San José State University

Math 261A: Regression Theory & Methods

Transformations and Weighting

Dr. Guangliang Chen

This lecture is based on the following part of the textbook:

- Sections 5.1 – 5.5

Outline of the presentation:

- **Variance-stabilizing transformations** (to deal with non-constant variance)
- **Transformations to linearize the model** (to deal with nonlinearity)
- **Generalized linear regression and weighting** (to deal with any kind of variance)

Introduction

In the last few lectures we introduced various graphical plots and a quantative test to check for different kinds of model inadequacy:

- **Residual plots**: outliers, constant variance, nonlinearity
- **Normal quantile plots (qq-plots)**: normality, outliers
- **Lack of fit test** (when having replicated observations): linearity

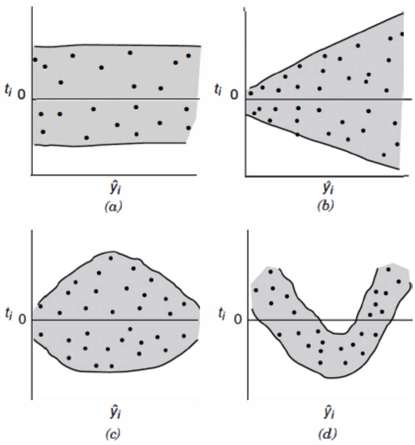
In this part we introduce some corrective procedures for unsatisfied model assumptions.

Variance-stabilizing transformations

When the constant variance assumption is violated as indicated in plots (b)-(d), we can transform the response variable in order to stabilize the variance.

Theorem 0.1. If $\mathrm{Var}(y) = h(\mathrm{E}(y))$ for some function h, then a variance-stabilizing transformation is

$$y' \propto \int \frac{1}{\sqrt{h(y)}}\,\mathrm{d}y$$



Transformations and weighting

For example, for the funnel shape with linear function $h(\mu) = \mu$, i.e.,

$$\mathrm{Var}(y) = h(\mathrm{E}(y)) = \mathrm{E}(y)$$

a variance-stabilizing transformation is

$$y' \propto \int \frac{1}{\sqrt{y}}\,\mathrm{d}y = 2\sqrt{y}$$

Transformations and weighting

Another example is the double bow shape with $h(\mu) = \mu(1-\mu)$, i.e.,

$$\mathrm{Var}(y) = h(\mathrm{E}(y)) = \mathrm{E}(y)(1-\mathrm{E}(y))$$

a variance-stabilizing transformation is

$$y' \propto \int \frac{1}{\sqrt{y(1-y)}}\,\mathrm{d}y = \int \frac{1}{\sqrt{1-y}} 2\,\mathrm{d}(\sqrt{y}) = 2\arcsin(\sqrt{y})$$

Transformations and weighting

Table 1: Common transformations and when to use them

When to use	**Which transformation**	
$\mathrm{Var}(y) \propto 1$	$y' = y$	(do nothing)
$\mathrm{Var}(y) \propto \mathrm{E}(y)$	$y' = \sqrt{y}$	(square root)
$\mathrm{Var}(y) \propto \mathrm{E}(y)(1 - \mathrm{E}(y))$	$y' = \arcsin(\sqrt{y})$	(arcsine)
$\mathrm{Var}(y) \propto \mathrm{E}(y)^2$	$y' = \log(y)$	(log)
$\mathrm{Var}(y) \propto \mathrm{E}(y)^3$	$y' = 1/\sqrt{y}$	(reciprocal sqrt)
$\mathrm{Var}(y) \propto \mathrm{E}(y)^4$	$y' = 1/y$	(reciprocal)

Transformations and weighting

Remark.

- The variance-stabilizing transformations are **empirical** methods, and they only transform the response to have **approximately constant** variance.
- Transforming the response to stabilize the residual variance is a trial-and-error procedure: You apply some of the common transformations to the response, refit the model with the transformed response and re-check the residual plots for each model.
- Weigh complicatedness of the transformation against “prettiness” of the residual plots: The simplest transformation that leads to acceptable residual plots should be preferred.

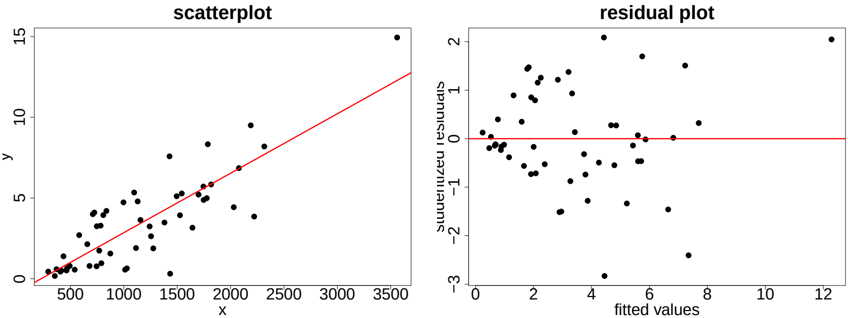
Simulation

Example: The Electric Utility Data

An electric utility company is interested in developing a model that relates peak-hour demand (y in KW) to total energy usage (x, in KWh) during the month. This relationship is important for the company, because they have to plan their generation system for the peak usage, while customers are charged for the total energy they use.

The data set *ElectricUtility.txt* contains observations on 53 residential customers for the month of August.

Transformations and weighting

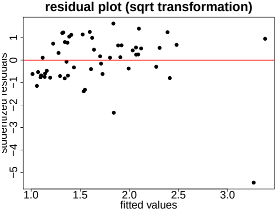


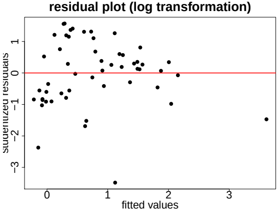
The residual plot shows a clear funnel-shaped pattern that suggests that the residual variance increases with the response.

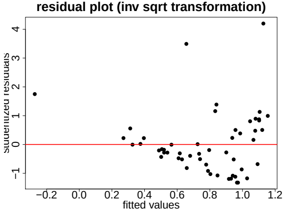
Transformations and weighting

We apply three different transformations on the respose and investigate their effects on the residuals:

- Linear growth ($h(\mu) = \mu$): square root
- Quadratic growth ($h(\mu) = \mu^2$): log
- Cubic growth ($h(\mu) = \mu^3$): reciprocal square root







Which plot seems to (approximately) have a constant variance?

Conclusion: The square root transformation seems to work the best.

With such a transformation, the model becomes

$$\sqrt{y} = \beta_0 + \beta_1 x + \epsilon$$

All the subsequent analysis will be about the transformed response $\sqrt{y}$, the original predictor x, and the corresponding error ϵ.

Transformations to linearize the model

Nonlinearity can occur not only in the response but also in one or more predictors in a multiple regression model.

In many cases, the model can be improved by replacing the predictor that is causing the problem with a non-linear function of the same variable, e.g., replace x_i by $\sqrt{x_i}$.

Such nonlinear models are called **intrinsically linear**.

Transformations and weighting

For example, if the scatter plot of y against x suggests an exponential relationship, then an appropriate model would be

$$y = \beta_0 e^{\beta_1 x} \epsilon.$$

This model is intrinsically linear, because it is equivalent to

$$\underbrace{\log y}_{y'} = \underbrace{\log \beta_0}_{\beta_0'} + \beta_1 x + \underbrace{\log \epsilon}_{\epsilon'}$$

Transformations and weighting

Common intrinsically linear models and required transformations

True relationship	**Transformation**	**Linearized model**
$y = \beta_0 x^{\beta_1}$	$y' = \log y, x' = \log x$	$y' = \log \beta_0 + \beta_1 x'$
$y = \beta_0 e^{\beta_1 x}$	$y' = \log y$	$y' = \log \beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 \log x$	$x' = \log x$	$y = \log \beta_0 + \beta_1 x'$
$y = \frac{x}{\beta_0 x - \beta_1}$	$y' = \frac{1}{y}, x' = \frac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

Remark. The scatterplot of y against x, or residual plot against x, can be used to infer true relationships.

Transformations and weighting

Determining which transformation is needed is also a trial-and-error process:

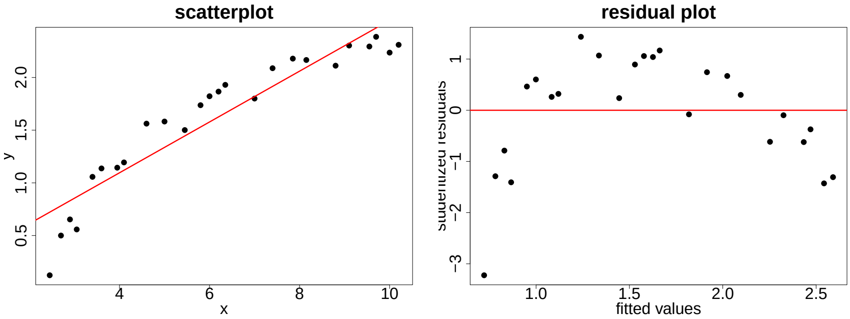
- Start with the most probable transformations based on the scatterplot or residual plots;
- Fit and compare the correspondinng models;
- You would want **a high** R^2, **a low** $\mathbf{MS_{Res}}$, and **a large** F**-statistic** (for significance of regression);
- Take also the context of the model into consideration.

Example: The Windmill Data

An engineer is using a windmill to generate electricity. She has collected data on the DC output of the windmill and the corresponding wind velocity (in mph). The data are available in *Windmill.txt* on the course website.

A scatter plot of the data shows that the relationship between predictor and response is clearly not linear. The curve pattern is even more pronounced in the corresponding residual plot.

Transformations and weighting

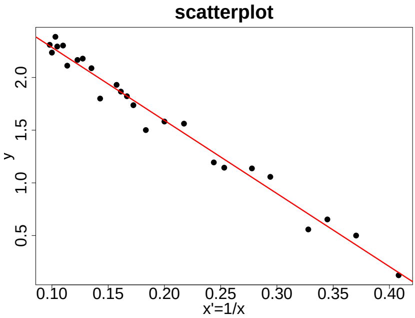


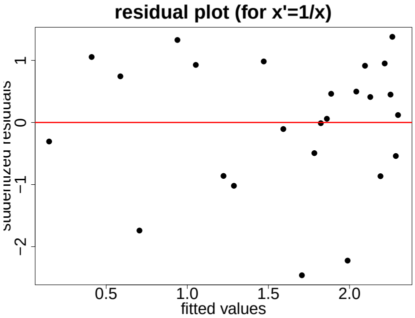
The following table displays the outcomes of the original model and three different transformations:

Model	$\mathbf{R^2}$	**MS**$_{\text{Res}}$	F **statistic**
$y = \beta_0 + \beta_1 x$	0.8745	0.2361^2	160.3
$y = \beta_0 + \beta_1 \sqrt{x}$	0.9219	0.1862^2	271.5
$y = \beta_0 + \beta_1 \log x$	0.9574	0.1376^2	516.6
$y = \beta_0 + \beta_1 \frac{1}{x}$	0.98	0.09417^2	1128.0

Conclusion: The reciprocal transformation is the best in all means.

Transformations and weighting





The Box-Cox method

In some problems, it is not obvious what the “best” transformation for a given data set would be.

George Box and David Cox (1964) came up with a method that would automatically select an optimal transformation of the response y.

The method makes use of a family of possible power transformations:

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda\,\tilde{y}^{\lambda-1}}, & \lambda \neq 0 \\ \tilde{y}\log y_i, & \lambda = 0 \end{cases}$$

where $\tilde{y} = (y_1 y_2 \cdots y_n)^{1/n}$ is the geometric mean.

Transformations and weighting

The optimal transformation parameter λ and the parameters of the least squares regression model

$$\mathbf{y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

are computed simultaneously through *maximum likelihood estimation*.

This method works only if the response y takes on only positive values, which is usually not a problem.

If the response takes on some negative values, add an appropriately large constant to all observations as a “pre-transformation”.

How to select λ

The optimal value of λ is determined by minimizing $SS_{Res}(\lambda)$, over a range of values of λ, of the correspondingly transformed data, or equivalently maximizing the likelihood function

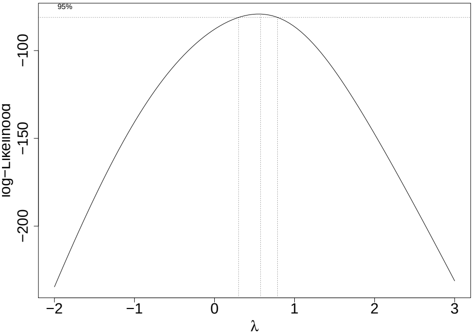
$$L(\lambda) = -\frac{n}{2}\log SS_{Res}(\lambda)$$

A simple λ is usually preferred.

λ	$y' = y^{(\lambda)}$
$\vdots$	$\vdots$
-2	$y' \propto 1/y^2$
-1	$y' \propto 1/y$
$-\frac{1}{2}$	$y' \propto 1/\sqrt{y}$
0	$y' \propto \log y$
$\frac{1}{2}$	$y' \propto \sqrt{y}$
1	$y' \propto y$
2	$y' \propto y^2$
$\vdots$	$\vdots$

Electric utility example revisited

Optimal λ is $\frac{1}{2}$ (consistent with slide 10)



Weighting

When the errors are uncorrelated but have unequal variances, i.e.,

$$\mathrm{Var}(\mathbf{e}) \text{ is a diagonal matrix, but } \mathrm{Var}(\mathbf{e}) \neq \sigma^2 \mathbf{I} \text{ (for any } \sigma^2\text{)}$$

weighting is another effective way of handling the non-constant variance (besides the variance-stabilizing transformation).

For example, consider the following model

$$y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, x\sigma^2)$$

where the predictor takes only positive values.

Transformations and weighting

Assume a sample of n data points (x_i, y_i) from it:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, x_i\,\sigma^2)$$

Then

$$\mathrm{Var}(\mathbf{y}) = \mathrm{Var}(\mathbf{e}) = \sigma^2\,\mathrm{diag}(x_1, x_2, \ldots, x_n).$$

To stabilize the variance, let

$$\epsilon_i' = \epsilon_i/\sqrt{x_i} \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

which are also uncorrelated.

Transformations and weighting

Then the original model can be rewritten as

$$\frac{y_i}{\sqrt{x}_i} = \beta_0 \frac{1}{\sqrt{x}_i} + \beta_1 \sqrt{x_i} + \epsilon_i', \quad i = 1, \ldots, n$$

The total least squares criterion is

$$\sum_{i=1}^{n} \left(\frac{y_i}{\sqrt{x}_i} - \beta_0 \frac{1}{\sqrt{x_i}} - \beta_1 \sqrt{x_i} \right)^2 = \sum_{i=1}^{n} \frac{1}{x_i} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

Correspondingly we have obtained a weighted least squares problem:

$$\min_{\hat\beta_0, \hat\beta_1} \sum_{i=1}^{n} w_i \left(y_i - \hat\beta_0 - \hat\beta_1 x_i \right)^2$$

where $w_i = \frac{1}{x_i}$ are the weights of different observations.

Transformations and weighting

Similarly, for the model

$$y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, x^2\sigma^2)$$

a sample of n data points (x_i, y_i) from it,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, x_i^2\,\sigma^2)$$

the correspondingly weighted least squares problem is

$$\min_{\hat\beta_0, \hat\beta_1} \sum_{i=1}^n w_i\,(y_i - \hat\beta_0 - \hat\beta_1 x_i)^2, \quad w_i = \frac{1}{x_i^2}$$

Weighted least squares with 1 predictor

Let

$$\bar{x}^{(w)} = \frac{\sum w_i x_i}{\sum w_i}, \quad \bar{y}^{(w)} = \frac{\sum w_i y_i}{\sum w_i}$$

and

$$\begin{aligned} S_{xx}^{(w)} &= \sum_{i=1}^{n} w_i (x_i - \bar{x}^{(w)})^2, \\ S_{xy}^{(w)} &= \sum_{i=1}^{n} w_i (x_i - \bar{x}^{(w)})(y_i - \bar{y}^{(w)}) \end{aligned}$$

Then we can prove the following result.

Transformations and weighting

Theorem 0.2. The solution of the weighted least squares problem (with only one predictor)

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} w_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

is given by

$$\begin{aligned}\hat{\beta}_1 &= \frac{S_{xy}^{(w)}}{S_{xx}^{(w)}},\\ \hat{\beta}_0 &= \bar{y}^{(w)} - \hat{\beta}_1 \bar{x}^{(w)}\end{aligned}$$

Simulation (cont'd)

Weighted least squares with k predictors

Assume a data set of n data points with k predictors: $\mathbf{X} \in \mathbb{R}^{n\times p}, \mathbf{y} \in \mathbb{R}^n$.

Let $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$ be the weights that we use for fitting a multiple linear regression model.

Equivalently, we are assuming the following sample regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \text{where} \quad \mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0},\ \boldsymbol{\epsilon} \sim \sigma^2\mathbf{W}^{-1}$$

Define $\mathbf{W}^{1/2} = \mathrm{diag}(w_1^{1/2}, \ldots, w_n^{1/2})$. Note that

$$\mathbf{W}^{1/2} \cdot \mathbf{W}^{1/2} = \mathbf{W}.$$

Transformations and weighting

The weighted least squares fitting problem is

$$\min_{\hat{\beta}} \|\mathbf{W}^{1/2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\|^2$$

and the least squares estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

Proof:

Transformations and weighting

Remark. If $\mathbf{W} = \mathbf{I}$, then the weighted least squares problem reduces to the ordinary least squares problem:

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \qquad \text{where} \quad \mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0},\ \boldsymbol{\epsilon} \sim \sigma^2\mathbf{I}.$$

Along the opposite direction, we can relax the diagonal covariance matrix to be any symmetric, positive definite matrix, leading to the so-called **generalized least squares** problem:

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \qquad \text{where} \quad \mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0},\ \boldsymbol{\epsilon} \sim \sigma^2\mathbf{V}$$

The least squares estimator is

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

Summary

- **Variance-stabilizing transformations** (on the response)
  - Diagnostic plots: residuals against fitted values
- **Linearizing transformations** (on response and/or predictor)
  - Diagnostic plots: response/residual against predictor
  - Want a high R^2, a low $\mathrm{MS}_{\mathrm{Res}}$, and a large F-statistic
- **The Box-Cox method** (to select the best power transformation for the response)
- **Weighted least squares**