Statistical Inference and Multivariate Analysis (MA324)

LECTURE SLIDES

Lecture 31

Regression: Model Adequacy Checking

Indian Institute of Technology Guwahati

Jan-May 2023

Prediction Interval

- Consider a level of regressor x_0.
- Let y_0 be the corresponding value of the response.
- We want a prediction interval for y_0.
- Let $\Psi = y_0 - \hat{y_0} = y_0 - \underset{\sim}{x_0}^T \hat{\underset{\sim}{\beta}} \sim N(0, \sigma^2(1 + \underset{\sim}{x_0}^T (X'X)^{-1} \underset{\sim}{x_0}))$.
- A pivot is, $\frac{\Psi - 0}{\sqrt{MS_{Res}(1 + \underset{\sim}{x_0}^T (X'X)^{-1} \underset{\sim}{x_0})}} \sim t_{n-p-1}$.
- A $100(1-\alpha)\%$ prediction interval of y_0 is, $[\hat{y_0} \mp t_{n-p-1;\frac{\alpha}{2}} \sqrt{MS_{Res}(1 + \underset{\sim}{x_0}^T (X'X)^{-1} \underset{\sim}{x_0})}]$

Standardized Regression Coefficients

- Difficult to compare regression coefficients. The magnitude of β_j reflects the unit of measurement of regressor x_j.
- For example, $y = 5 + x_1 + 1000x_2$, where y is measured in liters, x_1 in milliliters, and x_2 in liters. Here, $\beta_2 >> \beta_1$. But the effects of both regressor on y are identical.
- Wayout is to standardized the regressors and response so that they become unit free.

- A popular approach is as follows:
  - Define $W_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{S_{jj}}}$, $i = 1, 2, ..., n; j = 1, 2, ..., p$.
    $y_i^* = \frac{y_i - \overline{y}}{\sqrt{SS_T}}$, $i = 1, 2, ..., n$.
    Here $S_{jj} = \sum_{i=1}^{n} (x_{ij} - \overline{x_j})^2$, $j = 1, 2, ..., p$.
  - Clearly the mean $\overline{W}_j = 0$ and $\sqrt{\sum_{i=1}^{n} (W_{ij} - \overline{W}_j)^2} = \sqrt{\sum_{i=1}^{n} W_{ij}^2} = 1$
    $\overline{y}^* = 0$ and $(\sum_{i=1}^{n} (y_i^*)^2)^{\frac{1}{2}} = 1$
  - In terms of $y^*, W_1, ..., W_p$, the regression model becomes,
    $$y^* = b_1 W_1 + b_2 W_2 + ... + b_p W_p + \epsilon$$
  - In matrix notation, $\underset{\sim}{y}^* = W \underset{\sim}{b} + \underset{\sim}{\epsilon}$.
  - LSE, $\hat{\underset{\sim}{b}} = (W^{'} W)^{-1} W^{'} \underset{\sim}{y}^*$.

- $W'W = \begin{pmatrix} 1 & r_{12} & . & . & . & r_{1p} \\ r_{21} & 1 & . & . & . & r_{2p} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ r_{p1} & r_{p2} & . & . & . & 1 \end{pmatrix}_{p \times p}$ $W'\underset{\sim}{y}^* = \begin{pmatrix} r_{1y} \\ r_{2y} \\ . \\ . \\ . \\ r_{py} \end{pmatrix}_{p \times 1}$

where, $r_{ij} = \frac{\sum_{u=1}^{n}(x_{ui}-\bar{x}_i)(x_{uj}-\bar{x}_j)}{(S_{jj}S_{ii})^{\frac{1}{2}}} = \frac{S_{ij}}{(S_{jj}S_{ii})^{\frac{1}{2}}}$,

$r_{iy} = \frac{\sum_{u=1}^{n}(x_{ui}-\bar{x}_i)(y_u-\bar{y})}{(S_{jj}S_{ii})^{\frac{1}{2}}} = \frac{S_{jy}}{(S_{jj}SS_T)^{\frac{1}{2}}}$

- Drawback : $\hat{b_j}$ are affected by the range of values of regressor variables. Consequently, it may be dangerous to use the magnitude of $\hat{b_j}$ as a measure of relative importance of regressor x_j.

Model Adequacy Checking

- Major assumptions
  - linear relationship
  - Error mean zero
  - Error variance is constant
  - Error are uncorrelated
  - Error are normally distributed and independent
- Gross violation of the assumptions may lead to a totally different model with opposite conclusions.
- We perform the checking using residuals.

- We now define 3 types of residuals.
- Residual: $e_i = y_i - \hat{y}_i$.
- Standardized residual: $d_i = \frac{e_i}{\sqrt{MS_{Res}}}$
- Studentized residual:

  $\underset{\sim}{e} = (I - H)\epsilon \Rightarrow Var(\underset{\sim}{e}) = \sigma^2(I - H)$.

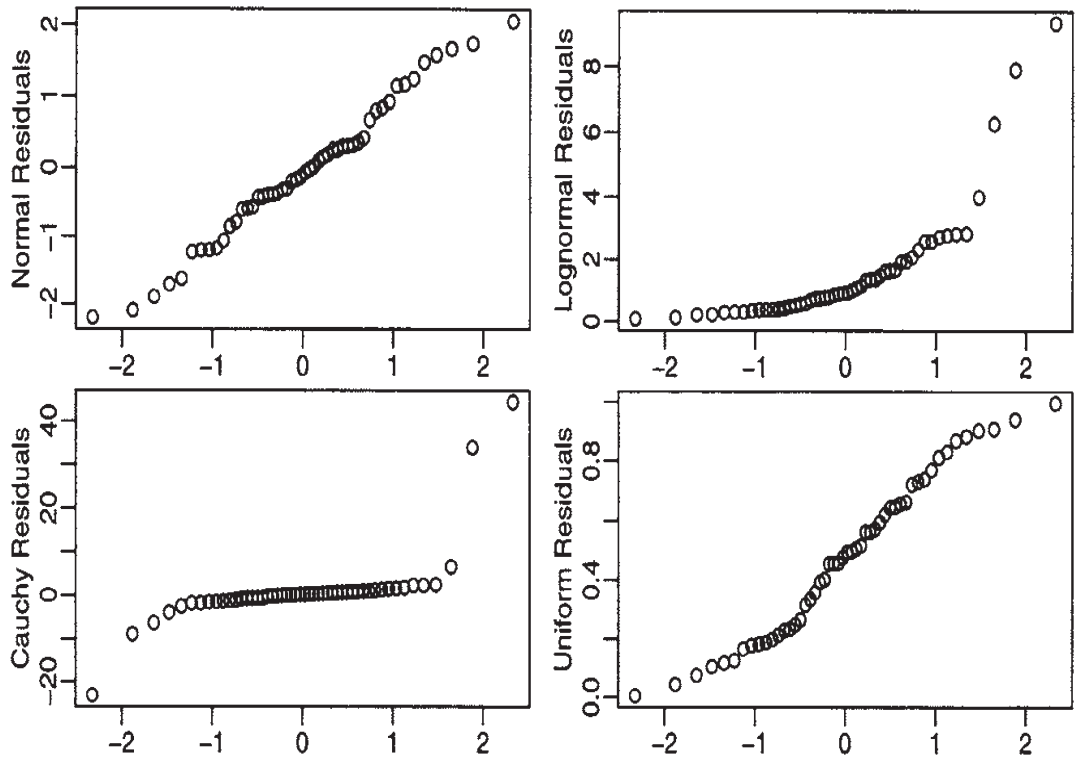
  $Var(e_i) = \sigma^2(1 - h_{ii})$ and $Cov(e_i, e_j) = -\sigma^2 h_{ij}$.

  It can be shown that $0 \leq h_{ii} \leq 1$.

  $r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}$.

Residual Plots: Q-Q Plot (Test for normality)

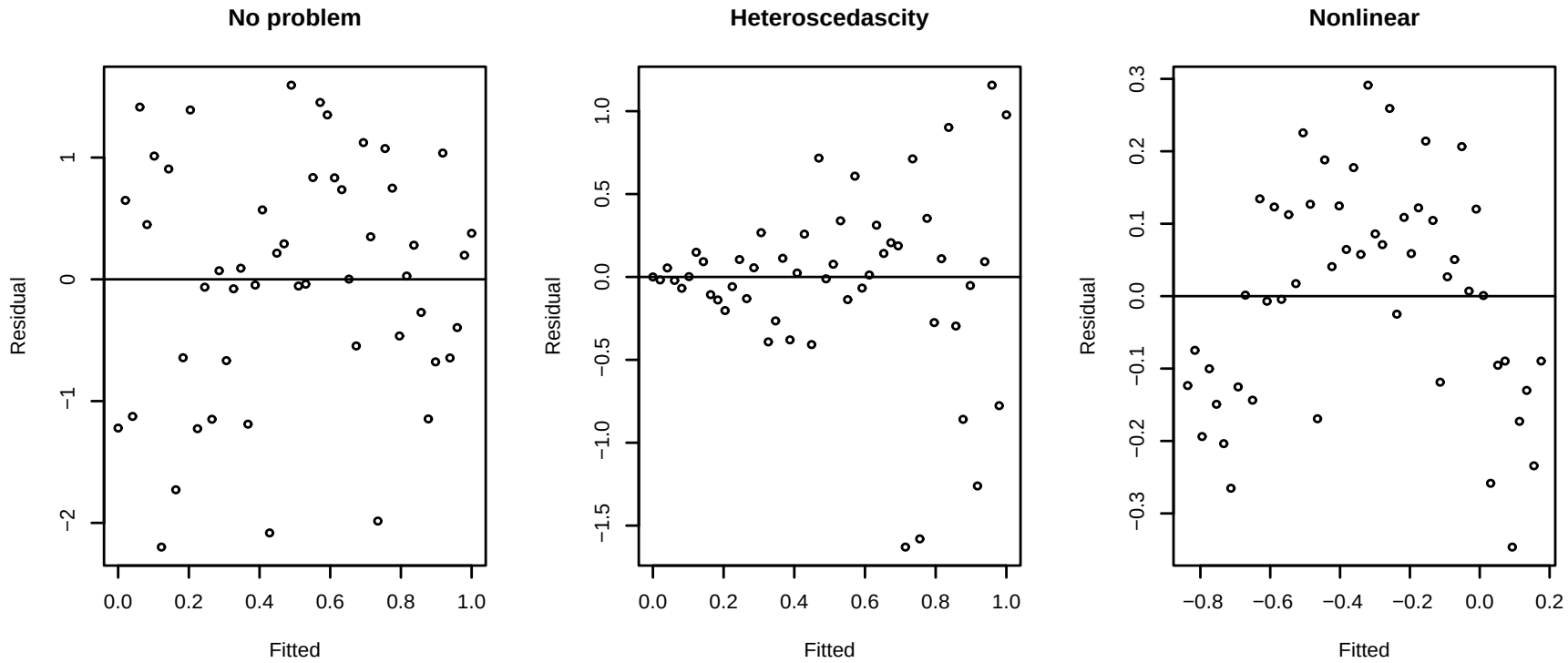
- The residuals can be **assessed for normality** using a Q–Q plot. This **compares the residuals to "ideal" normal** observations.
- We **plot the quantiles corresponding to sorted residuals (e_i) against** $\Phi^{-1}(\frac{i}{n+1})$ for $i = 1, \cdots, n$.
- Four Q-Q plots are:
  Normal: ideal;
  Lognormal: an example of a skewed distribution;
  Cauchy: an example of a long-tailed (platykurtic) distribution;



- **Uniform:** an example of a short-tailed (leptokurtic) distribution
- Source (book): Linear Models with R by Julian J.Faraway

Residual Plots: Plot of residual against Fitted values

- Plot $\hat{y}_i$ vs $e_i(or\ d_i\ or\ r_i)$: (Test of constant variance and non-linear relation)



Residuals vs Fitted plots - the **first** suggests **no change** to the current model while the **second** shows **non-constant variance** and the **third** indicates **some nonlinearity** which should prompt some change in the structural form of the model

Residual Plots

- Plot of residual against regressor
  - Plot x_{ij} vs $e_i \forall j$
  - In previous plot, replace 'fitted' by x_{ij} to find similar interpretation.
  - Repeat it for all the regressors.

- Partial Regression Plot:
  - Complete/correct marginal effect
  - Let we want to study the marginal effect of x_j on y
  - y is regressed on $x_1, ..., x_k$ except x_j.
  - x_j is regressed on $x_1, ..., x_k$ except x_j.
  - Plot y residual, $e_i(y|x_1, ..., x_{j-1}, x_{j+1}, ..., x_k)$ against x_j residual, $e_i(x_j|x_1, ..., x_{j-1}, x_{j+1}, ..., x_k)$
  - For ideal scenario, the partial regression plot should show a linear relationship (straight line with non-zero slop).
  - Curvilinear band: indicates higher order terms in x_j or it's transformation.
  - Horizontal band: indicates no additional useful information in x_j