Summary for Model Adequacy Checking

Methods for scaling residuals

- **Major assumptions**:
  1. The relationship between the response y and the regressors xs are linear, at least approximately;
  2. The error term ε has zero mean, i.e. $E(\varepsilon_i) = 0$, $i = 1, \ldots, n$;
  3. The error term ε has constant variance σ^2, i.e. $Var(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$;
  4. The errors are uncorrelated, i.e. $Cov(\varepsilon_i, \varepsilon_j) = 0$ $(i \neq j)$, $i, j = 1, \ldots, n$;
  5. The errors are normally distributed, i.e. $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$.
- **Residuals**: $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$.
  View:
  - A residual may be viewed as the **deviation** between the **data** and the **fit**;
  - It is also a measure of the variability in the response variable y not explained by the regression model;
  - It is also convenient to think of the residuals as the realized or observed values of the model errors.

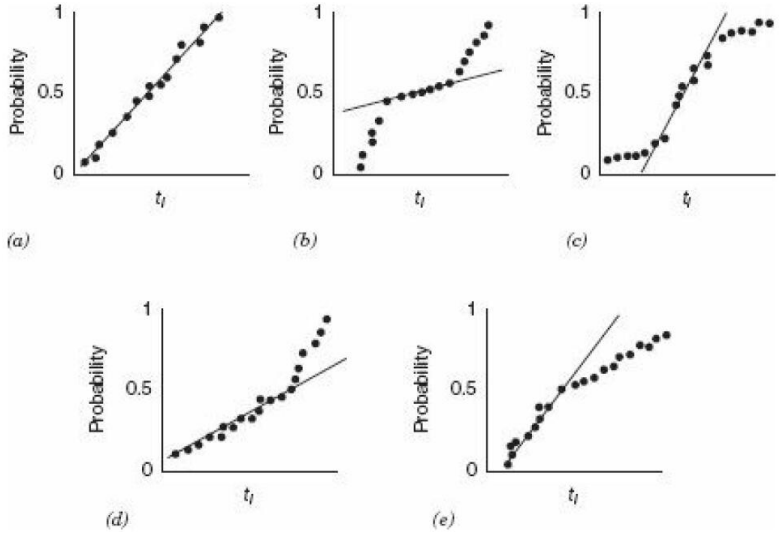
  Properties:
  - $\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$, $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$, $\sum_{i=1}^{n} x_i e_i = 0$, $\sum_{i=1}^{n} \hat{y}_i e_i = 0$
  - $E(e_i) = 0$, $i = 1, \ldots, n$; Unbiased estimator of σ^2: $\frac{\sum_{i=1}^{n}(e_i - \hat{e})^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_{Res}}{n-p} = MS_{Res}$;
  - The residuals are **NOT** independent.
- **Methods for scaling residuals**:
  - **Standardized residuals** d_i
  - **Studentized residuals** r_i
  - **PRESS residuals** e_i
  - R**-student** t_i

  Standardization procedure:
  - $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}$, $Var(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$;
  - $E(e_i) = 0$, $Var(e_i) = \sigma^2(1 - h_{ii})$ and $Cov(e_i, e_j) = -\sigma^2 h_{ij}$;
  - $\frac{e_i - E(e_i)}{\sqrt{Var(e_i)}} = \frac{e_i - 0}{\sqrt{\sigma^2(1-h_{ii})}}$

SCALED RESIDUAL	REMARK
$d_i = \frac{e_i}{\sqrt{MS_{Res}}},\ i = 1, \ldots, n;$	– $E(d_i) = 0$, $Var(d_i) \approx 1$; – A large standardized residual ($d_i > 3$, say) potentially indicates an outlier.
$r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}},$ $i = 1, \ldots, n$	– If there is only one regressor, the studentized residuals: $r_i = \frac{e_i}{\sqrt{MS_{Res}\left[1-\left(\frac{1}{n}+\frac{(x_i-\hat{x})^2}{S_{xx}}\right)\right]}},\ i = 1, \ldots, n.$
$e_{(i)} = y_i - \hat{y}_{(i)},\ i = 1, \ldots, n$	– They are also called the **prediction errors** or **deleted residuals**; – The ith PRESS residual is $e_{(i)} = \frac{e_i}{1-h_{ii}},\ i = 1, \ldots, n;$ – The variance of the ith PRESS residual is $Var(\varepsilon_{ii}) = \frac{\sigma^2}{1-h_{ii}}$, so that a **standardized PRESS residual** is $\frac{e_{(i)}}{\sqrt{Var[e_{(i)}]}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$ If we use MS_{Res} to estimate σ^2, is just the **studentized residual**.
$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}},\ i = 1, \ldots, n$	– It is also called **externally studentized residual**. – $S^2_{(i)} = \frac{(n-p)MS_{Res} - e_i^2/(1-h_{ii})}{n-p-1}$; – $t_i \sim t_{n-p-1}$.

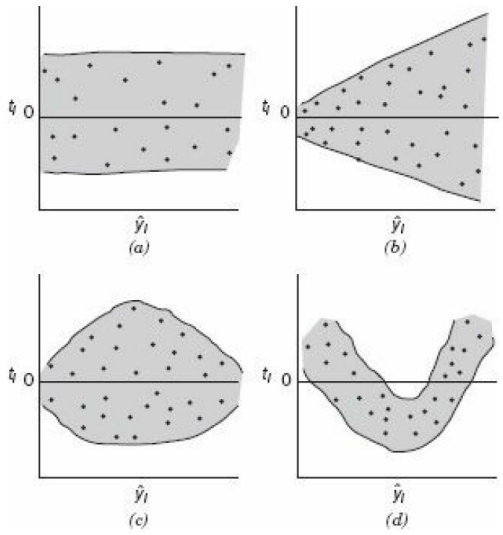
Residual plots

- **Normal probability plot**:



  – (a) "idealized" normal probability plot: approximately along a straight line;
  – (b) light-tailed: a sharp upward and downward curve at both extremes;
  – (c) heavy-tailed: flattening at the extremes;
  – (d) positive skew;
  – (e) negative skew.

- **Plot of residuals against the fitted values**:



  - (a) horizontal band: there are no obvious model defects.
  - (b) outward-opening funnel: the variance of the errors is not constant, and is an increasing function of y;
  - (c) double-bow: the variance of the errors is not constant, and often occurs when y is a proportion between 0 and 1;
  - (d) curve: nonlinearity, which means that other regressor variables/higher order regressor variables are needed in the model.