

回归分析


回归诊断

研究中国人 身高和体重是否呈现 线性关系：

如果是能否建出 回归模型 体重为Y变量，身高是X变量

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

$$\begin{aligned}\epsilon &\sim \text{Normal}(0, \sigma^2) \\ \text{cor}(\epsilon_i, \epsilon_j) &= 0\end{aligned}$$


$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\epsilon}$$

残差图： $\varepsilon \sim \text{Normal}(0, \sigma^2)$
 $\text{cor}(\varepsilon_i, \varepsilon_j) = 0$

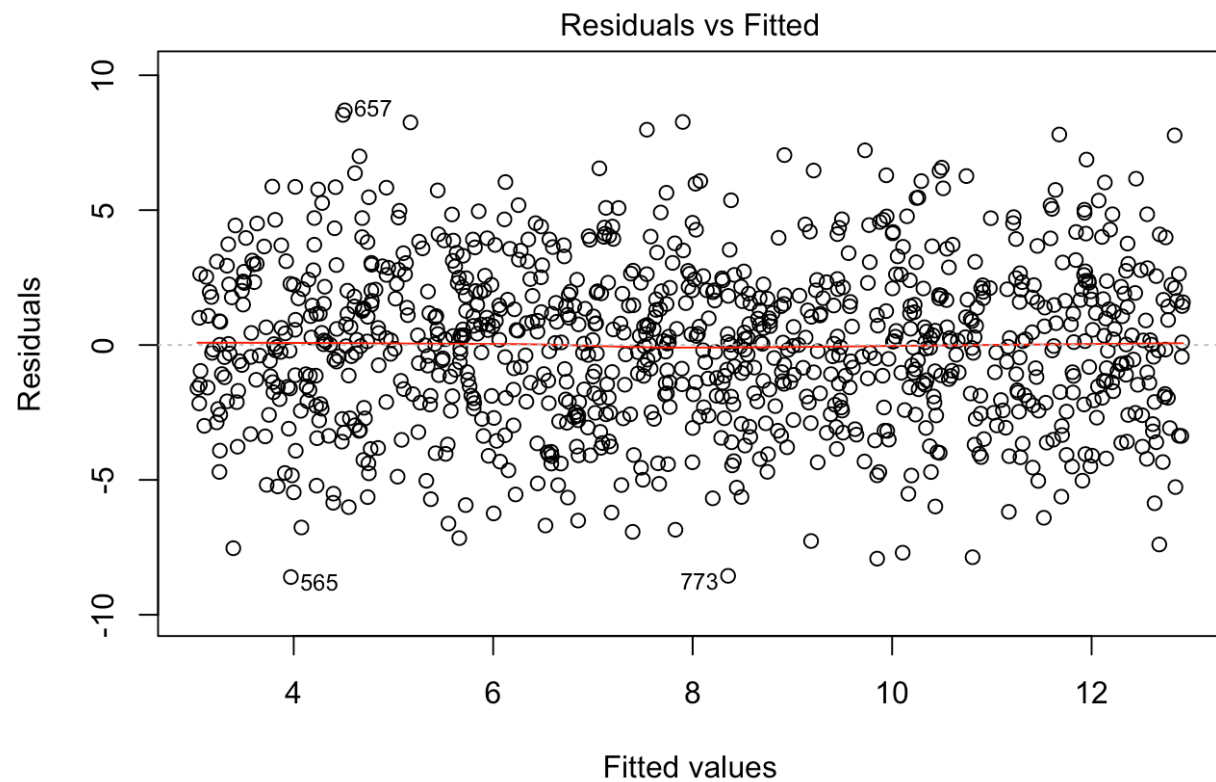
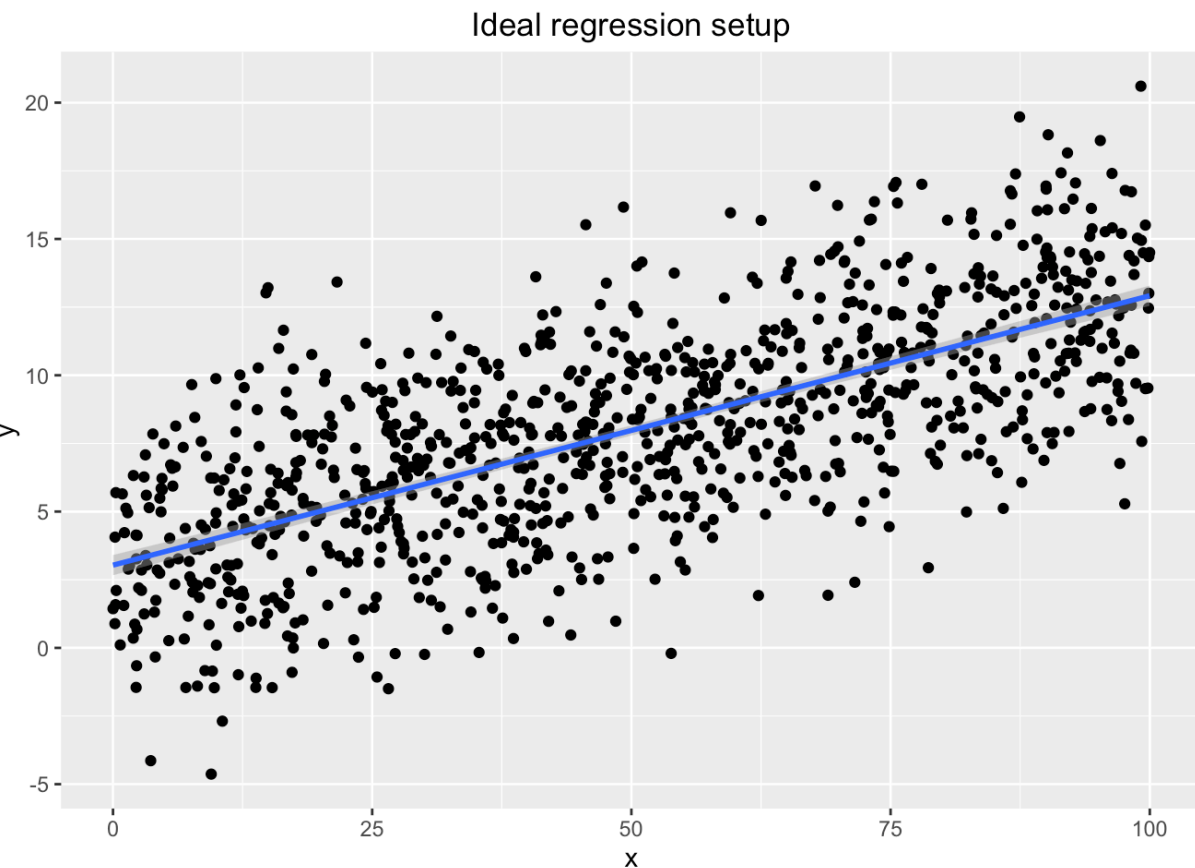
样本预测值 \hat{y}_i 为x轴， $\hat{\varepsilon}_i$ 为y轴 画的散点图

必须满足的三个性质：

随机性、正态性、等方差性

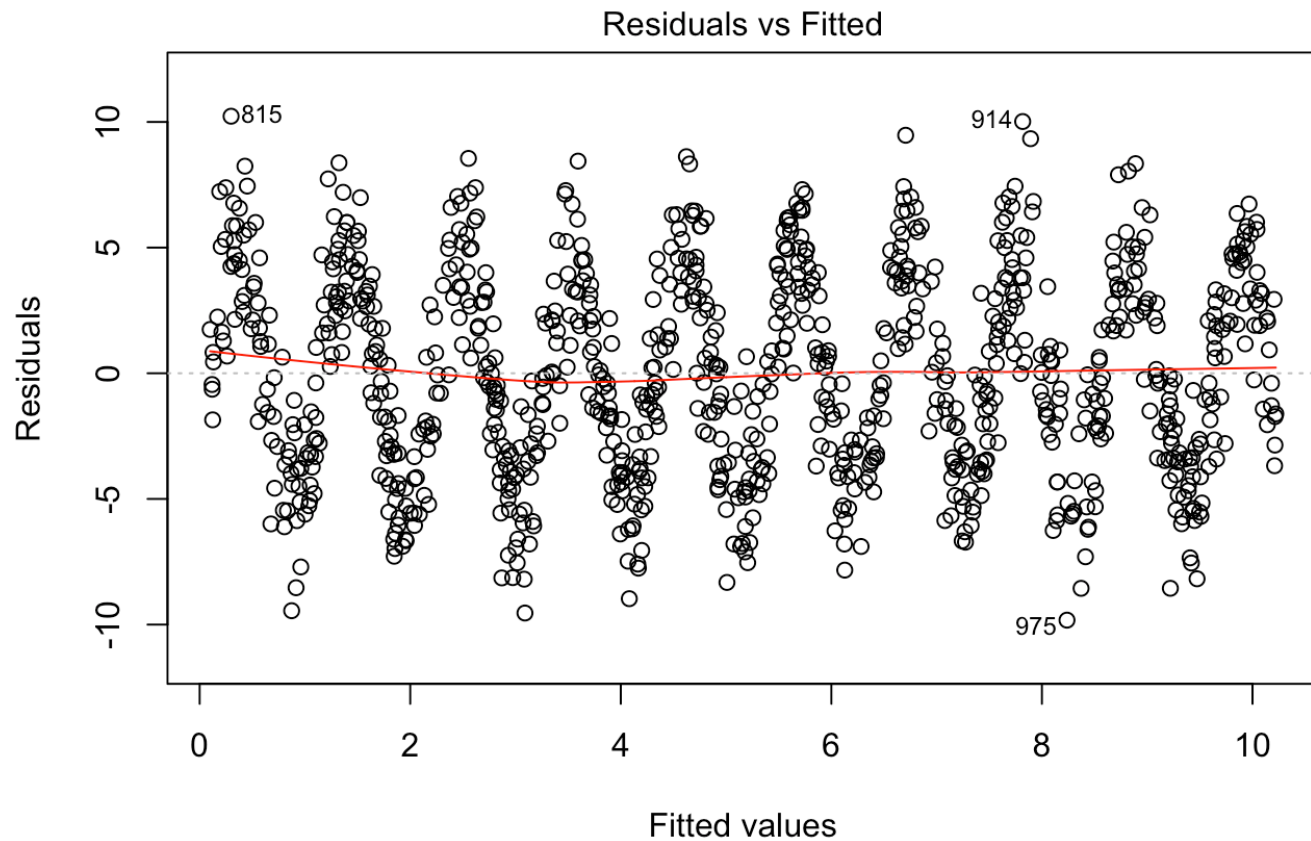
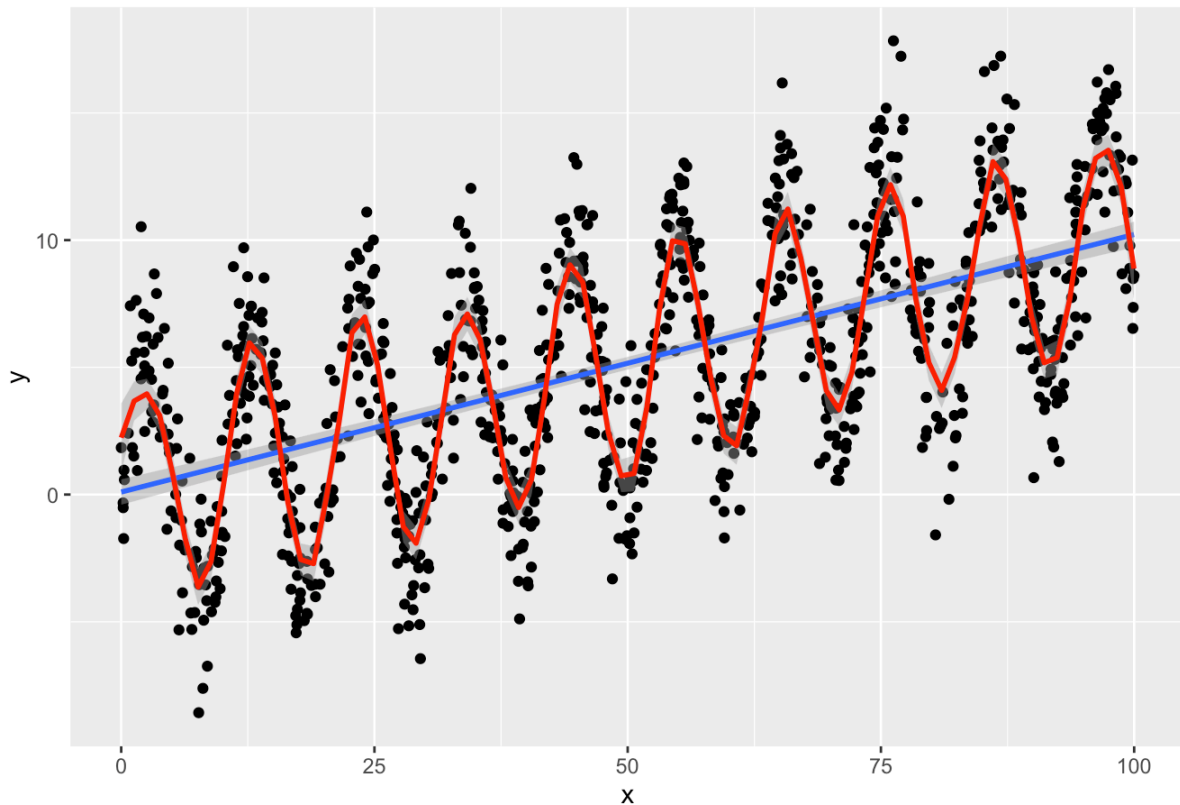
残差图： $\varepsilon \sim \text{Normal}(0, \sigma^2)$
 $\text{cor}(\varepsilon_i, \varepsilon_j) = 0$

样本预测值 \hat{y}_i 为x轴， $\hat{\varepsilon}_i$ 为y轴画的散点图



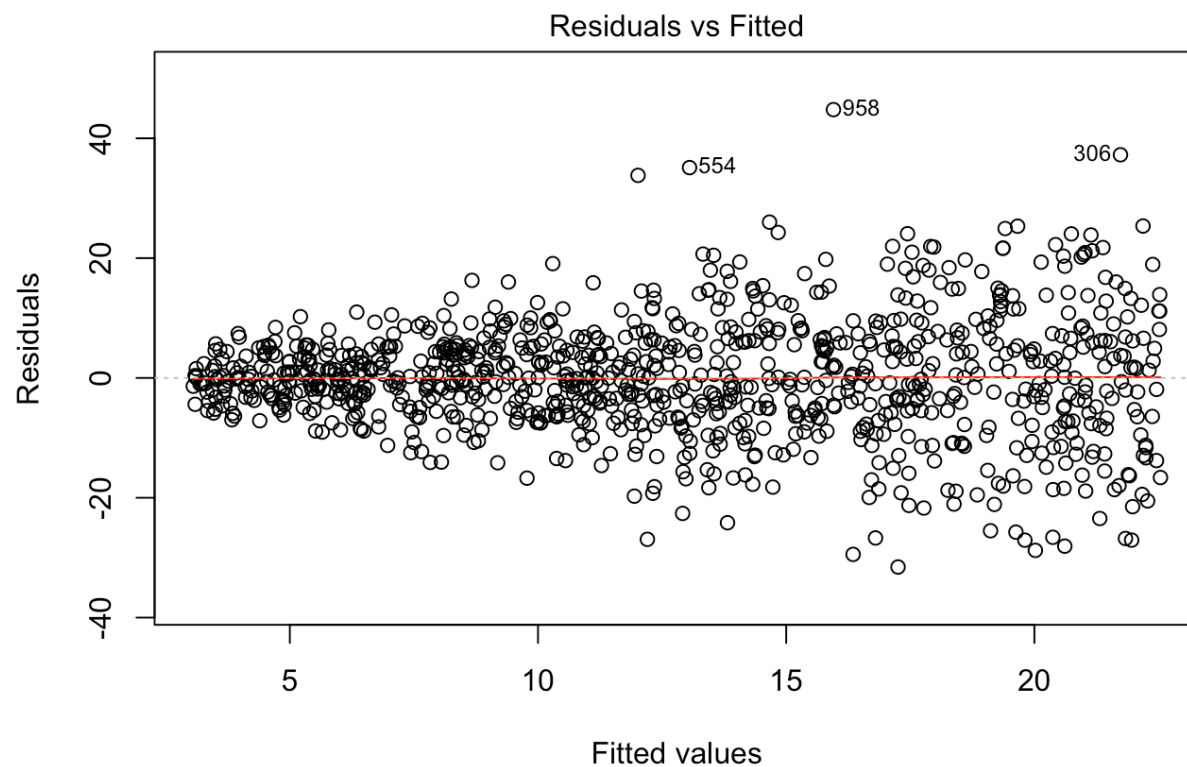
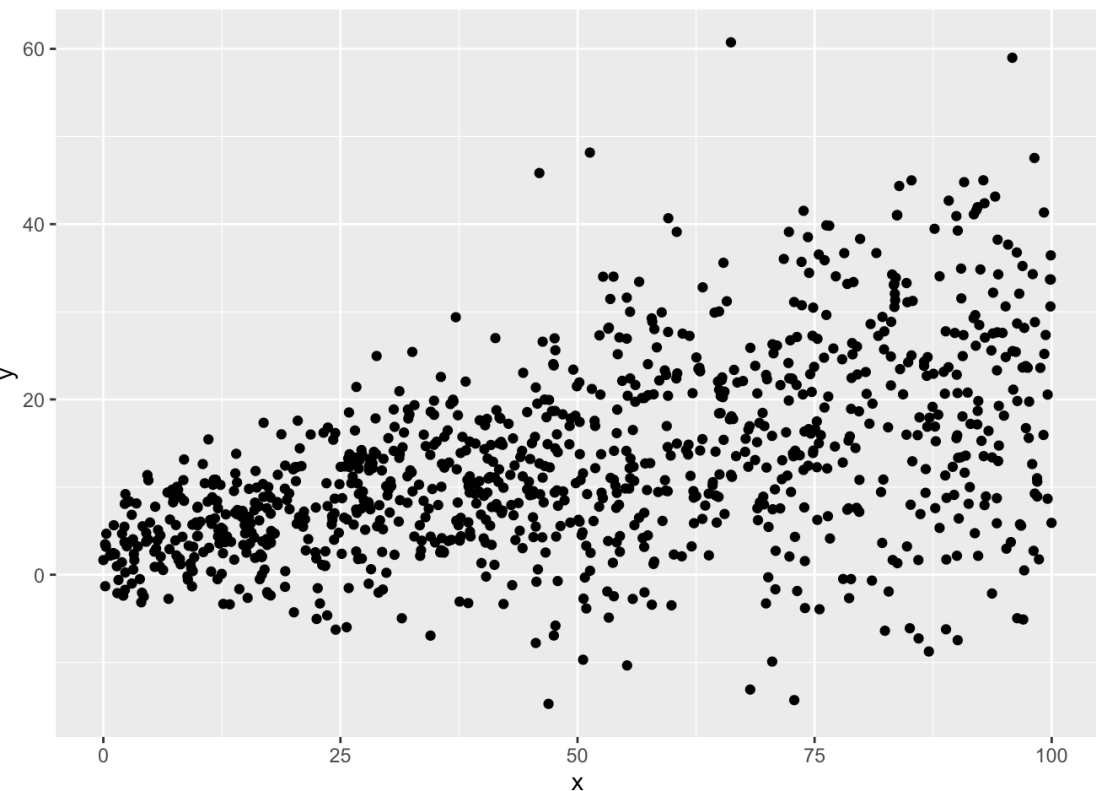
残差图： $\varepsilon \sim \text{Normal}(0, \sigma^2)$
 $\text{cor}(\varepsilon_i, \varepsilon_j) = 0$

必须满足的三个性质：随机性、正态性、等方差性



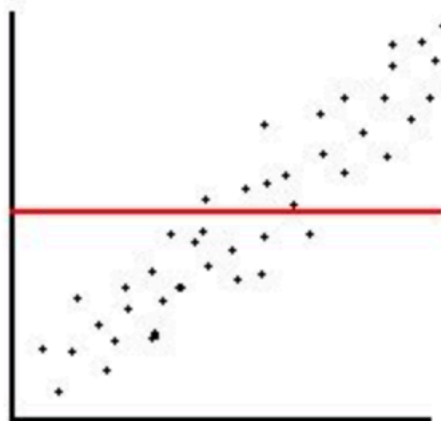
残差图： $\varepsilon \sim \text{Normal}(0, \sigma^2)$
 $\text{cor}(\varepsilon_i, \varepsilon_j) = 0$

必须满足的三个性质：随机性、正态性、等方差性

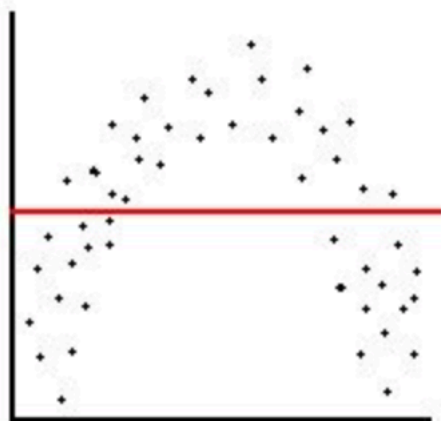


残差图：

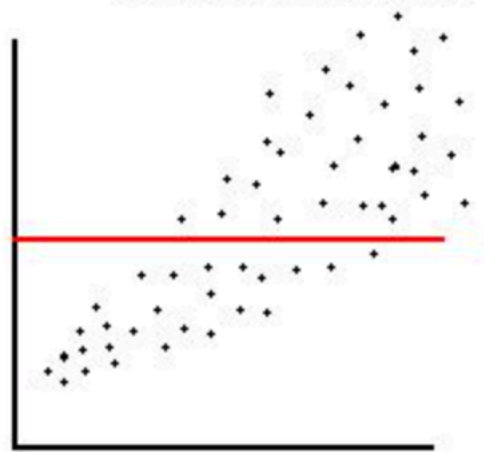
$$\varepsilon \sim \text{Normal}(0, \sigma^2)$$
$$\text{cor}(\varepsilon_i, \varepsilon_j) = 0$$



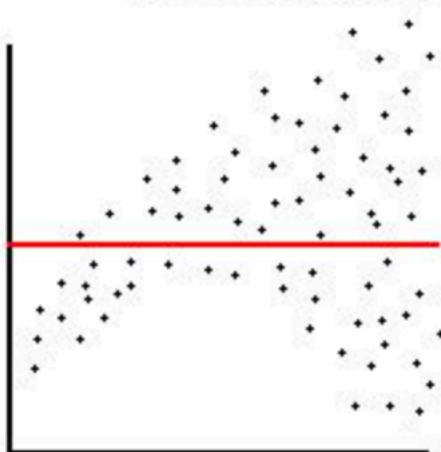
(b) Biased and Homoscedastic



(c) Biased and Homoscedastic



(e) Biased and Heteroscedastic

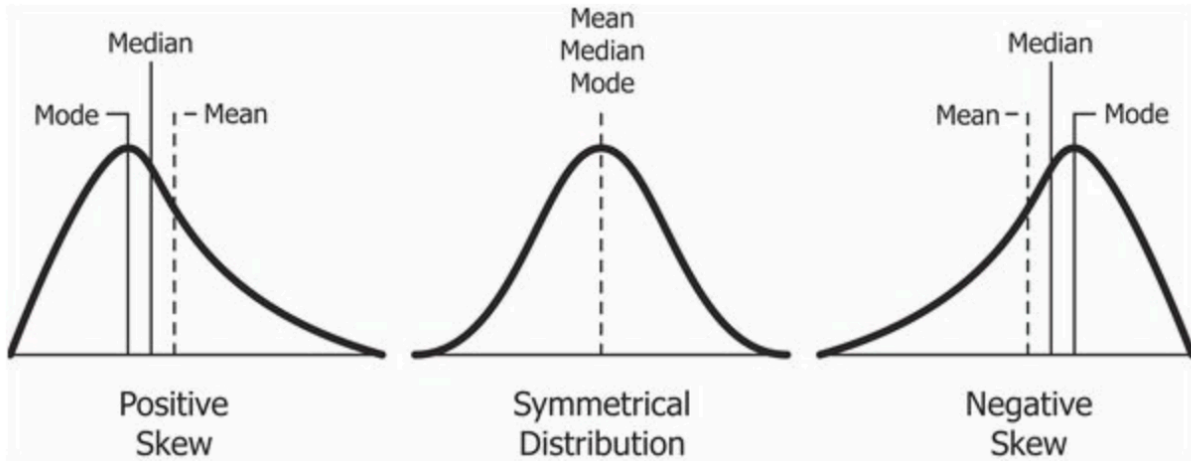


(f) Biased and Heteroscedastic

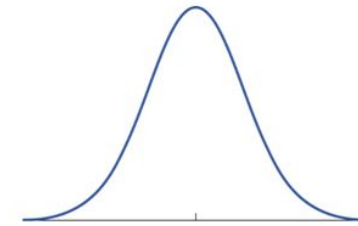
线性关系

- 1) 应采用非线性回归模型
- 2) 高度相关的自变量引起了共线性；
- 3) 模型缺少重要的自变量；

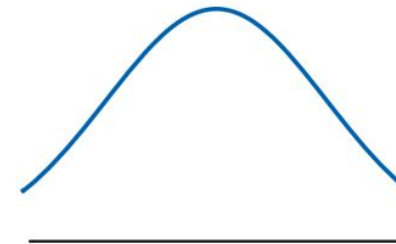
Q-Q图 (quantile-quantile plot)



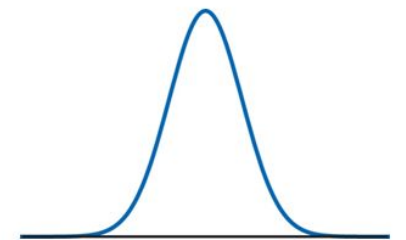
Shapes of Histograms (cont)



Normal distribution



Heavy Tails



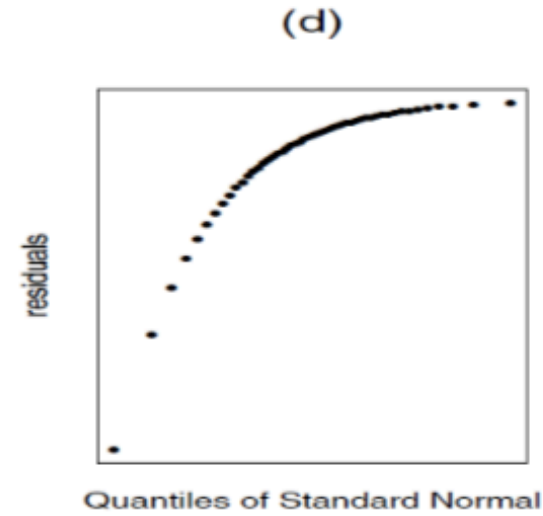
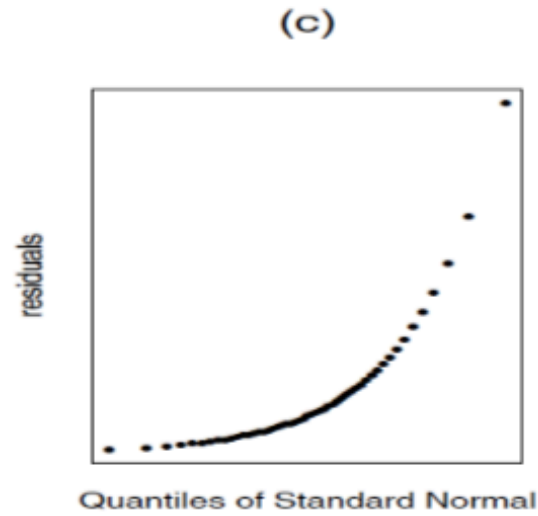
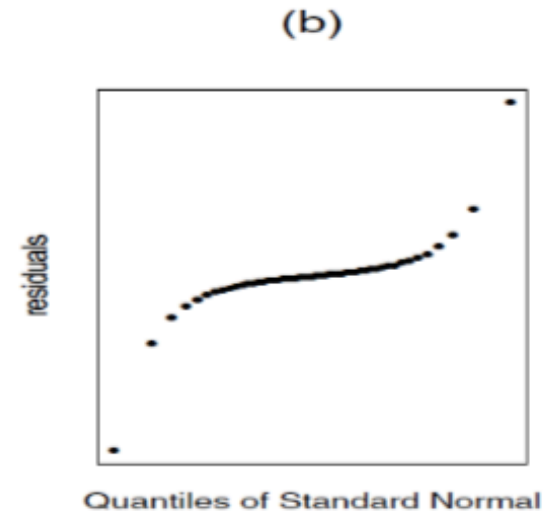
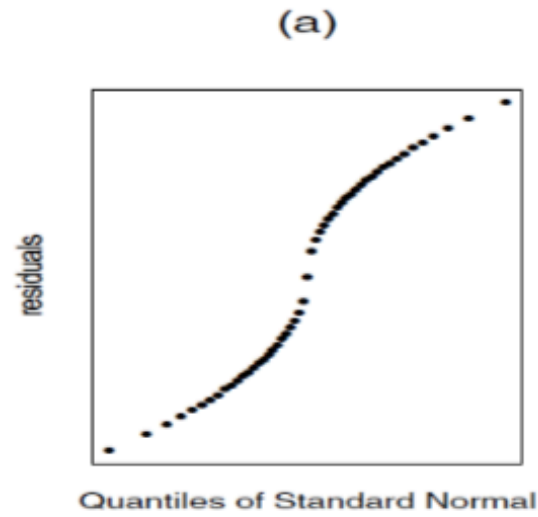
Light Tails

Q-Q图 (quantile-quantile plot)

- 根据残差自由度算出 $t_{(df.residual = n-2)}$ 分布对应 n 个样本的分位数

$$\text{theoretical quantile} = \phi_{df=n-2}^{-1}\left(\frac{0.5}{n}\right), \phi_{df=n-2}^{-1}\left(\frac{1.5}{n}\right) \dots \dots, \phi_{df=n-2}^{-1}\left(\frac{n-0.5}{n}\right)$$

- 把样本因变量 y 排序
- theoretical quantile 为x轴, 样本因变量(y 变量)排序过后为y轴, 做散点图



- light tailed (a), heavy tailed (b), positively skewed (c), negatively skewed (d).