–Cox transformation of P_n, the price of tractor 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 tractor 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(price, lambda) {

  # Calculate maximum likelighood estimates of the
     parameters.
  n <- length(price)
  lambda_price <- Lambda_Price(price, lambda)
  mu_0_lambda <- mean(lambda_price)
  sigma_2_lambda <- sum((lambda_price - 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_price -
     mu_0_lambda)^2)
  like <- like + (lambda - 1)*sum(log(price))
  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 $\lambda = -0.17$ 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(price =
     tractor_sales[, 'saleprice'],
                            lambda = lambda_grid[lambda_num])
}
```

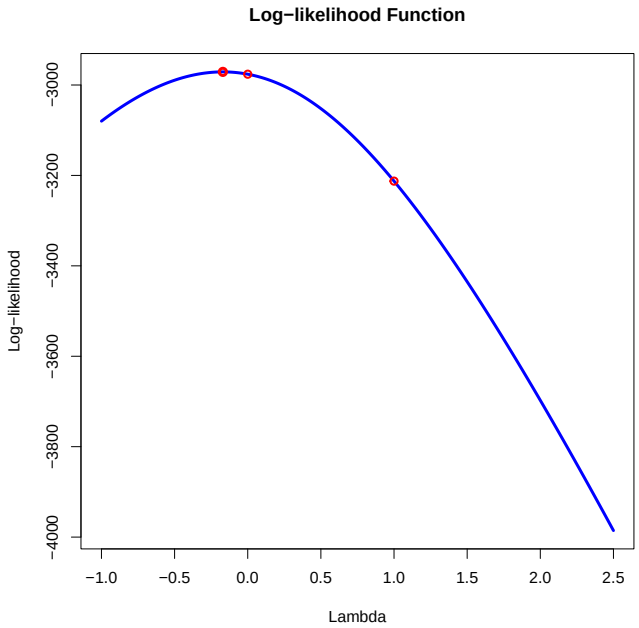


Figure 4: Log-likelihood Function for Box-Cox Transformation

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] 10.26956
> LR_stat_1 <- - 2*(like_mu_1 - like_MLE)
> print(LR_stat_1)
[1] 483.2539
>
>
> # 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] 0.001352434
> p_value_1 <- 1 - pchisq(q = LR_stat_1, df = 1)
> print(p_value_1)
[1] 0
>
```

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, since the Box-Cox transformation at the MLE offers only a marginal improvement over the log transformation. There exists a trade-of between interpretability and the accuracy of the statistical specification, and, perhaps, the log transformation is close enough for practical purposes.

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(saleprice ~ 1, data = tractor_sales))
bc_grid_MASS <- MASS::boxcox(saleprice ~ 1,
                      data = tractor_sales,
                      lambda = lambda_grid)
```

The output is plotted in Figure 5.

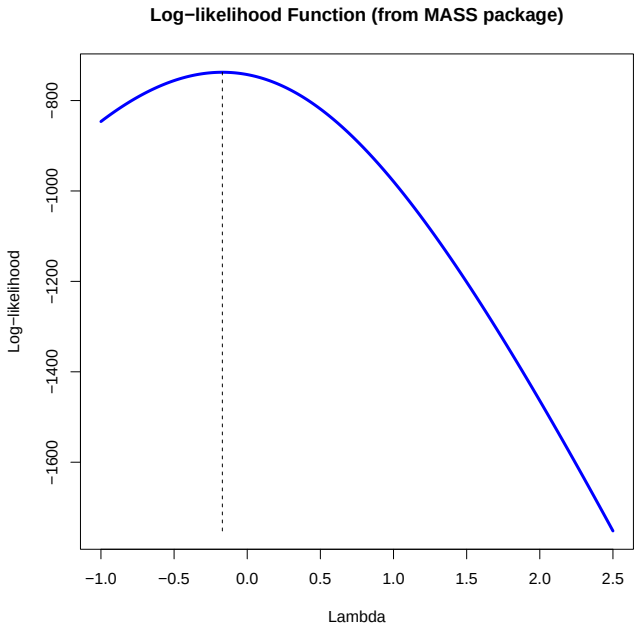


Figure 5: Log-likelihood Function for Box-Cox Transformation (`MASS` package)

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.

```
bc_grid_car <- car::boxCox(object = lm(data = tractor_sales,
                                       formula = saleprice ~ 1),
                           lambda = lambda_grid)
```

The output is plotted in Figure 6.

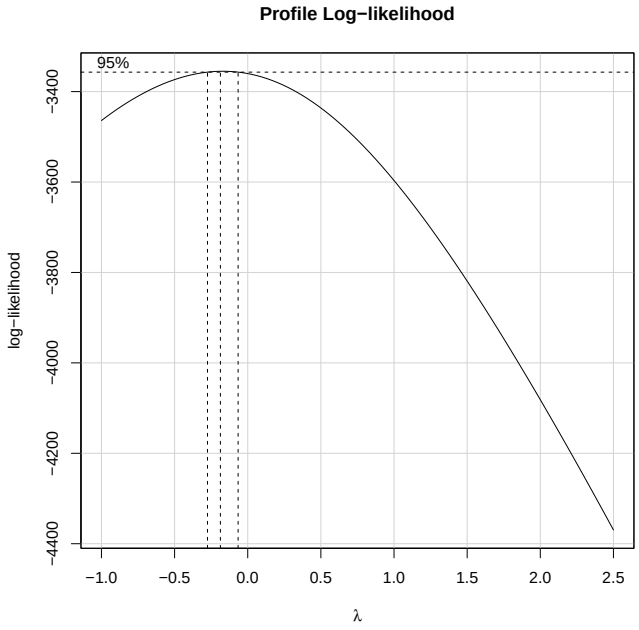


Figure 6: Log-likelihood Function for Box-Cox Transformation (`car` package)

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 = tractor_sales[,
    'saleprice'],
+                                    lambda = range(lambda_grid),
+                                    optimize = TRUE,
+                                    objective.name =
    "Log-Likelihood")
>
> bc_grid_ES_opt$lambda
[1] 0.4295551
>
```

The output is plotted in Figure 7.

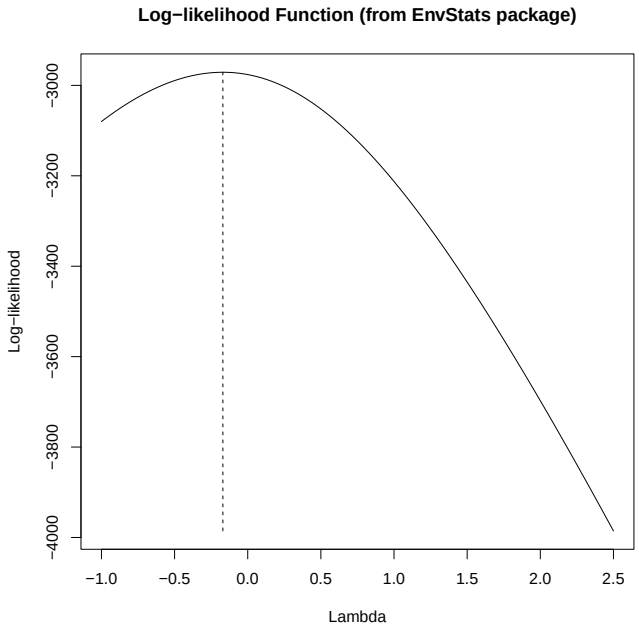


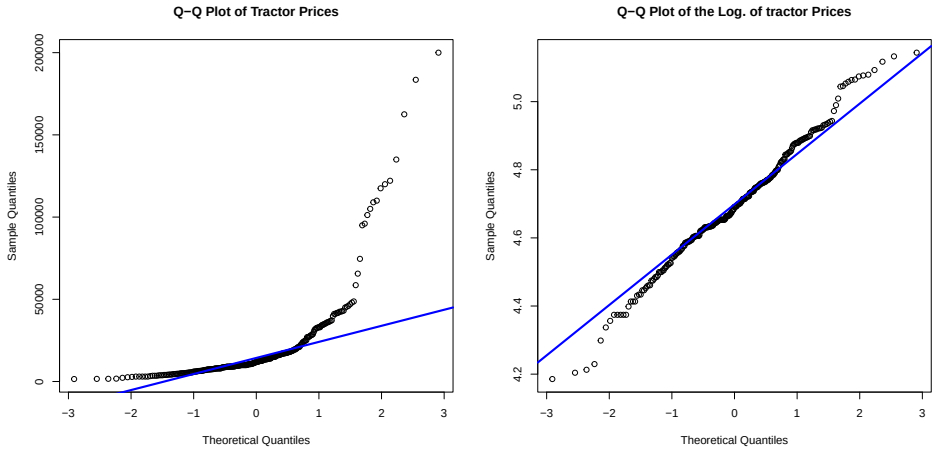
Figure 7: Log-likelihood Function for Box-Cox Transformation (`EnvStats` package)

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 tractor 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.
tractor_sales[, 'trans_saleprice'] <- Lambda_Price(price =
    tractor_sales[, 'saleprice'],
                                    lambda = lambda_hat)
```

Figure 8 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 $\lambda = -0.17$. From a practical point of view, however, the added complexity is not warranted when the log transformation is close enough.



(a) Tractor Prices (b) Transformed Tractor Prices

Figure 8: Q-QPlots of the Transformed Tractor Prices