residuals are useful for detecting lack of fit and checking model assumptions

\( Y = X \beta + \varepsilon = \hat{Y} + \hat{\varepsilon} \)

1. leverage: if there is some underlying heteroscedasticity (i.e., violation of \( \text{var}(\epsilon) = \sigma^2 I \)), then \( \text{var}(\epsilon) \) is non-constant variance removed, so is covariance

\[ \text{cov}(\epsilon) = \Sigma \]

2. studentized residuals are preferred in residual plots, because they are made by S.-W. Cheng (NTHU, Taiwan)

\[ (\hat{\varepsilon}_i) = (1-h_i)^{1/2} \sigma \]

\[ \text{cor}(\epsilon_i, \hat{\varepsilon}_i) \approx 0 \]

\[ \sigma^2 I \]

\( \Sigma \)

\( \hat{\varepsilon}_i \)

\( \varepsilon_i \)

\( \hat{Y} \)

\( \hat{\varepsilon}_i \)

\( \Sigma \)

\( \sigma^2 I \)

\( \hat{\varepsilon}_i \)

\( \varepsilon_i \)

\( \hat{Y} \)

\( \hat{\varepsilon}_i \)

\( \Sigma \)

\( \sigma^2 I \)

\( \hat{\varepsilon}_i \)

\( \varepsilon_i \)

\( \hat{Y} \)

\( \hat{\varepsilon}_i \)
• An outlier is a point that does not fit the current model (Q: possible cause?)

⇒ Usually, large residual (Q: why?)

• Q: Is there a problem if (raw or studentized) residuals are used to detect outliers?

⇒ Outliers may affect the fit (see plot)

• Idea: Exclude the \( i \)th observation and re-compute the estimates to get $\hat{\beta}(i)$ and $\hat{\sigma}^2(i)$, where \((i)\) denotes that the \( i \)th case has been excluded. Then, consider $\hat{\beta}(i)$ & $\hat{\sigma}^2(i)$ not influenced by the \( i \)th observation.

... (continued)
• **Q**: What should be done if some observations are identified as outliers?
  - check for a data entry error first
  - Examine the physical context (sometimes, outliers may have physical significance)
  - exclude the point from the analysis
    - try to re-include later if model changed
    - if exclude permanently, report
  - dangerous to exclude them in an automatic manner **e.g. NASA 1985**

  **Reading**: Faraway (2005, 1^st^ ed.), 4.2.2  **Further reading**: D&S, 8.1

• Each observations have different influence/contribution to the fitted model. Our fitted model should not change too much (i.e., robust) just because of adding/dropping a specific observation.

• **influential point**: one whose removal from data would cause large change in the fit.

• an influential point *may or may not* be an outlier and *may or may not* have large leverage but it will tend to have at least one of those two properties.

• **measures of influence** (**Q**: how to numerically characterize “large change in fit”?)  p. 7-7

  - change in coefficients: \( \hat{\beta} - \beta^{(i)} \)
  - change in fit: \( \hat{Y} - \hat{Y}^{(i)} \)

  **Cook’s statistics/distances** (scale and unit free):

  \[
  D_i = \frac{(\hat{Y} - \hat{Y}^{(i)})^T(X^T X)(\hat{\beta} - \beta^{(i)})}{(p \hat{\sigma}^2)} = \frac{(\hat{Y} - \hat{Y}^{(i)})^T(\hat{Y} - \hat{Y}^{(i)})}{(1/p) r_i^2 (h_i/(1-h_i))}
  \]

  ⇒ it's a combination of residual and leverage. (**Q**: what are the effects of residual and leverage on Cook’s statistic?)

  ⇒ **Q**: how large is large? If assume \( X \) is multivariate Normal, can do a test on \( D_i \). However, normality may not be a reasonable assumption in practice.

  Others: DFFITS, Atkinson's modified Cook's statistics

• **Residual plots** (**Q**: how large is large?)

  - residual plots: plot residuals (or absolute values of residuals) against (i) \( \hat{y}_i \), (ii) \( x_k \) (for predictors in model and not in model), (iii) combination (or transformation) of \( x_k \)’s, (iv) time order (if available), (v) any other quantities relevant to residuals
  - **Q**: why draw residual plots?

  **Reading**: Faraway (2005, 1^st^ ed.), 4.2.3  **Further reading**: D&S, 8.3, 8.4

  **Residual plots**

  \[
  y_i - \frac{X_i \beta}{(\sigma^2_x, \mu_x)} \sim N(\mu_x, \sigma_x^2)
  \]

  **Residual plots**

  better to 1) remove outliers & influential obs 2) use standardized residuals

  made by S.-W. Cheng (NTHU, Taiwan)
• in residual plots,
  ➢ find overall patterns from the shape of all points (cf., residuals used in checking outliers or influential points ⇒ identifying individually unusual point)
  ➢ check assumptions: (i) non-constant variance; (ii) incorrectly specified mean structure (i.e., $E(Y) = \mathbf{X}\beta \neq \text{too simple (lack of fit)}$
  ➢ rather subjective
• a satisfactory residual plot (null plot)
  ➢ constant variance $\checkmark$
  ➢ no curvature in the mean of residuals $\checkmark$

Note: one satisfactory residual plot cannot guarantee the residual plots for other variables will be satisfactory
• some possible patterns in unsatisfactory residual plots:
  ➢ evidence of non-constant variance
  ➢ curvature in the mean of residuals $\Rightarrow$ evidence of incorrectly specified mean structure
  ➢ evidence of non-constant variance and incorrectly specified mean structure

➤ unfortunately, in real data set, it’s rare the pattern is so clear (Q: what will you conclude from the residual plot on the right?)
  ➢ in models with many terms or models with complex non-linear mean structure, cannot necessarily associate shapes in a residual plot with a particular problem with the assumptions, e.g.,
    true model: $E(Y) = |x_1|/(2 + (1.5 + x_2)^2)$ with constant variance
    fitted model: $E(Y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

• possible remedies for unsatisfactory residual plots

<table>
<thead>
<tr>
<th>unsatisfactory residual plot</th>
<th>plot residuals against ...</th>
<th>$\hat{y}$</th>
<th>$x_k$</th>
<th>time order</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-constant variance</td>
<td>1. weighted least square</td>
<td>1. weighted least square</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. transform $y$</td>
<td>2. transform $y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>curvature in mean structure</td>
<td>1. add extra term</td>
<td>1. add extra term of $x_k$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. transform $y$</td>
<td>2. transform $y$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

➤ Reading: E, 4.1.1
➤ Further reading: D&S, 2.5

Non-constant variance
• if not sure, plot absolute values of residuals against $\hat{y}$, $x_k$’s, time order
• when non-constant variance exists, $\hat{\beta}_{\text{GLS}}$ will be more variable than the best estimates ($\hat{\beta}_{\text{OLS}}$ unbiased but not BLUE) and $\hat{\sigma}$ wrong ($\Rightarrow$ test and C.I. inaccurate)