Проверка на хипотези

Постановка: някаква хипотеза, се свежда до $F_X(x,\theta)$

$\vec{X}$; $W \subseteq \mathbb{R}^n$; $\vec{x} \in W$, то отхвърляме хипотезата

$\vec{x} \notin W$, то приемаме

– X, $F_X(x,\theta)$

$H_0: \theta = \theta_0$ – нулева хипотеза

$H_1: \theta = \theta_1$ – алтернативна хипотеза

} проста с/у проста хипотеза

– $\vec{X}$, W, ако $\vec{x} \in W$, то отхв. H_0 и приемаме H_1

$H_0: \theta = \theta_0$ $\quad W \subseteq \mathbb{R}^n$, $\vec{X} \in W$ отхвърляме H_1

$H_1: \theta = \theta_1$ $\quad \vec{X} \notin W$ приемаме H_0

Грешка от I$^{\text{ви}}$ род: $\mathbb{P}(\vec{X} \in W | H_0) = \alpha = \alpha_W$

Грешка от II$^{\text{ри}}$ род: $\mathbb{P}(\vec{X} \notin W | H_1) = \beta = \beta_W$

$W_1 \subseteq W_2 \quad \alpha_{W_1} \leq \alpha_{W_2}$

$\beta_{W_1} \geq \beta_{W_2}$

Контролира се грешката от I$^{\text{ви}}$ род: $\alpha = 0.05, 0.1, 0.01$

W е критична област за грешка от I$^{\text{ви}}$ род α

Деф) (Оптимална критична област) о.к.о

Наричаме W^* о.к.о. при фикс. $\alpha = \alpha_{W^*}$, ако $\beta_{W^*} = \min\limits_{\alpha_W = \alpha_{W^*} = \alpha} \beta_W$

$$\beta^* = \mathbb{P}(\vec{X} \notin W^* | H_1) = \min_{W:\, \mathbb{P}(\vec{X} \in W | H_0) = \alpha} \left(\mathbb{P}(\vec{X} \notin W | H_1)\right)$$

Допускаме X е сл. вел. и има пл. $f_X(x,\theta)$, $L(\vec{x},\theta)$. $H_0: \theta=\theta_0 \rightarrow$; $\vec{X}\in\mathbb{R}^n$
$H_1: \theta=\theta_1$

$L_1(\vec{x}) = L(\vec{x},\theta_1)$

$L_0(\vec{x}) = L(\vec{x},\theta_0)$

$$L(\vec{x},\theta) = \prod_{j=1}^{n} f_X(x_j,\theta)$$

Лема (Нейман-Пирсън)

Нека имаме сл. вел. X, с плътност $f_X(x,\theta)$, ф-я на правдопод. $L(\vec{x},\theta)$ и извадка $\vec{x}$. Ако $W^{*} \subseteq \mathbb{R}^n$ е критична област с грешка от първи род α, т.ч. $W^{*} \subseteq \{L_1(\vec{x}) \geq kL_0(\vec{x})\}$

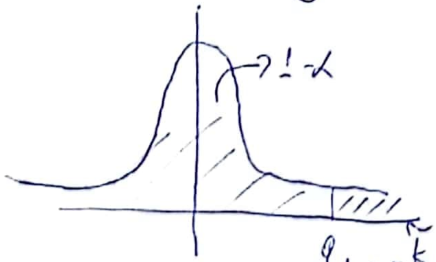
$(W^{*})^c \subseteq \{L_1(\vec{x}) \leq kL_0(\vec{x})\}$ за някое $k>0$

Тогава W^{*} е о.к.о. за $H_0: \theta=\theta_0$
$H_1: \theta=\theta_1$

⊕ $X \sim N(\mu,\sigma^2)$ ↪ изв. $H_0: \mu=\mu_0$, $H_1: \mu=\mu_1$ $(\mu_1>\mu_0)$, $\alpha=0.05$, извадка $\vec{X}$ с n набл.

$$L_1(\vec{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma^n} \prod_{j=1}^{n} e^{-\frac{(x_j-\mu_1)^2}{2\sigma^2}}$$

$$L_0(\vec{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma^n} \prod_{j=1}^{n} e^{-\frac{(x_j-\mu_0)^2}{2\sigma^2}}$$



$W^{*} = \{L_1 \geq kL_0\}$ и $P(\vec{X}\in W^{*} \mid H_0) = \alpha$

$$\{\vec{x}\in\mathbb{R}^n : L_1(\vec{x}) \geq k.L_0(\vec{x})\} = \left\{ \cancel{\left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma^n}} e^{-\frac{\sum_{j=1}^{n}(x_j-\mu_1)^2}{2\sigma^2}} \geq k \cancel{\left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma^n}} e^{-\frac{\sum(x_j-\mu_0)^2}{2\sigma^2}} \right\} \quad / \ln$$

$$= \left\{ -\frac{\sum_{j=1}^{n}(x_j-\mu_1)^2}{2\sigma^2} \geq \ln k - \frac{\sum_{j=1}^{n}(x_j-\mu_0)^2}{2\sigma^2} \right\} = \left\{ -\cancel{\sum_{j=1}^{n} x_j^2} + 2\sum_{j=1}^{n} x_j\mu_1 - \underbrace{\sum_{j=1}^{n}\mu_1^2}_{n.\mu_1^2} \geq 2\sigma^2 \ln k - \cancel{\sum_{j=1}^{n} x_j^2} + 2\sum_{j=1}^{n} x_j\mu_0 - \sum_{j=1}^{n}\mu_0^2 \right\}$$

$$= \left\{ 2(\mu_1-\mu_0)\sum_{j=1}^{n} x_j \geq 2\sigma^2 \ln k + n(\mu_1^2-\mu_0^2) \right\} =$$

$$= \left\{ \sum_{j=1}^{n} x_j \geq \underbrace{\frac{2\sigma^2 \ln k + n(\mu_1^2-\mu_0^2)}{2(\mu_1-\mu_0)}}_{\tilde{k}} \right\} \hookrightarrow W^{*}$$ Търсим k: $P(\vec{X}\in W^{*} \mid H_0) = P\left(\sum_{j=1}^{n} X_j \geq \tilde{k} \mid H_0\right) =$

Знаем, че $\frac{\tilde{k} - n.\mu_0}{\sigma\sqrt{n}} = q_{1-\alpha}$

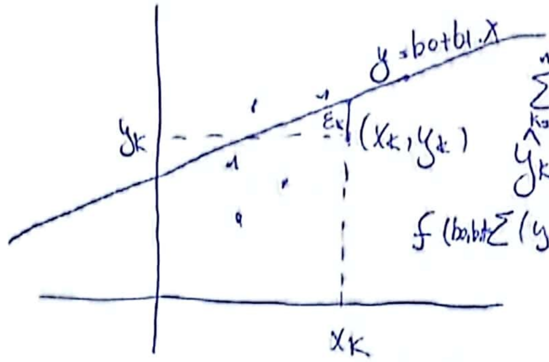
$$= P\left(\frac{\sum_{j=1}^{n} X_j - n\mu_0}{\sigma\sqrt{n}} \geq \frac{\tilde{k} - n.\mu_0}{\sigma\sqrt{n}} \,\middle|\, H_0\right) = P\left(Z \geq \frac{\tilde{k} - n\mu_0}{\sigma\sqrt{n}}\right)$$

Знаем n, μ_0, σ, μ_1 и по обратен път намираме $\tilde{k}$ и от там $\ln k \rightarrow k$

Линейна регресия

$$y_{син} = \alpha + \beta(y_{баща} - \alpha)$$

α — среден ръст мъже; $\beta \to 0.6$



$$\sum_{k=1}^{n} \varepsilon_k^2 \to min$$

$$\hat{y}_k = b_0 + b_1 x_k$$

$$f(b_0, b_1) = \sum (y_k - \hat{y}_k)^2 = \sum (y_k - b_0 - b_1 x_k)^2$$

$$y = \beta_0 + \beta_1 . X + \varepsilon$$

y — отклик; X — предиктор; $\varepsilon \to$ грешка

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n}$$

$$\frac{\partial f}{\partial b_0} = 0 = -2\sum(y_k - b_0 - b_1 x_k) = n\bar{y} - nb_0 - b_1 n\bar{x} \Rightarrow b_0 = \bar{y} - b_1\bar{x}$$

$$\frac{\partial f}{\partial b_1} = 0 = -2\sum x_k(y_k - b_0 - b_1 x_k) = \sum x_k y_k - b_0 n\bar{x} - b_1 \sum x_k^2 = 0$$

$$b_1 = \frac{\sum x_k y_k - n\bar{x}\bar{y}}{\sum x_k^2 - n(\bar{x})^2} = \frac{\sum (x_k - \bar{x})(y_k - \bar{y})}{\sum (x_k - \bar{x})^2} = \frac{\sum (x_k - \bar{x})}{\sum (x_k - \bar{x})^2} . y_k$$

$$b_0 = \bar{y} - b_1\bar{x} = \sum_1 \left(\frac{1}{n} - \bar{x}\frac{\sum_1 (x_k - \bar{x})}{\sum (x_k - \bar{x})^2}\right) y_k$$

$$A = \sum (x_k - \bar{x})^2$$

$$C_k = \frac{x_k - \bar{x}}{A}$$

$$b_1 = \sum C_k y_k$$

$$b_0 = \sum \left(\frac{1}{n} - \bar{x} . C_k\right) y_k$$

$$\sum_1 C_k = 0 = \frac{\sum x_k - n\bar{x}}{A} = 0$$

$$b_1 = \sum_{k=1}^{n} C_k y_k \qquad C_k = \frac{x_k - \bar{x}}{A}$$

$$b_0 = \sum_{k=1}^{n} \left(\frac{1}{n} - \bar{x} C_k\right) y_k \qquad A = \sum_{k=1}^{n} (x_k - \bar{x})^2$$

$\Rightarrow$ дават най-добрата права, т.е. кв. грешки до нея са мин.

Стохастичен модел

$$y_k = \beta_0 + \beta_1 X_k + \varepsilon_k$$

y_k — отклик; X_k — предиктор; $\varepsilon_k \to$ стохастична грешка

- $(\varepsilon_1, \dots, \varepsilon_n)$ – са независими в съвкупност
- $E[\varepsilon_j] = 0$, $j = \overline{1,n}$ – липсва систематична грешка
- $D[\varepsilon_j] = \sigma^2$ – хомоскедастичност
- $\varepsilon_j \sim N(0, \sigma^2)$, $j = \overline{1,n}$

$$\hat{\beta}_1 = b_1 = \sum_{k=1}^{n} \frac{(x_k - \bar{x})}{A} \cdot y_k = \sum_{k=1}^{n} c_k \cdot y_k \sim N\left(\beta_1, \frac{\sigma^2}{A}\right)$$

$$\hat{\beta}_0 = b_0 = \sum_{k=1}^{n} \left(\frac{1}{n} - \frac{(x_k - \bar{x})\bar{x}}{A} \right) \cdot y_k = \sum \left(\frac{1}{n} - \bar{x} \cdot c_k \right) y_k \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{A}\right)\right)$$

$b_1 = \hat{\beta}_1$ | $y_k = \beta_0 + \beta_1 \cdot x_k + \varepsilon_k$ $\mathbb{E}[y_k] = \beta_0 + \beta_1 x_k$

$$\mathbb{E}[b_1] = \sum_{k=1}^{n} c_k \mathbb{E}[y_k] = \sum_{k=1}^{n} c_k (\beta_0 + \beta_1 x_k) = \beta_1 \sum_{k=1}^{n} \frac{(x_k - \bar{x}) x_k}{A} + \frac{1}{A} \beta_0 \underbrace{\sum (x_k - \bar{x})}_{0}$$

не променя нищо

$$= \frac{\beta_1}{A} \sum_{k=1}^{n} (x_k - \bar{x})(x_k - \bar{x}) = \beta_1 \Rightarrow \text{неизместена оценка}$$

От $\varepsilon_k \sim N(0, \sigma^2) \Rightarrow y_k \sim N(\beta_0 + \beta_1 x_k, \sigma^2)$ и са нез., защото $(\varepsilon_k)_{k=1}^{n}$ са нез.

$$b_1 = \sum_{k=1}^{n} c_k y_k$$

$$\mathbb{D}[b_1] = \sum_{k=1}^{n} \mathbb{D}[c_k y_k] = \sum_{k=1}^{n} c_k^2 \sigma^2 = \sigma^2 \frac{\sum_{k=1}^{n} (x_k - \bar{x})^2}{\left(\sum_{k=1}^{n} (x_k - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2} = \frac{\sigma^2}{A}$$

$$\Rightarrow b_1 \sim N\left(\beta_1, \frac{\sigma^2}{A}\right)$$

$b_0 = \hat{\beta}_0$ | $= \bar{y} - b_1 \bar{x} = \sum_{k=1}^{n} \left(\frac{1}{n} - \bar{x} c_k \right) y_k$

$$\mathbb{E}[b_0] = \sum_{k=1}^{n} \left(\frac{1}{n} - \bar{x} c_k \right) \mathbb{E}[y_k] = \sum_{k=1}^{n} \left(\frac{1}{n} - \bar{x} c_k \right) (\beta_0 + \beta_1 x_k) =$$

$$= \beta_0 + \beta_1 \bar{x} - \beta_0 \bar{x} \underbrace{\sum_{k=1}^{n} c_k}_{0} - \beta_1 \bar{x} \sum_{k=1}^{n} c_k x_k = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \frac{\sum_{k=1}^{n} (x_k - \bar{x})(x_k - \bar{x})}{A} =$$

$$= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0 \rightarrow \text{неизместена}$$

$$\mathbb{D}[b_0] = \sum_{k=1}^{n} \left(\frac{1}{n} - \bar{x} c_k \right)^2 \sigma^2 = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{A}$$

$$\Rightarrow b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{A}\right)\right)$$

$\sigma^2 \rightarrow$ можем да го знаем, но е малко вероятно!

σ^2 е известно

$$\frac{y_k - \beta_0 - \beta_1 x_k}{\sigma} \sim N(0,1) =: Z_k$$

$$\sum_{k=1}^{n} Z_k^2 = \frac{1}{\sigma^2}\sum_{k=1}^{n}(y_k - \beta_0 - \beta_1 x_k)^2 \sim \chi^2(n)$$

$$\frac{1}{n}\cdot\frac{1}{\sigma^2}\sum_{k=1}^{n}(y_k - \beta_0 - \beta_1 x_k)^2 \xrightarrow[n\to\infty]{п.с.} 1 = \mathbb{E}[Z_1^2]$$

$$\sigma^2 \sim \frac{1}{n}\sum_{k=1}^{n}(y_k - \beta_0 - \beta_1 x_k)^2$$

σ^2 не е известно

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(y_k - \beta_0 - \beta_1 x_k)^2 \sim \sigma^2$$

$$\sum_{k=1}^{n}\frac{(y_k - b_0 - b_1 x_k)^2}{\sigma^2} \sim \chi^2(n-2)$$ – използваме 2 оценки за b_0 и b_1

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{k=1}^{n}(y_k - b_0 - b_1 \cdot x_k)^2 \approx \sigma^2$$

$$\lim_{n\to\infty}\sum_{k=1}^{n}\frac{(y_k - b_0 - b_1 x_k)^2}{(n-2)\sigma^2} \overset{п.с.}{=} 1$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{k=1}^{n}(y_k - b_0 - b_1 x_k)^2$$

Проверка на хипотези

$H_0: \beta_1 = 0$ → няма лин. зав. $\quad Y_k = \beta_0 + \varepsilon_k$

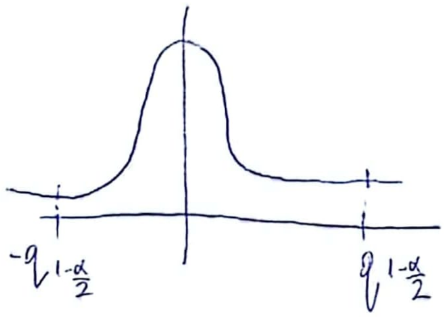
$H_1: \beta_1 = const$
или
$\beta_1 \neq 0$

$H_0: \beta_1 = \beta$

$H_1: \beta_1 = \beta^* \ (\beta_1 \neq \beta)$

σ^2 е известно

$$\frac{b_1 - \beta}{\sqrt{\frac{\sigma^2}{A}}} =: Z \sim N(0,1) \text{ при валидност на } H_0$$



α – грешка от първи род

$W = \{ |Z| \geq q_{1-\frac{\alpha}{2}} \}$

$\beta \in (I_1, I_2)$

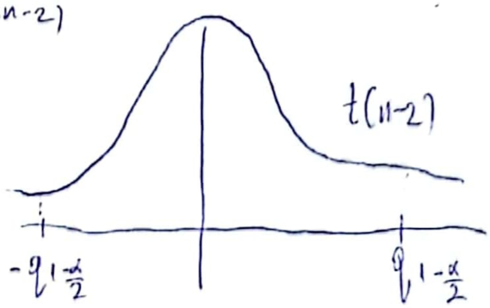
$I_1 = b_1 - q_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{A}}$

$I_2 = b_1 + q_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{A}}$

σ^2 не е известна

$$T = \frac{b_1 - \beta}{\sqrt{\frac{\hat{\sigma}^2}{A}}} \in t(n-2) \qquad \frac{\hat{\sigma}^2}{\sigma^2}(n-2) \in \chi^2(n-2)$$

$$T = \frac{\overbrace{\frac{b_1 - \beta}{\sigma/\sqrt{A}}}^{\sim N(0,1)}}{\frac{\sqrt{\frac{\hat{\sigma}^2}{A} \cdot \frac{(n-2)}{(n-2)}}}{\sigma/\sqrt{A}}} = \frac{Z}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2} \cdot \frac{(n-2)}{(n-2)}}} = \frac{Z}{\sqrt{\frac{\chi^2(n-2)}{(n-2)}}}$$



α – грешка от първи род

$\{ |T| \geq q_{1-\frac{\alpha}{2}} \}$

$I_1 = b_1 - q_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2}{A}}$

$I_2 = b_1 + q_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2}{A}}$

При $n \geq 32$

$t(n-2) \sim N(0,1)$

$\boxed{\text{известно } \sigma^2}$

$$z = \frac{b_0 - \beta}{\sqrt{\dfrac{\sigma^2}{\frac{1}{n} + \frac{\bar{X}^2}{A}}}} \sim N(0,1)$$

$\boxed{\text{неизвестно } \sigma^2}$

$$z = \frac{b_0 - \beta}{\sqrt{\dfrac{\hat{\sigma}^2}{\frac{1}{n} + \frac{\bar{X}^2}{A}}}} \sim t(n-2) \overset{n \geq 32}{\sim} N(0,1)$$