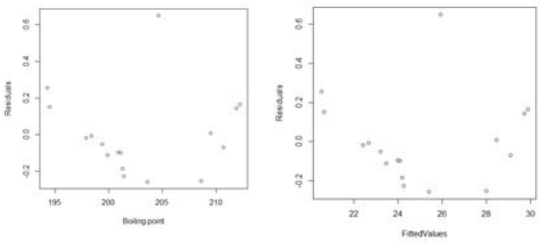
1

STAT2008/STAT6038

Transformations

Residual Plots from a SLR model on the Forbes data

2



Need to modify the model

3

- A noticeable pattern or structure in the residuals is an indication the underlying assumptions of the simple linear regression model do not hold.
- If there is a quadratic structure to the relationship which needs to be taken into account – we could try the following model (which is a actually a multiple regression model):

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon,$$

Transformations to change the scale

4

- Another approach to modifying the model is to change the scale of the axis for one or both the response and explanatory variables.
- We then assume that the linear model holds on the transformed scale (i.e. the relationship may not be linear using the original measurements, but it is linear using the transformed measurements).

Scale Transformations

5

- Sensible changes of scale include transformations that are monotonic as they keep the observations in the same relative order
- A useful transformation is the ln() or natural log transformation to the base e, which is the default base when you use the log() function in R
- There are weaker and stronger transformations, many of which can be expressed as power transformations:
  - square (power 2)
  - square root (power 0.5)

Types of transformations

6

The most common transformations are:

- Square root
- Log
- (some other) Power transformation

Which transformation to apply?

7

Common choices for the functions f and g are the natural logarithm function, and the square root function.

The choice of the functions f and g can often be based on scientific or mathematical reasoning.

For example, scientific theory leads to the following model:

$$\ln(\text{Pressure}) = \beta_0 + \beta_1 \text{Boiling Point} + \epsilon.$$

The data set of Forbes lists the boiling point of water (in Farenheit) and the atmospheric pressure (in inches of mercury) at different places in the alps.

Transformations

8

- We could simply try out several possible transformations and see which one yields the most suitable residual plot (an empirical approach)
- But be aware that transformations have consequences.
- Interpretations about the model parameters and the units they are measured in, are changed by transformation.
- For example, if we use a log() transformation on our data, then any predictions will need to be put through the inverse exp() transformation to interpret the results in the original units of measurement.

Another example

9

suppose we fit the model:

$$Z = \beta_0 + \beta_1 w + \epsilon$$

where $w = \ln(x)$ and $Z = \sqrt{Y}$. If we want to predict the value of the original response variable Y for a particular value, x_0, of the original predictor, we must note that:

$$w_0 = \ln(x_0)$$

$$\hat{Y}(x_0) = [\hat{Z}(w_0)]^2$$

$$= (\beta_0 + \beta_1 w_0)^2$$

$$= (\beta_0 + \beta_1 \ln(x_0))^2$$

Problems with the model assumptions which show up in residual diagnostic plots

10

Relation between Y and X is not linear

Errors have non-constant variance

Errors are not independent

Existence of Outlying Observations

Non-normal Errors

Missing predictor variable(s)

Non-linearity

11

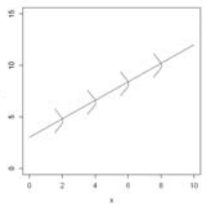
- In some situations, we can transform to achieve linearity. Some transformations can allow us to model relationship between *Y* and *X* that are not linear on the original scale, but become linear on the transformed scale.
- For example: $Y = cX^b$ is actually linear after we apply log transformations to both *Y* and *X*:

$$\log(Y) = \log(cX^b) = \log(c) + b * \log(X)$$

Visualise constant variance:

12

- The four Normal curves represent the Normally distributed outcomes (Y values) at each of four x values.
- The four Normal curves having the same spreads represents the equal variance assumption
- The four means of the Normal curves fall along a straight line representing the linearity assumption.



Non-constant Error Variance

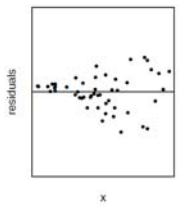
13

- If the variance is not constant, parameter estimates may be valid, but the confidence intervals are misleading.
- Transformations can help us achieve homogeneity of variance (constant variance about the regression equation, also called homoscedasticity)
- For example transformations such as log or square root will affect larger values more than the smaller values, so more help to correct for problems where the variance increases as the X or fitted values increase
- Transformations that correct both non-linearity and heteroscedasticity (non-constant variance) are called "linearising and variance-stabilising" transformations

Example: Square Root Transformation

14

- Square root of a very small number is only a moderately small number
- Square roots of moderate numbers remain moderate. The square roots transformations affects large numbers more dramatically.
- So, taking the square root of the response data can change the nature of the scale of the data in a differential way across the range of the predictors or X variables



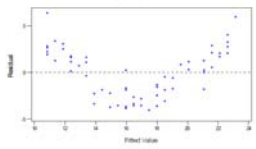
Non-Normality and other problems

15

- Transformations may also help to achieve normality of residuals in some situations or transformations which have a more dramatic affect on larger or smaller values may help bring the more extreme observations (potential outliers or influential observations) closer to the rest of the data

Example

16



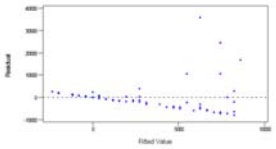
Fixed by fitting a model with X^2 (in addition to X)

OR by replacing X by $log(X)$

OR by replacing X with $\sqrt{(X)}$, or some other nonlinear function.

Another Example

17



- Possible transformation: replace Y by its logarithm.