Statistical Linear Models

Summary: It is often difficult to determine from diagnostic plots which transformation of Y is most appropriate for correcting unequal error variance and nonlinearity. Today we will discuss a procedure which automatically identifies a transformation Y based on the data. We will also discuss improving regression modeling through the use of and

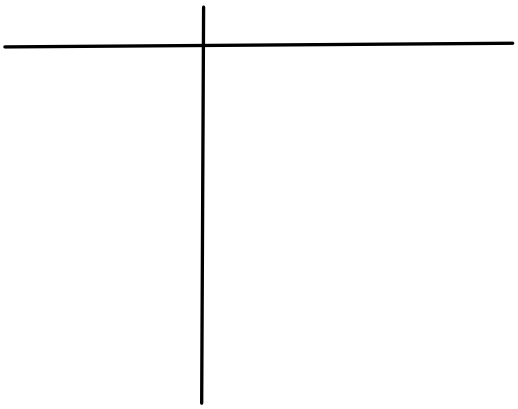
---

Motivation Transformation of the response variable is useful when trying to get a better linear fit of the data or trying to correct violations of model assumptions. We've discussed a variety of ways to transform the data ad hoc ( ). The Box Cox method allows us to more precisely determine the best transformation for the data.

Box Cox Transformations

The Box Cox procedure

This family encompasses the following simple transformations



we solve for λ by finding the maximum likelihood estimates of

Our likelihood function becomes: (for the simple linear regression case)

Maximizing the above with respect to $\beta_0, \beta_1, \sigma^2$, & λ yields the maximum likelihood estimators of these parameters.

Note: Your book defines $g_\lambda(Y)$ as:

Intuitively, what is Box Cox doing?

Why do we define $\log(Y)$ when $\lambda = 0$?

We can rewrite the Box Cox formula as

Note: we can get the same result by using from calculus.

Some considerations when using Box Cox

1.

2.

3.

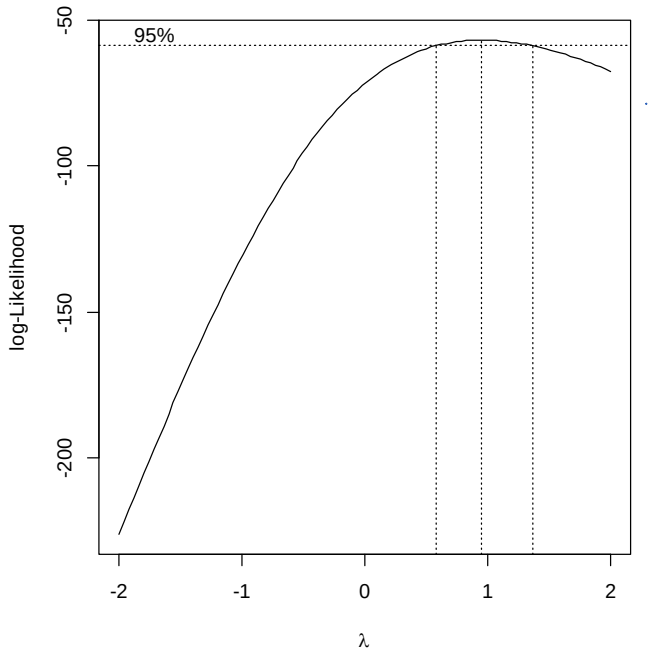
Questions?

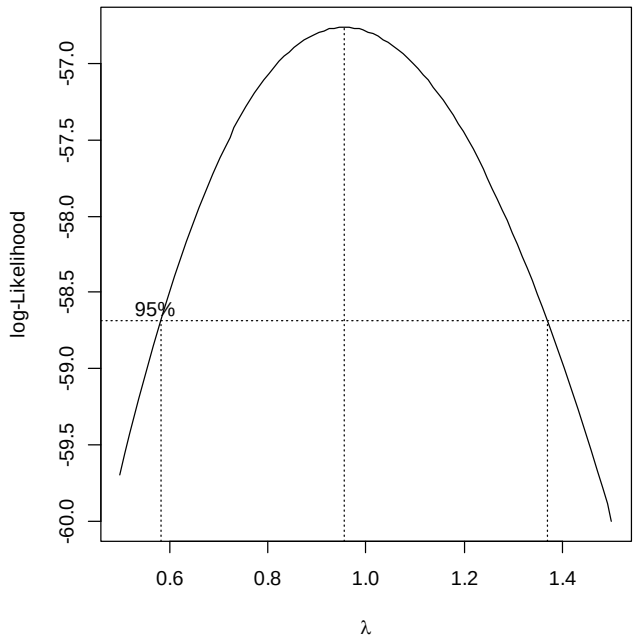
Random Statistics Trivia

George Box and Sir David Cox collaborated on one paper (Box, 1964). The story is that while Cox was visiting Box at Wisconsin, they decided they should write a paper together because of the similarity of their names (and that both are British). In fact, Professor Box is married to the daughter of Sir Ronald Fisher.

Example: Savings Dataset

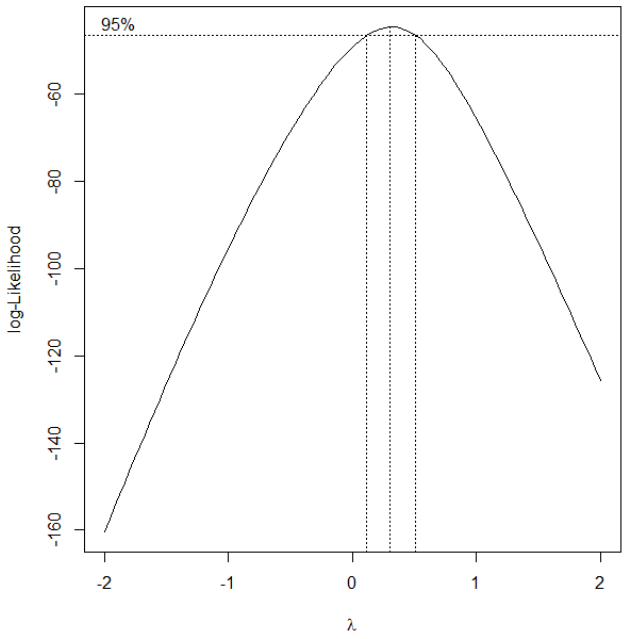
```
g ← lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
boxplot(g, plotit=T)
```

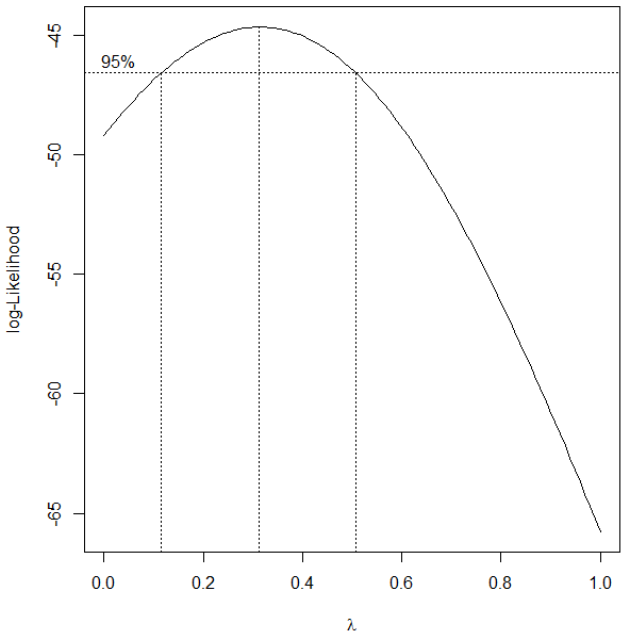




Galápagos Island dataset

```
g ← lm(Species ~ Area + Elevation + Nearest + Scruz
       + Adjacent, data)
```

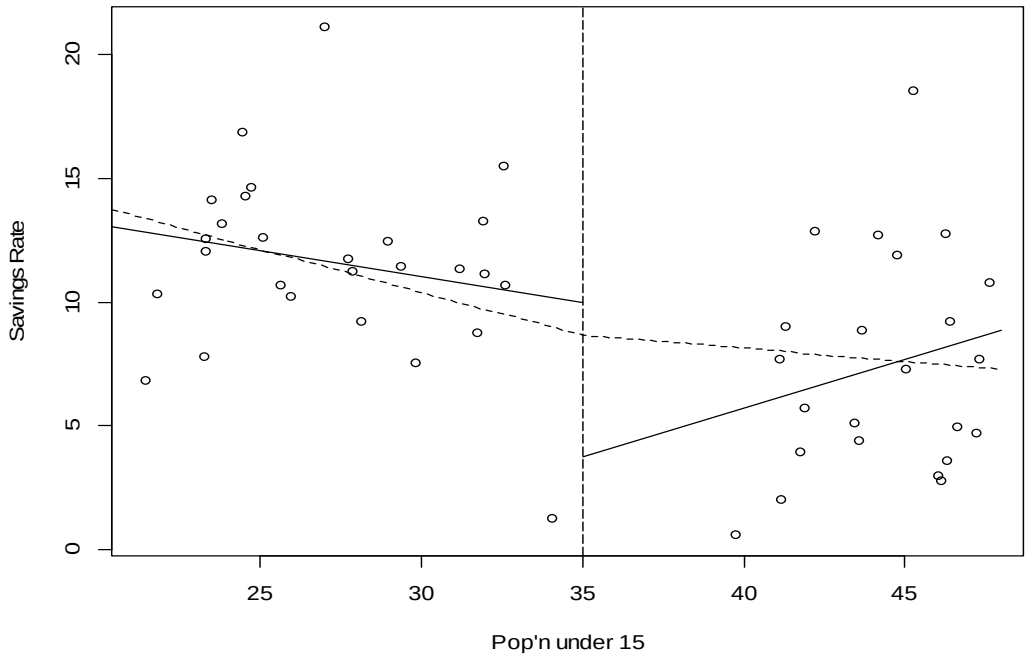




What value of λ would you choose?

Depends on interpretability.

Broken stick regression



Solid line: just fitting two seperate regression functions.

Dotted line: fitting using the basis functions.
(forces the lines to touch)

Why is "Linear Regression" refered to as linear?

Polynomial Regression Models

Among the most frequently used curvilinear response models because of their ease in handling the general linear regression model framework.

2nd order model (quadratic model)

K^{th} order model (in one variable)

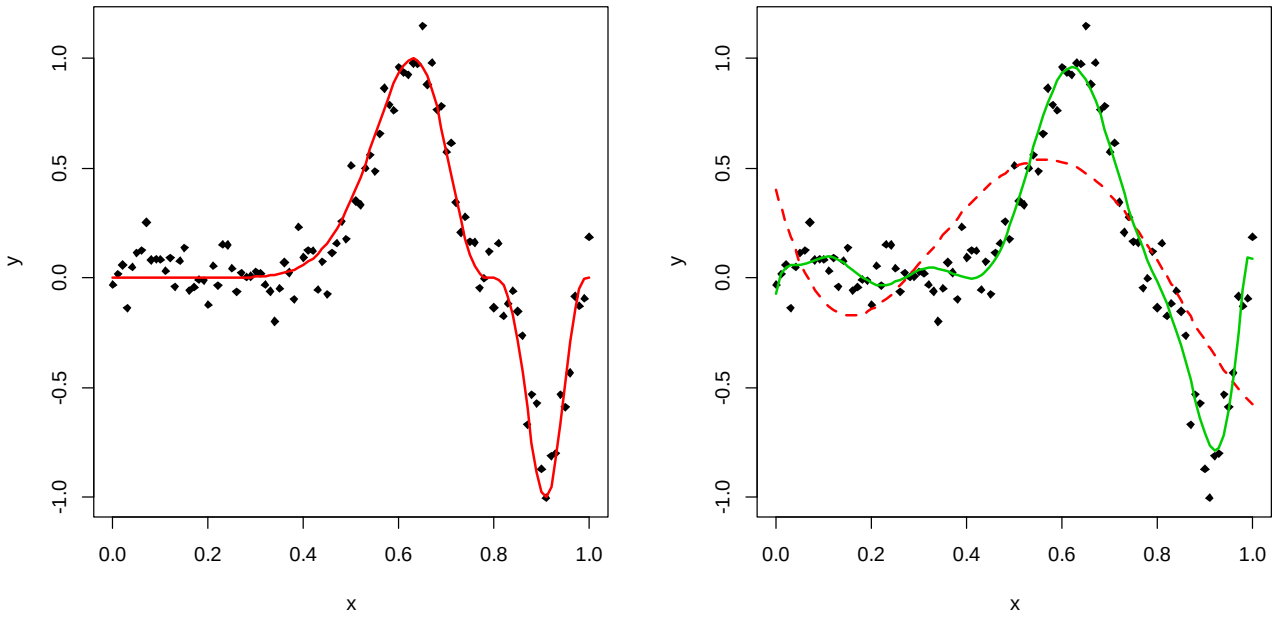
Polynomial regression models may contain one, two or more that two predictor variables and each predictor variable may be present in various powers.

Does anyone see a problem with including $X + X^2$ as predictor variables?

Example: Simulate data from the model

$$Y = \sin^3(2\pi x^3) + \varepsilon \qquad \varepsilon \sim N(0, (0.1)^2)$$

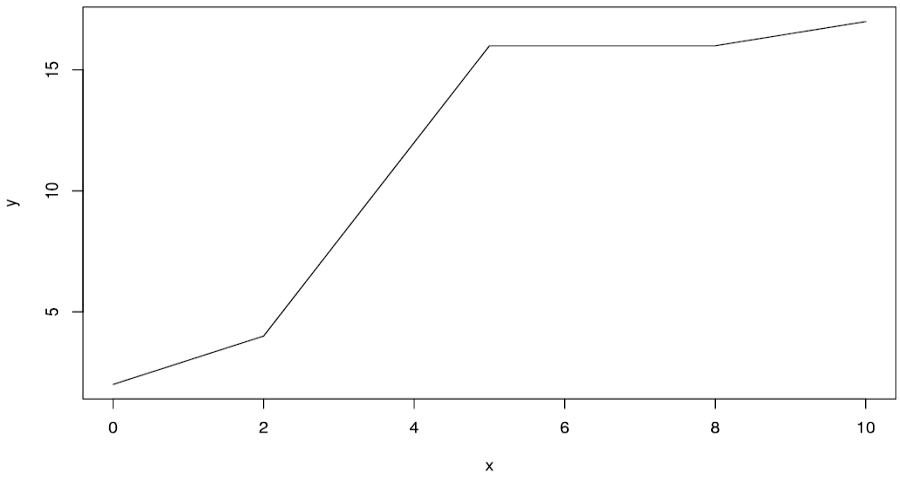
red line – true function
simulated data

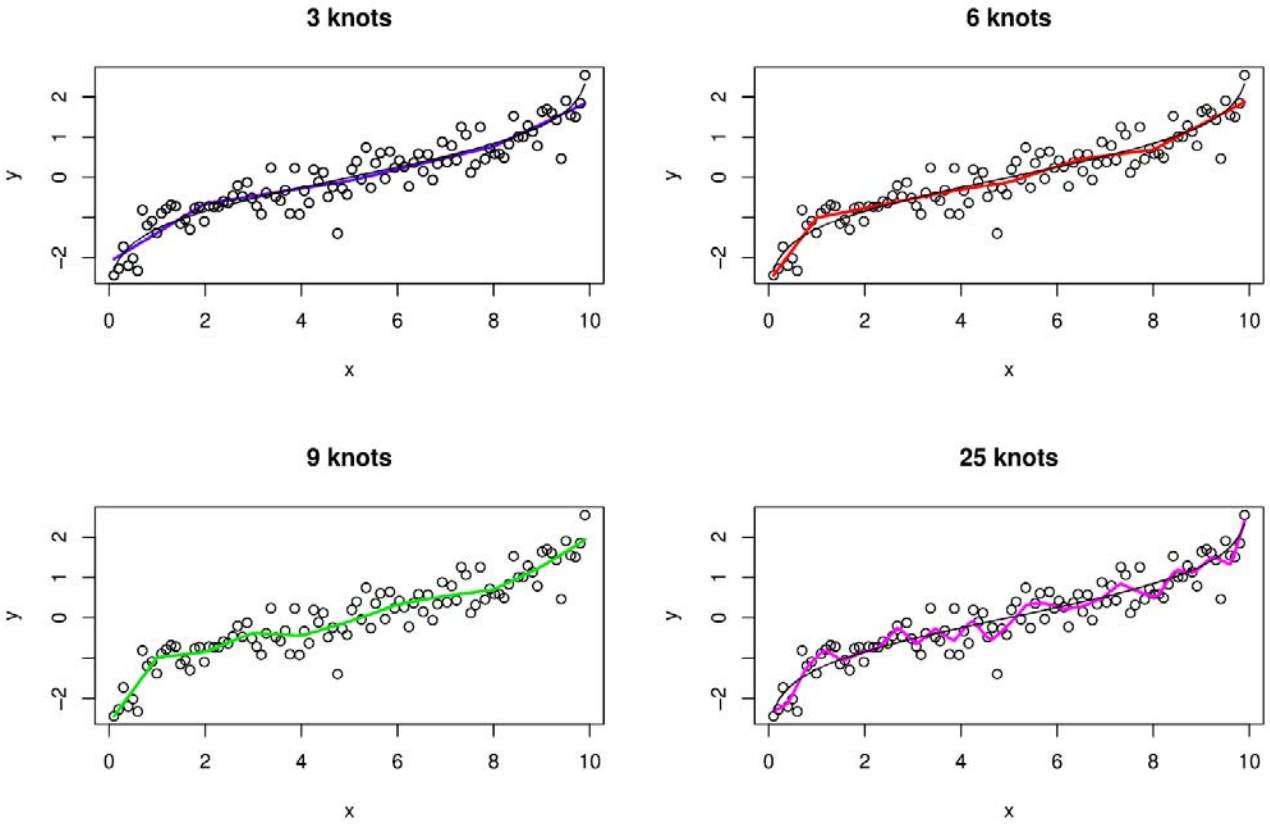


Regression Splines

A low order polynomial may provide a poor fit to the data and increasing the order of the polynomial may not help. One solution is to use

The piecewise linear spline function is given by:





Although linear splines may work well, they are not smooth and will not fit highly curved data well (unless many knots are used, which requires a lot of data)

More common to use

A cubic spline function with K knots:

1.

1.

2.

Given by: