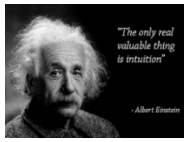
Transformation #3:
Box-Cox

LI(×) – Relationship between Y and X's are not linear or
YN(×) – Y is not normally distributed



Box-Cox: Intuition

- Box-Cox is a **family** of transformations for the response variable y^*
- You get a different transformation y^* for every value of λ in
  - For $\lambda = 0$ → $y^* = Log(y)$
  - For $\lambda \neq 0$ → $y^* = \frac{y^{\lambda}-1}{\lambda}$
- Box-Cox transformations are useful when one is having **difficulties** obtaining a transformed variable with a **normal distribution**.
- The idea is to **systematically** calculate y^* for $\lambda = 0, 1, 2,$ **etc.** and select the transformation that yields the most normally distributed y^*
- Box-Cox transformations are a useful statistical **feature engineering** technique, but the transformed variables are difficult to interpret. However, they may prove to be very useful when **predictive accuracy** is a more important goal than interpretation

Tips

`boxcox(){MASS}` → provides a maximum-likelihood plot showing which value of λ provides the best fit in a linear model

`boxcox(lm.fit)` → provides the maximum-likelihood plot for a wide range of λ's in the linear model `lm.fit` → pick the λ with the highest ML value

`boxcox(lm.fit,lambda=seq(-0.1, 0.1, 0.01))` → if, for example, the highest λ is around `0.04`, get a zoomed in plot around that area → in he example, the function provides a plot between λ `=-0.1` and `0.1` in `0.01` increments.

KOGOD SCHOOL
of
BUSINESS