Transformation and Weighting – Part I

Johns Hopkins Engineering

625.461 Statistical Models and Regression

Module 8 – Lecture 8C

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Impact of “Constance Variance” Assumption

“Constant Variance” is a required assumption such that the SS table in regression analysis and ANOVA table are usable.

If “Constant Variance” assumption is violated, the LS estimate will still be unbiased for the parameter it estimates, but the variance of the LS estimate would need a special care.

An Example: Poisson Distribution

If y is a Poisson random variable in a simple linear regression model, then the variance of y is equal to the mean. Since the mean of y is related to the regressor x, the variance of y will be proportional to x. To reach constance variance, variance-stablizing transformations are often useful, e.g.,

regress $y' = y^{1/2}$ on x

An Example: Bernoulli Distribution

If the response variable is a proportion ($0 \le y_i \le 1$) and the plot of the residuals versus the fitted value of y_i has the double-bow pattern (Figure 4.5c, p.140), the arcsin transformation

$$y' = \sin^{-1}(\sqrt{y})$$

is appropriate.

Variance-Stablizing Transformations

TABLE 5.1 Useful Variance-Stabilizing Transformations

Relationship of σ^2 to $E(y)$	Transformation
$\sigma^2 \propto$ constant	$y' = y$ (no transformation)
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}$ (square root; Poisson data)
$\sigma^2 \propto E(y)[1 - E(y)]$	$y' = \sin^{-1}(\sqrt{y})$ (arcsin; binomial proportions $0 \le y_i \le 1$)
$\sigma^2 \propto [E(y)]^2$	$y' = \ln(y)$(log)
$\sigma^2 \propto [E(y)]^3$	$y' = y^{-1/2}$ (reciprocal square root)
$\sigma^2 \propto [E(y)]^4$	$y' = y^{-1}$(reciprocal)

Impact of “Nonconstant Variance”

If the nonconstant variance is not corrected, then the LS estimators will still be unbiased but they will no longer have the minimum-variance property. The regression coefficients will have larger standard errors than necessary.

Impact of “Transformation”

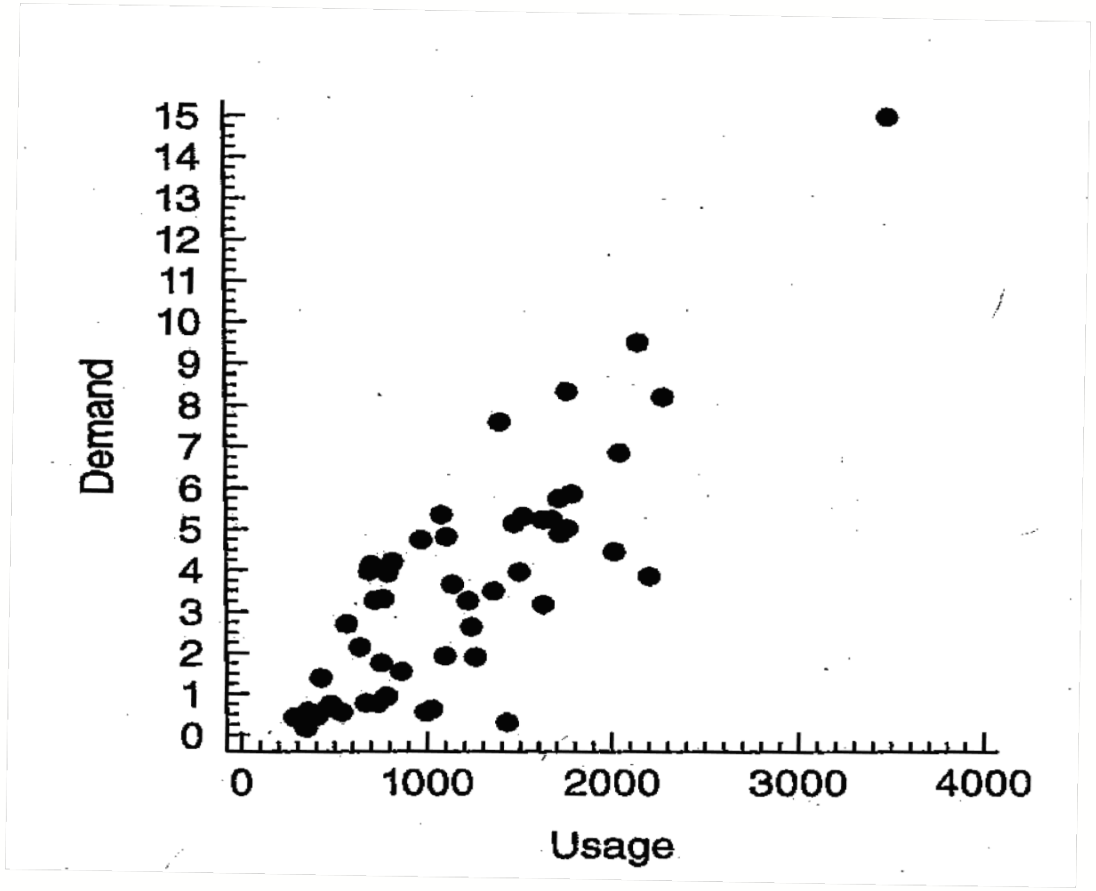
However, after transformation, the predicted value will be in the transformed scale. It is often necessary to convert the predicted values back to the original units. Applying inverse transformation directly to the predicted values will not give an estimate of the mean of the original response variable.

Impact of “Transformation”

Confidence or prediction intervals may be directly converted from one metric to another, since these interval estimates are percentiles of a distribution and percentiles are unaffected by transformation. But, there is no assurance that the resulting intervals in the original units are the shortest possible intervals.

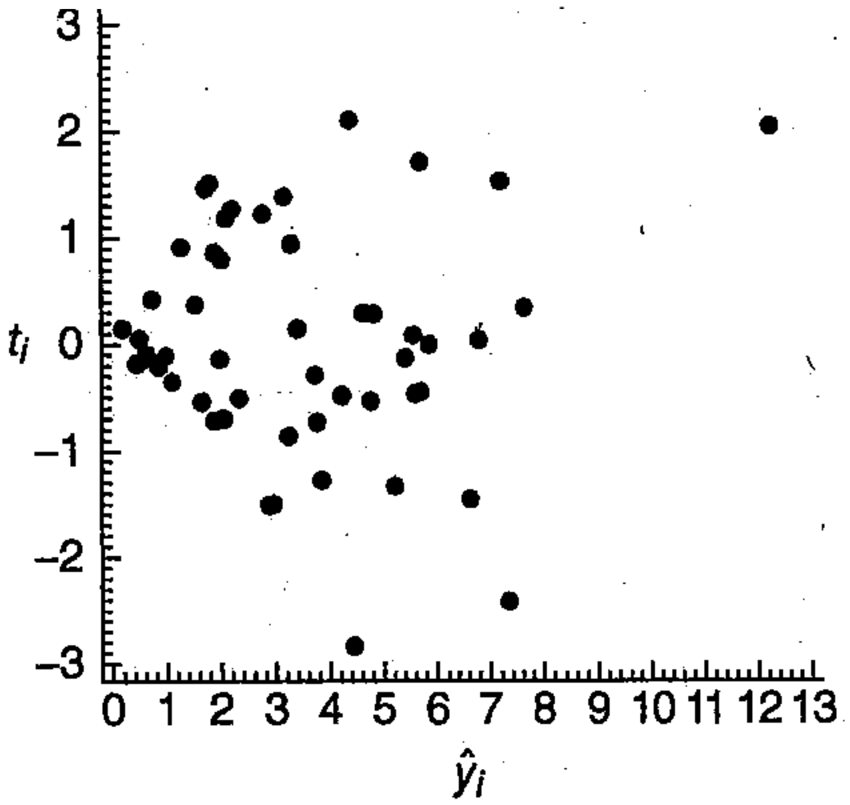
Electricity Utility Data (Ex 5.1, page 173 of Textbook)

Relate peak-hour demand (y) to total energy usage (x), based on Table 5.2 data



Electricity Utility Data (Ex 5.1, page 173 of Textbook)

$$\hat{y} = -0.8313 + 0.00368x$$



Outward-opening funnel

Figure 5.2 Plot of R-student values t_i versus fitted values $\hat{y}_i$, Example 5.1.

Electricity Utility Data (Ex 5.1, page 173 of Textbook)

May view y as a "count" of the number of kilowatts used by a customer during a particular hour. This suggests

$$y^* = \sqrt{y}$$

Electricity Utility Data (Ex 5.1, page 173 of Textbook)

$$\hat{y}^* = 0.5822 + 0.0009529x$$

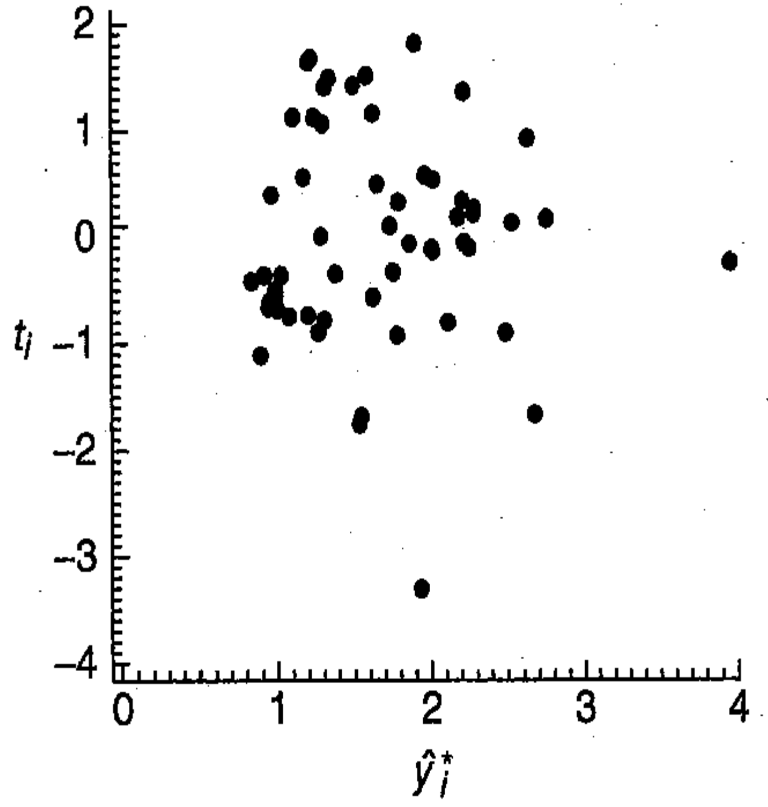


Figure 5.3 Plot of *R*-student values t_i versus fitted values $\hat{y}_i^*$ for the transformed data, Example 5.1.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING