Chapter 10. Assessing model adequacy

1. The major assumptions that we have used so far in the regression analysis:

   (1). The relationship between *y* and *x* is linear

   (2). The error term ε has zero mean

   (3). The error term ε has constant variance σ^2 (identical)

   (4). The errors are uncorrelated (independent)

   (5). The errors are normally distributed

   In other words, ε is iid $N(0, \sigma^2)$. In this chapter we will check if these assumptions are met given a data set.

2. Residuals (Deviation between the *observation* and the *fit*)

   $\varepsilon_i = y_i - \hat{y}_i, \qquad i = 1, 2, \ldots, n$

   It is a measure of the variability not explained by the regression model.

3. Properties of the residuals

   It has zero mean: $\bar{\varepsilon} = \dfrac{\sum \varepsilon_i}{n} = \dfrac{\sum (y_i - \hat{y}_i)}{n} = 0$

   Its approximate variance: $V(\varepsilon_i) = \dfrac{\sum (\varepsilon_i - \bar{\varepsilon})^2}{n-2} = \dfrac{\sum \varepsilon_i^2}{n-2} = \dfrac{SS_E}{n-2} = MS_E$

   Its exact variance: $V(\varepsilon_i) = V(y_i - \hat{y}_i) = \sigma^2 \left(1 - \dfrac{1}{n} - \dfrac{(x_i - \bar{x})^2}{S_{xx}} \right)$

4. Standardized residuals and outliers:

$$d_i = \frac{\varepsilon_i}{\sqrt{MS_E}}$$

The studentized residual is the residual divided by the exact variance $V(\varepsilon_i)$

$$r_i = \frac{\varepsilon_i}{\sqrt{MS_E\left(1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{S_{xx}}\right)}}, \qquad i = 1, 2, \ldots, n$$

An observation is considered as an outlier if its y_i has $r_i > 2$ or $r_i < -2$.

5. Q-Q plot

(1). Rank the residuals ε_i in ascending order: $\varepsilon_{(1)}, \varepsilon_{(2)}, \ldots, \varepsilon_{(n)}$

R-code: **z=sort(e)**

(2). Compute the corresponding cumulative probability: $P_i = (i\text{-}0.5)/n$

(3). Compute the expected normal quantile corresponding to P_i: $q_i = \Phi^{-1}(P_i)$

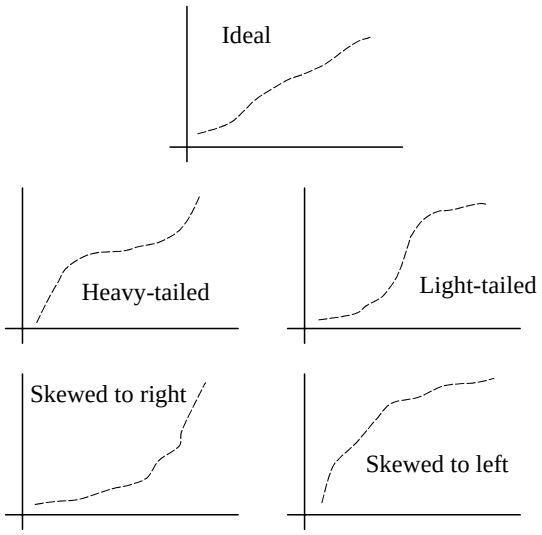
R-code: **qnorm(0.2)** if $P_i = 0.2$ whose quantile = -0.8416212.

(4). Plot q_i versus z

Using R, these steps can simply be implemented as **qqnorm(hl.lm$resid)**

Further, put a 1-1 line: **qqline(hl.lm$resid)**

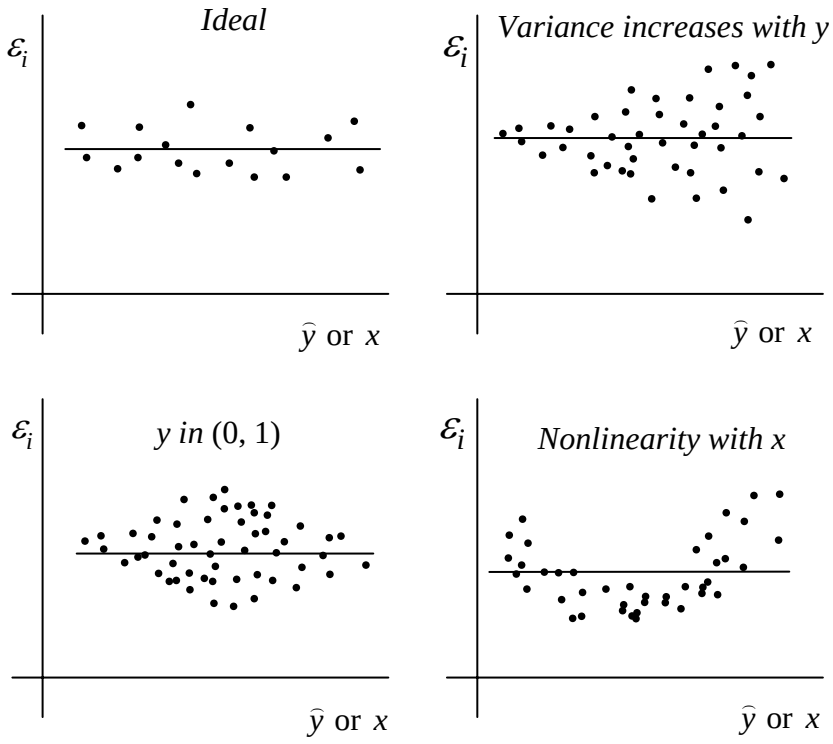
6. Interpretation of Q-Q plots:



7. Plots of residuals versus $\hat{y}$ (not y) or x

R-code: **plot(hl.lm$fitted.value, hl.lm$resid)** # residuals versus yhat

Or **plot(dbh, hl.lm$resid)** # residuals versus x



8. Cook's distance (diagnosing influential points):

   It measures how far an observation y_i is from the rest of data. (It is not a method to identify outliers but to identify influential points.)

$$D_i = \frac{\sum(\hat{y}_{-i} - \hat{y})}{(k+1)MS_E}$$

   where $\hat{y}$ is the estimate from all data, while $\hat{y}_{-i}$ is the estimate by deleting the i^{th} data point y_i. k is the number of regressors, in simple linear regression $k = 1$.

   R-code: The above assessments can be done in one shot using **plot.lm(lm.out)**

9. Variance-stabilizing transformations

Relationship of σ^2 to $E(y)$	**Transformation**
$\sigma^2 \propto consant$	No transformation
$\sigma^2 \propto E(y)$	$\sqrt{y}$ (Poisson data)
$\sigma^2 \propto E(y)[1-E(y)]$	$\sin^{-1}(\sqrt{y})$ (binomial data: $0 \le y \le 1$)
$\sigma^2 \propto [E(y)]^2$	log(y)
$\sigma^2 \propto [E(y)]^3$	$y^{-1/2}$
$\sigma^2 \propto [E(y)]^4$	$\frac{1}{y}$

10. Exercise: Model *htb ~ dbh*

    (1). Without transformation: **htb.lm=lm(htb~dbh,data=hl.dat)**

    Examine plots: **plot(htb.lm)**

    (2). With transformation: **loghtb.lm=lm(log(htb)~log(dbh),data=hl.dat)**

    Examine plots: **plot(loghtb.lm)**