Curve fitting

G. Palafox

October 20, 2020

Abstract

Various techniques of fitting curves to data are explored. First, computer generated numbers are used to try the models. Afterwards, real data is used.

1 Introduction

It is a common occurrence when handling data to have only observations of phenomena and not an exact relationship between the variables observed. In order to better understand the subject of study at hand, or to make predictions, it is useful to try and fit a curve to the data observed. In this work, performed on a Jupyter notebook [2] with R version 4.0.0 [5], some techniques for fitting curves to data are employed[1]. First, on computer-generated numbers, and then on real data of vehicles in circulation in Mexico, obtained from INEGI's website [1].

2 Curve fitting

The techniques employed here can be found on Navarro's [4] online book, or in the work of Lane et. al. [3]. To begin this work, two hundred numbers between 100 and 500 are generated uniformly, which are taken as the independent x values. Then, different y values dependent on x are generated, to which Gaussian noise N is added. A fragment of these data can be seen in Table 1, while graphics can be seen in Figure 1. An R function `choose_lambda` was created to compute a λ such that the correlation coefficient of x and $\tilde{y}_\lambda$, where

$$\tilde{y}_\lambda := \begin{cases} y^\lambda, & \text{if } \lambda > 0; \\ \log y, & \text{if } \lambda = 0; \\ -(y^\lambda), & \text{if } \lambda < 0, \end{cases} \tag{1}$$

is maximized. This function is shown in Listing 1.

Listing 1: Function for choosing λ in a Tukey transformation.

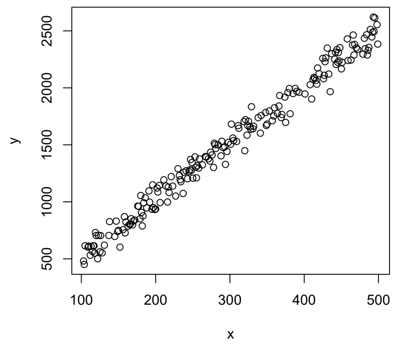
```
choose_lambda <- function(x,y){
if (min(y) < 0){
print("Error. Negative values")
return(NaN)
}

cors <- numeric()
lambdas <- seq(-10, 10, .01)

for (i in lambdas){
if (i == 0)
cors <- c(cors, cor(x, log(y)))
else if (i > 0)
cors <- c(cors, cor(x, y**i))
else
cors <- c(cors, cor(x, -(y**i)))
}

return (lambdas[which.max(cors)])
}
```

For each of the y values generated (see Table 1), the `choose_lambda` function is used to obtain a λ value, and the y values are transformed. Then, using R's `lm` function, a linear regression is performed on the $x, \tilde{y}_\lambda$ values, which gives a linear function $\tilde{y}_\lambda = ax + b$. Finally, an inverse transformation is applied to get a function $y = f(x)$ fitting the original values. The results of this process are plotted in Figure 2.

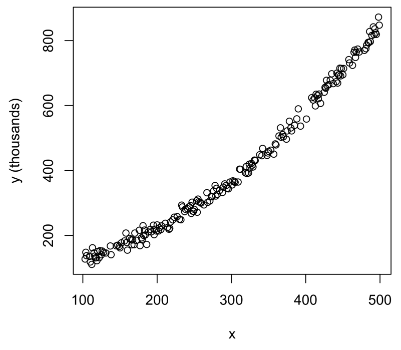
[1]The notebook with the code containing our analysis, as well as this report, can be found in the Github Repository: `https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment7`

Table 1: Fragment of the data generated.

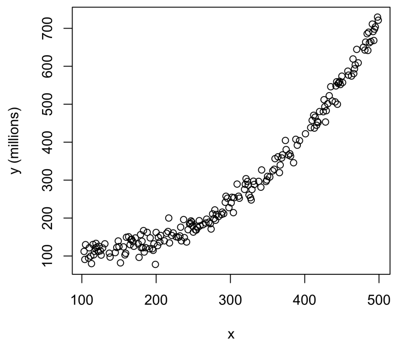
x	$y = 5x + 4 + N$	$y = 3x^2 + 50 + N$	$y = 5x^3 + .4x^2 + 1 + N$	$y = .8\log(x) + 8 + N$	$y = .7\sqrt{x} + 14 + N$
103.00	481.10	127,372.77	112,075,025.75	11.78	21.61
104.00	451.14	148,056.62	90,963,451.34	11.84	23.43
105.00	614.70	137,575.08	129,960,611.26	11.73	23.80
109.00	601.88	135,537.58	94,598,979.05	11.78	19.64
110.00	611.29	118,270.35	121,634,806.71	11.69	21.51
112.00	531.72	110,758.23	99,643,078.53	11.75	21.99



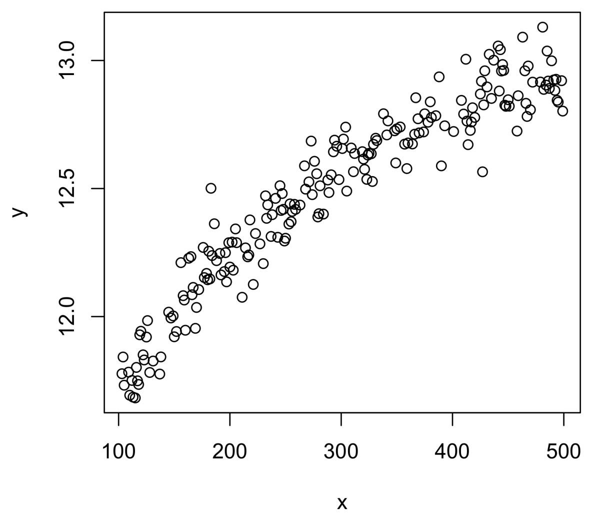
(a) $y = 5x + 4 + N$.



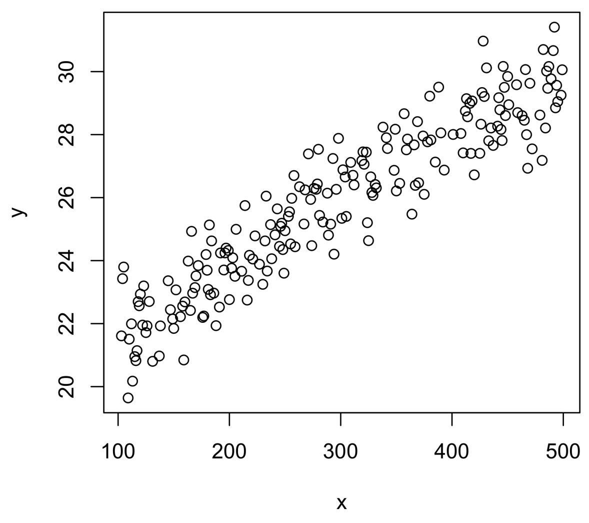
(b) $y = 3x^2 + 50 + N$.



(c) $y = 5x^3 + .4x^2 + 1 + N$.

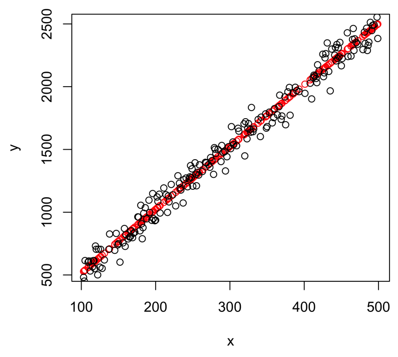


(d) $y = .8\log(x) + 8 + N$.

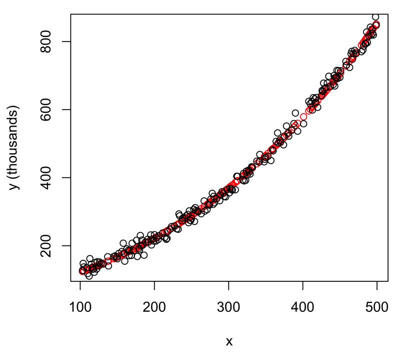


(e) $y = .7\sqrt{x} + 14 + N$.

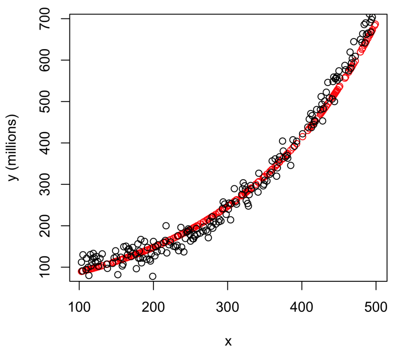
Figure 1: Plots of the data generated.



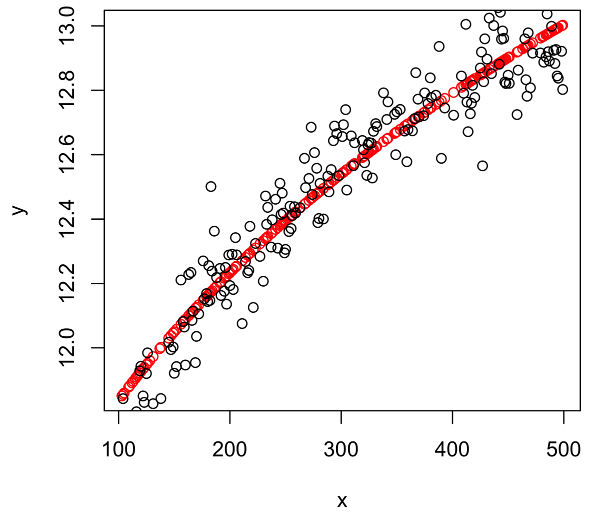
(a) In black, $y = 5x + 4 + N$, and in red, a curve fitted with $\lambda = 1.04$.



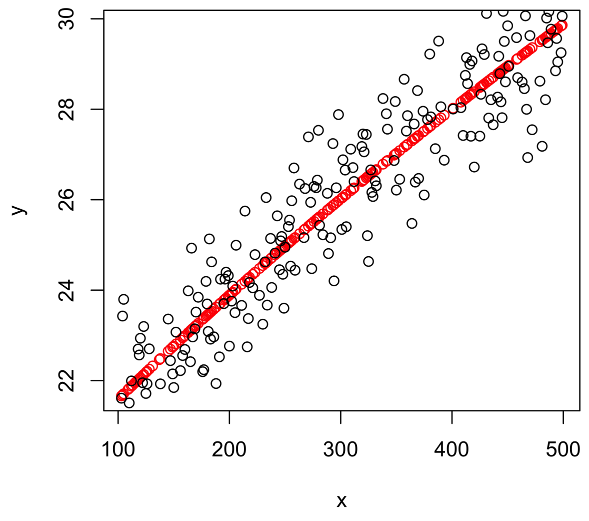
(b) In black, $y = 3x^2 + 50 + N$, and in red, a curve fitted with $\lambda = 0.29$.



(c) In black, $y = 5x^3 + .4x^2 + 1 + N$, and in red, a curve fitted with $\lambda = 0$.



(d) In black, $y = .8\log(x) + 8 + N$, and in red, a curve fitted with $\lambda = 10$.



(e) In black, $y = .7\sqrt{x} + 14 + N$, and in red, a curve fitted with $\lambda = 1.85$.

Figure 2: Data and curves fitted.

Table 2: Vehicles in circulation in Mexico.

	year	vehicles
1	1981	6,339,836
2	1982	6,695,164
3	1983	6,941,252
4	1984	7,305,066
5	1985	7,725,623
6	1986	7,732,012

The capabilities of R's `lm` function extends to multilinear regression. In order to exemplify this, three different sets of independent variables were created, and a variable dependent on these three was computed. The code where this is done, and its results, are shown in Listing 2.

Listing 2: Multilinear regression with `lm`.

```
x1 <- sample(x = 100:500, size = 200, replace = FALSE)

x2 <- sample(x = 200:600, size = 200, replace = FALSE)

x3 <- sample(x = 100:500, size = 200, replace = FALSE)

my2 <- 3*x1 + .5*log(x2) + 5*x3

lm(my2 ~ x1 + log(x2) + x3)

#Call:
#lm(formula = my2 ~ x1 + log(x2) + x3)

#Coefficients:
#(Intercept)           x1        log(x2)            x3
#2.187e-12     3.000e+00     5.000e-01     5.000e+00
```

2.1 Vehicles in circulation in Mexico

In this subsection, a curve is fitted to real data. The number of vehicles in circulation in Mexico per year, from 1986 to 2019, is downloaded from INEGI's webiste [1]. A fragment of the data can be seen in Table 2. The data is also shown in Figure 3a. Two distinct approaches are taken here. The first approach consists of using the function `choose_lambda`, which gives a value λ to transform the data as was done in Section 2. A value of $\lambda = -0.201$ is obtained. As before, a linear model is fitted for $(x, \tilde{y}_\lambda)$. The function obtained with this method is $y = (0.8393 - 0.0004x)^{-1/0.201}$, and it can be seen in Figure 3b. The second approach consists of assuming the number of vehicles in circulation has exponential growth, and fitting a curve with `lm(log (y) ∼ x)`. This gives a model $y = \exp(0.0568 - 97.1138x)$, and can be seen in Figure 3c. This second model is also used to plot a 99% confidence interval for the data, as seen in Figure 3d.

3 Conclusion

The theory and methods of curve fitting is larger than what was presented here. Many other techniques can be applied and studied further, for example, fitting a model on real data with more than one independent variable. Models including non-uniform independent variables can also be studied.

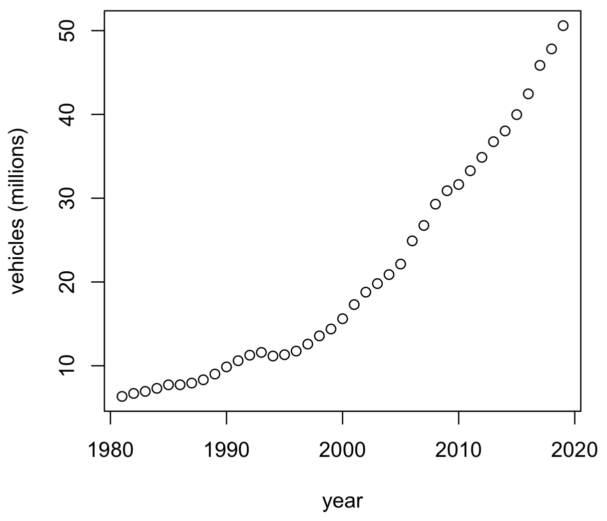
References

[1] Instituto Nacional de Estadística y Geografía, *Total nacional de vehículos. Vehículos de motor registrados en circulación.* `https://www.inegi.org.mx/temas/vehiculos/`.

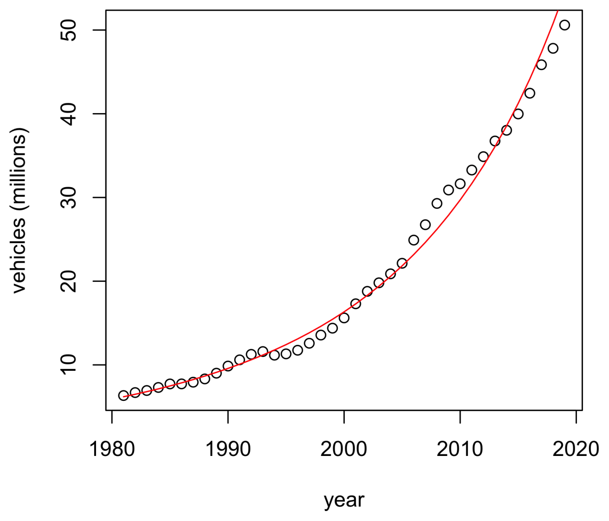
[2] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al., *Jupyter notebooks—a publishing format for reproducible computational workflows*, in Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing, IOS Press, 2016, p. 87.

[3] D. M. Lane, D. Scott, M. Hebl, R. Guerra, D. Osherson, and H. Zimmer, *Introduction to Statistics*, online ed. `http://onlinestatbook.com/Online_Statistics_Education.pdf`.

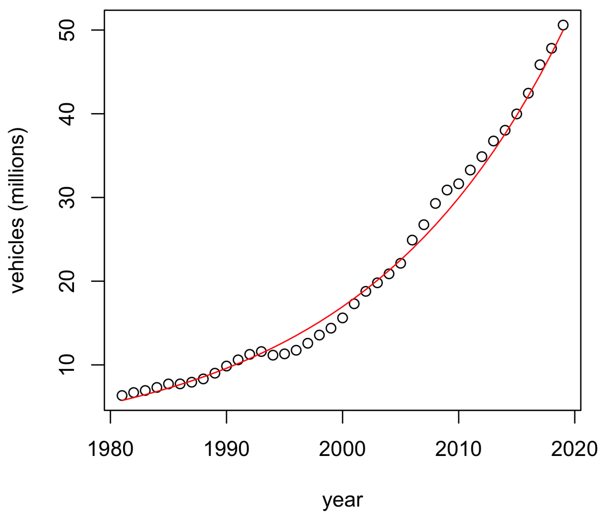
[4] D. Navarro, *Learning statistics with R.* `https://learningstatisticswithr.com/`.



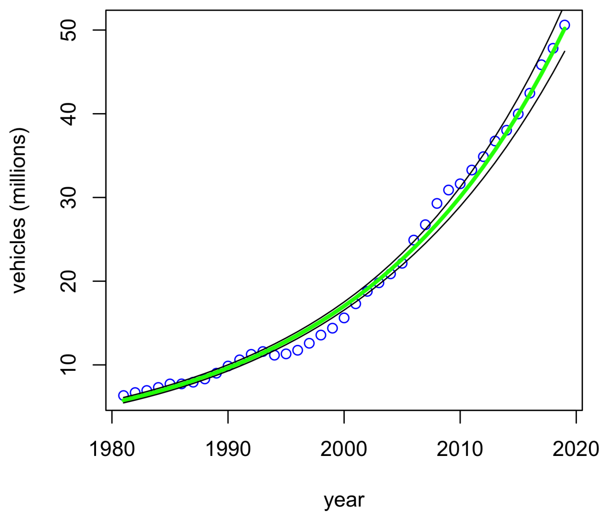
(a) Vehicles in circulation per year.



(b) Vehicle data (black) and $y = (0.8393 - 0.0004x)^{-1/0.201}$ (red).



(c) Vehicle data (black) and $y = \exp(0.0568 - 97.1138x)$ (red).



(d) Vehicle data (blue) and predicted intervals (black and green).

Figure 3: Models for vehicle data.

[5] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.