Homework 6, MATH 5261

Zongyi Liu

Thu, Oct 16, 2025

Github Repository Directory: `https:/github.com/zongyiliu/STAT5261/tree/main/Homework_6`

1 Question 1

1.1 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.

Answer

Besides the square root, I also used 1/4 (if we regard square root as 1/2 power) power and 1/8 power in this case, and I added those graphs together.

```
fourthroot.earnings <- male.earnings^(1/4)   # 4th root
  eighthroot.earnings <- male.earnings^(1/8)   # 8th root
```

Due to the refine of length, I can not show full codes here, but I've put them in the repository listed above, for the reference, please check that.

For the qqplots, the untransformed looks convex, and the log-transformed data looks concave, it is little bit over-transforms to left skewness. The square-root transformed plot looks to be the straightest curve of the three. The 1/4 power transformed looks similar to square-root, and the 1/8 power transformed looks little bit more concave than the 1/4 one. Moreover, the left-tail of the distribution is not modeled very well in all of those five transformations.

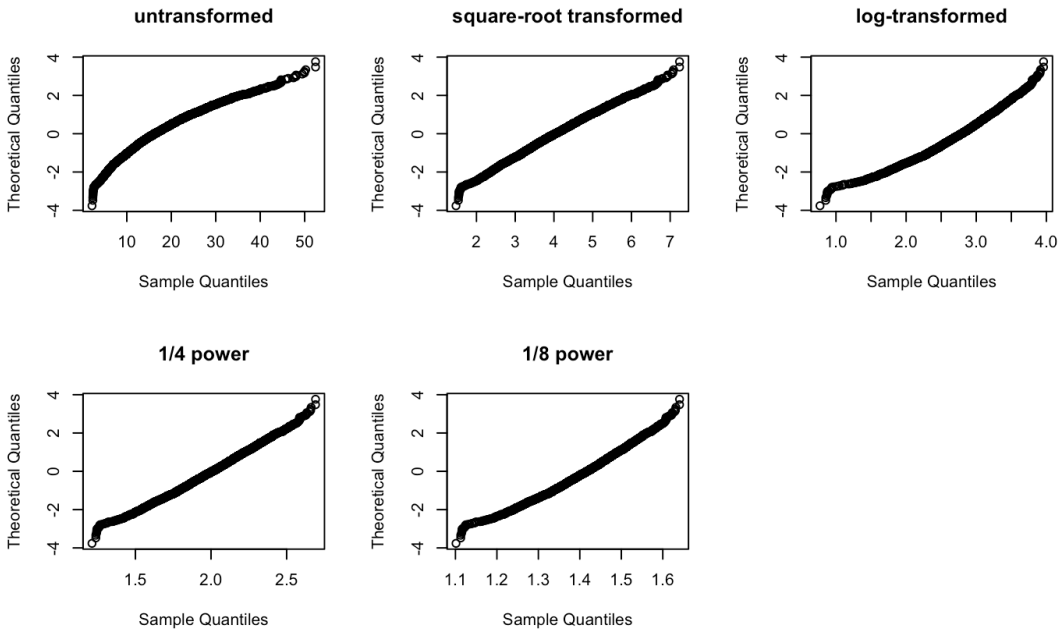

Figure 1: QQ plot of five transformations

For the boxplots, we can see that all of them have a large number of outliers, it is hard to conclude which method is the best because their performances are all not very well.

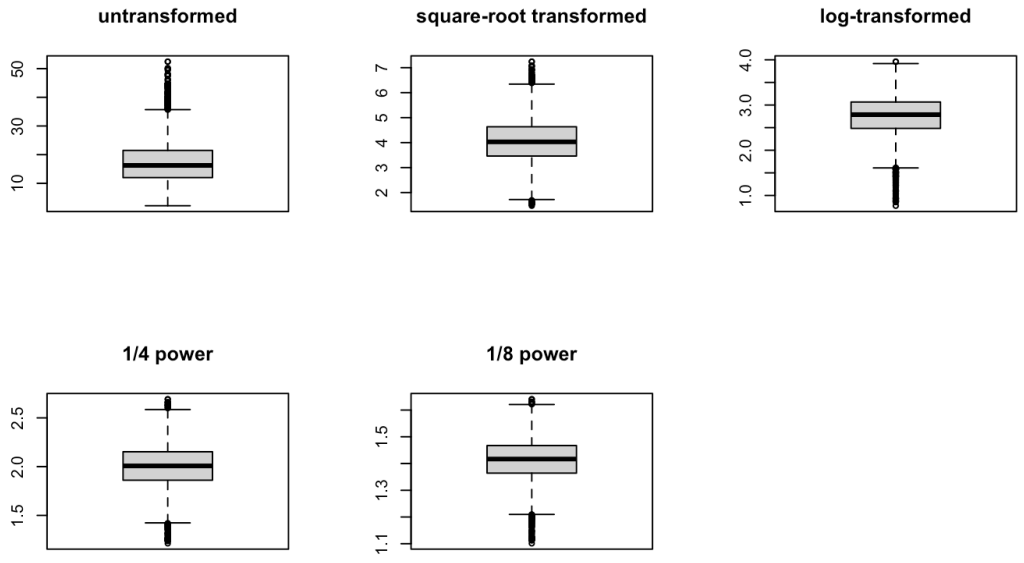

Figure 2: Box plot of five transformations

The kernel density estimate show that the untransformed and the log- transformed variables look skewed to right and left, respectively, while the square-root transformed variables look to have the most normal and symmetric form.

The 1/4 power transformed one is as centered as the square-root transformed, and 1/8 power transformed is more skewed than previous two.

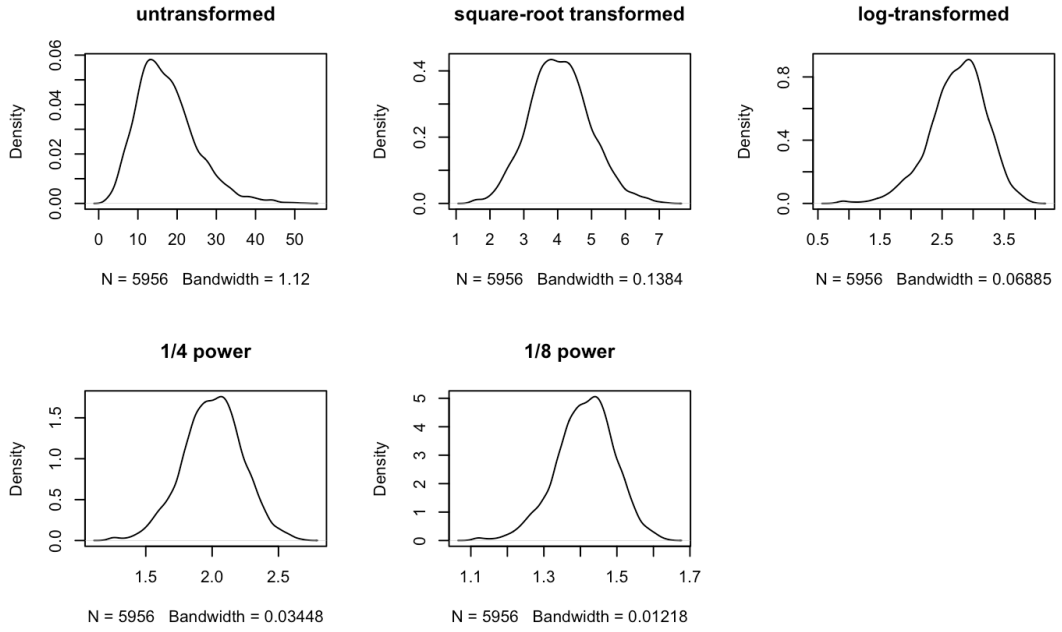

Figure 3: Kernel density plot of five transformations

1.2 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.

Answer

Part a

In R, `ind` are just index variables to store values. Here `ind` helped us to find the MLE. The logical vector `ind` holds a `TRUE` for the index in x (and y) that is the location of the maximum in the Box-Cox likelihood plot.

As for `ind2`, it holds the locations where we are within the 95% confidence interval for λ.

Part b

In the `boxcox()` function, the argument `interp` controls whether interpolation is used to find the optimal value of λ:

- `interp = TRUE` (default): the function uses smooth interpolation (e.g., spline fitting) over the specified range of λ values to estimate the maximum more precisely, rather than being restricted to the discrete grid.
  For example, the result might be $\lambda = 0.376$.
- `interp = FALSE`: the function only computes the log-likelihood for the discrete λ values provided, without performing any interpolation.
  For example, the result might be $\lambda = 0.37$.

Thus, in this case:

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

the function computes the log-likelihood only at the discrete points $\lambda = 0.3, 0.31, \ldots, 0.45$, without interpolating between them.

Part c

The optimal value of λ to use in the Box-Cox transformation is found to be:

```
[1] "Maxlikihood lambda value= 0.360000"
```

This can also be double-checked with the plot of location of MLE.

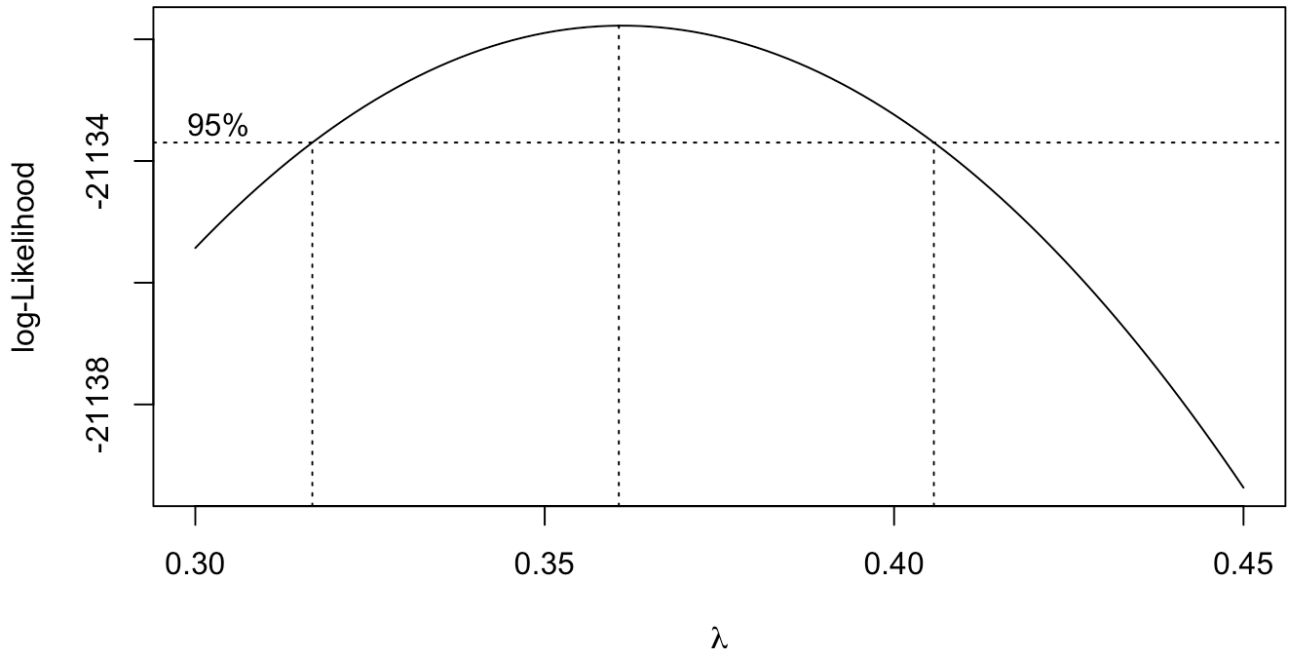

Figure 4: The output from the Box-Cox transformation showing the location of the MLE

Part d

A 95% confidence interval for λ is given as below, rounded to 2 digits:

```
> range(bc$x[ind2])
[1] 0.32 0.40
```

Part e

We change 0.95 into 0.99 in this part and thus can get the 99% confidence interval.

```
ind2 = (bc$y > max(bc$y) - qchisq(0.95, df = 1) / 2)
```

Similarly, a 99% confidence interval for λ is given as below, rounded to 2 digits:

```
> range(bc$x[ind2])
[1] 0.31 0.41
```

1.3 Problem 3

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

Answer

We can find the `fit`:

```
> fit
$minimum
[1] 20121.41

$estimate
mean        sd        nu        xi
17.322933   7.492441  21.599882   1.651652
```

These are the parameters in a Fernandez–Steel (F–S) skewed t-distribution. In this case, degrees-of-freedom, $\nu = 21.599949$, it is large, indicating the tail approximates normal but still pretty heavy; $\xi = 1.651652$, which is greater than 1, and this indicates right skewness of the data.

2 Question 2

Use the data set in HW5 to identify which distribution fits better each set of returns (try t, normal, ged ect)

Answer

I loaded the dataset for HW 5 and excluded the column for risk-free asset, and then fit the three distributions. With AIC and BIC criterias; For full codes to implement this, please see my repository. It turns out that normal distribution fits the best for all five cases, however, as states in the book, typically the assets are followed by the Student's t-distribution, here the factor leading to this result might be the dates are just 120, which does not have too many extreme data to distort the overall distribution.

Asset <chr>	Best_Distribution <chr>	AIC <dbl>	LogLikelihood <dbl>	KS_Statistic <dbl>
GE	Normal	-294.4946	149.24732	0.05344125
GM	Normal	-226.0596	115.02980	0.07724434
IBM	Normal	-213.1882	108.59410	0.07291649
MPORT	Normal	-392.8029	198.40147	0.10672905
MSOFT	Normal	-173.1735	88.58673	0.08424863

5 rows

Figure 5: Results 1

	Distribution <chr>	LogLikelihood <dbl>	AIC <dbl>	BIC <dbl>	KS_Statistic <dbl>	KS_PValue <dbl>	Asset <chr>
D	Normal	88.58673	-173.1735	-167.5985	0.08424863	0.3619076	MSOFT
D1	Normal	149.24732	-294.4946	-288.9197	0.05344125	0.8829799	GE
D2	Normal	115.02980	-226.0596	-220.4846	0.07724434	0.4711576	GM
D3	Normal	108.59410	-213.1882	-207.6132	0.07291649	0.5461612	IBM
D4	Normal	198.40147	-392.8029	-387.2280	0.10672905	0.1299002	MPORT

5 rows

Figure 6: Results 2

For the distributions of four assets and the market portfolio can be plotted as below. We can see that `MSOFT`, `GE` and `GM` are centered near zero, and `IBM` is left-skewed whereas `MPORTM` is a little bit right skewed.

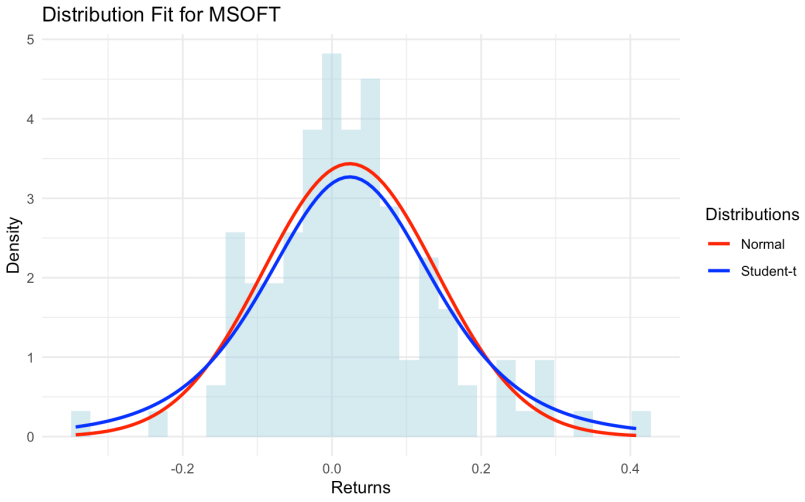

Figure 7: Distribution fit for `MSOFT`

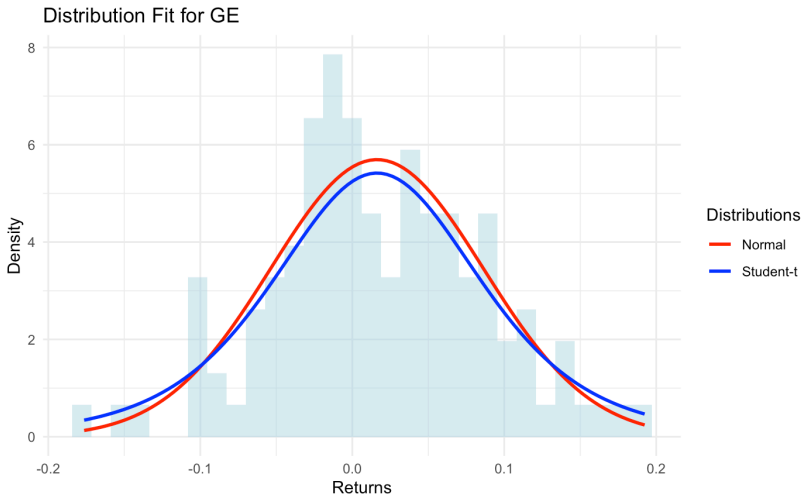

Figure 8: Distribution fit for `GE`

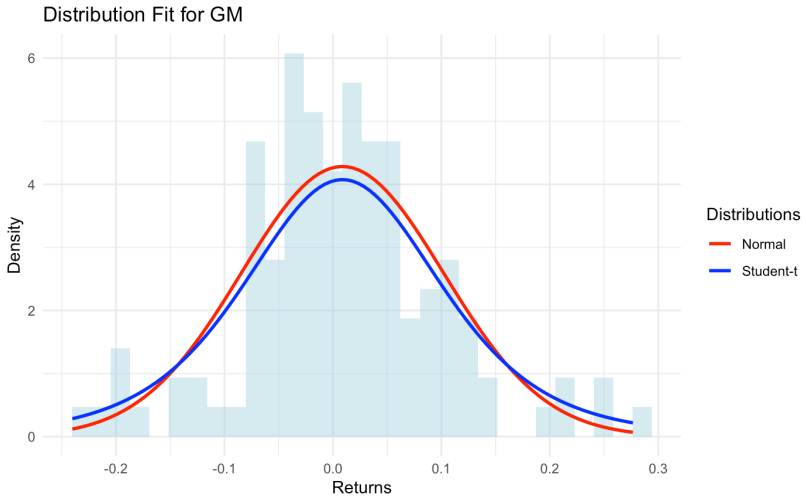

Figure 9: Distribution fit for `GM`

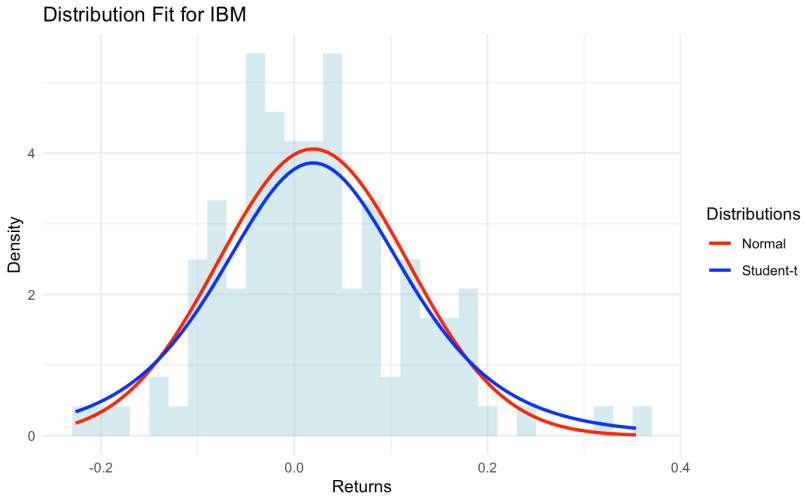

Figure 10: Distribution fit for `IBM`

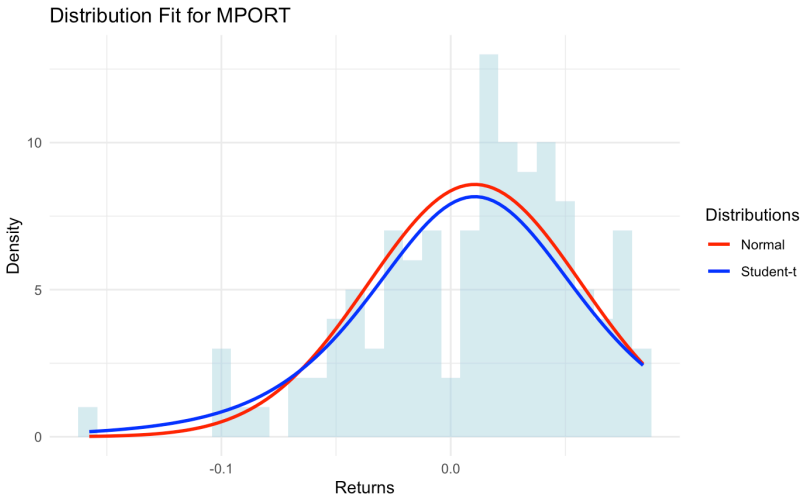

Figure 11: Distribution fit for `MPORT`

3 Question 3

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?

Answer

Part a

The printout results are as follows:

```
[1] "year" "month" "day" "ge" "ibm" "mobil" "crsp"
[1] "numeric"
[1] "ts"
[1] "numeric"

            ge          ibm         mobil
ge     1.882164e-04 8.007660e-05 5.270394e-05
ibm    8.007660e-05 3.061309e-04 3.588748e-05
mobil  5.270394e-05 3.588748e-05 1.670265e-04

          ge        ibm      mobil
ge     1.0000000 0.3335979 0.2972499
ibm    0.3335979 1.0000000 0.1587072
mobil  0.2972499 0.1587072 1.0000000

ge         ibm        mobil
0.0010713801 0.0007000767 0.0007788801
```

We can see that the mean value of Mobil is 0.0007789; and values for GE and IBM are 0.00107 and 0.0007 respectively.

Part b

We can see that the variance of GE is 1.882164e-04, and variance for IBM and Mobil are 3.061309e-04 and 1.670265e-04 respectively.

Part c

The covariance of GE and Mobil returns is 5.270394e-05.

Part d

The correlation between GE and Mobil returns is 0.2972499.

4 Question 4

Show that why Eq. (5.15) can be intergrated into $\frac{1}{2}(\xi^{-1} + \xi)$.

Answer

Let f be a symmetric probability density function about 0, so that

$$\int_{-\infty}^{\infty} f(y)\,dy = 1, \qquad \int_{-\infty}^{0} f(y)\,dy = \int_{0}^{\infty} f(y)\,dy = \frac{1}{2}.$$

Define

$$f^*(y \mid \xi) = \begin{cases} f(\xi y), & y < 0, \\ f(y/\xi), & y \geq 0, \end{cases} \qquad \xi > 0.$$

We compute the unnormalized integral:

$$\int_{-\infty}^{\infty} f^*(y \mid \xi)\,dy = \underbrace{\int_{-\infty}^{0} f(\xi y)\,dy}_{I_1} + \underbrace{\int_{0}^{\infty} f(y/\xi)\,dy}_{I_2}.$$

For the first term, let $u = \xi y$, so $dy = du/\xi$, and $y \in (-\infty, 0) \Rightarrow u \in (-\infty, 0)$. Hence

$$I_1 = \int_{-\infty}^{0} f(\xi y)\,dy = \frac{1}{\xi}\int_{-\infty}^{0} f(u)\,du = \frac{1}{\xi}\cdot\frac{1}{2}.$$

For the second term, let $v = y/\xi$, so $dy = \xi\,dv$, and $y \in (0, \infty) \Rightarrow v \in (0, \infty)$. Hence

$$I_2 = \int_{0}^{\infty} f(y/\xi)\,dy = \xi\int_{0}^{\infty} f(v)\,dv = \xi\cdot\frac{1}{2}.$$

Adding both parts gives

$$\int_{-\infty}^{\infty} f^*(y \mid \xi)\,dy = \frac{1}{2}(\xi^{-1} + \xi).$$