Model Diagnostics and Remedial Measures in SLR

Jeffrey Woo

MSDS, University of Virginia

Assumptions for Linear Regression Model

In mathematical form: $\epsilon_1, \ldots, \epsilon_n$ i.i.d. $\sim \mathrm{N}(0, \sigma^2)$ (i.i.d. means idependent and identically distributed)

Assumptions for Linear Regression Model

1. The errors, for each fixed value of x, have mean 0. This implies that the relationship as specified in the regression equation is appropriate.
2. The errors, for each fixed value of x, have constant variance. That is, the variation in the errors is theoretically the same regardless of the value of x (or $\hat{y}$).
3. The errors are independent.
4. The errors, for each fixed value of x, follow a normal distribution.

Assumptions for Linear Regression Model

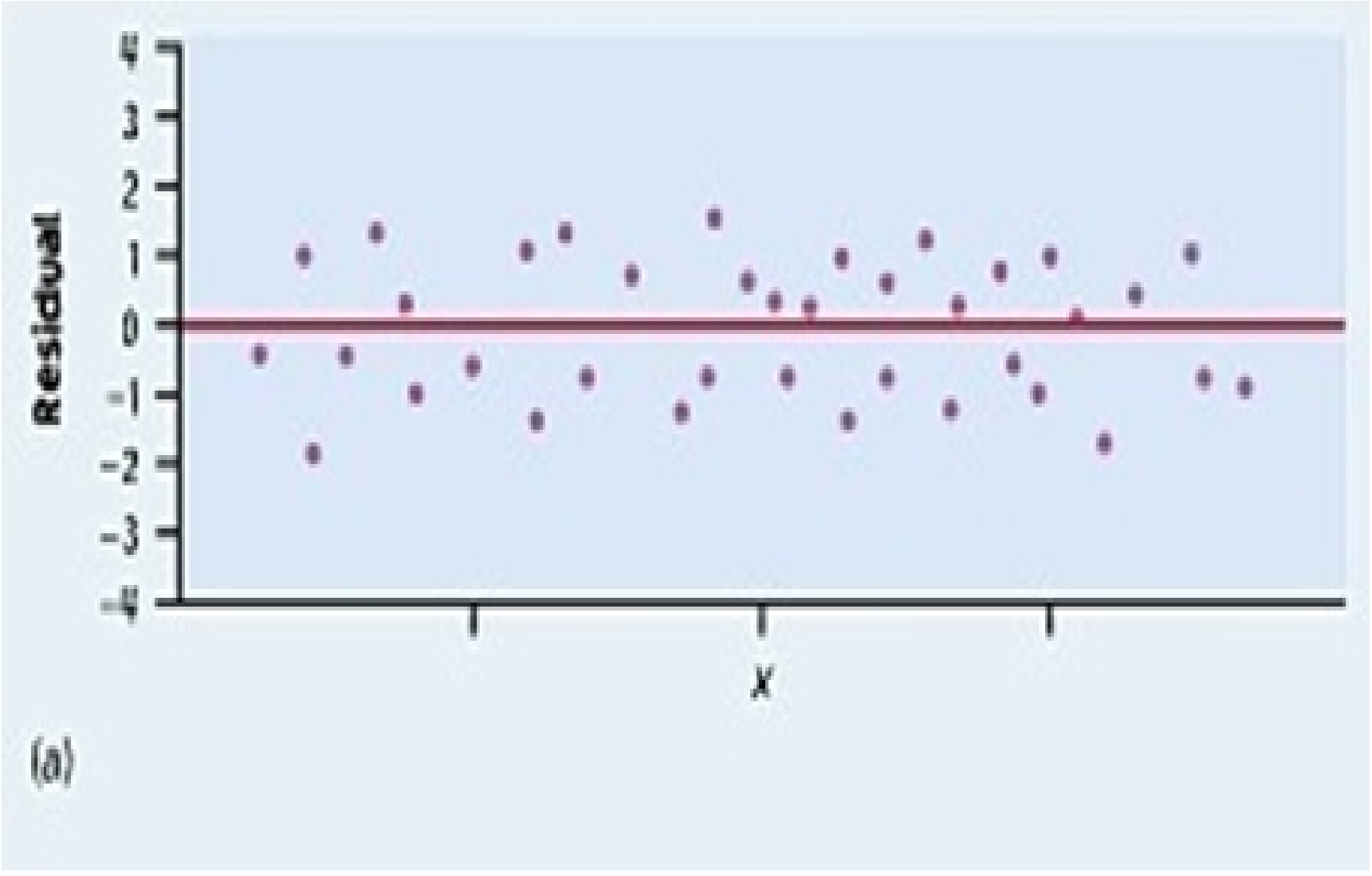
If assumptions are not met, results from hypothesis tests and confidence intervals are no longer reliable. **Note:** It has been shown that the model is fairly robust to the normality assumption; i.e. results are still reliable with some deviation from this assumption.

Residual Plot to Assess Assumptions

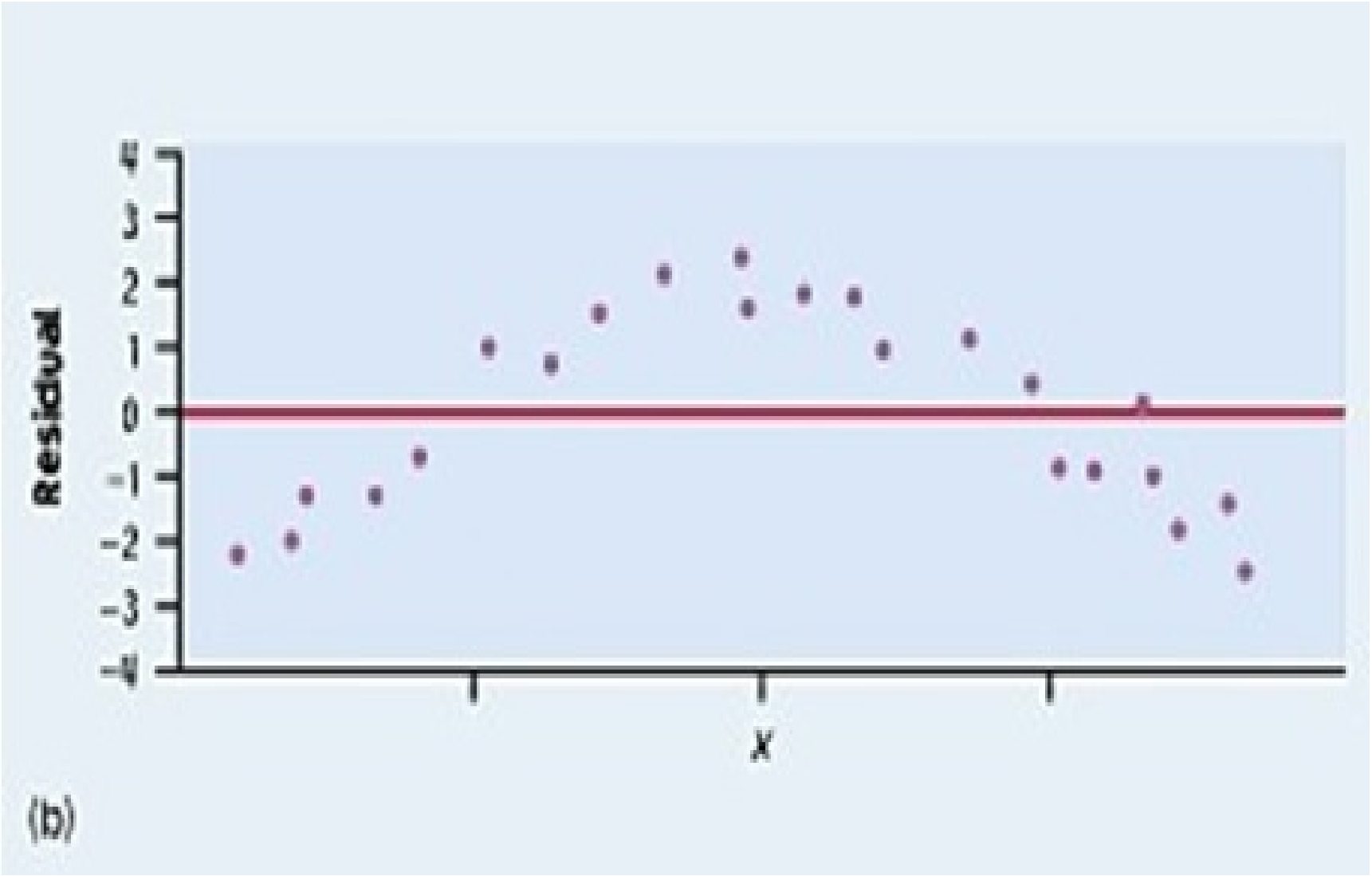
Residual plot: residuals against predictor (in SLR) or $\hat{y}$.

- Residuals should be randomly scattered across the horizontal axis (mean 0).
- Residuals should not display any pattern (mean 0).
- Spread of residuals for each fitted value or x is constant (constant variance).

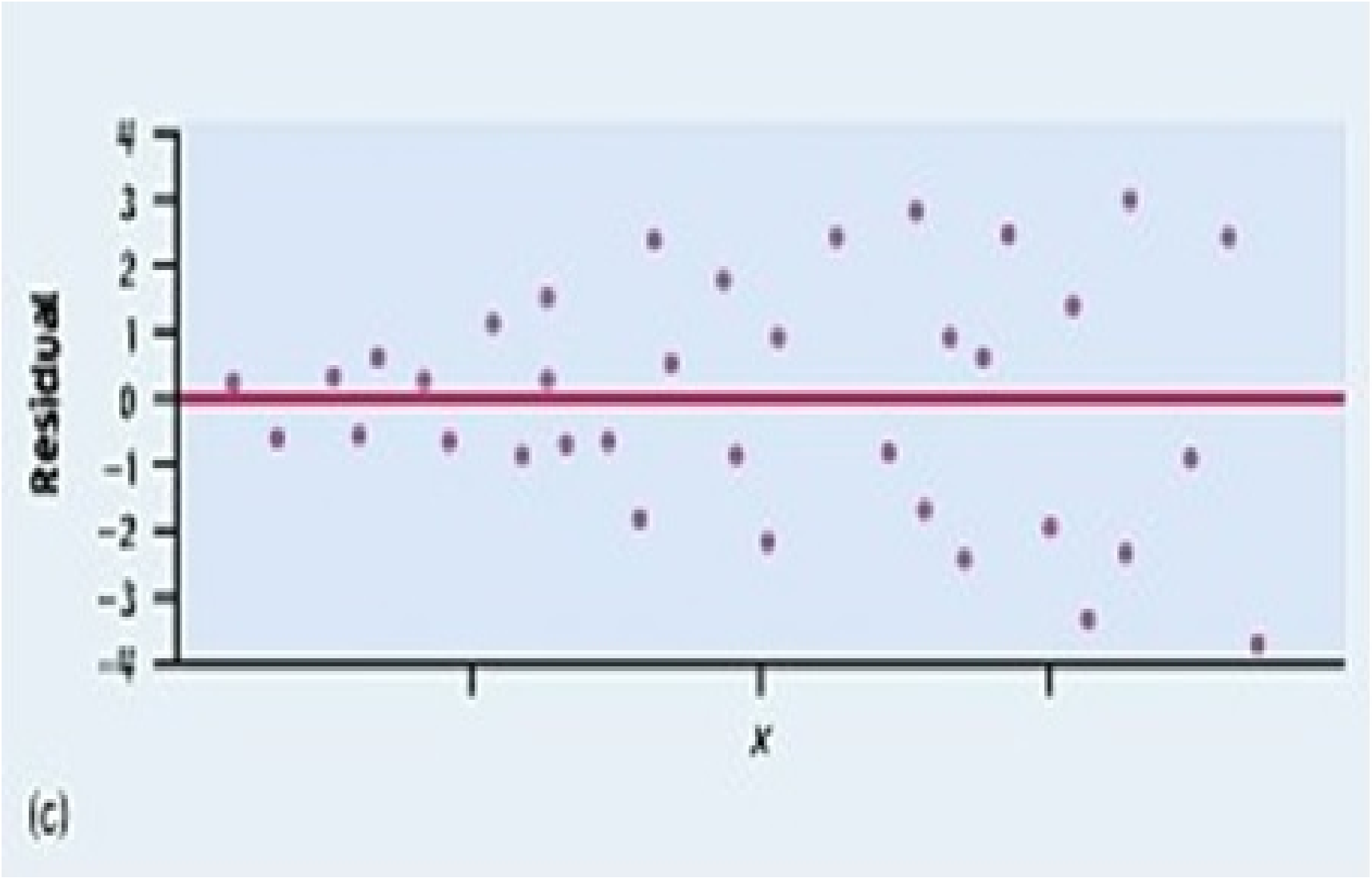
Residual Plot (a)



Residual Plot (b)



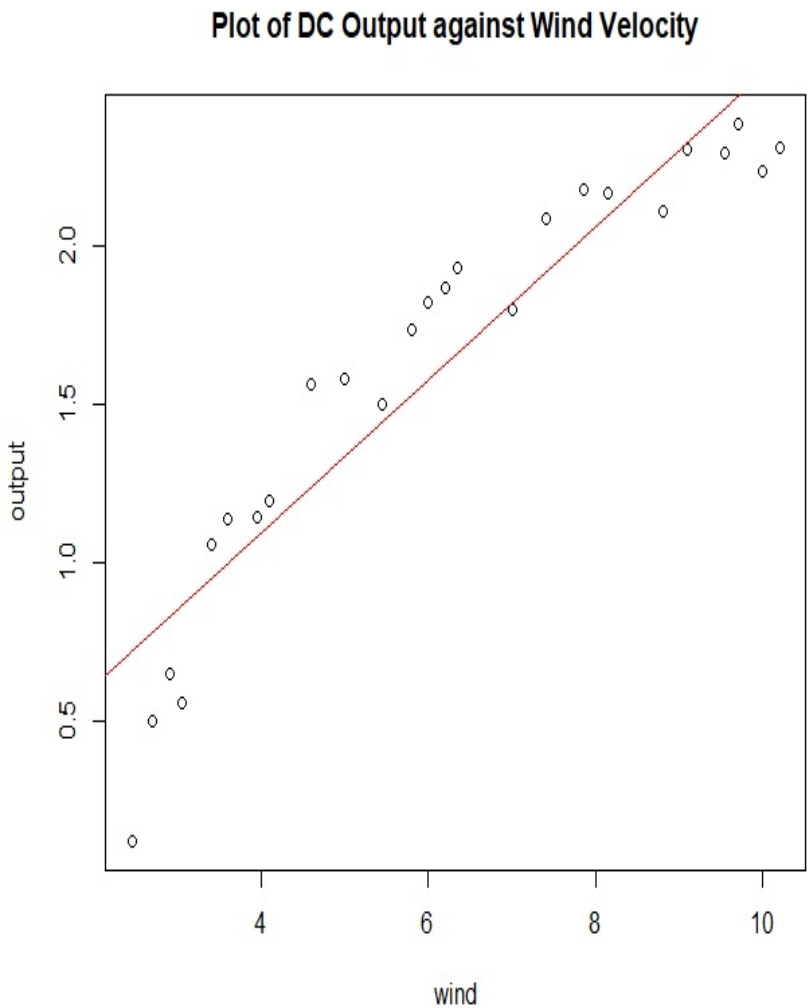
Residual Plot (c)

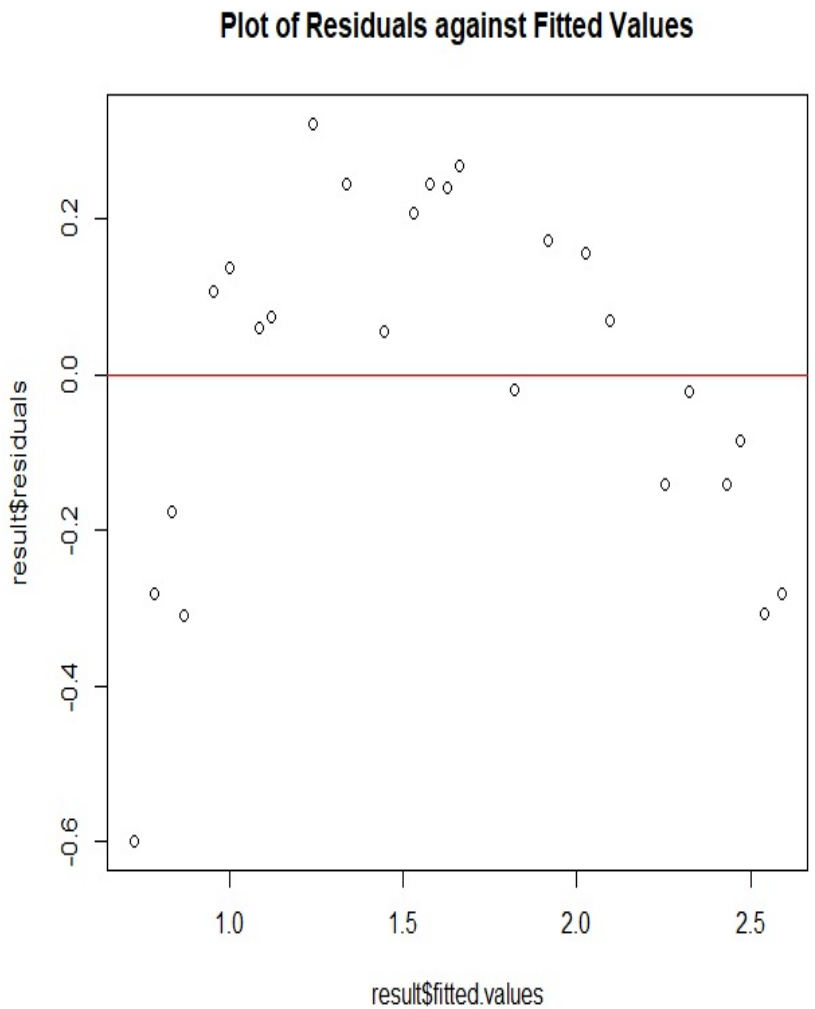


Scatterplot

In simple linear regression, one can actually use the scatterplot of the variables as well.

Example from Tutorial





General Rule for Data Transformation

Residual plots can help detect issues 1 and 2. Data transformations can be used to deal with issues 1 and 2.

- Transforming the response is performed to handle issue 2. A successful transformation of the response will result in a residual plot with constant variance.
- Transforming the response may also influence issue 1; however, the choice of how to transform the response is chosen to solve issue 2.
- Transforming the predictor is performed to handle issue 1. Transforming the predictor does not, theoretically, help with issue 2.
- When both issues 1 and 2 are present, we transform the response first, to handle issue 2. Then we transform the predictor if issue 1 is still present.

Transforming the Response Variable for Non-Constant Variance

From page 12 of textbook,

$$Var(y|x) = Var(\beta_0 + \beta_1 x + \epsilon|x) = Var(\epsilon|x) = \sigma^2$$

since β_0, β_1 are parameters and so are fixed, and the distribution of the response is conditioned on the predictor (predictor viewed as fixed).

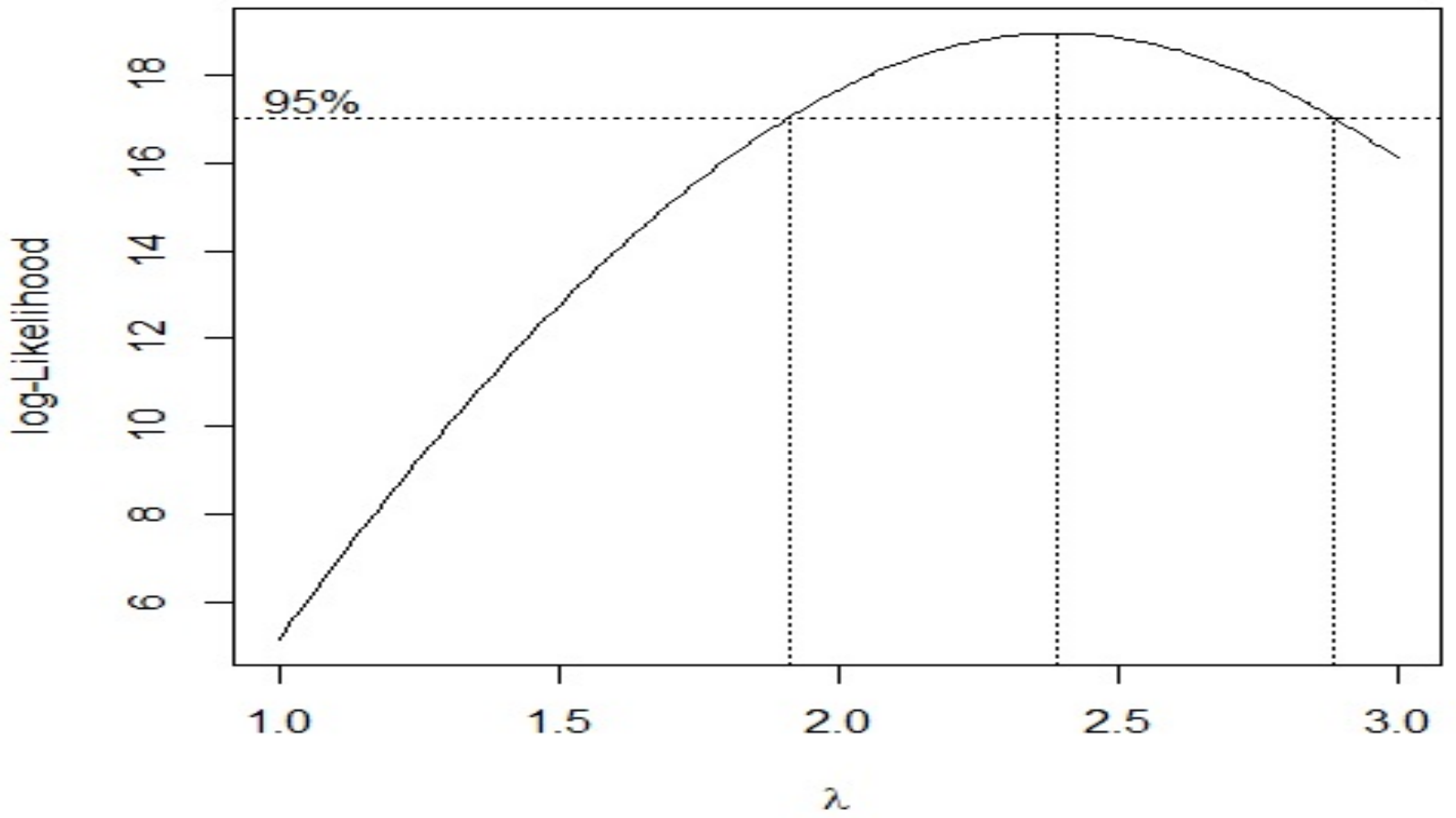
Therefore, to fix issues with constant variance of the errors, we transform the response variable.

Box Cox Transformation

Residual plot is an empirical way to evaluate assumptions. The Box Cox method is an analytical way to evaluate the constant variance and normality assumptions.

- $y^{(\lambda)} = y^{\lambda}$ if $\lambda \neq 0$
- $y^{(\lambda)} = \log y$ if $\lambda = 0$

Example



Data Transformation Procedure

1. A residual plot should always be produced for each model. A Box Cox plot could also be produced.
2. Assess the plots to decide which variables need to be transformed, and how. The choice of transformation should be guided by what you see in the plots, and not by trial and error.
3. Transform one variable at a time. Refit the model. Produce a residual plot (and maybe also a Box Cox plot) after each transformation. The plots should then be assessed if the transformation helped in the way you intended.
4. Assess if another transformation is needed; if so, repeat steps 2 and 3. Stop when the residual plot is fine.