20.02.2018 Приложна статистика (лекции)

Линейна регресия - $\overset{\text{отклик}}{Y} = a\underset{\text{predictor}}{X} + b + \varepsilon^{\text{грешка}}$

линейна по X, a и b - линеен модел спрямо параметрите

X и Y са непрекъснати (количествени) - регресионен анализ

X е дискретна (качествена) и Y е количествена - дисперсионен анализ

X е количествена и Y е качествена - класификация

0) Събиране на данни
1) Образуване на вариационен ред

$$X_{(1)} < X_{(2)} < \ldots < X_{(k)}, \quad k \leq n$$
(повторенията не се броят)

$m_1 \quad m_2 \quad \ldots \quad m_k$ - честоти

$$\sum_{i=1}^{k} m_i = n$$

$p_1 \quad p_2 \quad \ldots \quad p_k$ - относителни честоти, $\frac{m_i}{n}$

$f_i = \sum_{j=1}^{n} m_j$ - кумулативни честоти

$\bar{f}_i = \frac{f_i}{n}$ - отн. кумул. честоти

2) Изобразяване на данните графично

* честотен полигон - $X_{(i)}$ по абсцисата, p_i или m_i по ординатата.

* кумулативни честоти (отн. или не)

* хистограма - Размахът се дели на m инт. и се сумират m_i във всеки инт.

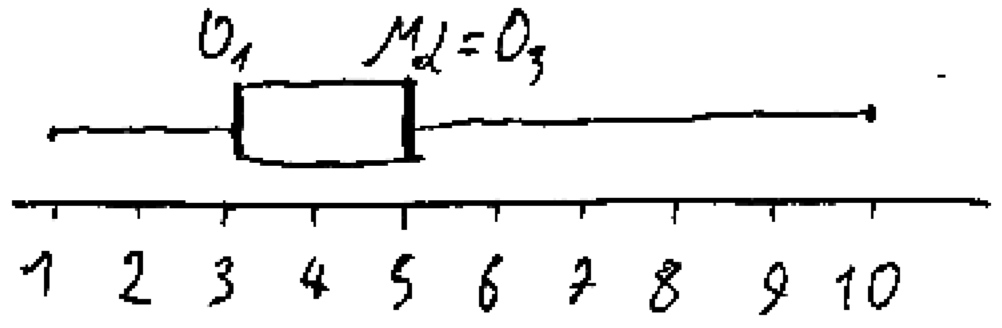
<u>Опр.</u> Ако имаме 1 max в плътността на дадено разпр., то се нарича унимодално.

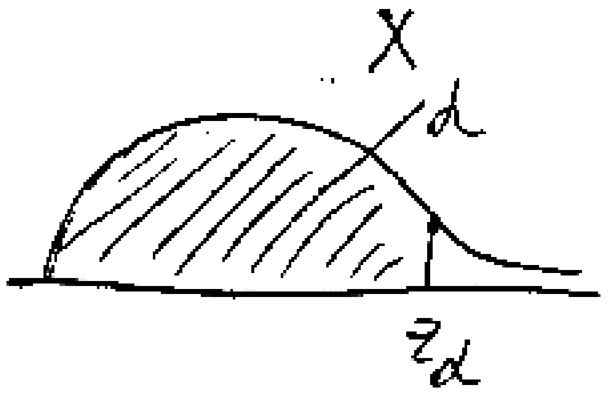
27.02.2018

5 3 1 5 5 3 1 5 6 10

1 1 3 3 5 | 5 5 5 6 10

Q_1 $\quad M_d \quad$ Q_3





$P(X < z_\alpha) = \alpha$

Q_1 = 1-ви квартил - 25%

$Q_2 = M_d$ - 2-ри кв. - 50%

Q_3 = 3-ти квартил 75%

$X_1, \dots, X_n \in F(x, \theta)$

<u>Пр.</u> За $N_0(\mu, \sigma^2)$ $\quad \vec{\theta} = (\mu, \sigma^2)$

$\hat{\theta} = t_n(X_1, \dots, X_n)$ - произв. ф-я на разпределение

$t_n(\overline{X}) \xrightarrow[n \to \infty]{P} \theta$ - състоятелна оценка

---

*Сходимост по разпределение

$\{\xi_n\}$ - случайни величини над $(\Omega, \mathcal{F}, P)$

$\xi_n \xrightarrow[n \to \infty]{d} \xi$ ако $F_{\xi_n}(x) \to F_\xi(x)$ $\forall x$, в което F е непр.

*Сходимост по вероятност: $\xi_n \xrightarrow[n \to \infty]{P} \xi$

$\forall \varepsilon > 0 \ \lim\limits_{n \to \infty} P(|\xi_n - \xi| < \varepsilon) = 1$

*функционална сходимост (п.с.): $\xi_n \xrightarrow[n \to \infty]{п.с.} \xi$

$P(\lim\limits_{n \to \infty} \xi_n = \xi) = 1$

*Сходимост в L_2 (в средноквадратичен смисъл): $\xi_n \xrightarrow[n\to\infty]{L_2} \xi$

$\lim\limits_{n\to\infty} \mathbb{E}(\xi_n - \xi)^2 = 0$

$$\begin{matrix} \text{п.с.} \searrow & \\ & P \to d \\ L_2 \nearrow & \end{matrix}$$

$\xi_1, \xi_2 \ldots$ $\qquad \mathbb{E}\xi_i = \mu$

$\qquad \mathbb{D}\xi_i < C$

$\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i \xrightarrow[n\to\infty]{\text{п.с.}} \mu$ (Закон за големите числа)

$P(|\xi - \mathbb{E}\xi| \geq \varepsilon) \leq \frac{\mathbb{D}\xi}{\varepsilon^2}$ (Неравенство на Чебишов)

$\lim\limits_{n\to\infty} P(|\xi_n - \xi| \geq \varepsilon) = 0$

$\lim\limits_{n\to\infty} P(|t_n(\vec{X}) - \theta| \geq \varepsilon) = 0$ (от нер. на Чебишов) (1)

$\mathbb{E}\, t_n(\vec{X}) = \theta$ - неизместеност на оценка

От (1) => оценката е слабосъстоятелна

$\lim\limits_{n\to\infty} \mathbb{E}\, t_n(\vec{X}) = \theta$ - асимптотична неизместеност

Тр. $\lim\limits_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \stackrel{п.с.}{=} \mathbb{E}X = \mu$

Зад. От силна състоятелност => слаба състоятелност

$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \frac{1}{n}.n.\mu = \mu$ - неизм. оценка за средното

$\mathbb{D}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\Big[\sum_{i=1}^{n}\mathbb{D}X_i + \sum_{i\neq j}\underbrace{\Sigma\, Cov(X_i, X_j)}_{0 \text{ за независими сл. вел.}}\Big] = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$

$\overline{S_n}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ - неизместена

$\overline{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ - изместена

$B = \mathbb{E}t_n(\vec{X}) = \theta$ - изместване на оценката

$\mathbb{E}(t_n(\vec{X}) - \theta)^2 = MSE(t_n(\vec{X}))$ - квадратична грешка

$MSE(t_n(\vec{X})) = \mathbb{D}(t_n(\vec{X})) + B^2$

$\vec{\theta} = (\theta_1, \dots, \theta_k)$

$\mathbb{E}X^i, i = 1, \dots, k$ - начален момент

$\mathbb{E}\widehat{X^i} = \frac{1}{n}\sum_{j=1}^{n} X_j^i \xrightarrow[n\to\infty]{P} \mathbb{E}X^i = g_i(\theta_1, \dots, \theta_k), i = 1, \dots, k \iff$

$\theta_j = t_j(\mathbb{E}X, \mathbb{E}X^2, \dots, \mathbb{E}X^k), j = 1, \dots, k$

Пр.
$X \in Ge(p)$. Търсим оценка за p

$P(X=k) = pq^k = p(1-p)^k,\ k=0,1,\dots \qquad (q=1-p)$

$\mathbb{E}X = \frac{q}{p} = \frac{1-p}{p}$

$p = \frac{1}{1+\mathbb{E}X}$; $\hat{p} = \frac{1}{1+\overline{X}_n} \xrightarrow[n\to\infty]{P} p$; $\widehat{\mathbb{E}X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Метод на max правдоподобие

$X_1, \dots, X_n$ - независими, едн. разпр. сл. вел.

$f(x_i;\theta)$ - плътност

$P(X=X_i) = P_{X_i}(\theta)$

Совместната плътност на $X_1, \dots, X_n$ се нарича ф-я на правдоподобие и се бележи с $L(X_1, \dots, X_n;\theta)$

$$L(X_1, \dots, X_n;\theta) = \prod_{i=1}^{n} f(X_i,\theta) = \prod_{i=1}^{n} P_{X_i}(\theta)$$

Пр. (продължение)

$$L(X_1, \dots, X_n;p) = \prod_{i=1}^{n} p(1-p)^{X_i} = p^n(1-p)^{\sum_{i=1}^{n} X_i}$$

$$\ln L(X_1, \dots, X_n;p) = n\ln p + \sum_{i=1}^{n} X_i \ln(1-p)$$

Търсим нулите на производната:

$$\frac{\partial}{\partial p} \ln L(X_1, \ldots, X_n; p) = \frac{n}{p} - \frac{\sum_{i=1}^{n} X_i}{1-p} = 0$$

$$n(1-p) = pn\overline{X}_n \;/n$$

$$p = \frac{1}{1+\overline{X}_n}$$

06.03.2018

$X_i \in Bernoulli \; (P(X_i=1)=1-P(X_i=0)=p)$

$Y=\sum_{i=1}^{n} X_i \in Bi(n,p)$

$P(X_1=x_1, X_2=x_2, \dots, X_n=x_n \mid Y=y)=$

$$= \frac{P(X_1=x_1, \dots, X_n=x_n, \sum_{i=1}^{n} X_i = y)}{P(\sum_{i=1}^{n} X_i = y)} = \frac{\prod_{i=1}^{n} P(X_i=x_i)}{\binom{n}{y} p^y (1-p)^{n-y}} =$$

$$= \frac{\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}}{\binom{n}{y} p^y (1-p)^{n-y}} = \binom{n}{y}^{-1}$$

<u>Т-ма</u> (За факторизацията)

U - достатъчна $\iff g(U,\theta)\, h(\underbrace{X_1, \dots, X_n}_{\text{няма } \theta})$

<u>Опр.</u> $X_1, \dots, X_n$ iid, $X \in f(x,\theta)$

$U = g(X_1, \dots, X_n)$

$P(X_1, \dots, X_n \mid U)$ - не зависи от θ - <u>достатъчна статистика</u>

<u>Пр.</u> $X_1, \dots, X_n \in Exp(\frac{1}{a})$ iid

$\mathbb{E}X = a$

$f(x) = \frac{1}{2} e^{-\frac{x}{a}}$

$$L(X_1,\dots,X_n;\alpha)=\prod_{i=1}^{n}\frac{1}{\alpha}e^{-\frac{X_i}{\alpha}}=\frac{1}{\alpha^n}e^{-\frac{1}{\alpha}\underbrace{\sum_{i=1}^{n}X_i}_{\text{достатъчна статистика}}}$$

Т-ма (Lehmann - Scheffè)

$X_1,\dots,X_n$ и $Y_1,\dots,Y_n$ iid

$\dfrac{L(X_1,\dots,X_n;\theta)}{L(Y_1,\dots,Y_n;\theta)}$ не зависи от θ $\Leftrightarrow$ $g(X_1,\dots,X_n)=g(Y_1,\dots,Y_n)$

<u>минимална дост. статистика</u>

Пр. 1

$X_1,\dots,X_n$ и $Y_1,\dots,Y_n\in$ Bernoulli(p) i.i.d.

$$\frac{L(X_1,\dots,X_n)}{L(Y_1,\dots,Y_n)}=\frac{p^{\sum_{i=1}^{n}X_i}(1-p)^{n-\sum_{i=1}^{n}X_i}}{p^{\sum_{i=1}^{n}Y_i}(1-p)^{n-\sum_{i=1}^{n}Y_i}}=$$

$=p^{\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}Y_i}(1-p)^{\sum_{i=1}^{n}Y_i-\sum_{i=1}^{n}X_i}$ - не зависи от p при $\overline{X}_n=\overline{Y}_n$

Пр. 2

$X_1,\dots,X_n$ и $Y_1,\dots,Y_n\in No(\mu,\sigma^2)$ iid

$\vec{\theta}=(\mu,\sigma^2)$

$$\frac{L(X_1,\dots,X_n;\vec{\theta})}{L(Y_1,\dots,Y_n;\vec{\theta})}=\frac{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{(X_i-\mu)^2}{2\sigma^2}}}{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{(Y_i-\mu)^2}{2\sigma^2}}}=$$

$= e^{-\frac{1}{2\sigma^2}\left[\left(\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} Y_i^2\right) - 2\mu\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right)\right]}$

Не зависи от μ при $\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$

Не зависи от σ^2 при $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i^2$

$\hat{\theta} = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$

<u>Пр. 3</u> $X_1, \dots, X_n$ и $Y_1, \dots, Y_n$ са iid с плътност

$$f_X(x) = \begin{cases} \frac{2x}{\theta} e^{-\frac{x^2}{\theta}}, & x > 0 \\ 0, & x \le 0 \end{cases}$$ – разпр. на Релей.

Търсим ефективна оценка за θ.

$$\frac{L(X_1, \dots, X_n | \theta)}{L(Y_1, \dots, Y_n | \theta)} = \frac{\prod_{i=1}^{n} \frac{2X_i}{\theta} e^{-\frac{X_i^2}{\theta}}}{\prod_{i=1}^{n} \frac{2Y_i}{\theta} e^{-\frac{Y_i^2}{\theta}}} = \frac{X_1 \dots X_n}{Y_1 \dots Y_n} e^{-\frac{1}{\theta}\left[\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} Y_i^2\right]}$$

$\sum_{i=1}^{n} X_i^2$ е дост. статистика за θ

<u>Т-ма</u> Неизместени оценки, които са ф-ции на дост. стат., са ефективни

$\frac{1}{n} \mathbb{E} \sum_{i=1}^{n} X_i^2 = \frac{1}{n} n \mathbb{E} X_1^2 = \mathbb{E} T$, където $T = X_1^2$

$$f_T(x) = \begin{cases} \frac{1}{2\sqrt{x}} \frac{2\sqrt{x}}{\theta} e^{-\frac{x}{\theta}} = \frac{1}{\theta} e^{-\frac{x}{\theta}}, & x > 0 \\ 0, & x \le 0 \end{cases} \Rightarrow T \in Exp\left(\frac{1}{\theta}\right) \Rightarrow \mathbb{E} T = \theta$$

$\Rightarrow \hat{\theta} = \sum_{i=1}^{n} X_i^2$ е неизместена $\Rightarrow \hat{\theta}$ е ефективна

ефективна оценка - efficient estimator
достатъчна статистика - sufficient statistic
MVUE - minimum variance ~~un~~ unbiased estimator
максимално правдоподобие - maximum likelihood

оценка $\begin{cases} \text{теоретична - estimator} \\ \text{емпирична - estimate} \end{cases}$

<u>Т-ма</u> (Rao-Blackwell) Нека U е дост. за $X_1, \dots, X_n$ и нека T е неизместена и не е ф-я на U. Тогава $g(U) = \mathbb{E}(T|U)$ е ефективна и $\mathbb{D}g(U) \leq \mathbb{D}T$

<u>Пр.1</u> $X_1, \dots, X_n \in Bernoulli(p)$

Търсим ефективна оценка за $\mathbb{D}X = p(1-p)$

Нека T е сл. вел., за която $P(T=1) = p(1-p)$ ако $X_1 = 1$ и $X_2 = 0$ и $P(T=0) = 1 - P(T=1)$ (т.е. T е бинарна)

$\mathbb{E}T = p(1-p)$

$$P(T=1 | \sum_{i=1}^n X_i = y) = \frac{P(X_1 = 1, X_2 = 0, \sum_{i=1}^n X_i = y)}{P(\sum_{i=1}^n X_i = y)} = \frac{P(X_1=1)P(X_2=0)P(\sum_{i=3}^n X_i = y-1)}{P(\sum_{i=1}^n X_i = y)}$$

$$= \frac{p(1-p)\left[\binom{n-2}{y-1} p^{y-1}(1-p)^{n-y-1}\right]}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{\binom{n-2}{y-1}}{\binom{n}{y}} = \frac{(n-2)! \cdot y!(n-y)!}{(y-1)!(n-y-1)!\, n!} =$$

$$= \frac{y(n-y)}{n(n-1)} = \frac{n}{n-1}\left[\frac{y}{n}\left(1-\frac{y}{n}\right)\right] = \underbrace{\frac{n}{n-1}\left(\overline{X}_n(1-\overline{X}_n)\right)}_{\text{ефективна оценка}}$$

Т-ма (Rao-Cramer)

$$\frac{\partial}{\partial \theta} \ln L(X_1, \dots, X_n; \theta) = K(\theta)\Big[\underbrace{t_n(\vec{X})}_{\text{ефективна оценка за } r(\theta)} - r(\theta)\Big]$$

Заб. Ако θ и $r(\theta)$ са линейно зависими, θ също е ефективна

Заб. Ефективна оценка, получена от т-мата на Рао-Крамер, е единствена

13.03.2018

$X_1, \dots, X_n \in F(x, \theta)$

$\hat{\theta} = t_n(X_1, \dots, X_n)$ - точкова оценка

Доверителен интервал:

$I_\gamma = [t_n^1(X_1, \dots, X_n), t_n^2(X_1, \dots, X_n)]$

$t_n^1 \quad \theta \quad t_2^2$

$t_n^1 \le t_n^2$

ниво на съгласие

$P(\theta \in I_\gamma) = \underbrace{\gamma}_{\text{ниво на доверие}} = 1 - \alpha$

$X_1, \dots, X_n \in No(\mu, \sigma^2)$

$\underbrace{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2}_{(n-1)\bar{S}_n^2} \cdot \frac{1}{\sigma^2} \in \chi^2(n-1)$

$$\frac{\left(\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}\right)}{\sqrt{\frac{(n-1)\bar{S}_n^2}{(n-1)\sigma^2}}} = \frac{(\bar{X} - \mu)\sqrt{n}}{\bar{S}_n} \in t(n-1)$$

$$P\left(-z_{\frac{1+\gamma}{2}} < \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} < z_{\frac{1+\gamma}{2}}\right) = P\left(\bar{X}_n - z_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\bar{X}_n - t_{\frac{1+\gamma}{2}} \frac{\bar{S}_n}{\sqrt{n}} < \mu < \bar{X}_n + t_{\frac{1+\gamma}{2}} \cdot \frac{\bar{S}_n}{\sqrt{n}}\right)$$

Пример 1

$\bar{X}_n = 0,5$ $\quad \begin{cases} H_0: \mu = 0 \\ H_1: \mu \neq 0 \end{cases}$

$n = 8$

$\bar{S}_n = 1$

$\alpha = 0,05$ Реш.

$$0,5 - 2,365 \cdot \frac{1}{\sqrt{8}} < \mu < 0,5 + 2,365 \cdot \frac{1}{\sqrt{8}}$$

$$-0,34 < \mu < 1,34$$

с 95% доверителен интервал $\mu = 0$

Метод на максималното правдоподобие:

$X_1, \dots, X_n \in F(x, \theta)$

$\hat{\theta}$ - максимално правдоподобна оценка (м.п.о.)

- състоятелна, ако $\hat{\theta}_n \xrightarrow[n \to \infty]{P} \theta$
- асимптотично неизместена: $\lim\limits_{n \to \infty} \mathbb{E}\,\theta_n = \theta$
- асимптотично ефективна: $\mathbb{D}\,\theta_n \to \min$
- асимптотично нормална:

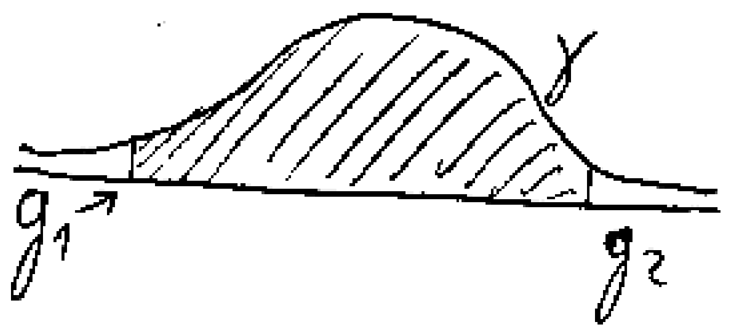
$\sqrt{n I(\theta)}\left(\hat{\theta} - \theta\right) \xrightarrow[n \to \infty]{d} Y \in N_0(0,1)$, където

$\underbrace{I(\theta)}_{\text{инф. на Фишер}} = \mathbb{E}\left[\frac{\partial}{\partial \theta} \ln f(x, \theta)\right]^2 = \mathbb{D}\left[\frac{\partial}{\partial \theta} \ln f(x, \theta)\right]$, защото

$\left(\mathbb{E}\left[\frac{\partial}{\partial \theta} \ln f(x, \theta)\right]\right)^2 = 0$. Наистина, $\mathbb{E}\left[\frac{\partial}{\partial \theta} \ln f(x, \theta)\right] =$

$= \int_{\mathbb{R}} \frac{\partial}{\partial \theta} \ln f(x, \theta) \cdot f(x, \theta)\, dx = \int_{\mathbb{R}} \frac{\partial}{\partial \theta} f(x, \theta) \cdot \frac{1}{f(x, \theta)} f(x, \theta)\, dx = \frac{\partial}{\partial \theta} \overbrace{\int_{\mathbb{R}} f(x, \theta)\, dx}^{=1} = 0$

$G(x,\theta) := \sqrt{n I(\theta)}\,(\hat{\theta} - \theta)$



$P(g_1 \le G(x,\theta) \le g_2) = \gamma$

$g_1 = -g_2$ и $g_2 = z_{\frac{1+\gamma}{2}}$

$\sqrt{n I(\theta)}\,(\hat{\theta} - \theta) = g_1 \Rightarrow \theta - \hat{\theta} - \frac{g_1}{\sqrt{n I(\theta)}} = 0$

Получаваме интервал

$$\left(\hat{\theta} - z_{\frac{1+\gamma}{2}} / \sqrt{n I(\theta)},\ \hat{\theta} + z_{\frac{1+\gamma}{2}} / \sqrt{n I(\theta)}\right)$$

<u>Пр.2</u> $X_1, \dots, X_2$, $X_n \in Ge_1(p)$ $\quad (\mathbb{D}X_n = (1-p)/p^2)$

$I(p) = \mathbb{D}\frac{\partial}{\partial p}\ln\left(p(1-p)^k\right) = \mathbb{D}\left(\frac{\partial}{\partial p}\left(\ln p + k\ln(1-p)\right)\right) =$

$= \mathbb{D}\left(\underbrace{\frac{1}{p}}_{=0} + \frac{k}{(1-p)}\right) = \frac{\mathbb{D}k}{(1-p)^2} = \frac{\cancel{p^2}\,(1-p)}{(1-p)^2 p^2} = \left[(1-p)p^2\right]^{-1}$

$$L(X_1, \dots, X_n; p) = \prod_{i=1}^{n} p(1-p)^{X_i} = p^n (1-p)^{\sum_{i=1}^{n} X_i} \Rightarrow$$

$$\Rightarrow \ln L = n \ln p + \sum_{i=1}^{n} X_i \ln(1-p) \Rightarrow \frac{\partial}{\partial p} \ln L = \frac{n}{p} - \frac{\sum^{n} X_i}{1-p}$$

$$\Rightarrow n(1-p) = p \sum_{i=1}^{n} X_i \Rightarrow \underbrace{\hat{p} = \frac{1}{1+\overline{X}_n}}_{\text{заместваме в } I(p)} \text{ - максимално правдоподобна оценка}$$

$$I(\hat{p}) = \frac{(1+\overline{X}_n)^2}{\left(\frac{\overline{X}_n}{1+\overline{X}_n}\right)} = \frac{(1+\overline{X}_n)^3}{\overline{X}_n}$$

$\Rightarrow$ дов. инт. в явна форма е

$$\frac{1}{1+\overline{X}_n} \pm z_{\frac{1+\gamma}{2}} \frac{\sqrt{\overline{X}_n}}{\sqrt{n(1+\overline{X}_n)^3}}$$

20.09.2018

$(X_1, \dots, X_n)$, $X \in F(x, \theta)$

$\theta \in \Theta$

$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta = \theta_1 \end{cases}$
проста хипотеза
проста алтернатива

$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \,/\, \theta > \theta_0 \,/\, \theta < \theta_0 \end{cases}$
проста хипотеза
сложна алтернатива

$\begin{cases} H_0: \theta \in \Theta_0 \subseteq \Theta \\ H_1: \theta \in \Theta \setminus \Theta_0 \end{cases}$
сложна хипотеза
сложна алтерн.

$\alpha = P(H_1 | H_0) = P(\vec{X} \in W | H_0)$ - грешка от първи род

$\beta = P(H_0 | H_1)$ - грешка от втори род

$\left.\begin{matrix} W \text{ - критична} \\ \overline{W} \text{ - допустима} \end{matrix}\right\}$ област

При фиксирано α, можем единствено да намалим β, или, еквивалентно, да максимизираме $\pi = P(H_1 | H_1) = 1 - \beta$.

Пр. 1

М	Ж
$n_1 = 50$	$n_2 = 50$
$\overline{X}_1 = 3{,}60$	$\overline{X}_2 = 3{,}80$
$\overline{S}_1^2 = 0{,}18$	$\overline{S}_2^2 = 0{,}14$

$X_1, \dots, X_{n_1}$, $X \in No(\mu_1, \sigma_1^2)$ за М

$Y_1, \dots, Y_{n_2}$, $Y \in No(\mu_2, \sigma_2^2)$ за Ж

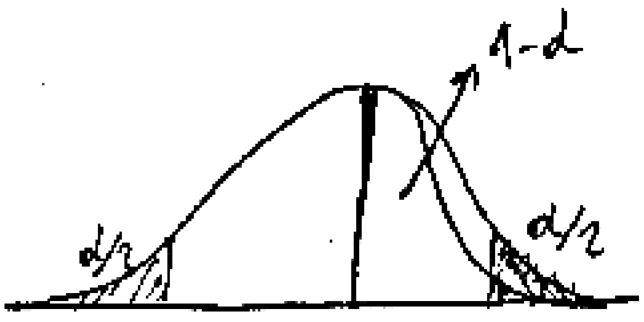
$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases} \iff \begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \neq 0 \end{cases}$

$\alpha = 0{,}05$

$\begin{cases} H_0: \mu_1 - \mu_2 = d_0 \\ H_1: \mu_1 - \mu_2 \neq d_0 \end{cases}$

1 сл.) σ_1^2, σ_2^2 - известни

$\overline{X}_1 - \overline{X}_2 \in No\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$



$T = \dfrac{\overline{X}_1 - \overline{X}_2 - \overbrace{(\mu_1 - \mu_2)}^{d_0}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$; $\alpha = P(H_1 | H_0) = P(|T| > z_{\frac{\alpha}{2}})$

2) σ_1^2 и σ_2^2 са неизвестни, но имаме основание да твърдим, че са равни

$$T = \frac{\overline{X}_1 - \overline{X}_2 - d_0}{\hat{\sigma}^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$, където $\hat{\sigma}^2$ е обобщена оценка; $T \in t(n_1+n_2-2)$

$$\hat{\sigma}^2 = S_p^2 = \frac{(n_1-1) S_1^2 + (n_2-1) S_2^2}{n_1+n_2-2}$$

3) σ_1^2 и σ_2^2 са неизвестни, но имаме основание да твърдим, че $\sigma_1^2 \neq \sigma_2^2$

$$T = \frac{\overline{X}_1 - \overline{X}_2 - d_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \in t(V), \quad V = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}$$

продължение на зад. 1 (решение)

$$T = \frac{3.6 - 3.8 - 0}{\sqrt{\frac{0.18}{50} + \frac{0.14}{50}}} \approx -2{,}5$$

$$W: \underset{2,5}{|T|} > z_{1-\frac{\alpha}{2}} = z_{0,975} = 1{,}96 \Rightarrow$$ отхвърляме H_0

Зад. 2

$n = 36$

$\overline{X}_n = 17$

$S_1^2 = 9$

$H_0: \mu = 15$

$H_1: \mu > 15$

Реш.

$$T = \frac{\overline{X}_n - \mu}{\sqrt{\sigma / n}} > z_{1-\alpha} \quad ; \quad \frac{17-15}{\sqrt{9/36}} = \frac{2}{2^{-1}} = 4 > 1{,}645 \Rightarrow$$ Приемаме H_1

27.03.2018

<u>Зад. 1</u> $n = 36$

$$\begin{cases} H_0: \mu = 15 \\ H_1: \mu > 15 \end{cases}$$

$\alpha = 0,05$

$\overline{X}_n = 19$

$\overline{S}_n^2 = 9$

$$\beta = \alpha = P(\overline{X}_n > c \mid \mu = \overset{15}{\mu_0}) = P\left(\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} > \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = P(Z > Z_{1-\alpha})$$

$$\frac{c - \mu_0}{\sigma/\sqrt{n}} = Z_{1-\alpha} \Rightarrow c = \mu_0 + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

$$P = P(H_0 | H_1) = P(\overline{X}_n \le c \mid \mu = \overset{16}{\mu_a}) = P\left(\frac{\overline{X}_n - \mu_a}{\sigma\sqrt{n}} < \frac{c - \mu_a}{\sigma\sqrt{n}}\right) = P(Z < Z_{1-\alpha})$$

$$\frac{c - \mu_0}{\sigma/\sqrt{n}} = Z_\beta = Z_{1-\beta} \Rightarrow c = \mu_a - Z_{1-\beta}\frac{\sigma}{\sqrt{n}}$$

$$\mu_0 + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = \mu_a + Z_{1-\beta}\frac{\sigma}{\sqrt{n}} \Rightarrow \mu_0 - \mu_a = \frac{\sigma}{\sqrt{n}}(Z_{1-\beta} + Z_{1-\alpha})$$

$$n = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 \sigma^2}{(\mu_a - \mu_0)^2} = \frac{(1,645 + 1,645)^2 \cdot 9}{(16-15)^2} = 97,4$$

Критерий на знаците

$X_1, \dots, X_n$, $X \in F_X(x)$ $\qquad p = P(X > Y)$

$Y_1, \dots, Y_n$, $Y \in F_Y(y)$

$\begin{cases} H_0 : p = \frac{1}{2} \\ H_1 : p \neq \frac{1}{2} \end{cases}$

определяме α от
$\mu = 0$ или $\mu = 10$:

$$\alpha = P(\mu = 0 \mid p = \tfrac{1}{2}) + P(\mu = 10 \mid p = \tfrac{1}{2}) = \binom{10}{0}(0,5) + \binom{10}{10}(0,5)^{10} = 0,002$$

$$\mu \in \{0, 1, 9, 10\} \Rightarrow \alpha = 0,022$$

$$|D_{(1)}| \leq |D_{(2)}| \leq \dots \leq |D_{(k)}| \qquad \text{(Уилкоксън)}$$

T^+ = сумата на положителните рангове в D_i

T^- = —— || —— отрицателните —— || ——

$$\mathbb{E}T^+ = \frac{n(n+1)}{4} \qquad \mathbb{D}T^+ = \frac{n(n+1)(2n+1)}{24}$$

<u>Зад. 2</u> За един и същ образец на чиста повърхност са проведени по 12 измервания с два двойни микроскопа с номера 61 и 263. Измерванията са дадени в таблица 1.

i	№	1	2	3	4	5	6	7	8	9	10	11	12
(ξ)	61	0,8	1,9	3,0	3,5	3,8	2,5	1,7	0,9	1,0	2,3	3,3	3,4
(η)	263	1,4	2,1	3,1	3,6	2,7	1,8	1,1	0,2	1,6	2,8	4,0	4,7

При ниво на съгласие $\alpha = 0,05$ може ли да се счита, че между показанията на уредите няма систематични различия и че те имат еднаква разделителна способност?

$\underline{\text{Реш.}}$ Вар. ред на $\{\xi_i\} \cup \{\eta_i\}$

$0{,}2 < \overset{1}{\underline{0{,}8}} < \overset{1}{\underline{0{,}9}} < \overset{1}{\underline{1{,}0}} < 1{,}1 < 1{,}4 < 1{,}6 < \overset{4}{\underline{1{,}7}} < 1{,}8 < \overset{5}{\underline{1{,}9}} < 2{,}1 < \overset{6}{\underline{2{,}3}} < \overset{6}{\underline{2{,}5}} < 2{,}7 < 2{,}8 < \overset{8}{\underline{3{,}0}} < 3{,}1 < \overset{9}{\underline{3{,}3}} < \overset{9}{\underline{3{,}4}} < \overset{9}{\underline{3{,}5}} <$
$< 3{,}6 < \underset{12}{\underline{3{,}8}} < 4{,}0 < 4{,}\ldots$

(Подчертани са стойностите от $\{\xi_i\}$)

$$\overline{U} = 3.1 + 4 + 5 + 2.6 + 8 + 3.9 + 12 = 71$$

$$\begin{cases} H_0: U = \frac{12.12}{2} = 72 \\ H_1: U \neq 72 \end{cases}$$

$$W: \left| \frac{\overline{U} - 72}{\underset{17{,}3}{\underset{\shortparallel}{DU}}} \right| = 0{,}058 < 1{,}96 \quad (\alpha = 0{,}05)$$

$\underline{\text{Зад. 3}}$ По статистически данни разпределението на доходите по възрастовите групи на 40-годишните (на възраст от 40 до 50) и на 50-годишните (от 50г. до 60г.) работници е показано на таблица 2

	<1000	1000-2000	2000-3000	3000-4000	4000-6000	>6000
40	7831	26740	35572	20009	11527	6919
50	7558	20685	24185	12280	6776	4222

Може ли да се счита значителна разликата в разпределението на доходите между 40-годишните и 50-годишните работници?

$\underline{\text{Реш.}}$ $\begin{cases} H_0: F_X = F_Y \\ H_1: F_X \neq F_Y \end{cases}$ $\quad \alpha = 0{,}01$

$$\hat{\chi}^2 = n_1 n_2 \sum_{i=1}^{r} \frac{\left(\frac{n_1^i}{n_1} - \frac{n_2^i}{n_2} \right)^2}{n_1^i + n_2^i} \sim \chi^2(r-1)$$

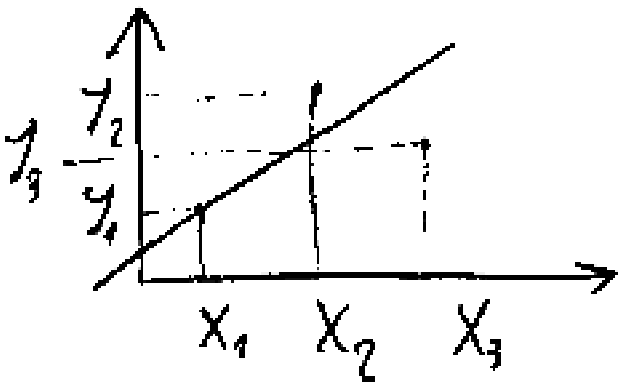
23.04.2018

$\underline{Y} = \beta_0 + \beta_1 \underline{X} + \underline{\varepsilon}$, грешка (адитивна)

отклик — predictor

$X_1, X_2, \ldots, X_n$

$Y_1, Y_2, \ldots, Y_n$



$\mathbb{E}\varepsilon = 0, \mathbb{D}\varepsilon = \sigma^2$

$\mathbb{E}Y = \beta_0 + \beta_1 X$

$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X$

β_0, β_1 - параметри на модела

σ^2 - „скрит" параметър

$Y = \beta_0 x^{\beta_1} \cdot \varepsilon$ — мултипликативна грешка — нелинеен модел

$\ln Y = \ln \beta_0 + \beta_1 \ln x + \ln \varepsilon$ - линеаризиран модел

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$

$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$ - полиномна регресия

$$\left.\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_1^1 + \beta_2 x_2^1 + \ldots + \beta_n x_n^1 + \varepsilon_1 \\ &\vdots \\ Y_m &= \beta_0 + \beta_1 x_1^m + \beta_2 x_2^m + \ldots + \beta_n x_n^m + \varepsilon_m \end{aligned}\right\} \iff \vec{Y} = X\vec{\beta} + \vec{\varepsilon}$$

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

sum of squares for error

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_1) = 0 \Rightarrow \hat{\beta}_0 = \bar{Y}_n - \beta_1 \bar{X}_n$$

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_n \qquad \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_n$$

$$\frac{\partial SSE}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_1)x_i = 0 \Rightarrow \hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$\hat{\vec{\beta}} = (X'X)^{-1} X' \vec{Y}$$

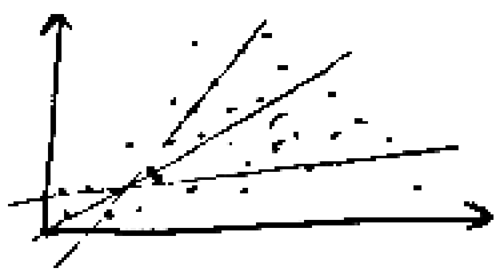
$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(y_i - \overline{y}_n)}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \sum_{i=1}^{n}(y_i - \overline{y}_n)^2}}$$

$$\mathbb{E}\hat{\beta}_1 = \beta_1$$

$$\mathbb{E}\hat{\beta}_0 = \mathbb{E}(\overline{y}_n - \hat{\beta}_1 \overline{X}_n)$$

17.04.2018

$$\vec{y} = \underbrace{\beta_0 + \beta_1 \vec{x}}_{\perp\!\!\perp} + \vec{\varepsilon}$$



$$\underbrace{\sum_{i=1}^{n} \varepsilon_i^2}_{SSE} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \longrightarrow \min_{\beta_0, \beta_1}$$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

Предполагаме, че $\vec{X}$ са числа, а $\vec{\varepsilon}$ - случайни величини

$\Rightarrow \vec{Y}$ също е случайна величина

$$\widehat{\rho(X,Y)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \sum_{i=1}^{n} (y_i - \bar{y}_n)^2}}$$

$$\Rightarrow \hat{\beta}_1 = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y}_n)^2}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}}\, \widehat{\rho(X,Y)}$$

Освен това $\hat{\beta}_1 = \dfrac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x}_n) y_i}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$

$\dfrac{\partial SSE}{\partial \beta_0} = 0$ и $\dfrac{\partial SSE}{\partial \beta_1} = 0$ ако $\left|\begin{array}{l} n \hat{\beta}_0 + \sum_{i=1}^{n} x_i \hat{\beta}_1 = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i \hat{\beta}_0 + \sum_{i=1}^{n} x_i \hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i \end{array}\right\}$ нормални уравнения

$\Updownarrow$

$$(X^T X)\hat{\vec{\beta}} = X^T \vec{Y}$$

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{Y}$$

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\vec{y}} = \underbrace{\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{X} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\vec{y} = X \vec{\beta} + \vec{\varepsilon}$$

$\mathbb{E}\hat{\beta}_0 = \beta_0$ и $\mathbb{E}\hat{\beta}_1 = \beta_1$

β_0 и β_1 - линейни по y

$$\mathbb{D}\hat{\beta}_0 = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x}_n)^2} \qquad \mathbb{D}\hat{\beta}_1 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\overline{X}_n \sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

Нека сега да намерим доверителни интервали за β_0 и β_1

$$P(\underbrace{\hat{\beta}_{0,1} \le \beta_0 \le \hat{\beta}_{0,2}}_{S_0}) = 1 - \alpha$$

$$P(\underbrace{\hat{\beta}_{1,1} \le \beta_1 \le \hat{\beta}_{1,2}}_{S_1}) = 1 - \alpha$$

$$P(S_0 \cap S_1) = 1 - P(\overline{S}_0 \cup \overline{S}_1) = 1 - \overset{\alpha}{P(\overline{S}_0)} - \overset{\alpha}{P(\overline{S}_1)} + \underset{\overset{\vee\!|}{0}}{P(\overline{S}_1 \cap \overline{S}_2)} \ge 1 - 2\alpha$$

Ако $\overline{X}_n = 0 \Rightarrow Cov(\hat{\beta}_0, \hat{\beta}_1) = 0 \Leftrightarrow \hat{\beta}_0 \perp\!\!\!\perp \hat{\beta}_1$

Ако $\varepsilon \in \mathcal{N} \Rightarrow \hat{\beta}_0 \perp\!\!\!\perp \hat{\beta}_1$

$\vec{\beta} = (\beta_0 \ \ \beta_1)^T$

$\hat{\vec{\beta}} = (\hat{\beta}_0 \ \ \hat{\beta}_1)^T$

$$\mathbb{E}\hat{\vec{\beta}} = \begin{pmatrix} \mathbb{E}\hat{\beta}_0 \\ \mathbb{E}\hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \vec{\beta}$$

$$\mathbb{D}\hat{\vec{\beta}} = Cov(\hat{\vec{\beta}}, \hat{\vec{\beta}}) = \begin{pmatrix} \dfrac{\sigma^2 \sum_{i=1}^{n} X_i}{n \sum_{i=1}^{n}(X_i - \overline{X}_n)} & -\dfrac{\overline{X}_n \sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} \\ -\dfrac{\overline{X}_n \sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} & \dfrac{\sigma^2}{\sum(X_i - \overline{X}_n)^2} \end{pmatrix}$$

$\mathbb{D}\hat{\vec{\beta}}$ е симетрична и неотрицателно определена

$\mathbb{D}\vec{\beta} = \sigma^2 (X^T X)^{-1}$

Пр. 1

						Σ
x	-2	-1	0	1	2	0
y	0	0	1	1	3	5

$\hat{y} = 1 + 0{,}7X$

$$\begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$$

$\varepsilon = y - \beta_0 - \beta_1 X$

$\hat{\varepsilon} = y - \hat{\beta}_0 - \hat{\beta}_1 X$

$\hat{\varepsilon} = 1 - 1 - 0{,}7X = 1 - 1 - 0{,}7.1 = 0{,}7$

$$\mathbb{D}\vec{\beta} = \sigma^2 (X^T X)^{-1} = \sigma^2 \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}^{-1} = \begin{pmatrix} 1/5 & 0 \\ 0 & 1/10 \end{pmatrix} \sigma^2 \Rightarrow \begin{array}{l} \mathbb{D}\hat{\beta}_0 = \frac{\sigma^2}{5} \\ \mathbb{D}\hat{\beta}_1 = \frac{\sigma^2}{10} \\ Cov(\hat{\beta}_0, \hat{\beta}_1) = 0 \end{array}$$

Оценка на σ^2

$\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$

$SSE := \sum_{i=1}^{n} \hat{\varepsilon}_i^2$

sum of squares for errors

$$\mathbb{E}\,SSE = \mathbb{E} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \mathbb{E} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 =$$

$$= \mathbb{E} \sum_{i=1}^{n} (y_i - \bar{y}_n + \hat{\beta}_1 \bar{X}_n - \hat{\beta}_1 X_i)^2 = \mathbb{E} \sum_{i=1}^{n} \left((y_i - \bar{y}_n) - \hat{\beta}_1 (X_i - \bar{X}_n)\right)^2 =$$

$$= \mathbb{E}\left(\sum_{i=1}^{n} (y_i - \bar{y}_n)^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X}_n)(y_i - \bar{y}_n)\right) =$$

$$= \mathbb{E}\left(\sum_{i=1}^{n} (y_i - \bar{y}_n)^2 - \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right) = \sum_{i=1}^{n} \mathbb{E} Y_i^2 - n\,\mathbb{E}(\bar{y}_n)^2 - (\mathbb{E}\hat{\beta}_1^2) \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 =$$

$$= \sum_{i=1}^{n} \left(\sigma^2 + (\beta_0 + \beta_1 \cancel{x})^2\right) - n\left(\frac{\sigma^2}{n} + (\beta_0 + \beta_1 x_n)^2\right) - \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \left[\frac{\sigma^2}{\sum (X_i - \bar{X}_n)^2} + \beta_1^2\right] =$$

$$= (n-2)\sigma^2.$$

$\Rightarrow \underbrace{\frac{SSE}{n-2}}_{MSE}$ е неизместена оценка за σ^2

(Mean square error)

$$SSE = (\vec{y} - X\hat{\vec{\beta}})^T (\vec{y} - X\hat{\vec{\beta}}) = \vec{y}^T\vec{y} - \vec{y}^T X\hat{\vec{\beta}} - \hat{\vec{\beta}}^T X^T \vec{y} + \hat{\vec{\beta}}^T X^T X \hat{\vec{\beta}} =$$

$$= \vec{y}^T\vec{y} - 2\hat{\vec{\beta}}^T X^T \vec{y} + \hat{\vec{\beta}}^T X^T \vec{y} = \vec{y}^T\vec{y} - \hat{\vec{\beta}}^T X^T \vec{y}$$

$$SSE = (0\ 0\ 1\ 1\ 3)\begin{pmatrix}0\\0\\1\\1\\3\end{pmatrix} - (1\quad 0{,}7)\begin{bmatrix}1 & 1 & 1 & 1 & 1\\ -2 & -1 & 0 & 1 & 2\end{bmatrix}\begin{pmatrix}0\\0\\1\\1\\3\end{pmatrix} = 11 - 9{,}9 = 1{,}1$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1.1}{3} \approx 0{,}367$$

24.04.2018

$\vec{Y} = \beta_0 + \beta_1 \vec{x} + \underbrace{\vec{\varepsilon}}_{\in N(0,\sigma^2)}$ $\qquad \hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{Y}$

$\hat{\beta}_i \in N(\beta_i, c_{ii}\sigma^2)$

$E\hat{\beta}_i = \beta_i$

$D\hat{\beta}_i = c_{ii}\sigma^2$, където $C = (X^T X)^{-1}$

$\Rightarrow Y \in N(\beta_0 + \beta_1 x, \sigma^2)$

книга
Север. лин. регресия

$\begin{cases} H_0: \beta_i = \beta_{i0} \\ H_1: \beta_i \neq \beta_{i0} \end{cases}$

$Z = \dfrac{\hat{\beta}_i - \beta_{i0}}{\sigma\sqrt{c_{ii}}} \overset{H_0}{\in} N(0,1)$ - σ е известно

$W: |Z| > Z_{\alpha/2}$

$t = \dfrac{\hat{\beta}_i - \beta_{i0}}{\sqrt{MSE \cdot c_{ii}}} \overset{H_0}{\in} t(n-2)$ - σ е неизвестно

$\hat{\beta}_i \perp\!\!\!\perp MSE$

Пр.1 $\begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$ $\qquad \alpha = 0,05$

$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$

$MSE = 0,367 = \hat{\sigma}^2$

$t = \dfrac{\hat{\beta}_1 - 0}{\sqrt{MSE\, c_{ii}}} = \dfrac{0,7}{\sqrt{0,367 \cdot \frac{1}{10}}} = 3,65$

$W: |3,65| > t_{\frac{0,05}{2}, 3} = 3,182$

отхвърляме H_0, тоест $\beta_1 \neq 0$

p-value $= 2P(T > 3,65)$

Дов. интервал за $\quad W: |t| = \left| \frac{\hat{\beta}_i - \beta_{i0}}{\sqrt{c_{ii} MSE}} \right| > t_{\alpha \cdot \frac{1}{2}, n-1}$

$$\beta_i \in \left(\hat{\beta}_i \pm t_{\alpha/2, n-2} \sqrt{c_{ii}} \sqrt{MSE} \right.$$

За пример 1 $\quad \beta_1 \in (0,7 \pm 3,182 \sqrt{0,167 \cdot \frac{1}{16}}) = (0,072, 1,31) \Rightarrow \beta_1 \neq 0 \notin (0,072, 1,31)$

Избираме a_0 и $a_1 \in \mathbb{R}$

$$U = a_0 \beta_0 + a_1 \beta_1 \in \mathcal{N}\Big(\underbrace{a_0\beta_0 + a_1\beta_1}_{\vec{a}^T \cdot \vec{\beta}}, \sigma^2 [\vec{a}^T (X^T X)^{-1} \vec{a}]\Big)$$

$$\mathbb{E}(a_0 \hat{\beta}_0 + a_1 \hat{\beta}_1) = a_0 \beta_0 + a_1 \beta_1$$

$$\mathbb{D}(a_0 \hat{\beta}_0 + a_1 \hat{\beta}_1) = a_0^2 \mathbb{D}\hat{\beta}_0 + a_1^2 \mathbb{D}\hat{\beta}_1 + a_0 a_1 \operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) + a_1 a_0 \operatorname{cov}(\hat{\beta}_1, \hat{\beta}_0)$$

$$= a_0^2 c_{00} \sigma^2 + a_1 c_{11} \sigma^2 + a_0 a_1 c_{10} \sigma^2 + a_1 a_0 c_{01} \sigma^2 = \sigma^2 [\vec{a}^T (X^T X)^{-1} \vec{a}]$$

$$\frac{U - \mathbb{E}U}{\sqrt{\mathbb{D}U}} \in \mathcal{N}(0,1)$$

$$\begin{cases} H_0: \vec{a}^T \vec{\beta} = \vec{a}^T \vec{\beta}_0 \\ H_1: \vec{a}^T \vec{\beta} \neq \vec{a}^T \vec{\beta}_0 \end{cases}$$

$$Z = \frac{U - \mathbb{E}_0 U}{\sqrt{\vec{a}^T (X^T X)^{-1} \vec{a}}\, \sigma} \qquad a_0 \beta_0 + a_1 \beta_1 \in \left(U \pm z_{\alpha/2} \sigma \sqrt{\vec{a}^T (X^T X)^{-1} \vec{a}} \right.$$

$$W: |Z| > z_{\alpha/2}$$

$$T = \frac{U - \mathbb{E}_0 U}{\sqrt{\vec{a}^T (X^T X)^{-1} \vec{a}\, MSE}} \qquad a_0 \beta_0 + a_1 \beta_1 \in \left(U \pm t_{\alpha/2, n-2} \sqrt{\vec{a}^T (X^T X)^{-1} \vec{a}\, MSE} \right.$$

$$W: |T| > t_{\alpha/2, n-2}$$

90% доверителен интервал при $x=1$

$\mathbb{E}Y = \beta_0 + \beta_1 x$

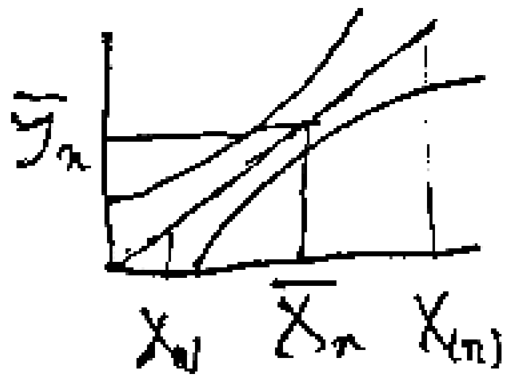
$\mathcal{U} = \hat{Y} = 1 + 0{,}7x_0,\ x_0 = 1$

$\vec{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 1 \\ x_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$$\mathbb{D}\hat{Y} = \sigma^2 \vec{a}^T (X^TX)^{-1}\vec{a} = \sigma^2 (1, 1)\begin{pmatrix} 1/5 & 0 \\ 0 & 1/10 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \sigma^2 \left(1/5 \quad 1/10\right)\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{3\sigma^2}{10}$$

$$\mathbb{E}Y \in \left(\hat{Y} \pm t_{\alpha/2,3}\sqrt{MSE\,\vec{a}^T(X^TX)^{-1}\vec{a}}\right) = \left(1 + 0{,}7x_0 \pm 2{,}353\sqrt{0{,}367 \cdot 0{,}3}\right) = (0{,}919,\ 2{,}481)$$

$$\vec{a}^T(X^TX)^{-1}\vec{a} = \frac{1}{n} + \frac{(X_0 - \bar{X}_n)^2}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$



$\hat{\varepsilon} = Y - \hat{Y}$

$\mathbb{E}\hat{\varepsilon} = \mathbb{E}(Y - \hat{Y}) = 0$

$\mathbb{D}\hat{\varepsilon} = \mathbb{D}(Y - \hat{Y}) = \mathbb{D}Y + \mathbb{D}\hat{Y} - 2\,cov(Y, \hat{Y})$

$Y \perp\!\!\!\perp \hat{Y}$, ако Y не е в данните

$\Rightarrow \mathbb{D}\varepsilon = \mathbb{D}Y + \mathbb{D}\hat{Y} = \sigma^2\left(1 + \vec{a}^T(X^TX)^{-1}\vec{a}\right)$

$$Z = \frac{Y - \hat{Y}}{\sigma\sqrt{1 + \vec{a}^T(X^TX)^{-1}\vec{a}}} \in \mathcal{N}o(0,1)$$

$$T = \frac{Y - \hat{Y}}{\sqrt{(1 + \vec{a}^T(X^TX)^{-1}\vec{a})MSE}} \in t(n - \underbrace{(k+1)}_{2})$$

$Y = a_0 + a_1X_1 + \dots + a_kX_k + \dots$

08.05.2018

Модел 1: $Y = \beta_0 + \beta_1 x_1 + \dots + \beta_g x_g + \varepsilon$ $\quad K > g$

Модел 2: $Y = \beta_0 + \beta_1 x_1 + \dots + \beta_g x_g + \dots + \beta_K x_K$ $\quad \varepsilon \in N(0, \sigma^2)$

$$\begin{cases} H_0: \beta_{g+1} = \dots = \beta_K = 0 \\ H_1: \text{Поне } 1\ \beta_i \neq 0 \end{cases}$$

$SSE_2 \to n-(K+1)$ $\qquad MSE_2 = \dfrac{SSE_2}{n-(K+1)}, \ \mathbb{E}\, MSE_2 = \sigma^2$

$SSE_1 \to n-(g+1)$ $\qquad MSE_1 = \dfrac{SSE_1}{n-(g+1)}, \ \mathbb{E}\, MSE_1 = \sigma^2$

$SSE_1 = SSE_2 + \underbrace{(SSE_1 - SSE_2)}$

sum of squares associated with $x_{g+1}, \dots, x_K$ adjusteg for $x_1, \dots, x_g$

$$\mathbb{E}(SSE_1 - SSE_2) = \mathbb{E}\, SSE_1 - \mathbb{E}\, SSE_2 = \sigma^2(n-(g+1) - n + (K+1)) = \sigma^2(K-g)$$

$\Rightarrow \dfrac{SSE_1 - SSE_2}{K-g}$ - неизместена за σ^2

$$\frac{SSE_2}{n-(K+1)} = MSE_2$$

$$\frac{SSE_2}{\sigma^2} \in \chi^2(n-(K+1)) \quad \perp\!\!\!\perp \quad \frac{SSE_1 - SSE_2}{\sigma^2} \in \chi^2(K-g)$$

$$F = \frac{\dfrac{SSE_1 - SSE_2}{K-g}}{\dfrac{SSE_2}{n-(K+1)}} \overset{H_0}{\in} F(K-g, n-(K+1))$$

Критичната област има вида $W: F > F_{\alpha, K-g, n-(K+1)}$

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

$\vec{y} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix} \quad \vec{x} = \begin{bmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{bmatrix}$

$\begin{cases} H_0\ \beta_1 = \beta_2 = 0 \\ H_1 : \beta_1 \neq 0 \text{ или } \beta_2 \neq 0 \end{cases}$

$\vec{y} = X\vec{\beta} + \vec{\varepsilon}$

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix} ; (X^T X)^{-1} = \begin{bmatrix} 17/35 & 0 & -\frac{1}{7} \\ 0 & 1/10 & 0 \\ -1/7 & 0 & 1/14 \end{bmatrix}$$

$X^T y = \begin{bmatrix} 5 \\ 7 \\ 17 \end{bmatrix}$

$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y} \approx (0{,}571 \quad 0{,}7 \quad 0{,}214)^T$

$SSE_2 = \vec{y}^T \vec{y} - \hat{\vec{\beta}}^T X^T \vec{y} \approx 0{,}463 \Rightarrow MSE_2 \approx 0{,}232 ; \hat{\sigma} = \sqrt{MSE_2} \approx 0{,}482$

Първо тестваме $\begin{cases} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \\ \alpha = 0{,}05 \end{cases}$; $t = \dfrac{\hat{\beta}_2 - 0}{\sqrt{MSE_2 \, 1/14}} \approx 1{,}67$

$W : |t| > t_{\alpha/2, 2} = t_{0{,}025, 2} = 4{,}303$

$\Rightarrow$ Включването на квадратен член няма особен смисъл

p-value $= 2P(T(2) > 1{,}67) \approx 0{,}2$

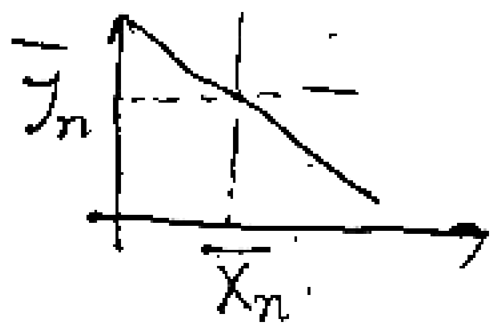
$\beta_2 \in (\hat{\beta}_2 \pm t_{0{,}025,2} \sqrt{MSE_2 \, 1/14} \approx (-0{,}34 ; 0{,}97)$

Редуцираният модел е $y = \hat{\beta}_0 + \varepsilon$

$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix} = e\beta_0 + \vec{\varepsilon} \Rightarrow X^T X = 5$ и $\beta_0 = (X^T X)^{-1} X^T \vec{y} = 1$ и $SSE_1 = 6$

$$F = \frac{\frac{SSE_1 - SSE_2}{k-g}}{MSE_2} \approx 11,93 < F_{0,05,2,2} = 19$$

Прост линеен модел



$$y - \bar{y}_n = \hat{y} - \bar{y}_n + y - \hat{y}.$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}_{SStot} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y})^2}_{SSE}.$$

$$n-1 = k + n - (k+1)$$

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-(k+1)}} \overset{?}{>} F_{\alpha, k, (n-(k+1))}$$ – частен случай на горепоменатия тест

Разглеждаме общ линеен модел $\vec{y} = X\vec{\beta} + \vec{\varepsilon}$

$SSE(\vec{\beta}) = \vec{\varepsilon}^T\vec{\varepsilon} = [\vec{y} - X\vec{\beta}]^T[\vec{y} - X\beta]$

Т-ма 1 Ако $\hat{\vec{\beta}}$ е решение на системата нормални уравнения

$$X^TX\vec{\beta} = X^T\vec{y},$$

то за $\hat{\vec{\beta}}$ се достига минимум на $SSE(\vec{\beta})$, т.е. $SSE(\vec{\beta}) \geq SSE(\hat{\vec{\beta}})$

Д-во $SSE(\vec{\beta}) = [\vec{y} - X\vec{\beta}]^T[\vec{y} - X\vec{\beta}] = [\vec{y} - X\hat{\vec{\beta}} + X\hat{\vec{\beta}} - X\vec{\beta}]^T[\vec{y} - \ldots - X\vec{\beta}] =$

$= [\vec{y} - X\hat{\vec{\beta}}]^T[\vec{y} - X\hat{\vec{\beta}}] + [X\hat{\vec{\beta}} - X\vec{\beta}]^T[X\hat{\vec{\beta}} - X\vec{\beta}] + 2[X\hat{\vec{\beta}} - X\beta]^T[\vec{y} - X\hat{\vec{\beta}}] =$

$= SSE(\hat{\vec{\beta}}) + \|X\hat{\vec{\beta}} - X\vec{\beta}\|^2 + 2[\hat{\vec{\beta}} - \vec{\beta}]^T\underbrace{X^T(\vec{y} - X\hat{\vec{\beta}})}_{0} \geq SSE(\hat{\vec{\beta}})$ □

Д-во 2:

$$\frac{\partial SSE(\vec\beta)}{\partial\vec\beta} = \frac{\partial[\vec y - X\vec\beta]^T[\vec y - X\vec\beta]}{\partial\vec\beta} = \frac{\partial[\vec y^T\vec y - 2\vec y^T X\vec\beta + \vec\beta^T X^T X\vec\beta]}{\partial\vec\beta} =$$

$$= \frac{\partial(\vec\beta^T X^T - 2\vec y^T)X\vec\beta}{\partial\vec\beta} = X^T X\vec\beta + (\vec\beta^T X^T - 2\vec y^T)X = 2X^T X\vec\beta - 2\vec y^T \cancel{X}$$

$$= 2(X^T X\vec\beta - X^T\vec y) = 0 \qquad \square$$

Т.в. Ако $\varepsilon_i \in N(0, \sigma^2)$ и са некорелирани, то оценките по метода на най-малките квадрати и оценките по метода на максималното правдоподобие съвпадат

Д-во $\vec z \in N(\vec a, \Sigma)$

$$\varphi(\vec t) = \mathbb{E} e^{i(\vec t, \vec z)} = \exp\{i(\vec t, \vec a) - \frac{1}{2}(\Sigma\vec t, \vec t)\}$$

$$f_{z_1, \dots z_n}(z_1, \dots z_n) = \frac{1}{(2\pi)^{n/2}(\det\Sigma)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(z-a)^T\Sigma^{-1}(z-a)\right\}$$

Ако $\det\Sigma = 0$, плътността на $\vec z$ не $\exists$

$$f_{y_1, \dots y_n}(y_1, \dots y_n) = C\exp\left\{-\frac{1}{2}(y - X\vec\beta)^T\Sigma^{-1}(y - X\vec\beta)\right\}$$

$f_{y_1, \dots, y_n}$ се максимизира $\Leftrightarrow$ $SSE(\vec\beta)$ се минимизира $\square$

15.05.2018

Т-ма (Гаус-Марков)

$\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$

$$\left.\begin{aligned} &1)\ \mathbb{E}\vec{\varepsilon} = 0 \\ &2)\ \mathbb{D}\vec{\varepsilon} = \sigma^2 > 0 \\ &3)\ cov(\varepsilon_i, \varepsilon_j) = 0\ \forall i \neq j \end{aligned}\right\} \Rightarrow \mathbb{D}\vec{Y} = \mathbb{D}\vec{\varepsilon} = \sigma^2$$

Нека $\hat{\vec{\beta}}$ е решение на $X^T X \vec{\beta} = X^T \vec{Y}$

$\Rightarrow \hat{\vec{\beta}}$ е BLUE (best linear unbiased estimator).

Изпълнени са:

* $\hat{\vec{\beta}}$ е линейна по $\vec{Y} \Rightarrow \hat{\vec{\beta}} = C\vec{Y}$, $C \in \mathbb{R}^{p \times n}$ (1)
* $\hat{\vec{\beta}}$ е неизместена на $\vec{\beta}$ (2)
* $\hat{\vec{\beta}}$ е с минимална дисперсия (3)

$D = \mathbb{D}\tilde{\vec{\beta}} - \mathbb{D}\hat{\vec{\beta}}$ - неотрицателно определена

неизместена оценка за $\vec{\beta}$

<u>Д-во</u> За (1) имаме $\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{Y} \Rightarrow C = (X^T X)^{-1} X^T$

За (2) имаме $\mathbb{E}\hat{\vec{\beta}} = (X^T X)^{-1} X^T \mathbb{E}\vec{Y} = (X^T X)^{-1} X^T (X\vec{\beta} + \underbrace{\mathbb{E}\vec{\varepsilon}}_{0}) = \vec{\beta}$

За (3) имаме $\mathbb{D}C\vec{Y} = C\mathbb{D}\vec{Y}C^T = (X^T X)^{-1} X^T \sigma^2 E X (X^T X)^{-1} =$
$\sigma^2 (X^T X)^{-1} (X^T X)(X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$

Нека $\tilde{\vec{\beta}}$ също е неизместена $\tilde{\vec{\beta}} = B\vec{Y}$, $B \in \mathbb{R}^{p \times n}$

$\mathbb{D}\tilde{\vec{\beta}} = \sigma^2 B B^T$

Представяме $B = M + C \Leftrightarrow M = B - C$

Ако вземем $\mathbb{E} M\vec{Y} = \mathbb{E}(B-C)\vec{Y} = \mathbb{E} B\vec{Y} - \mathbb{E} C\vec{Y} = \vec{\beta} - \vec{\beta} = \vec{0}$

или $\mathbb{E} M\vec{Y} = M\mathbb{E}\vec{Y} = MX\vec{\beta} \Rightarrow MX\vec{\beta} = \vec{0}\ \forall \vec{\beta} \Rightarrow MX = \mathbb{O}$

$\Rightarrow MX(X^TX)^{-1} = \vec{0} \Rightarrow MC^T = 0 \Leftrightarrow CM^T = 0$

Имаме, че $\mathbb{D}\tilde{\vec{\beta}} = \sigma^2 BB^T = \sigma^2(M+C)(M+C)^T = \sigma^2 MM^T + \underbrace{\sigma^2 M^TC + \sigma^2 C^TM}_{0} + \underbrace{\sigma^2 CC^T}_{\mathbb{D}\hat{\vec{\beta}}}$

Тогава $D = \mathbb{D}\tilde{\vec{\beta}} - \mathbb{D}\hat{\vec{\beta}} = \sigma^2 \underbrace{MM^T}_{D}$ - неотрицателно опр. $\square$

Оценката по МНМК:

(1) линейна

(2) неизместена

(3) има минимална дисперсия

Ако към т-мата на Гаус-Марков добавим $\varepsilon \in N(\mu, \sigma^2 I)$, то получаваме ефективна оценка

Дисперсионен анализ

Изучава се влиянието на една или повече факторни променливи върху количествена величина

Нива	Наблюдения	$\bar{Y}_{i\cdot}$
1	$y_{11}\ y_{12}\ \dots\ y_{1r_1}$	$\bar{Y}_{1\cdot}$
2	$y_{21}\ y_{22}\ \dots\ y_{2r_2}$	$\bar{Y}_{2\cdot}$
$\vdots$	$\vdots\ \ \vdots\ \ \ddots\ \ \vdots$	$\vdots$
v	$y_{v1}\ y_{v2}\ \dots\ y_{vr_v}$	$\bar{Y}_{v\cdot}$

$$\bar{Y}_{i\cdot} = \frac{1}{r_i}\sum_{t=1}^{r_i} y_{it}$$

$$\bar{Y}_{\cdot\cdot} = \frac{1}{\underbrace{n}_{n = r_1 + \dots + r_v}} \sum_{i=1}^{v}\sum_{t=1}^{r_i} y_{it}$$

Модел: $Y_{it} = \underset{\text{общо средно}}{\mu} + \underset{\text{ефект}}{\tau_i} + \underset{\text{случайна грешка}}{\varepsilon_{it}}$

$$\left.\begin{matrix} \hat{\mu}_i = \overline{Y}_{i\cdot} \\ \mu_i = \mu + \tau_i \end{matrix}\right\} \Rightarrow Y_{it} = \mu_i + \varepsilon_{it}$$

Обикновено $\left| \begin{matrix} H_0: \tau_1 = \dots = \tau_v = 0 \\ H_1: \text{поне едно } \tau_i \neq 0 \end{matrix} \right.$

$$Y_{it} - \overline{Y}_{\cdot\cdot} = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot} + Y_{it} - \overline{Y}_{i\cdot} \Rightarrow \underbrace{\sum_{i=1}^{v} \sum_{t=1}^{r_i} (Y_{it} - \overline{Y}_{\cdot\cdot})^2}_{SST \text{ - total sum of squares}} =$$

$$= \underbrace{\sum_{i=1}^{v} \sum_{t=1}^{r_i} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})}_{SSF \text{ - factor}} + \underbrace{\sum_{i=1}^{v} \sum_{t=1}^{r_i} (Y_{it} - Y_{i\cdot})}_{SSE \text{ - errors}}$$

$\varepsilon_{it} \in N(0, \sigma^2)$ - некорелирани

$Y_{it} - \overline{Y}_{\cdot\cdot}$ имат $n-1$ степени на свобода

$\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}$ имат $v-1$ степени на свобода

$Y_{it} - \overline{Y}_{i\cdot}$ имат $n-v$ —"—

При H_0:
$$\left.\begin{matrix} \frac{SST}{\sigma^2} \in \chi^2(n-1) \\ \frac{SSF}{\sigma^2} \in \chi^2(v-1) \\ \frac{SSF}{\sigma^2} \in \chi^2(n-v) \end{matrix}\right\} \Rightarrow F = \frac{MSF}{MSE} \in F(v-1, n-v)$$

$$MSF = \frac{SSF}{v-1}, \quad MSE = \frac{SSE}{n-v}$$

22.05.2018

<u>Зад. 1</u>

A	B
6,1	9,1
7,1	8,2
7,8	8,6
6,9	6,9
7,6	7,5
8,5	7,9
47,3	48,2

има ли разлика в средните?

$\alpha = 0,05$

t-test: $\left|\dfrac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right| > t_{\alpha/2,\, n_1+n_2-2}$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X}_1) + \sum_{j=1}^{n}(X_j - \overline{X}_2)}{n_1+n_2-2} =$$

$$= \frac{1}{n_1+n_2-2}\left(\sum_{i=1}^{2}\sum_{k=1}^{n_i}(X_{ik} - \overline{X}_i)^2\right) = \frac{SSE}{n_1+n_2-2} \approx \frac{5,86}{10} \approx 0,586$$

$$t \approx \left|\frac{47,3/6 - 48,2/6}{\sqrt{0,586 \cdot \frac{1}{3}}}\right| = \left|\frac{8,03 - 7,88}{0,44}\right| \approx 0,34 < t_{0,05,10} \approx 2,28 \Rightarrow \text{няма разлика}$$

$$SSF = \sum_{i=1}^{v}\sum_{t=1}^{r_2}\left(\overline{y}_{i*} - \overline{y}_{**}\right)^2 \approx 1,68$$

$MSF = SSF$

$$\Rightarrow F = \frac{MSF}{MSE} \approx \frac{1,68}{0,568} \approx 2,88 < F_{0,05,1,10} \approx 4,86 \Rightarrow \text{няма разлика}$$

$y_{it} = \mu + \tau_i + \varepsilon_{it}$, $i = 1, 2, \dots, v$, $t = 1, 2, \dots, r_i$, $\varepsilon_{it} \in N(0, \sigma^2)$

$$\min \sum_{i=1}^{v} \sum_{t=1}^{r_i} (y_{it} - \mu - \tau_i)^2 = \sum_{i=1}^{v} \sum_{t=1}^{r_i} \varepsilon_{it}^2$$

$$\widehat{\mu + \tau_i} = \overline{y}_{i\cdot} = \frac{1}{r_i} \sum_{t=1}^{r_i} y_{it}, \; i = 1, 2, \dots, v$$

Ако $\hat{\mu} = 0 \Rightarrow \hat{\tau}_i = \overline{y}_{i\cdot}$ - избягваме да търсим оценки за $\hat{\mu}$ и $\hat{\tau}_i$ поотделно

Опр. Оценим параметър е параметър, който може да се представи като очакване на линейна комбинация на откликите

$$\mathbb{E}\left(\sum_{i=1}^{v} \sum_{t=1}^{r_i} a_{it} y_{it}\right) = \sum_{i=1}^{v} \sum_{t=1}^{r_i} a_{it} \underbrace{\mathbb{E} y_{it}}_{=\mu + \tau_i} = \sum_{i=1}^{v} (\mu + \tau_i) \underbrace{\sum_{t=1}^{r_i} a_{it}}_{\text{озн. } b_i} = \sum_{i=1}^{v} b_i (\mu + \tau_i)$$

Ако $b_1 = 1$, $b_i = 0 \; \forall i = 2, \dots, v \Rightarrow \mu + \tau_1$

Ако $\underbrace{\sum_{i=1}^{v} b_i = 0}_{\text{полагаме } b_i = c_i}$, то оценката $\sum_{i=1}^{v} c_i \tau_i$ - контраст

Оценка по МНМК: $\widehat{\sum_{i=1}^{v} c_i \tau_i} = \sum_{i=1}^{v} c_i \overline{y}_{i\cdot}$

Пр. 2 $\underbrace{\overset{\tau_1}{\delta} \; \overset{\tau_2}{x}}$ $\underbrace{\overset{\tau_3}{c} \; \overset{\tau_4}{z} \; \overset{\tau_5}{\tau}}$ $\qquad H_0: \dfrac{\tau_1 + \tau_2}{2} = \dfrac{\tau_3 + \tau_4 + \tau_5}{3}$

$\vec{c} = \left(\frac{1}{2} \;\; \frac{1}{2} \;\; -\frac{1}{3} \;\; -\frac{1}{3} \;\; -\frac{1}{3}\right)^T$ задава контраст

$$\mathbb{E} \overline{y}_{i\cdot} = \frac{1}{r_i} \sum_{t=1}^{r_i} \mathbb{E} y_{it} = \mu + \tau_i \Rightarrow \mathbb{E} \widehat{\sum_{i=1}^{v} c_i \tau_i} = \sum_{i=1}^{v} c_i \tau_i$$

$$\mathbb{D}\bar{Y}_{i\bullet} = \frac{1}{r_i^2}\sum_{i=1}^{v}\mathbb{D}Y_{it} = \frac{\sigma^2}{r_i} \Rightarrow \mathbb{D}\widehat{\sum_{i=1}^{v} c_i\tau_i} = \mathbb{D}\sum_{i=1}^{v} c_i\bar{Y}_{i\bullet} = \sigma^2\sum_{i=1}^{v}\frac{c_i^2}{r_i}$$

$$\widehat{\sum_{i=1}^{v} c_i\tau_i} \in N\left(\sum_{i=1}^{v} c_i\tau_i,\ \sigma^2\sum_{i=1}^{v}\frac{c_i^2}{r_i}\right)$$

Не знаем σ^2, затова можем да използваме $\hat{\sigma}^2 = MSE$

$$\left.\begin{array}{ll} \dfrac{\widehat{\sum_{i=1}^{v} c_i\tau_i} - \sum_{i=1}^{v} c_i\tau_i}{\sigma\sqrt{\sum_{i=1}^{v}\frac{a_i^2}{r_i}}} & \in No(0,1) \\ \hline \dfrac{\frac{\sqrt{SSE}}{\sigma}}{n-v} & \in \sqrt{\chi^2(n-v)} \end{array}\right\} \in t(n-v)$$

$$P\left(-t_{\frac{\alpha}{2},n-v} < \frac{\sum_{i=1}^{v} c_i\bar{Y}_{i\bullet} - \sum_{i=1}^{v} c_i\tau_i}{\sqrt{MSE\sum_{i=1}^{v}\frac{c_i^2}{r_i}}} < t_{\frac{\alpha}{2},n-v}\right) = 1-\alpha$$

$$SS\overset{\text{contrast}}{C} = \frac{\left(\sum_{i=1}^{v} c_i\bar{Y}_{i\bullet}\right)^2}{\sum_{i=1}^{v}\frac{a_i^2}{r_i}}$$

$H_0: \sum_{i=1}^{v} c_i\tau_i = 0$ можем да проверим $\frac{SSC}{MSE} \overset{?}{>} F_{\alpha,1,n-v}$

$H_1: \sum_{i=1}^{v} c_i\tau_i \neq 0$

Пр.3 Дадени са m дов. интервали

$S_1, \ldots, S_m$

S_j = j-ти ДИ е коректен

$P(S_j) = 1-\alpha \Rightarrow P(\bar{S}_j) = \alpha$

$$P(\bar{S}_1 \cup \bar{S}_2) = P(\bar{S}_1) + P(\bar{S}_2) - P(\overline{S_1 S_2}) \le P(\bar{S}_1) + P(\bar{S}_2) = 2\alpha$$

$$P(S_1 \cap \ldots \cap S_m) = 1 - P(\bar{S}_1 \cup \bar{S}_2 \ldots \cup \bar{S}_m) \ge 1 - \underbrace{m\alpha}$$

общо ниво на доверие
overall significance level

интервалите се различават по този квантил

Bonferroni - $\sum_{i=1}^{v} c_i \tau_i \in \left(\sum_{i=1}^{v} c_i \bar{y}_{i\cdot} \pm \overbrace{t_{\frac{\alpha}{2m}, n-v}}\right)$

Scheffe' - $\sqrt{(v-1)F_{\alpha, v-1, n-v}}$

Tukey, Dunnett } - $\tau_i - \tau_j$ $\quad q_{\alpha, v, n-v}/\sqrt{2}$ - рангово разпр. на Стюдънт

$|t|^{(0,5)}_{\alpha, v-1, n-v}$ - многомерно t разпределение

29.05.2018

Зад. 1

A	B
6,1	9,1
7,1	8,2
7,8	8,6
6,9	6,9
7,6	7,5
8,2	7,9

$$X^T X = \begin{pmatrix} 12 & 6 & 6 \\ 6 & 6 & 0 \\ 6 & 0 & 6 \end{pmatrix} ; \det X^T X = 0$$

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y}$$

$$y_{it} = \mu + \tau_i + \varepsilon_{it}, \ i = 1,2,3$$

$$y_{it} = \mu + \tau_1 X_1 + \tau_2 X_2 + \tau_3 X_3 + \varepsilon_{it}$$

$\zeta \in Bi(1, p)$

$\frac{p}{1-p}$ - odds

$\left.\ln \frac{p}{1-p}\right\}$ logit(p) / logodds

$$\mathrm{logit}(p_1) - \mathrm{logit}(p_2) = \ln \frac{p_1(1-p_2)}{(1-p_1)p_2} = \ln \underset{\text{odds ratio}}{OR_{12}}$$

$$OR_{12} = e^{\mathrm{logit}\, p_1 - \mathrm{logit}\, p_2}$$

$X_1, \dots, X_n$ - predictors

Y - категорна сл. вел. $1, 2, \dots, G$

$$Y = \mu + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon \quad , \quad \sum_{i=1}^{n} \beta_i = 1$$

$$P(y = g \mid \vec{X}) = p_g$$

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

$$d_{gi} = \begin{cases} 1, & \text{если в группе } g \\ 0, & \text{иначе} \end{cases}$$

$$\ln L = \sum_{i=1}^{N} \sum_{g=1}^{G} d_{gi} \ln \pi_{gi} + \sum_{i=1}^{N} \sum_{g=1}^{G} d_{gi} \ln \frac{e^{\gamma_g \beta' x}}{\sum_{s=1}^{G} e^{\gamma_s \beta' x}} =$$

$$= \sum_{i=1}^{N} \sum_{g=1}^{G} d_{gi} \gamma_g \beta' x - \sum_{g=1}^{G} d_{gi} \sum_{i=1}^{N} \ln \sum_{s=1}^{G} e^{\gamma_s \beta' x}$$

$$\frac{\partial \ln L}{\partial \beta_{il}} = \sum_{j=1}^{N}\sum_{g=1}^{G} y_{gj} \frac{\partial \vec{X}_j \vec{\beta}_g}{\partial \beta_{il}} - \sum_{j=1}^{N} \frac{\partial}{\partial \beta_{il}} \ln \sum_{s=1}^{G} e^{\vec{X}_j \vec{\beta}_s} =$$

$l = 1, \dots, K$

$i = 2, \dots, G$

$$= \sum_{j=1}^{N} y_{ij} X_{lj} - \sum_{j=1}^{N} \frac{X_{lj} e^{\vec{X}_j^T \vec{\beta}_i}}{\sum_{s=1}^{G} e^{\vec{X}_j \vec{\beta}_s}} = \sum_{j=1}^{N} X_{lj}(y_{ij} - \pi_{ij})$$

likelihood ratio

$$LR = -2[\log L_{submodel} - \log L_{model}] = -2\ln \frac{L_{submodel}}{L_{model}} \in \chi^2 (\text{разлика в брой регресионни коефициенти})$$

$$\left.\begin{array}{l} y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ \left.\begin{array}{l} y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ y_2 = \beta_0 + \beta_1 X_1 \end{array}\right\} LR \end{array}\right\} LR \text{ се нарича deviance } (D)$$

$$z_{gj} = \frac{\hat{\beta}_{gj}}{SE(\hat{\beta}_{gj})} \in N$$

$\hat{p}_1, \hat{p}_2, \dots, \hat{p}_g$

$p_i(\vec{X}) = P(y = i \mid \vec{X})$

Обобщена линейна регресия

$\vec{y} = X\vec{\beta} + \vec{\varepsilon}$, може и да не бъде единична матрица, но трябва да бъде симетрична и неотрицателно определена

$\mathbb{D}\vec{\varepsilon} = \sigma^2 V$

$V = KK^T$

$K^{-1}\vec{y} = K^{-1}X\vec{\beta} + K^{-1}\vec{\varepsilon}$

$\vec{z} = B\vec{\beta} + \eta$

$\mathbb{E}\vec{\eta} = \mathbb{E}K^{-1}\vec{\varepsilon} = 0 \; ; \; \mathbb{D}\vec{\eta} = K^{-1}\underbrace{\mathbb{D}\vec{\varepsilon}}_{\sigma^2 V} K^{-1} = \sigma^2 K^{-1} K K^T (K^{-1})^T = \sigma^2 E$

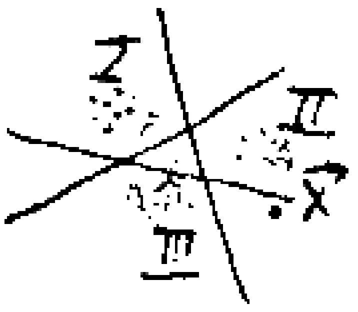
$$\vec{\beta}^* = (B^T B)^{-1} B^T \vec{z} = ((K^{-1}X)^T (K^{-1}X))^{-1} (K^{-1}X)^T K^{-1} \vec{y} =$$

$$= (X^T (K^{-1})^T K^{-1} X)^{-1} X^T (K^{-1})^T K^{-1} \vec{y} = (X^T V^{-1} X)^{-1} X^T V^{-1} \vec{y}$$

$$SSE = (\vec{z} - \vec{B}\vec{\beta}^*)(\vec{z} - \vec{B}\vec{\beta}^*) = (\vec{y} - X\beta^*) V^{-1} (y - X\beta^*)$$

Mahalanobis distance

05.06.2018



$\vec{X} = (X_1, \dots, X_p)$

$\vec{X}_1, \dots, \vec{X}_N$

$w_1, \dots, w_N$

G - класове

π_k - априорна вероятност

$$P(G=k \mid \vec{X}=\vec{x}) = \frac{P(G=k)\,P(\vec{X}=\vec{x} \mid G=k)}{\sum_{i=1}^{n} P(G=i)\,P(\vec{X}=\vec{x} \mid G=i)} = \frac{\pi_k f_k(\vec{x})}{\sum_{i=1}^{k} \pi_i f_i(\vec{x})}$$

$$f_k(\vec{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_k)^T \Sigma_k^{-1} (\vec{x}-\vec{\mu}_k)}$$

$\vec{\mu}_k = \mathbb{E}\vec{X} = (\mathbb{E}X_1, \dots, \mathbb{E}X_p)$

Σ_k - ковариационна матрица

$$\Sigma_k = \mathbb{D}\vec{X} = \begin{pmatrix} \mathbb{D}X_1 & & \\ & \ddots & \\ & & \end{pmatrix}$$

I. $\Sigma_k = \Sigma_m$

$$\ln \frac{P(G=k \mid \vec{X}=\vec{x})}{P(G=m \mid \vec{X}=\vec{x})} = \ln \frac{\pi_k f_k(\vec{x})}{\sum_{j=1}^{g} \pi_j f_j(\vec{x})} \cdot \frac{\sum_{j=1}^{g} \pi_j f_j(\vec{x})}{\pi_m f_m(\vec{x})} = \ln \frac{\pi_k}{\pi_m} + \ln \frac{f_k(\vec{x})}{f_m(\vec{x})} =$$

$$= \ln \frac{\pi_k}{\pi_m} - \frac{1}{2}(\vec{x}-\vec{\mu}_k)^T \Sigma^{-1} (\vec{x}-\vec{\mu}_k) + \frac{1}{2}(\vec{x}-\vec{\mu}_m)^T \Sigma^{-1} (\vec{x}-\mu_m) =$$

$$= \ln \frac{\pi_k}{\pi_m} \cancel{- \frac{1}{2}\vec{x}^T \Sigma^{-1} \vec{x}} + \frac{1}{2}\vec{x}^T \Sigma^{-1} \vec{\mu}_k + \frac{1}{2}\vec{\mu}_k^T \Sigma^{-1} \vec{x} - \frac{1}{2}\vec{\mu}_k^T \Sigma^{-1} \vec{\mu}_k$$
$$\cancel{+ \frac{1}{2}\vec{x}^T \Sigma^{-1} \vec{x}} - \frac{1}{2} x^T \Sigma^{-1} \mu_m - \frac{1}{2}\vec{\mu}_m^T \Sigma^{-1} \vec{x} + \frac{1}{2}\vec{\mu}_m^T \Sigma^{-1} \vec{\mu}_m =$$

$$= \ln \frac{\pi_k}{\pi_m} + \frac{1}{2} \vec{x}^T \Sigma^{-1} (\vec{\mu}_k - \vec{\mu}_m) + \frac{1}{2} \vec{x} \Sigma^{-1} (\mu_k - \mu_m)^T - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \frac{1}{2}\vec{\mu}_m^T \Sigma^{-1} \vec{\mu}_m$$

$$P(G=k \mid \vec{X}=\vec{x}) = \frac{\pi_k f_k(\vec{x})}{\sum \pi_m f_m(\vec{x})} \iff \min \pi_k f_k(\vec{x})$$

$$\ln \pi_k f_k(\vec{x}) = \ln \pi_k - \frac{1}{2}(\vec{x}-\vec{\mu}_k)^T \Sigma^{-1} (\vec{x}-\vec{\mu}_k) - \ln (2\pi)^{p/2} |\Sigma|^{1/2} =$$

$$= \underbrace{\ln \pi_k + \vec{x} \Sigma^{-1} \vec{\mu}_k - \frac{1}{2} \vec{\mu}_k^T \Sigma^{-1} \vec{\mu}_k}_{\delta_k(\vec{x})} - \frac{1}{2} \vec{x}^T \Sigma^{-1} \vec{x} - \ln 2\pi^{p/2} |\Sigma|^{1/2}$$

linear discriminant f-n

$$\delta_k(\vec{x}) = -\frac{1}{2} \ln |\Sigma| - \frac{1}{2}(\vec{x}-\mu_k)^T \Sigma_k^{-1} (\vec{x}-\vec{\mu}_k) + \ln \pi_k$$

QDA (quadratic discriminant function)

$$G(\vec{x}) = \arg\max_k \delta_k(\vec{x})$$