Chapter 3 Diagnostics and Remedial Measures

3.2, 3.3 Diagnostics for Residuals

The residuals are defined as

\[ e_i = y_i - \hat{y}_i = y_i - b_0 - b_1x_{i1} - \cdots - b_{p-1}x_{ip-1}, \quad i = 1, 2, \ldots, n. \]

If the least square estimate is a sensible estimate, \( e_i \) can be regarded as the “estimate” of \( \varepsilon_i \),

\[ e_i = \beta_0 + \beta_1x_{i1} + \cdots + \beta_{p-1}x_{ip-1} + \varepsilon_i - b_0 - b_1x_{i1} - \cdots - b_{p-1}x_{ip-1} \]

\[ \approx \varepsilon_i \]

Thus, the residuals should exhibit tendencies that tend to confirm the assumptions about \( \varepsilon_i \), such as independence, having zero mean, having a constant variance \( \sigma^2 \), and following a normal distribution. On the other hand, as the residuals do not exhibit tendencies, this might imply the assumptions might be violated. Usually, we can plot the residuals and then examine the tendencies. The principal ways of plotting residuals \( e_i \) are

I. To check for non-normality.
II. To check for time effects if the time order of the data is known.
III. To check for nonconstant variance.
IV. To check for curvature of higher order than fitted in the X’s

I. Non-Normality Checks on Residuals:

A simple histogram or a stem and leaf plot of the residuals can be used to check the normality. If these residuals are bell-shaped distributed, this might imply the normality assumption might not be violated.

Histogram:
Stem-and leaf plot:
-2 : 2
-1 : 9755
-1 : 33322111110
-0 : 998887776665555
-0 : 44444333333332210
 0 : 0111122222234444
 0 : 5666677778889999
 1 : 011223
 1 : 5556788
 2 : 01
 2 : 89

The better method to check normality is the normal probability plot. The residuals and their corresponding normal scores are plotted. If the plot look straight, then the normality assumption might not be violated.

II, III, IV:
The 3 properties can be checked by plotting the residuals versus
(a) fitted value $\hat{y}_i$
(b) the covariate $x_{ij}$
The typical satisfactory plots are as follows:
Three typical unsatisfactory plots are as follows:

(i)

(ii)
(iii)

(a) versus fitted value $\hat{y}_i$:

Why $\hat{y}_i$, but not $y_i$:

$$\Rightarrow \quad r_{ey} = \text{correlation coefficient of } e \text{ and } y = \left\{1 - R^2\right\}^{1/2}$$

$$r_{ey} = 0$$

Therefore, $e_i's$ and $y_i's$ are correlated. If we plot $e_i's$ versus $y_i's$, then a linear
trend might be due to the positive correlation of $e_i$'s and $y_i$'s rather than the violation of the assumptions about the random error. On the other hand, since $e_i$'s and $\hat{y}_i$'s are uncorrelated, an unsatisfactory residual plot might do imply the violation of the assumptions.

**Derivation of** $r_{e\hat{y}} = 0$:

Let

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i \quad \text{and} \quad \bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i.$$  

Since

$$\sum_{i=1}^{n} e_i \bar{\hat{y}} = \bar{\hat{y}} \sum_{i=1}^{n} e_i = 0$$

and

$$X'e = X'(y - \bar{\hat{y}}) = X'(y - Xb) = X'y - X'Xb$$

$$= X'y - X'X(X'X)^{-1}X'y = X'y - X'y = 0$$

then

$$S_{e\hat{y}} = \sum_{i=1}^{n} (e_i - \bar{e})(\hat{y}_i - \bar{\hat{y}}) = \sum_{i=1}^{n} e_i (\hat{y}_i - \bar{\hat{y}}) = \sum_{i=1}^{n} e_i \hat{y}_i \quad \left(\text{since } \sum_{i=1}^{n} e_i \bar{\hat{y}} = 0\right)$$

$$= \hat{y}'e = (Xb)'e = b'X'e = 0 \quad \left(\text{since } X'e = 0\right)$$

Thus,

$$r_{e\hat{y}} = \frac{S_{e\hat{y}}}{(S_{ee}S_{\hat{y}\hat{y}})^{\frac{1}{2}}} = \frac{\sum_{i=1}^{n} (e_i - \bar{e})(\hat{y}_i - \bar{\hat{y}})}{\left[\sum_{i=1}^{n} (e_i - \bar{e})^2 \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2\right]^{\frac{1}{2}}} = 0$$

**Derivation of** $r_{e\hat{y}} = \left(1 - R^2\right)^{\frac{1}{2}}$:

Since
\[
S_{ey} = \sum_{i=1}^{n} (e_i - \bar{e})(y_i - \bar{y}) = \sum_{i=1}^{n} e_i (y_i - \bar{y}) = \sum_{i=1}^{n} e_i y_i \quad \text{(since } \sum_{i=1}^{n} e_i \bar{y} = \bar{y} \sum_{i=1}^{n} e_i = 0) \\
\sum_{i=1}^{n} e_i (y_i - \hat{y}_i) \quad \text{(since } \sum_{i=1}^{n} e_i \hat{y}_i = 0) \\
= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

thus,

\[
r_{ey} = \frac{S_{ey}}{(S_{ee} S_{yy})^{1/2}} = \frac{\sum_{i=1}^{n} (e_i - \bar{e})(y_i - \bar{y})}{\left[ \sum_{i=1}^{n} (e_i - \bar{e})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 \right]^{1/2}} \\
= \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\left[ \sum_{i=1}^{n} y_i^2 - \bar{y}^2 \right]^{1/2}} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\left[ \sum_{i=1}^{n} (y_i - \bar{y})^2 \right]^{1/2}} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = (1 - R^2)^{1/2}
\]

\[
\left( \text{since } \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \right)
\]

- The residual plot (i): nonconstant variance.

Remedy: use weighted least square or transformation of \( Y_i \).

Note:

Transformation of the response \( Y_i \) can stabilize the variance (make
the variance of the transformed $Y_i$ constant).

- **The residual plot (ii): errors in analysis, some extra terms highly correlated to the other terms are missing or wrongful omission of $\beta_0$.**

  [Justification:]
  Suppose the postulated model is
  \[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i,
  \]
  and the true model is
  \[
  Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + \beta_p x_{ip} + \varepsilon_i,
  \]
  \[i = 1, 2, \ldots, n.\]
  Then,
  \[
  \hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_{p-1} x_{ip-1}
  \]
  and
  \[
  e_i \approx \beta_p x_{ip} + \varepsilon_i.
  \]
  If $\beta_p > 0$ and $x_{ip}$ is correlated to $\hat{y}_i$ (since $x_{ip}$ is correlated to some covariates in the postulated model), then the residual plot will be like (ii).

  **Remedy:** rechecking the analysis process, adding some extra terms to the model or adding $\beta_0$ back to the model.

- **The residual plot (iii): some extra terms (might be second order terms of existing variables or some variables not in the model) are missing or unequal variance**

  **Remedy:** adding extra terms to the model or transformation of $Y_i$. 
(b) versus the covariate $x_j$:

- **The residual plot (i): nonconstant variance.**

Remedy: use weighted least square or transformation of $Y_i$.

- **The residual plot (ii): errors in analysis or some extra terms highly correlated to $X_j$ are missing.**

Suppose the true model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \beta_p z_i + \epsilon_i,$$

and the covariate $X_j$ is correlated to $Z$. Thus, heuristically,

$$e_i \approx \beta_p z_i + \epsilon_i.$$

As $\beta_p \geq 0$ and $X_j$ is positively correlated to $Z$, the residual plot would look like (ii).

Remedy: rechecking the analysis process or adding some extra terms to the model.

- **The residual plot (iii): some extra terms are missing or the variance of $y_i$ are not equal.**

Remedy: adding extra terms (second order terms of the existing variables or the terms not in the original model) to the model or transformation of $Y_i$

### 3.9 Transformation

**Box-Cox Transformation:**

Suppose the tentatively used model is $y = X\beta + \epsilon$. Two transformations can be used. The first transformation is
The above transformation is also called **Box-Cox transformation**.

**Note:** \( \lim_{\lambda \to 0} W_i(\lambda) = \lim_{\lambda \to 0} \frac{y_i^{\lambda} - 1}{\lambda} = \log(y_i) \).

The other transformation is to scale \( W_i(\lambda) \) by

\[
V_i(\lambda) = \frac{(y_i^{\lambda} - 1)}{\lambda \hat{y}^{\lambda - 1}}, \quad \lambda \neq 0
\]

\[\hat{y} \log(y_i), \quad \lambda = 0\]

where

\[
\hat{y} = \sqrt[n]{y_1 y_2 \cdots y_n} = \left(\prod_{i=1}^{n} y_i\right)^{1/n}
\]

is the geometric mean of \( y_1, y_2, \ldots, y_n \).

The model based on the transformed response is

\[
W(\lambda) = \begin{bmatrix}
W_1(\lambda) \\
W_2(\lambda) \\
\vdots \\
W_n(\lambda)
\end{bmatrix} = X\beta + \varepsilon
\]

or

\[
V(\lambda) = \begin{bmatrix}
V_1(\lambda) \\
V_2(\lambda) \\
\vdots \\
V_n(\lambda)
\end{bmatrix} = X\beta + \varepsilon.
\]

**Note:** \( V \) transformation is usually preferred!!

### 11.1 Weighted Least Squares

Suppose the linear regression model is
\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_i^2), i = 1, \ldots, n. \]

Let \( W_i = \frac{1}{\sigma_i^2} \) and

\[
W = \begin{bmatrix}
w_1 & 0 & \cdots & 0 \\
0 & w_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & w_n
\end{bmatrix}.
\]

To estimate \( \beta \), we need to minimize the \textit{weighted least square criterion},

\[
S_w(\beta) = S_w(\beta_0, \beta_1, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{ip-1})^2 = (y - X\beta)^t W (y - X\beta)
\]

The weighted least square estimate is

\[
b = \left[ X^t W X \right]^{-1} X^t W Y
\]