Biometrické metody v genetice, odhadů genetických parametrů

- lineární modely

prof. Ing. Tomáš Urban, Ph.D.
urban@mendelu.cz

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Proč biometrické metody v genetice

Cíle

- Popsat genetickou strukturu populací (odhad komponent variance a kovariance) a popsat změny genetické výstavby populací
- Na znalosti genetické struktury populací jsou založeny šlechtitelské programy

Možnosti biometrických metod:

1. Odhady výkonnosti populací – čistokrevné i hybridní
2. odhady genetických parametrů - h^2, r_{op}, r_G, …
3. odhady plemenné hodnoty (PH) – rozdíly mezi jedincem a vrstevníky, očištěný od negenetických vlivů (realizace šlecht. programů)
4. Stanovení selekčního (genetického) zisku
5. Optimalizace selekčních a hybridizačních programů

Uplatnění poznatků: molekulární a biochemické genetiky, cytogenetiky, imunogenetiky a genové manipulace v genetice populací

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Kvantitativní genetika – hodnocení pomocí modelů

Biometrika v genetice (≈ *kvantitativní genetika*)

Účinek polygenů se sleduje na základě počtu pravděpodobnosti (hromadné jevy).

Společné efekty více genů vytváří proměnlivost, většinou s normálním rozdělením, kterou lze analyzovat matematicko-statistickými operacemi.

Teorie: přenos GI u kvantitativních vlastností je **polygenní** (**velký počet lokusů** s mendelistickým přenosem + větší či menší vliv **prostředí -** vnitřní a vnější).

Operační metody pro analýzu přenosu této GI: **biometrické**.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Analýza variance (ANOVA)

Funkce ANOVA (Fisher 1918):

$$\sigma^2_{celková} = \sigma^2_1 + \sigma^2_2 + ... + \sigma^2_k$$

1. odhad pevných efektů
2. odhad komponent (složek) variance – podíl jednotlivých variancí, např. varianci genotypovou nebo prostředí
3. testování hypotéz o příčinách variance modelem (jak vznikla, velikost vlivu faktorů)

ANOVA nebalancované metody	speciální případ nebalancovaných metod	Balancované metody –výjimečné –speciální případ nebalancované metody
1. velké systém rovnic s využitím matic 2. nelze realizovat podle plánu – náhodný efekt (využití u zvířat) 3. hodnotí se chovy, šlechtění (software: Harvey, SAS, BMPD) – ***metody nejmenších čtverců, maximální věrohodnosti***		1. přesnější 2. plánované pokusy (u zvířat toho nelze dosáhnout)
1. otec má 100 potomků, 2. jich má 50 a 3. 10 –> to je nebalancované		- stejný počet pozorování ve všech podtřídách

Biometrické modely - lineární

Biometrické metody spočívají na lineárních biometrických modelech.

Pravdivý (skutečný, teoretický) model

popisuje data přesně, bez reziduální nebo nevysvětlené variance. Variance P je vyčerpána faktory. Pravdivý model není nikdy přesně znám.

Ideální (praktický) model

je vytvořen výzkumníkem, který je tak blízký skutečnému modelu, jak jen to je možné. Takový model by se měl používat k analýzám, ale často není dostatek informací (chybí).

Operační (pracovní, proveditelný) model

je zjednodušená forma ideálního modelu a je využíván výzkumníky v analýzách. Na této úrovni se vede široká diskuse o nejlepší operační model.

Pozorování

Vektor pozorování **y** obsahuje prvky vyplývající z měření vlastnosti v daných jednotkách

- předpoklad – že se jedná o náhodný výběr z nekonečně velké populace

Efekty

- Efekty (faktory) se vztahují k proměnným, které mohou ovlivňovat nebo být ve vztahu k prvkům ve vektoru pozorování
- *Diskrétní efekty* mají obvykle třídy nebo úrovně
- „obtěžující efekty“ - musí být zahrnuty → minimalizace **e**

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Pevné a náhodné efekty

Pevné efekty (fixní) jsou ty, v kterých úrovně zahrnují všechny možné úrovně, které lze pozorovat.

Náhodné efekty jsou efekty, jejichž úrovně jsou považovány za náhodně vybrané z nekonečně velké populace úrovní.

1. *Kolik úrovní má efekt v modelu?* Jestliže málo, pak je to pravděpodobně pevný efekt, jestliže mnoho, pak se jedná o náhodný efekt.
2. *Je počet úrovní efektu v populaci dost velký na to, aby mohla být považována za nekonečnou?* Jestliže ano, pak je pravděpodobně efekt náhodný.
3. *Budou použity opět stejné úrovně, jestliže by byl experiment opakován podruhé?* Jestliže ano, pak se jedná pravděpodobně o pevný efekt.
4. *Byly úrovně efektu určeny nenáhodným způsobem?* Jestliže ano, pak by měl být efekt určen jako pevný.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Modely

Lineární modely obsahují řadu efektů (faktorů), které aditivně ovlivňují pozorování

V tradičním smyslu jsou lineární modely složeny ze tří částí:

1. Rovnice.
2. Matice očekávaných hodnot a variančně kovarianční matice náhodných proměnných.
3. Předpoklady a omezení.

ad 1. **Rovnice**

Rovnice modelu definuje efekty, které mohou mít vliv na pozorovanou vlastnost. Čím více faktorů pokryjeme, tím je výpočet přesnější, tím více se blížíme k variabilitě způsobenou genotypem.

Lineární funkce určitých parametrů a proměnných:

$$\mathbf{y}_{ij} = \mu + \mathbf{b}_i + \boldsymbol{u}_j + \mathbf{e}_{ijk} \qquad \mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

ad 2. **Matice očekávaných hodnot a VCV**

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{Xb}\\ 0\\ 0\end{bmatrix} \qquad V\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & 0\\ 0 & \mathbf{R}\end{bmatrix}$$

kde G a R jsou základní čtvercové matice s předpokladem nesingularity a pozitivní definovanosti a s prvky, které jsou známé. Takže: V(y) = ZGZ´+ R.

ad 3. Předpoklady a omezení

informace o datech nebo způsob jejich sběru, náhodnost výběru, podmínkách chovu apod.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Typy lineárních modelů

Lineární modely (obecně)

$y_{ij} = \mu + a_i + e_{ij}$ e_{ij} $N(0, \sigma^2_e)$; a_i = faktor s i-tými úrovněmi

Regresní modely – funkční vztahy

$y_i = a + bX_i + e_i$ a – konstanta, b, regresní koef., a, b odhadujeme MNČ nebo MV)

Mnohonásobné regresní vztahy

$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + e_i$

Modely s pevnými efekty (více faktorové)

$y_{ijkl} = \mu + a_i + b_j + c_k + e_{ijk,}$ $y_{ijk} = \mu + a_i + b_{ij} + e_{ijk}$

Modely s náhodnými efekty

$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk}$ α_i $N(0, \sigma^2_\alpha)$

Modely se smíšenými efekty

$y_{ijk} = \mu + a_i + \beta_j + e_{ijk}$

smíšené modely se používají k odhadu PH

Komplikují odhad komponent variance

Komplikují odhad fixních efektů

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Vyjádření modelů maticovým zápisem

Skalární zápis modelu s pevnými efekty:

$$y_{ijk} = \mu + a_i + b_j + e_{ijk}$$

jedna pozorovaná hodnota (zastupuje všechny pozor. hodnoty) je symbolicky znázorněna

Maticový model s pevnými efekty, kde jsou vyjádřeny všechny pozorované hodnoty

$$\mathbf{y = Xb} + \mathbf{e}$$

y – vektor pozorování

X – incidenční matice (designová, strukturní matice) – uvádí, které pevné efekty jsou obsaženy v **y**)

b – vektor odhadovaných parametrů

c – vektor náhodných efektů: $e \sim N(0, I\,\sigma^2_e)$

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Vybalancovaný pokus

Analýza množství tuku v mléce u 18 dojnic s vlivem efektů stáda a věku:
a_i – stádo (i = 1, 2) ; b_j – věk (j = 1, 2, 3)

		věk			
		b_1	**b_2**	**b_3**	**průměr**
stádo	**a_1**	165 154 148	136 116 128	161 157 165	**147,78**
	a_2	168 154 120	115 142 186	112 118 128	**138,11**
	průměr	**151,50**	**137,17**	**140,17**	**142,94**

$$\hat{a}_i = \bar{y}_{i}.. - \bar{y}... \qquad \hat{b}_j = \bar{y}._j. - \bar{y}... \qquad \hat{\mu}_j = \bar{y}...$$

$a_1 =$ 4,83

$a_2 =$ -4,83

$b_1 =$ 8,56

$b_2 =$ -5,78

$b_3 =$ -2,78

Ověřit v SAS !

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

The GLM Procedure

Class Level Information

Class	Levels	Values
a	2	1 2
b	3	1 2 3

Number of observations 18

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1106.277778	368.759259	0.71	0.5608
Error	14	7250.666667	517.904762		
Corrected Total	17	8356.944444			

R-Square	Coeff Var	Root MSE	y Mean
0.132378	15.92054	22.75752	142.9444

Source	DF	Type IV SS	Mean Square	F Value	Pr > F
a	1	420.5000000	420.5000000	0.81	0.3828
b	2	685.7777778	342.8888889	0.66	0.5312

Aritm. průměr		BLUE / GLM	
142,9444	μ	142,94444	
147,7778	A1	4,8333333	147,778
138,1111	A2	-4,833333	138,111
151,5	B1	8,5555556	151,500
137,1667	B2	-5,777778	137,167
140,1667	B3	-2,777778	140,167

GLM Procedure

Least Squares Means

a	y LSMEAN
1	147.777778
2	138.111111

b	y LSMEAN
1	151.500000
2	137.166667
3	140.166667

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
ZDĚLÁVÁNÍ
polufinancována
ciálním fondem
a státním rozpočtem České republiky

Nevybalancovaný pokus

Analýza množství tuku v mléce u 8 dojnic s vlivem efektů stáda a věku: a_i – stádo (i = 1, 2) ; b_j – věk (j = 1, 2, 3)

		věk		
		b_1	**b_2**	**b_3**
stádo	**a_1**	165 154	136	161
	a_2		115 142 186	112

$$
\begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{131} \\ y_{221} \\ y_{222} \\ y_{223} \\ y_{231} \end{bmatrix}
\approx
\begin{bmatrix} 165 \\ 154 \\ 136 \\ 161 \\ 115 \\ 142 \\ 186 \\ 112 \end{bmatrix}
=
\begin{bmatrix} \mu + a_1 + b_1 + e_{111} \\ \mu + a_1 + b_1 + e_{112} \\ \mu + a_1 + b_2 + e_{121} \\ \mu + a_1 + b_3 + e_{131} \\ \mu + a_2 + b_2 + e_{221} \\ \mu + a_2 + b_2 + e_{222} \\ \mu + a_2 + b_2 + e_{223} \\ \mu + a_2 + b_3 + e_{231} \end{bmatrix}
=
\begin{array}{c}
\begin{array}{cccccc} \mu & a_1 & a_2 & b_1 & b_2 & b_3 \end{array} \\
\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}
\end{array}
\cdot
\begin{bmatrix} \mu \\ a_1 \\ a_2 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}
+
\begin{bmatrix} e_{111} \\ e_{112} \\ e_{121} \\ e_{131} \\ e_{221} \\ e_{222} \\ e_{223} \\ e_{231} \end{bmatrix}
$$

y ≈ **y =** **Xb + e** **= X** **. b** **+ e**

b = ?

2. disperzní (**variančně kovarianční, VCV**) **matice** pozorování:

Předpoklad: každý náhodný efekt e_{ijk} je vybrán ze základního souboru s nulovým průměrem a variancí např. 30 kg

$$
\mathbf{V_e} = \sigma^2 \mathbf{I} =
\begin{bmatrix}
\sigma^2_{e_1} & \sigma_{e_1e_2} & \sigma_{e_1e_3} & . & . & . \\
\sigma_{e_2e_1} & \sigma^2_{e_2} & \sigma_{e_2e_3} & . & . & . \\
\sigma_{e_3e_1} & \sigma_{e_3e_2} & \sigma^2_{e_3} & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & .
\end{bmatrix}
=
\begin{bmatrix}
30 & & & & & & & \\
& 30 & & & & & & \\
& & 30 & & & & & \\
& & & 30 & & & & \\
& & & & 30 & & & \\
& & & & & 30 & & \\
& & & & & & 30 & \\
& & & & & & & 30
\end{bmatrix}
= 30.
\begin{bmatrix}
1 & & & & & & & \\
& 1 & & & & & & \\
& & 1 & & & & & \\
& & & 1 & & & & \\
& & & & 1 & & & \\
& & & & & 1 & & \\
& & & & & & 1 & \\
& & & & & & & 1
\end{bmatrix}
$$

$= 30\, I_8 = 30I = \sigma^2 I$

- Maticový zápis:
  - je méně názorný než data zapsaná v tabulce
  - ALE je kratší a úplnější než model skalární
  - musí se definovat matice X (Ta však při větším objemu dat může nabývat velikých rozměrů – nutná výkonná výpočetní technika a softwarové zázemí)

The GLM Procedure

The GLM Procedure

Class Level Information

Class	Levels	Values
a	2	1 2
b	3	1 2 3

Number of observations 8

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	748.575000	249.525000	0.27	0.8465
Error	4	3733.300000	933.325000		
Corrected Total	7	4481.875000			

R-Square	Coeff Var	Root MSE	y Mean
0.167023	20.87130	30.55037	146.3750

Source	DF	Type IV SS	Mean Square	F Value	Pr > F
a	1	198.4500000	198.4500000	0.21	0.6687
b	2	283.4500000	141.7250000	0.15	0.8638

Aritm. průměry		BLUE / GLM	
146,375	μ		145,867
154,00	**A1**	6,3	**152,167**
138,75	**A2**	-6,3	**139,567**
159,50	**B1**	7,33	**153,200**
144,75	**B2**	2,03	**147,900**
136,50	**B3**	-9,37	**136,500**

GLM Procedure

Least Squares Means

a	y LSMEAN
1	152.166667
2	139.566667

b	y LSMEAN
1	153.200000
2	147.900000
3	136.500000

Řešení nejmenších čtverců pro zobecněný lineární model (GLM)

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

$$(\mathbf{y} - \mathbf{Xb})` (\mathbf{y} - \mathbf{Xb}) = \mathbf{e}`\mathbf{e}$$

$$\mathbf{y}`\mathbf{y} - 2(\mathbf{Xb})`\mathbf{y} + (\mathbf{Xb})`\mathbf{Xb} = \mathbf{e}`\mathbf{e}$$

derivace s ohledem, že b = 0 → *normální rovnice*

$$(\mathbf{X}`\mathbf{X})\ \mathbf{b} = \mathbf{X}`\mathbf{y}$$

$$\mathbf{b} = (\mathbf{X}`\mathbf{X})^{-1}\ \mathbf{X}`\mathbf{y} \qquad (V = I\ \sigma^2_E)$$

Modifikace (Jsou-li pozorování korelovaná a nemají-li stejné variance)

$$(\mathbf{X}`\mathbf{V}^{-1}\mathbf{X})\ \mathbf{b} = \mathbf{X}`\mathbf{V}^{-1}\mathbf{y}$$

$$\mathbf{b} = (\mathbf{X}`\mathbf{V}^{-1}\mathbf{X})^{-1}\ \mathbf{X}`\mathbf{V}^{-1}\mathbf{y} \qquad (V = V)$$

Řešení poslední rovnice se nazývá řešení „*zobecněných nejmenších čtverců*“ → minimalizuje **e`e**.

Biometrické odhady genetických parametrů

Problémy aplikace kvantitativní genetiky na populace zvířat jsou ve skutečnosti problémy **statistických odhadů**

Šlechtění je založeno na **znalosti genetické struktury populací**, kterou *zatím* pro kvant. vlastnosti nelze určovat přímo (frekvence alel a genotypů)

$\Rightarrow$ nutné analyzovat efekty, příčiny genetické a prostřeďové, které se podílejí na celkové proměnlivosti

2 parametrů $\Rightarrow$ **variance** a **kovariance**.

Realizace

odhad PH jedince (**OPH**)

(*Estimate of Breeding Value* – **EBV**)

odhad genotypových hodnot skupin jedinců

- který z odhadů je nejlepší odhad ?!?

Nejlepší odhady

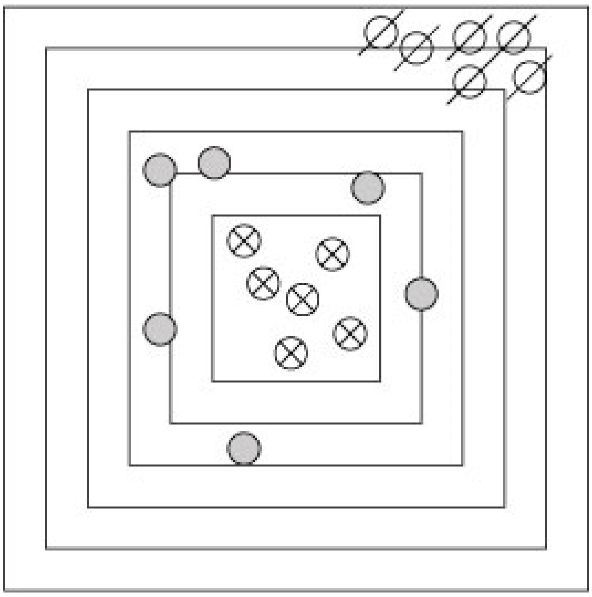
BLUE	***B****est* ***L****inear* ***U****nbiased* ***E****stimators* - nejlepší lineární nevychýlené odhady (nejmenších čtverců)
Nejlepší - *Best*	- **nejlepší odhad průměru populace** = náhodný vzorek (reprezentativní, dostatečný počet), pak je nejlepším odhadem - **nejlepší odhad PH** - souhrnná PH = vložit do selekčního indexu, který hodnotí všechny PH pro všechny hodnocené vlastnosti; nejlepším odhadem je hodnota, která maximalizuje genetický zisk - **minimální variance** = metodou nejmenších čtverců (metoda odhadu), které minimalizují varianci, tyto odhady jsou nejlepší, ale i nestranné (nevychýlené) a lineární
Využíváme:	**lineární modely** – každý odhad je počítán jako lineární kombinace pozorovaných hodnot **nevychýlený** – při opakovaném odhadu je střední hodnota odhadu identická se skutečnými parametry

odhad $\hat{b}$ **je nevychýleným parametrem b, když** $E(\hat{b}) = b$

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Nevychýlenost (vyrovnanost) a přesnost (variabilita)

- (model terče)



- nepřesná (vychýlená) s nízkou variabilitou

- přesná (nevychýlená) s velkou variabilitou

- přesná (nevychýlená) s nízkou variabilitou
- nejlepší odhad

⇒ použít metodu BLUE - metoda odhadu nejmenších čtverců s pevnými efekty

Nejlepší předpovědi

BLUP	- ***B****est* ***L****inear* ***U****nbiased* ***P****rediction* - **n**ejlepší **l**ineární **n**evychýlená **p**ředpověď **NLNP** (metoda nejmenších čtverců) - metoda odhadu nejmenších čtverců náhodných nebo smíšených modelů
smíšený model: mnohovlastnostní (multitrait)	$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$
	X, **Z** – incidenční matice, udávající, které efekty jsou obsaženy v pozorování **b** – vektor obsahující všechny fixní efekty (fixní genetické rozdíly a systematické vlivy prostředí) **u** – vektor všech náhodných systematických efektů (stádo, rok, sezóna); obsahuje také OPH **e** – náhodné nesystematické zbytkové efekty
Metody	*Metoda nejmenších čtverců (LS) nebo zobecněných nejmenších čtverců (GLM), metoda maximální věrohodnosti (ML) nebo restringované maximální věrohodnosti (REML)*

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Způsob řešení pro výběr odhadců je mnoho

Ve šlechtění se v současné době využívá metoda

- nejmenších čtverců (*least square* – LS)
- zobecněných nejmenších čtverců (*generalized least square* – GLM)
- metoda maximální věrohodnosti (*maximum likelihood* – ML)
- či její modifikovaná metoda restringované maximální věrohodnosti (REML)

Metody založené na ML

Maximum Likelihood (ML)

REstricted Maximum Likelihood (REML)

maximilizuje pravděpodobnost pozorovaných dat daných parametrů

nebalancovaná data

komplexní rodokmenová struktura (matice příbuznosti)

simultánní korekce pro fixní efekty

Vyžaduje známou distribuci (normální)

Odhady jsou nevychýlené a jsou vždy v parametrovém prostoru

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Funkce hustoty pravděpodobnosti normálního rozdělení:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

Očekávané průměry E(y) = Xb a var(y) = V

Logaritmus věrohodnostní funkce:

$$L(b,V \mid X, y) = -\tfrac{1}{2} N \log(2\pi) - \tfrac{1}{2}\log(|V|) \boxed{- \tfrac{1}{2}(y - Xb)' V^{-1}(y - Xb)}$$

Rovnice dává pravděpodobnost parametrů (*b, V*) daných dat (*X, y*)

Na pravé straně

první dva výrazy jsou očekávané hodnoty

poslední výraz je součet čtverců

První derivace: $\delta(\log L)/\delta \mathbf{b} = -2\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b})$

Derivace = 0 $\hat{\mathbf{b}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$ **Stejné jako pro LS odhady**

Příklad algoritmu REML

1 Řešení rovnic smíšeného modelu s a priory hodnotou komponent variance (poměr)

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'Y \end{bmatrix}$$

2 Řešení komponent variance z MME

$$\sigma_a^2 = [\hat{a}' A^{-1} a + tr(A^{-1}C)\sigma_e^2] / q$$

$$\sigma_e^2 = [y'y - \hat{b}'X'y - \hat{a}'Z'y] / (N - r(X))$$

Nové λ (= σ_e^2/σ_a^2) a iterovat mezi 1 a 2