Variance-Stabilizing Transformation

Dr. Kiah Wah Ong

ILLINOIS TECH

Introduction

Previously, we have learned how to use various graphical plots to check for different kinds of model inadequacy.

Now we will introduce some corrective procedures for unsatisfied model assumptions.

Model Assumptions

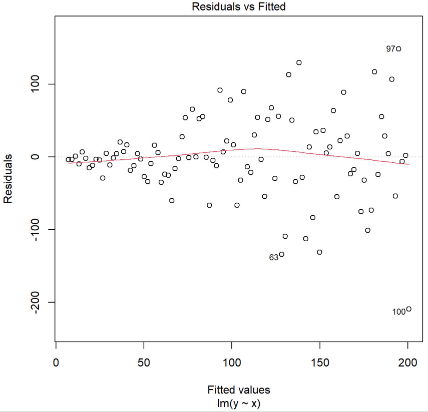
Recall the major assumptions we have made in linear regression models

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \cdots, n$$

are

- The relationship between the response and regressors is linear.
- The error terms ϵ_i have mean zero.
- $\boxed{\text{The error terms } \epsilon_i \text{ have constant variance } \sigma^2 \text{ (homoscedasticity)}}$
- The error terms ϵ_i are normally distributed.
- The error terms ϵ_i and ϵ_j are uncorrelated for $i \neq j$.
- The regressors $x_1, \cdots, x_k$ are nonrandom.
- The regressors $x_1, \cdots, x_k$ are measured without error.
- The regressors are linearly independent.

Variance-Stabilizing Transformation



The plot on the left is an example of a case when the homoscedasticity assumption of a regression model has been violated.

We are going to introduce one type of transformation method (in the response variable) in order to stabilize the variance.

Variance-Stabilizing Transformation

Suppose we have a random variable y whose variance depends on its mean:

$$E(y) = \mu \quad \text{and} \quad \mathrm{Var}(y) = g(\mu)$$

for some function g. We can then write

$$y = \mu + e$$

where $E(e) = 0$ and $\mathrm{Var}(e) = g(\mu)$.

Now we want to find a transformation h such that $h(y)$ has constant variance.

Variance-Stabilizing Transformation

Using a first-order Taylor expansion (centered at μ), we see that

$$h(y) = h(\mu + e) \approx h(\mu) + eh'(\mu).$$

This implies that

$$E(h(y)) \approx h(\mu)$$

and

$$\mathrm{Var}(h(y)) \approx 0 + [h'(\mu)]^2 \mathrm{Var}(\mathrm{e}) = [h'(\mu)]^2 g(\mu).$$

Variance-Stabilizing Transformation

From $\mathrm{Var}(h(y)) = [h'(\mu)]^2 g(\mu)$, we see that the variance stabilizing transformation is a function h that will make $\mathrm{Var}(h(y))$ approximately constant.

To do this, we can solve $[h'(\mu)]^2 g(\mu) = 1$, that is,

$$h'(\mu) = \frac{1}{\sqrt{g(\mu)}} \quad \text{and} \quad h(\mu) = \int^{\mu} \frac{1}{\sqrt{g(s)}}\, ds$$

Variance-Stabilizing Transformation

Example

Let y be the number of fatal car accidents and x be the speed at impact. Suppose the relationship between x and y can be modeled by

$$y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is a random error. Because y is a counting variable, it often have a Poisson distribution with both mean and variance equal to λ. Let us assume here that $y \sim \mathrm{Poisson}(\lambda)$.

Now that the variance of y is equal to its mean, this example fit the discussion we have on the previous slides, with g as an identity map. (Recall previously, we have $E(y) = \mu$ and $\mathrm{Var}(y) = g(\mu)$.)

Not only this, in the next slide, we will show that the variance of y is a function of x.

Variance-Stabilizing Transformation

From

$$y = \beta_0 + \beta_1 x + \epsilon$$

we see that $E(y) = \beta_0 + \beta_1 x$, i.e. the mean of y is a function of x and will increase with x.

Since the variance of y is the same as the mean of y (both equal to λ), we conclude that the variance of y is also going to increase with x, therefore we will not have constant variance in this model.

The assumption on homoscedasticity will be violated. In fact, the variance will increase, when x increases.

Variance-Stabilizing Transformation

Since in our example, we have

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

Hence from our previous discussion, we obtain

$$h(x) = \int \frac{1}{\sqrt{x}}\, dx$$

since g in our example is an identity function. (Recall previously, we have $E(y) = \mu$ and $\mathrm{Var}(y) = g(\mu)$.)

We can now regress $y' = \sqrt{y}$ against x, with $y' = \sqrt{y}$ as our variance-stabilizing transformation. The model we fit is

$$y' = \sqrt{y} = \beta_0' + \beta_1' x.$$

Note that y' is used to indicate the new variable $\sqrt{y}$ and it doesn't mean the derivative of y.

Variance-Stabilizing Transformation

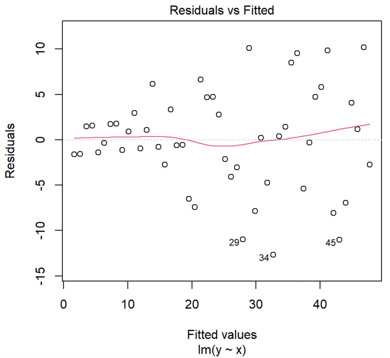
Let us import a simulated data set (where y follows a Poisson distribution) called $\mathrm{VST1.CSV}$ and run the regression model as

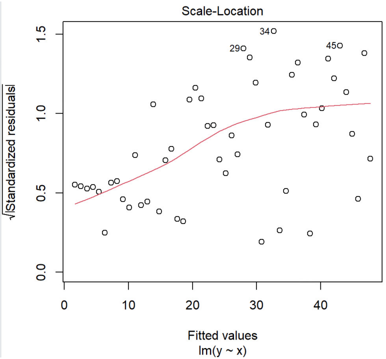
$$y = \beta_0 + \beta_1 x.$$

We then call the regression model as $\mathrm{model1}$

```
VST1<-read.csv("VST1.CSV", header=TRUE, sep=",")
x<-VST1$x
y<-VST1$y
plot(x,y)
model1=lm(y~x)
abline(model1)
plot(model1)
```

Variance-Stabilizing Transformation





Notice that the residuals vs fitted plot and the scale-location plot both show sign of heteroskedasticity. In fact, there is a clear indication that the variances are increasing as $\hat{y}$ increases.

Variance-Stabilizing Transformation

Using the transformation

$$y \to y' = \sqrt{y}$$

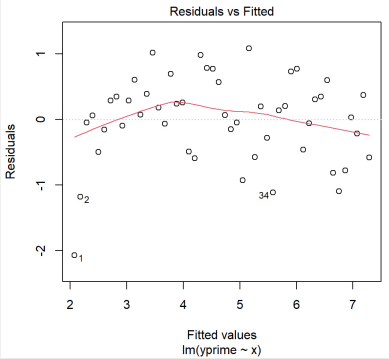
we constructed the second regression model

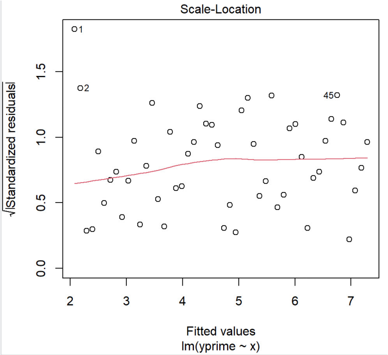
$$\sqrt{y} = \beta_0' + \beta_1' x$$

and call it $\mathrm{model2}$ in R.

```
VST1<-read.csv("VST1.CSV", header=TRUE, sep=",")
x<-VST1$x
y<-VST1$y
plot(x,y)
model1=lm(y~x)
abline(model1)
plot(model1)
yprime<-sqrt(y)
model2=lm(yprime~x)
plot(model2)
```

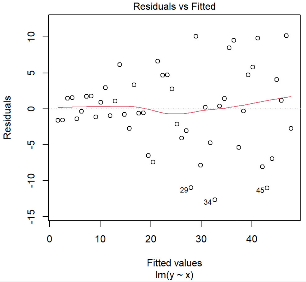
Variance-Stabilizing Transformation

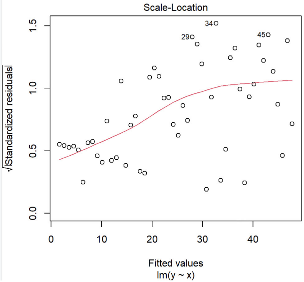


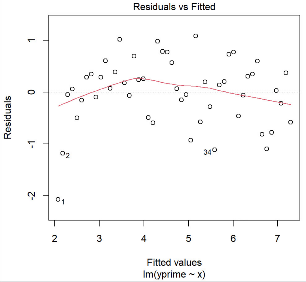


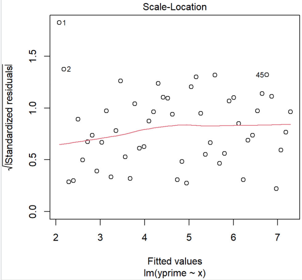
Notice that the residuals vs fitted plot and the scale-location plot both suggest that homoskedasticity condition has been met.

Variance-Stabilizing Transformation









Variance-Stabilizing Transformation

Similar method discussed earlier gives these common transformations:

Relationship of σ^2 to $E(y)$	Transformation
$\sigma^2 \propto \text{constant}$	$y' = y$
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}$
$\sigma^2 \propto E(y)(1 - E(y))$	$y' = \sin^{-1}(\sqrt{y})$
$\sigma^2 \propto [E(y)]^2$	$y' = \ln(y)$
$\sigma^2 \propto [E(y)]^3$	$y' = y^{-1/2}$
$\sigma^2 \propto [E(y)]^4$	$y' = y^{-1}$

Variance-Stabilizing Transformation

Remarks:

The variance-stabilizing transformation like the one shown here is empirical and is a trial-and-error procedure.

You may have to perform some of the common transformations to the response variable y, check the diagnostic plots and refit the model until you found a transformation that leads to acceptable diagnostic plots.