3. Inference for the regression parameters and model

Remark： 3.2 – 3.4 ，3.8 基于假设：

① 不相关

② 正态分布

3.1 Three important distributions （正态总体下的三大抽样分布）

(a). χ^2 Distribution：

定义1 若 $X_1, X_2, \cdots, X_n \sim N(0,1)$ 相互独立，则称

$$Y = \sum_{i=1}^{n} X_i^2$$ [redacted]

服从"自由度"(degrees of freedom) 为 n 的卡方分布，记为 $Y \sim \chi^2(n)$

(1) Y 的 p.d.f.

$$p(y) = \begin{cases} \dfrac{1}{\Gamma(\frac{n}{2})} \cdot \left(\frac{1}{2}\right)^{\frac{n}{2}} \cdot y^{\frac{n}{2}-1} \cdot e^{-\frac{y}{2}} & y > 0 \\ 0 & y \le 0 \end{cases}$$ [redacted]

其中 $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt \quad (x>0)$ [redacted]

(2) $E(Y) = n$ [redacted]

$var(Y) = 2n$ [redacted]

(3) 若 $Y_1 \sim \chi^2(n)$ $Y_2 \sim \chi^2(m)$ 且 $Y_1 \perp\!\!\!\perp Y_2$ 则

$$Y_1 + Y_2 \sim \chi^2(n+m)$$ [redacted]

(b). t Distribution :

定义 2 若 $X \sim N(0,1)$ $Y \sim \chi^2(n)$ 且 $X \perp Y$ 则称

$$Z = \frac{X}{\sqrt{Y/n}}$$

服从"自由度"为 n 的 t 分布，记为 $Z \sim t(n)$

(1) Z 的 p.d.f.

$$p(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\cdot\Gamma(\frac{n}{2})}\cdot\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

且关于 y 轴对称

(2) 当 $m \geq n$ 时，$t(n)$ 的 m 阶及高于 m 阶的原点矩、中心矩均不存在

(3) $E(Z) = 0$, $n > 1$

$\mathrm{var}(Z) = \frac{n}{n-2}$, $n > 2$

(4) $t(1)$ 即 Cauchy 分布

$n \longrightarrow \infty$ 时，$t(n)$ 逐渐接近 $N(0,1)$

(c). F Distribution

定义 3 若 $X \sim \chi^2(m)$ $Y \sim \chi^2(n)$ 且 $X \perp Y$ 则称：

$$W = \frac{X/m}{Y/n}$$

服从自由度为 m,n 的 F 分布，记作 $W \sim F(m,n)$

(1) W 的 p.d.f.

$$p(x) = \begin{cases} \dfrac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\cdot\Gamma(\frac{n}{2})}\cdot\left(\dfrac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1}\cdot\left(1+\dfrac{m}{n}x\right)^{-\frac{m+n}{2}} & x>0 \\ 0 & \end{cases}$$ [redacted]

[redacted]

(2) $F(1,n) = t^2(n)$ 即

若 $Z = \dfrac{X}{\sqrt{Y/n}} \sim t(n)$ 则 $Z^2 = \dfrac{X^2}{Y/n} \sim F(1,n)$ [redacted]

(3) 若 $W \sim F(m,n)$ 则 $\dfrac{1}{W} \sim F(n,m)$ [redacted]

(4) $E(W) = \dfrac{n}{n-2}$, $n>2$ [redacted]

$var(W) = \dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$, $n>4$ [redacted]

定理 1 设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本，样本均值和方差为：

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(X_i-\overline{X})^2$$

则有：

(1) $\overline{X}$ 与 S^2 相互独立 [redacted]

(2) $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ [redacted]

(3) $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ [redacted]

(4) $\dfrac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t(n-1)$ [redacted]

定理 2 设 $X_1, X_2, \cdots, X_m$ 是来自正态总体 $N(\mu_1, \sigma_1^2)$ 的样本；$Y_1, Y_2, \cdots, Y_n$ 是来自正态总体 $N(\mu_2, \sigma_2^2)$ 的样本；且两样本相互独立，记

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$S_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i-\overline{X})^2$$

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2$$

则有

$$F = \frac{S_x^2/\sigma_1^2}{S_y^2/\sigma_2^2} \sim F(m-1, n-1)$$ [redacted]

特别地：

若 $\sigma_1^2 = \sigma_2^2$ 则：$F = \dfrac{S_x^2}{S_y^2} \sim F(m-1, n-1)$ [redacted]

3.2 t-test for β_1

$$H_0: \beta_1 = 0$$ [redacted]

$$H_1: \beta_1 \neq 0$$ [redacted]

① 已知：$\hat{\beta}_1 \sim N(\beta_1, \dfrac{\sigma^2}{L_{xx}})$

② 在假设 H_0 为真时，$\beta_1 = 0$ 故 $\hat{\beta}_1 \sim N(0, \dfrac{\sigma^2}{L_{xx}})$

③ 但 σ^2 未知（unknown parameter），用 σ^2 的 Unbiased Estimator：

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$ [redacted]

$$= \frac{1}{n-2}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ [redacted]

④ 由 [定理 1 (3)]

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$ Remind [redacted]

⑤ 又因为 $\hat{\sigma}^2 \perp \hat{\beta}_1$

故在 $H_0: \beta_1 = 0$ 的假设下有:

$$t = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}} \sim t(n-2)$$ [redacted]

检验:

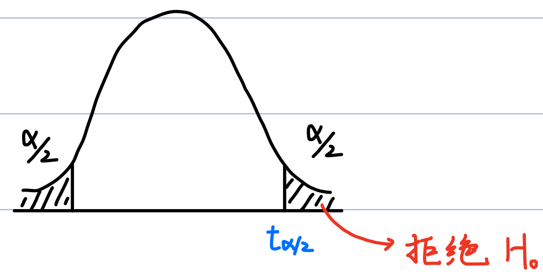
① 当原假设 $H_0: \beta_1 = 0$ 成立时, $t = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}} \sim t(n-2)$

② 给定显著性水平 α

③ 当 $|t| > t_{\alpha/2}$ 时, 拒绝 $H_0: \beta_1 = 0$

认为 β_1 显著不为 0,

y 对 x 的一元线性回归成立



p-value test:

① 当原假设 $H_0: \beta_1 = 0$ 成立时, $t = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}} \sim t(n-2)$

② 给定显著性水平 α

③ $P(|t| > t_{\alpha/2}) = p$ 当 $p \le \alpha$ 时, 拒绝 $H_0: \beta_1 = 0$

认为 β_1 显著不为 0,

y 对 x 的一元线性回归成立

3.3 F-test for regression model

H_0 : the regression model is insignificant [redacted]

H_1 : the regression model is significant [redacted]

平方和分解式

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$ [redacted]

$$SST = SSR + SSE$$

(1) Sum of Squres Total : 总离差平方和

(2) Sum of Squres Regression : 回归平方和

(3) Sum of Squres Error : 残差平方和

回归的效果

SSR 占比越大，回归模型效果越佳

故，下面构造 F 检验

$y_i \sim N(\mu, \sigma^2)$

$E(y_i) = \beta_0 + \beta_1 x_i$

F-test for $\frac{SSR}{SSE}$

H_0 : $\beta_1 = 0$

$\Rightarrow \beta_0 = E(y_i) = \mu$

① $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = (n-2)\hat{\sigma}^2$

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$ [redacted]

② $SSR = \sum_{i=1}^{n}(\hat{y}_i-\bar{y}_i)^2$

$\frac{SSR}{\sigma^2} \sim \chi^2(1)$ under H_0 : $\beta_1 = 0$ [redacted]

or H_0 : $SSR = 0$

③ SSE ⊥ SSR

故构造

$$F = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$$

④ 当 $F > F_{1-\alpha}(1, n-2)$ 时，拒绝 H_0，说明回归方程显著

3.4 Significance test of the correlation coefficient 相关系数

相关系数的显著性检验

correlation coefficient 相关系数

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{L_{xy}}{\sqrt{L_{xx} \cdot L_{yy}}}$$

$$= \hat{\beta}_1 \cdot \sqrt{\frac{L_{xx}}{L_{yy}}}$$

Under H_0

$$\frac{\sqrt{n-2} \cdot r}{\sqrt{1-r^2}} \sim t(n-2)$$

then, 拒绝域 $|t| > t_{\alpha/2}(n-2)$

Rule of thumb 相关系数检验规则

(1) highly correlated : $|r| \geq 0.8$

(2) moderately correlated : $0.5 \leq |r| < 0.8$ [redacted]

(3) weakly correlated : $0.3 \leq |r| < 0.5$

(4) very weakly correlated : $|r| < 0.3$

Remark : $r \longrightarrow$ only <u>linear</u> association

3.5 Coefficient of determination 决定系数

The proportion of the response's variation that can be explained by the regression model, that is

$$\frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$ [redacted]

$$= \frac{\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{x})^2}{L_{yy}}$$

$$= \hat{\beta}_1^2 \cdot \frac{L_{xx}}{L_{yy}} = \frac{L_{xy}^2}{L_{xx} L_{yy}}$$

$$= r^2$$ [redacted]

Remark $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

检验： r^2 越接近 1，回归拟合优度好

3.6 Relationship among test and Summary 检验的关系

(1) t and r

$$t = \hat{\beta}_1 \cdot \frac{\sqrt{L_{xx}}}{\hat{\sigma}} = \frac{\sqrt{n-2}\ r}{\sqrt{1-r^2}}$$

(2) t and F

$$F = \frac{SSR/1}{SSE/(n-2)} = \hat{\beta}_1^2 \cdot \frac{L_{xx}}{\hat{\sigma}^2} = t^2$$

(3) SSR SST and r

$$\frac{SSR}{SST} = r^2$$

四种检验 $\{ t,\ F,\ r,\ SSR/SST \}$ 等价

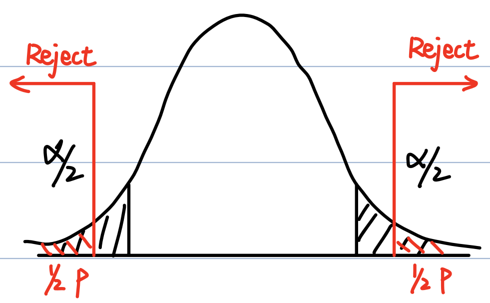
补: $\sqrt{F(1,m)} = t(m)$

3.7 P-Value P值检验

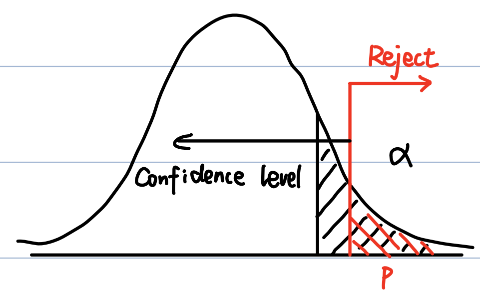
$P > \alpha$ do not reject H_0

$P \le \alpha$ reject H_0

p-value of a two-sided test



p-value of a one-sided test



3.8 Confidence Interval 置信区间

由于 ① $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$

② $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$

③ $\hat{\sigma}^2 \perp \hat{\beta}_1$

所以有

$$\frac{(\hat{\beta}_1 - \beta_1)\cdot\sqrt{L_{xx}}}{\hat{\sigma}} \sim t(n-2)$$

[redacted]

由 $P\left(\left| \frac{(\hat{\beta}_1 - \beta_1)\cdot\sqrt{L_{xx}}}{\hat{\sigma}} \right| < t_{1-\alpha/2}(n-2) \right) = 1-\alpha$

得 置信区间 $(1-\alpha)$ CI of β_1 is

$$\left[\hat{\beta}_1 - t_{1-\alpha/2}(n-2)\cdot\frac{\hat{\sigma}}{\sqrt{L_{xx}}} , \hat{\beta}_1 + t_{1-\alpha/2}(n-2)\cdot\frac{\hat{\sigma}}{\sqrt{L_{xx}}} \right]$$

[redacted]

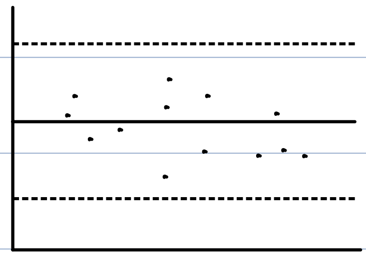
4. Residual Analysis 残差分析

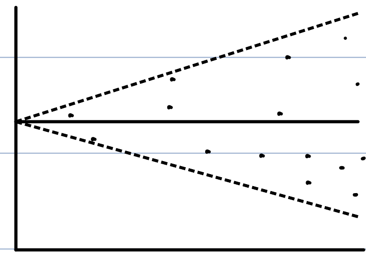
4.1 Definitions of residuals and residual plots

定义 残差 $e_i = y_i - \hat{y}_i$

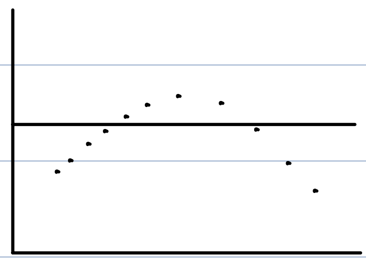
可视为对 $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$ 的估计 (可视化，但并不是)

残差图 以 x_i 为横轴 (或 $\hat{y}_i$)，e_i 为纵轴

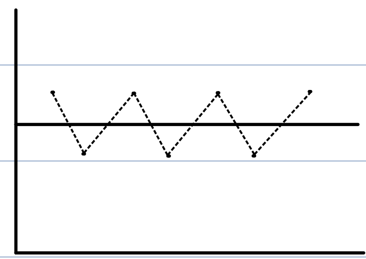




与 $var(\varepsilon_i) = \sigma^2 \in const$ 矛盾



曲线 or $cov(\varepsilon_i, \varepsilon_j) \neq 0$ 自相关



$cov(\varepsilon_i, \varepsilon_j) \neq 0$ 自相关

4.2 Properties of residuals 残差性质

(1) Expectation :

$$E(e_i) = 0$$ [redacted]

(2) Variance :

$$var(e_i) = (1 - h_{ii})\sigma^2$$ [redacted]

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{L_{xx}} \quad \text{is "Leverage"}$$ [redacted]

当 x_i 靠近 $\bar{x}$，h_{ii} 越近 0，残差方差越大

(3) Equation :

$$\sum_{i=1}^{n} e_i = 0$$ [redacted]

$$\sum_{i=1}^{n} x_i e_i = 0$$ [redacted]

表明残差 $e_1, e_2, \cdots, e_n$ 是相关的，不是独立的

(4) $\hat{\sigma}^2$ is unbiased estimator

$$E(\hat{\sigma}^2) = \frac{1}{n-2}\sum_{i=1}^{n} E(e_i^2) = \frac{1}{n-2}\cdot\sum_{i=1}^{n} var(e_i) = \sigma^2$$ [redacted]

Proof (2) 由 $\hat{y}_i \sim N\left(\beta_0+\beta_1 x_i,\ \left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{L_{xx}}\right)\sigma^2\right)$

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{x_i-\bar{x}}{L_{xx}} y_i$$

$$\hat{\beta}_0 = \sum_{i=1}^{n}\left[\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{L_{xx}}\right]\cdot y_i$$

有 $var(e_i) = var(y_i-\hat{y}_i)$

$$= var(y_i)+var(\hat{y}_i)-2cov(y_i,\hat{y}_i)$$

$$= \sigma^2 + \left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{L_{xx}}\right)\sigma^2 - 2\underbrace{cov(y_i,\ \hat{\beta}_0+\hat{\beta}_1 x_i)}_{①}$$

$$① = cov(y_i,\hat{\beta}_0) + x_i\, cov(y_i,\hat{\beta}_1)$$

$$= \left(\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{L_{xx}}+\frac{x_i(x_i-\bar{x})}{L_{xx}}\right)\sigma^2$$

故原式 $= \sigma^2 + \left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{L_{xx}}\right)\sigma^2 - 2\left(\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{L_{xx}}+\frac{x_i(x_i-\bar{x})}{L_{xx}}\right)\sigma^2$

$$= \left(1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{L_{xx}}\right)\sigma^2$$

4.3 Modified Residuals 改进的残差

标准化残差：

$$ZRE_i = \frac{e_i}{\hat{\sigma}}$$ [redacted]

学生化残差 (Studentized residuals)：

$$SRE_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$ [redacted]

$$= \frac{e_i}{\hat{\sigma}\cdot\sqrt{1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{L_{xx}}}}$$ [redacted]

判断异常值

$$|SRE_i| > 3$$

的观测值，视为异常值.