OBSAH

ÚVOD

V praxi jsme často postaveni před problém zařadit jisté objekty do předem vymezených skupin. K tomuto účelu máme k dispozici naměřené určité znaky na těchto objektech a naším úkolem je na základě znalosti hodnot těchto znaků zařadit objekt do některé skupiny. K řešení tohoto problému lze přistupovat několika způsoby, které si dále popíšeme pro situaci, kdy každý objekt patří do jedné ze dvou skupin. K sestavení rozhodovacího pravidla máme k dispozici obvykle několik testovacích objektů, na kterých máme naměřeny příslušné znaky a o kterých buď víme anebo nevíme, do které skupiny patří. Práce se zabývá třemi možnými přístupy, pomocí kterých lze provádět diskriminaci. Jedná se o modely logistické regrese, normální diskriminační analýzy a směsi normálních rozdělení.

První kapitola je věnována definici studovaných modelů a odvození teoretických diskriminačních funkcí. V druhé kapitole je popsána jedna z možností, jak odhadnout neznámé parametry a jak sestavit diskriminační funkce v praktických situacích. Ve třetí kapitole jsou ukázány některé souvislosti mezi studovanými modely, zejména který model platí při splnění předpokladů některého jiného. Čtvrtá kapitola se zabývá problematikou ověřování předpokladů používaných modelů. Problém volby správného modelu v praktické situaci by měla alespoň částečně řešit pátá kapitola. Numerické aspekty výpočtu odhadů neznámých parametrů v modelu směsi normálních rozdělení jsou nastíněny v šesté kapitole. Jak porovnávat jednotlivé modely v konkrétních situacích, resp. jak odhadovat pravděpodobnosti chybné klasifikace, se může čtenář dozvědět v rámci sedmé kapitoly. Použití studovaných modelů na reálných datech je ukázáno v kapitole číslo osm. Dodatek A je stručným přehledem teorie statistických rozhodovacích funkcí, jež je využita k odvození teoretických diskriminačních funkcí. V dodatku B je dokázáno tvrzení použité v páté kapitole. S přiloženým programovým vybavením je možné se seznámit v dodatku C. Jako dodatek D jsou zařazeny barevné grafy ilustrující jeden z praktických příkladů.

V celém textu je používáno značení v následujícím významu.

(LR) ..... model logistické regrese,
(NDA) ..... model normální diskriminační analýzy,
(MND) ..... model směsi normálních rozdělení,
$\mathrm{r}(A)$ ..... hodnost matice A,
$\mathrm{tr}(A)$ ..... stopa matice A,
$|A|$ ..... determinant matice A,
$\mathfrak{L}(\mathbf{X})$ ..... rozdělení náhodného vektoru $\mathbf{X}$,
$N_p(\boldsymbol{\mu}, \Sigma)$ .. p-rozměrné normální rozdělení se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí Σ,

$\xrightarrow[n\to\infty]{s.j.}$ ..... konvergence skoro jistě,
$\xrightarrow[n\to\infty]{\mathfrak{D}}$ ..... konvergence v distribuci,
$\mathfrak{o}(a_n)$ ..... malé o posloupnosti a_n,
$\delta_{i,j}$ ..... Kroneckerovo delta, je rovno jedné, pokud $i = j$ a nule jinak.

Odhady v modelu (LR) jsou značeny symbolem vlnky (např. $\tilde{\boldsymbol{\beta}}$), v modelu (NDA) symbolem stříšky (např. $\hat{\boldsymbol{\beta}}$) a v modelu (MND) symbolem půlobloučku (např. $\breve{\boldsymbol{\beta}}$).

I. DEFINICE MODELŮ

Úkolem, před kterým stojíme, je zařadit jisté objekty do předem vymezených skupin. K dispozici máme naměřené určité znaky na těchto objektech. V naší práci se budeme zabývat pouze situací, kdy každý objekt patří do jedné ze dvou skupin (označme je čísly nula a jedna).

K sestavení rozhodovacího pravidla máme k dispozici obvykle n testovacích objektů, na kterých máme naměřeny příslušné znaky a o kterých buď víme anebo nevíme, do které skupiny patří. Naměřené znaky nechť jsou reprezentovány p-rozměrnými náhodnými vektory $\mathbf{X}_1, \ldots, \mathbf{X}_n$ a příslušnost i-tého objektu k dané skupině nechť je vyjádřena hodnotou náhodné veličiny Y_i, která nabývá hodnot nula nebo jedna podle toho, do které skupiny daný objekt náleží. U nového objektu, který chceme zařadit na základě vytvořeného rozhodovacího pravidla, nechť jsou naměřené znaky reprezentovány p-rozměrným náhodným vektorem $\mathbf{X}$ a rozhodnutí hodnotou náhodné veličiny Y.

1.1. Logistická regrese

Logistická regrese nebyla původně vytvořena pro účely diskriminace, ale jak si ukážeme, lze ji pro ni s úspěchem použít.

Model logistické regrese, který je upravený pro účely diskriminace, je definován následovně. Nechť $Y_1, \ldots, Y_n$ je posloupnost nezávislých náhodných veličin s alternativním rozdělením, jehož parametr splňuje

$$P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i) = \left[1 + \exp(-\beta_0 - \boldsymbol{\beta}'\mathbf{x}_i)\right]^{-1},$$
$$P(Y_i = 0|\mathbf{X}_i = \mathbf{x}_i) = \left[1 + \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i)\right]^{-1},$$

kde $(\beta_0, \boldsymbol{\beta}')'$ je neznámý $p+1$ rozměrný parametr a $\mathbf{X}_1, \ldots, \mathbf{X}_n$ je posloupnost nezávislých náhodných veličin.

Tento model má tzv. učící fázi, ve které známe u každého objektu jak hodnoty $\mathbf{X}_i$, tak hodnoty Y_i (tj. víme, do které skupiny ten který objekt patří). Na základě této znalosti odhadneme parametry $\beta_0, \boldsymbol{\beta}$ a poté dostaneme odhad $\tilde{\pi}(\mathbf{x})$ funkce $\pi(\mathbf{x})$, kde

$$\pi(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x}) = \left[1 + \exp(-\beta_0 - \boldsymbol{\beta}'\mathbf{x})\right]^{-1}.$$

Další objekt, u kterého neznáme jeho zařazení a u něhož jsme naměřili hodnotu $\mathbf{X}$ pomocných znaků, přiřadíme do jedné ze skupin podle hodnoty rozhodovací funkce.

Tuto funkci sestavíme podle kapitoly A o statistických rozhodovacích funkcích pomocí bayesovského přístupu k nalezení optimální rozhodovací funkce. Při značení z této kapitoly hraje roli neznámého parametru θ náhodná veličina Y, která nabývá hodnot $0, 1$ podle toho, do které skupiny daný objekt zařadíme. Tedy parametrický prostor je

$$\Omega = \{0, 1\},$$

prostorem hodnot náhodné veličiny $\mathbf{X}$ je

$$\mathfrak{X} = \mathbb{R}^p.$$

Zde sice neznáme apriorní hustotu parametru θ, ani podmíněnou hustotu náhodné veličiny $\mathbf{X}$ za podmínky θ, ale pro výpočet optimální rozhodovací funkce nám postačí znalost podmíněné hustoty parametru θ za podmínky $\mathbf{X} = \mathbf{x}$, kterou zde označíme $p(\theta|\mathbf{x})$ a která má tvar

$$\begin{aligned} p(1|\mathbf{x}) = \pi(\mathbf{x}) &= \big[1 + \exp(-\beta_0 - \boldsymbol{\beta}'\mathbf{x})\big]^{-1}, \\ p(0|\mathbf{x}) = 1 - \pi(\mathbf{x}) &= \big[1 + \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{x})\big]^{-1}. \end{aligned}$$

Jde přitom o hustotu vzhledem k čítací míře. Množinou $\mathfrak{D}$ rozhodovacích funkcí je množina funkcí $\delta : \mathbb{R}^p \to \{0, 1\}$. Ztrátovou funkci zvolíme následujícím způsobem:

$$L(i, j) = 1 - \delta_{i,j}, \quad i,\ j = 0,\ 1,$$

tj. při správném zařazení objektu nulová a při chybném zařazení objektu jednotková ztráta.

Podle kapitoly o statistických rozhodovacích funkcích lze optimální rozhodovací pravidlo δ^* získat následujícím způsobem:

$$\delta^*(\mathbf{x}) = \operatorname*{argmin}_{\delta \in \mathfrak{D}} E\big[L\big(Y, \delta(\mathbf{X})\big)\big|\mathbf{X} = \mathbf{x}\big],$$

pokud toto minimum existuje alespoň pro $\nu(\mathbf{x})$–skoro všechna $\mathbf{x} \in \mathfrak{X}$. V našem případě hraje roli míry ν míra Lebesgueova.

Je-li $\delta(\mathbf{x}) = j$, potom

$$\begin{aligned} E\big[L\big(Y, \delta(\mathbf{X})\big)\big|\mathbf{X} = \mathbf{x}\big] &= \sum_{i=0}^{1} L\big(i, \delta(\mathbf{x})\big) p(i|\mathbf{x}) = \sum_{i=0}^{1} L(i, j) p(i|\mathbf{x}) \\ &= L(1 - j, j) p(1 - j|\mathbf{x}) = \begin{cases} \pi(\mathbf{x}), & j = 0, \\ 1 - \pi(\mathbf{x}), & j = 1. \end{cases} \end{aligned}$$

Potom

$$\min_{\delta \in \mathfrak{D}} E\big[L\big(Y, \delta(\mathbf{X})\big)\big|\mathbf{X} = \mathbf{x}\big] = \min\big\{\pi(\mathbf{x}), 1 - \pi(\mathbf{x})\big\}.$$

Toto minimum existuje $\forall \mathbf{x} \in \mathbb{R}^p$ a tudíž můžeme psát

$$\delta^*(\mathbf{x}) = \underset{j=0,1}{\operatorname{argmin}} L(1-j,j)p(1-j|\mathbf{x}).$$

Tedy objekt, na němž jsme naměřili hodnotu $\mathbf{X}$ pomocných znaků, zařadíme do první skupiny, pokud

$$\begin{aligned} \pi(\mathbf{X}) &\geq 1-\pi(\mathbf{X}) \\ \pi(\mathbf{X}) &\geq \frac{1}{2} \\ \left[1+\exp(-\beta_0-\boldsymbol{\beta}'\mathbf{X})\right]^{-1} &\geq \frac{1}{2} \\ \beta_0+\boldsymbol{\beta}'\mathbf{X} &\geq 0. \end{aligned}$$

Tudíž objekt zařadíme do první skupiny, pokud $S(\mathbf{X}) \geq 0$ a do nulté skupiny, pokud $S(\mathbf{X}) < 0$. Přitom $S(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}'\mathbf{x}$. Dodejme, že pokud $S(\mathbf{X}) = 0$ (tj. $\pi(\mathbf{X}) = \frac{1}{2}$), můžeme objekt zařadit do libovolné skupiny, aniž bychom zvýšili hodnotu bayesovské rizikové funkce $\rho_q(\delta) = E\Big\{E\big[L\big(\theta,\delta(\mathbf{X})\big)\big|\theta\big]\Big\}$, kterou lze v této situaci interpretovat jako pravděpodobnost špatného zařazení daného objektu. K vlastní diskriminaci však musíme použít odhad $\tilde{S}(\mathbf{x})$ funkce $S(\mathbf{x})$, ve kterém jsou neznámé parametry $\beta_0, \boldsymbol{\beta}$ nahrazeny odhady $\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}$ (jejich nalezení bude popsáno v následující kapitole). Tedy

$$\tilde{S}(\mathbf{x}) = \tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}'\mathbf{x}.$$

Ekvivalentně můžeme rozhodování založit na hodnotě

$$\tilde{\pi}(\mathbf{x}) = \left[1+\exp(-\tilde{\beta}_0-\tilde{\boldsymbol{\beta}}'\mathbf{x})\right]^{-1},$$

přičemž nyní zařadíme objekt s naměřenými znaky $\mathbf{X}$ do první skupiny, pokud $\tilde{\pi}(\mathbf{X}) \geq \frac{1}{2}$.

Hlavní výhodou tohoto modelu je fakt, že neklade žádné podmínky na rozdělení náhodných vektorů $\mathbf{X}_1, \ldots, \mathbf{X}_n$, na rozdíl od následujících dvou modelů.

Poznámka 1.1.1. V regresních modelech bývají obvykle veličiny $\mathbf{X}_1, \ldots, \mathbf{X}_n$, jež jsou naměřeny na objektech učící skupiny, nenáhodné, resp. jejich hodnoty jsou nastaveny experimentátorem. Může se také stát, že i v případě spojitého rozdělení veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$ se některé z naměřených hodnot $\mathbf{X}_1, \ldots, \mathbf{X}_n$ opakují (v důsledku zaokrouhlování apod.). Nic z právě řečeného však není na závadu. Stále můžeme na veličiny $\mathbf{X}_1, \ldots, \mathbf{X}_n$ pohlížet jako na náhodné. Pro určení teoretické diskriminační funkce nepotřebujeme znát hustotu veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$, postačuje nám znalost podmíněné hustoty veličin Y_i za podmínky $\mathbf{X}_i = \mathbf{x}_i$, $i = 1, \ldots, n$. Jak si ukážeme v následující kapitole, znalost hustoty veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$ není nutná ani pro výpočet odhadů neznámých parametrů β_0 a $\boldsymbol{\beta}$. Navíc nám nebude vadit ani opakování se některých hodnot $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

1.2. Normální diskriminační analýza

Model normální diskriminační analýzy je definován následujícím způsobem. Nechť $Y_1, \ldots, Y_n$ je posloupnost nezávislých náhodných veličin s alternativním rozdělením, kde

$$P(Y_i = 1) = \lambda \in (0, 1).$$

Nechť $\mathbf{X}_1, \ldots, \mathbf{X}_n$ je posloupnost nezávislých náhodných vektorů, jejichž podmíněné rozdělení je

$$\mathfrak{L}(\mathbf{X}_i | Y_i = 0) = N_p(\boldsymbol{\mu}_0, \Sigma),$$
$$\mathfrak{L}(\mathbf{X}_i | Y_i = 1) = N_p(\boldsymbol{\mu}_1, \Sigma).$$

Přitom λ, $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, Σ jsou neznámé parametry, které je potřeba odhadnout. Budeme navíc předpokládat, že matice Σ je pozitivně definitní.

Zde máme opět učící fázi, ve které pro n objektů známe jak hodnoty pomocných znaků $\mathbf{X}_1, \ldots, \mathbf{X}_n$, tak skupinu, do které náleží (vyjádřenou hodnotou $Y_1, \ldots, Y_n$) a na základě této informace sestavíme odhad diskriminační funkce $S(\mathbf{x})$, pomocí něhož budeme klasifikovat další objekty. K sestavení teoretické diskriminační funkce využijeme teorie statistických rozhodovacích funkcí a konkrétně bayesovského přístupu pro nalezení optimální diskriminační funkce, stejně jako u modelu logistické regrese. Při značení z kapitoly o statistických rozhodovacích funkcích dostáváme pro model normální diskriminační analýzy:

$$\Omega = \{0, 1\}$$

je parametrický prostor,

$$\mathfrak{X} = \mathbb{R}^p$$

je prostor hodnot náhodné veličiny $\mathbf{X}$, pomocí níž budeme provádět rozhodnutí o hodnotě parametru $\theta \in \Omega$, tj. o zařazení objektu do skupiny (roli parametru θ tedy hraje náhodná veličina Y, stejně jako u modelu logistické regrese).

Apriorní hustota parametru $\theta \in \Omega$ je

$$q(0) = 1 - \lambda, \qquad q(1) = \lambda,$$

přičemž se jedná o hustotu vzhledem k čítací míře. Podmíněná hustota náhodné veličiny $\mathbf{X}$ za podmínky $\theta = Y = y$ má tvar

$$r(\mathbf{x}|0) = g_0(\mathbf{x}), \quad r(\mathbf{x}|1) = g_1(\mathbf{x}),$$

kde g_0, resp. g_1 jsou hustoty $N_p(\boldsymbol{\mu}_0, \Sigma)$, resp. $N_p(\boldsymbol{\mu}_1, \Sigma)$ vzhledem k Lebesgueově míře. Množinou $\mathfrak{D}$ rozhodovacích funkcí je opět množina funkcí $\delta : \mathbb{R}^p \to \{0, 1\}$. Ztrátovou funkci zvolíme stejným způsobem jako u modelu logistické regrese:

$$L(i, j) = 1 - \delta_{i,j}, \quad i,\ j = 0,\ 1.$$

Užitím Bayesovy věty dostaneme podmíněnou hustotu parametru θ za podmínky známé hodnoty náhodné veličiny $\mathbf{X}$.

$$\pi(i|\mathbf{x}) = \frac{g_i(\mathbf{x})q(i)}{\sum_{k=0}^{1} g_k(\mathbf{x})q(k)}, \quad i = 0,\ 1.$$

Podle kapitoly o statistických rozhodovacích funkcích má optimální rozhodovací pravidlo δ^* tvar

$$\delta^*(\mathbf{x}) = \operatorname*{argmin}_{\delta \in \mathfrak{D}} E\big[L\big(Y, \delta(\mathbf{X})\big)\big|\mathbf{X} = \mathbf{x}\big],$$

pokud toto minimum existuje alespoň pro $\nu(\mathbf{x})$–skoro všechna $\mathbf{x} \in \mathfrak{X}$. V našem případě hraje roli míry ν míra Lebesgueova.

Je-li $\delta(\mathbf{x}) = j$, dostáváme

$$\begin{aligned} E\big[L\big(Y, \delta(\mathbf{X})\big)\big|\mathbf{X} = \mathbf{x}\big] = \sum_{i=0}^{1} L\big(i, \delta(\mathbf{x})\big)\pi(i|\mathbf{x}) &= \sum_{i=0}^{1} L(i, j)\frac{g_i(\mathbf{x})q(i)}{\sum_{k=0}^{1} g_k(\mathbf{x})q(k)} \\ &= \frac{1}{\sum_{k=0}^{1} g_k(\mathbf{x})q(k)} g_{1-j}(\mathbf{x})q(1-j). \end{aligned}$$

Potom

$$\min_{\delta \in \mathfrak{D}} E\big[L\big(Y, \delta(\mathbf{X})\big)\big|\mathbf{X} = \mathbf{x}\big] = \min_{j=0,1} \frac{g_{1-j}(\mathbf{x})q(1-j)}{\sum_{k=0}^{1} g_k(\mathbf{x})q(k)}.$$

Toto minimum existuje $\forall \mathbf{x} \in \mathbb{R}^p$ a tedy můžeme psát

$$\begin{aligned} \delta^*(\mathbf{x}) = \operatorname*{argmin}_{j=0,1} \frac{g_{1-j}(\mathbf{x})q(1-j)}{\sum_{k=0}^{1} g_k(\mathbf{x})q(k)} &= \operatorname*{argmin}_{j=0,1} g_{1-j}(\mathbf{x})q(1-j) \\ &= \operatorname*{argmax}_{j=0,1} g_j(\mathbf{x})q(j). \end{aligned}$$

Tedy objekt, na němž jsme naměřili hodnotu $\mathbf{X}$ pomocných znaků, zařadíme do první skupiny, pokud

$$\begin{aligned} g_1(\mathbf{X})q(1) &\geq g_0(\mathbf{X})q(0) \\ g_1(\mathbf{X})\lambda &\geq g_0(\mathbf{X})(1-\lambda) \\ \ln(g_1(\mathbf{X})) + \ln\lambda &\geq \ln(g_0(\mathbf{X})) + \ln(1-\lambda) \\ \ln(\frac{\lambda}{1-\lambda}) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\Sigma^{-1}\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) &\geq 0. \end{aligned}$$

Tudíž objekt zařadíme do první skupiny, pokud $S(\mathbf{X}) \geq 0$, kde

$$\begin{aligned} S(\mathbf{X}) &= \ln(\frac{\lambda}{1-\lambda}) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\Sigma^{-1}\mathbf{X} \\ &= \ln(\frac{\lambda}{1-\lambda}) - \frac{1}{2}\boldsymbol{\mu}_1'\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0'\Sigma^{-1}\boldsymbol{\mu}_0 + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\Sigma^{-1}\mathbf{X} \\ &= \beta_0 + \boldsymbol{\beta}'\mathbf{X}, \end{aligned}$$

přitom

(1.2.1)

$$\begin{aligned} \boldsymbol{\beta} &= \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \\ \beta_0 &= \ln(\frac{\lambda}{1-\lambda}) - \frac{1}{2}\boldsymbol{\mu}_1'\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0'\Sigma^{-1}\boldsymbol{\mu}_0 = \ln(\frac{\lambda}{1-\lambda}) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)'\boldsymbol{\beta}. \end{aligned}$$

Později ukážeme, že $\beta_0, \boldsymbol{\beta}$ jsou rovny stejně označeným parametrům z logistického modelu. Objekt zařadíme do nulté skupiny, pokud $S(\mathbf{X}) < 0$. Pro úplnost dodejme, že pokud $S(\mathbf{X}) = 0$, můžeme objekt zařadit do libovolné skupiny, aniž bychom zvýšili hodnotu bayesovské rizikové funkce $\rho_q(\delta) = E\Big\{E\big[L\big(\theta, \delta(\mathbf{X})\big)\big|\theta\big]\Big\}$, jejímuž explicitnímu vyjádření se budeme nyní věnovat. Podle kapitoly A o statistických rozhodovacích funkcích je

$$\begin{aligned} \rho_q(\delta) &= P\big(\delta(\mathbf{X}) = 1, \theta = 0\big) + P\big(\delta(\mathbf{X}) = 0, \theta = 1\big) \\ &= P(\theta = 0)P\big(\delta(\mathbf{X}) = 1\big|\theta = 0\big) + P(\theta = 1)P\big(\delta(\mathbf{X}) = 0\big|\theta = 1\big) \\ &= P(Y = 0)P\big(\delta(\mathbf{X}) = 1\big|Y = 0\big) + P(Y = 1)P\big(\delta(\mathbf{X}) = 0\big|Y = 1\big) \\ &= (1-\lambda)P\big(\delta(\mathbf{X}) = 1\big|Y = 0\big) + \lambda P\big(\delta(\mathbf{X}) = 0\big|Y = 1\big) \\ &= (1-\lambda)\int_{\{\mathbf{x}:\delta(\mathbf{x})=1\}} g_0(\mathbf{x})dx + \lambda\int_{\{\mathbf{x}:\delta(\mathbf{x})=0\}} g_1(\mathbf{x})dx. \end{aligned}$$

Funkce $\rho_q(\delta)$ vyjadřuje pravděpodobnost toho, že daný objekt nesprávně zařadíme. Tedy výše popsané rozhodovací pravidlo minimalizuje pravděpodobnost nesprávné klasifikace. K samotné diskriminaci nových objektů však funkci $S(\mathbf{X})$ použít nemůžeme, neboť obsahuje neznámé parametry. Použijeme proto odhadnutou diskriminační funkci $\hat{S}(\mathbf{X})$, ve které neznámé parametry nahradíme vhodnými odhady, jež získáme s použitím známých hodnot $\mathbf{X}_1, \ldots, \mathbf{X}_n$ a $Y_1, \ldots, Y_n$. Tím, že neznámé parametry nahradíme jejich odhady však obvykle zvyšujeme pravděpodobnost nesprávné klasifikace.

Poznámka 1.2.1. Jak v modelu logistické regrese, tak v modelu normální diskriminační analýzy, nebylo pro účely sestavení teoretické diskriminační funkce nutně potřeba, aby veličiny $Y_1, \ldots, Y_n$, jež určují zařazení objektů učící sady do jednotlivých skupin, byly náhodné. Do této situace se dostaneme, pokud si předem mezi učící objekty vybereme n_1 příslušníků první skupiny a n_0 příslušníků skupiny nulté (např. n_1 osob, jež prodělali určitou chorobu a n_0 osob, které danou chorobou nikdy netrpěli). Tuto skutečnost bychom však měli zohlednit při výpočtu odhadů některých neznámých parametrů

(zejména parametru λ). K této problematice se ještě vrátíme v následující kapitole, která bude věnována právě výpočtu odhadů.

1.3. Směs normálních rozdělení

Model směsi normálních rozdělení je definován tímto způsobem. Nechť $\mathbf{X}_1, \ldots, \mathbf{X}_n$ je posloupnost nezávislých náhodných veličin, které jsou rozděleny následovně:

$$\mathfrak{L}(\mathbf{X}_i) = N_p(\boldsymbol{\mu}_1, \Sigma) \text{ s pravděpodobností } \lambda,$$
$$\mathfrak{L}(\mathbf{X}_i) = N_p(\boldsymbol{\mu}_0, \Sigma) \text{ s pravděpodobností } 1 - \lambda.$$

Tedy $\mathbf{X}_1, \ldots, \mathbf{X}_n$ jsou stejně rozdělené s hustotou

$$g(\mathbf{x}) = \lambda g_1(\mathbf{x}) + (1 - \lambda) g_0(\mathbf{x}) \tag{1.3.1}$$

vzhledem k Lebesgueově míře. Opět předpokládáme, že matice Σ je pozitivně definitní a $\lambda \in (0, 1)$.

V tomto modelu neznáme zařazení n testovacích objektů do skupin a nemůžeme tedy této znalosti využít k sestavení odhadnuté diskriminační funkce. Přesto si zavedeme náhodné veličiny $Y_1, \ldots, Y_n$, které budou nezávislé a nabývají hodnot 0, 1. $Y_i = 1$, pokud $\mathfrak{L}(\mathbf{X}_i) = N_p(\boldsymbol{\mu}_1, \Sigma)$ a $Y_i = 0$, pokud $\mathfrak{L}(\mathbf{X}_i) = N_p(\boldsymbol{\mu}_0, \Sigma)$. Tímto jsme model směsi normálních rozdělení převedli na model normální diskriminační analýzy a tudíž diskriminaci můžeme založit na totožné teoretické diskriminační funkci $S(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}'\mathbf{x}$, kde $\beta_0, \boldsymbol{\beta}$ jsou parametry mající stejný význam jako v modelu normální diskriminační analýzy. Jediným rozdílem je, že odhadnutá diskriminační funkce $\breve{S}(\mathbf{x}) = \breve{\beta}_0 + \breve{\boldsymbol{\beta}}'\mathbf{x}$ závisí na odhadech parametrů $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, Σ, λ založených pouze na hodnotách $\mathbf{X}_1, \ldots, \mathbf{X}_n$ a bude se tudíž obecně lišit od diskriminační funkce $\hat{S}(\mathbf{x})$.

Poznámka 1.3.1. Model směsi normálních rozdělení je implicitně platný též při platnosti modelu normální diskriminační analýzy. Za platnosti modelu (NDA) je totiž nepodmíněná hustota veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$ tvaru (1.3.1).

II. ODHADY PARAMETRŮ

Ve všech třech uvažovaných modelech je třeba odhadnout jisté parametry. K jejich odhadu zvolíme metodu maximální věrohodnosti.

2.1. Logistická regrese

V modelu logistické regrese budeme maximalizovat logaritmus sdružené podmíněné hustoty vektoru $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ za podmínky $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Tato je rovna:

$$f_{\beta_0,\boldsymbol{\beta}}(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i} (1 - \pi(\mathbf{x}_i))^{1-y_i}. \tag{2.1.1}$$

Potom logaritmická věrohodnostní funkce je rovna:

$$l(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[Y_i(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_i) - \ln\big(1 + \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_i)\big) \Big]. \tag{2.1.2}$$

$$\frac{\partial l}{\partial(\beta_0, \boldsymbol{\beta}')'} = \mathbb{X}'\mathbf{Y} - \mathbb{X}'\pi(\beta_0, \boldsymbol{\beta}),$$

kde

$$\mathbb{X} = \begin{pmatrix} 1 & \mathbf{X}_1{}' \\ \vdots & \vdots \\ 1 & \mathbf{X}_n{}' \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \pi(\beta_0, \boldsymbol{\beta}) = \begin{pmatrix} \pi(\mathbf{X}_1) \\ \vdots \\ \pi(\mathbf{X}_n) \end{pmatrix}.$$

Tedy odhad $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}')'$ musí splňovat rovnici

$$\mathbb{X}'\mathbf{Y} = \mathbb{X}'\pi(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}). \tag{2.1.3}$$

Nevýhodou je, že zmíněnou rovnici je nutné řešit iteračně.

Při splnění jistých podmínek jsou $\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}$ maximálně věrohodnými odhady parametrů $\beta_0, \boldsymbol{\beta}$. Připomeňme, že matice $\mathbb{X}$ je typu $n \times (p+1)$. Dále budeme vždy předpokládat $n > p + 1$.

Věta 2.1.1. *Nechť* $\mathrm{r}(\mathbb{X}) = p+1$. *Potom jsou* $\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}$ *splňující výše uvedenou rovnici maximálně věrohodnými odhady parametrů* $\beta_0, \boldsymbol{\beta}$.

Důkaz.

$$\mathfrak{l}(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[Y_i(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_i) - \ln\big(1 + \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_i)\big) \Big].$$

$$\frac{\partial \mathfrak{l}}{\partial \beta_j} = \sum_{i=1}^{n} \Big[Y_i X_{i,j} - \big[1 + \exp(-\beta_0 - \boldsymbol{\beta}'\mathbf{X}_i)\big]^{-1} X_{i,j} \Big], \quad j = 0, 1, \ldots, p,$$

přičemž $\mathbb{X} = \big(X_{i,j}\big)_{i,j=0,\ldots p}$.

$$\begin{aligned} \frac{\partial^2 \mathfrak{l}}{\partial \beta_j \partial \beta_k} &= -\sum_{i=1}^{n} \Big[\big[1 + \exp(-\beta_0 - \boldsymbol{\beta}'\mathbf{X}_i)\big]^{-1} \cdot \big[1 + \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_i)\big]^{-1} X_{i,j} X_{i,k} \Big] \\ &= -\sum_{i=1}^{n} \Big[\pi(\mathbf{X}_i)\big(1 - \pi(\mathbf{X}_i)\big) X_{i,j} X_{i,k} \Big] \\ &= -v_{j,k}(\mathbb{X}). \end{aligned}$$

Nechť $V(\mathbb{X}) = \big(v_{j,k}(\mathbb{X})\big)_{j,k=0,\ldots,p}$. Matici $V(\mathbb{X})$ můžeme rozložit na součin tvaru

$$V(\mathbb{X}) = \mathbb{X}' V_{\beta_0, \boldsymbol{\beta}} \mathbb{X},$$

kde $V_{\beta_0,\boldsymbol{\beta}}$ je diagonální matice typu $n \times n$ s diagonálou $\pi(\mathbf{X}_1)\big(1-\pi(\mathbf{X}_1)\big), \ldots, \pi(\mathbf{X}_n)\big(1-\pi(\mathbf{X}_n)\big)$. Všechny prvky na diagonále této matice jsou kladné a tedy matice $V_{\beta_0,\boldsymbol{\beta}}$ je pozitivně definitní $\forall \beta_0 \in \mathbb{R}, \forall \boldsymbol{\beta} \in \mathbb{R}^p$. Ukážeme, že za podmínky $\mathrm{r}(\mathbb{X}) = p+1$ je též matice $V(\mathbb{X})$ pozitivně definitní.

Nechť $\mathbf{c} \in \mathbb{R}^{p+1}$, $\mathbf{c} \neq (0, 0, \ldots, 0)'$. Označme $\mathbf{z} = \mathbb{X}\mathbf{c}$.

$$\mathbf{c}' V(\mathbb{X}) \mathbf{c} = \mathbf{c}' \mathbb{X}' V_{\beta_0,\boldsymbol{\beta}} \mathbb{X}\mathbf{c} = \mathbf{z}' V_{\beta_0.\boldsymbol{\beta}} \mathbf{z} \ > 0,$$

neboť z pozitivní definitnosti matice $V_{\beta_0,\boldsymbol{\beta}}$ plyne

$$\mathbf{z}' V_{\beta_0,\boldsymbol{\beta}} \mathbf{z} = 0 \Leftrightarrow \mathbf{z} = (0, 0, \ldots, 0)' \Leftrightarrow \mathbb{X}\mathbf{c} = (0, 0, \ldots, 0),$$

což je ve sporu s požadavkem $\mathrm{r}(\mathbb{X}) = p+1$. Tedy matice $V(\mathbb{X})$ je pozitivně definitní $\forall \beta_0 \in \mathbb{R}, \forall \boldsymbol{\beta} \in \mathbb{R}^p$, z čehož plyne negativní definitnost matice

$$\Big(\frac{\partial^2 \mathfrak{l}}{\partial \beta_j \partial \beta_k} \Big)_{j,k=0,1,\ldots,p} \forall \beta_0 \in \mathbb{R}, \forall \boldsymbol{\beta} \in \mathbb{R}^p.$$

Vzhledem k tomu, že matice druhých parciálních derivací funkce $\mathfrak{l}(\beta_0, \boldsymbol{\beta})$ je negativně definitní $\forall \beta_0 \in \mathbb{R}$, $\forall \boldsymbol{\beta} \in \mathbb{R}^p$, je tato funkce ryze konkávní na $\mathbb{R}^{p+1}$. Podle vět matematické analýzy je tedy každý bod $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}')' \in \mathbb{R}^{p+1}$, splňující $\frac{\partial \mathfrak{l}}{\partial \beta_j}(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}')' = 0 \quad \forall j = 0, 1, \ldots, p$, bodem lokálního maxima dané funkce. Věty konvexní analýzy dokončují důkaz tvrzením, že bod lokálního maxima ryze konkávní funkce je bodem globálního maxima této funkce. □

Poznámka 2.1.2. Prospektivní studie

Jak je uvedeno v kapitole 1.1, mohou se naměřené hodnoty $\mathbf{X}_1,\dots,\mathbf{X}_n$ opakovat a být nastaveny experimentátorem, tj. být nenáhodné (tzv. prospektivní studie). Nechť I je počet různých hodnot veličin $\mathbf{X}_1,\dots,\mathbf{X}_n$ v učící skupině a $\mathbf{x}_1,\dots,\mathbf{x}_I$ jsou tyto hodnoty. Nechť nyní $Y_{i,j}$, $i=1,\dots,I$, $j=1,\dots,m_i$, vyjadřují zařazení objektů do skupin. Přitom m_i je počet objektů s hodnotou vysvětlujících znaků $\mathbf{x}_i$, celkový počet objektů je tedy nyní roven $n=\sum_{i=1}^{I} m_i$. Nechť $Y_{i\cdot}=\sum_{j=1}^{m_i} Y_{i,j}$. Jestliže jsou hodnoty vysvětlujících veličin nenáhodné a nenáhodná jsou i čísla $m_1,\dots,m_I$, měli bychom při hledání maximálně věrohodných odhadů parametrů β_0 a $\boldsymbol{\beta}$ maximalizovat sdruženou hustotu veličin $Y_{1\cdot},\dots,Y_{I\cdot}$ za podmínky $\mathbf{X}_1=\mathbf{x}_1,\dots,\mathbf{X}_I=\mathbf{x}_I$. Rozdělení veličin $Y_{i\cdot}$ za podmínky $\mathbf{X}_i=\mathbf{x}_i$ je binomické s parametry m_i a $\pi(\mathbf{x}_i)$. Uvedená podmíněná sdružená hustota je potom rovna

$$f^{*}_{\beta_0,\boldsymbol{\beta}}(y_{1\cdot},\dots,y_{I\cdot}|\mathbf{x}_1,\dots,\mathbf{x}_I)=\prod_{i=1}^{I}\binom{m_i}{y_{i\cdot}}\pi(\mathbf{x}_i)^{y_{i\cdot}}\big(1-\pi(\mathbf{x}_i)\big)^{1-y_{i\cdot}}. \tag{2.1.4}$$

Logaritmická věrohodnostní funkce je rovna

(2.1.5)

$$\begin{aligned}\ell^{*}(\beta_0,\boldsymbol{\beta}) &= \sum_{i=1}^{I}\ln\binom{m_i}{Y_{i\cdot}}+\sum_{i=1}^{I}\Big[Y_{i\cdot}(\beta_0+\boldsymbol{\beta}'\mathbf{X}_i)-m_i\ln\big(1+\exp(\beta_0+\boldsymbol{\beta}'\mathbf{X}_i)\big)\Big]\\ &= \sum_{i=1}^{I}\ln\binom{m_i}{Y_{i\cdot}}+\sum_{i=1}^{I}\sum_{j=1}^{m_i}\Big[Y_{i,j}(\beta_0+\boldsymbol{\beta}'\mathbf{X}_i)-\ln\big(1+\exp(\beta_0+\boldsymbol{\beta}'\mathbf{X}_i)\big)\Big].\end{aligned}$$

Logaritmická věrohodnost (2.1.5) se tedy od logaritmické věrohodnosti (2.1.2) liší pouze o člen $\sum_{i=1}^{I}\ln\binom{m_i}{Y_{i\cdot}}$, který nezávisí na β_0 ani na $\boldsymbol{\beta}$. Tudíž obě tyto funkce nabývají svého maxima ve stejném bodě.

Poznámka 2.1.3. Retrospektivní studie

Situace, že by naopak veličiny $Y_1,\dots,Y_n$ byly pevně zvolené a vektory $\mathbf{X}_1,\dots,\mathbf{X}_n$ náhodné (tzv. retrospektivní studie), není pro regresní modely příliš typická, avšak v praxi se vyskytuje poměrně často.

Předpokládejme, že $n_1=\sum_{i=1}^{n} Y_i$, $n_0=\sum_{i=1}^{n}(1-Y_i)$. Hodnoty n_1 a n_0 jsou tedy nyní pevně zvoleny. Zaveďme dále náhodnou veličinu V, která bude indikovat zařazení objektu z dané populace do našeho výběru, tj. $V=1$, pokud je objekt z populace ve výběru a $V=0$ jinak. Nechť $s(\mathbf{x}|y,v)$ je podmíněná hustota vektoru $\mathbf{X}$ za podmínky $Y=y$, $V=v$. Při hledání maximálně věrohodných odhadů bychom nyní měli maximalizovat podmíněnou sdruženou hustotu vektorů $\mathbf{X}_1,\dots,\mathbf{X}_n$ za podmínky $Y_i=y_i$, $V_i=v_i=1$, $i=1,\dots,n$, neboť náš výběr je podmíněn volbou hodnot Y_i a také tím, které objekty ze studované populace jsou obsaženy v našem výběru. Nejprve jsme totiž museli z populace vybrat n_1 objektů se znakem $Y=1$ a n_0 objektů se znakem $Y=0$. Tato sdružená hustota je tvaru:

$$f^{\star}_{\beta_0,\boldsymbol{\beta}}(\mathbf{x}_1,\dots,\mathbf{x}_n|y_1,\dots,y_n,v_1,\dots,v_n)=\prod_{i=1}^{n}s(\mathbf{x}_i|y_i,1).$$

Podle Bayesovy věty je při označení podmíněné hustoty vektoru $\mathbf{X}$ za podmínky $V = v$ jako $t(\mathbf{x}|v)$:

$$s(\mathbf{x}|y,1) = \frac{P(Y=y|\mathbf{X}=\mathbf{x},V=1)\ t(\mathbf{x}|1)}{P(Y=y|V=1)}.$$

Budeme-li předpokládat, že pravděpodobnost zahrnutí objektu do výběru nezávisí na vektoru $\mathbf{X}$, tj.

$$P(V=1|Y=1,\mathbf{X}=\mathbf{x}) = P(V=1|Y=1) = \vartheta_1,$$
$$P(V=1|Y=0,\mathbf{X}=\mathbf{x}) = P(V=1|Y=0) = \vartheta_0$$

a že platí model logistické regrese, tj. $P(Y=1|\mathbf{X}=\mathbf{x}) = \big[1+\exp(-\beta_0-\boldsymbol{\beta}'\mathbf{x})\big]^{-1} = \pi(\mathbf{x})$, je opět podle Bayesovy věty

$$\begin{aligned}
&P(Y=1|\mathbf{X}=\mathbf{x},V=1) = \\
&\quad = \frac{P(V=1|Y=1,\mathbf{X}=\mathbf{x})\pi(\mathbf{x})}{P(V=1|Y=1,\mathbf{X}=\mathbf{x})\pi(\mathbf{x}) + P(V=1|Y=0,\mathbf{X}=\mathbf{x})\big(1-\pi(\mathbf{x})\big)} = \\
&\quad = \frac{\vartheta_1\pi(x)}{\vartheta_1\pi(x)+\vartheta_0\big(1-\pi(x)\big)} = \frac{\frac{\vartheta_1}{\vartheta_0}\frac{\pi(x)}{1-\pi(x)}}{1+\frac{\vartheta_1}{\vartheta_0}\frac{\pi(x)}{1-\pi(x)}} = \\
&\quad = \Big[1+\exp\big(-(\beta_0+\ln\frac{\vartheta_1}{\vartheta_0})-\boldsymbol{\beta}'\mathbf{x}\big)\Big]^{-1} = \big[1+\exp(-\beta_0{}^\star-\boldsymbol{\beta}^{\star\prime}\mathbf{x})\big]^{-1} = \pi^\star(\mathbf{x}),
\end{aligned}$$

kde $\beta_0{}^\star = \beta_0 + \ln\frac{\vartheta_1}{\vartheta_0}$, $\boldsymbol{\beta}^\star = \boldsymbol{\beta}$. Podobně $P(Y=0|\mathbf{X}=\mathbf{x},V=1) = 1-\pi^\star(\mathbf{x})$.

Při využití předpokladu nezávislosti $\mathbf{X}$ a V je dále $t(\mathbf{x}|v) = t(\mathbf{x})$ a tedy

$$\begin{aligned}
&f^\star_{\beta_0,\boldsymbol{\beta}}(\mathbf{x}_1,\dots,\mathbf{x}_n|y_1,\dots,y_n,v_1,\dots,v_n) = \\
&\quad = \prod_{i:y_i=1}\pi^\star(\mathbf{x}_i)\frac{t(\mathbf{x}_i)}{P(Y=1|V=1)}\prod_{i:y_i=0}\big(1-\pi^\star(\mathbf{x}_i)\big)\frac{t(\mathbf{x}_i)}{P(Y=0|V=1)} = \\
&\quad = \prod_{i=1}^n \pi^\star(\mathbf{x}_i)^{y_i}\big(1-\pi^\star(\mathbf{x}_i)\big)^{1-y_i}\prod_{i=1}^n\frac{t(\mathbf{x}_i)}{P(Y=y_i|V=1)}.
\end{aligned}$$

Věrohodnostní funkce je tedy tvaru

$$L(\beta_0,\boldsymbol{\beta}) = L^\star(\beta_0{}^\star,\boldsymbol{\beta}^\star)\prod_{i=1}^n\frac{t(\mathbf{X}_i)}{P(Y=Y_i|V=1)},$$

kde $L^\star(\beta_0{}^\star,\boldsymbol{\beta}^\star) = \prod_{i=1}^n \pi^\star(\mathbf{X}_i)^{Y_i}\big(1-\pi^\star(\mathbf{X}_i)\big)^{1-Y_i}$. Předpokládejme navíc, že funkce $t(\mathbf{x})$ nezávisí na parametrech β_0, $\boldsymbol{\beta}$, resp. $\beta_0{}^\star$, $\boldsymbol{\beta}^\star$, což je v praxi obvykle splněno. Při hledání maximálně věrohodných odhadů budeme maximalizovat funkci $L(\beta_0,\boldsymbol{\beta})$ za podmínek $P(Y=1|V=1) = \frac{n_1}{n}$ a $P(Y=0|V=1) = \frac{n_0}{n}$. Po zderivování funkce $L^\star$ podle $\beta_0{}^\star$ zjistíme, že body $\tilde{\beta_0}{}^\star$ a $\tilde{\boldsymbol{\beta}}^\star$, které maximalizují funkci $L^\star$, tyto podmínky splňují.

Jelikož podle učiněného předpokladu je $t(\mathbf{x})$ nezávislé na parametrech β_0, $\boldsymbol{\beta}$, resp. ${\beta_0}^\star$, $\boldsymbol{\beta}^\star$, stačí při hledání maximálně věrohodných odhadů parametrů β_0, $\boldsymbol{\beta}$ maximalizovat funkci $L^\star$ a hledané odhady $\tilde{\beta}_0$, $\tilde{\boldsymbol{\beta}}$ jsou potom rovny

$$\tilde{\beta}_0 = {\tilde{\beta}_0}^\star - \ln\frac{\vartheta_1}{\vartheta_0}, \qquad \tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}^\star.$$

Tedy, abychom u retrospektivní studie získali správné odhady koeficientů v diskriminační funkci, je nutné upravit odhad absolutního členu, pokud jsme s daty pracovali jako kdyby veličiny $Y_1, \ldots, Y_n$ i $\mathbf{X}_1, \ldots, \mathbf{X}_n$ byly náhodné.

2.2. Normální diskriminační analýza

V modelu normální diskriminační analýzy získáme maximálně věrohodné odhady pomocí maximalizace logaritmu sdružené hustoty vektoru $(\mathbf{X}_1', \ldots, \mathbf{X}_n', Y_1, \ldots, Y_n)'$. Sdružená hustota dvojice $(\mathbf{X}', Y)'$ je rovna:

$$f(\mathbf{x}, y) = P(Y = y) f(\mathbf{x}|y).$$

Tedy $f(\mathbf{x}, 1) = \lambda g_1(\mathbf{x})$ a $f(\mathbf{x}, 0) = (1-\lambda) g_0(\mathbf{x})$, kde $g_1(\mathbf{x})$ je hustota rozdělení $N_p(\boldsymbol{\mu}_1, \Sigma)$ a $g_0(\mathbf{x})$ hustota rozdělení $N_p(\boldsymbol{\mu}_0, \Sigma)$. Můžeme tudíž psát $f(\mathbf{x}, y) = \big[\lambda g_1(\mathbf{x})\big]^y \big[(1-\lambda) g_0(\mathbf{x})\big]^{(1-y)}$. Potom sdružená hustota vektoru $(\mathbf{X}_1', \ldots, \mathbf{X}_n', Y_1, \ldots, Y_n)'$ je:

(2.2.1)
$$f_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma, \lambda}(\mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n) = \lambda^{\sum_{i=1}^n y_i} (1-\lambda)^{\sum_{i=1}^n (1-y_i)} \prod_{i: y_i = 1} g_1(\mathbf{x}_i) \prod_{i: y_i = 0} g_0(\mathbf{x}_i).$$

Logaritmická věrohodnostní funkce má tedy tvar:

(2.2.2)
$$\begin{aligned} \ell(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma, \lambda) &= \ln f_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma, \lambda}(\mathbf{X}_1, \ldots, \mathbf{X}_n, Y_1, \ldots, Y_n) \\ &= \sum_{i=1}^n Y_i \ln\lambda + \sum_{i=1}^n (1 - Y_i) \ln(1-\lambda) - \frac{np}{2} \ln 2\pi + \frac{n}{2} \ln |\Sigma|^{-1} + \\ &\quad + \sum_{i=1}^n \big[-\frac{1}{2} (\mathbf{X}_i - \boldsymbol{\mu}_1)' \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu}_1) Y_i\big] + \sum_{i=1}^n \big[-\frac{1}{2} (\mathbf{X}_i - \boldsymbol{\mu}_0)' \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu}_0)(1 - Y_i)\big]. \end{aligned}$$

$$\begin{aligned} \frac{\partial \ell}{\partial \lambda} &= \frac{\sum_{i=1}^n Y_i}{\lambda} - \frac{\sum_{i=1}^n (1 - Y_i)}{(1-\lambda)}, \\ \frac{\partial \ell}{\partial \boldsymbol{\mu}_1} &= \sum_{i=1}^n Y_i \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu}_1), \\ \frac{\partial \ell}{\partial \boldsymbol{\mu}_0} &= \sum_{i=1}^n (1 - Y_i) \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu}_0), \end{aligned}$$

$$\frac{\partial \mathfrak{l}}{\partial \Sigma^{-1}} = \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n} Y_i(\mathbf{X}_i - \boldsymbol{\mu}_1)(\mathbf{X}_i - \boldsymbol{\mu}_1)' - \frac{1}{2}\sum_{i=1}^{n}(1 - Y_i)(\mathbf{X}_i - \boldsymbol{\mu}_0)(\mathbf{X}_i - \boldsymbol{\mu}_0)'.$$

Kořeny věrohodnostních rovnic tedy splňují následující vztahy:

(2.2.3)

$$\begin{aligned}
\hat{\lambda} &= \frac{1}{n}\sum_{i=1}^{n} Y_i, \\
\hat{\boldsymbol{\mu}}_1 &= \frac{\sum_{i=1}^{n} Y_i \mathbf{X}_i}{\sum_{i=1}^{n} Y_i} = \frac{1}{\sum_{i=1}^{n} Y_i}\sum_{i:Y_i=1} \mathbf{X}_i, \\
\hat{\boldsymbol{\mu}}_0 &= \frac{\sum_{i=1}^{n}(1 - Y_i)\mathbf{X}_i}{\sum_{i=1}^{n}(1 - Y_i)} = \frac{1}{\sum_{i=1}^{n}(1 - Y_i)}\sum_{i:Y_i=0} \mathbf{X}_i, \\
\hat{\Sigma} &= \frac{1}{n}\Big[\sum_{i=1}^{n} Y_i(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)' + \sum_{i=1}^{n}(1 - Y_i)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_0)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_0)'\Big] \\
&= \frac{1}{n}\Big[\sum_{i:Y_i=1}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)' + \sum_{i:Y_i=0}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_0)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_0)'\Big].
\end{aligned}$$

Dále ukážeme, že kořeny věrohodnostních rovnic jsou skutečně maximálně věrohodnými odhady parametrů $\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma, \lambda$. Zaveďme následující označení:

$$n_1 = \sum_{i=1}^{n} Y_i, \quad n_0 = \sum_{i=1}^{n}(1 - Y_i),$$
$$\mathbb{S}_1 = \sum_{i:Y_i=1}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)', \quad \mathbb{S}_0 = \sum_{i:Y_i=0}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_0)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_0)'.$$

Při tomto značení můžeme tedy psát:

$$\hat{\lambda} = \frac{n_1}{n}, \quad \hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1}\sum_{i:Y_i=1} \mathbf{X}_i, \quad \hat{\boldsymbol{\mu}}_0 = \frac{1}{n_0}\sum_{i:Y_i=0} \mathbf{X}_i, \quad \hat{\Sigma} = \frac{1}{n}(\mathbb{S}_1 + \mathbb{S}_0).$$

Lemma 2.2.1. *$\forall \boldsymbol{\mu}_1, \boldsymbol{\mu}_0 \in \mathbb{R}^p$ a pro všechny matice Σ typu $p \times p$ platí*

(1) $\sum_{i:Y_i=1}(\mathbf{X}_i - \boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_1) = \operatorname{tr}(\Sigma^{-1}\mathbb{S}_1) + n_1(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\Sigma^{-1}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)$.

(2) $\sum_{i:Y_i=0}(\mathbf{X}_i - \boldsymbol{\mu}_0)'\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_0) = \operatorname{tr}(\Sigma^{-1}\mathbb{S}_0) + n_1(\hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0)'\Sigma^{-1}(\hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0)$.

Důkaz. Dokážeme pouze první část tvrzení, neboť druhá část by se dokazovala analogicky.

$$\begin{aligned}
\sum_{i:Y_i=1}(\mathbf{X}_i - \boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_1) &= \sum_{i:Y_i=1}\operatorname{tr}\big[(\mathbf{X}_i - \boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_1)\big] \\
&= \sum_{i:Y_i=1}\operatorname{tr}\big[\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_1)(\mathbf{X}_i - \boldsymbol{\mu}_1)'\big]
\end{aligned}$$

$$
\begin{aligned}
&= \operatorname{tr}\big[\Sigma^{-1} \sum_{i:Y_i=1} (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\big] \\
&= \operatorname{tr}\big[\Sigma^{-1} \sum_{i:Y_i=1} (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)' + \Sigma^{-1} n_1 (\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\big] \\
&= \operatorname{tr}(\Sigma^{-1}\mathbb{S}_1) + \operatorname{tr}\big[n_1(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\Sigma^{-1}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)\big] \\
&= \operatorname{tr}(\Sigma^{-1}\mathbb{S}_1) + n_1(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\Sigma^{-1}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1). \quad \square
\end{aligned}
$$

Poznámka. Užitím lemmatu 2.2.1 snadno odvodíme

$$
f_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_0,\Sigma,\lambda}(\mathbf{X}_1,\ldots,\mathbf{X}_n,Y_1,\ldots,Y_n) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \lambda^{n_1}(1-\lambda)^{n-n_1} \exp\left[-\frac{1}{2}\operatorname{tr}(n\Sigma^{-1}\hat{\Sigma})\right] \cdot
$$

$$
\cdot \exp\left\{-\frac{1}{2}\big[n_1(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\Sigma^{-1}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1) + (n-n_1)(\hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0)'\Sigma^{-1}(\hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0)\big]\right\},
$$

z čehož plyne, že n_1, $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_0$ a $\hat{\Sigma}$ jsou postačujícími statistikami pro model normální diskriminační analýzy.

Lemma 2.2.2. *Matice $\mathbb{S}_1$ a $\mathbb{S}_0$ jsou pozitivně semidefinitní.*

Důkaz. Dokážeme pouze pozitivní semidefinitnost matice $\mathbb{S}_1$, neboť pozitivní semidefinitnost matice $\mathbb{S}_0$ by se dokazovala analogicky. Nechť $\mathbf{c} \in \mathbb{R}^p$. Potom

$$
\begin{aligned}
\mathbf{c}'\mathbb{S}_1\mathbf{c} &= \mathbf{c}'\big[\sum_{i:Y_i=1} (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)'\big]\mathbf{c} = \sum_{i:Y_i=1} \mathbf{c}'(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)'\mathbf{c} \\
&= \sum_{i:Y_i=1} \mathbf{z}_i'\mathbf{z}_i \geq 0,
\end{aligned}
$$

kde $\mathbf{z}_i = (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)'\mathbf{c}$. $\square$

Lemma 2.2.3. $\forall \lambda \in (0,1) \quad n_1 \ln \frac{n_1}{\lambda} + n_0 \ln \frac{n_0}{1-\lambda} - n \ln n \geq 0.$

Důkaz. Nechť $h(\lambda) = n_1 \ln \frac{n_1}{\lambda} + n_0 \ln \frac{n_0}{1-\lambda} - n \ln n \quad$ pro $\lambda \in (0,1)$.

$$
h'(\lambda) = -(\frac{n_1}{\lambda} - \frac{n_0}{1-\lambda}) \quad \forall \lambda \in (0,1).
$$

$$
h'(\lambda) = 0 \Leftrightarrow \frac{n_1}{\lambda} = \frac{n_0}{1-\lambda} \Leftrightarrow \lambda = \frac{n_1}{n} = \hat{\lambda}.
$$

Pro $0 < \lambda < \hat{\lambda}$ je $h'(\lambda) < 0$ a pro $\hat{\lambda} < \lambda < 1$ je $h'(\lambda) > 0$. $h(\hat{\lambda}) = 0$, na intervalu $(0, \hat{\lambda})$ je funkce h klesající a na intervalu $(\hat{\lambda}, 1)$ je funkce h rostoucí, z čehož přímo plyne platnost tvrzení. $\square$

Lemma 2.2.4. *Nechť Σ je pozitivně definitní matice typu $p \times p$. Nechť η je vlastní číslo matice $\frac{1}{n}\Sigma^{-1}(\mathbb{S}_1 + \mathbb{S}_0)$, kterému přísluší vlastní vektor $\mathbf{z}$. Potom $\eta \geq 0$.*

Důkaz. Z definice vlastních vektorů a vlastních čísel je $\mathbf{z}$ nenulový vektor, který splňuje

$$
\eta\mathbf{z} = \frac{1}{n}\Sigma^{-1}(\mathbb{S}_1 + \mathbb{S}_0)\mathbf{z}.
$$

Vynásobíme–li tuto rovnici zleva výrazem $\mathbf{z}'\Sigma$, dostaneme následující vztah:

$$\begin{aligned}\eta\mathbf{z}'\Sigma\mathbf{z} &= \frac{1}{n}\mathbf{z}'(\mathbb{S}_1+\mathbb{S}_0)\mathbf{z}\\ \eta &= \frac{1}{n}\cdot\frac{1}{\mathbf{z}'\Sigma\mathbf{z}}\mathbf{z}'(\mathbb{S}_1+\mathbb{S}_0)\mathbf{z} \ \geq 0,\end{aligned}$$

neboť $\mathbf{z}'\Sigma\mathbf{z} > 0$ kvůli pozitivní definitnosti matice Σ a $\mathbf{z}'(\mathbb{S}_1+\mathbb{S}_0)\mathbf{z} \geq 0$ podle lemmatu 2.2.2. □

Lemma 2.2.5. $\forall\eta \geq 0 \quad \ln\eta - \eta + 1 \leq 0.$

Důkaz.

$$\begin{aligned}\forall\eta \geq 0 \qquad \eta &\leq \exp(\eta-1)\\ \ln\eta &\leq \eta - 1\\ \ln\eta - \eta + 1 &\leq 0. \quad \square\end{aligned}$$

Věta 2.2.6. *V modelu normální diskriminační analýzy jsou* $\hat\lambda, \hat{\boldsymbol\mu}_1, \hat{\boldsymbol\mu}_0, \hat\Sigma$ *maximálně věrohodnými odhady parametrů* $\lambda, \boldsymbol\mu_1, \boldsymbol\mu_0, \Sigma$.

Důkaz. Nechť $\boldsymbol\mu_1, \boldsymbol\mu_0 \in \mathbb{R}^p, \lambda \in (0,1)$ a Σ je pozitivně definitní matice typu $p \times p$. Užitím lemmatu 2.2.1 dostáváme:

$$\begin{aligned}\mathfrak{l}(\boldsymbol\mu_1,\boldsymbol\mu_0,\Sigma,\lambda) &= \sum_{i:Y_i=1}\ln\lambda + \sum_{i:Y_i=0}\ln(1-\lambda) - \frac{np}{2}\ln 2\pi + \frac{n}{2}\ln|\Sigma|^{-1} -\\ &\quad - \frac{1}{2}\sum_{i:Y_i=1}(\mathbf{X}_i-\boldsymbol\mu_1)'\Sigma^{-1}(\mathbf{X}_i-\boldsymbol\mu_1) - \frac{1}{2}\sum_{i:Y_i=0}(\mathbf{X}_i-\boldsymbol\mu_0)'\Sigma^{-1}(\mathbf{X}_i-\boldsymbol\mu_0)\\ &= n_1\ln\lambda + n_0\ln(1-\lambda) - \frac{np}{2}\ln 2\pi + \frac{n}{2}\ln|\Sigma|^{-1} -\\ &\quad - \frac{1}{2}\operatorname{tr}(\Sigma^{-1}\mathbb{S}_1) - \frac{1}{2}\operatorname{tr}(\Sigma^{-1}\mathbb{S}_0) -\\ &\quad - \frac{n_1}{2}(\hat{\boldsymbol\mu}_1-\boldsymbol\mu_1)'\Sigma^{-1}(\hat{\boldsymbol\mu}_1-\boldsymbol\mu_1) - \frac{n_0}{2}(\hat{\boldsymbol\mu}_0-\boldsymbol\mu_0)'\Sigma^{-1}(\hat{\boldsymbol\mu}_0-\boldsymbol\mu_0),\end{aligned}$$

$$\begin{aligned}\mathfrak{l}(\hat{\boldsymbol\mu}_1,\hat{\boldsymbol\mu}_0,\hat\Sigma,\hat\lambda) &= n_1\ln\frac{n_1}{n} + n_0\ln\frac{n_0}{n} - \frac{np}{2}\ln 2\pi + \frac{n}{2}\ln\Big|\frac{1}{n}(\mathbb{S}_1+\mathbb{S}_0)\Big|^{-1} -\\ &\quad - \frac{1}{2}\operatorname{tr}\Big[\big(\frac{1}{n}(\mathbb{S}_1+\mathbb{S}_0)\big)^{-1}\mathbb{S}_1\Big] - \frac{1}{2}\operatorname{tr}\Big[\big(\frac{1}{n}(\mathbb{S}_1+\mathbb{S}_0)\big)^{-1}\mathbb{S}_0\Big]\\ &= n_1\ln n_1 + n_0\ln n_0 - n\ln n - \frac{np}{2}\ln 2\pi + \frac{n}{2}\ln\big(n^p|\mathbb{S}_1+\mathbb{S}_0|^{-1}\big) -\\ &\quad - \frac{1}{2}\operatorname{tr}\Big[n(\mathbb{S}_1+\mathbb{S}_0)^{-1}\mathbb{S}_1\Big] - \frac{1}{2}\operatorname{tr}\Big[n(\mathbb{S}_1+\mathbb{S}_0)^{-1}\mathbb{S}_0\Big]\\ &= n_1\ln n_1 + n_0\ln n_0 - n\ln n - \frac{np}{2}\ln 2\pi + \frac{n}{2}\ln n^p + \frac{n}{2}\ln|\mathbb{S}_1+\mathbb{S}_0|^{-1} -\\ &\quad - \frac{1}{2}\operatorname{tr}\Big[n(\mathbb{S}_1+\mathbb{S}_0)^{-1}(\mathbb{S}_1+\mathbb{S}_0)\Big]\end{aligned}$$

$$= n_1 \ln n_1 + n_0 \ln n_0 - n \ln n - \frac{np}{2} \ln 2\pi + \frac{n}{2} \ln n^p - \frac{n}{2} \ln |\mathbb{S}_1 + \mathbb{S}_0| - \frac{np}{2}.$$

Nechť $\eta_1, \ldots, \eta_p$ jsou vlastní čísla matice $\frac{1}{n}\Sigma^{-1}(\mathbb{S}_1 + \mathbb{S}_0)$. Podle lemmatu 2.2.4 platí $\eta_1 \geq 0, \ldots, \eta_p \geq 0$ a dále

$$\Big|\frac{1}{n}\Sigma^{-1}(\mathbb{S}_1 + \mathbb{S}_0)\Big| = \prod_{j=1}^{p} \eta_j,$$

$$\operatorname{tr}\Big[\frac{1}{n}\Sigma^{-1}(\mathbb{S}_1 + \mathbb{S}_0)\Big] = \sum_{j=1}^{p} \eta_j.$$

Užitím lemmat 2.2.3, 2.2.5 a s využitím pozitivní definitnosti matice Σ, která implikuje $(\hat{\mu}_k - \mu_k)'\Sigma^{-1}(\hat{\mu}_k - \mu_k) \geq 0$, $k = 0, 1$, dostaneme následující rovnosti a nerovnosti:

$$\begin{aligned}
&\mathfrak{l}(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_0, \hat{\Sigma}, \hat{\lambda}) - \mathfrak{l}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma, \lambda) = \\
&\quad = n_1 \ln n_1 + n_0 \ln n_0 - n \ln n + \\
&\qquad + \frac{n}{2} \ln n^p - \frac{n}{2} \ln |\mathbb{S}_1 + \mathbb{S}_0| - \frac{np}{2} - n_1 \ln \lambda - n_0 \ln(1 - \lambda) + \\
&\qquad + \frac{n}{2} \ln |\Sigma| + \frac{1}{2} \operatorname{tr}(\Sigma^{-1}\mathbb{S}_1) + \frac{1}{2} \operatorname{tr}(\Sigma^{-1}\mathbb{S}_0) + \\
&\qquad + \frac{n_1}{2}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\Sigma^{-1}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1) + \frac{n_0}{2}(\hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0)'\Sigma^{-1}(\hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0) \\
&\quad = \Big\{ n_1 \ln \frac{n_1}{\lambda} + n_0 \ln \frac{n_0}{1 - \lambda} - n \ln n \Big\} + \\
&\qquad + \Big\{ \frac{n_1}{2}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\Sigma^{-1}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1) + \frac{n_0}{2}(\hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0)'\Sigma^{-1}(\hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0) \Big\} + \\
&\qquad + \Big\{ -\frac{n}{2} \ln \frac{|\mathbb{S}_1 + \mathbb{S}_0|}{n^p |\Sigma|} + \frac{1}{2} \operatorname{tr}\big[\Sigma^{-1}(\mathbb{S}_1 + \mathbb{S}_0)\big] - \frac{np}{2} \Big\} \\
&\quad \geq -\frac{n}{2} \ln \frac{|\mathbb{S}_1 + \mathbb{S}_0|}{|n\Sigma|} + \frac{1}{2} \operatorname{tr}\big[\Sigma^{-1}(\mathbb{S}_1 + \mathbb{S}_0)\big] - \frac{np}{2} \\
&\quad = -\frac{n}{2} \Big\{ \ln \Big|\frac{1}{n}\Sigma^{-1}(\mathbb{S}_1 + \mathbb{S}_0)\Big| - \operatorname{tr}\big[\frac{1}{n}\Sigma^{-1}(\mathbb{S}_1 + \mathbb{S}_0)\big] + p \Big\} \\
&\quad = -\frac{n}{2} \Big\{ \ln \prod_{j=1}^{p} \eta_j - \sum_{j=1}^{p} \eta_j + p \Big\} \\
&\quad = -\frac{n}{2} \sum_{j=1}^{p} \{ \ln \eta_j - \eta_j + 1 \} \\
&\quad \geq 0. \quad \square
\end{aligned}$$

Poznámka 2.2.7. Jak již bylo řečeno v poznámce 1.2.1, veličiny $Y_1, \ldots, Y_n$ nemusejí být nutně náhodné. V této situaci bychom však měli předpokládat, že parametr λ, který vyjadřuje poměrné zastoupení jedinců dvou uvažovaných skupin v populaci, je známý.

Pro odhad diskriminační funkce v modelu (NDA) tedy místo $\hat{\lambda}$ použijeme přesnou hodnotu λ. Ostatní parametry ($\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$ a Σ) odhadneme stejným způsobem jako v případě, že $Y_1, \ldots, Y_n$ jsou náhodné.

Jsou-li nenáhodné naopak vektory $\mathbf{X}_1, \ldots, \mathbf{X}_n$ a veličiny $Y_1, \ldots, Y_n$ zůstávají náhodné (např. v medicíně podáváme části populace předem určenou dávku léku, druhé části populace jinou předem určenou dávku léku a sledujeme, u kterých jedinců se zlepšil a u kterých nezlepšil zdravotní stav), je vhodné k diskriminaci použít jiného modelu (např. logistické regrese). V této situaci totiž nemůžeme λ odhadovat pomocí výrazu $\frac{1}{n}\sum_{i=1}^{n} Y_i$, neboť tento výraz odhaduje pravděpodobnost, že $Y = 1$, podmíněnou ovšem realizovanými, předem určenými, hodnotami vektorů $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

2.3. Směs normálních rozdělení

V modelu směsi normálních rozdělení získáme maximálně věrohodné odhady maximalizací logaritmu sdružené hustoty náhodných vektorů $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Tato hustota je tvaru:

$$f_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_0,\Sigma,\lambda}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n}[\lambda g_1(\mathbf{x}_i) + (1-\lambda) g_0(\mathbf{x}_i)]. \tag{2.3.1}$$

Odtud dostaneme logaritmickou věrohodnostní funkci

$$\begin{aligned} \ell(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma, \lambda) &= \sum_{i=1}^{n} \ln[\lambda g_1(\mathbf{X}_i) + (1-\lambda) g_0(\mathbf{X}_i)] \\ &= n \ln(2\pi)^{-\frac{p}{2}} + n \ln |\Sigma|^{-\frac{1}{2}} + \\ &\quad + \sum_{i=1}^{n} \ln\Big[\lambda \exp\big(-\frac{1}{2}(\mathbf{X}_i - \boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_1)\big) \\ &\qquad + (1-\lambda) \exp\big(-\frac{1}{2}(\mathbf{X}_i - \boldsymbol{\mu}_0)'\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_0)\big)\Big]. \end{aligned} \tag{2.3.2}$$

Nechť $g(\mathbf{X}) = \lambda g_1(\mathbf{X}) + (1-\lambda) g_0(\mathbf{X})$.

$$\frac{\partial \ell}{\partial \lambda} = \sum_{i=1}^{n}\Big[\frac{1}{\lambda}\frac{\lambda g_1(\mathbf{X}_i)}{g(\mathbf{X}_i)} - \frac{1}{1-\lambda}\frac{(1-\lambda) g_0(\mathbf{X}_i)}{g(\mathbf{X}_i)}\Big] = \frac{\sum_{i=1}^{n} w_i^1}{\lambda} - \frac{\sum_{i=1}^{n} w_i^0}{1-\lambda},$$

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}_1} = \sum_{i=1}^{n} \frac{\lambda g_1(\mathbf{X}_i)}{g(\mathbf{X})}\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_1) = \sum_{i=1}^{n} w_i^1 \Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_1),$$

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}_0} = \sum_{i=1}^{n} \frac{(1-\lambda) g_0(\mathbf{X}_i)}{g(\mathbf{X})}\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_0) = \sum_{i=1}^{n} w_i^0 \Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_0),$$

$$\frac{\partial \mathfrak{l}}{\partial \Sigma^{-1}} = n\Sigma - \sum_{i=1}^{n}\Big[\frac{\lambda g_1(\mathbf{X}_i)}{g(\mathbf{X})}(\mathbf{X}_i-\boldsymbol{\mu}_1)(\mathbf{X}_i-\boldsymbol{\mu}_1)' + \frac{(1-\lambda)g_0(\mathbf{X}_i)}{g(\mathbf{X}_i)}(\mathbf{X}_i-\boldsymbol{\mu}_0)(\mathbf{X}_i-\boldsymbol{\mu}_0)'\Big]$$
$$= n\Sigma - \sum_{i=1}^{n}\big[w_i^1(\mathbf{X}_i-\boldsymbol{\mu}_1)(\mathbf{X}_i-\boldsymbol{\mu}_1)' + w_i^0(\mathbf{X}_i-\boldsymbol{\mu}_0)(\mathbf{X}_i-\boldsymbol{\mu}_0)'\big],$$

$$\text{kde } w_i^1 = \frac{\lambda g_1(\mathbf{X}_i)}{g(\mathbf{X}_i)},$$
$$w_i^0 = \frac{(1-\lambda)g_0(\mathbf{X}_i)}{g(\mathbf{X}_i)}, \quad i=1,\dots,n,$$

přitom platí $w_i^0 = 1 - w_i^1 \quad \forall i = 1,\dots,n$. Kořeny věrohodnostních rovnic tedy splňují vztahy:

(2.3.3)

$$\breve{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\breve{w}_i^1,$$
$$\breve{\boldsymbol{\mu}}_1 = \frac{\sum_{i=1}^{n}\breve{w}_i^1\mathbf{X}_i}{\sum_{i=1}^{n}\breve{w}_i^1},$$
$$\breve{\boldsymbol{\mu}}_0 = \frac{\sum_{i=1}^{n}(1-\breve{w}_i^1)\mathbf{X}_i}{\sum_{i=1}^{n}(1-\breve{w}_i^1)},$$
$$\breve{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\big[\breve{w}_i^1(\mathbf{X}_i-\breve{\boldsymbol{\mu}}_1)(\mathbf{X}_i-\breve{\boldsymbol{\mu}}_1)' + (1-\breve{w}_i^1)(\mathbf{X}_i-\breve{\boldsymbol{\mu}}_0)(\mathbf{X}_i-\breve{\boldsymbol{\mu}}_0)'\big].$$

Povšimněme si, že věrohodnostní rovnice pro směs normálních rozdělení získáme formálně z věrohodnostních rovnic pro model normální diskriminační analýzy nahrazením hodnot $Y_1,\dots,Y_n$ vahami $w_1^1,\dots,w_n^1$. Na rozdíl od modelu diskriminační analýzy však neznáme hodnoty $Y_1,\dots,Y_n$ a váhy $w_1^1,\dots,w_n^1$ závisejí na neznámých parametrech. Odhady musíme tudíž hledat iteračně.

Váhy $w_1^1,\dots,w_n^1$ jsou navíc logistickými pravděpodobnostmi $\pi(\mathbf{x}_i)$, které se vyskytují v modelu logistické regrese, neboť užitím Bayesovy věty dostáváme vztah (při definování veličin $Y_1,\dots,Y_n$ jako v kapitole 1.3)

$$P(Y_i=1|\mathbf{X}_i=\mathbf{x}_i) = \frac{g_1(\mathbf{x}_i)P(Y_i=1)}{g_1(\mathbf{x}_i)P(Y_i=1)+g_0(\mathbf{x}_i)P(Y_i=0)} = \frac{\lambda g_1(\mathbf{x}_i)}{g(\mathbf{x}_i)} = w_i^1,\ i=1,\dots,n.$$

V případě směsi normálních rozdělení se nám nepodaří dokázat, že odhady splňující věrohodnostní rovnice jsou vždy skutečně maximálně věrohodnými odhady neznámých parametrů. Logaritmická věrohodnost totiž může být neomezená pro $\Sigma \to \mathbf{0}$ (nulová matice) (viz [13]).

Příklad. Nechť $p = 1$, $n = 2$ a $X_1 = 1$, $X_2 = 0$. Potom při $\mu_1 = 1$, $\mu_0 = 0$ a $\lambda = \frac{1}{2}$ je

$$\mathfrak{l}(\mu_1, \mu_0, \Sigma, \lambda) = \mathfrak{l}(1, 0, \Sigma, \frac{1}{2}) = h(\Sigma) = -\ln(2\pi) + 2\ln\frac{1}{2} - \ln(\Sigma) + 2\ln(1 + e^{-1/2\Sigma})$$

a

$$\lim_{\Sigma \to 0_+} h(\Sigma) = \infty.$$

III. SOUVISLOSTI MEZI MODELY

Odpověď na otázku, zda lze přejít od jednoho modelu k druhému, by měla podat právě tato kapitola. Ukážeme, že při splnění jistých předpokladů lze některý uvažovaný model získat z jiného.

3.1. Model logistické regrese a směsi normálních rozdělení

Nejprve ukážeme, jak model směsi normálních rozdělení souvisí s modelem logistické regrese. Budeme používat značení uvedené v první kapitole. Opět si v modelu směsi normálních rozdělení zavedeme pomocné veličiny $Y_1, \ldots, Y_n$, které budou indikovat správné rozdělení náhodných veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Tedy $Y_i = 1$, pokud $\mathfrak{L}(X_i) = N_p(\boldsymbol{\mu}_1, \Sigma)$ a $Y_i = 0$, pokud $\mathfrak{L}(X_i) = N_p(\boldsymbol{\mu}_0, \Sigma)$, $\quad i = 1, \ldots, n$. Připomeňme pouze, že veličiny $\mathbf{X}_1, \ldots, \mathbf{X}_n$ mají hustotu $g(\mathbf{x}) = \lambda g_1(\mathbf{x}) + (1-\lambda) g_0(\mathbf{x})$ vzhledem k Lebesgueově míře, kde g_1, resp. g_0 jsou hustoty $N_p(\boldsymbol{\mu}_1, \Sigma)$, resp. $N_p(\boldsymbol{\mu}_0, \Sigma)$. $P(Y_i = 1) = \lambda$ a $P(Y_i = 0) = 1 - \lambda$, $\quad i = 1, \ldots, n$.

Vyjdeme-li nyní od modelu směsi normálních rozdělení, dostáváme užitím Bayesovy věty následující podmíněné rozdělení pomocných veličin Y_i, $i = 1, \ldots, n$ za podmínky $\mathbf{X}_i = \mathbf{x}_i$.

(3.1.1)

$$
\begin{aligned}
P(Y_i = 1 | \mathbf{X}_i = \mathbf{x}_i) &= \frac{g_1(\mathbf{x}_i) P(Y_i = 1)}{g_1(\mathbf{x}_i) P(Y_i = 1) + g_0(\mathbf{x}_i) P(Y_i = 0)} \\
&= \frac{1}{1 + \exp\left(- \ln \frac{\lambda}{1-\lambda} - \frac{1}{2} \boldsymbol{\mu}_0' \Sigma^{-1} \boldsymbol{\mu}_0 + \frac{1}{2} \boldsymbol{\mu}_1' \Sigma^{-1} \boldsymbol{\mu}_1 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)' \Sigma^{-1} \mathbf{x}_i \right)} \\
&= \left[1 + \exp(-\beta_0 - \boldsymbol{\beta}' \mathbf{x}_i) \right]^{-1}, \qquad i = 1, \ldots, n,
\end{aligned}
$$

kde

$$
\begin{aligned}
\boldsymbol{\beta} &= \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \\
\beta_0 &= \ln \frac{\lambda}{1-\lambda} - \frac{1}{2} \boldsymbol{\mu}_1' \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_0' \Sigma^{-1} \boldsymbol{\mu}_0 = \ln \frac{\lambda}{1-\lambda} - \frac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)' \boldsymbol{\beta}.
\end{aligned}
$$

Vidíme tedy, že od modelu směsi normálních rozdělení lze přejít k logistickému modelu s výše uvedenými parametry β_0 a $\boldsymbol{\beta}$, neboť v modelu logistické regrese požadujeme, aby

podmíněné rozdělení veličin, které indikují zařazení objektů do jednotlivých skupin, bylo právě tvaru (3.1.1). Opačný přechod možný není (ani kdybychom v modelu logistické regrese předpokládali normalitu náhodných veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n, \quad i = 1, \ldots, n$) již z toho důvodu, že v modelu směsi normálních rozdělení máme $(\frac{p^2}{2} + \frac{5p}{2} + 1)$ nezávislých neznámých parametrů a v modelu logistické regrese pouze $(p+1)$ nezávislých neznámých parametrů.

3.2. Model logistické regrese a normální diskriminační analýzy

Vzhledem k tomu, že model diskriminační analýzy se liší od modelu směsi normálních rozdělení pouze tím, že pomocné náhodné veličiny $Y_1, \ldots, Y_n$ jsou již zahrnuty v modelu, lze naprosto stejným způsobem ukázat, že je možné přejít od modelu diskriminační analýzy k logistickému modelu s parametry totožnými těm, které byly odvozeny v předcházející části. Ani zde není možný opačný přechod.

Ukážeme si však, že logistický model platí též při obecnějších předpokladech o rozdělení vysvětlujících veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$, než těch, které jsou požadovány pro model normální diskriminační analýzy. Nechť

$$P(Y_i = 1) = \lambda \in (0,1), \quad i = 1, \ldots, n,$$
$$P(Y_i = 0) = 1 - \lambda \in (0,1), \quad i = 1, \ldots, n.$$

Nechť vysvětlující veličiny $\mathbf{X}_1, \ldots, \mathbf{X}_n$ mají následující podmíněnou hustotu za podmínky $Y_i = k, \quad k = 0, 1$ vzhledem k Lebesgueově míře:

$$r(\mathbf{x}|1) = f_1(\mathbf{x}) = C(\vartheta_1, \eta) h(\mathbf{x}, \eta) \exp\big(\mathbf{x}' T(\vartheta_1, \eta)\big),$$
$$r(\mathbf{x}|0) = f_0(\mathbf{x}) = C(\vartheta_0, \eta) h(\mathbf{x}, \eta) \exp\big(\mathbf{x}' T(\vartheta_0, \eta)\big),$$

kde ϑ_1, ϑ_0 jsou parametry, jejichž hodnoty mohou být rozdílné v závislosti na hodnotě Y_i a η je parametr, jehož hodnota je stejná pro obě podmíněná rozdělení. C, h jsou známé funkce, T je též známá, avšak vektorová funkce. Navíc předpokládejme $f_1(\mathbf{x}) > 0, \ f_0(\mathbf{x}) > 0 \ \forall \mathbf{x} \in \mathbb{R}^p$.

Za těchto předpokladů dostáváme užitím Bayesovy věty následující vztah:

$$\begin{aligned} P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i) &= \frac{\lambda f_1(\mathbf{x}_i)}{\lambda f_1(\mathbf{x}_i) + (1-\lambda) f_0(\mathbf{x}_i)} \\ &= \frac{1}{1 + \exp\Big(-\ln\frac{\lambda}{1-\lambda} - \ln\frac{C(\vartheta_1,\eta)}{C(\vartheta_0,\eta)} - \big(T(\vartheta_1,\eta) - T(\vartheta_0,\eta)\big)' \mathbf{x}_i\Big)} \\ &= \frac{1}{1 + \exp(-\beta_0 - \boldsymbol{\beta}' \mathbf{x}_i)}, \quad i = 1, \ldots, n, \end{aligned}$$

kde

$$\beta_0 = \ln\frac{\lambda}{1-\lambda} + \ln\frac{C(\vartheta_1,\eta)}{C(\vartheta_0,\eta)},$$
$$\boldsymbol{\beta} = T(\vartheta_1,\eta) - T(\vartheta_0,\eta).$$

Model normální diskriminační analýzy je speciálním případem výše popsané situace, neboť

$$\begin{aligned}\vartheta_1 &= \boldsymbol{\mu}_1,\\ \vartheta_0 &= \boldsymbol{\mu}_0,\\ \eta &= \Sigma,\end{aligned}$$

$$\begin{aligned}f_k(\mathbf{x}) &= (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\exp\big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\big)\\ &= \exp(-\frac{1}{2}\boldsymbol{\mu}_k'\Sigma^{-1}\boldsymbol{\mu}_k)(2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\exp(-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x})\exp(\mathbf{x}'\Sigma^{-1}\boldsymbol{\mu}_k), \quad k=0,1.\end{aligned}$$

Tedy

$$\begin{aligned}C(\boldsymbol{\mu}_k,\Sigma) &= \exp(-\frac{1}{2}\boldsymbol{\mu}_k'\Sigma^{-1}\boldsymbol{\mu}_k), \quad k=0,1,\\ h(\mathbf{x},\Sigma) &= (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\exp(-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x}),\\ T(\boldsymbol{\mu}_k,\Sigma) &= \Sigma^{-1}\boldsymbol{\mu}_k, \quad k=0,1.\end{aligned}$$

3.3. Model směsi normálních rozdělení a diskriminační analýzy

Vztah mezi těmito dvěma modely není nutné podrobně rozebírat, neboť jak již bylo řečeno, liší se tyto modely pouze tím, jestli jsou do nich zahrnuty pomocné veličiny $Y_1,\ldots,Y_n$, či nikoliv (tj. tyto modely se odlišují znalostí příslušnosti daných objektů do jednotlivých skupin). Respektive model (NDA) je definován pomocí podmíněného rozdělení veličin $\mathbf{X}_1,\ldots,\mathbf{X}_n$ za podmínky $Y_1=y_1,\ldots,Y_n=y_n$ a model (MND) pomocí nepodmíněného rozdělení $\mathbf{X}_1,\ldots,\mathbf{X}_n$. Lze tedy od jednoho modelu přejít k druhému, aniž by se jejich parametry změnily. Lišit se ovšem budou odhady těchto parametrů i při použití stejných metod odhadu, jak bylo ukázáno v druhé kapitole.

IV. OVĚŘOVÁNÍ PŘEDPOKLADŮ STUDOVANÝCH MODELŮ

Proč testovat předpoklady modelů? Logistický model sice neklade žádné zvláštní požadavky na rozdělení náhodných veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$, ale místo toho předpokládá specifický tvar pravděpodobnosti $P(Y = 1|\mathbf{X} = \mathbf{x})$. Skutečnost, že opravdu platí $P(Y = 1|\mathbf{X} = \mathbf{x}) = \left[1+\exp(-\beta_0-\boldsymbol{\beta}'\mathbf{x})\right]^{-1}$, bychom měli ověřit nejlépe pomocí nějakého statistického testu.

U zbývajících dvou modelů by naopak mělo postačovat ověření jejich předpokladů, jimiž jsou v první řadě mnohorozměrná normalita veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$ a dále shodnost variančních matic vysvětlujících veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$ v obou uvažovaných skupinách, do kterých dané objekty zařazujeme (tyto podmínky mj. zajišťují platnost modelu logistické regrese, jak bylo ukázáno ve třetí kapitole). Neznáme-li přitom zařazení jednotlivých objektů do skupin, je ověření předpokladu normality poměrně obtížné, neboť $\mathbf{X}_1, \ldots, \mathbf{X}_n$ není výběrem z jednoho normálního rozdělení, ale ze směsi dvou normálních rozdělení s různými středními hodnotami. Hustota této směsi se bude lišit od hustoty normálního rozdělení a obvykle nebude symetrická, jako je tomu u hustoty normálního rozdělení. Z tohoto důvodu většina běžných testů jednorozměrné normality zamítne hypotézu, že jednotlivé složky vektorů $\mathbf{X}_1, \ldots, \mathbf{X}_n$ pocházejí z normálního rozdělení. Proto se při použití modelu směsi normálních rozdělení musíme obvykle spokojit pouze s nepřesným ověřením předpokladů pomocí různých grafů a obrázků (histogramů apod.).

Dále uvedeme některé testy dobré shody pro model logistické regrese publikované v [8] a test pro ověření shodnosti variančních matic několika výběrů.

4.1. Základní testy dobré shody pro model logistické regrese

Pro popis testů dobré shody použijme značení, jež bylo zavedeno v rámci poznámky 2.1.2. Tj. nechť učící skupina obsahuje n objektů, I je počet různých hodnot veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$ v učící skupině a $\mathbf{x}_1, \ldots, \mathbf{x}_I$ jsou tyto hodnoty. Jak již bylo řečeno v poznámkách 1.1.1 a 2.1.2, veličiny $\mathbf{X}_1, \ldots, \mathbf{X}_n$ nemusejí být nutně náhodné (obvyklý jev u regresních modelů). Mohou se tedy některé z hodnot $\mathbf{X}_1, \ldots, \mathbf{X}_n$ opakovat, i když jsou veličiny $\mathbf{X}_1, \ldots, \mathbf{X}_n$ spojitě rozdělené. Celý model logistické regrese pracuje totiž s podmíněným rozdělením veličin $Y_1, \ldots, Y_n$ za podmínky $\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_n = \mathbf{x}_n$. Hodnoty $Y_1, \ldots, Y_n$ vyjadřující zařazení i-tého objektu do jedné ze dvou skupin přeznačme na $Y_{i,j}$, $i = 1, \ldots, I$, $j = 1, \ldots, m_i$, kde m_i je počet objektů s hodnotou vysvětlujících znaků $\mathbf{X}_i$ a $Y_{i,j}$, $j = 1, \ldots, m_i$ označuje zařazení objektů, u nichž mají vysvětlující

znaky hodnotu $\mathbf{X}_i = \mathbf{x}_i$. Stejně jako dříve označme

$$n_1 = \sum_{i=1}^{I} \sum_{j=1}^{m_i} Y_{i,j},$$

$$n_0 = \sum_{i=1}^{I} \sum_{j=1}^{m_i} (1 - Y_{i,j}).$$

Metodou maximální věrohodnosti, která je popsána v kapitole 2.1, získáme odhady $\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}$ parametrů $\beta_0, \boldsymbol{\beta}$. Pomocí těchto odhadů spočítáme odhady logistických pravděpodobností

$$\tilde{\pi}(\mathbf{x}_i) = \tilde{\pi}_i = \left[1 + \exp(-\tilde{\beta}_0 - \tilde{\boldsymbol{\beta}}' \mathbf{x}_i)\right]^{-1}.$$

Přímo z věrohodnostních rovnic uvedených v kapitole 2.1 plyne vztah

$$n_1 = \sum_{i=1}^{I} \sum_{j=1}^{m_i} Y_{i,j} = \sum_{i=1}^{I} m_i \tilde{\pi}_i,$$

$$n_0 = \sum_{i=1}^{I} \sum_{j=1}^{m_i} (1 - Y_{i,j}) = \sum_{i=1}^{I} m_i (1 - \tilde{\pi}_i).$$

Nechť

$$Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{i,j}.$$

Popíšeme test dobré shody založený na Pearsonově χ^2 statistice a test poměrem věrohodností.

4.1.A. Pearsonův χ^2 test

Celou situaci lze popsat kontingenční tabulkou typu $2 \times I$ s danými marginálními sloupcovými četnostmi. Přitom i-tý sloupec tabulky reprezentuje binomické rozdělení s parametry m_i, $\pi(\mathbf{x}_i) = \pi_i$.
Tabulka odhadnutých teoretických četností má tedy tvar:

	$\mathbf{X}$			
Y	$\mathbf{x}_1$	...	$\mathbf{x}_I$	
1	$m_1\tilde{\pi}_1$	...	$m_I\tilde{\pi}_I$	n_1
0	$m_1(1-\tilde{\pi}_1)$	...	$m_I(1-\tilde{\pi}_I)$	n_0
	m_1	...	m_I	n

Tabulka empirických (pozorovaných) četností je tvaru:

	$\mathbf{X}$			
Y	$\mathbf{x}_1$	...	$\mathbf{x}_I$	
1	$Y_{1\cdot}$	...	$Y_{I\cdot}$	n_1
0	$m_1 - Y_{1\cdot}$	...	$m_I - Y_{I\cdot}$	n_0
	m_1	...	m_I	n

Protože máme dány marginální sloupcové četnosti, jsou hodnoty v naší tabulce vázány I podmínkami $(m_1\pi_1 + m_1(1-\pi_1) = m_1, \ldots, m_I\pi_I + m_I(1-\pi_I) = m_I)$. Tento údaj budeme potřebovat pro výpočet stupňů volnosti v Pearsonově testové statistice, která má tvar

$$\chi^2 = \sum_{i=1}^{I} \frac{(Y_{i\cdot} - m_i\tilde{\pi}_i)^2}{m_i\tilde{\pi}_i} + \sum_{i=1}^{I} \frac{(m_i - Y_{i\cdot} - m_i(1-\tilde{\pi}_i))^2}{m_i(1-\tilde{\pi}_i)} = \sum_{i=1}^{I} \frac{(Y_{i\cdot} - m_i\tilde{\pi}_i)^2}{m_i\tilde{\pi}_i(1-\tilde{\pi}_i)}.$$

Pomocí této statistiky lze testovat shodu dat v pozorované tabulce s tabulkou teoretickou, která je v našem případě založena na modelu logistické regrese (podrobnější informace o testech dobré shody lze získat např. v [12]). Při hypotéze H_0: platí logistický model, má statistika χ^2 asymptoticky rozdělení χ^2 s následujícím počtem stupňů volnosti: velikost kontingenční tabulky − počet vazeb v teoretické tabulce − počet odhadnutých prametrů $= 2I - I - (p+1) = I - (p+1)$.
Tedy při platnosti H_0 je

$$\mathfrak{L}(\chi^2) \approx \chi^2\big(I - (p+1)\big).$$

Samozřejmě musí být splněna podmínka $I > p + 1$.

4.1.B. Test poměrem věrohodností

Uvažujme nadmodel logistického modelu, ve kterém nebudou pravděpodobnosti $\pi_i = P(Y = 1|\mathbf{X} = \mathbf{x}_i)$ vázány žádným funkcionálním vztahem, tj. $\pi_i = \pi(\mathbf{x}_i) = p_i$, $i = 1, \ldots, I$. Takto definovaný model je obvykle nazýván jako saturovaný model. Připomeňme, že maximálně věrohodné odhady parametrů získáme maximalizací logaritmu sdružené podmíněné hustoty vektoru $\mathbf{Y} = (Y_{1,1}, \ldots, Y_{1,m_1}, \ldots, Y_{I,1}, \ldots, Y_{I,m_I})'$ za podmínky $\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_I = \mathbf{x}_I$, která je:

$$f(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_I) = \prod_{i=1}^{I} \prod_{j=1}^{m_i} \pi_i^{y_{i,j}} \big(1-\pi_i\big)^{1-y_{i,j}}.$$

Odtud logaritmická věrohodnostní funkce je rovna:

$$\mathfrak{l}(\pi_1, \ldots, \pi_I) = \sum_{i=1}^{I} \Big[\sum_{j=1}^{m_i} Y_{i,j} \ln \pi_i + \big(m_i - \sum_{j=1}^{m_i} Y_{i,j}\big) \ln(1-\pi_i) \Big].$$

Z vyjádření logaritmické věrohodnostní funkce již snadno získáme maximálně věrohodné odhady parametrů $\pi_1, \ldots, \pi_I$ v saturovaném modelu, které jsou:

$$\check{\pi}_1 = \frac{Y_{1\cdot}}{m_1} = y_{1\cdot}$$

$$\ldots$$

$$\check{\pi}_I = \frac{Y_{I\cdot}}{m_I} = y_{I\cdot}.$$

Maximálně věrohodné odhady v logistickém modelu, jež je podmodelem modelu saturovaného, již známe: $\tilde{\pi}_1, \ldots, \tilde{\pi}_I$. Nyní můžeme vyjádřit testovou statistiku D pro test hypotézy H_0: platí logistický model, získanou metodou poměrem věrohodností.

$$\begin{aligned} D &= -2\ln\Big(\frac{\text{věrohodnost podmodelu}}{\text{věrohodnost modelu}}\Big) \\ &= -2\big(l(\tilde{\pi}_1, \ldots, \tilde{\pi}_I) - l(\check{\pi}_1, \ldots, \check{\pi}_I)\big) \\ &= 2\sum_{i=1}^{I}\Big[Y_{i\cdot}\ln\frac{Y_{i\cdot}}{m_i\tilde{\pi}_i} + (m_i - Y_{i\cdot})\ln\frac{m_i - Y_{i\cdot}}{m_i(1-\tilde{\pi}_i)}\Big]. \end{aligned}$$

Statistika D je obvykle nazývána jako deviance logistického modelu a má při platnosti hypotézy H_0 asymptoticky rozdělení χ^2 s počtem stupňů volnosti, který je roven rozdílu dimenzí prostoru parametrů v modelu a v podmodelu, což je v našem případě $I-(p+1)$. Tudíž při platnosti H_0 je

$$\mathfrak{L}(D) \approx \chi^2\big(I - (p+1)\big).$$

Opět musí být splněna podmínka $I > p+1$.

Poznámka 4.1.1. Jak je uvedeno v poznámce 2.1.2, je v případě nenáhodných čísel $m_1, \ldots, m_I$ nutné při hledání maximálně věrohodných odhadů maximalizovat sdruženou hustotu veličin $Y_{1\cdot}, \ldots, Y_{I\cdot}$ za podmínky $\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_I = \mathbf{x}_i$. Z ní odvozená logaritmická věrohodnost se však od té námi použité liší pouze o člen $\sum_{i=1}^{I}\ln\binom{m_i}{Y_{i\cdot}}$, který nezávisí na odhadovaných parametrech. Nezmění se tedy ani odhady v saturovaném modelu, ani deviance, neboť o stejný člen se liší od námi použitých věrohodností věrohodnost jak saturovaného, tak zkoumaného modelu.

Nyní by bylo vhodné upozornit na některé problémy spojené s použitím výše uvedených testů dobré shody. Prvně, tyto testy lze použít pouze v případě, že hodnoty $\mathbf{x}_1, \ldots, \mathbf{x}_I$ jsou skutečně pevně dány a číslo I je zvoleno předem. Pokud má však alespoň jedna složka vysvětlujících náhodných vektorů $\mathbf{X}_1, \ldots, \mathbf{X}_n$ spojitý charakter a chceme-li, aby hodnoty $\mathbf{x}_1, \ldots, \mathbf{x}_I$ byly dostatečně reprezentativní pro dané rozdělení, bude obvykle nutné volit $I \approx n$. Rozdělení testových statistik je však získáno asymptoticky pro $n \to \infty$ a tedy pokud $I \approx n$, roste s rozsahem výběru též počet stupňů volnosti testových statistik. V publikaci [10] je uvedeno, že pro $I \approx n$ je při platnosti H_0 : platí logistický model, $E\chi^2 < I - (p+1)$ a též $ED < I - (p+1)$.

Dalším problémem, který je spojen s použitím Pearsonovy χ^2 statistiky, je požadavek na dostatečně velké odhadnuté teoretické četnosti, např. $m_i\tilde{\pi}_i \geq 5,\ m_i(1-\tilde{\pi}_i) \geq 5,\ i = 1, \ldots, I$, který též nebude obvykle splněn, pokud $I \approx n$.

Oba výše uvedené problémy lze vyřešit, pokud bude počet sloupců dříve uvedených kontingenčních tabulek předem pevně daný a menší než rozsah výběru n. Popíšeme si testovou statistiku navrženou v [8], která je založena právě na této myšlence.

4.2. Hosmerovy-Lemeshowovy testy

Statistiky vhodné pro testy dobré shody navržené Hosmerem a Lemeshowem v [8] jsou založeny na seskupení některých sloupců kontingenčních tabulek uvedených v kapitole 4.1. Je dobré připomenout, že tyto testy jsou vhodným zlepšením testů dobré shody uvedených v téže kapitole pro situaci, v níž jsou veličiny $\mathbf{X}_1, \ldots, \mathbf{X}_n$ spojité a s rostoucím rozsahem výběru roste počet možných kombinací jejich hodnot. Nyní nemusí být číslo I nutně zvoleno předem, stejně jako hodnoty $\mathbf{x}_1, \ldots, \mathbf{x}_I$.

Nejprve zvolíme $g < n$ počet požadovaných sloupců kontingenční tabulky. Pozorování přeznačíme tak, aby platilo $\tilde{\pi}_1 \leq \tilde{\pi}_2 \leq \cdots \leq \tilde{\pi}_I$. Autoři uvádějí dvě možné metody seskupování sloupců.

(1) Metoda, jejímž výsledkem jsou skupiny obsahující přibližně stejný počet pozorování, pracuje následujícím způsobem. Do prvního sloupce zařadíme přibližně n/g pozorování $Y_{1,1}, \ldots, Y_{1,m_1}, \ldots, Y_{n_1',1}, \ldots, Y_{n_1',m_{n_1'}}$, kterým náleží nejmenší odhadnuté pravděpodobnosti $\tilde{\pi}_i,\ i = 1, \ldots, n_1'$. Naší snahou je, aby $m_1^* = \sum_{i=1}^{n_1'} m_i$ bylo co možná nejblíže hodnotě n/g. Postupně vytváříme další sloupce, až konečně v g-tém sloupci je přibližně n/g pozorování $Y_{t,1}, \ldots, Y_{t,m_t}, \ldots, Y_{I,1}, \ldots, Y_{I,m_I}$, kterým náleží největší odhadnuté pravděpodobnosti $\tilde{\pi}_i,\ i = t, \ldots, I$ $(t = \sum_{k=1}^{g-1} n_k' + 1)$, přitom $n_1', \ldots, n_g'$ označují počty různých hodnot vysvětlujících veličin $\mathbf{X}_1, \ldots, \mathbf{X}_I$ v jednotlivých sloupcích (tedy platí $\sum_{k=1}^{g} n_k' = I$). Nechť $t_0 = 0,\ t_k = \sum_{j=1}^{k} n_j',\ k = 1, \ldots, g$ a nechť $m_1^*, \ldots, m_g^*$ jsou počty pozorování v jednotlivých sloupcích, tedy splňují vztahy $m_k^* = \sum_{i=t_{k-1}+1}^{t_k} m_i,\ k = 1, \ldots, g$. Snažíme se, aby m_k^* bylo co nejblíže hodnotě $n/g\ \forall k = 1, \ldots, g$. Je-li $g = 10$, nazývají se hodnoty odhadnutých pravděpodobností, jež oddělují jednotlivé sloupce jako *decily rizika*. Samotné sloupce kontingenční tabulky budeme v naší práci nazývat *decilovými skupinami.*⋆

(2) Metoda založená na pevných dělících bodech. Pomocí této metody zařadíme do prvního sloupce ta pozorování $Y_{1,1}, \ldots, Y_{1,m_1}, \ldots, Y_{n_1',1}, \ldots, Y_{n_1',m_{n_1'}}$, která splňují $0 \leq \tilde{\pi}_i \leq 1/g,\ i = 1, \ldots, n_1'$, do k-tého sloupce pozorování $Y_{t_{k-1}+1,1}, \ldots, Y_{t_{k-1}+1,m_{t_{k-1}+1}}, \ldots, Y_{t_k,1}, \ \ldots, \ Y_{t_k,m_{t_k}}$, jež splňují $(k-1)/g < \tilde{\pi}_i \leq k/g,\ i = t_{k-1}+1, \ldots, t_k,\ k = 2, \ldots, g$. Hodnoty $n_1', \ldots, n_g', t_0, \ldots, t_g, m_1^*, \ldots, m_g^*$ mají stejný význam jako u předcházející metody seskupování.

Pro novou kontingenční tabulku typu $2 \times g$ nyní spočítáme odhadnuté teoretické a empirické četnosti. Odhadnutá teoretická četnost pro řádek $Y = 1$ a k-tý sloupec je

$$c_k = \sum_{i=t_{k-1}+1}^{t_k} m_i \tilde{\pi}_i,\ k = 1, \ldots, g,$$

pro řádek $Y = 0$ a k-tý sloupec:

$$m_k^* - c_k = \sum_{i=t_{k-1}+1}^{t_k} m_i (1 - \tilde{\pi}_i),\ k = 1, \ldots, g.$$

⋆O decilech se mluví i v situacích, kdy není v každém sloupci přesně desetina všech pozorování.

Empirická četnost pro řádek $Y = 1$ a k-tý sloupec je

$$o_k = \sum_{i=t_{k-1}+1}^{t_k} \sum_{j=1}^{m_i} Y_{i,j},\ k = 1, \ldots, g,$$

pro řádek $Y = 0$ a k-tý sloupec:

$$m_k^* - o_k = \sum_{i=t_{k-1}+1}^{t_k} \sum_{j=1}^{m_i} (1 - Y_{i,j}),\ k = 1, \ldots, g.$$

Dále nechť

$$\overline{\pi}_k = \frac{1}{m_k^*} \sum_{i=t_{k-1}+1}^{t_k} m_i \tilde{\pi}_i = \frac{c_k}{m_k^*},\ k = 1, \ldots, g$$

je odhad pravděpodobnosti $P\big(Y = 1 | \mathbf{X} \in \{\mathbf{x}_{t_{k-1}+1}, \ldots, \mathbf{x}_{t_k}\}\big)$.
Tabulka odhadnutých teoretických četností má tedy tvar:

	X			
Y	1. sloupec	...	g-tý sloupec	
1	$m_1^* \overline{\pi}_1$	...	$m_g^* \overline{\pi}_g$	n_1
0	$m_1^*(1 - \overline{\pi}_1)$	...	$m_g^*(1 - \overline{\pi}_g)$	n_0
	m_1^*	...	m_g^*	n

Tabulka empirických (pozorovaných) četností je tvaru:

	X			
Y	1. sloupec	...	g-tý sloupec	
1	o_1	...	o_g	n_1
0	$m_1^* - o_1$	...	$m_g^* - o_g$	n_0
	m_1^*	...	m_g^*	n

Testová statistika Hosmerova-Lemeshowova testu pro ověřování shody s modelem logistické regrese má tvar běžné Pearsonovy χ^2 statistiky pro ověřování shody teoretické a empirické tabulky, tedy

$$\begin{aligned} \hat{C} &= \sum_{k=1}^{g} \frac{\big(o_k - m_k^* \overline{\pi}_k\big)^2}{m_k^* \overline{\pi}_k} + \sum_{k=1}^{g} \frac{\big(m_k^* - o_k - m_k^*(1 - \overline{\pi}_k)\big)^2}{m_k^*(1 - \overline{\pi}_k)} \\ &= \sum_{k=1}^{g} \frac{(o_k - m_k^* \overline{\pi}_k)^2}{m_k^* \overline{\pi}_k (1 - \overline{\pi}_k)}. \end{aligned}$$

Užitím rozsáhlých simulací (podrobněji viz [9]) bylo ukázáno, že pro $I = n$ má při platnosti hypotézy H_0 : platí logistický model, statistika $\hat{C}$ přibližně rozdělení χ^2 o $g - 2$ stupních volnosti. Podle [8] lze při platnosti hypotézy H_0 dobře aproximovat rozdělení

statistiky $\hat{C}$ rozdělením $\chi^2(g-2)$ též v situaci, kdy $I \approx n$. Dále autoři [8] doporučují používat první metodu spojování sloupců založenou na percentilech z důvodu lepší dosažené shody testové statistiky s rozdělením $\chi^2(g-2)$, obzvláště v situacích, kdy je většina z odhadnutých pravděpodobností $\tilde{\pi}_1, \ldots, \tilde{\pi}_I$ blízkých 0.

Aby bylo možné použít výše uvedenou statistiku, měli bychom ještě ověřit např. podmínku $m_k^* \overline{\pi}_k \geq 5$, $m_k^*(1-\overline{\pi}_k) \geq 5$, $k = 1, \ldots, g$. Není-li tato podmínka splněna, měli bychom sloučit některé sloupce tabulky a tedy snížit hodnotu čísla g. Autoři [8] však tvrdí, že porušení této podmínky není příliš na závadu. V [8] je dále doporučováno volit $g \geq 6$, neboť pro $g < 6$ je již statistika $\hat{C}$ málo citlivá na rozdíly mezi teoretickými a empirickými četnostmi a téměř vždy indikuje shodu s modelem.

4.3. Shodnost variančních matic více výběrů

Na tomto místě se seznámíme s testem, který slouží k ověření shody variančních matic několika populací. Tento test, jež je zobecněním známého Bartlettova testu, byl publikován např. v [7].

Nechť $\mathbf{X}_1^1, \ldots, \mathbf{X}_{n_1}^1, \ldots, \mathbf{X}_1^K, \ldots, \mathbf{X}_{n_K}^K$ jsou vzájemně nezávislé náhodné výběry z K p-rozměrných rozdělení s variančními maticemi $\Sigma^1, \ldots, \Sigma^K$. Je třeba ověřit hypotézu $H_0 : \Sigma^1 = \cdots = \Sigma^K$. Nechť

$$\overline{\mathbf{X}}^k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{X}_i^k, \qquad k = 1, \ldots, K,$$

$$\mathfrak{S}^k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\mathbf{X}_i^k - \overline{\mathbf{X}}^k)(\mathbf{X}_i^k - \overline{\mathbf{X}}^k)', \qquad k = 1, \ldots, K$$

je průměr a výběrová varianční matice pro k-tou populaci. Nechť dále

$$n = \sum_{k=1}^{K} n_k,$$

$$\mathfrak{S} = \frac{1}{n-K} \sum_{k=1}^{K} (n_k - 1)\mathfrak{S}^k.$$

Postup užívaný k ověření hypotézy H_0 je znám jako *Boxův test* a pracuje s následující testovou statistikou:

$$B = \frac{1}{C_p} \left[(n-K) \ln |\mathfrak{S}| - \sum_{k=1}^{K} (n_k - 1) \ln |\mathfrak{S}^k| \right], \text{ kde}$$

$$C_p = 1 + \frac{2p^2 + 3p - 1}{6(K-1)(p+1)} \left(\sum_{k=1}^{K} \frac{1}{n_k - 1} - \frac{1}{n-K} \right).$$

Statisika B má přitom při platnosti hypotézy H_0 přibližně rozdělení χ^2 o $(K-1)\frac{p(p+1)}{2}$ stupních volnosti.

V. VOLBA SPRÁVNÉHO MODELU

První otázkou, kterou si položí každý, kdo je nucen zabývat se diskriminací, nejspíše bude, který z nabízených modelů bude v dané situaci optimální. Patrně neexistuje žádný obecný návod pro volbu toho nejlepšího modelu, ale pokusíme se shrnout některé argumenty pro a proti jednotlivým modelům.

Nebudeme podrobně uvádět argumenty týkající se modelu směsi normálních rozdělení, neboť tento model se od modelu normální diskriminační analýzy liší pouze neznalostí zařazení objektů z učící skupiny do jednotlivých tříd. Není-li tedy k dispozici tato informace, musíme použít tento model, neboť v těch dvou zbylých je k sestavení odhadu diskriminační funkce nutná znalost příslušnosti objektů z učící skupiny do jednotlivých tříd. Měli bychom však mít na paměti předpoklady tohoto modelu, především normalitu náhodných veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$, pomocí nichž chceme provádět diskriminaci. Jsme-li však v situaci, že známe zařazení objektů z učící skupiny do jednotlivých tříd, máme na výběr mezi modelem logistické regrese a modelem normální diskriminační analýzy, kterému za těchto podmínek dáme vždy přednost před modelem směsi normálních rozdělení. Volbě mezi logistickým modelem a modelem normální diskriminační analýzy se nyní budeme věnovat podrobněji.

5.1. Výhody modelu logistické regrese

Jak již bylo řečeno, model logistické regrese neklade žádné požadavky na rozdělení vysvětlujících veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Naproti tomu, model normální diskriminační analýzy požaduje v prvé řadě mnohorozměrnou normalitu veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$, což v praxi často nebývá splněno. Obvykle mají alespoň některé složky náhodných vektorů $\mathbf{X}_1, \ldots, \mathbf{X}_n$ diskrétní rozdělení (binomické apod.), nebo spojité rozdělení jiné než normální. Speciálně výskyt binomických veličin mezi složkami náhodných vektorů $\mathbf{X}_1, \ldots, \mathbf{X}_n$ není žádným vzácným jevem.

Shrneme zde některá fakta, uvedená v článcích [11] a [6], která podporují použití odhadu diskriminační funkce, získaného užitím maximálně věrohodných odhadů parametrů v logistickém modelu oproti odhadu z modelu normální diskriminační analýzy. Připomeňme, že odhady parametrů β_0, $\boldsymbol{\beta}$ v modelu logistické regrese značíme $\tilde{\beta}_0$, $\tilde{\boldsymbol{\beta}}$ a v modelu normální diskriminační analýzy $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$.

(1) Není-li rozdělení vysvětlujících veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$ mnohorozměrné normální, nejsou odhady $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$ parametrů β_0, $\boldsymbol{\beta}$ z diskriminační funkce $S(\mathbf{x})$ obecně konzistentní.

Tudíž ani poměrně velký rozsah učící skupiny (tj. $n \to \infty$) nezajistí správný odhad diskriminační funkce $S(\mathbf{x})$. Vyskytuje-li se tedy mezi vysvětlujícími proměnnými nějaká jasně nenormální veličina, nejčastěji kvalitativní znak, je vhodnější k odhadu diskriminační funkce použít logistického modelu, který zajistí konzistenci odhadů $\tilde{\beta}_0$, $\tilde{\boldsymbol{\beta}}$ parametrů β_0, $\boldsymbol{\beta}$ v diskriminační funkci $S(\mathbf{x})$.

Příklad. Nechť vysvětlující náhodné veličiny $X_1, \ldots, X_n$ mají alternativní rozdělení,

$$\begin{aligned} \text{tj. } & P(X_i = 1) = \gamma_1 \in (0,1), \\ & P(X_i = 0) = \gamma_0 = 1 - \gamma_1 \in (0,1), \quad i = 1, \ldots, n. \end{aligned}$$

Logistický model má tedy dva parametry β_0, β a je tvaru

$$\begin{aligned} P(Y_i = 1 | X_i = 1) &= \left[1 + \exp(-\beta_0 - \beta)\right]^{-1} = \pi_1, \\ P(Y_i = 1 | X_i = 0) &= \left[1 + \exp(-\beta_0)\right]^{-1} = \pi_0, \quad i = 1, \ldots, n. \end{aligned}$$

Zaveďme následující označení:

$$\begin{aligned} n_1 &= \sum_{i=1}^{n} Y_i, & n_0 &= \sum_{i=1}^{n} (1 - Y_i), \\ r_1 &= \sum_{i:X_i=1} Y_i, & r_0 &= \sum_{i:X_i=0} Y_i, \\ N_1 &= \sum_{i=1}^{n} X_i, & N_0 &= \sum_{i=1}^{n} (1 - X_i). \end{aligned}$$

Schématicky můžeme vše zapsat do tabulky četností:

	X		
Y	1	0	
1	r_1	r_0	n_1
0	$N_1 - r_1$	$N_0 - r_0$	n_0
	N_1	N_0	n

Odhady parametrů v modelu normální diskriminační analýzy jsou podle kapitoly 2.2 následující:

$$\hat{\mu}_1 = \frac{r_1}{n_1}, \qquad \hat{\mu}_0 = \frac{N_1 - r_1}{n_0} = \frac{N_1 - r_1}{n - n_1}, \qquad \hat{\lambda} = \frac{n_1}{n}.$$

Nechť

$$\begin{aligned} \mathbb{S}_1 &= \sum_{i:Y_i=1} (X_i - \hat{\mu}_1)^2 = \left(r_1 - \frac{r_1^2}{n_1}\right), \\ \mathbb{S}_0 &= \sum_{i:Y_i=0} (X_i - \hat{\mu}_0)^2 = \left(N_1 - r_1 - \frac{(N_1 - r_1)^2}{n - n_1}\right). \end{aligned}$$

Potom

$$\hat{\Sigma} = \frac{1}{n}(\mathbb{S}_1 + \mathbb{S}_0) = \frac{1}{n}\Big(\frac{r_1 r_0}{n_1} + \frac{(N_1 - r_1)(N_0 - r_0)}{n - n_1}\Big).$$

Odhady parametrů β_0, β pomocí modelu normální diskriminační analýzy jsou opět podle kapitoly 2.2:

$$\begin{aligned}
\hat{\beta} = \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) &= n\frac{r_1 N_0 - r_0 N_1}{r_0 N_1(N_0 - r_0) + r_1 N_0(N_1 - r_1)} \\
&= \frac{\frac{r_1}{n}\frac{N_0}{n} - \frac{r_0}{n}\frac{N_1}{n}}{\frac{r_0}{n}\frac{N_1}{n}(\frac{N_0}{n} - \frac{r_0}{n}) + \frac{r_1}{n}\frac{N_0}{n}(\frac{N_1}{n} - \frac{r_1}{n})}, \\
\hat{\beta}_0 = -\frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_0)\hat{\beta} - \ln\frac{1-\hat{\lambda}}{\hat{\lambda}} &= -\frac{1}{2}\Big(\frac{r_1}{n_1} + \frac{N_1 - r_1}{n - n_1}\Big)\hat{\beta} - \ln\frac{n - n_1}{n_1} \\
&= -\frac{1}{2}\Big(\frac{r_1}{n}\frac{n}{n_1} + (\frac{N_1}{n} - \frac{r_1}{n})\frac{n}{n_0}\Big)\hat{\beta} - \ln\big(\frac{n}{n_1} - 1\big).
\end{aligned}$$

Podle silného zákona velkých čísel platí:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{s.j.} EX_1 = \gamma_1, \qquad \text{tj. } \frac{N_1}{n} \xrightarrow[n\to\infty]{s.j.} \gamma_1,$$

$$\frac{1}{n}\sum_{i=1}^{n}(1 - X_i) \xrightarrow[n\to\infty]{s.j.} E(1 - X_1) = \gamma_0, \qquad \text{tj. } \frac{N_0}{n} \xrightarrow[n\to\infty]{s.j.} \gamma_0.$$

Tedy též platí:

$$N_1 \xrightarrow[n\to\infty]{s.j.} \infty, \qquad N_0 \xrightarrow[n\to\infty]{s.j.} \infty.$$

Dalším užitím silného zákona velkých čísel dostáváme platnost následujících tvrzení:

$$\frac{1}{N_1}\sum_{i:X_i=1} Y_i \xrightarrow[n\to\infty]{s.j.} E[Y_1|X_1 = 1] = \pi_1, \qquad \text{tj. } \frac{r_1}{N_1} \xrightarrow[n\to\infty]{s.j.} \pi_1,$$

$$\frac{1}{N_0}\sum_{i:X_i=0} Y_i \xrightarrow[n\to\infty]{s.j.} E[Y_1|X_1 = 0] = \pi_0, \qquad \text{tj. } \frac{r_0}{N_0} \xrightarrow[n\to\infty]{s.j.} \pi_0,$$

z nichž přímo plyne:

$$\frac{r_1}{n} \xrightarrow[n\to\infty]{s.j.} \pi_1\gamma_1, \qquad \frac{r_0}{n} \xrightarrow[n\to\infty]{s.j.} \pi_0\gamma_0.$$

Dále pomocí silného zákona velkých čísel dostáváme následující výsledky:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow[n\to\infty]{s.j.} EY_1 = P(Y_1 = 1) = \pi_1\gamma_1 + \pi_0\gamma_0,$$

$$\frac{1}{n}\sum_{i=1}^{n}(1 - Y_i) \xrightarrow[n\to\infty]{s.j.} E(1 - Y_1) = 1 - EY_1 = (1 - \pi_1)\gamma_1 + (1 - \pi_0)\gamma_0,$$

tj.

$$\frac{n_1}{n} \xrightarrow[n\to\infty]{s.j.} \pi_1\gamma_1 + \pi_0\gamma_0,$$
$$\frac{n_0}{n} \xrightarrow[n\to\infty]{s.j.} (1-\pi_1)\gamma_1 + (1-\pi_0)\gamma_0.$$

Užitím výše uvedených tvrzení již získáváme asymptotické hodnoty odhadů $\hat{\beta}_0,\ \hat{\beta}$:

$$\hat{\beta} \xrightarrow[n\to\infty]{s.j.} \beta^\star, \qquad \hat{\beta}_0 \xrightarrow[n\to\infty]{s.j.} {\beta_0}^\star,$$

kde

$$\beta^\star = \frac{\pi_1 - \pi_0}{\pi_0(1-\pi_0)\gamma_0 + \pi_1(1-\pi_1)\gamma_1},$$
$${\beta_0}^\star = -\frac{\gamma_1}{2}\Big(\frac{\pi_1}{\pi_1\gamma_1 + \pi_0\gamma_0} + \frac{1-\pi_1}{(1-\pi_1)\gamma_1 + (1-\pi_0)\gamma_0}\Big)\frac{\pi_1 - \pi_0}{\pi_0(1-\pi_0)\gamma_0 + \pi_1(1-\pi_1)\gamma_1} +$$
$$+ \ln\frac{\pi_1\gamma_1 + \pi_0\gamma_0}{(1-\pi_1)\gamma_1 + (1-\pi_0)\gamma_0}.$$

Skutečné hodnoty parametrů $\beta_0,\ \beta$ splňují:

$$\ln\frac{\pi_1}{1-\pi_1} = \beta_0 + \beta, \qquad \ln\frac{\pi_0}{1-\pi_0} = \beta_0,$$
$$\text{tedy} \qquad \beta = \ln\frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}, \qquad \beta_0 = \ln\frac{\pi_0}{1-\pi_0}.$$

Z výše uvedeného je vidět, že $\hat{\beta}_0,\ \hat{\beta}$ nejsou obecně konzistentními odhady parametrů $\beta_0,\ \beta$. Konzistence je zaručena, pokud $\pi_1 = \pi_0 \Leftrightarrow \beta = 0$. Tento případ je však z hlediska diskriminace poměrně nezajímavý, neboť pokud $\beta = 0$, je uvažovaný kvalitativní znak X pro diskriminaci nepřínosný.

Maximálně věrohodné odhady parametrů $\beta_0,\ \beta$ v modelu logistické regrese, které jsou konzistentní, získáme řešením věrohodnostních rovnic:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \pi(\mathbf{X}_i)$$
$$\sum_{i=1}^{n} Y_i X_i = \sum_{i=1}^{n} X_i \pi(X_i),$$

které mají v tomto případě tvar:

$$r_1 + r_0 = N_1\big[1 + \exp(-\beta_0 - \beta)\big]^{-1} + N_0\big[1 + \exp(-\beta_0)\big]^{-1}$$
$$r_1 = N_1\big[1 + \exp(-\beta_0 - \beta)\big]^{-1}.$$

Jejich řešením dostáváme konzistentní odhady parametrů $\beta_0,\ \beta$:

$$\tilde{\beta}_0 = \ln \frac{r_0}{N_0 - r_0}, \qquad \tilde{\beta} = \ln \frac{r_1(N_0 - r_0)}{r_0(N_1 - r_1)}.$$

Pro ilustraci byla pomocí programu MATLAB generována následující data (označíme je jako `Matlab data`, neboť si na nich budeme ilustrovat i některé další vlastnosti zkoumaných modelů) při hodnotách parametrů

$$\gamma_1 = 0{,}75\ ,$$
$$\pi_1 = 0{,}5\ , \quad \pi_0 = 0{,}25\ :$$

Matlab data

	X		
Y	1	0	
1	37314	6344	43658
0	37417	18925	56342
	74731	25269	100000

Pomocí dříve uvedených vzorců získáme následující výsledky (zaokrouhlené na čtyři desetinná místa):
skutečné hodnoty parametrů $\beta_0,\ \beta$

$$\beta_0 = -1{,}0986\ , \quad \beta = 1{,}0986\ ,$$

jejich odhady v modelu logistické regrese

$$\tilde{\beta}_0 = -1{,}0930\ , \quad \tilde{\beta} = 1{,}0902\ ,$$

asymptotické hodnoty odhadů v modelu normální diskriminační analýzy

$$\beta_0^\star = -1{,}0640\ , \quad \beta^\star = 1{,}0667\ ,$$

a odhady parametrů $\beta_0,\ \beta$ pomocí modelu normální diskriminační analýzy

$$\hat{\beta}_0 = -1{,}0595\ , \quad \hat{\beta} = 1{,}0594.$$

Vidíme, že odhady získané pomocí modelu normální diskriminační analýzy jsou mírně vychýlené, na rozdíl od odhadů spočítaných pomocí modelu logistické regrese, jež jsou téměř rovny teoretickým hodnotám parametrů $\beta_0,\ \beta$.

(2) Při porušení podmínek normality vysvětlujících veličin byla prováděna numerická porovnání odhadů $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}')'$ a $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}')'$ parametru $(\beta_0, \boldsymbol{\beta}')'$ získaných z modelů normální diskriminační analýzy a logistické regrese. Zmíněné odhady byly použity k získání odhadů $\hat{\pi}(\mathbf{x})$ a $\tilde{\pi}(\mathbf{x})$ logistické pravděpodobnosti $\pi(\mathbf{x}) = \left[1 + \exp(-\beta_0 - \boldsymbol{\beta}'\mathbf{x})\right]^{-1}$ pro různé

sady dat. Zkoumala se shoda empirických a odhadnutých teoretických četností mezi jednotlivými decilovými skupinami* pro oba dva modely. K vyjádření shody byl pro každý model a každou sadu dat vypočítán regresní koeficient v závislosti pozorovaných četností na četnostech odhadnutých (běžná lineární regrese bez absolutního členu). Lepší shodu teoretických a pozorovaných četností indikuje tento regresní koeficient bližší hodnotě 1. Bylo zjištěno (viz [6]), že lepší shody je obvykle dosaženo, pokud jsou použity odhady $\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}$, jež jsou získány užitím modelu logistické regrese.

Jako ilustraci si pomocí náhodného generátoru programu MATLAB vytvoříme dále popsaný datový soubor. Při značení kapitol 4.1 a 4.2 byly použity následující hodnoty:

$$I = 100, \qquad m_i = m = 10,\ i = 1, \ldots, I,$$
$$X_1 = 1,\ X_2 = 2, \ldots,\ X_{100} = 100.$$

Data byla generována při hodnotách parametrů β_0, β :

$$\beta_0 = -5, \qquad \beta = 0{,}1\ ,$$

tj.

$$\pi(x) = P(Y = 1|X = x) = [1 + \exp(5 - 0{,}1x)]^{-1}\,.$$

Hodnoty pozorovaných četností $\sum_{j=1}^{m_i} Y_{i,j}$, $i = 1, \ldots, I$ je možné nalézt na připojené disketě v souboru `pr5_1_2_95.xls` (formát pro EXCEL 95) ve sloupci označeném jako *počet zdarů, je-li* $X = X_i$.

Odhady parametrů β_0, β jsou (zaokrouhleno na čtyři desetinná místa) v modelu logistické regrese:

$$\tilde{\beta}_0 = -5{,}3187, \qquad \tilde{\beta} = 0{,}1064$$

a v modelu normální diskriminační analýzy:

$$\hat{\beta}_0 = -6{,}5383, \qquad \hat{\beta} = 0{,}1304.$$

Vzhledem k tomu, že $\beta > 0$, $\tilde{\beta} > 0$ a $\hat{\beta} > 0$, jsou funkce $\pi(x)$ a také $\tilde{\pi}(x) = [1+\exp(-\tilde{\beta}_0 - \tilde{\beta}x)]^{-1}$, stejně jako $\hat{\pi}(x) = [1 + \exp(-\hat{\beta}_0 - \hat{\beta}x)]^{-1}$ rostoucí. Hodnoty vysvětlujících veličin $X_1, \ldots, X_{100}$ jsou voleny monotónně od jedné do sta a pro každou hodnotu veličiny X máme stejný počet deseti pozorování. S využitím právě řečeného zjišťujeme, že decilové skupiny budou shodné pro oba uvažované modely, přičemž v každé skupině bude obsaženo právě sto pozorování. V první skupině pozorování s hodnotami veličiny X od jedné do deseti, v druhé skupině s hodnotami od jedenácti do dvaceti atd.

Pomocí vzorců z kapitoly 4.2 spočítáme odhady teoretických četností (označíme je $\tilde{c}_1, \ldots \tilde{c}_{10}$, resp. $\hat{c}_1, \ldots, \hat{c}_{10}$), pozorované četnosti (označené $o_1, \ldots, o_{10}$) a skutečné teoretické četnosti (označené $c_1, \ldots, c_{10}$) v jednotlivých decilových skupinách (zaokrouhleno na dvě desetinná místa).

*Konstrukce decilových skupin je popsána v kapitole 4.2. Jedná se o první variantu seskupování sloupců vzniklé kontingenční tabulky pro $g = 10$.

Četnosti v decilových skupinách

decilová skupina (i)	1	2	3	4	5	6	7	8	9	10
c_i	1,20	3,19	8,20	19,39	39,15	63,16	82,11	92,52	97,10	98,91
o_i	1	5	7	16	40	58	88	93	99	98
$\tilde{c}_i$ (model LR)	0,91	2,59	7,13	18,06	38,52	63,94	83,46	93,54	97,66	99,18
$\hat{c}_i$ (model NDA)	0,32	1,16	4,12	13,51	35,81	66,34	87,58	96,25	98,95	99,71

V naší situaci známe též skutečné hodnoty parametrů β_0 a β. Můžeme tedy spočítat nejen regresní koeficient v závislosti pozorovaných četností na četnostech odhadnutých, ale také regresní koeficient závislosti teoretických četností na odhadnutých.

Uvažujme regresní model $U = \xi V + \varepsilon$, kde za U dosadíme postupně teoretické (c) a pozorované četnosti (o), za V postupně odhady četností v modelu logistické regrese ($\tilde{c}$) a modelu normální diskriminační analýzy ($\hat{c}$). ε je náhodná chyba. Metodou nejmenších čtverců získáme následující odhady parametru ξ.

Odhady parametru ξ

	V	
U	$\tilde{c}$ (LR)	$\hat{c}$ (NDA)
c (teor.)	0,9931	0,9743
o (pozor.)	0,9997	0,9813

(3) Užijeme-li odhadů koeficientů β_0, $\boldsymbol{\beta}$ získaných pomocí modelu normální diskriminační analýzy, můžeme přehlédnout některá problematická místa, neboť odhady $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$ nás na rozdíl od odhadů $\tilde{\beta}_0$, $\tilde{\boldsymbol{\beta}}$ před možnými potížemi nemusejí varovat. Uvedené problémy budeme ilustrovat na příkladu uvedeném v [11].

Příklad. Předpokládejme, že máme jednu veličinu, podle které provádíme diskriminaci a mějme následující data:

i	1	2	3	4	4	6	7	8
X_i	-4	-3	-2	-1	1	2	3	4
Y_i	0	0	0	0	1	1	1	1

Diskriminační funkce je tvaru $S(x) = \beta_0 + \beta x$ a obsahuje dva neznámé parametry. Model logistické regrese předpokládá vztah

$$\pi(x) = P(Y = 1|X = x) = [1 + \exp(-\beta_0 - \beta x)]^{-1}.$$

Maximálně věrohodné odhady v tomto modelu by měly splňovat rovnice

$$\sum_{i=1}^{8} Y_i = \sum_{i=1}^{8} \pi(X_i), \qquad \sum_{i=1}^{8} Y_i X_i = \sum_{i=1}^{8} X_i \pi(X_i),$$

které mají v naší situaci tvar

$$4 = \sum_{i=1}^{8} [1 + \exp(-\beta_0 - \beta X_i)]^{-1}$$

$$10 = -\sum_{i=1}^{4} (-X_i) [1 + \exp(-\beta_0 - \beta X_i)]^{-1} + \sum_{i=5}^{8} X_i [1 + \exp(-\beta_0 - \beta X_i)]^{-1}.$$

Druhá rovnice však nemůže být nikdy splněna, neboť její pravá strana je vždy menší než deset. Maximálně věrohodné odhady parametrů β_0, β tedy v modelu logistické regrese neexistují. Logaritmická věrohodnostní funkce je přitom tvaru $l(\beta_0, \beta) = 4\beta_0 + 10\beta - \sum_{i=1}^{8} \ln[1 + \exp(\beta_0 + \beta X_i)]$. Některé její vlastnosti dále podrobněji rozebereme.

V modelu normální diskriminační analýzy se předpokládá platnost vztahu

$$\mathfrak{L}(X|Y=1) = N_1(\mu_1, \sigma^2), \quad \mathfrak{L}(X|Y=0) = N_1(\mu_0, \sigma^2),$$
$$P(Y=1) = \lambda, \quad P(Y=0) = 1-\lambda$$

a maximálně věrohodné odhady neznámých parametrů získáme snadno jako

$$\hat{\mu}_1 = \frac{1}{4}\sum_{i=5}^{8} X_i = \frac{5}{2}, \quad \hat{\mu}_0 = \frac{1}{4}\sum_{i=1}^{4} X_i = -\frac{5}{2},$$
$$\hat{\sigma}^2 = \frac{1}{8}\Big[\sum_{i=1}^{4}(X_i - \hat{\mu}_0)^2 + \sum_{i=5}^{8}(X_i - \hat{\mu}_1)^2\Big] = \frac{5}{4},$$
$$\hat{\lambda} = \frac{1}{8}\sum_{i=1}^{8} Y_i = \frac{1}{2}.$$

Tyto odhady přitom maximalizují sdruženou hustotu veličin $Y_1, \ldots, Y_8$, $X_1, \ldots, X_8$. Z nich podle kapitoly 1.2 nalezneme odhady $\hat{\beta}_0$, $\hat{\beta}$ koeficientů v diskriminační funkci:

$$\hat{\beta} = \frac{1}{\hat{\sigma}^2}(\hat{\mu}_1 - \hat{\mu}_0) = 4, \quad \hat{\beta}_0 = \ln\left(\frac{\hat{\lambda}}{1-\hat{\lambda}}\right) - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_0)\hat{\beta} = 0.$$

Tedy model normální diskriminační analýzy nám zajišťuje odhad diskriminační funkce $\hat{S}(x) = 4x$, aniž bychom byli jakkoliv varováni před možnými nesrovnalostmi. Podle kapitoly 3.2 totiž při platnosti modelu normální diskriminační analýzy platí též model logistické regrese, jehož parametry můžeme odhadnout pomocí maximálně věrohodných odhadů z modelu normální diskriminační analýzy, které jsou rovny právě $\hat{\beta}_0$, $\hat{\beta}$. Tudíž podle modelu normální diskriminační analýzy je odhad pro $\pi(x) = P(Y=1|X=x)$ roven $\hat{\pi}(x) = [1 + \exp(-4x)]^{-1}$. Přitom logaritmická věrohodnostní funkce, s níž se operuje v modelu logistické regrese (jde o logaritmus podmíněné hustoty $Y_1, \ldots, Y_8$ za podmínky $X_1, \ldots, X_8$), má v bodě $(\beta_0, \beta) = (0, 4)$ hodnotu $l(0, 4) \doteq -0{,}037$ a tuto hodnotu můžeme libovolně přiblížit k 0, pokud zvětšíme koeficient β, neboť funkce $f(\beta) = l(0, \beta)$ je rostoucí a $\lim_{\beta\to\infty} f(\beta) = 0$.

V modelu logistické regrese nelze jednoznačně určit nejlepší křivku požadovaného tvaru $\big[1 + \exp(-\beta_0 - \beta x)\big]^{-1}$, která by vysvětlovala vztah $\pi(x) = P(Y=1|X=x)$. K tomu, aby toto bylo možné, je zapotřebí přidat další objekty do učící skupiny (pokud to je prakticky proveditelné), které mají hodnotu vysvětlujících znaků X z intervalu $(-1, 1)$. Tento fakt ovšem nezjistíme, pokud použijeme k odhadu parametrů pouze modelu normální diskriminační analýzy.

(4) S modelem logistické regrese je spojena $(p+1)$-rozměrná postačující statistika

$$\mathbf{U}(Y_1,\ldots,Y_n|\mathbf{X}_1,\ldots,\mathbf{X}_n)=\mathbf{U}(\mathbf{Y}|\mathbb{X})=\sum_{i=1}^{n}\begin{pmatrix}1\\ \mathbf{X}_i\end{pmatrix}Y_i,$$

neboť podmíněnou hustotu (2.1.1) lze postupně upravit následujícím způsobem:

$$\begin{aligned}f_{\beta_0,\boldsymbol{\beta}}(\mathbf{y}|\mathbf{x}_1,\ldots,\mathbf{x}_n)&=\prod_{i=1}^{n}\pi(\mathbf{x}_i)^{y_i}\big(1-\pi(\mathbf{x}_i)\big)^{1-y_i}\\ &=\prod_{i=1}^{n}\left[\frac{\exp(\beta_0+\boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\beta_0+\boldsymbol{\beta}'\mathbf{x}_i)}\right]^{y_i}\left[\frac{1}{1+\exp(\beta_0+\boldsymbol{\beta}'\mathbf{x}_i)}\right]^{1-y_i}\\ &=\exp\left\{(\beta_0,\boldsymbol{\beta}')\mathbf{U}(\mathbf{y}|\mathbf{x}_1,\ldots,\mathbf{x}_n)-\sum_{i=1}^{n}\ln\big[1+\exp(\beta_0+\boldsymbol{\beta}'\mathbf{x}_i)\big]\right\}.\end{aligned}$$

Z teorie je známo, že maximálně věrohodné odhady (tj. $\tilde{\beta}_0$, $\tilde{\boldsymbol{\beta}}$) jsou funkcí této postačující statistiky. Naproti tomu odhady $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$ parametrů β_0, $\boldsymbol{\beta}$, nalezené pomocí modelu normální diskriminační analýzy, u něhož nejsou splněny všechny potřebné předpoklady (zejména normalita), jsou získány maximalizací jakési funkce (2.2.1), jež není ani podmíněnou hustotou $Y_1,\ldots,Y_n$ za podmínky $\mathbf{X}_1,\ldots,\mathbf{X}_n$, ani sdruženou hustotou $Y_1,\ldots,Y_n,\mathbf{X}_1,\ldots,\mathbf{X}_n$. Tudíž tyto odhady nemusejí být funkcí statistiky $\mathbf{U}$. Podle Raovy-Blackwellovy věty mají tedy odhady $\tilde{\beta}_0$, $\tilde{\boldsymbol{\beta}}$ stejnou nebo menší střední čtvercovou chybu než odhady $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$.

(5) Maximálně věrohodné odhady parametrů $\beta_0,\boldsymbol{\beta}$ v modelu logistické regrese (tj. $\tilde{\beta}_0,\tilde{\boldsymbol{\beta}}$) splňují rovnost

$$\sum_{i=1}^{n}Y_i=\sum_{i=1}^{n}\tilde{\pi}(\mathbf{X}_i), \tag{5.1.1}$$

kde $\tilde{\pi}(\mathbf{X}_i)=\big[1+\exp(-\tilde{\beta}_0-\tilde{\boldsymbol{\beta}}'\mathbf{X}_i)\big]^{-1}$. Pohlížejme nyní na náhodné veličiny $Y_1,\ldots,Y_n$ jako na příznaky nějakého stavu (např. nemoci), tj. $Y_i=1$, pokud i-tý jedinec trpí danou chorobou a $Y_i=0$, je-li zdráv. Mezi složkami náhodných vektorů $\mathbf{X}_1,\ldots,\mathbf{X}_n$ může být zahrnut např. krevní tlak, tělesná hmotnost apod. Při tomto pohledu $\sum_{i=1}^{n}Y_i$ označuje počet nemocných jedinců v učící skupině (tj. pozorovanou četnost) a $\sum_{i=1}^{n}\tilde{\pi}(\mathbf{X}_i)$ je odhadem střední hodnoty počtu nemocných jedinců v učící skupině (tj. odhadem teoretické četnosti). Tedy podle vztahu (5.1.1) zajišťují odhady $\tilde{\beta}_0$, $\tilde{\boldsymbol{\beta}}$ rovnost mezi pozorovanou četností a odhadem četnosti teoretické. Vztah (5.1.1) je přitom nultou věrohodnostní rovnicí (viz kapitola 2.1). Odhady $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$ získané užitím modelu normální diskriminační analýzy dávají často odhad teoretické četnosti dosti odlišný od četnosti pozorované.

Pro ilustraci využijeme opět datového souboru `Matlab data`. V tomto případě je

$$\sum_{i=1}^{n}Y_i=\sum_{i=1}^{n}\tilde{\pi}(X_i)=\sum_{i=1}^{n}\Big[1+\exp(-\tilde{\beta}_0-\tilde{\beta})\Big]^{-1}=43658$$

a po zaokrouhlení na celá čísla

$$\sum_{i=1}^{n} \hat{\pi}(X_i) = \sum_{i=1}^{n} \left[1 + \exp(-\hat{\beta}_0 - \hat{\beta})\right]^{-1} = 43867.$$

Vidíme, že odhad teoretické četnosti v modelu normální diskriminační analýzy je vyšší než je četnost pozorovaná.

5.2. Výhody modelu normální diskriminační analýzy

Podle předcházejících stránek by se mohlo zdát, že model normální diskriminační analýzy je vždy horší než logistický model. Na následujících řádcích se pokusíme ukázat, že tomu tak není a že při splnění předpokladu normality vysvětlujících náhodných vektorů $\mathbf{X}_1, \ldots, \mathbf{X}_n$ je výhodnější k diskriminaci použít modelu normální diskriminační analýzy. Popisované výsledky se přitom opírají o článek [4].

Vycházejme z modelu normální diskriminační analýzy, tj. $Y_1, \ldots, Y_n$ je posloupnost nezávislých náhodných veličin s alternativním rozdělením, kde

$$P(Y_i = 1) = \lambda \in (0,1), \quad i = 1, \ldots, n. \tag{5.2.1}$$

Přitom Y_i vyjadřuje zařazení i-tého objektu. Dále $\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*$ je posloupnost nezávislých náhodných veličin, jejichž podmíněné rozdělení je

$$\begin{aligned} \mathfrak{L}(\mathbf{X}_i^* | Y_i = 1) &= N_p(\boldsymbol{\mu}_1, \Sigma), \\ \mathfrak{L}(\mathbf{X}_i^* | Y_i = 0) &= N_p(\boldsymbol{\mu}_0, \Sigma), \quad i = 1, \ldots, n, \end{aligned} \tag{5.2.2}$$

kde $\boldsymbol{\mu}_1, \boldsymbol{\mu}_0 \in \mathbb{R}^p$, $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_0$, Σ je pozitivně definitní matice typu $p \times p$.

Podle kapitoly 3.2 platí za výše uvedených předpokladů též logistický model, tj.

$$\begin{aligned} P(Y_i = 1 | \mathbf{X}_i^* = \mathbf{x}_i) &= \left[1 + \exp(-\beta_0 - \boldsymbol{\beta}' \mathbf{x}_i)\right]^{-1}, \\ P(Y_i = 0 | \mathbf{X}_i^* = \mathbf{x}_i) &= \left[1 + \exp(\beta_0 + \boldsymbol{\beta}' \mathbf{x}_i)\right]^{-1}, \quad i = 1, \ldots, n, \end{aligned} \tag{5.2.3}$$

kde

$$\begin{aligned} \boldsymbol{\beta} &= \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \\ \beta_0 &= \ln \frac{\lambda}{1-\lambda} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)' \boldsymbol{\beta} \\ &= \ln \frac{\lambda}{1-\lambda} - \frac{1}{2} \boldsymbol{\mu}_1' \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_0' \Sigma^{-1} \boldsymbol{\mu}_0. \end{aligned} \tag{5.2.4}$$

Nový objekt, na němž jsme naměřili hodnotu $\mathbf{X}^*$ pomocných znaků zařadíme do jedné ze dvou daných skupin podle teorie uvedené v kapitolách 1.1 a 1.2. Tj. objekt zařadíme do první skupiny, pokud

$$S^*(\mathbf{X}^*) \geq 0, \quad \text{kde } S^*(\mathbf{X}^*) = \beta_0 + \boldsymbol{\beta}' \mathbf{X}^*.$$

Přitom $S^*(\mathbf{x})$ je teoretická diskriminační funkce, kterou k vlastnímu rozlišování nových objektů nemůžeme použít. Nechť

$$\mathfrak{B}^* = \{\mathbf{x}^* : S^*(\mathbf{x}^*) = 0\}$$

je hranice mezi oblastmi $\mathfrak{R}_1^* = \{\mathbf{x}^* : S^*(\mathbf{x}^*) > 0\}$ a $\mathfrak{R}_0^* = \{\mathbf{x}^* : S^*(\mathbf{x}^*) < 0\}$. Při zařazování nového objektu do skupiny se potom řídíme tím, do které z oblastí $\mathfrak{R}_1^*$, $\mathfrak{R}_0^*$ náleží naměřená hodnota $\mathbf{X}^*$. Leží-li $\mathbf{X}^*$ na hranici $\mathfrak{B}^*$, je lhostejno, do které skupiny nový objekt zařadíme, jak bylo ukázáno v kapitolách 1.1 a 1.2.

K vlastní diskriminaci využíváme odhadnutých hranic $\hat{\mathfrak{B}}^*$, resp. $\tilde{\mathfrak{B}}^*$ a odhadnutých oblastí $\hat{\mathfrak{R}}_1^*$, $\hat{\mathfrak{R}}_0^*$, resp. $\tilde{\mathfrak{R}}_1^*$, $\tilde{\mathfrak{R}}_0^*$, v jejichž vyjádření jsou neznámé parametry nahrazeny odhady $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$, resp. $\tilde{\beta}_0$, $\tilde{\boldsymbol{\beta}}$. Přitom $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$ se spočítají podle vzorce (5.2.4) nahrazením parametrů $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, Σ, λ jejich odhady $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_0$, $\hat{\Sigma}$, $\hat{\lambda}$, které získáme podle kapitoly 2.2, odhady $\tilde{\beta}_0$, $\tilde{\boldsymbol{\beta}}$ najdeme užitím postupů z kapitoly 2.1. Nechť $\hat{S}^*(\mathbf{x})$, resp. $\tilde{S}^*(\mathbf{x})$ jsou příslušné odhadnuté diskriminační funkce, získané z modelu normální diskriminační analýzy, resp. logistické regrese.

Kvalitu diskriminační procedury můžeme posuzovat podle hodnoty bayesovské rizikové funkce $\rho_q(\delta)$, kterou lze v naší situaci interpretovat jako pravděpodobnost nesprávné klasifikace, jak bylo ukázáno v kapitolách 1.1, 1.2 a dodatku A. Nazvěme tuto veličinu jako *pravděpodobnost chyby* a označme ji *err*. Nechť je p-rozměrný eukleidovský prostor $\mathbb{R}^p$ rozdělen na dvě oblasti $\mathfrak{D}_1$ a $\mathfrak{D}_0$ tak, že nový objekt, na kterém jsme naměřili hodnotu $\mathbf{X}^*$ pomocných znaků, zařadíme do první skupiny, pokud $\mathbf{X}^* \in \mathfrak{D}_1$ a do nulté skupiny, pokud $\mathbf{X}^* \in \mathfrak{D}_0$. Podle zmíněných kapitol lze *err* vyjádřit následovně:

$$\begin{aligned} err &= P(\mathbf{X}^* \in \mathfrak{D}_1, Y = 0) + P(\mathbf{X}^* \in \mathfrak{D}_0, Y = 1) \\ &= P(Y = 0)P(\mathbf{X}^* \in \mathfrak{D}_1 | Y = 0) + \\ &\quad + P(Y = 1)P(\mathbf{X}^* \in \mathfrak{D}_0 | Y = 1) \\ &= (1 - \lambda)P\big(\mathbf{X}^* \in \mathfrak{D}_1 \big| E\mathbf{X}^* = \boldsymbol{\mu}_0\big) + \\ &\quad + \lambda P\big(\mathbf{X}^* \in \mathfrak{D}_0 \big| E\mathbf{X}^* = \boldsymbol{\mu}_1\big). \end{aligned} \tag{5.2.5}$$

Pravděpodobnost chyby je přitom minimální, užíváme-li k diskriminaci teoretickou diskriminační funkci $S^*(\mathbf{x})$. Tj. $\mathfrak{D}_1 = \mathfrak{R}_1^*$, $\mathfrak{D}_0 = \mathfrak{R}_0^*$ a hranice $\mathfrak{B}^*$ je připojena k libovolné z oblastí $\mathfrak{D}_1$, $\mathfrak{D}_0$. Přitom volba oblasti, ke které hranici $\mathfrak{B}^*$ připojíme, nemá vliv na hodnotu *err*. Tuto minimální hodnotu pravděpodobnosti chyby označíme err_0.

Používáme-li k diskriminaci odhadů diskriminační funkce, stává se z pravděpodobnosti chyby náhodná veličina, neboť rozdělení eukleidovského prostoru $\mathbb{R}^p$ na oblasti závisí na hodnotách náhodných veličin $Y_1, \ldots, Y_n$ a $\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*$, pomocí nichž počítáme odhady koeficientů v diskriminační funkci. Kvalitu takto získané diskriminační procedury lze potom posuzovat např. podle střední hodnoty pravděpodobnosti chyby nebo podle odchylky této střední hodnoty od minimální hodnoty err_0. Po provedení předběžných úvah vyjádříme pravděpodobnost chyby a střední pravděpodobnost chyby pro diskriminační procedury založené na odhadech v modelu normální diskriminační analýzy a modelu logistické regrese.

Nejprve převedeme obecný model (5.2.2) na tzv. standardní situaci. Jak je ukázáno v dodatku B, lze užitím lineární transformace $\mathbf{X}_i = \mathbf{a} + \mathbb{A}\mathbf{X}_i^*$, $i = 1, \ldots, n$, $\mathbf{a} \in \mathbb{R}^p$, $\mathbb{A} \in \mathbb{R}^{p \times p}$, (5.2.2) převést na

$$\begin{aligned} \mathfrak{L}(\mathbf{X}_i | Y_i = 1) &= N_p(\frac{\Delta}{2}\mathbf{e}_1, I_p), \\ \mathfrak{L}(\mathbf{X}_i | Y_i = 0) &= N_p(-\frac{\Delta}{2}\mathbf{e}_1, I_p), \quad i = 1, \ldots, n, \end{aligned} \tag{5.2.6}$$

kde $\Delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}$ je zobecněná Mahalanobisova vzdálenost vektorů $\boldsymbol{\mu}_1$ a $\boldsymbol{\mu}_0$, $\mathbf{e}_1 = (1, 0, \ldots, 0)' \in \mathbb{R}^p$ a I_p je jednotková matice typu $p \times p$. Přitom $\mathbf{X}_1, \ldots, \mathbf{X}_n$ jsou stále nezávislé náhodné veličiny. Nový objekt, na kterém jsme naměřili $\mathbf{X}^*$ potom zařazujeme na základě hodnoty $\mathbf{X} = \mathbf{a} + \mathbb{A}\mathbf{X}^*$ následovně. Při použití teoretické diskriminační funkce zařadíme nový objekt do první skupiny, pokud $\mathbf{X} \in \mathfrak{R}_1 = \{\mathbf{x} : \mathbf{x} = \mathbf{a} + \mathbb{A}\mathbf{x}^*, \mathbf{x}^* \in \mathfrak{R}_1^*\}$, do nulté skupiny v případě, že $\mathbf{X} \in \mathfrak{R}_0 = \{\mathbf{x} : \mathbf{x} = \mathbf{a} + \mathbb{A}\mathbf{x}^*, \mathbf{x}^* \in \mathfrak{R}_0^*\}$ a do libovolné skupiny, pokud $\mathbf{X} \in \mathfrak{B} = \{\mathbf{x} : \mathbf{x} = \mathbf{a} + \mathbb{A}\mathbf{x}^*, \mathbf{x}^* \in \mathfrak{B}^*\}$. Analogicky získáme transformace oblastí $\hat{\mathfrak{R}}_1$, $\hat{\mathfrak{R}}_0$, resp. $\tilde{\mathfrak{R}}_1$, $\tilde{\mathfrak{R}}_0$ a hranic $\hat{\mathfrak{B}}$, resp. $\tilde{\mathfrak{B}}$, popisujících diskriminaci založenou na odhadech diskriminační funkce v modelech normální diskriminační analýzy, resp. logistické regrese. Lze říci, že diskriminační procedury založené na transformovaných datech jsou totožné s procedurami založenými na transformacích oblastí, které určují diskriminační procedury založené na datech původních. Pravděpodobnost chyby pro jednotlivé modely má zřejmě stejné rozdělení jak při modelu (5.2.2), tak při modelu (5.2.6). Z tohoto důvodu budeme dále pracovat pouze s modelem (5.2.6), který budeme nazývat jako standardní model.

5.2.A. Minimální pravděpodobnost chyby err_0

Nechť

$$\omega = \ln \frac{\lambda}{1 - \lambda}.$$

Při platnosti standardního modelu má teoretická diskriminační funkce tvar:

$$\begin{aligned} S(\mathbf{x}) &= \omega - \frac{1}{2}\frac{\Delta^2}{4}\mathbf{e}_1{}' I_p \mathbf{e}_1 + \frac{1}{2}\frac{\Delta^2}{4}\mathbf{e}_1{}' I_p \mathbf{e}_1 + \Delta \mathbf{e}_1{}' I_p^{-1}\mathbf{x} \\ &= \omega + \Delta x^1. \end{aligned} \tag{5.2.7}$$

Přitom x^1 je první složka vektoru $\mathbf{x}$, tj. $\mathbf{x} = (x^1, \ldots, x^p)'$. Geometricky lze hranici $\mathfrak{B} = \{\mathbf{x} : S(\mathbf{x}) = 0\} = \{\mathbf{x} : x^1 = -\omega/\Delta\}$ interpretovat jako $(p-1)$-rozměrnou nadrovinu v $\mathbb{R}^p$, která je kolmá na osu x_1 a protíná ji v bodě

$$\tau = -\frac{\omega}{\Delta}.$$

Potom

$$
\begin{aligned}
err_0 &= (1-\lambda)P\left(S(\mathbf{X}) \geq 0 \middle| E\mathbf{X} = -\frac{\Delta}{2}\mathbf{e}_1\right) + \\
&\quad + \lambda P\left(S(\mathbf{X}) < 0 \middle| E\mathbf{X} = \frac{\Delta}{2}\mathbf{e}_1\right) \\
&= (1-\lambda)P\left(X^1 \geq \tau \middle| EX^1 = -\frac{\Delta}{2}\right) + \\
&\quad + \lambda P\left(X^1 < \tau \middle| EX^1 = \frac{\Delta}{2}\right) \\
&= (1-\lambda)\Phi\left(-(\tau + \frac{\Delta}{2})\right) + \lambda\Phi\left(\tau - \frac{\Delta}{2}\right) \\
&= (1-\lambda)\Phi(-D_0) + \lambda\Phi(-D_1),
\end{aligned} \tag{5.2.8}
$$

neboť za všech okolností je $\operatorname{var}\mathbf{X} = I_p$ a $\operatorname{var}X^1 = 1$. Přitom Φ je distribuční funkce normovaného normálního rozdělení $N_1(0,1)$ a

$$
\begin{aligned}
D_1 &= \frac{\Delta}{2} - \tau, \\
D_0 &= \frac{\Delta}{2} + \tau.
\end{aligned} \tag{5.2.9}
$$

5.2.B. Aproximace střední pravděpodobnosti chyby v modelech logistické regrese a normální diskriminační analýzy

Nechť $S(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}'\mathbf{x}$ je teoretická diskriminační funkce, $\hat{S}(\mathbf{x}) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}$ je odhad diskriminační funkce v modelu normální diskriminační analýzy a $\tilde{S}(\mathbf{x}) = \tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}'\mathbf{x}$ odhad diskriminační funkce v modelu logistické regrese, vždy pro standardní situaci. Dále nechť

$$
\begin{aligned}
d\hat{\beta}_0 &= \hat{\beta}_0 - \beta_0, \qquad & d\tilde{\beta}_0 &= \tilde{\beta}_0 - \beta_0, \\
d\hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \qquad & d\tilde{\boldsymbol{\beta}} &= \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}
\end{aligned} \tag{5.2.10}
$$

a nechť je splněno

$$
\sqrt{n}\begin{pmatrix} d\hat{\beta}_0 \\ d\hat{\boldsymbol{\beta}} \end{pmatrix} \xrightarrow[n\to\infty]{\mathfrak{D}} \hat{\mathbf{Z}}, \qquad \sqrt{n}\begin{pmatrix} d\tilde{\beta}_0 \\ d\tilde{\boldsymbol{\beta}} \end{pmatrix} \xrightarrow[n\to\infty]{\mathfrak{D}} \tilde{\mathbf{Z}}, \tag{5.2.11}
$$

kde

$$
\begin{aligned}
\mathfrak{L}(\hat{\mathbf{Z}}) &= N_{p+1}(\mathbf{0}, \hat{\mathbb{V}}), \qquad & \hat{\mathbb{V}} &= (\hat{v}_{i,j})_{i,j=0,1,\ldots,p}, \\
\mathfrak{L}(\tilde{\mathbf{Z}}) &= N_{p+1}(\mathbf{0}, \tilde{\mathbb{V}}), \qquad & \tilde{\mathbb{V}} &= (\tilde{v}_{i,j})_{i,j=0,1,\ldots,p}.
\end{aligned}
$$

Při určitých metodách odhadu je možné tento předpoklad splnit (např. užitím metody maximální věrohodnosti).

Označme $\widehat{err}$, resp. $\widetilde{err}$ pravděpodobnosti chyb diskriminačních procedur získaných užitím odhadů neznámých parametrů metodou maximální věrohodnosti v modelu normální diskriminační analýzy, resp. logistické regrese (podle druhé kapitoly). Střední pravděpodobnost chyby lze vyjádřit dle článku [4] jako

(5.2.12)
$$E(\widehat{err} - err_0) = \frac{\lambda\varphi(D_1)}{2\Delta n}\left(\hat{v}_{0,0} - 2\frac{\omega}{\Delta}\hat{v}_{0,1} + \frac{\omega^2}{\Delta^2}\hat{v}_{1,1} + \hat{v}_{2,2} + \cdots + \hat{v}_{p,p}\right) + R_n,$$
$$E(\widetilde{err} - err_0) = \frac{\lambda\varphi(D_1)}{2\Delta n}\left(\tilde{v}_{0,0} - 2\frac{\omega}{\Delta}\tilde{v}_{0,1} + \frac{\omega^2}{\Delta^2}\tilde{v}_{1,1} + \tilde{v}_{2,2} + \cdots + \tilde{v}_{p,p}\right) + R_n,$$

kde $R_n = \mathfrak{o}(1/n)$, $n \to \infty$, $\varphi(t) = (2\pi)^{-\frac{1}{2}}\exp(-t^2/2)$ je hustota normovaného normálního rozdělení. Přitom matice $\hat{\mathbb{V}}$, resp. $\tilde{\mathbb{V}}$ mají následující tvar:

$$\hat{\mathbb{V}} = \frac{1}{\lambda(1-\lambda)}\begin{pmatrix} 1+\frac{\Delta^2}{4} & -\frac{\Delta}{2}(1-2\lambda) & 0 & \dots & 0 \\ -\frac{\Delta}{2}(1-2\lambda) & 1+2\Delta^2\lambda(1-\lambda) & 0 & \dots & 0 \\ 0 & 0 & 1+\Delta^2\lambda(1-\lambda) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1+\Delta^2\lambda(1-\lambda) \end{pmatrix},$$

$$\tilde{\mathbb{V}} = \frac{1}{\lambda(1-\lambda)}\begin{pmatrix} \frac{A_2}{A_0A_2-A_1^2} & -\frac{A_1}{A_0A_2-A_1^2} & 0 & \dots & 0 \\ -\frac{A_1}{A_0A_2-A_1^2} & \frac{A_0}{A_0A_2-A_1^2} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{A_0} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{A_0} \end{pmatrix},$$

kde

$$A_i = A_i(\lambda,\Delta) = \int_{-\infty}^{\infty} \frac{\exp(-\frac{\Delta^2}{8})x^i\varphi(x)}{\lambda\exp(x\frac{\Delta}{2}) + (1-\lambda)\exp(-x\frac{\Delta}{2})}dx, \quad i = 0,1,2.$$

Užitím těchto výsledků můžeme spočítat *asymptotickou relativní vydatnost* $Eff_p(\omega,\Delta)$ logistické regrese vzhledem k normální diskriminační analýze, kterou definujeme jako

(5.2.13)
$$Eff_p(\omega,\Delta) = \lim_{n\to\infty}\frac{E(\widehat{err} - err_0)}{E(\widetilde{err} - err_0)}$$

a jež udává, kolikrát je odchylka střední pravděpodobnosti chyby od její minimální hodnoty v modelu normální diskriminační analýzy nižší (nebo vyšší) než v modelu logistické regrese, jde-li rozsah výběru učící skupiny do nekonečna. Z (5.2.12) dostáváme

(5.2.14)
$$Eff_p(\omega,\Delta) = \frac{1+\frac{\Delta^2}{4}+2\frac{\omega}{\Delta}\frac{\Delta}{2}(1-2\lambda)+\frac{\omega^2}{\Delta^2}\left[1+2\Delta^2\lambda(1-\lambda)\right]+(p-1)\left[1+\Delta^2\lambda(1-\lambda)\right]}{(A_0A_2-A_1^2)^{-1}\left(A_2+2\frac{\omega}{\Delta}A_1+\frac{\omega^2}{\Delta^2}A_0\right)+(p-1)A_0^{-1}}$$
$$= \frac{Q_1+(p-1)Q_2}{Q_3+(p-1)Q_4},$$

kde

$$
\begin{aligned}
Q_1 &= \left(1, \frac{\omega}{\Delta}\right) \begin{pmatrix} 1+\frac{\Delta^2}{4} & \frac{\Delta}{2}(1-2\lambda) \\ \frac{\Delta}{2}(1-2\lambda) & 1+2\Delta^2\lambda(1-\lambda) \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\omega}{\Delta} \end{pmatrix}, \\
Q_2 &= 1+\Delta^2\lambda(1-\lambda), \\
Q_3 &= \frac{1}{A_0A_2 - A_1^2} \left(1, \frac{\omega}{\Delta}\right) \begin{pmatrix} A_2 & A_1 \\ A_1 & A_0 \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\omega}{\Delta} \end{pmatrix}, \\
Q_4 &= \frac{1}{A_0}.
\end{aligned}
$$

Protože $Eff_1(\omega, \Delta) = Q_1/Q_3$ a $Eff_\infty(\omega, \Delta) = \lim_{p\to\infty} Eff_p(\omega, \Delta) = Q_2/Q_4$, lze psát

$$
Eff_p(\omega, \Delta) = \frac{q(\omega, \Delta)\, Eff_1(\omega, \Delta) + (p-1)\, Eff_\infty(\omega, \Delta)}{q(\omega, \Delta) + (p-1)}, \tag{5.2.15}
$$

kde $q(\omega, \Delta) = Q_3/Q_4$. Tj. $Eff_p(\omega, \Delta)$ je váženým průměrem $Eff_1(\omega, \Delta)$ a $Eff_\infty(\omega, \Delta)$.

Poznamenejme, že v případě $\omega = 0$ (tj. $\lambda = 1/2$) platí vztah

$$
Eff_p(\omega, \Delta) = Eff_\infty(\omega, \Delta) = A_0 \left(1 + \frac{\Delta^2}{4}\right) \ \forall p \in \mathbb{N},
$$

neboť pro $\omega = 0$ je $A_1 = 0$, $q(0, \Delta) = 1$ a $Q_1/Q_3 = Q_2/Q_4$.

V následující tabulce je numericky vyjádřeno $Eff_p(\omega, \Delta)$ pro některé hodnoty ω, Δ, $p = 1$ a $p = \infty$. Vzhledem k tomu, že $q(\omega, \Delta)$ se výrazně neliší od jedné, bude pro vyšší hodnoty dimenze p ($p \geq 3$) $Eff_p(\omega, \Delta)$ blízké $Eff_\infty(\omega, \Delta)$. Pro námi spočítané hodnoty ω, Δ je vždy $Eff_p(\omega, \Delta) < 1$ a klesá s rostoucí zobecněnou Mahalanobisovou vzdáleností středních hodnot měřených znaků u dvou uvažovaných populací. Tedy vyšší zobecněná Mahalanobisova vzdálenost středních hodnot $\boldsymbol{\mu}_1$ a $\boldsymbol{\mu}_0$ znamená vyšší střední pravděpodobnost chyby modelu logistické regrese oproti modelu normální diskriminační analýzy (uvažováno relativně). Vše, s výjimkou řádků, kde $\Delta = 10$, je zaokrouhleno na čtyři desetinná místa. Výpočty byly provedeny pomocí programu MAPLE. Obdobnou tabulku, ovšem pro jiné hodnoty λ a Δ je možné nalézt též v [4].

Asymptotická relativní vydatnost (LR) vzhledem k (NDA)

λ	ω	Δ	$Eff_1(\omega,\Delta)$	$Eff_\infty(\omega,\Delta)$	$q(\omega,\Delta)$
0,5	0	0,5	0,9999	0,9999	1
0,5	0	1	0,9949	0,9949	1
0,5	0	2	0,8992	0,8992	1
0,5	0	5	0,1374	0,1374	1
0,5	0	10	$2,32 \cdot 10^{-5}$	$2,32 \cdot 10^{-5}$	1
0,75	1,0986	0,5	0,9972	0,9987	5,5426
0,75	1,0986	1	0,9835	0,9840	1,9887
0,75	1,0986	2	0,9145	0,8545	1,1773
0,75	1,0986	5	0,1530	0,1222	1,0116
0,75	1,0986	10	$2,65 \cdot 10^{-5}$	$2,02 \cdot 10^{-5}$	1,0010
0,9	2,1972	0,5	0,9951	0,9983	19,1166
0,9	2,1972	1	0,9479	0,9768	4,9058
0,9	2,1972	2	0,8043	0,8006	1,6968
0,9	2,1972	5	0,1839	0,0954	1,0460
0,9	2,1972	10	$3,65 \cdot 10^{-5}$	$1,46 \cdot 10^{-5}$	1,0040
0,95	2,9444	0,5	0,9964	0,9988	33,5987
0,95	2,9444	1	0,9483	0,9811	7,9937
0,95	2,9444	2	0,7062	0,8015	2,2329
0,95	2,9444	5	0,2033	0,0834	1,0822
0,95	2,9444	10	$4,74 \cdot 10^{-5}$	$1,13 \cdot 10^{-5}$	1,0072

5.3. Doporučení pro volbu správného modelu

Nyní se pokusíme stručně shrnout některá doporučení, kterými by se měl řídit každý, kdo chce použít některý ze studovaných modelů k diskriminaci. V jednotlivých odstavcích bude uvedeno, za jakých podmínek je daný model nejvhodnější, případně kdy není vhodný vůbec.

5.3.A. Model směsi normálních rozdělení

Model (MND) musíme použít, jestliže neznáme zařazení učících objektů do skupin. Měla by být ovšem splněna normalita vysvětlujících veličin v každé skupině — obtížně ověřitelný předpoklad — tento požadavek nebude v žádném případě splněn, je-li rozdělení některých složek vektoru $\mathbf{X}$ diskrétní. Pokud známe zařazení učících objektů, je lepší používat některý ze zbývajících dvou modelů.

5.3.B. Model logistické regrese

Základním předpokladem pro použití modelu (LR) je znalost zařazení učících objektů do skupin. Je-li rozdělení vysvětlujících veličin normální, je vhodnější použít model (NDA) (viz kapitola 5.2). Modelu (LR) užijeme v případě, že některé vysvětlující veličiny nejsou normálně rozděleny (obvykle mají nějaké diskrétní rozdělení) (viz kapitola 5.1).

Další situace, ve které bude volba modelu (LR) nejvhodnější, je případ, kdy jsou veličiny $\mathbf{X}_1, \ldots, \mathbf{X}_n$ nenáhodné (nastaveny experimentátorem) (viz poznámky 2.1.2 a 2.2.7). V případě, že jsou pevné naopak hodnoty $Y_1, \ldots, Y_n$, lze též použít modelu (LR), ale je nutné upravit odhad absolutního členu v diskiminační funkci (viz poznámka 2.1.3).

5.3.C. Model normální diskriminační analýzy

Též model (NDA) vyžaduje znalost zařazení učících objektů do skupin. Modelu (NDA) dáme přednost před (LR), jestliže vektory vysvětlujících veličin jsou normálně rozděleny (viz kapitola 5.2).

VI. VÝPOČTY ODHADŮ V MODELU SMĚSI NORMÁLNÍCH ROZDĚLENÍ

Maximálně věrohodné odhady v modelu normální diskriminační analýzy získáme prostým dosazením do dříve uvedených vzorců. Odhady z modelu logistické regrese lze snadno spočítat pomocí některého z numerických nebo statistických souborů programů, neboť logaritmická věrohodnostní funkce je ryze konkávní, jak bylo ukázáno v kapitole 2.1 a tudíž při její maximalizaci nenastávají žádné podstatnější potíže. Na rozdíl od dvou výše uvedených modelů se při výpočtu maximálně věrohodných odhadů v modelu směsi normálních rozdělení setkáváme s jistými problémy, neboť zde může být věrohodnostní funkce neomezená pro $\Sigma \to \mathbf{0}$ (nulová matice) (viz kapitola 2.3). K získání odhadů budeme využívat tzv. EM algoritmu, jehož princip si dále popíšeme.

6.1. EM algoritmus

Podstatu EM algoritmu si ukážeme na modelu směsi normálních rozdělení. Nechť náhodné veličiny $\mathbf{X}_1, \ldots, \mathbf{X}_n$ a $Y_1, \ldots, Y_n$ mají stejný význam jako v kapitole 1.3 ($\mathbf{X}_1, \ldots, \mathbf{X}_n$ označují naměřené znaky na daných objektech a $Y_1, \ldots, Y_n$ příslušnost objektů do jednotlivých skupin). Hodnoty $Y_1, \ldots, Y_n$ přitom u tohoto modelu neznáme. Nechť θ označuje neznámé parametry (v našem případě tedy $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, λ, Σ) a Ω nechť značí prostor přípustných parametrů. Pro získání maximálně věrohodných odhadů je nutné přes $\theta \in \Omega$ maximalizovat funkci

$$\mathfrak{l}(\theta) = \ln f_\theta(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \ln f_\theta(\mathbb{X}) = \sum_{i=1}^{n} \ln\big[\lambda g_1(\mathbf{X}_i) + (1-\lambda) g_0(\mathbf{X}_i)\big],$$

kde g_1, resp. g_0 jsou hustoty $N_p(\boldsymbol{\mu}_1, \Sigma)$, resp. $N_p(\boldsymbol{\mu}_0, \Sigma)$. Funkce $\mathfrak{l}(\theta)$ je tedy logaritmem sdružené hustoty neúplných dat (tj. dat, která neposkytují informaci o zařazení jednotlivých objektů do skupin).

Pokud bychom znali též $Y_1, \ldots, Y_n$, maximalizovali bychom funkci

$$\begin{aligned} &\ln h_\theta(\mathbf{X}_1, \ldots, \mathbf{X}_n, Y_1, \ldots, Y_n) = \ln h_\theta(\mathbb{X}, \mathbf{Y}) \\ &\quad = \ln\lambda \sum_{i=1}^{n} Y_i + \ln(1-\lambda) \sum_{i=1}^{n} (1 - Y_i) + \sum_{i:Y_i=1} g_1(\mathbf{X}_i) + \sum_{i:Y_i=0} g_0(\mathbf{X}_i), \end{aligned}$$

což je logaritmická věrohodnostní funkce v modelu normální diskriminační analýzy (jde o logaritmus sdružené hustoty veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n, Y_1, \ldots, Y_n$, tj. úplných dat).

EM algoritmus se řadí mezi iterační postupy. Každá iterace se přitom skládá ze dvou kroků. Z výpočtu podmíněné střední hodnoty (E krok) a z maximalizace jisté funkce (M krok). Nechť $\theta^{(1)}, \ldots, \theta^{(r)}$ jsou hodnoty parametru θ po r krocích algoritmu. Nový odhad $\theta^{(r+1)}$ získáme pomocí zmíněných dvou kroků.

E krok

Spočítáme

$$Q(\theta|\theta^{(r)}) = E\big[\ln h_\theta(\mathbb{X}, \mathbf{Y})\big|\mathbb{X}, \theta^{(r)}\big] \quad \forall\theta \in \Omega.$$

M krok

$$\theta^{(r+1)} = \operatorname*{argmax}_{\theta\in\Omega} Q(\theta|\theta^{(r)}).$$

Vyjádřeno slovy, EM algoritmus v každé iteraci maximalizuje podmíněnou střední hodnotu logaritmické věrohodnostní funkce z modelu, ve kterém máme k dispozici úplnou informaci (model normální diskriminační analýzy) při daných hodnotách $\mathbf{X}_1, \ldots, \mathbf{X}_n$ a poslední známé hodnotě parametru, kterou je $\theta^{(r)}$.

Nyní uvedeme některé základní vlastnosti právě popsaného algoritmu. Nechť je každá iterace EM algoritmu vyjádřena pomocí zobrazení

$$\begin{aligned} M &: \Omega \to \Omega \\ &\theta \mapsto M(\theta), \\ \text{tj. } &\theta^{(r+1)} = M(\theta^{(r)}). \end{aligned}$$

Potom platí

(1) $\forall\theta \in \Omega \quad \mathfrak{l}\big(M(\theta)\big) \geq \mathfrak{l}(\theta)$.

(2) Pokud pro $\theta^* \in \Omega$ je $\forall\theta \in \Omega \quad \mathfrak{l}(\theta^*) \geq \mathfrak{l}(\theta)$ (tj. θ^* je globálním maximem funkce $\mathfrak{l}$ na Ω), pak $\mathfrak{l}\big(M(\theta^*)\big) = \mathfrak{l}(\theta^*)$.

(3) Pokud pro $\theta^* \in \Omega$ je $\forall\theta \in \Omega \setminus \{\theta^*\} \quad \mathfrak{l}(\theta^*) > \mathfrak{l}(\theta)$, pak $M(\theta^*) = \theta^*$ (tj. bod jednoznačného globálního maxima funkce $\mathfrak{l}$ na Ω je pevným bodem EM algoritmu).

Důkazy výše uvedených tvrzení spolu s dalšími vlastnostmi EM algoritmu je možné nalézt v [3] a [2].

6.2. EM algoritmus pro směs normálních rozdělení

V modelu směsi normálních rozdělení dostáváme při označení

$$\begin{aligned} \mathbf{Z}_i &= (Y_i, 1 - Y_i)', \quad i = 1, \ldots, n, \\ \mathbf{c}(\lambda) &= (\ln\lambda, \ln(1-\lambda))', \\ \mathbf{d}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma) &= \begin{pmatrix} (\mathbf{X}_i - \boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_1) \\ (\mathbf{X}_i - \boldsymbol{\mu}_0)'\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_0) \end{pmatrix}, \end{aligned}$$

$$\ln h_\theta(\mathbb{X}, \mathbf{Y}) = -\frac{np}{2}\ln 2\pi - \frac{n}{2}\ln|\Sigma| + \sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{c}(\lambda) - \frac{1}{2}\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{d}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma).$$

Vzhledem k lineární závislosti této funkce na $Y_1, \ldots, Y_n$ získáme užitím Jensenovy nerovnosti následující vztah:

$$Q(\theta|\theta^*) = E\big[\ln h_\theta(\mathbb{X}, \mathbf{Y})\big|\mathbb{X}, \theta^*\big]$$
$$= -\frac{np}{2}\ln 2\pi - \frac{n}{2}\ln|\Sigma| + \sum_{i=1}^{n}E[\mathbf{Z}_i|\mathbb{X}, \theta^*]'\mathbf{c}(\lambda) + \sum_{i=1}^{n}E[\mathbf{Z}_i|\mathbb{X}, \theta^*]'\mathbf{d}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma).$$

Přitom podle kapitoly 2.3 platí:

$$E[\mathbf{Z}_i|\mathbb{X}, \theta^*] = \begin{pmatrix} E[Y_i|\mathbb{X}, \theta^*] \\ 1 - E[Y_i|\mathbb{X}, \theta^*] \end{pmatrix} = \begin{pmatrix} P(Y_i = 1|\mathbb{X}, \theta^*) \\ P(Y_i = 0|\mathbb{X}, \theta^*) \end{pmatrix} = \begin{pmatrix} w_i^1(\theta^*) \\ w_i^0(\theta^*) \end{pmatrix} = \mathbf{W}_i(\theta^*),$$

kde

$$w_i^1(\theta^*) = \frac{\lambda^* g_1^*(\mathbf{X}_i)}{\lambda^* g_1^*(\mathbf{X}_i) + (1-\lambda^*)g_0^*(\mathbf{X}_i)},$$
$$w_i^0(\theta^*) = \frac{(1-\lambda^*)g_0^*(\mathbf{X}_i)}{\lambda^* g_1^*(\mathbf{X}_i) + (1-\lambda^*)g_0^*(\mathbf{X}_i)} = 1 - w_i^1(\theta^*),$$

g_1^*, resp. g_0^* jsou hustoty $N_p(\boldsymbol{\mu}_1^*, \Sigma^*)$, resp. $N_p(\boldsymbol{\mu}_0^*, \Sigma^*)$. Tedy

$$Q(\theta|\theta^*) = -\frac{np}{2}\ln 2\pi - \frac{n}{2}\ln|\Sigma| + \sum_{i=1}^{n}\mathbf{W}_i(\theta^*)'\mathbf{c}(\lambda) + \sum_{i=1}^{n}\mathbf{W}_i(\theta^*)'\mathbf{d}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma).$$

Nyní si můžeme povšimnout, že $Q(\theta|\theta^*)$ je logaritmickou věrohodnostní funkcí pro model normální diskriminační analýzy, ve které jsou v této situaci neznámé hodnoty $Y_1, \ldots, Y_n$ nahrazeny vahami $w_1^1(\theta^*), \ldots, w_n^1(\theta^*)$. Maximalizace funkce $Q(\theta|\theta^*)$ přes $\theta \in \Omega$ tedy odpovídá výpočtu maximálně věrohodných odhadů v modelu normální diskriminační analýzy s tím, že hodnoty $Y_1, \ldots, Y_n$ jsou nahrazeny příslušnými vahami. Z výše uvedeného vyplývá, že bod $M(\theta^*) \in \Omega$, který maximalizuje $Q(\theta|\theta^*)$ přes $\theta \in \Omega$ splňuje rovnice (2.2.3), ve kterých jsou veličiny $Y_1, \ldots, Y_n$ nahrazeny vahami $w_1^1(\theta^*), \ldots, w_n^1(\theta^*)$. Takto upravené rovnice (2.2.3) přesně odpovídají rovnicím (2.3.3).

Tedy jestliže $\theta^{(1)}, \ldots, \theta^{(r)}$ jsou hodnoty parametru θ po r iteracích algoritmu, potom $\theta^{(r+1)}$ získáme z následujících vztahů:

(6.2.1)
$$\lambda^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} w_i^1(\theta^{(r)}),$$
$$\boldsymbol{\mu}_1^{(r+1)} = \frac{\sum_{i=1}^{n} w_i^1(\theta^{(r)})\mathbf{X}_i}{\sum_{i=1}^{n} w_i^1(\theta^{(r)})},$$
$$\boldsymbol{\mu}_0^{(r+1)} = \frac{\sum_{i=1}^{n}\big(1 - w_i^1(\theta^{(r)})\big)\mathbf{X}_i}{\sum_{i=1}^{n}\big(1 - w_i^1(\theta^{(r)})\big)},$$
$$\Sigma^{(r+1)} = \frac{1}{n}\Big\{\sum_{i=1}^{n}\Big[w_i^1(\theta^{(r)})\big(\mathbf{X}_i - \boldsymbol{\mu}_1^{(r+1)}\big)\big(\mathbf{X}_i - \boldsymbol{\mu}_1^{(r+1)}\big)' +$$
$$+ \big(1 - w_i^1(\theta^{(r)})\big)\big(\mathbf{X}_i - \boldsymbol{\mu}_0^{(r+1)}\big)\big(\mathbf{X}_i - \boldsymbol{\mu}_0^{(r+1)}\big)'\Big]\Big\}.$$

VII. NUMERICKÁ POROVNÁNÍ NĚKTERÝCH MODELŮ

Na tomto místě se budeme věnovat numerickému porovnání diskriminačních procedur založených na modelu logistické regrese a modelu normální diskriminační analýzy u daných konkrétních dat. Procedury můžeme porovnávat podle hodnoty bayesovské rizikové funkce, kterou jsme v kapitole 5.2 nazvali jako *pravděpodobnost chyby* a označili err. Odhadu pravděpodobnosti chyby v praktické situaci bude věnována následující část.

7.1. Odhad pravděpodobnosti chyby

Na úvod připomeňme značení. $(\mathbf{X}_1', Y_1)', \ldots, (\mathbf{X}_n', Y_n)'$ jsou data v učící skupině, na jejichž základě sestavujeme diskriminační proceduru (Y_i přitom nabývají hodnot nula a jedna podle skutečné příslušnosti i-tého objektu do jedné ze dvou skupin). Ve shodě s kapitolou 1 a dodatkem A označme $L(i,j) = 1 - \delta_{i,j}$ ztrátovou funkci při zařazení objektu do j-té skupiny, patří-li ve skutečnosti do skupiny i-té, $\delta : \mathbb{R}^p \to \{0,1\}$ rozhodovací pravidlo v diskriminační proceduře sestavené na základě hodnot $(\mathbf{X}_1', Y_1)', \ldots, (\mathbf{X}_n', Y_n)'$. Podle dodatku A je pravděpodobnost chyby err definována jako

$$\begin{aligned} err &= E\left\{E\left[L\big(Y, \delta(\mathbf{X})\big)\middle| Y\right]\right\} \\ &= E\left\{L\big(Y, \delta(\mathbf{X})\big)\right\} \end{aligned} \tag{7.1.1}$$

Střední hodnota je přitom počítána vzhledem k distribuční funkci vektoru $(\mathbf{X}', Y)'$, z jehož rozdělení je $(\mathbf{X}_1', Y_1)', \ldots, (\mathbf{X}_n', Y_n)'$ náhodným výběrem. Zřejmým odhadem pro err bude

$$\overline{err} = \frac{1}{n} \sum_{i=1}^{n} L\big(Y_i, \delta(\mathbf{X}_i)\big). \tag{7.1.2}$$

Tento odhad ovšem skutečnou pravděpodobnost chyby podhodnocuje, neboť při jeho výpočtu zařazujeme do skupin objekty pomocí procedury, která byla vytvořena na jejich základě. Označme tuto chybu jako pd, tj.

$$pd = err - \overline{err} \tag{7.1.3}$$

a nazvěme ji *podhodnocení.*

Jiným odhadem pro *err* může být odhad získaný metodou *křížového porovnávání*

$$err^{+} = \frac{1}{n}\sum_{i=1}^{n} L\big(Y_i, \delta_{(i)}(\mathbf{X}_i)\big), \tag{7.1.4}$$

kde $\delta_{(i)}$ je diskriminační procedura vytvořená pomocí učící skupiny $(\mathbf{X}_1', Y_1)', \ldots, (\mathbf{X}_{i-1}', Y_{i-1})'$, $(\mathbf{X}_{i+1}', Y_{i+1})', \ldots, (\mathbf{X}_n', Y_n)'$. Odhad *podhodnocení* potom dostaneme jako

$$\omega^{+} = err^{+} - \overline{err} = \frac{1}{n}\sum_{i=1}^{n} \left\{ L\big(Y_i, \delta_{(i)}(\mathbf{X}_i)\big) - L\big(Y_i, \delta(\mathbf{X}_i)\big) \right\}. \tag{7.1.5}$$

Podle článku [5] je ω^{+} téměř nevychýleným odhadem pro $\omega = Epd$, ale má bohužel příliš velkou variabilitu na to, aby byl dobrým odhadem. Ve zmíněném článku jsou dále porovnávány odhady metodami *bootstrap* a *jackknife*. Odhad získaný metodou *bootstrap* má sice poměrně nízkou variabilitu, ale zase je příliš vychýlený směrem dolů. Nejlepším se jeví odhad ω^{j} získaný metodou *jackknife,* který se spočítá podle vzorce

$$\omega^{j} = \frac{1}{n}\sum_{i=1}^{n} \left\{ L\big(Y_i, \delta_{(i)}(\mathbf{X}_i)\big) - \frac{1}{n}\sum_{j=1}^{n} L\big(Y_i, \delta_{(j)}(\mathbf{X}_i)\big) \right\}. \tag{7.1.6}$$

Odhad pravděpodobnosti chyby spočítaný metodou *jackknife* potom dostaneme jako

$$err^{j} = \omega^{j} + \overline{err}. \tag{7.1.7}$$

VIII. ILUSTRAČNÍ PŘÍKLADY

Závěrem si ukážeme na několika souborech dat praktické použití dříve popsaných modelů. Prvně popisovaný datový soubory je možné nalézt na přiložené disketě.

8.1. Určení pohlaví jedince při archeologickém výzkumu

Úkolem je určit pohlaví jedince, jehož ostatky byly nalezeny na archeologickém nalezišti. Pohlaví je určováno na základě měr některých kostí. K dispozici jsou vykopávky z pohřebišť tří různých kultur (únětické, šňůrové a zvonové). Pro objekty učící skupiny určil pohlaví odborník na archeologii. Pohlaví dalších vykopávek si však chceme určovat sami, bez pomoci tohoto odborníka. Pokusíme se tedy sestavit diskriminační funkci pro účely určení pohlaví. Vzhledem k tomu, že známe zařazení objektů učící skupiny, použijeme k dosažení tohoto cíle modelů logistické regrese a normální diskriminační analýzy.

Použitá data jsou uložena ve formátu pro programy STATISTICA 4.5 a NCSS 6.0 v souborech `kosti.sta`, resp. `kosti.s0` a `kosti.s1`. Jednotlivé proměnné mají následující význam.

Kultura ... slovní označení kultury, z jejíhož pohřebiště pochází daná vykopávka (une, snu, zvo).

Sex_mm ... slovně označené pohlaví dané vykopávky (*femme* = žena, *homme* = muž).

FOS_8 a **UPD_51** ... míry kostí popsané v dalším textu.

Sex_mm_c ... číselně označené pohlaví (0 = žena, 1 = muž).

Kult_une ... alternativní veličina indikující příslušnost jedince k únětické kultuře (tj. má hodnotu 1, pokud *Kultura* = *une* a hodnotu 0 jinak).

Kult_snu ... alternativní veličina indikující příslušnost jedince ke kultuře šňůrové (tj. nabývá hodnoty 1, pokud *Kultura* = *snu* a hodnoty 0 jinak).

Míry kostí, na jejichž základě budeme určovat pohlaví jedince, jsou označeny FOS_8 a UPD_51, uvedeny jsou v milimetrech a mají následující význam: FOS_8 je obvod středu femuru (tj. obvod středu stehenní kosti), UPD_51 označuje délku proximální epifýzy ulny (tj. délku hlavice loketní kosti, která se nachází u loketního kloubu). Obecně se jedná o jednotlivé osteologicky přesně definované rozměry kostry. Sestavíme diskriminační procedury, které buď budou, anebo nebudou využívat znalosti pohřebiště, z kterého daná vykopávka pochází.

Nejprve uvedeme některé popisné statistiky pro veličiny udávající míry kostí, otestujeme zvlášť pro každé pohlaví jednorozměrnou normalitu těchto veličin a shodu va-

riančních matic uvedených veličin u žen a mužů. Vše bylo spočítáno pomocí programu NCSS 6.0, Boxův test potom užitím programu MATLAB.

Muži – popisné statistiky
počet pozorování $n_1 = 54$

veličina	průměr	výběr. rozptyl	minimum	maximum
FOS_8	$89{,}185 = \bar{X}_1^1$	$34{,}616 = S_{v,1}^1$	76,4	99,0
UPD_51	$42{,}456 = \bar{X}_1^2$	$6{,}644 = S_{v,1}^2$	36,4	50,1

Ženy – popisné statistiky
počet pozorování $n_0 = 36$

veličina	průměr	výběr. rozptyl	minimum	maximum
FOS_8	$78{,}864 = \bar{X}_0^1$	$24{,}857 = S_{v,0}^1$	65,8	87,7
UPD_51	$37{,}842 = \bar{X}_0^2$	$5{,}326 = S_{v,0}^2$	33,6	41,9

Přitom

$$S_{v,j}^k = \frac{1}{n_j - 1} \sum_{i:Y_i=j} (X_i^k - \bar{X}_j^k)^2, \qquad j = 0,\ 1, \quad k = 1,\ 2.$$

Nyní otestujeme shodu variančních matic vektoru složeného z uvažovaných veličin FOS_8 a UPD_51 u mužské a ženské populace pomocí Boxova testu, jež je popsán v kapitole 4.3. Při označení z této kapitoly je $K = 2$, $p = 2$. Odhad varianční matice pro mužskou populaci je

$$\mathfrak{S}^1 = \begin{pmatrix} 34{,}616 & 8{,}208 \\ 8{,}208 & 6{,}644 \end{pmatrix},$$

pro ženskou populaci

$$\mathfrak{S}^0 = \begin{pmatrix} 24{,}857 & 6{,}346 \\ 6{,}346 & 5{,}326 \end{pmatrix}.$$

Odhad společné varianční matice pro jedince obou pohlaví vyjde

$$\mathfrak{S} = \begin{pmatrix} 30{,}734 & 7{,}467 \\ 7{,}467 & 6{,}120 \end{pmatrix}.$$

Hodnota konstanty C_p je v našem případě $C_2 = 1{,}026$. Z uvedených údajů již spočítáme hodnotu testové statistiky

$$B = 1{,}708\ .$$

Tato statistika má při platnosti hypotézy shodnosti variančních matic přibližně rozdělení χ^2 o $(K-1)\frac{p(p+1)}{2} = 3$ stupních volnosti. Dosažená hladina testu je potom $p_{level} = 0{,}64$. Shodu variančních matic tedy nemůžeme na 5% hladině významnosti zamítnout.

K testování normality byly použity následující testy: Lillieforsovo zobecnění Kolmogorovova-Smirnovova testu a D'Agostinovy testy využívající výběrové šikmosti, výběrové špičatosti a omnibusu (podrobnější informace o těchto testech lze nalézt např. v [1]). Dosažené hladiny testů jsou uvedeny v následujících tabulkách. Pro Lillieforsův test uvádíme hodnotu testové statistiky lomenou kritickou hodnotou na 5% hladině.

Muži – testy normality

test	FOS_8	UPD_51
Lillieforsův	0,079/0,120	0,071/0,120
D'Agostinova šikmost	0,30	0,41
D'Agostinova špičatost	0,09	0,27
D'Agostinův omnibus	0,14	0,39

Ženy – testy normality

test	FOS_8	UPD_51
Lillieforsův	0,100/0,146	0,079/0,146
D'Agostinova šikmost	0,32	0,84
D'Agostinova špičatost	0,99	0,15
D'Agostinův omnibus	0,61	0,36

Vidíme, že jak pro mužskou, tak pro ženskou populaci není normalita zamítnuta na 5% hladině žádným z uvedených testů ani pro jednu z uvažovaných měr kostí.

8.1.A. Diskriminační procedury, které nezohledňují kulturu, srovnání modelů (NDA) a (LR)

Nejdříve sestavíme diskriminační funkce, které budou určovat pohlaví pouze na základě znalosti veličin FOS_8 a UPD_51 a nebudou využívat informace o kultuře. Vektor $\mathbf{X} = (X^1, X^2)'$, na jehož základě zjišťujeme pohlaví, je tedy dvousložkový, přitom X^1 přísluší veličině FOS_8 a X^2 veličině UPD_51. Na základě výše uvedených testů jednorozměrné normality složek vektoru $\mathbf{X}$ v jednotlivých populacích se můžeme domnívat, že pro obě pohlaví je $\mathbf{X}$ normálně rozděleno. Shoda variančních matic vektoru $\mathbf{X}$ u mužské a ženské populace byla ověřena též. Lze se tedy domnívat, že jsou splněny předpoklady modelu normální diskriminační analýzy. Podle třetí kapitoly jsou splněny také předpoklady modelu logistické regrese.

Nyní můžeme přistoupit k výpočtu odhadů neznámých parametrů. Není-li řečeno jinak, je vše počítáno pomocí programu MATLAB. Odhady parametrů v modelu (NDA) jsou podle (2.2.3) následující:

$$\hat{\lambda} = 0{,}600\ ,$$

$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 89{,}185 \\ 42{,}456 \end{pmatrix}, \qquad \hat{\boldsymbol{\mu}}_0 = \begin{pmatrix} 78{,}864 \\ 37{,}842 \end{pmatrix},$$

$$\hat{\Sigma} = \begin{pmatrix} 30{,}051 & 7{,}302 \\ 7{,}302 & 5{,}984 \end{pmatrix}.$$

Odtud podle (1.2.1) spočítáme odhady koeficientů diskriminační funkce:

$$\hat{\beta}_0 = -38{,}326\ , \qquad \hat{\boldsymbol{\beta}} = \begin{pmatrix} 0{,}222 \\ 0{,}500 \end{pmatrix}.$$

Odhady koeficientů diskriminační funkce v modelu (LR) získáme maximalizací logaritmické věrohodnosti (2.1.2):

$$\tilde{\beta}_0 = -38{,}388\ , \qquad \tilde{\boldsymbol{\beta}} = \begin{pmatrix} 0{,}206 \\ 0{,}537 \end{pmatrix}.$$

Tedy u nové vykopávky, u které naměříme $(X^1, X^2)'$, identifikujeme při použití modelu (NDA) pohlaví jako mužské, pokud

$$0{,}222X^1 + 0{,}500X^2 > 38{,}326$$

a jako ženské jinak. Při použití modelu (LR) je rozhodovací pravidlo nepatrně odlišné a určuje pohlaví jako mužské, pokud

$$0{,}206X^1 + 0{,}537X^2 > 38{,}388$$

a jako ženské jinak.

Podle (5.2.8) byl dále spočítán odhad minimální pravděpodobnosti chyby $\widehat{err}_0$ (do vzorce (5.2.8) byly místo skutečných hodnot parametrů λ, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, Σ dosazeny jejich odhady $\hat{\lambda}$, $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_0$, $\hat{\Sigma}$). Uvádíme i hodnoty některých veličin nutných k výpočtu $\hat{err}_0$.

$$\hat{\Delta} = \sqrt{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)'\hat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)} = 2{,}1445\ ,$$

$$\hat{\omega} = \ln \frac{\hat{\lambda}}{1 - \hat{\lambda}} = 0{,}4055\ ,$$

$$\widehat{err}_0 = 0{,}1376.$$

Odhad pravděpodobnosti chyby jednotlivých diskriminačních procedur byl proveden metodami *jackknife* a *křížového porovnání* podle sedmé kapitoly. Veličiny $\overline{err}$, ω^+, err^+, ω^j a err^j jsou počítány podle vzorců (7.1.2), (7.1.4), (7.1.5) (7.1.6) a (7.1.7).

Odhad pravděpodobnosti chyby

Použitý model	$\overline{err}$	ω^+	err^+	ω^j	err^j
NDA	0,1667	0	0,1667	0	0,1667
LR	0,1667	0	0,1667	0,0011	0,1678

Vidíme, že odhad pravděpodobnosti chyby (metodou jackknife), neboli pravděpodobnosti chybné klasifikace je nepatrně nižší u diskriminační funkce, jež je založena na modelu normální diskriminační analýzy. Tento poznatek je v souladu s teorií uvedenou v předcházejících kapitolách.

Užitím odhadů pravděpodobnosti chyby je možné nyní spočítat odhad relativní vydatnosti $\widehat{Eff}$ logistické regrese vzhledem k normální diskriminační analýze, který je roven

$$\widehat{Eff} = \frac{err^j(NDA) - \widehat{err}_0}{err^j(LR) - \widehat{err}_0} = 0{,}9632.$$

Tuto hodnotu můžeme porovnat s asymptotickou relativní vydatností $Eff_2(\hat{\omega}, \hat{\Delta})$, jež je spočítána podle vzorce (5.2.15) pomocí programu MAPLE. Její hodnotu uvádíme s některými mezivýsledky, nutnými k jejímu vyjádření.

$$q(\hat{\omega}, \hat{\Delta}) = 1{,}0201 \ ,$$
$$Eff_1(\hat{\omega}, \hat{\Delta}) = 0{,}8779 \ , \qquad Eff_\infty(\hat{\omega}, \hat{\Delta}) = 0{,}8630 \ ,$$
$$Eff_2(\hat{\omega}, \hat{\Delta}) = 0{,}8705.$$

Asymptotická relativní vydatnost se liší od relativní vydatnosti spočítané s využitím odhadů pravděpodobnosti chyby metodou jackknife. Rozsah našeho výběru však nebyl příliš velký a tedy není na této skutečnosti nic zvláštního.

V příloze uvádíme graf D.1, na kterém jsou znázorněny objekty učící skupiny spolu s přímkami, jež určují poloroviny, které rozlišují vykopávky mužského pohlaví od vykopávek pohlaví ženského.

8.1.B. Diskriminační procedury, které využívají znalosti kultury, srovnání modelů (NDA) a (LR)

V následujících diskriminačních procedurách budeme využívat též znalosti kultury, ke které patřil jedinec, jehož pohlaví chceme určit. Vektor $\mathbf{X} = (X^1, X^2, X^3, X^4)'$ bude tedy čtyřsložkový, přitom jeho první dvě složky budou spojitého charakteru a budou mít stejný význam jako v předcházejícím případě. Dále $X^3 = 1$, pokud daný jedinec patřil k únětické kultuře a $X^3 = 0$ jinak, $X^4 = 1$, pokud jedinec příslušel ke kultuře šňůrové a $X^3 = 0$ jinak. Příslušnost jedince ke zvonové kultuře tedy indikují nulové hodnoty veličin X^3 a X^4. Všechny dále uváděné výpočty jsou prováděny opět pomocí programu MATLAB, není-li řečeno jinak.

Nyní není žádný důvod se domnívat, že by vektor $\mathbf{X}$ mohl být normálně rozdělen, neboť dvě jeho složky jsou rozděleny diskrétně. Jistě tedy nejsou splněny předpoklady modelu normální diskriminační analýzy. Sestavíme přesto diskriminační procedury založené jak na tomto modelu, tak na modelu logistické regrese. Vhodnost modelu logistické regrese ověříme pomocí Hosmerova-Lemeshowova testu, který je uveden v sedmé kapitole. K tomuto účelu je nejprve nutné spočítat odhady koeficientů β_0 a $\boldsymbol{\beta}$ v modelu logistické regrese, abychom mohli určit odhad logistické pravděpodobnosti $\pi(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$. Tyto odhady, které získáme maximalizací logaritmické věrohodnosti (2.1.2), mají hodnotu:

$$\tilde{\beta}_0 = -46{,}750 \ , \qquad \tilde{\boldsymbol{\beta}} = \begin{pmatrix} 0{,}251 \\ 0{,}710 \\ -2{,}263 \\ -3{,}822 \end{pmatrix} .$$

V dalším kroku sloučíme pozorování podle decilů rizika (jedná se o první metodu slučování popsanou v kapitole 4.2 s hodnotou čísla $g = 10$).

Dostaneme následující odhady počtu mužů (c_i) a počtu žen $(m_i^* - c_i)$ v jednotlivých decilových skupinách, pozorované počty mužů (o_i) a počty žen $(m_i^* - o_i)$ v jednotlivých

skupinách a odhady pravděpodobností, že daná vykopávka byla mužského pohlaví za podmínky, že hodnota vektoru $\mathbf{X}$ na ní naměřená je stejná jako u některého z objektů učící skupiny, který leží v daném decilové skupině ($\overline{\pi}_i$). Vše je samozřejmě zaokrouhleno. Podrobněji se lze o významu těchto veličin dočíst v kapitole 4.2.

Decilové skupiny

Decilová skupina (i)	1	2	3	4	5	6	7	8	9	10	Součet
c_i	0,04	0,39	1,24	3,05	6,01	7,92	8,57	8,82	8,97	8,99	54
o_i	0	1	2	1	6	8	9	9	9	9	54
$m_i^* - c_i$	8,96	8,61	7,76	5,95	2,99	1,08	0,43	0,18	0,04	0,01	36
$m_i^* - o_i$	9	8	7	8	3	1	0	0	0	0	36
$\overline{\pi}_i$	0,005	0,043	0,138	0,339	0,668	0,880	0,952	0,981	0,996	0,999	

Z uvedené tabulky vyplývá, že v deseti políčkách teoretické kontingenční tabulky je odhad teoretické četnosti nižší než pět. Jak však již bylo řečeno ve čtvrté kapitole, autoři [8] uvádějí, že tato skutečnost není příliš na závadu. Testová statistika Hosmerova-Lemeshowova testu $\hat{C}$ má v naší situaci hodnotu

$$\hat{C} = 4{,}374.$$

Rozdělení této statistiky při platnosti hypotézy, že platí model logistické regrese je χ^2 o $10 - 2 = 8$ stupních volnosti. Dosažená hladina testu je tedy rovna $p = 0{,}82$, z čehož plyne, že na 5% hladině významnosti nelze zamítnout platnost modelu logistické regrese.

Nyní přistoupíme k výpočtu odhadů neznámých parametrů. Odhady parametrů v modelu (NDA) jsou podle (2.2.3) následující:

$$\hat{\lambda} = 0{,}600\ ,$$

$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 89{,}185 \\ 42{,}456 \\ 0{,}611 \\ 0{,}204 \end{pmatrix}, \qquad \hat{\boldsymbol{\mu}}_0 = \begin{pmatrix} 78{,}864 \\ 37{,}842 \\ 0{,}500 \\ 0{,}333 \end{pmatrix},$$

$$\hat{\Sigma} = \begin{pmatrix} 30{,}051 & 7{,}302 & 0{,}319 & 0{,}427 \\ 7{,}302 & 5{,}984 & 0{,}182 & 0{,}171 \\ 0{,}319 & 0{,}182 & 0{,}243 & -0{,}141 \\ 0{,}427 & 0{,}171 & -0{,}141 & 0{,}186 \end{pmatrix}.$$

Poznamenejme, že poslední dvě složky odhadů $\boldsymbol{\mu}_1$ a $\boldsymbol{\mu}_0$ udávají poměrné zastoupení jedinců únětické a šňůrové kultury mezi muži a ženami v učícím souboru.

Podle (1.2.1) spočítáme odhady koeficientů diskriminační funkce v modelu (NDA):

$$\hat{\beta}_0 = -45{,}401\ , \qquad \hat{\boldsymbol{\beta}} = \begin{pmatrix} 0{,}275 \\ 0{,}629 \\ -2{,}664 \\ -3{,}925 \end{pmatrix}.$$

Odhady koeficientů diskriminační funkce v modelu (LR), které jsme získali maximalizací logaritmické věrohodnosti (2.1.2) jsme uvedli již dříve.

Pro model normální diskriminační analýzy tedy dostáváme následující rozhodovací pravidlo (vždy je uvedeno, kdy zařadíme jedince, u kterého bylo naměřeno X^1 a X^2 mezi muže). Absolutní člen v rozhodovací funkci je nyní vypočítán zvlášť pro jednotlivé kultury.

únětická kultura: $0{,}275X^1 + 0{,}629X^2 > 48{,}065$,
šňůrová kultura: $0{,}275X^1 + 0{,}629X^2 > 49{,}326$,
zvonová kultura: $0{,}275X^1 + 0{,}629X^2 > 45{,}401$.

Analogická rozhodovací pravidla mají v modelu logistické regrese tvar:

únětická kultura: $0{,}251X^1 + 0{,}710X^2 > 49{,}013$,
šňůrová kultura: $0{,}251X^1 + 0{,}710X^2 > 50{,}572$,
zvonová kultura: $0{,}251X^1 + 0{,}710X^2 > 46{,}750$.

Odhad pravděpodobnosti chyby jednotlivých diskriminačních procedur byl stejně jako v situaci, kdy jsme do modelů nezahrnovali informaci o kultuře proveden metodami *jackknife* a *křížového porovnání* podle sedmé kapitoly. Veličiny $\overline{err}$, ω^+, err^+, ω^j a err^j jsou opět počítány podle vzorců (7.1.2), (7.1.4), (7.1.5), (7.1.6) a (7.1.7).

Odhad pravděpodobnosti chyby

Použitý model	$\overline{err}$	ω^+	err^+	ω^j	err^j
NDA	0,0889	0,0222	0,1111	0,0210	0,1099
LR	0,1667	0	0,0889	0,0110	0,0999

Vidíme, že odhad pravděpodobnosti chyby, neboli pravděpodobnosti chybné klasifikace, založený na obou metodách, je nyní nepatrně nižší u diskriminační funkce, jež je sestavena na základě modelu logistické regrese.

V příloze jsou uvedeny grafy D.2 až D.4, na kterých jsou znázorněny objekty učící skupiny spolu s přímkami, jež určují poloroviny, které rozlišují vykopávky mužského pohlaví od vykopávek pohlaví ženského. Grafy jsou nyní tři, každý pro jedince jedné kultury.

8.1.C. Diskriminační procedury, které nezohledňují kulturu, srovnání modelů (NDA) a (MND)

Na závěr srovnáme model normální diskriminační analýzy s modelem směsi normálních rozdělení. Diskriminaci přitom založíme na vektoru $\mathbf{X}$, který neobsahuje informaci o kultuře. Podle předcházejícího textu jsou předpoklady obou uvažovaných modelů splněny.

Odhady neznámých parametrů v modelu (MND) získáme pomocí EM algoritmu (viz kapitola 6) využitím programu MATLAB, konkrétně funkce `MNDest` s argumenty: $tolerance = 10^{-5}$, $rovnat = 1$ a $\lambda^0 = 5/9$. Dostaneme následující výsledky:

$$\breve{\lambda} = 0{,}490\ ,$$

$$\breve{\boldsymbol{\mu}}_1 = \begin{pmatrix} 90{,}944 \\ 42{,}798 \end{pmatrix}\ , \qquad \breve{\boldsymbol{\mu}}_0 = \begin{pmatrix} 79{,}406 \\ 38{,}510 \end{pmatrix}\ ,$$

$$\breve{\Sigma} = \begin{pmatrix} 22{,}350 & 6{,}368 \\ 6{,}368 & 6{,}499 \end{pmatrix}.$$

Odtud podle (1.2.1) spočítáme odhady koeficientů v diskriminační funkci:

$$\breve{\beta}_0 = -47{,}512\ , \qquad \breve{\boldsymbol{\beta}} = \begin{pmatrix} 0{,}455 \\ 0{,}213 \end{pmatrix}.$$

Tedy u nově nalezené kostry, u které naměříme $(X^1, X^2)'$, identifikujeme pohlaví jako mužské při použití modelu (MND), pokud

$$0{,}455X^1 + 0{,}213X^2 > 47{,}512\ .$$

Pomocí modelu (NDA) identifikujeme muže, jestliže

$$0{,}222X^1 + 0{,}500X^2 > 38{,}326\ ,$$

jak bylo ukázáno v části 8.1.A.

Jelikož známe pohlaví objektů učící skupiny, můžeme jak metodou *jackknife,* tak metodou *křížového porovnání* určit odhady pravděpodobnosti chyby též pro model (MND). Jejich hodnoty jsou uvedeny v následující tabulce.

Odhad pravděpodobnosti chyby

Použitý model	$\overline{err}$	ω^+	err^+	ω^j	err^j
NDA	0,1667	0	0,1667	0	0,1667
MND	0,1667	0	0,1667	-0,0025	0,1642

Vidíme, že odhad pravděpodobnosti špatné klasifikace je metodou *křížového porovnání* stejný pro oba modely a metodou *jackknife* je nepatrně nižší u modelu (MND). Toto zjištění neodpovídá plně závěrům uvedeným na předcházejících stranách, ale musíme si uvědomit, že se jedná pouze o odhady pravděpodobnosti chyby, jejichž přesnost může být ovlivněna mnoha skutečnostmi, např. ne příliš vysokým rozsahem učícího souboru.

Také pro tyto dva modely uvádíme v příloze graf D.5 se znázorněním objektů učící skupiny spolu s diskriminačními přímkami.

8.2. Přijímací zkoušky na Právnické fakultě UK v Praze

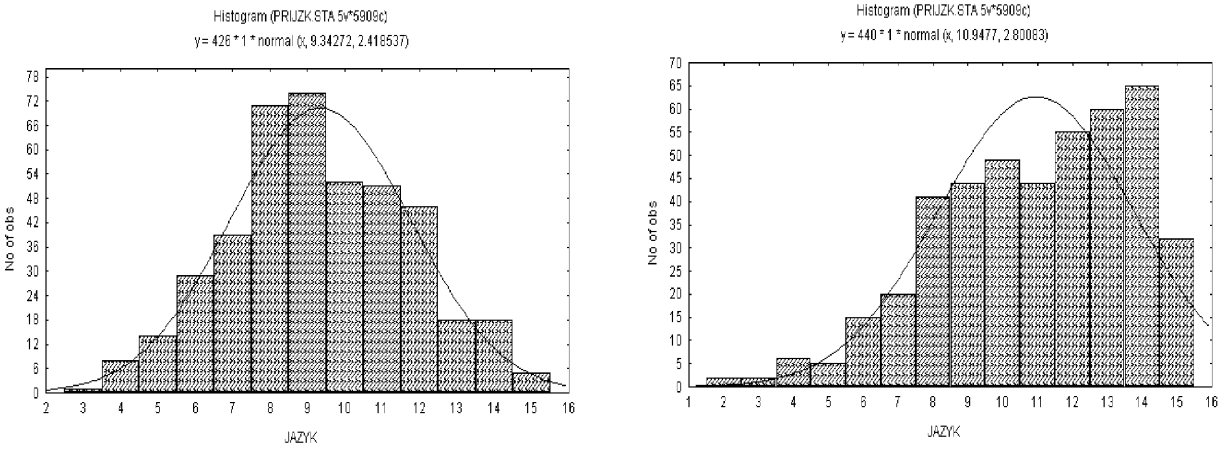
V následujícím příkladě se pokusíme zanalyzovat výsledky přijímacích zkoušek na Právnické fakultě UK v Praze v roce 1999. Tyto přijímací zkoušky jsou nechvalně známy možností, že někteří uchazeči o studium na zmíněné fakultě znali znění přijímacích testů před vlastní přijímací zkouškou. Pomocí studovaných modelů se pokusíme rozlišit studenty, kteří neznali zadání přijímacích testů (běžní studenti), a studenty, kteří mohli znát předem znění těchto testů (zvýhodnění studenti).

K dispozici jsou výsledky jednotlivých uchazečů v následující podobě: počet bodů za test z cizího jazyka (proměnná *jazyk*), z historie a všeobecného přehledu (proměnná

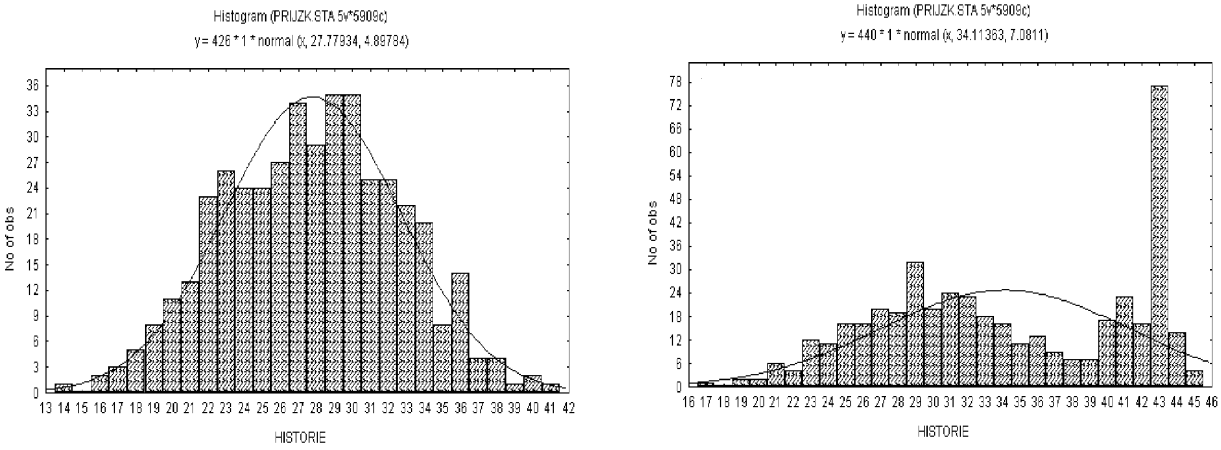
historie) a za test z logiky (proměnná *logika*). Dále je u každého uchazeče uvedeno pořadové číslo termínu zkoušky, kterého se zúčastnil. Termínů bylo dohromady třináct, přitom ten třináctý byl náhradní za termín číslo dvanáct, který byl anulován kvůli podezření na podvodné jednání některých uchazečů. V analýze nebudeme tedy pracovat s daty z třináctého termínu, neboť se ho zúčastnili studenti, kteří již přijímací zkoušku absolvovali v termínu dvanáctém. Přidání dat ze třináctého termínu do celého souboru by mohlo způsobit porušení nezávislosti jednotlivých pozorování. Každého z prvních dvanácti termínů se zúčastnil přibližně stejný počet uchazečů v rozmezí od 426 do 488. Za test z jazyka bylo přitom možné získat maximálně patnáct bodů, za test z historie a všeobecného přehledu maximálně čtyřicet pět bodů a za test z logiky maximálně čtyřicet bodů.

Veličina Y, jež indikuje zařazení jednotlivých uchazečů, bude nabývat hodnoty jedna pro zvýhodněné a hodnoty nula pro běžné uchazeče. Diskriminaci budeme provádět na základě vektoru $\mathbf{X}$, jehož složky budou odpovídat po řadě proměnným *jazyk, historie, logika*. Učící skupinu v tomto případě tvoří všichni uchazeči, kteří se zúčastnili jednoho z prvních dvanácti termínů. U žádného z nich nevíme, zda ho zařadit mezi běžné nebo zvýhodněné studenty. K sestavení diskriminační funkce tedy musíme nyní použít model směsi normálních rozdělení. Pro podpoření domněnky, že zkoumaná data jsou skutečně směsí dvou normálních rozdělení, uvedeme histogramy dosažených bodů u jednotlivých testů zvlášť pro první a dvanáctý termín. Výsledky uchazečů z prvního termínu by směs tvořit neměly, naopak výsledky dvanáctého termínu by měly tvořit směs z rozdělení, z něhož pocházejí data u otatních termínů a rozdělení, z něhož pocházejí data zvýhodněných uchazečů. Histogramy pro druhý až jedenáctý termín se od toho pro termín číslo jedna příliš neliší a proto nejsou uvedeny. Vlevo je vždy graf odpovídající prvnímu a vpravo graf odpovídající dvanáctému termínu. Na závěr je ještě uveden histogram celkového součtu bodů ze všech tří testů.

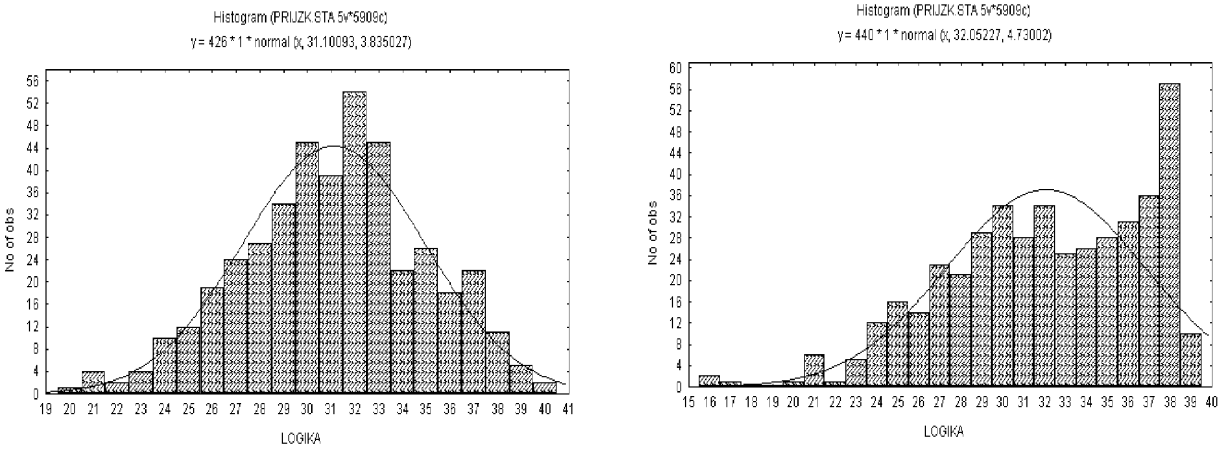
Obr. 8.2.1. Histogramy – test z cizího jazyka



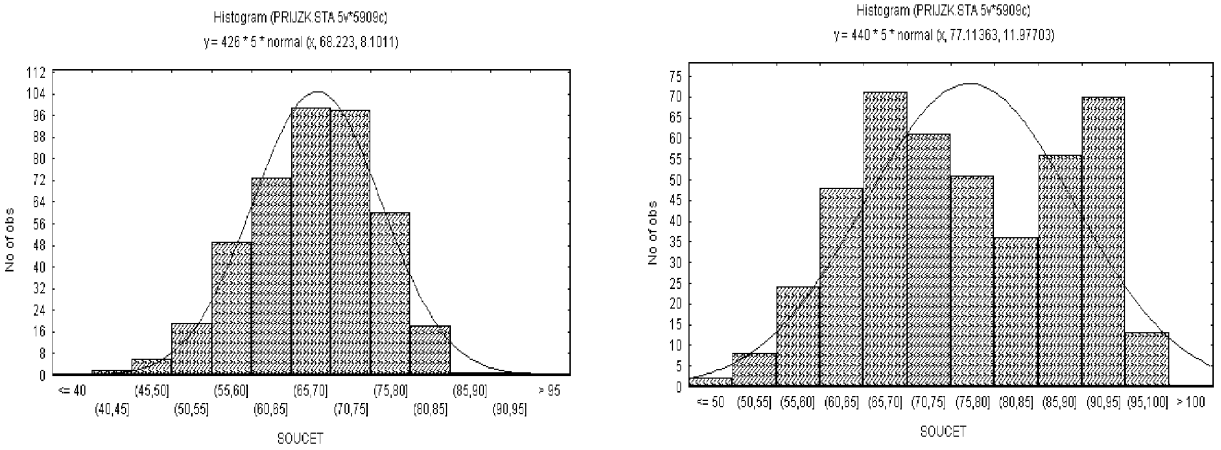
Obr. 8.2.2. Histogramy – test z historie a všeobecného přehledu



Obr. 8.2.3. Histogramy – test z logiky



Obr. 8.2.4. Histogramy – bodový součet



Směs dvou rozdělení lze odhalit v podstatě na všech histogramech odpovídajících dvanáctému termínu, přitom nejvíce se promíchání dat ze dvou výběrů projevuje u testu z historie a všeobecného přehledu. Naproti tomu histogramy prvního termínu poměrně dobře odpovídají hustotě normálního rozdělení. Na závěr ještě uvádíme tabulku s průměry výsledků jednotlivých testů a celkového bodového součtu zvlášť pro prvních jedenáct termínů a pro termín dvanáctý.

Průměry získaných bodů

	1. - 11. termín (5110 studentů)	12. termín (440 studentů)
jazyk	10,11	10,95
historie	27,51	34,11
logika	28,64	32,05
bodový součet	66,27	77,11

Z tabulky vidíme, že průměry dosažených bodů jsou u dvanáctého termínu vždy vyšší. Přitom rozdíl je téměř zanedbatelný pro jazyk a nejvyšší pro historii. Také tato zjištění nás utvrzují v domněnce, že máme co do činění se směsí dvou rozdělení.

Podrobněji se lze s důvody, jež vedou k domněnce, že data jsou směsí dvou rozdělení, seznámit na síti Internet na adrese `http://www.prf.cuni.cz/pr99/Komentar.html`, kde je zveřejněn *Komentář ke statistickému zpracování výsledků přijímacích zkoušek na Právnické fakultě UK v Praze v roce 1999.*

Pro vlastní sestavení diskriminační funkce použijeme výsledky všech uchazečů, kteří se zúčastnili prvních dvanácti termínů. Takto získáme výběr ze směsi dvou rozdělení, přičemž nyní již promíchanost nevynikne tolik, jako v případě dvanáctého termínu. Prvních dvanácti termínů se zúčastnilo 5550 uchazečů. Odhady budeme počítat pomocí funkce `MNDest` v MATLABU. Počáteční odhad λ^0 poměru zvýhodněných uchazečů zvolíme následujícím způsobem. U dvanáctého termínu se směs projevuje nejvíce v testu z historie a všeobecného přehledu. Zběžným pohledem na histogram bodového zisku z historie v tomto termínu (obr. 3.2.2 vpravo) se zdá, že hranice mezi běžnými a zvýhodněnými uchazeči by mohla být blízká hodnotě 42. Tohoto anebo vyššího počtu bodů dosáhlo v rámci dvanáctého termínu 111 uchazečů. Za λ^0 zvolíme tedy hodnotu $111/5550 = 0{,}02$. Předpokládáme totiž, že zbylých termínů se zúčastnili pouze běžní uchazeči. Data dále necháme setřídit podle proměnné odpovídající výsledku testu z historie. Toleranci nastavíme na hodnotu 10^{-5}. Po provedení 56 iterací EM algoritmu získáme následující odhady:

$$\breve{\lambda} = 0{,}062\ ,$$

$$\breve{\boldsymbol{\mu}}_1 = \begin{pmatrix} 11{,}57 \\ 38{,}84 \\ 33{,}61 \end{pmatrix}\ , \qquad \breve{\boldsymbol{\mu}}_0 = \begin{pmatrix} 10{,}09 \\ 27{,}32 \\ 28{,}60 \end{pmatrix}\ ,$$

$$\breve{\Sigma} = \begin{pmatrix} 7{,}07 & 2{,}48 & 2{,}31 \\ 2{,}48 & 20{,}84 & 3{,}61 \\ 2{,}31 & 3{,}61 & 15{,}70 \end{pmatrix}.$$

Vidíme, že odhad střední hodnoty bodových zisků běžných uchazečů je téměř shodný

s průměry bodových zisků studentů, kteří se zúčastnili prvních jedenácti termínů. Odhad střední hodnoty bodových zisků zvýhodněných uchazečů je o něco vyšší než průměr bodových zisků dosažených v rámci dvanáctého termínu. Tento fakt je způsoben skutečností, že dvanáctého termínu se zúčastnili též běžní studenti. Vzhledem k uvedenému se potvrzuje domněnka, že prvních jedenácti termínů se patrně nezúčastnil žádný zvýhodněný student.

Z uvedených odhadů pomocí funkce `MNDbetas` spočítáme odhady koeficientů v diskriminační funkci:

$$\breve{\beta}_0 = -25{,}92\ , \qquad \breve{\boldsymbol{\beta}} = \begin{pmatrix} -0{,}04 \\ 0{,}52 \\ 0{,}20 \end{pmatrix}.$$

Tedy uchazeče, který u přijímací zkoušky dosáhnul bodového zisku reprezentovaného vektorem $\mathbf{X} = (jazyk, historie, logika)'$, zařadíme mezi zvýhodněné, pokud

$$-0{,}04 \cdot jazyk + 0{,}52 \cdot historie + 0{,}20 \cdot logika > 25{,}92.$$

Pokud aplikujeme toto rozhodovací pravidlo na výsledky uvažovaných uchazečů, získáme následující odhady počtu běžných a zvýhodněných uchazečů na jednotlivých termínech přijímací zkoušky. Odhady počtu běžných uchazečů jsou ve sloupci označeném nulou, počtu zvýhodněných uchazečů ve sloupci označeném jedničkou.

Odhady počtu běžných a zvýhodněných uchazečů

	zařazení			poměr
termín	0	1	součet	zvýhodněných (%)
1.	420	6	426	1,4
2.	451	0	451	0
3.	468	13	481	2,7
4.	443	7	450	1,6
5.	467	2	469	0,4
6.	464	5	469	1,1
7.	458	7	465	1,5
8.	470	3	473	0,6
9.	460	3	463	0,6
10.	468	7	475	1,5
11.	478	10	488	2,0
12.	279	161	440	36,6
součet	5326	224	5550	4,0

Samozřejmě, že ne každý uchazeč, který je podle našeho diskriminačního pravidla označen za zvýhodněného, jím skutečně je. Diskriminační funkce musí totiž pomocí roviny rozdělit jednoznačně trojrozměrný eukleidovský prostor na dvě části. Takto se do části se zvýhodněnými uchazeči může dostat i ten, který přirozeným způsobem (vlastními vědomostmi) dosáhnul vyššího bodového zisku. Proto se mezi „zvýhodněnými“ uchazeči objevují též studenti, kteří se zúčastnili jednoho z prvních jedenácti termínů, nikdy jich však není mnoho (maximálně 2,7 %). Naproti tomu v případě dvanáctého termínu bylo za zvýhodněné označeno 161 studentů, tj. 36,6 %, což podporuje domněnku,

že někteří uchazeči, kteří se zúčastnili tohoto termínu přijímacích zkoušek, znali zadání testů předem.

Pro srovnání ještě spočítáme odhady neznámých parametrů pouze s využitím dat z kritického dvanáctého termínu. Za počáteční odhad parametru λ nyní zvolíme $\lambda^0 = 111/440$ a při toleranci 10^{-5} dostaneme po 16 iteracích EM algoritmu následující odhady neznámých parametrů:

$$\breve{\lambda}^{12} = 0{,}431 \ ,$$

$$\breve{\boldsymbol{\mu}}_1^{12} = \begin{pmatrix} 12{,}09 \\ 41{,}19 \\ 35{,}18 \end{pmatrix} \ , \qquad \breve{\boldsymbol{\mu}}_0^{12} = \begin{pmatrix} 10{,}08 \\ 28{,}75 \\ 29{,}68 \end{pmatrix} \ ,$$

$$\breve{\Sigma}^{12} = \begin{pmatrix} 6{,}83 & 1{,}47 & 1{,}64 \\ 1{,}47 & 12{,}12 & 2{,}74 \\ 1{,}64 & 2{,}74 & 14{,}92 \end{pmatrix} .$$

Odhady $\breve{\boldsymbol{\mu}}_1^{12}$, $\breve{\boldsymbol{\mu}}_0^{12}$ a $\breve{\Sigma}^{12}$ jsou poměrně blízké odhadům $\breve{\boldsymbol{\mu}}_1$, $\breve{\boldsymbol{\mu}}_0$, Σ. Odhad $\breve{\lambda}^{12}$ s odhadem $\breve{\lambda}$ srovnávat nemůžeme, neboť se vztahuje k podílu zvýhodněných uchazečů v rámci dvanáctého termínu, který byl podstatně vyšší než v rámci celého přijímacího řízení. Odhady koeficientů v diskriminační funkci jsou následující:

$$\breve{\beta}_0^{12} = -40{,}94 \ , \qquad \breve{\boldsymbol{\beta}}^{12} = \begin{pmatrix} 0{,}04 \\ 0{,}98 \\ 0{,}18 \end{pmatrix} .$$

Pokud pomocí této diskriminační procedury zařadíme uchazeče, kteří se zúčastnili dvanáctého termínu, bude jich 185 označeno za zvýhodněné, což je o 24 více, než při diskriminaci prováděné pomocí původní procedury. Přitom žádný z uchazečů, který byl původní procedurou označen za zvýhodněného, nebude nyní nezvýhodněný. Nová procedura tedy pouze k původním zvýhodněným studentům přidala dalších 24 uchazečů. Tato skutečnost může být způsobena faktem, že nyní byl podíl zvýhodněných uchazečů v učícím souboru podstatně vyšší, než při sestavování původní procedury. Zařazovat uchazeče z ostatních termínů pomocí procedury určené koeficienty ${\beta_0}^{12}$ a $\boldsymbol{\beta}^{12}$ nebude mít příliš velký smysl kvůli chybnému odhadu podílu zvýhodněných uchazečů v souboru všech studentů, kteří se zúčastnili přijímacích zkoušek. Upravíme-li tento odhad do tvaru

$$\breve{\lambda}^{12,v\check{s}e} = \frac{\breve{\lambda}^{12} \cdot \text{počet uchazečů v 12. termínu}}{\text{počet všech uchazečů}} = \frac{0{,}431 \cdot 440}{5550} = 0{,}034$$

a spočítáme pomocí $\breve{\boldsymbol{\mu}}_1^{12}$, $\breve{\boldsymbol{\mu}}_0^{12}$, $\breve{\Sigma}^{12}$ a $\breve{\lambda}^{12,v\check{s}e}$ koeficienty ${\breve{\beta}_0}^{12,v\check{s}e}$, $\breve{\boldsymbol{\beta}}^{12,v\check{s}e}$, jež vyjdou

$$\breve{\beta}_0^{12,v\check{s}e} = -44{,}00 \ , \qquad \breve{\boldsymbol{\beta}}^{12,v\check{s}e} = \begin{pmatrix} 0{,}04 \\ 0{,}98 \\ 0{,}18 \end{pmatrix} ,$$

získáme diskriminační proceduru, pomocí níž již můžeme zařazovat též studenty z ostatních termínů. Tato procedura označí studenta za zvýhodněného, pokud

$$0{,}04 \cdot jazyk + 0{,}98 \cdot historie + 0{,}18 \cdot logika > 44{,}00.$$

Toto rozhodovací pravidlo se na první pohled poměrně liší od původního pravidla založeného na $\breve{\beta}_0$, $\breve{\boldsymbol{\beta}}$, ale pokud porovnáme rozhodnutí učiněná na základě těchto dvou procedur, zjistíme, že odlišnost není příliš velká, jak je možné se přesvědčit v následující tabulce, která obě procedury porovnává. Ve sloupci označeném 0 – 1 je počet uchazečů označených novou procedurou za zvýhodněné, ale starou za běžné, sloupec označený 1 – 0 obsahuje naopak počet uchazečů označených za zvýhodněné pouze původní procedurou. Sloupce původní a nová procedura přinášejí počty uchazečů, kteří byli označeni za zvýhodněné užitím příslušné diskriminační funkce.

Porovnání dvou procedur

termín	původní procedura	nová procedura	0 – 1	1 – 0	počet odlišně zařazených
1.	6	5	0	1	1
2.	0	0	0	0	0
3.	13	11	1	3	4
4.	7	7	1	1	2
5.	2	2	0	0	0
6.	5	4	0	1	1
7.	7	9	2	0	2
8.	3	3	0	0	0
9.	3	3	0	0	0
10.	7	8	1	0	1
11.	10	12	3	1	4
12.	161	160	1	2	3
součet	224	224	9	9	18

A. STATISTICKÉ ROZHODOVACÍ FUNKCE

V této kapitole se budeme zabývat problematikou statistických rozhodovacích funkcí. Nechť $\theta \in \Omega$ je neznámý parametr, jehož hodnotu chceme zjistit, $\Omega \subset \mathbb{R}^k$ je parametrický prostor. Nechť $\mathbf{X} \in \mathfrak{X}$ je náhodná veličina s hustotou $r(\mathbf{x})$ vzhledem k σ–konečné míře $\nu(\mathbf{x})$. Nechť $\delta : \mathfrak{X} \to \Omega$ je *rozhodovací funkce.* Dále nechť $L\big(\theta, \delta(\mathbf{X})\big)$ je *ztrátová funkce* ohodnocující naše rozhodnutí, je-li skutečná hodnota parametru rovna θ. Naším úkolem je najít takovou rozhodovací funkci δ z předem dané množiny $\mathfrak{D}$, aby hodnota ztrátové funkce byla v jistém smyslu minimální.

V ideálním případě bude optimální rozhodovací pravidlo δ^* splňovat následující vztah:

$$\delta^* = \operatorname*{argmin}_{\delta \in \mathfrak{D}} L\big(\theta, \delta(\mathbf{X})\big) \qquad \forall \theta \in \Omega, \forall \mathbf{X} \in \mathfrak{X}.$$

Pravidlo δ^* s touto vlastností však nalezneme pouze v nemnoha případech. Pro nalezení optimálního rozhodovacího pravidla v nějakém smyslu zavedeme *rizikovou funkci* $R(\theta, \delta) = E\big[L\big(\theta, \delta(\mathbf{X})\big)\big|\theta\big]$.

Definice. Řekneme, že rozhodovací pravidlo δ' je R–lepší než rozhodovací pravidlo δ, pokud platí

$$R(\theta, \delta') \leq R(\theta, \delta) \quad \forall \theta \in \Omega$$

a alespoň pro jedno $\theta \in \Omega$ je nerovnost ostrá.

Popíšeme si dvě možnosti nalezení optimální rozhodovací funkce.

a) minimax

$$R(\theta, \delta^*) = \min_{\delta \in \mathfrak{D}} \max_{\theta \in \Omega} R(\theta, \delta)$$

b) bayesovský přístup

Předpokládejme, že θ má apriorní hustotu q vzhledem k σ–konečné míře $\eta(\theta)$ a že $\mathbf{X}$ má podmíněnou hustotu $r(\mathbf{x}|\theta)$ vzhledem k σ–konečné míře $\nu(\mathbf{x})$. Nechť $\rho_q(\delta) = ER(\theta, \delta) = \int_\Omega R(\theta, \delta) q(\theta) d\eta(\theta)$ je *bayesovská riziková funkce.*

$$ER(\theta, \delta^*) = \rho_q(\delta^*) = \min_{\delta \in \mathfrak{D}} \rho_q(\delta),$$

tedy

$$\delta^* = \operatorname*{argmin}_{\delta \in \mathfrak{D}} \rho_q(\delta).$$

V tomto případě se budeme tvaru optimální rozhodovací funkce věnovat podrobněji.

(1) Užitím Bayesovy věty dostaneme podmíněnou hustotu θ za podmínky $\mathbf{X} = \mathbf{x}$

$$\pi(\theta|\mathbf{x}) = \begin{cases} \frac{r(\mathbf{x}|\theta)\ q(\theta)}{\int_\Omega r(\mathbf{x}|\theta)\ q(\theta)\ d\eta(\theta)}, & \text{je–li } \mathbf{x} \in \mathfrak{X}^+, \\ 0 & \text{jinak,} \end{cases}$$

kde $\mathfrak{X}^+ = \{\mathbf{x} : \int_\Omega r(\mathbf{x}|\theta)\ q(\theta)\ d\eta(\theta) \neq 0\}$.

(2) Nepodmíněnou hustotu $\mathbf{X}$ dostaneme integrací sdružené hustoty $r(\mathbf{x}, \theta)$ přes Ω. Přitom

$$r(\mathbf{x}, \theta) = r(\mathbf{x}|\theta)\ q(\theta) \quad (\nu \otimes \eta)\text{-skoro všude}$$

a tedy

$$r(\mathbf{x}) = \int_\Omega r(\mathbf{x}|\theta)\ q(\theta)\ d\eta(\theta) \quad \nu\text{-skoro všude.}$$

Potom

$$\begin{aligned}
\rho_q(\delta) &= ER(\theta, \delta) \\
&= \int_\Omega R(\theta, \delta) q(\theta)\ d\eta(\theta) \\
&= \int_\Omega \Big[\int_{\mathfrak{X}} L\big(\theta, \delta(\mathbf{x})\big) r(\mathbf{x}|\theta)\ d\nu(\mathbf{x})\Big] q(\theta)\ d\eta(\theta) \\
&= \int_{\mathfrak{X}} \int_\Omega L\big(\theta, \delta(\mathbf{x})\big) r(\mathbf{x}|\theta)\ q(\theta)\ d\eta(\theta) d\nu(\mathbf{x}) \\
&= \int_{\mathfrak{X}^+} \Big[\int_\Omega L\big(\theta, \delta(\mathbf{x})\big) \frac{r(\mathbf{x}|\theta)\ q(\theta)}{\int_\Omega r(\mathbf{x}|\theta)\ q(\theta)\ d\eta(\theta)} d\eta(\theta)\Big] \cdot \Big[\int_\Omega r(\mathbf{x}|\theta)\ q(\theta)\ d\eta(\theta)\Big] d\nu(\mathbf{x}) \\
&= \int_{\mathfrak{X}} \Big[\int_\Omega L\big(\theta, \delta(\mathbf{x})\big) \pi(\theta|\mathbf{x})\ d\eta(\theta)\Big] \cdot \Big[\int_\Omega r(\mathbf{x}|\theta)\ q(\theta)\ d\eta(\theta)\Big] d\nu(\mathbf{x}) \\
&= \int_{\mathfrak{X}} E\big[L\big(\theta, \delta(\mathbf{X})\big)\big|\mathbf{X} = \mathbf{x}\big] r(\mathbf{x})\ d\nu(\mathbf{x}).
\end{aligned}$$

Nechť $\hat{\delta}(\mathbf{x}) = \operatorname{argmin}_{\delta \in \mathfrak{D}} E\big[L\big(\theta, \delta(\mathbf{X})\big)\big|\mathbf{X} = \mathbf{x}\big]$ a nechť toto minimum existuje pro $\nu(\mathbf{x})$–skoro všechna $\mathbf{x} \in \mathfrak{X}$. Potom

$$\begin{aligned}
\min_{\delta \in \mathfrak{D}} \rho_q(\delta) &= \min_{\delta \in \mathfrak{D}} \int_{\mathfrak{X}} E\big[L\big(\theta, \delta(\mathbf{X})\big)\big|\mathbf{X} = \mathbf{x}\big] r(\mathbf{x})\ d\nu(\mathbf{x}) \\
&\geq \int_{\mathfrak{X}} \min_{\delta \in \mathfrak{D}} E\big[L\big(\theta, \delta(\mathbf{X})\big)\big|\mathbf{X} = \mathbf{x}\big] r(\mathbf{x})\ d\nu(\mathbf{x}) \\
&= \int_{\mathfrak{X}} E\big[L\big(\theta, \hat{\delta}(\mathbf{X})\big)\big|\mathbf{X} = \mathbf{x}\big] r(\mathbf{x})\ d\nu(\mathbf{x}) \\
&= \rho_q(\hat{\delta}).
\end{aligned}$$

Tedy $\rho_q(\delta) \geq \rho_q(\hat{\delta}) \qquad \forall \delta \in \mathfrak{D}$ a tudíž $\delta^* = \hat{\delta}$.

Na závěr si uveďme explicitní tvar rizikové a bayesovské rizikové funkce pro případ, že

$$\begin{aligned}\Omega &= \{0,1\},\\ \mathfrak{D} &= \big\{\delta, \delta : \mathfrak{X} \to \{0,1\}\big\},\\ L(i,j) &= 1 - \delta_{i,j}, \quad i,j = 0,1.\end{aligned}$$

V tomto případě hraje roli míry η čítací míra. Pro rizikovou funkci $R(\theta,\delta)$ postupně dostáváme:

$$\begin{aligned}R(i,\delta) &= E\big[L\big(\theta,\delta(\mathbf{X})\big)\big|\theta = i\big]\\ &= \int_{\{\mathbf{x}:\delta(\mathbf{x})=1\}} L(i,1) r(\mathbf{x}|i) d\nu(\mathbf{x}) + \int_{\{\mathbf{x}:\delta(\mathbf{x})=0\}} L(i,0) r(\mathbf{x}|i) d\nu(\mathbf{x})\\ &= \int_{\{\mathbf{x}:\delta(\mathbf{x})=1\}} (1-\delta_{i,1}) r(\mathbf{x}|i) d\nu(\mathbf{x}) + \int_{\{\mathbf{x}:\delta(\mathbf{x})=0\}} (1-\delta_{i,0}) r(\mathbf{x}|i) d\nu(\mathbf{x})\\ &= \begin{cases} \int_{\{\mathbf{x}:\delta(\mathbf{x})=1\}} r(\mathbf{x}|i) d\nu(\mathbf{x}) = P\big(\delta(\mathbf{X}) = 1\big|\theta = 0\big), & i = 0,\\ \int_{\{\mathbf{x}:\delta(\mathbf{x})=0\}} r(\mathbf{x}|i) d\nu(\mathbf{x}) = P\big(\delta(\mathbf{X}) = 0\big|\theta = 1\big), & i = 1.\end{cases}\end{aligned}$$

Z vyjádření rizikové funkce již snadno dostáváme bayesovskou rizikovou funkci:

$$\begin{aligned}\rho_q(\delta) &= ER(\theta,\delta)\\ &= P(\theta=0)R(0,\delta) + P(\theta=1)R(1,\delta)\\ &= P(\theta=0)P\big(\delta(\mathbf{X}) = 1\big|\theta=0\big) + P(\theta=1)P\big(\delta(\mathbf{X})=0\big|\theta=1\big)\\ &= P\big(\delta(\mathbf{X}) = 1, \theta = 0\big) + P\big(\delta(\mathbf{X}) = 0, \theta = 1\big).\end{aligned}$$

Vidíme tedy, že bayesovskou rizikovou funkci lze v této situaci interpretovat jako pravděpodobnost špatného rozhodnutí o hodnotě parametru θ.

B. STANDARDNÍ SITUACE V MODELU NORMÁLNÍ DISKRIMINAČNÍ ANALÝZY

Cílem tohoto dodatku je dokázat platnost následující věty.

Věta B.1. *Nechť* $\boldsymbol{\mu}_1, \boldsymbol{\mu}_0 \in \mathbb{R}^p$, $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_0$. *Nechť* $\lambda \in (0,1)$. *Nechť* Σ *je pozitivně definitní matice typu* $p \times p$. *Potom lze užitím lineární transformace* $\mathbf{X}^\star = \mathbf{a} + \mathbb{A}\mathbf{X}$, $\mathbf{a} \in \mathbb{R}^p$, $\mathbb{A}$ *matice typu* $p \times p$, *převést případ*

$$\mathfrak{L}(\mathbf{X}) = N_p(\boldsymbol{\mu}_1, \Sigma) \text{ s pravděpodobností } \lambda,$$
$$\mathfrak{L}(\mathbf{X}) = N_p(\boldsymbol{\mu}_0, \Sigma) \text{ s pravděpodobností } 1-\lambda$$

na situaci

$$\mathfrak{L}(\mathbf{X}^\star) = N_p(\frac{\Delta}{2}\mathbf{e}_1, I_p) \text{ s pravděpodobností } \lambda,$$
$$\mathfrak{L}(\mathbf{X}^\star) = N_p(-\frac{\Delta}{2}\mathbf{e}_1, I_p) \text{ s pravděpodobností } 1-\lambda,$$

kde

$$\mathbf{e}_1 = (1, 0, \ldots, 0)' \in \mathbb{R}^p,$$

$$I_p \text{ je jednotková matice typu } p \times p,$$

$$\Delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}$$

je zobecněná Mahalanobisova vzdálenost vektorů $\boldsymbol{\mu}_1$ *a* $\boldsymbol{\mu}_0$.

Před provedením vlastního důkazu věty B.1 uvedeme několik pomocných tvrzení.

Věta B.2. *Nechť* $\mathbb{M}$, $\mathbb{N}$ *jsou reálné symetrické matice typu* $p \times p$. *Nechť matice* $\mathbb{N}$ *je pozitivně definitní. Potom existuje matice* $\mathbb{P}$ *typu* $p \times p$, *která je regulární a lze psát*

$$\mathbb{M} = \mathbb{P}'\Lambda\mathbb{P},$$
$$\mathbb{N} = \mathbb{P}'\mathbb{P},$$

přitom matice Λ *typu* $p \times p$ *je diagonální.*

Důkaz. Viz [12]. □

Lemma B.3. *Nechť* $\mathbf{u} \in \mathbb{R}^p$ *je nenulový vektor. Potom matice* $\mathbb{U} = \mathbf{u}\mathbf{u}'$ *je nenulová a platí* $\mathrm{r}(\mathbb{U}) = 1$.

Důkaz. Nechť $\mathbf{u} = (u_1, \ldots, u_p)'$. Kdyby matice $\mathbb{U}$ byla nulová, muselo by mj. platit $u_i^2 = 0 \ \forall i = 1, \ldots, p$, což vede ke sporu s nenulovostí vektoru $\mathbf{u}$. Tedy matice $\mathbb{U}$ je skutečně nenulová.

$$\mathrm{r}(\mathbb{U}) = \mathrm{r}(\mathbf{u}\mathbf{u}') \leq \mathrm{r}(\mathbf{u}) = 1.$$

Přitom $\mathrm{r}(\mathbb{U}) = 0$ tehdy a jen tehdy, je-li matice $\mathbb{U}$ nulová, což nemůže nastat. Tudíž $\mathrm{r}(\mathbb{U}) = 1$. □

Lemma B.4. *Nechť* $p \geq 2$. *Buď* $\mathbf{u} \in \mathbb{R}^p$ *nenulový vektor. Nechť* $\mathbb{N}$ *je pozitivně definitní matice typu* $p \times p$. *Potom* $\mathbf{u}\mathbf{u}'\mathbf{z} = \eta\mathbb{N}\mathbf{z}$ *pro nějaký nenulový vektor* $\mathbf{z} \in \mathbb{R}^p$ *tehdy a jen tehdy, je-li* $\eta = 0$ *nebo* $\eta = \Delta^2$, *kde* $\Delta = \sqrt{\mathbf{u}'\mathbb{N}^{-1}\mathbf{u}}$.

Důkaz. (1a) Nechť $\eta = 0$. $\mathbf{u}\mathbf{u}'\mathbf{z} = \mathbf{0}$ pro nějaký nenulový vektor $\mathbf{z}$ tehdy a jen tehdy, je-li $\det(\mathbf{u}\mathbf{u}') = 0$, což je splněno, neboť matice $\mathbf{u}\mathbf{u}'$ je typu $p \times p$, $p \geq 2$ a podle lemmatu B.3 je $\mathrm{r}(\mathbf{u}\mathbf{u}') = 1$.

(1b) Nechť $\eta = \Delta^2$. $\Delta \neq 0$, neboť $\mathbf{u}$ je nenulový vektor a matice $\mathbb{N}$ je pozitivně definitní. $\mathbf{u}\mathbf{u}'\mathbf{z} = \Delta^2\mathbb{N}\mathbf{z}$ pro vektor

$$\mathbf{z} = \frac{1}{\Delta}\mathbb{N}^{-1}\mathbf{u}.$$

Přitom vektor $\mathbf{z}$ je nenulový díky pozitivní definitnosti matice $\mathbb{N}$ a nenulovosti vektoru $\mathbf{u}$.

(2) Užitím věty B.2 dostáváme existenci regulární matice $\mathbb{P}$ a diagonální matice $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$, které splňují:

$$\mathbf{u}\mathbf{u}' = \mathbb{P}'\Lambda\mathbb{P},$$
$$\mathbb{N} = \mathbb{P}'\mathbb{P}.$$

Tohoto rozkladu záhy využijeme.

Nechť $\mathbf{u}\mathbf{u}'\mathbf{z} = \eta\mathbb{N}\mathbf{z}$ pro nějaký nenulový vektor $\mathbf{z}$.
$\Leftrightarrow (\mathbf{u}\mathbf{u}' - \eta\mathbb{N})\mathbf{z} = \mathbf{0}$ pro nějaký nenulový vektor $\mathbf{z}$.
$\Leftrightarrow \det(\mathbf{u}\mathbf{u}' - \eta\mathbb{N}) = 0$.
$\Leftrightarrow \det(\mathbb{P}'\Lambda\mathbb{P} - \eta\mathbb{P}'\mathbb{P}) = 0$.
$\Leftrightarrow \det(\mathbb{P}')\det(\Lambda - \eta I_p)\det(\mathbb{P}) = 0$.
$\Leftrightarrow \det(\Lambda - \eta I_p) = 0$, neboť $\det(\mathbb{P}') = \det(\mathbb{P}) \neq 0$ díky regularitě matice $\mathbb{P}$.
$\Leftrightarrow \prod_{i=1}^{p}(\lambda_i - \eta) = 0$.
$\Leftrightarrow \eta = \lambda_i$ pro nějaké $i = 1, \ldots, p$.

Dále platí, že právě jediný prvek na diagonále matice Λ je nenulový. Pravdivost tohoto tvrzení dostáváme z následujícího faktu:

$$\mathrm{r}(\Lambda) = \mathrm{r}(\mathbb{P}'\Lambda\mathbb{P}) = \mathrm{r}(\mathbf{u}\mathbf{u}') = 1.$$

Označme tento nenulový prvek ϑ.

Tedy $\mathbf{u}\mathbf{u}'\mathbf{z} = \eta\mathbb{N}\mathbf{z}$ pro nějaký nenulový vektor $\mathbf{z}$ tehdy a jen tehdy, když $\eta = 0$ nebo $\eta = \vartheta \neq 0$. Protože vztah $\mathbf{u}\mathbf{u}'\mathbf{z} = \eta\mathbb{N}\mathbf{z}$ pro nějaký nenulový vektor $\mathbf{z}$ splňuje nenulové číslo Δ^2, je nutně $\vartheta = \Delta^2$, čímž je důkaz dokončen. □

Lemma B.5. *Nechť $p \geq 2$. Buď $\mathbf{u} \in \mathbb{R}^p$ nenulový vektor. Nechť $\mathbb{N}$ je pozitivně definitní matice typu $p \times p$. Nechť $\mathbb{P}'\mathbb{P} = \mathbb{N}$ a $\mathbb{P}'\Lambda\mathbb{P} = \mathbf{u}\mathbf{u}'$ jsou rozklady matic $\mathbb{N}$ a $\mathbf{u}\mathbf{u}'$ podle věty B.2. Nechť $\eta = 0$ nebo $\eta = \Delta^2 = \mathbf{u}'\mathbb{N}^{-1}\mathbf{u}$. Nechť $\mathbf{z} \in \mathbb{R}^p$, $\mathbf{z} \neq \mathbf{0}$. Potom $\mathbf{u}\mathbf{u}'\mathbf{z} = \eta\mathbb{N}\mathbf{z}$ tehdy a jen tehdy, je-li $\mathbf{z} = \mathbb{P}^{-1}\mathbf{w}$, kde $\Lambda\mathbf{w} = \eta\mathbf{w}$.*

Důkaz. (1) Nechť $\mathbf{u}\mathbf{u}'\mathbf{z} = \eta\mathbb{N}\mathbf{z}$. S využitím rozkladu získaného z věty B.2 postupně upravujme:

$$\begin{aligned}
(\mathbf{u}\mathbf{u}' - \eta\mathbb{N})\mathbf{z} &= \mathbf{0} \\
(\mathbb{P}'\Lambda\mathbb{P} - \eta\mathbb{P}'\mathbb{P})\mathbf{z} &= \mathbf{0} \\
\mathbb{P}'(\Lambda - \eta I_p)\mathbb{P}\mathbf{z} &= \mathbf{0} \quad | \cdot (\mathbb{P}')^{-1} \text{ (zleva)} \\
(\Lambda - \eta I_p)\mathbb{P}\mathbf{z} &= \mathbf{0}.
\end{aligned}$$

Položíme-li $\mathbf{w} = \mathbb{P}\mathbf{z}$, platí $\mathbf{z} = \mathbb{P}^{-1}\mathbf{w}$ a navíc $(\Lambda - \eta I_p)\mathbf{w} = \mathbf{0}$.

(2) Nechť $\mathbf{z} = \mathbb{P}^{-1}\mathbf{w}$ a $(\Lambda - \eta I_p)\mathbf{w} = \mathbf{0}$. Z prvně jmenovaného vztahu plyne $\mathbf{w} = \mathbb{P}\mathbf{z}$. Z druhého vztahu po dosazení za $\mathbf{w}$ a jeho vynásobení maticí $\mathbb{P}'$ zleva dostáváme $\mathbb{P}'(\Lambda - \eta I_p)\mathbb{P}\mathbf{z} = \mathbf{0}$, z čehož snadnou úpravou s využitím rozkladu získaného z věty B.2 získáváme $(\mathbf{u}\mathbf{u}' - \eta\mathbb{N})\mathbf{z} = \mathbf{0}$. □

Z lemmat B.4 a B.5 tedy vidíme, že nenulové vektory $\mathbf{z}$ splňující $\mathbf{u}\mathbf{u}'\mathbf{z} = \eta\mathbb{N}\mathbf{z}$ pro $\eta = 0$, resp. $\eta = \Delta^2$ získáme transformací vlastních vektorů matice Λ, příslušejících vlastnímu číslu 0, resp. Δ^2.

Nyní již můžeme přistoupit k slibovanému důkazu věty B.1.

Důkaz věty B.1.

Nechť $\mathbf{u} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. Podle předpokladů je $\mathbf{u} \neq 0$, $\mathbf{u} \in \mathbb{R}^p$ a $\Delta = \sqrt{\mathbf{u}'\Sigma^{-1}\mathbf{u}}$.

$$E\mathbf{X}^\star = \mathbf{a} + \mathbb{A}E\mathbf{X} = \begin{cases} \mathbf{a} + \mathbb{A}\boldsymbol{\mu}_1 & \text{s pravděpodobností } \lambda \\ \mathbf{a} + \mathbb{A}\boldsymbol{\mu}_0 & \text{s pravděpodobností } 1 - \lambda. \end{cases}$$

$$\operatorname{var}\mathbf{X}^\star = \mathbb{A}\operatorname{var}\mathbf{X}\mathbb{A}' = \mathbb{A}\Sigma\mathbb{A}'.$$

Rozdělení náhodného vektoru $\mathbf{X}^\star$ je zřejmě směsí dvou p-rozměrných normálních rozdělení. K provedení důkazu tedy stačí ukázat, že existují $\mathbf{a} \in \mathbb{R}^p$ a matice $\mathbb{A}$ typu $p \times p$ takové, že

$$\begin{aligned}
\mathbf{a} + \mathbb{A}\boldsymbol{\mu}_1 &= \frac{\Delta}{2}\mathbf{e}_1, \\
\mathbf{a} + \mathbb{A}\boldsymbol{\mu}_0 &= -\frac{\Delta}{2}\mathbf{e}_1, \\
\mathbb{A}\Sigma\mathbb{A}' &= I_p.
\end{aligned}$$

Výše uvedená soustava rovnic je ekvivalentní soustavě:

$$\begin{aligned}
\mathbf{a} &= \frac{\Delta}{2}\mathbf{e}_1 - \mathbb{A}\boldsymbol{\mu}_1, \\
\mathbb{A}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) &= \Delta\mathbf{e}_1 \qquad (\text{tj. } \mathbb{A}\mathbf{u} = \Delta\mathbf{e}_1), \\
\mathbb{A}\Sigma\mathbb{A}' &= I_p.
\end{aligned}$$

Dále se budeme věnovat řešitelnosti posledních dvou rovnic.
Nechť

$$\mathbb{A} = \begin{pmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix}, \quad \mathbf{a}_i \in \mathbb{R}^p,\ i = 1, \ldots, p.$$

Pomocí tohoto označení je možné zmíněné dvě rovnice přepsat do tvaru:

$$\begin{aligned} \mathbf{a}_1'\mathbf{u} &= \Delta, \\ \mathbf{a}_i'\mathbf{u} &= 0, \quad i = 2, \ldots, p, \\ \mathbf{a}_i'\Sigma\mathbf{a}_j &= \delta_{i,j}, \quad i, j = 1, \ldots, p. \end{aligned}$$

Z poslední sady rovnic vyplývá při volbě $i = j,\ i = 1, \ldots, p$, že vektory $\mathbf{a}_1, \ldots, \mathbf{a}_p$ musejí být nenulové, neboť matice Σ je pozitivně definitní.

Nyní rozlišíme případy $p = 1$ a $p \geq 2$.

(1) $p = 1$

Nechť $\Sigma = \sigma^2 > 0$. $\mathbf{u} = u \neq 0$. Zřejmě $\Delta = \frac{|u|}{\sqrt{\sigma^2}}$. Hledáme číslo a_1 takové, že

$$\begin{aligned} a_1 u &= \Delta, \\ a_1^2\sigma^2 &= 1. \end{aligned}$$

Tyto dvě rovnosti ovšem splňuje $a_1 = \frac{\Delta}{u} \neq 0$.
Tedy lze volit:

$$\begin{aligned} \mathbb{A} &= \frac{\Delta}{u} = \frac{\Delta}{\mu_1 - \mu_0}, \\ \mathbf{a} &= \frac{\Delta}{2} - \frac{\Delta}{u}\mu_1 = \frac{\Delta}{2} - \frac{\Delta}{\mu_1 - \mu_0}\mu_1. \end{aligned}$$

(2) $p \geq 2$

Podle věty B.2 a lemmat B.3 a B.4 existuje regulární matice $\mathbb{P}$ taková, že

$$\begin{aligned} \mathbf{u}\mathbf{u}' &= \mathbb{P}'\Lambda\mathbb{P}, \\ \Sigma &= \mathbb{P}'\mathbb{P}, \end{aligned}$$

kde Λ je diagonální matice s jediným nenulovým prvkem na diagonále, který je roven Δ^2.

Nechť

$$\mathbf{a}_1 = \frac{1}{\Delta}\Sigma^{-1}\mathbf{u} \neq \mathbf{0}.$$

Potom jsou splněny rovnice:

$$\begin{aligned} \mathbf{a}_1'\mathbf{u} &= \Delta, \\ \mathbf{a}_1'\Sigma\mathbf{a}_1 &= 1. \end{aligned}$$

Navíc platí $\mathbf{u}\mathbf{u}'\mathbf{a}_1 = \Delta^2\Sigma\mathbf{a}_1$. Tedy podle lemmatu B.5 lze $\mathbf{a}_1$ napsat jako $\mathbf{a}_1 = \mathbb{P}^{-1}\mathbf{b}_1$, kde $\mathbf{b}_1$ je vlastní vektor matice Λ, příslušný vlastnímu číslu $\Delta^2 \neq 0$. Dále $\mathbf{b}_1 = \mathbb{P}\mathbf{a}_1$, z čehož dostaneme $\mathbf{b}_1'\mathbf{b}_1 = \mathbf{a}_1'\mathbb{P}'\mathbb{P}\mathbf{a}_1 = \mathbf{a}_1'\Sigma\mathbf{a}_1 = 1$. Tudíž $\mathbf{b}_1$ je normovaný vlastní vektor matice Λ.

Podle lemmat B.4 a B.5 má matice Λ pouze jediné nenulové vlastní číslo Δ^2. Nechť $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p$ jsou ortonormální vlastní vektory matice Λ, přitom $\mathbf{b}_2, \ldots, \mathbf{b}_p$ příslušejí vlastnímu číslu 0.
Tedy je splněno:

$$\begin{aligned} \mathbf{b}_i'\mathbf{b}_j &= \delta_{i,j}, \quad i,j = 1,2,\ldots,p, \\ \Lambda\mathbf{b}_i &= \mathbf{0}, \quad i = 2,\ldots,p. \end{aligned}$$

Podle lemmatu B.5 jsou $\mathbf{a}_i = \mathbb{P}^{-1}\mathbf{b}_i,\ i = 2,\ldots,p$ nenulové vektory, pro které platí:

$$\mathbf{u}\mathbf{u}'\mathbf{a}_i = \mathbf{0}, \quad i = 2,\ldots,p$$

a navíc

$$\mathbf{a}_i'\Sigma\mathbf{a}_j = \mathbf{b}_i'(\mathbb{P}^{-1})'\mathbb{P}'\mathbb{P}\mathbb{P}^{-1}\mathbf{b}_j = \mathbf{b}_i'\mathbf{b}_j = \delta_{i,j}, \quad i,j = 1,2,\ldots,p.$$

Přitom vztah $\mathbf{u}\mathbf{u}'\mathbf{a}_i = \mathbf{0},\ i = 2,\ldots,p$, je ekvivalentní vztahu

$$\mathbf{a}_i'\mathbf{u} = 0, \quad i = 2,\ldots,p.$$

Kdyby $(\mathbf{a}_i'\mathbf{u} \neq 0\ \&\ \mathbf{u}\mathbf{u}'\mathbf{a}_i = \mathbf{0})$, potom $(\sum_{k=1}^p a_{i,k}u_k \neq 0\ \&\ u_l(\sum_{k=1}^p a_{i,k}u_k) = 0\ \forall l = 1,\ldots,p)$, což je ve sporu s předpokladem $\mathbf{u} \neq \mathbf{0}$. Implikace $(\mathbf{a}_i'\mathbf{u} = 0,\ i = 2,\ldots,p \Rightarrow \mathbf{u}\mathbf{u}'\mathbf{a}_i = \mathbf{0},\ i = 2,\ldots,p)$ je snadná.

Tedy skutečně existují nenulové vektory $\mathbf{a}_1,\ldots,\mathbf{a}_p \in \mathbb{R}^p$, které splňují vztahy:

$$\begin{aligned} \mathbf{a}_1'\mathbf{u} &= \Delta, \\ \mathbf{a}_i'\mathbf{u} &= 0, \quad i = 2,\ldots,p, \\ \mathbf{a}_i'\Sigma\mathbf{a}_j &= \delta_{i,j}, \quad i,j = 1,2,\ldots,p. \end{aligned}$$

Tudíž matice

$$\mathbb{A} = \begin{pmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix}$$

a vektor

$$\mathbf{a} = \frac{\Delta}{2}\mathbf{e}_1 - \mathbb{A}\boldsymbol{\mu}_1$$

mají požadované vlastnosti. □

C. PROGRAMOVÉ VYBAVENÍ

Na tomto místě uvádíme dokumentaci k programům, jež byly použity k některým dříve uvedeným výpočtům a seznam a popis souborů uložených na přiložené disketě.

C.1. Programy pro Matlab

Na následujících řádcích je uveden seznam vytvořených funkcí, které jsou určeny k výpočtům pomocí programu MATLAB. Vše je uvedeno se stručným popisem a charakterizací vstupních a výstupních argumentů ve tvaru, v kterém se zadávají při samotných výpočtech, tj. [*výstupní argumenty*] = název funkce (*vstupní argumenty*). Popis některých častěji se opakujících se argumentů uvedeme na začátku. V závorce je vždy uvedena dimenze příslušného objektu, místo 1×1 je uvedeno pouze 1. Rozsah učící skupiny je označen jako n, počet vysvětlujících veličin jako p.

$\mathbb{X}_{(-1)}(n \times p)$ ... matice dat vysvětlujících náhodných vektorů. Její i-tý řádek odpovídá hodnotám znaků naměřených na i-tém objektu učící skupiny. Matice $\mathbb{X}_{(-1)}$ je rovna matici $\mathbb{X}$ z kapitoly 2.1, ve které je vynechán první sloupec obsahující jedničky.
$\mathbf{Y}(n \times 1)$ ... vektor obsahující nuly a jedničky. Jeho i-tá složka určuje zařazení i-tého objektu učící skupiny.
$\mathbb{Z}(n \times (p+1))$ ... $\mathbb{Z} = (\mathbb{X}_{(-1)}, \mathbf{Y})$.

C.1.A. Pomocné funkce

■ g = norhust $(\boldsymbol{\mu}, \Sigma, \mathbf{x})$
$g(1), \boldsymbol{\mu}(p \times 1), \Sigma(p \times p), \mathbf{x}(p \times 1)$
Funkce pro výpočet hustoty p-rozměrného normálního rozdělení se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí Σ v bodě $\mathbf{x}$. Dimenze p je určena počtem složek vektoru $\boldsymbol{\mu}$.

■ l = LRloglik $(\boldsymbol{\gamma}, \mathbb{Z})$
$l(1), \boldsymbol{\gamma}((p+1) \times 1)$
Funkce pro výpočet logaritmické věrohodnostní funkce v modelu (LR) v bodě $\boldsymbol{\gamma} = (\beta_0, \boldsymbol{\beta}')'$ při daných hodnotách vysvětlujících veličin $\mathbf{X}_1, \ldots, \mathbf{X}_n$ a daném $\mathbf{Y}$ určených maticí $\mathbb{Z}$. Počítáno podle vzorce (2.1.2).

■ *minusl* = LRloglik_minus $(\boldsymbol{\gamma}, \mathbb{Z})$
minusl$(1), \boldsymbol{\gamma}((p+1) \times 1)$
minusl je rovno $(-1)\cdot$LRloglik$(\boldsymbol{\gamma}, \mathbb{Z})$.

C.1.B. Funkce pro výpočet odhadů neznámých parametrů

■ $[\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}, \mathfrak{l}] =$ LRest $(\mathbb{X}_{(-1)}, \mathbf{Y})$

$\tilde{\beta}_0(1)$, $\tilde{\boldsymbol{\beta}}(p \times 1)$, $\mathfrak{l}(1)$

Funkce pro výpočet maximálně věrohodných odhadů parametrů β_0, $\boldsymbol{\beta}$ v modelu (LR). Odhady se počítají maximalizací logaritmické věrohodnosti (2.1.2) pomocí funkce Matlabu `fminu`. Výsledná hodnota logaritmické věrohodnosti je uložena v proměnné $\mathfrak{l}$.

■ $[\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}] =$ LRbetas $(\mathbb{X}_{(-1)}, \mathbf{Y})$

$\tilde{\beta}_0(1)$, $\tilde{\boldsymbol{\beta}}(p \times 1)$

Funkce totožná funkci LRest, pouze s tím rozdílem, že nevrací hodnotu logaritmické věrohodnosti v bodě maxima.

■ $[\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_0, \hat{\lambda}, \hat{\Sigma}] =$ NDAest $(\mathbb{X}_{(-1)}, \mathbf{Y})$

$\hat{\boldsymbol{\mu}}_1(p \times 1)$, $\hat{\boldsymbol{\mu}}_0(p \times 1)$, $\hat{\lambda}(1)$, $\hat{\Sigma}(p \times p)$

Funkce počítá podle vzorců (2.2.3) odhady neznámých parametrů v modelu (NDA).

■ $[\hat{\beta}_0, \hat{\boldsymbol{\beta}}] =$ NDAbetas $(\mathbb{X}_{(-1)}, \mathbf{Y})$

$\hat{\beta}_0(1)$, $\hat{\boldsymbol{\beta}}(p \times 1)$

Podle vzorce (1.2.1) jsou počítány odhady koeficientů v diskriminační funkci pro model (NDA). Do (1.2.1) se dosazují maximálně věrohodné odhady získané podle (2.2.3).

■ $[\breve{\boldsymbol{\mu}}_1, \breve{\boldsymbol{\mu}}_0, \breve{\lambda}, \breve{\Sigma}, \mathfrak{l}, pocit, tolerkonec, normal] =$ MNDest $(\mathbb{X}_{(-1)}, tolerance, rovnat, \lambda^0)$

$\breve{\boldsymbol{\mu}}_1(p\times1)$, $\breve{\boldsymbol{\mu}}_0(p\times1)$, $\breve{\lambda}(1)$, $\breve{\Sigma}(p\times p)$, $\mathfrak{l}(1)$, $pocit(1)$, $tolerkonec(1)$, $normal(1)$, $tolerance(1)$, $rovnat(1)$, $\lambda^0(1)$

Funkce pro výpočet maximálně věrohodných odhadů v modelu (MND) pomocí EM algoritmu. Proměnná *tolerance* určuje ukončovací pravidlo iteračního procesu. Výpočet končí v okamžiku, kdy rozdíl hodnot logaritmických věrohodností ve dvou po sobě jdoucích iteracích je menší než *tolerance*. Rozdíl hodnot logaritmických věrohodností v posledním a předposledním kroku algoritmu je uložen v proměnné *tolerkonec*.

Řádky matice $\mathbb{X}_{(-1)}$ se setřídí sestupně podle sloupce s indexem *rovnat*. Nechceme-li řádky matice $\mathbb{X}_{(-1)}$ uspořádávat, zvolíme $rovnat = 0$. Z počátečních odhadů postačuje zadat pouze počáteční hodnotu parametru λ, tj. λ^0. Zbylé počáteční odhady $(\boldsymbol{\mu}_1^0, \boldsymbol{\mu}_0^0, \Sigma^0)$ se potom počítají podle vzorců (1.2.1) ze setříděné matice $\mathbb{X}_{(-1)}$, k níž by příslušel vektor $\mathbf{Y}^0$, který obsahuje $\lambda^0 \cdot n$ jedniček a $(1-\lambda^0)\cdot n$ nul (počty jedniček a nul jsou zaokrouhleny na celá čísla). Využívá se zde předpokladu, že skupině, kterou identifikujeme pomocí $Y = 1$, příslušejí velké hodnoty vysvětlujícího znaku s indexem *rovnat*.

V proměnné $\mathfrak{l}$ je uložena hodnota logaritmické věrohodnostní funkce po poslední iteraci, *pocit* označuje počet provedených iterací. Proměnná *normal* nabývá hodnoty jedna, pokud $pocit \leq maxiter$ a nuly jinak, *maxiter* přitom určuje maximální počet prováděných iterací a je nastavena uvnitř funkce na hodnotu 500. Dosáhne-li počet provedených iterací hodnoty *maxiter*, výpočet skončí bez ohledu na rozdíl logaritmických věrohodností v posledních dvou krocích.

V průběhu výpočtu se na obrazovce zobrazuje postupně počet provedených iterací, hodnota logaritmické věrohodnosti po každém kroku a rozdíl vždy posledních dvou logaritmických věrohodností.

■ $[\breve{\beta}_0, \breve{\boldsymbol{\beta}}] =$ MNDbetas $(\breve{\boldsymbol{\mu}}_1, \breve{\boldsymbol{\mu}}_0, \breve{\lambda}, \breve{\Sigma})$
$\breve{\beta}_0(1)$, $\breve{\boldsymbol{\beta}}(p \times 1)$, $\breve{\boldsymbol{\mu}}_1(p \times 1)$, $\breve{\boldsymbol{\mu}}_0(p \times 1)$, $\breve{\lambda}(1)$, $\breve{\Sigma}(p \times p)$
Podle vzorce (1.2.1) jsou počítány odhady koeficientů v diskriminační funkci pro model (MND). Do (1.2.1) se dosazují odhady $\breve{\boldsymbol{\mu}}_1, \breve{\boldsymbol{\mu}}_0, \breve{\lambda}, \breve{\Sigma}$.

C.1.C. Další výpočtové funkce

■ $[\overline{err}, err^j, err^+, \omega^j, \omega^+] =$ LRerr $(\mathbb{X}_{(-1)}, \mathbf{Y})$
$\overline{err}(1)$, $err^j(1)$, $err^+(1)$, $\omega^j(1)$, $\omega^+(1)$
Funkce pro výpočet veličin $\overline{err}$, err^j, err^+, ω^j, ω^+ definovaných v kapitole 7.1, pro model logistické regrese s učící skupinou určenou maticemi $\mathbb{X}_{(-1)}$, $\mathbf{Y}$.

■ $[\overline{err}, err^j, err^+, \omega^j, \omega^+] =$ NDAerr $(\mathbb{X}_{(-1)}, \mathbf{Y})$
$\overline{err}(1)$, $err^j(1)$, $err^+(1)$, $\omega^j(1)$, $\omega^+(1)$
Funkce pro výpočet veličin $\overline{err}$, err^j, err^+, ω^j, ω^+ definovaných v kapitole 7.1, pro model normální diskriminační analýzy s učící skupinou určenou maticemi $\mathbb{X}_{(-1)}$, $\mathbf{Y}$.

■ $[\overline{err}, err^j, err^+, \omega^j, \omega^+] =$ MNDerr $(\mathbb{X}_{(-1)}, \mathbf{Y}, tolerance, rovnat, \lambda^0)$
$\overline{err}(1)$, $err^j(1)$, $err^+(1)$, $\omega^j(1)$, $\omega^+(1)$, $tolerance(1)$, $rovnat(1)$, $\lambda^0(1)$
Funkce pro výpočet veličin $\overline{err}$, err^j, err^+, ω^j, ω^+ definovaných v kapitole 7.1, pro model směsi normálních rozdělení s učící skupinou určenou maticemi $\mathbb{X}_{(-1)}$, $\mathbf{Y}$. Funkce je použitelná pouze v případě, že známe zařazení objektů učící skupiny. Diskriminační procedury nutné k výpočtu $\overline{err}$, err^j, err^+, ω^j, ω^+ se získají pomocí funkce MNDest s argumenty $tolerance$, $rovnat$ a λ^0.

■ $[\mathbf{c}, \mathbf{d}, \mathbf{o}, \overline{\boldsymbol{\pi}}, \mathbf{m}^*, \hat{C}, p_{level}] =$ HL_test $(\mathbb{X}_{(-1)}, \mathbf{Y}, \beta_0, \boldsymbol{\beta}, g)$
$\mathbf{c}(g \times 1)$, $\mathbf{d}(g \times 1)$, $\overline{\boldsymbol{\pi}}(g \times 1)$, $\mathbf{m}^*(g \times 1)$, $\hat{C}(1)$, $p_{level}(1)$, $\beta_0(1)$, $\boldsymbol{\beta}(p \times 1)$, $g(1)$
Funkce určená k provedení Hosmerova-Lemeshowova testu (viz kapitola 4.2). β_0, $\boldsymbol{\beta}$ jsou parametry logistického modelu, jehož shodu s našimi daty určenými maticemi $\mathbb{X}_{(-1)}$, $\mathbf{Y}$ chceme prokázat. Číslo g označuje počet sloupců kontingenční tabulky, do které lze napsat seskupená data. Vektor $\mathbf{c}$ označuje odhady teoretických četností v jednotlivých sloupcích tabulky v řádku $Y = 1$, $\mathbf{d}$ v řádku $Y = 0$. Pozorované četnosti v jednotlivých sloupcích pro řádek $Y = 1$ obsahuje vektor $\mathbf{o}$. Počty pozorování v jednotlivých sloupcích tabulky je možné zjistit z vektoru $\mathbf{m}^*$. Vektor $\overline{\boldsymbol{\pi}}$ obsahuje odhady pravděpodobností $Y = 1$ za podmínky, že vektor vysvětlujících veličin leží v i-té decilové skupině. Číslo $\hat{C}$ udává hodnotu testové statistiky Hosmerova-Lemeshowova testu a p_{level} dosaženou hladinu tohoto testu. Procedura předpokládá, že žádné dva řádky matice $\mathbb{X}_{(-1)}$ nejsou shodné.

■ $[\mathfrak{S}^1, \mathfrak{S}^0] =$ varmatice $(\mathbb{X}_{(-1)}, \mathbf{Y})$
$\mathfrak{S}^1(p \times p)$, $\mathfrak{S}^0(p \times p)$
Výpočet odhadů variančních matic pro Boxův test (viz kapitola 4.3).

■ $[Box, p_{level}] =$ Box_test $(\mathbb{X}_{(-1)}, \mathbf{Y})$
$Box(1)$, $p_{level}(1)$
Provedení Boxova testu (viz kapitola 4.3). Box je hodnota testové statistiky, p_{level} dosažená hladina testu.

C.1.D. Funkce pro zadávání dat

■ $[\mathbb{X}_{(-1)}, \mathbf{Y}]$ = kosti

Funkce zajišťující vytvoření matic $\mathbb{X}_{(-1)}$, $\mathbf{Y}$ z datového souboru `kosti.sta`. Sloupce matice $\mathbb{X}_{(-1)}$ jsou dva a odpovídají proměnným fos_8 a upd_51.

■ $[\mathbb{X}_{(-1)}, \mathbf{Y}]$ = kosti_k

Funkce zajišťující vytvoření matic $\mathbb{X}_{(-1)}$, $\mathbf{Y}$ z datového souboru `kosti.sta`. Sloupce matice $\mathbb{X}_{(-1)}$ jsou nyní čtyři a odpovídají postupně proměnným fos_8, upd_51, $kult\_une$ a $kult\_snu$.

C.2. Přiložená disketa

Přiložená disketa obsahuje tyto adresáře.

- `DATA`
- `EXCEL`
- `MATLAB`
- `TISK`

Obsah těchto adresářů je následující. Případné zabalení bylo provedeno pomocí programu WinZip.

DATA

- `kosti.s0`
- `kosti.s1`
- `kosti.sta`
  - datové soubory pro příklad z kapitoly 8.1.

EXCEL

- `pr5_1_2_95.xls`
  - soubor pro příklad v kapitole 5.1, část (2), formát pro Excel95.
- `pr5_1_2_97.xls`
  - soubor pro příklad z kapitoly 5.1, část (2), formát pro Excel97.

MATLAB

- `FUNKCE`
  - adresář s funkcemi pro Matlab, popsanými v kapitolách C.1.A až C.1.C.
- `DATA_M`
  - adresář s funkcemi pro zadávání dat v Matlabu, popsanými v kapitole C.1.D.

TISK

- `postscript.zip`
  - zabalené postscriptové soubory s jednotlivými kapitolami diplomové práce.

– `kosti_gr.xls`
  – soubor s grafy z dodatku D, ve formátu pro Excel97.

C.3. Použité programy

K výpočtům v rámci diplomové práce byly použity následující programy.

Matlab 5
Statistica 4.5
Ncss 6.0
Maple V
Microsoft Excel 7.0/95 a 97

D. GRAFY

Na následujících stránkách jsou zařazeny grafy D.1 až D.5, ilustrující určování pohlaví jedince při archeologickém výzkumu (viz kapitola 8.1).