Model diagnostics and remedies. I

Miaoyan Wang

Department of Statistics
UW Madison

Model Assumptions

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

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

Robustness of Model Assumptions

Departure	$\hat{\beta}/\hat{\mu}_h$	s.e.	$\hat{Y}_{h(new)}$	s.e.
Linearity	S	S	S	S
Equal variance	R	S	R	S
Independence	R	S	R	S
Normality	R	R	R	S
Outliers	S	S	S	S

S = sensitive; R = robust.

Model Diagnostics

- Correct inference hinges on model assumptions.
- **Model diagnostics** are to evaluate the model assumptions and determine how reasonably they are met.
- A main idea for model diagnostics is to examine the residuals.
- Consider graphical approaches: Subjective but informative.

Graphical Techniques

- Exploratory data analysis (EDA).
  - Exploration of X and Y.
  - May not be as effective for model diagnostics.
- Recall for $i = 1, \ldots, n$
  - the ith fitted value: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - the ith residual: $e_i = Y_i - \hat{Y}_i$

What does e_i estimate/predict:

$$\varepsilon_i = Y_i - \mathbb{E}(Y_i) \sim_{\text{i.i.d}} N(0, \sigma^2)$$

Properties of Residuals

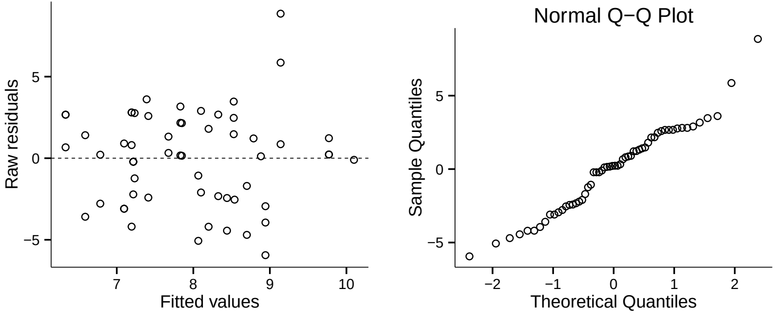
- Sample mean: $\bar{e} = 0$.
  Why?
  $$\bar{e} = \frac{\sum_{i=1}^{n} e_i}{n} = 0.$$
- Residual variance estimate $\hat{\sigma}^2$.
  Why?
  $$\text{MSE} = \frac{\text{SSE}}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \hat{\sigma}^2.$$
- Dependence (HW2)
  Why?
  $$\sum_{i=1}^{n} e_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} X_i e_i = 0.$$

Residual Plots

- Departures from model assumptions can be difficult to detect directly from X and Y.
- Thus consider residual plots.
  - Plot e_i against X_i.
  - Plot $|e_i|$ against X_i.
  - Plot e_i^2 against X_i.
  - Plot e_i against $\hat{Y}_i$.
  - Plot e_i against time.
  - Box plot of e_i.
  - Normal QQ plot of e_i.

Example: Wetland Species Richness

Raw residuals:



Types of Residuals

- **Raw residual** (or, **ordinary least squares residual)**:

$$e_i = Y_i - \hat{Y}_i.$$

- **standardized residual**:

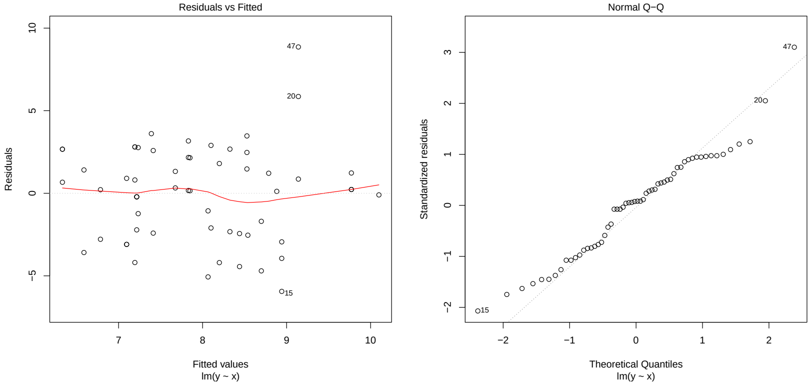
$$r_i = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}\sqrt{1 - p_{ii}}}, \quad \text{where} \quad p_{ii} \text{ is the } (i, i)\text{-th value of "hat matrix" } \boldsymbol{H}.$$

where $\hat{\sigma}^2 = \text{MSE}$ based on the entire sample. Why?

$$\begin{aligned}
\text{Var}(\boldsymbol{e}) &= \text{Var}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \text{Var}(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) \\
&= \text{Var}(\boldsymbol{Y} - \boldsymbol{X}\underbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}}_{\hat{\beta}}) \\
&= \text{Var}\left(\underbrace{(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)}_{\text{non-random}}\boldsymbol{Y}\right) \\
&= \sigma^2(\boldsymbol{I} - \boldsymbol{H})
\end{aligned}$$

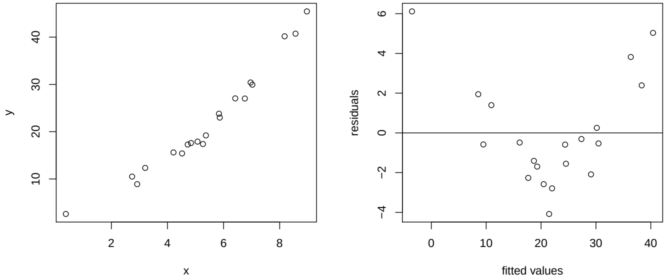
Example: Wetland Species Richness

Standardized residual



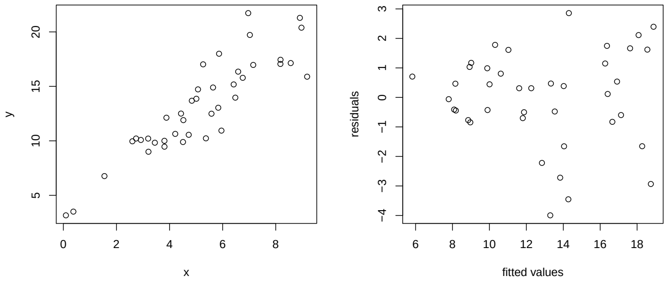
Nonlinearity of Regression Function

- Plot e_i against $\hat{Y}_i$ (or X_i).
- Random scatter indicates no serious departure from linearity.
- Example of departure from linearity:
  Curved relationship.



Non-equal Error Variance

- Plot e_i against $\hat{Y}_i$ (or X_i).
- Plot $|e_i|$ against $\hat{Y}_i$ (or X_i).
- Plot e_i^2 against $\hat{Y}_i$ (or X_i).
- Random scatter indicates no serious departure from constant variance.
- Example of departure from constant variance: Megaphone/funnel shape.



Nonindependence of Error Terms

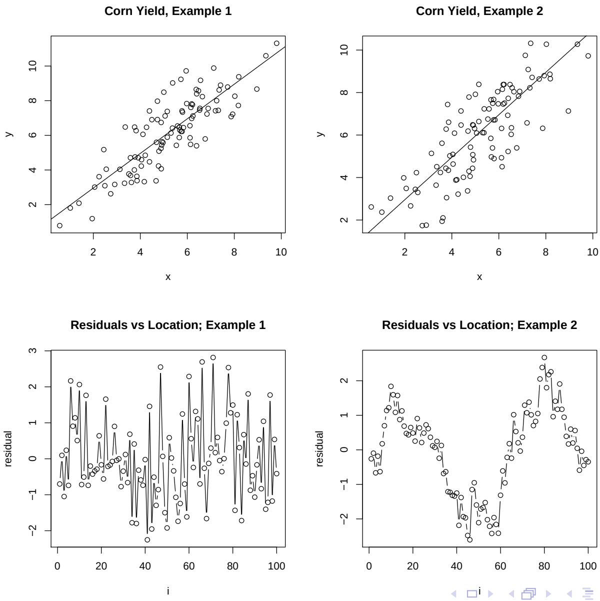
- Possible forms of nonindependence.
  - Observations collected over time and/or across space.
  - Study done on sets of siblings.
- Example of departure from independence:
  - Trend effect
  - Cyclical non-independence

Examples: Corn Yield

For $i = 1, \ldots, n$,

- $i =$ the index of the patch planted to corn.
- Patches are arranged in a long line at the edge of a field.
- $X_i =$ the amount of fertilizer applied to the ith patch.
- $Y_i =$ the corn yield in the ith patch.
- Plot e_i against location i.

Examples: Corn Yield



Nonnormality of Error Terms

Assess whether the residuals $\{e_i\}$ follow from normal.

- Box plot, histogram of e_i.
- Normal QQ plot: compared sorted residuals $e_{(1)}, \ldots, e_{(n)}$ to quantiles from standard normal $N(0, 1)$.
- If the residuals are approximately normal, the normal QQ plot should be approximately linear.
- It is a good idea to examine other departures first.
- Other departure affects the distribution, e.g., distribution of $\{e_i\}$ is subject to independence assumption especially in small sample size

Presence of Outliers

- An outlier refers to an extreme observation.
- Box plot, histogram plot of $\{e_i\}$.
- Plot e_i against $\hat{Y}_i$ (or X_i).
- Random scatter indicates absence of outliers.
- Outliers may convey important information. An error. A different mechanism is at work. A significant discovery.

Graphical Techniques: Remarks

- We generally do not plot residuals (e_i) against response (Y_i). Why not?
- Residual plots may provide evidence against model assumptions, but do not generally validate assumptions.
- For data analysis in practice: Fit model and check model assumptions (an iterative process).
- For this class, please include representative residual plots in homework assignments and reports.
- As much art as science. No golden rules. No magic formulas. Decision may be difficult for small sample size.