Practice 7: Curve fitting

Mayra Cristina Berrones Reyes 6291

October 20, 2020

1 Introduction

Part of the subject of curve fitting is the concept of correlation. In the field of statistics, correlation or dependence is any statistical relationship that two random variables have between each other. In commonly refers to the degree in which the pair is linearly related [1].

Correlations are useful because they can indicate the predictive side of a relationship, that help explore the data even further. The most familiar correlation measure is the Pearson coefficient, commonly known as the correlation coefficient. The product of this correlation attempts to establish a line of the best fit between two variables.

There is also a way to calculate the correlation between two variables, reveling non-linear interactions. They are called transformations. In correlation coefficient, there is no need for its values to have a normal shape, but it certainly helps to make them more clear to understand if the data is rearranged. This is where some transformations come in handy, depending on the type of data we are working on [2].

One transformation that we rely on for this experimentation, is the Tukey ladder of powers. A brief summary of this transformation is that we assume we have a collection of data, and we are interested to know the relationship between this variables. As we said before, a good way to understand our data is to re expressing its variables, in this case, using the power transformation [2].

Table 1 gives an example of the Tukey ladder of transformations.

2 Experimentation

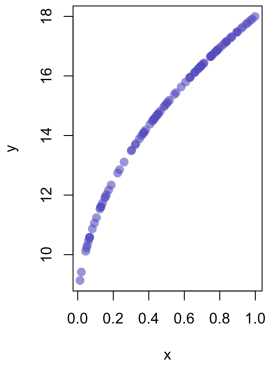
For this experimentation we are working with 4 different equations that have an `x` of random uniform numbers, with a `y` dependent of the value of x. Then we have a fifth experiment with an `x` with a different distribution, and a `y` dependent on that `x` but with a random value.

Table 1: Tukey ladder of transformations.

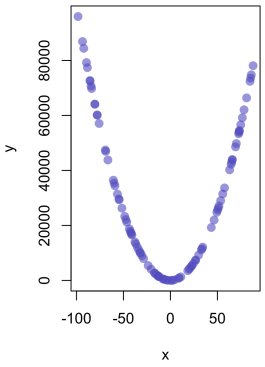
λ	-2	-1	-1/2	0	1/2	1	2
y	$\frac{1}{x^2}$	$\frac{1}{x}$	$\frac{1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

In Figure 1 we can see the scatter plots of all of this equations. In Figure 1a we have a polynomial equation. Figure 1b is a quadratic equation. Figure 1c is a exponential equation and Figure 1d is a logarithmic equation. Figure 1e is the one experimenting with the value of `x` and `y` .

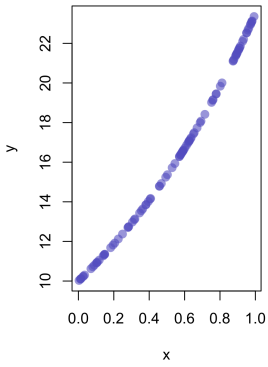
Visualization is a great tool to understand how some transformations work, so we use the `mosaic` and `manipulate` libraries in R to plot the behavior in each equation of the Tukey ladder of powers.



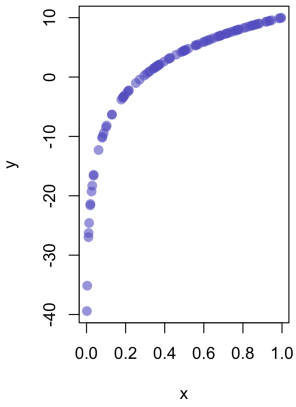
(a) $y = a * sqrt(x) + b$



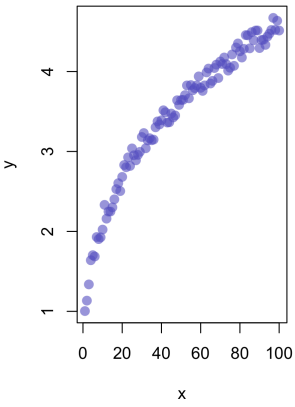
(b) $y = a * x^2 + b * x + c$



(c) $y = a * (exp(p * x))$

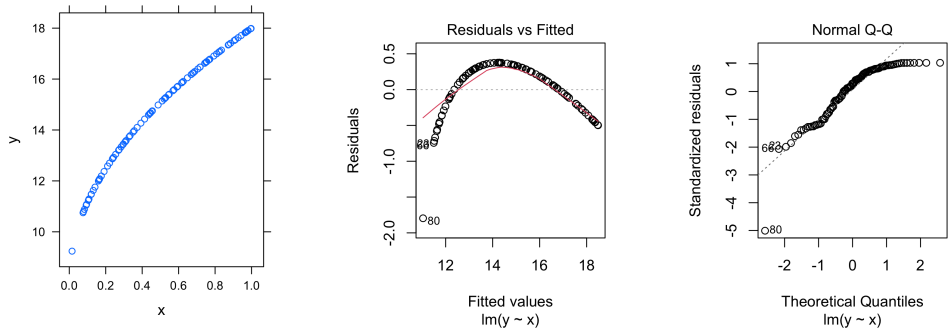


(d) $y = a + b * log(x)$



(e) y = jitter($x^p, factor = length(x)/2$)

Figure 1: Scatter plots of each equation.



(a) (b) (c)

Figure 2: Scatter plot of equation (a), Residuals vs. Fitted plot (b), and Normal Q-Q plot (c) of polynomial equation.

2.1 Polynomial

Polynomial equations are the most flexible tool to describe linear parameters, and can also be fitted with linear regression. The equation used in this part is Equation 1,

$$y = a * sqrt(x) + b \tag{1}$$

where:

`a` and `b` are fixed variables,

`x` random uniform variable.

In Figure 2 we can see the scatter plot of the equation 2a, the residuals versus fitted values 2b, and the normal Q-Q plot 2c. We used both values of `x, y` to build a data frame. Then, we pretend that we do not now the fixed variables we put in the equation. With the plots in 2 we can asees if the fit for a linear model is appropriate or not.

As stated before, we are using the Tukey ladder of powers to determine an approximate transformation of the `y` variable. With the `manipulate` library in R we automate the process of searching for an appropriate value for lambda. Figure 3 shows 13 iterations.

The best fit for a linear model is Figure 3a with λ= -3.

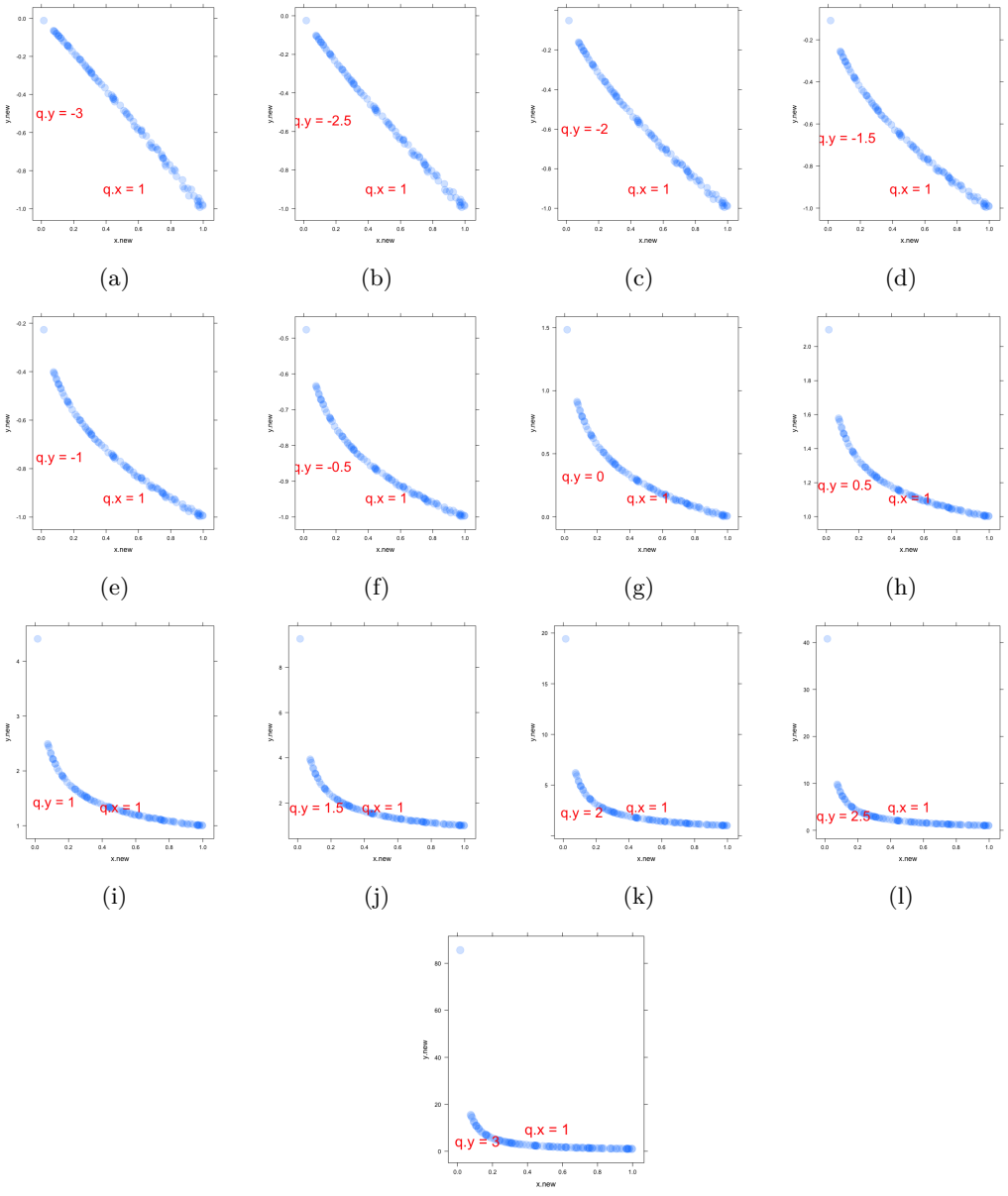
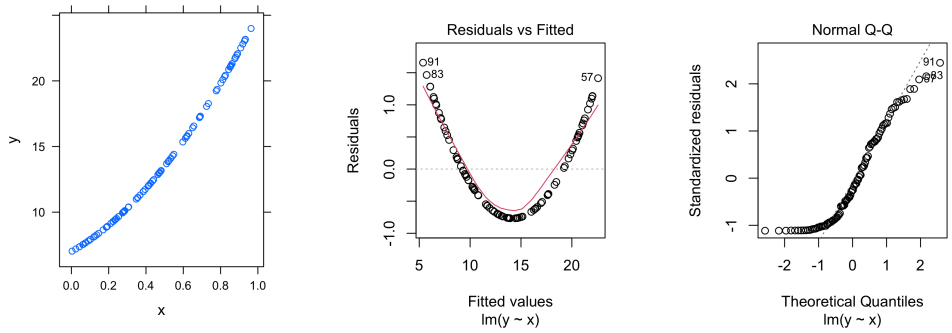


Figure 3: Iterations of the different values of lambda for the Tukey ladder of powers for the polynomial equation.



(a) (b) (c)

Figure 4: Scatter plot of equation (a), Residuals vs. Fitted plot (b), and Normal Q-Q plot (c) of quadratic equation.

2.2 Quadratic

The equation used for this is Equation 2,

$$y = a * x^2 + b * x + c \tag{2}$$

where all the fixed values are the same as the one used in Section 2.1.

In Figure 5 we have the iterations of the Tukey ladder. Examining this plot, Figure 5i is the one that shows a linear behaviour. In this case, $\lambda = 1$.

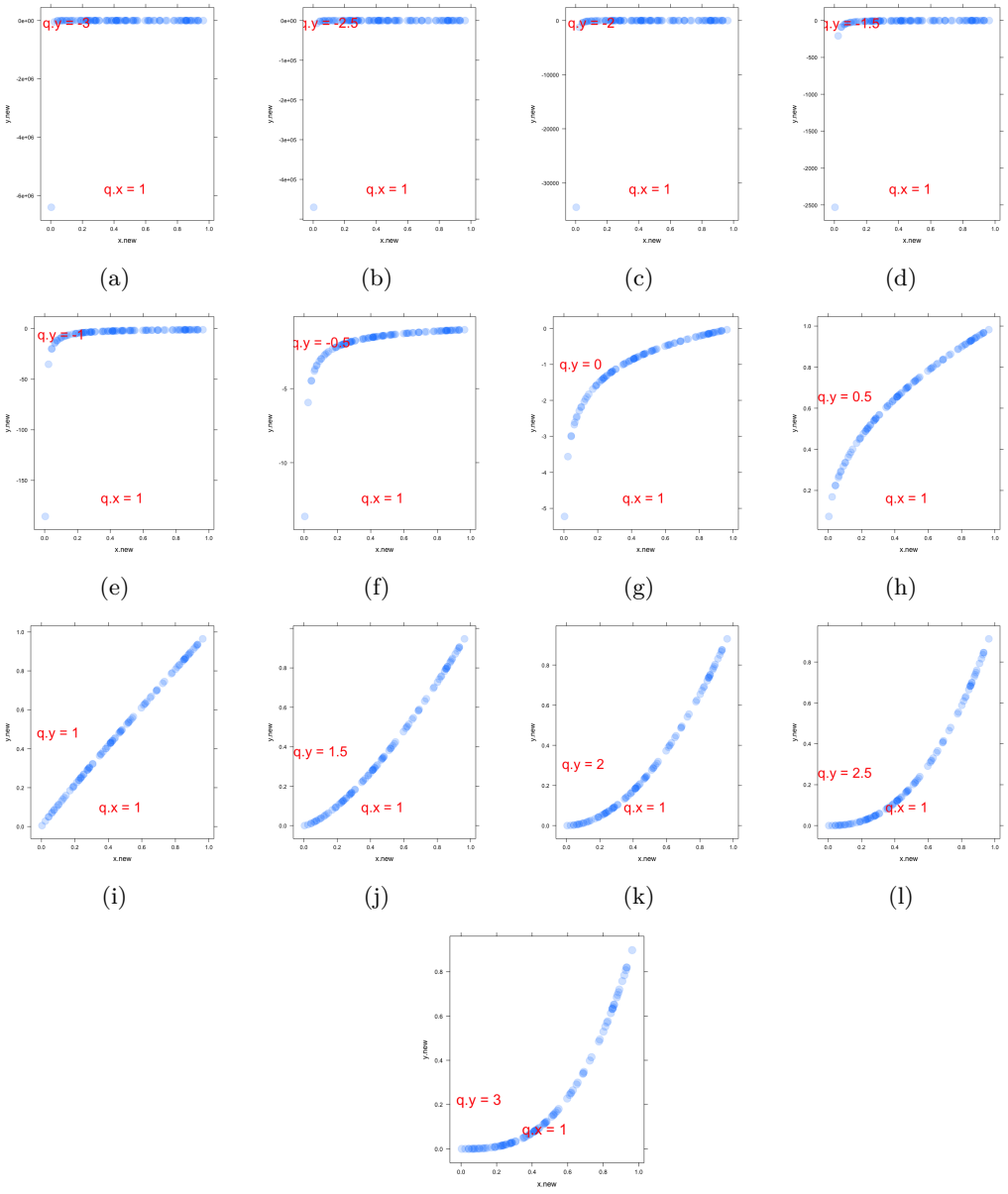
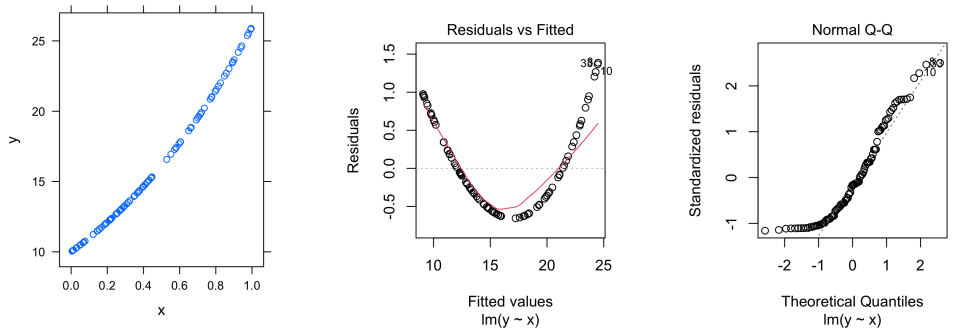


Figure 5: Iterations of the different values of lambda for the Tukey ladder of powers for the quadratic equation.



(a) $y = a * sqrt(x) + b$ (b) $y = a * x^2 + b * x + c$ (c) $y = a * (exp(p * x))$

Figure 6: Scatter plot of equation (a), Residuals vs. Fitted plot (b), and Normal Q-Q plot (c) of exponential equation.

2.3 Exponential

In the case of the exponential equation, it describes an increasing or decreasing trend, with a constant relative rate. Equation 3 shows the parameters used,

$$y = a * (exp(p * x)) \tag{3}$$

where all the constant variables are the same used in the polinomial and quadratic equations. The value of `p` is a standard normal random variable.

Seeing Figure 7 we see that the closer linear plot is between the values of Figure 7f and Figure 7g, so instead of showing the value of $\lambda = 3$, we plot the value between λ = -0.5 and λ = 0, and in Figure 7m we use the $\lambda = 0.25$. This shows to be the more linear in counts of behavior.

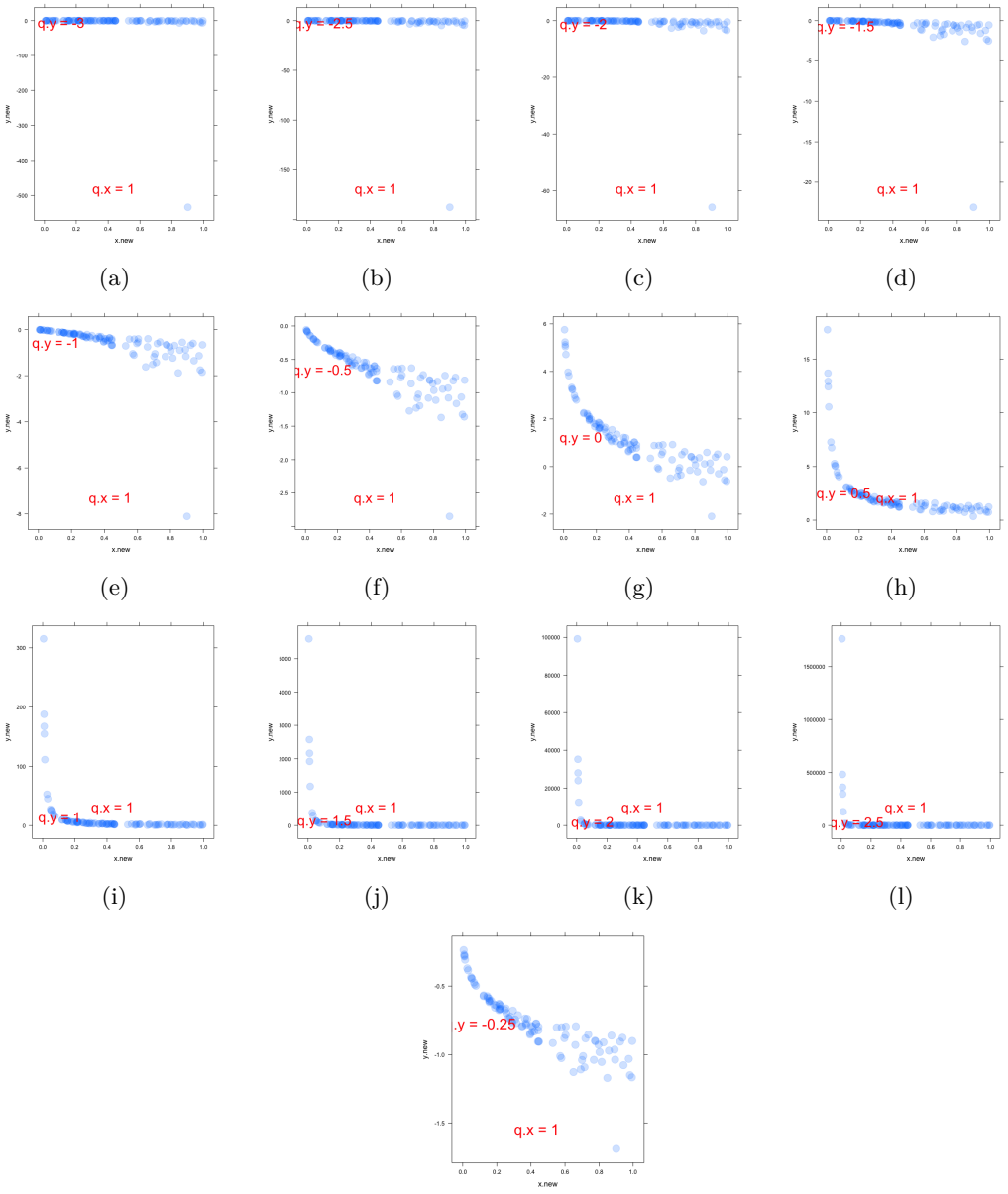
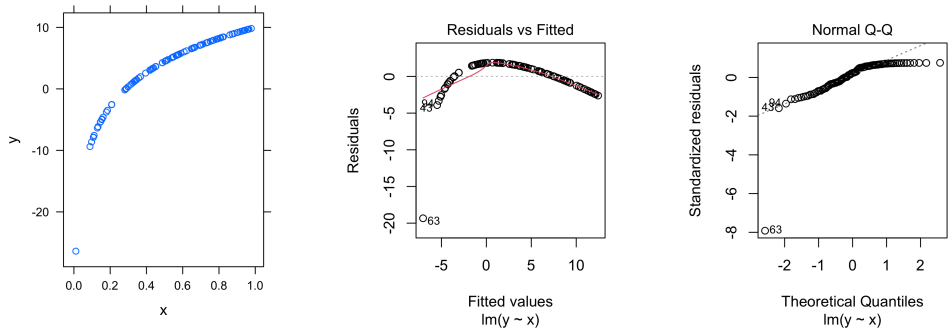


Figure 7: Iterations of the different values of lambda for the Tukey ladder of powers for the exponential equation.



(a) (b) (c)

Figure 8: Scatter plot of equation (a), Residuals vs. Fitted plot (b), and Normal Q-Q plot (c) of logarithmic equation.

2.4 Logarithmic

In this case, because of the logarithmic properties, `x` must be > to 0. Equation 4 shows the parameters used,

$$y = a + b * log(x) \tag{4}$$

where b is the parameter that determines the shape of the plot, and there are the same fixed variables used in the polynomial, quadratic and exponential experiments.

For the logarithmic equation, in Figure 9 we analyze the iterations of lambda and see that Figure 9e is the one with a linear behaviour, being λ = -1.5.

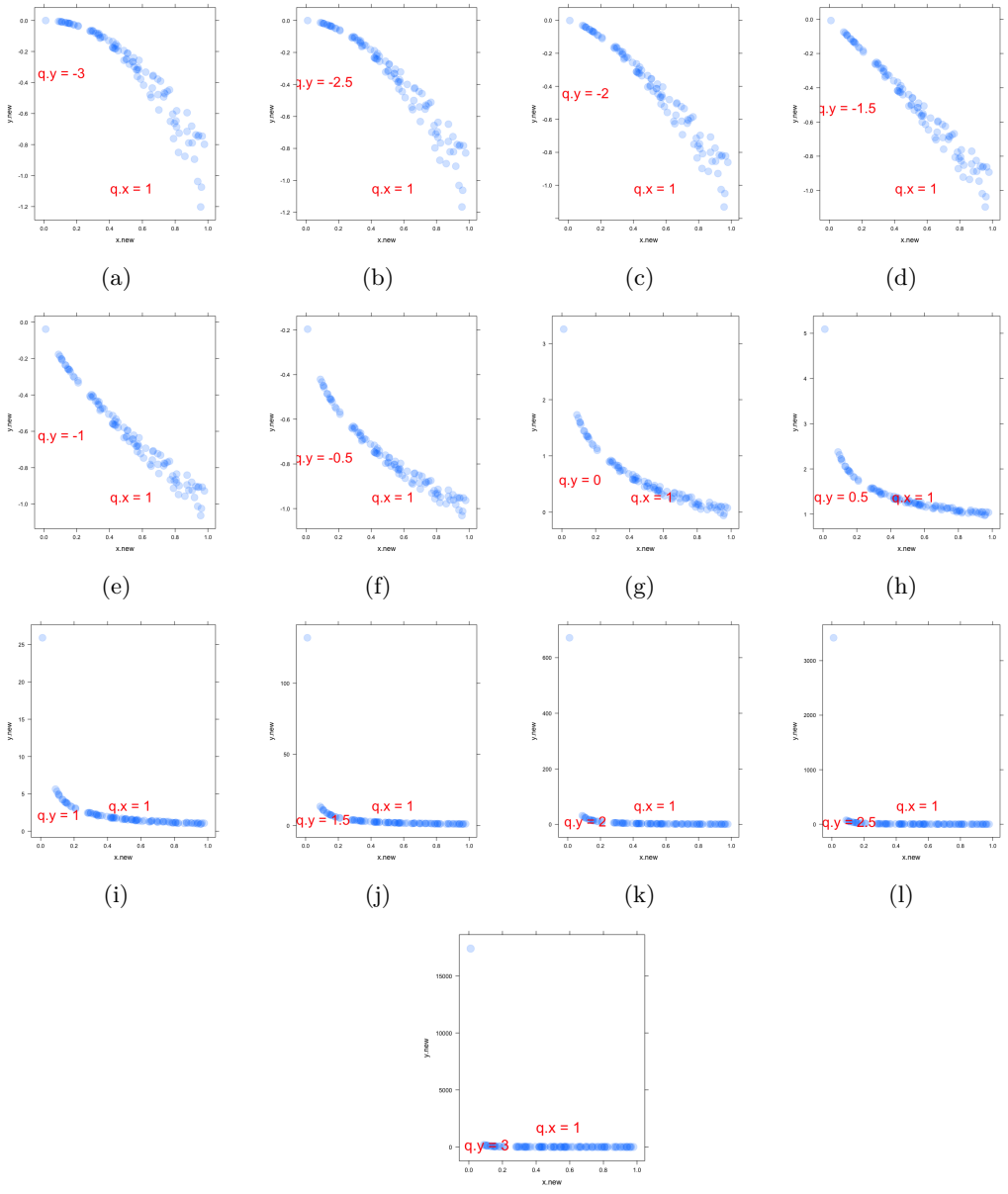
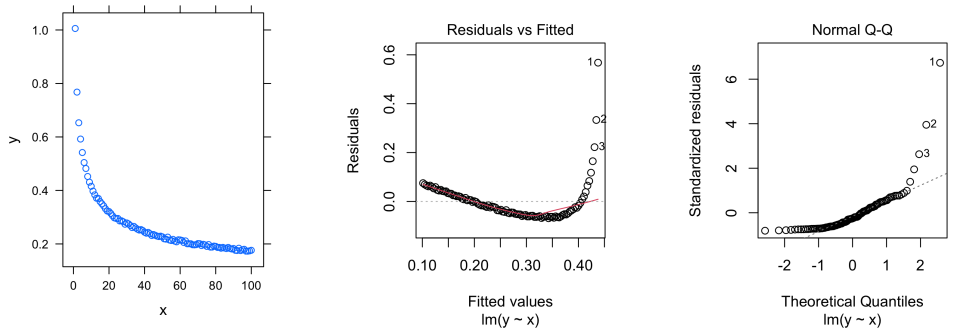


Figure 9: Iterations of the different values of lambda for the Tukey ladder of powers for the logarithmic equation.



(a) (b) (c)

Figure 10: Scatter plot of equation (a), Residuals vs. Fitted plot (b), and Normal Q-Q plot (c) of experiment with different `x`, and `y` with a random variable.

2.5 X with a different distribution.

For this experimentation we use the Equation 5,

$$y = jitter(x^p, factor = length(x)/2) \tag{5}$$

`p` is a standard normal random variable,

`x` is a number of 1:100,

`jitter` returns a numeric value of the same lenght as `x`, but with an amount of noise added in order to break ties.

In Figure 11 we apreciate that from all the iterations, Figure 11c is the one showing linear behaviour, with a λ = -2.

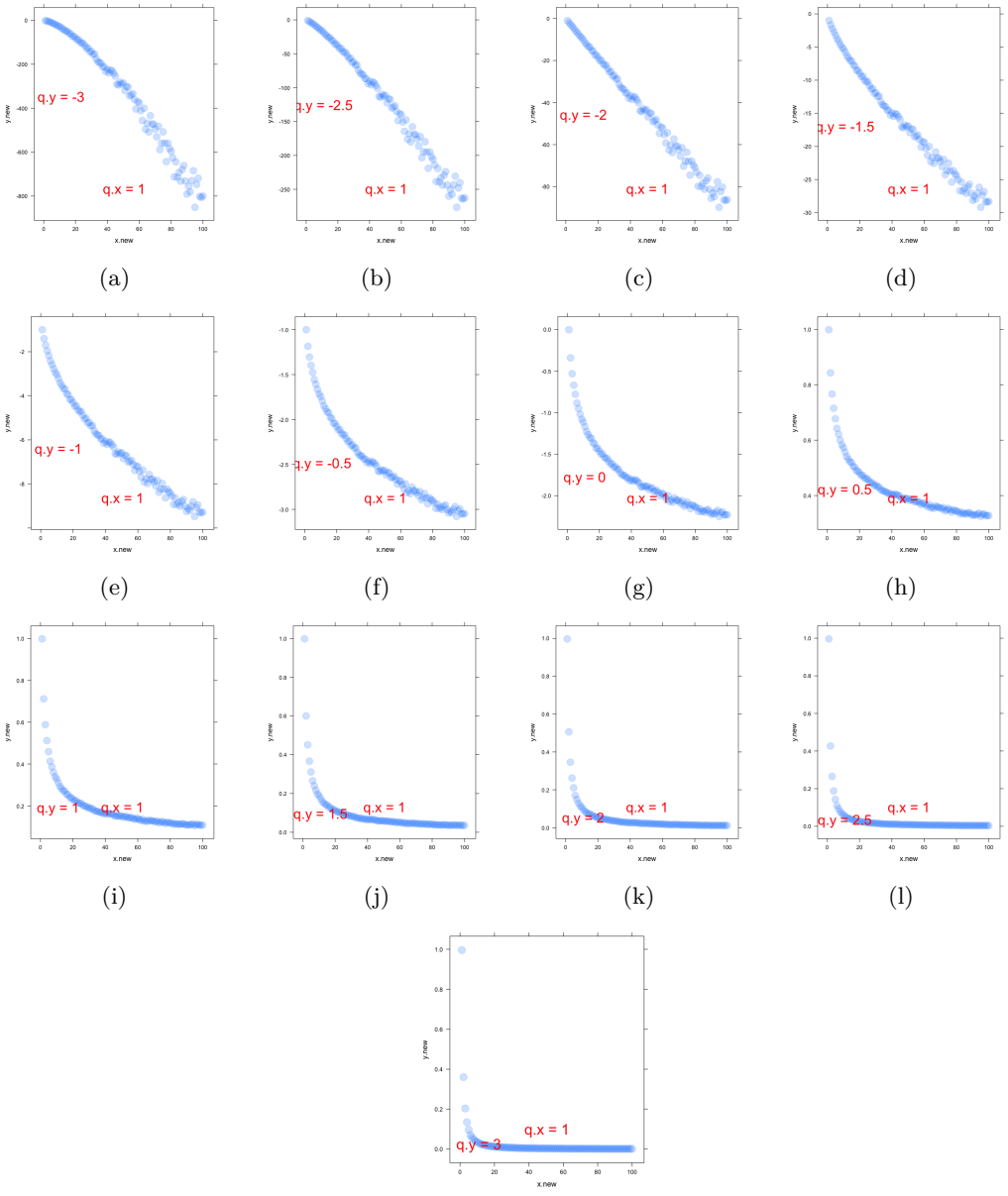
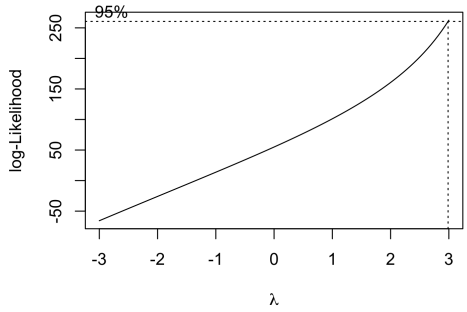
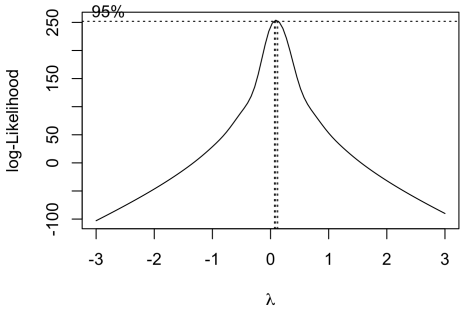


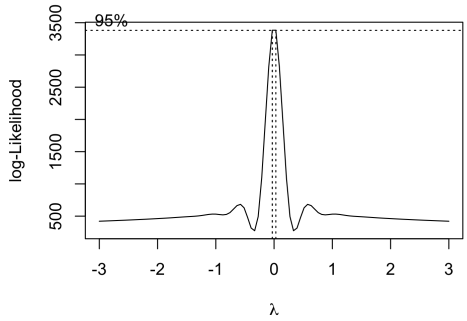
Figure 11: Iterations of the different values of lambda for the Tukey ladder of powers for the equation with different `x` distribution, and `y` with added noise.



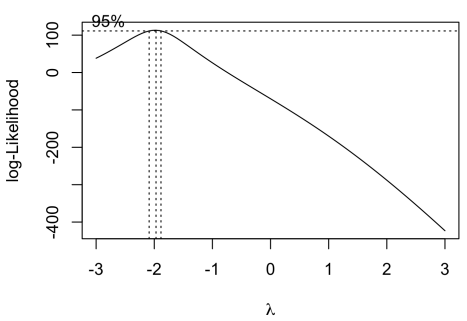
(a) $y = a * sqrt(x) + b$



(b) $y = a * x^2 + b * x + c$



(c) $y = a * (exp(p * x))$



(d) $y = jitter(x^p, factor = length(x)/2)$

Figure 12: Correlations with the different values of λ.

Figure 12 shows the correlations of the different values of lambda working with the different equations used for this work.

3 Conclusions

Starting this work was a bit complicated for me, because I did not fully understand the transformations, but seeing them as an example first helped me understand what it was happening on each iteration. I had to use a built in library to work with the Tukey ladder, and did not use the box-cox transformation, but I think I could grasp the concept of either one a bit better having finished this work.

References

[1] Correlation test between two variables in r. http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r. Accessed: 2020-10-20.

[2] David M. Lane. *Online Statistics Education: An Interactive Multimedia Course of Study*. Rice University.