Homework Assignment 7: Applied Probabilistic Models

Curve Fitting

5273

1 Introduction

Fitting a distribution from a dataset consists of finding the parameters' value, which, with greater probability, that distribution could have generated the observed data [7]. For example, the normal distribution has two parameters (mean and variance); once these two parameters are known, the entire distribution is known.

For the analysis, the R software is used in its version 4.0.2 [8], and the code used is available on the GitHub repository of [5]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Data

For this work, four functions have been created. Independent variables x_1, x_2 are the result of generated values using the R function `runif()`, x_3 is generated by the function `rchisq()` with two degrees of freedom, and x_4, using `rnorm` with its defuault values of mean 0 and standard deviation of 1.

$$y = 4x_1 + 5x_2, \tag{1}$$

$$y = (x_1)^2 + \texttt{rexp(1)}, \tag{2}$$

$$y = e^2(x_1)^2(x_2)^3, \tag{3}$$

$$y = (4x_1)^3(20x_2) + \frac{x_3}{5} + \log(x_4)^2. \tag{4}$$

In real case scenarios, the relation of the variables is unknown most of the time, meaning that a model that best describes this relation has to be found.

3 Experiments

The experiments for this work are based on the generated functions described in the previous section. The parameter used is the number of repetitions or sample size (n), with a value of 100. The previous functions are used in the transformations described in this report and analyzed one of them in each section of the experiments.

3.1 Multiple Linear Regression

For the Function 1, and assuming that the relationship between variables is unknown, it is considered to find the parameters or coefficient that best described this relation. In this case, it could be useful to perform a regression analysis.

data.txt

```
Call:
lmformula = y ~ x1 + x2, data = df1

Residuals:
       Min         1Q     Median         3Q        Max
-9.265e-15 -7.650e-17  9.510e-17  2.450e-16  6.130e-15

Coefficients:
             Estimate Std. Error   t value Pr>t
Intercept 1.421e-15  3.546e-16 4.008e+00  0.00012 ***

x1          4.000e+00  4.560e-16 8.771e+15  < 2e-16 ***

x2          5.000e+00  4.619e-16 1.083e+16  < 2e-16 ***

---

Residual standard error: 1.305e-15 on 97 degrees of freedom
Multiple R-squared:      1,        Adjusted R-squared:      1
F-statistic: 8.979e+31 on 2 and 97 DF,  p-value: < 2.2e-16
```

Regression in R software can be performed with the `lm()` function, which can be obtained the coefficients that best described a relation given by the form $Y = \beta_1 + \beta_2 X + \epsilon$ [6]. In this case, lets assume that it is known that function y depends on x_1 and x_2. For this reason, a multiple linear regression analysis is needed. This is an extension of simple linear regression used to predict an outcome variable (y) based on multiple distinct predictor variables (x). With two predictor variables (x), the prediction of y is expressed by the equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ [2]. Once the experiment is performed it throws in the column "Estimate" of the R output these β coefficients. In this case $\beta_1 = 4$ and $\beta_2 = 5$ which corresponds to the values previously declared when generating the function. The intercept value correspond with an error when estimate this relation.

3.2 Box-Cox Transformation

In some cases, if assumptions of the simplicity of structure for $E(y)$), the constancy of error variance and normality of distributions are not satisfied in terms of the original observations, a non-linear transformation of y may improve the analysis [1]. Box-Cox transformation is given by the equation:

$$y^{\lambda} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & if \quad \lambda \neq 0, \\ \\ \log y & if \quad \lambda = 0. \end{cases} \tag{5}$$

Figure 1 show plots of the diagnosis of function 2, fitting a model without doing a transformation first. The "Residual vs Fitted" plot shows a more concentrate residual points between 0 and -1 lines.

The "Quantile-Quantile" plot shows the normality of the errors, and it is not great specially at the upper tail, which goes off from the straight line. The "Scale-Location" plot, which shows there is no homoscedasticity and the "Residuals vs. Leverage" plot which shows if there are outliers in the residuals.

Then, Figure 2 shows a Maximum Likelihood plot for the values of λ with a 95% confidence interval. This plot gives a threshold with bounds of the recommended λ values for the Box-Cox transformation. Then the best value can be extracted, which represents the value when the curve is higher. This threshold can be seen by simple inspection; it moves from 0.1 to 0.5 approximately. Finally, the calculated value of λ is 0.263.

Figure 2 represents the same model diagnosis, but after performing the Box-Cox transformation with the previously calculated parameter, λ=0.263. Here main changes can be appreciated in the "Quantile-Quantile" Plot, which shows a better approximation to the normal line.

3.3 Log Transformation

Log transformations are another method that can be valuable for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics [4]. This log transformation it is used in Function 3, which depends of the variables x_1 and x_2.

data.txt

```
Intercept     logx1     logx2
  1.4739959   0.3864703   0.4155402
```

In the R output show for this section, the logarithm's coefficient for each independent variable and the intercept's value can be seen.

3.4 Tukey's Ladder of Power

Tukey [9] describes an orderly way of re-expressing variables using a power transformation suggesting exploring simple relationships such as $y^{\lambda} = b_0 + b_1 X$ where λ is a parameter chosen to make the relationship as close to a straight line as possible. Table 1 shows examples of the Tukey's ladder of transformations.

data.txt

```
    lambda      W Shapiro.p.value
428  0.675 0.9584        0.003112

if lambda >  0{TRANS = x ^ lambda}
if lambda == 0{TRANS = logx}
if lambda <  0{TRANS = -1 * x ^ lambda}
```

In the above output corresponding to the `transformTukey()` function of R, applied to Function 2 it shows the value of λ which data could be transformed, which is $\lambda = 0.675$ and the transformation should be x^{λ}. Figure 4 shows a histogram of the transformation.

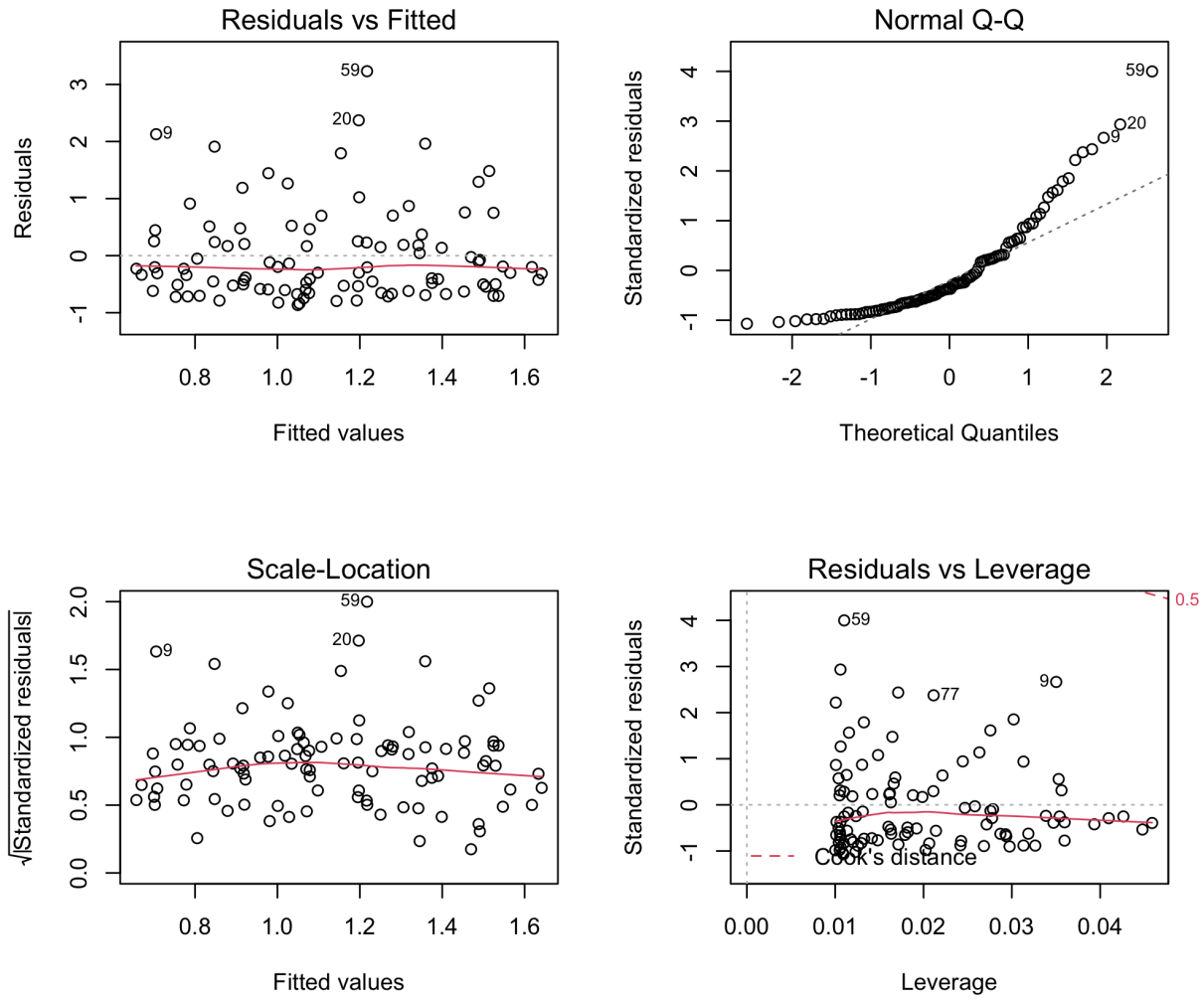

Figure 1: Plots of the Function 2 without transformations

Table 1: Tukey's Ladder of Transformations

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

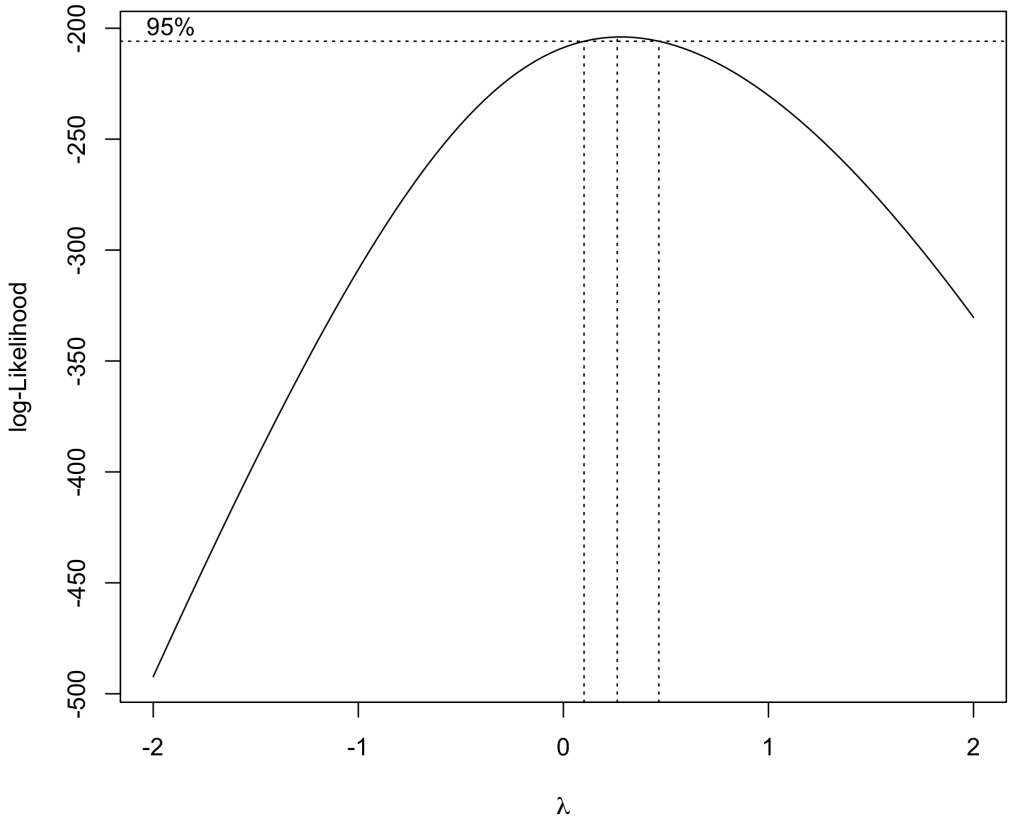

Figure 2: Max-Likelihood plot to determine the best λ for the Box-Cox Transformation

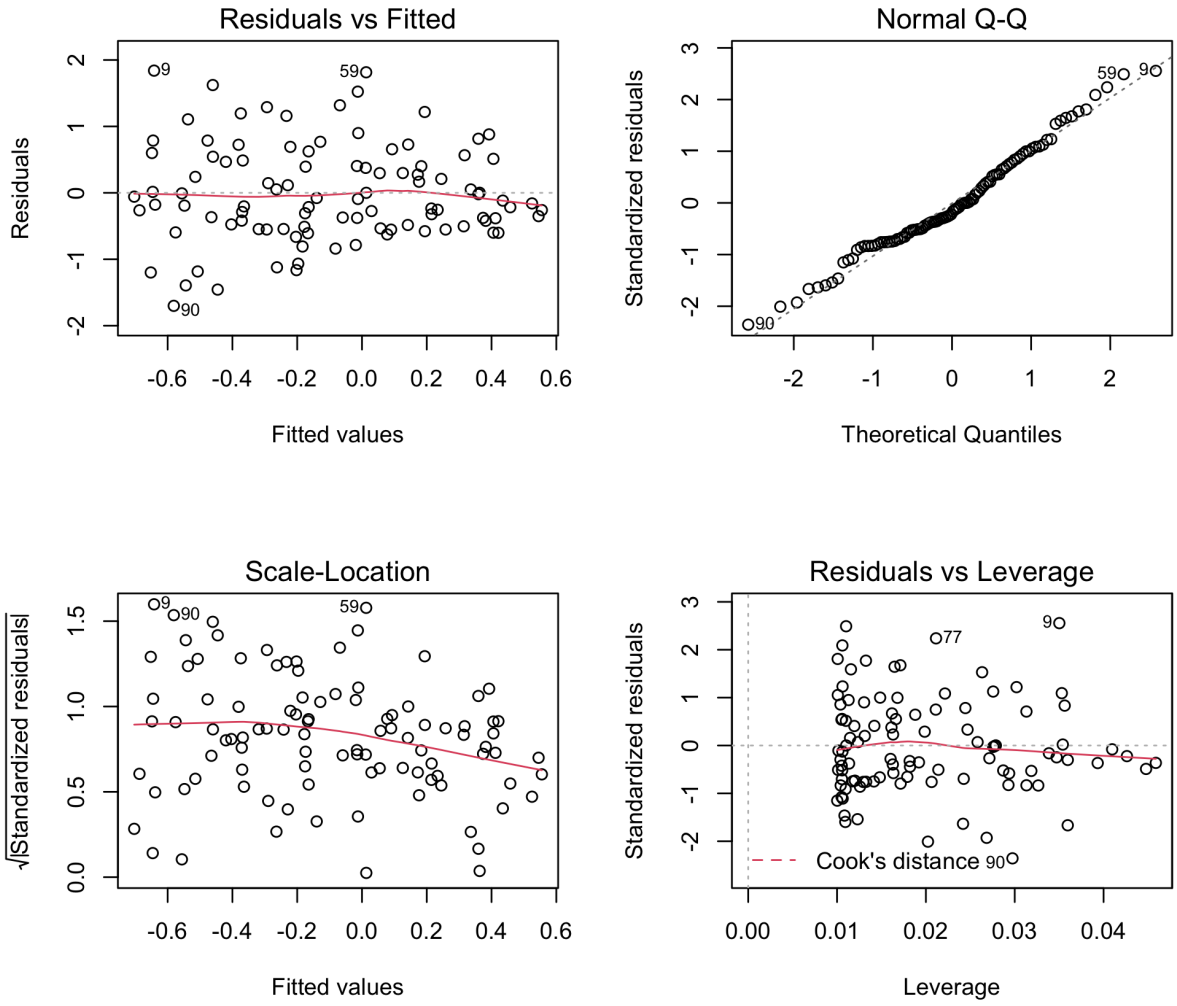

Figure 3: Plots of the Function 2 using the Box-Cox Transformation

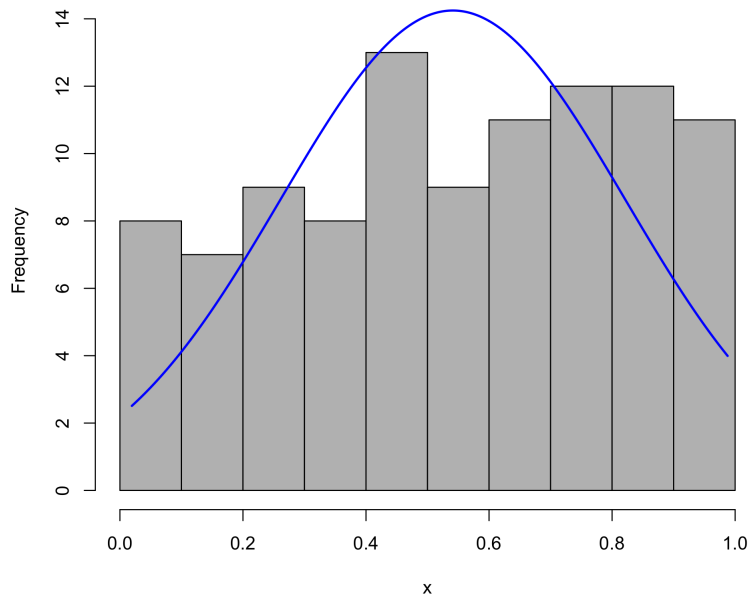

Figure 4: Histogram of Function 2 using the Tukey Ladder of Power Transformation

3.5 Stepwise Regression

The stepwise regression (or stepwise selection) consists of iteratively adding and removing predictors in the predictive model to find the subset of variables in the data set, resulting in the best performing model, that is, a model that lowers prediction error [3].

Here a comparison of the resulting fit model of multiple linear regression with the stepwise regression is performed. The R output showed below shows the results of fitting data applying multiple linear regression, indicating the intercept and values of the coefficient of all x variables.

data.txt

```
Call:
lmformula = y ~ ., data = df5

Residuals:
    Min      1Q  Median      3Q     Max
-15.162  -5.996  -2.152   3.966  34.388

Coefficients:
            Estimate Std. Error t value Pr>t
Intercept -16.9595     2.2949  -7.390 5.67e-11 ***

x1           31.8100     2.8789  11.049  < 2e-16 ***
x2           20.0134     2.8394   7.049 2.87e-10 ***
x3           -0.2517     0.4446  -0.566    0.573
x4            0.3228     0.8442   0.382    0.703
---
```

Table 2: Resulting Models from the two methods for Function 4

Multiple Linear Regression Model	**Stepwise Regression Model**
$31.81x_1 + 20.02x_2 - 0.25x_3 + 0.32x_4 - 16.96$	$43.49x_1 + 26.48x_2 - 27.09$

```
Residual standard error: 8.5 on 95 degrees of freedom
Multiple R-squared:  0.6305,        Adjusted R-squared:  0.6149
F-statistic: 40.52 on 4 and 95 DF,  p-value: < 2.2e-16
```

This last R output shows the final model results by performing the stepwise regression, where it can be seen it removes variables x_3 and x_4 from the initial model to predict y, in contrast with the multiple linear regression. Finally, it can be expressed the final models resulting form this two methods in Table 2.

data.txt

```
Subset selection object
4 Variables  and intercept
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
1 subsets of each size up to 2
Selection Algorithm: backward
         x1  x2  x3  x4
1   1  "*" " " " " " "
2   1  "*" "*" " " " "
Intercept          x1          x2
  -27.09467    43.49111    26.48317
```

References

[1] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[2] Alboukadel Kassambara. Multiple linear regression in R, 2018. `http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/`, Last accessed on 2020-10-17.

[3] Alboukadel Kassambara. Stepwise regression essentials in R, 2018. `http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/`, Last accessed on 2020-10-17.

[4] David Lane, David Scott, Mikki Hebl, Rudy Guerra, Dan Osherson, and Heidi Zimmer. *Introduction to statistics*. David Lane, 2003.

[5] Oscar Alejandro Hernadez Lopez. Probability in R. `https://github.com/oscaralejandro1907/probability-in-R/blob/master/assignment1/t1.R`, 2020.

[6] Selva Prabhakaran. Linear regression, 2016. `http://r-statistics.co/Linear-Regression.html`, Last accessed on 2020-10-17.

[7] Joaquín Amat Rodrigo. Ajuste de distribuciones con R, 2020. `https://www.cienciadedatos.net/documentos/55_ajuste_distribuciones_con_r.html`, Last accessed on 2020-10-15.

[8] The R Foundation. The R Project for Statistical Computing. `https://www.r-project.org/`, 2020.

[9] John W Tukey. *Exploratory data analysis*, volume 2. Reading, MA, 1977.