Model diagnostics and remedies. II

Miaoyan Wang

Department of Statistics
UW Madison

Model Assumptions

- The relationship between the response variable Y and the explanatory variables $X_1, X_2, \ldots, X_{p-1}$ is

$$E(Y_i|\boldsymbol{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} \qquad E(\varepsilon_i) = 0$$

- Equal variance:

$$Var(Y_i|\boldsymbol{X}_i) = Var(\varepsilon_i) = \sigma^2.$$

- Independence:

$$Cov(Y_i, Y_{i'}|\boldsymbol{X}_i, \boldsymbol{X}_{i'}) = Cov(\varepsilon_i, \varepsilon_{i'}) = 0 \quad \text{for} \quad i \neq i'.$$

- Normal distribution:

$$Y_i|\boldsymbol{X}_i \sim N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}, \sigma^2) \qquad \varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2$$

Remedial Measures

Basic approaches: replace with a more complex model or transform so that the model is appropriate.

- Nonlinearity of regression function:
  - Transformation.
  - Polynomial regression, nonlinear regression.
- Nonequal error variance:
  - Transformation.
  - Weighted least squares.
- Nonindependence of error terms:
  - Models with correlated error terms (STAT 850).
- Nonnormality of error terms.
  - Transformation.
  - Nonparametric methods.
  - Generalized linear models (STAT 850).
- Presence of outliers:
  - Removal of outliers (with caution).
  - Detection. Robust estimation.

Example: Surviving Bacteria

Data consist of number of surviving bacteria after exposure to X-rays for different periods of time.

- Let t denote time (in number of 6-minute intervals)
- let n denote number of surviving bacteria after exposure to X-rays for t time.

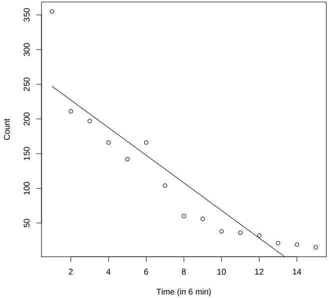
t	1	2	3	4	5	6	7	8
n	355	211	197	166	142	166	104	60

t	9	10	11	12	13	14	15
n	56	38	36	32	21	19	15

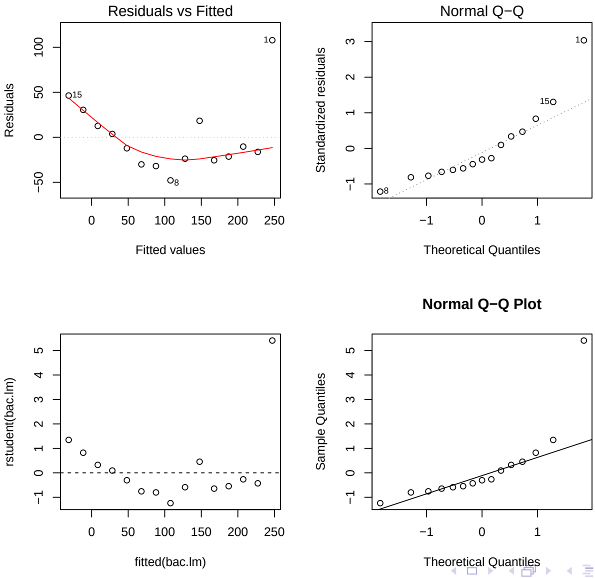
Example: Surviving Bacteria

	Estimate	Std. Error	t value	$Pr(\geq \|t\|)$
(Intercept)	267.010	22.170	12.044	2.0e-08 ***
t	-19.893	2.438	-8.158	1.8e-06 ***

Residual standard error: 40.8 on 13 degrees of freedom
Multiple R-squared: 0.8366, Adjusted R-squared: 0.824
F-statistic: 66.56 on 1 and 13 DF, p-value: 1.804e-06



Example: Surviving Bacteria



Example: Surviving Bacteria

- Here there is a theoretical model:

$$n_t = n_0 e^{\beta t},$$

  where
  - t is time,
  - n_t is the number of bacteria at time t,
  - n_0 is the number of bacteria at the start ($t = 0$), and
  - β is a decay rate with $\beta < 0$.

- Consider a log transformation:

$$\ln(n_t) = \ln(n_0) + \beta t = \alpha + \beta t,$$

  by setting $\alpha = \ln(n_0)$.
  That is, we log-transformed n_t and the result is a linear model.

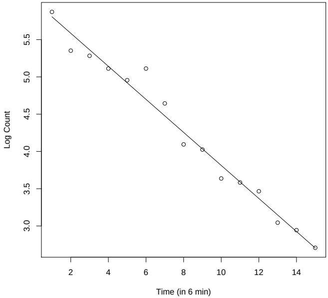
Example: Surviving Bacteria

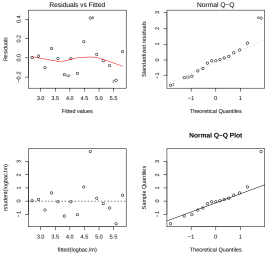
The transformed data are as follows.

t	1	2	3	4	5	6	7	8
$\ln(n)$	5.87	5.35	5.28	5.11	4.96	5.11	4.64	4.09

t	9	10	11	12	13	14	15
$\ln(n)$	4.03	3.64	3.58	3.47	3.04	2.94	2.71

Example: Surviving Bacteria





Example: Surviving Bacteria

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	6.028695	0.088259	68.31	$<$ 2e-16 ***
t	-0.221629	0.009707	-22.83	7.1e-12 ***

Residual standard error: 0.1624 on 13 degrees of freedom
Multiple R-squared: 0.9757, Adjusted R-squared: 0.9738
F-statistic: 521.3 on 1 and 13 DF, p-value: 7.103e-12

How to interpret β ? linear time trend in log count
How to interpret α ? expected log count at the start
Inference for n_0 is not straightforward.

$$\hat{n}_0 = e^{\hat{\alpha}} = 415.30 \quad \text{but} \quad E(\hat{n}_0) \neq n_0.$$

Transformation: Remarks

- Ideally, theory should dictate what transformation to use.
- In practice, transformation is usually chosen empirically based on data analysis.
- Usually it is best to start with a simple transformation and experiment.
  - To meet the linearity assumption, transformation could be that of X, or Y, or both.
  - Common transformations are $\log_{10}, \ln, \sqrt{\cdot}$. Less common transformations are $Y^2, 1/Y, 1/Y^2, \arcsin\sqrt{Y}$.
- Another advantage of transformation is to control unequal variance.

Transformation: Remarks

- Consider a transformation ladder for $Z = X$ or Y.

λ	$\cdots$	-2	-1	-0.5	0	0.5	1	2	$\cdots$
Z^λ	$\cdots$	$\frac{1}{Z^2}$	$\frac{1}{Z}$	$\frac{1}{\sqrt{Z}}$	$\log(Z)$	$\sqrt{Z}$	Z	Z^2	$\cdots$

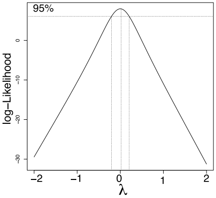
- Transforming Y can affect both linearity and equal variance, but transforming X can affect only linearity.
- Sometimes solving one problem can create another.

Box-Cox Transformation

- **Box-Cox method** is a formal approach to selecting λ to transform Y.
- The idea is to consider

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \epsilon_i.$$

- Estimate λ (along with $\beta_0, \beta_1, \sigma^2$) using ML.
- Choose an interpretable $\hat{\lambda}$ within a 95% CI. In the surviving bacteria example, the Box-Cox method gives $\hat{\lambda} = -0.0202$.
  Implication: ____________________



- R command: boxcox(object, ...) in the MASS library