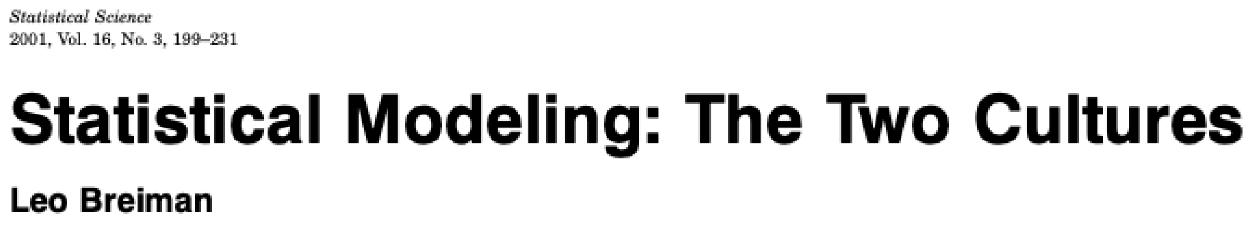
Lecture 2 Two Cultures and Simple Models

Li Xing

June 13, 2023

Statistical Modeling: The Two Cultures



Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Leo Breiman (January 27, 1928 – July 5, 2005) was a distinguished statistician at the University of California, Berkeley. He was the recipient of numerous honors and awards, and was a member of the United States National Academy of Sciences. (From Wiki)

Purpose of Statistical Modeling

- Date Generation

x -> nature -> y

- There are two goals to analyze the data.
    - **Information**: how nature is associating y to x
    - **Prediction**: predict y for future x
- Correspondingly there are two cultures for statistical modelling.

The Data Modeling Culture

Modeling Format:
response variables = f(predictor variables, random noise, parameters)

Model validation: goodness-of-fit test, residual examination

Examples: *linear regression, logistic regression, cox regression, etc*

Characters: model-based approach, usually simple and easy to interpret parameters

The Algorithmic Modeling Culture

Modeling Format:
usually a complex and algorithmic function to predict y for future x

Model validation: predication accuracy

Examples: *penalized regression, ensembles learning, neural network, etc*

Characters: data-driven, usually complex and hard to interpret parameters

A “New” Research Community

Targeted at predictive accuracy, there is a new research community consisted of **computer scientists, physicists, engineers, statisticians**.

They are working on complex prediction problems where data models were not applicable.

i.e. speech recognition, image recognition, nonlinear time series prediction, handwriting recognition, prediction in financial markets, etc.

Back to this course,

1. Statistical learning methods focus on the second culture. Be prepared that there is a shift of target from traditional methods.

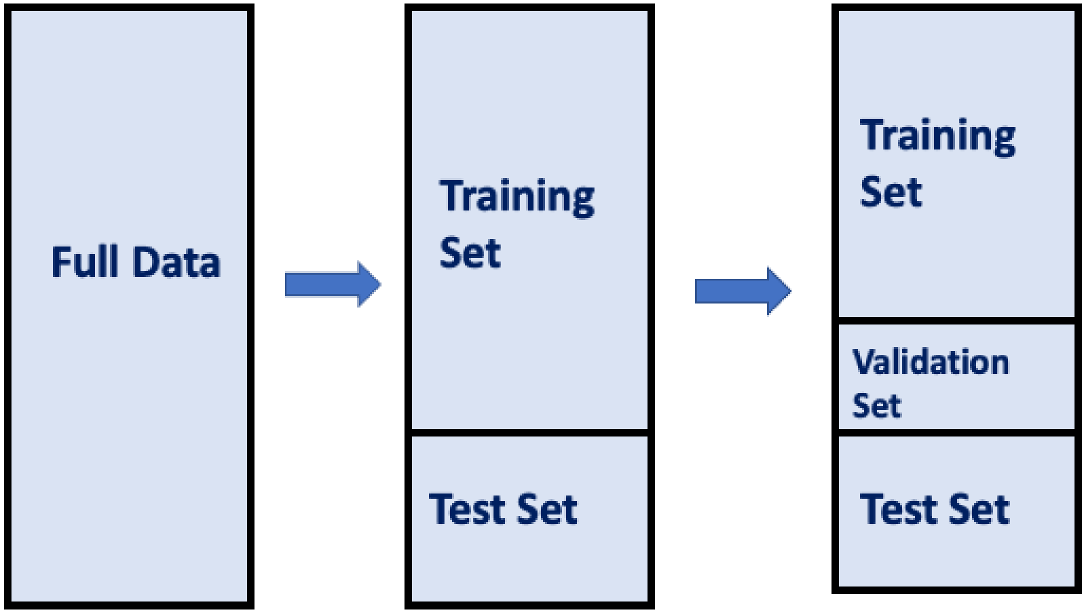
2. Be aware of the two cultures in reality. So get prepared!

Difference in analysis steps between those two models

- Traditional methods use all the observations and target at association parameters. We employ residuals to check goodness of the model fitting. The estimated coefficients are the point of attraction for most analysis.
- Machine learning methods are evaluated by prediction accuracy. Therefore, we will use the data differently.

A Critical Pre-Step For Building A Good Predictive Model

Validation Study



Linear Regression

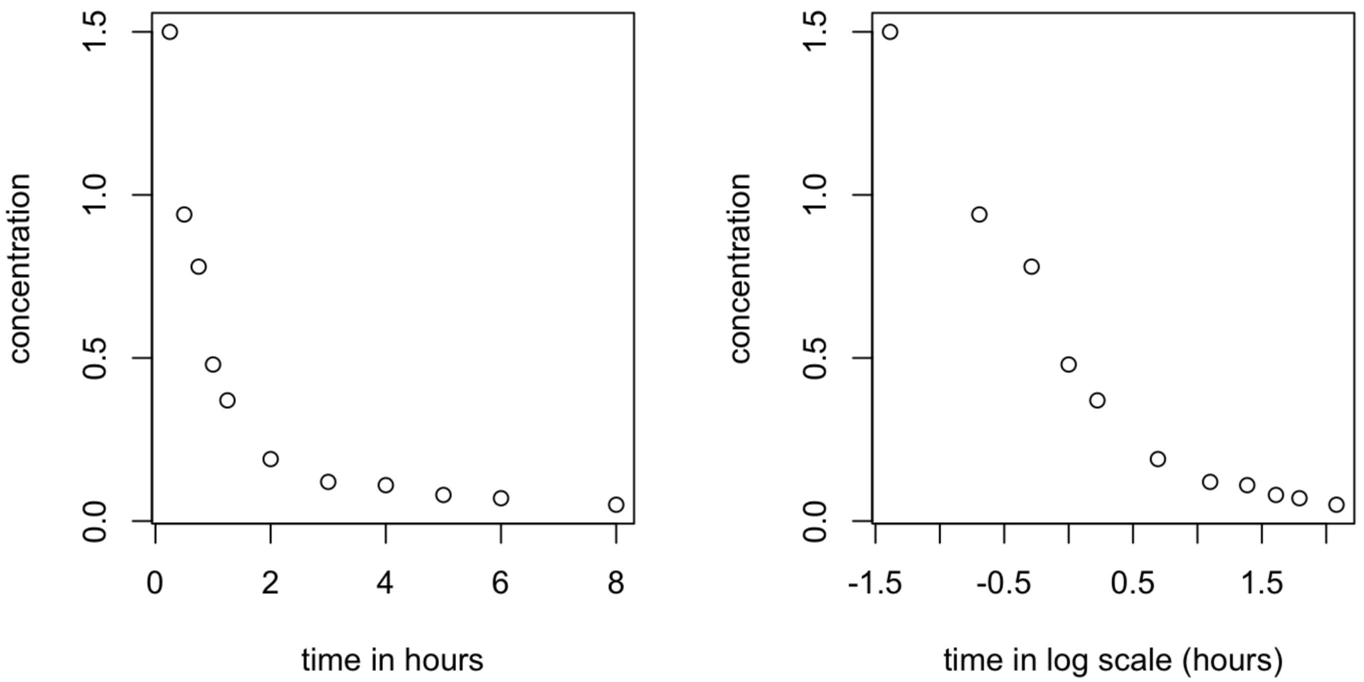
Pharmacokinetics of Indomethacin Data

```
data(Indometh)
mydata <- data.frame(subset(Indometh, Subject == 1))
head(mydata)
```

```
##   Subject time conc
## 1       1 0.25 1.50
## 2       1 0.50 0.94
## 3       1 0.75 0.78
## 4       1 1.00 0.48
## 5       1 1.25 0.37
## 6       1 2.00 0.19
```

Subject: subject index;
time: a numeric vector of times at which blood samples were drawn (hr);
conc: a numeric vector of plasma concentrations of indometacin (mcg/ml).

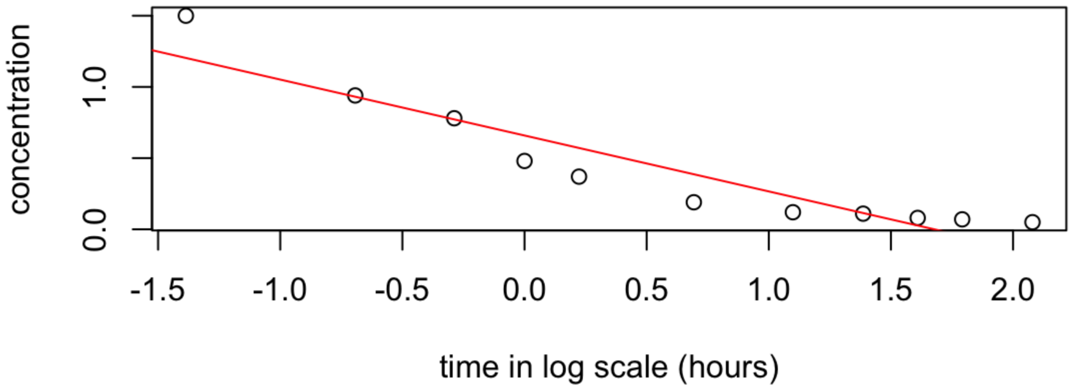
Check the Association between conc and time



Simple Linear Regression for conc and log(time)

We fit a simple linear regression between log(time) and conc.

```
myfit <- lm(mydata$conc~mydata$logtime)

plot(mydata$logtime, mydata$conc, xlab="time in log scale (hours)", ylab = "concentration")

abline(myfit$coef, col = "red")
```



Simple Linear Regression Definition

Assume y_i is the response variable for the ith subject and x_i is the predictor for the ith subject. A simple linear regression model is defined as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i ,$$

where β_0 is the parameter for the interpret term, β_1 is the parameter for the slope term, ϵ_i is the random error for the ith subject, and $\epsilon_i \sim N(0, \sigma)$ with $i = 1, 2, \dots , n$.

Simple Linear Regression Properties

- Mean

$$E(y|x) = \beta_0 + \beta_1 x$$

- Variance

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

Simple Linear Regression Terms

- fitted value

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- residual

$$e_i = y_i - \hat{y}_i$$

- coefficient of determination

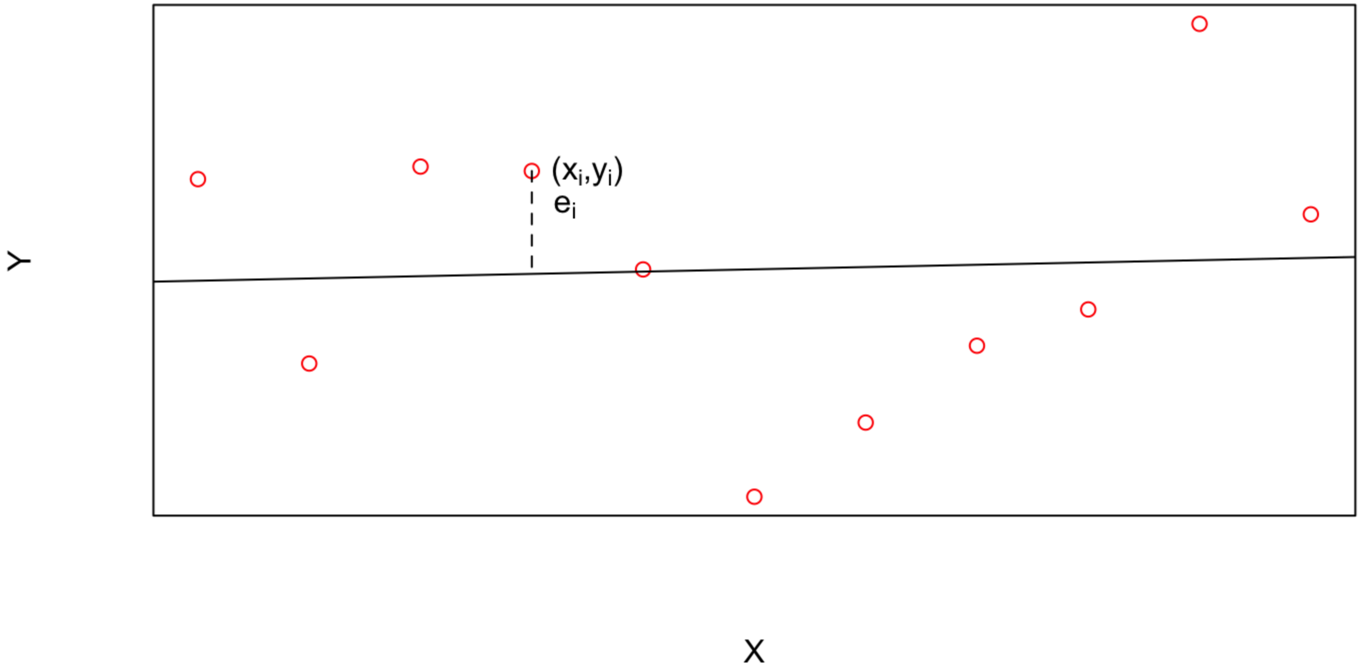
$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

Meaning: Proportion of variation explained by the predictor, x

Simple Linear Regression Model

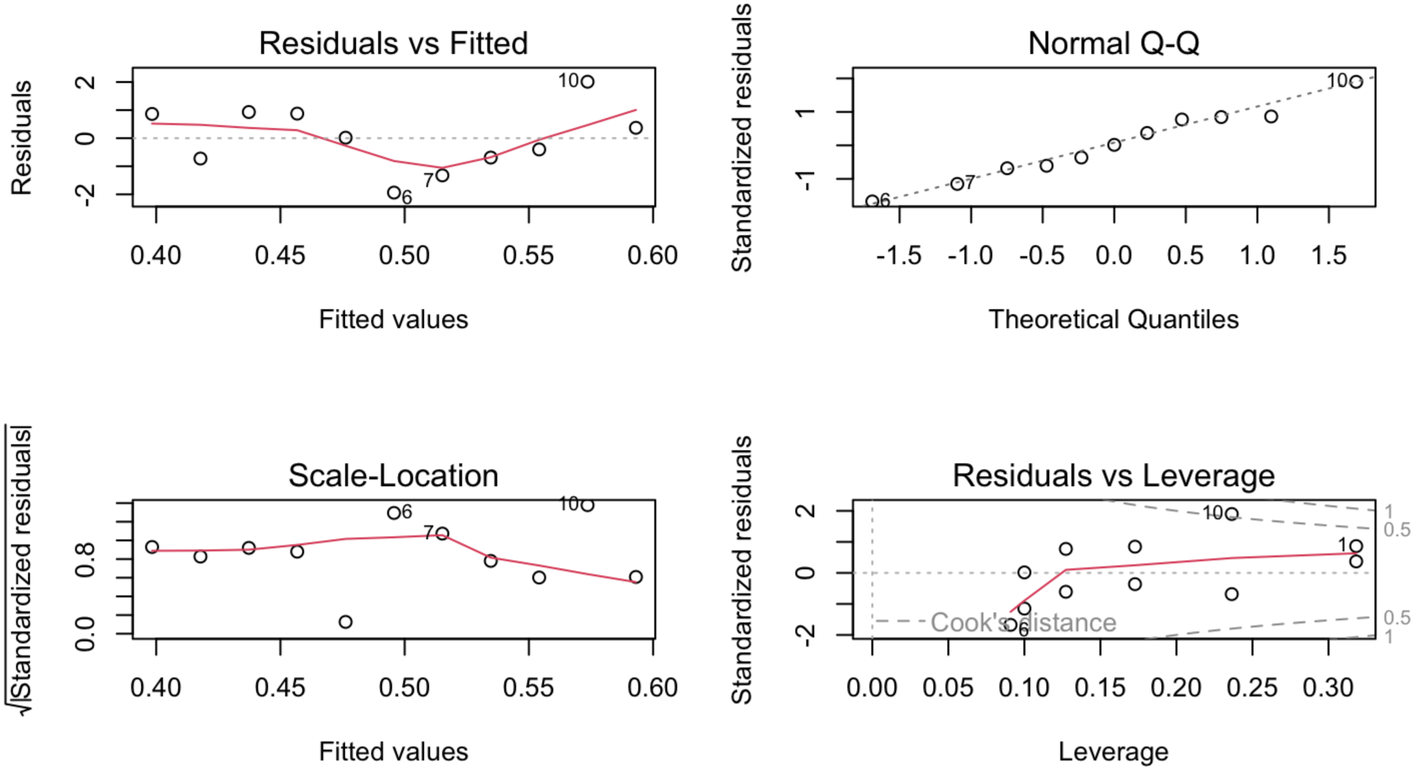
```r
myx <- seq(0, 1, by = 0.1)

set.seed(0)

mye <- rnorm(length(myx), 0, 1)

beta0 <- 0

beta1 <- 0.2

myy <- beta0 + beta1*myx + mye

mymodel <- lm(myy ~ myx)

plot(myx, myy, col = "red", ylab = "Y", yaxt="n", xlab = "X", xaxt="n")

abline(coef(mymodel))

text( myx[4] + 0.05, myy[4], expression(paste("("*x[i]*","*y[i]*")")) )

segments(myx[4], myy[4], myx[4], coef(mymodel)%*%t(t(c(1, myx[4]))), lty = 2)

text( myx[4] + 0.03, myy[4]-0.3, expression(e[i]))
```

Simple Linear Regression Model



Model Checking

```
par(mfrow = c(2, 2))

plot(mymodel)
```



Multiple Linear Regression

- Simple linear regression: one predictor;
- Multiple linear regression: at least two predictors.

We have n subjects (i.e. patients) with p predictors (also called features). A multiple linear regression can be written as below.

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i \,,$$

where $i = 1, 2, \cdots, n$ and $p \geq 1$.

Multiple Linear Regression Example

- with different predictors
- with same predictors such as polynomial regressions

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i ,$$

Note that coefficient of determination R^2 keeps on increasing by adding more predictors in. Alternative ways of model checking: residula plots, adjusted R^2, AIC, BIC, etc.

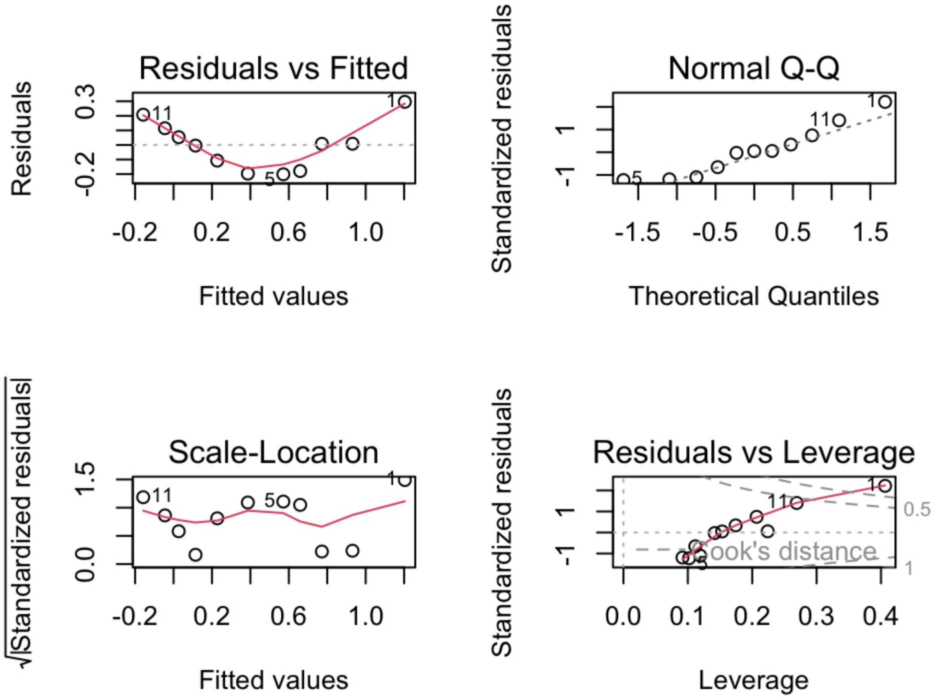
Back to Our Original Example on Indomethacin

The Previous Fitted Model of the Indomethacin

```
summary(myfit)
```

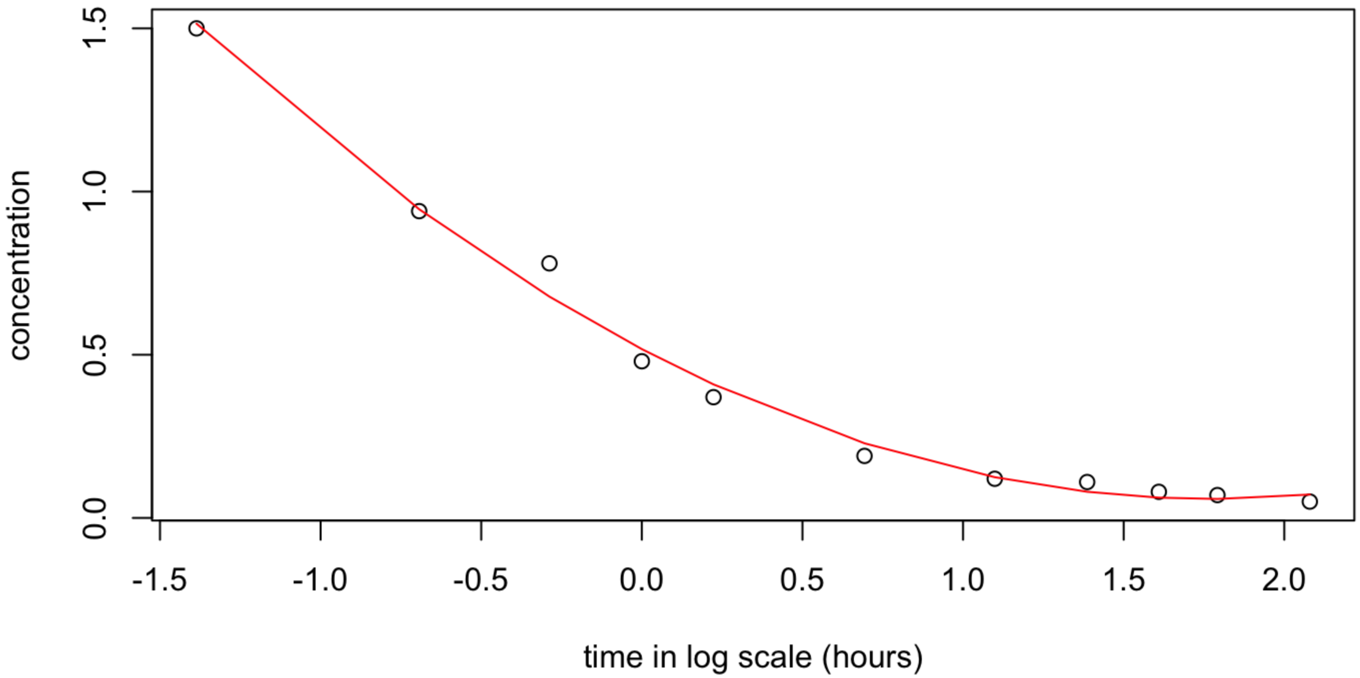
Model Checking Based on Residual Plots

```
par(mfrow = c(2, 2))

plot(myfit)
```



Fit a Quaratic Model

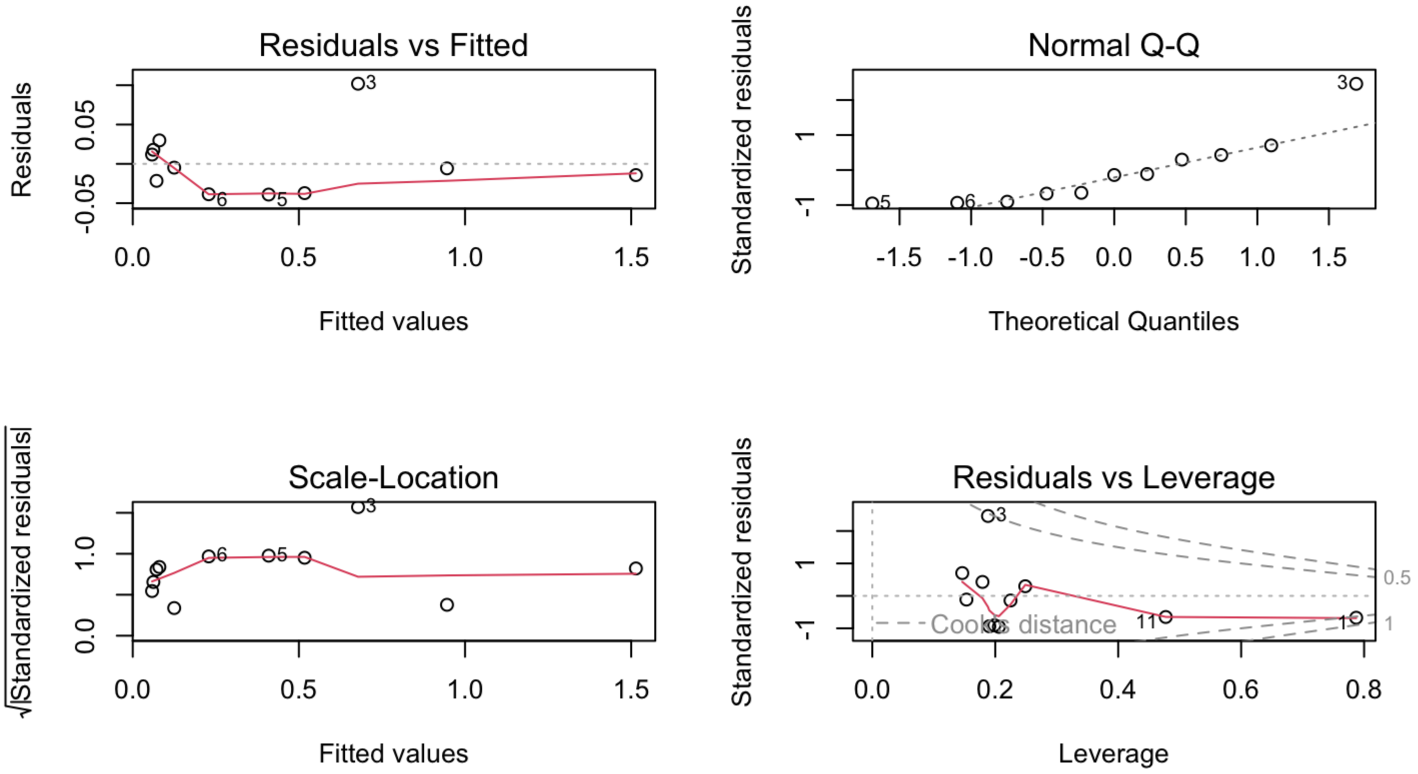
```
myfit2 <- lm(mydata$conc~mydata$logtime + I(mydata$logtime^2))

plot(mydata$logtime, mydata$conc, xlab="time in log scale (hours)", ylab = "concentration")

lines(mydata$logtime, myfit2$coef%*%t(cbind(1,mydata$logtime, mydata$logtime^2)), col = "red")
```



Residual Checking

```
par(mfrow = c(2, 2))
```

```
plot(myfit2)
```



Validation Study about Model Selection

We will use validation study and the criteria for selection is Mean Squared Error in the validation set.

-For the training data, $(x_i, y_i, \hat{y}_i)$ (fitted)
-For the validation/test data, $(x_i, y_i, \tilde{y}_i)$ (predicted)

-The MSE for the training set is $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$,
-The MSE for the test set is $\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$.