METODY STATYSTYCZNE W BIOLOGII

1. Wykład wstępny
2. Populacje i próby danych
3. Testowanie hipotez i estymacja parametrów
4. Planowanie eksperymentów biologicznych
5. Najczęściej wykorzystywane testy statystyczne
6. Podsumowanie materiału, wspólna analiza przykładów, dyskusja
7. Regresja liniowa
8. Regresja nieliniowa
9. Określenie jakości dopasowania równania regresji liniowej i nieliniowej
10. Korelacja
11. Elementy statystycznego modelowania danych - EDA
12. Porównywanie modeli
13. Analiza wariancji
14. Analiza kowariancji
15. Podsumowanie materiału, wspólna analiza przykładów, dyskusja

Wstęp

1. Najlepszy model

2. Porównywanie modeli zagnieżdżonych
   - Likelihood Ratio Test
   - Deviance

3. Porównywanie modeli niezagnieżdżonych
   - Akaike Information Criterion
   - Bayesian Information Criterion

najlepszy model

najlepszy model

- modele statystyczne opisują zmienność danych
- jakość modelu określa prawdopodobieństwo jego dopasowania do danych
- zadaniem modelu jest uogólnienie fluktuacji danych = eliminacje efektów błędu

najlepszy model

przykładowe modele

1. m.ciała = m + płeć + e → prawdopodobieństwo
2. m.ciała = m + płeć + wiek + e → prawdopodobieństwo
3. m.ciała = m + płeć + wiek + sport + e → prawdopodobieństwo
4. m.ciała = m + płeć + wiek + sport + palenie + e → prawdopodobieństwo

najlepszy model

najwyższe prawdopodobieństwo

1. m.ciała = m + płeć + e → prawdopodobieństwo
2. m.ciała = m + płeć + wiek + e → prawdopodobieństwo
3. m.ciała = m + płeć + wiek + sport + e → prawdopodobieństwo
4. m.ciała = m + płeć + wiek + sport + palenie + e → prawdopodobieństwo

najlepszy model

najlepszy model

1. m.ciała = m + płeć + e → prawdopodobieństwo
2. m.ciała = m + płeć + wiek + e → prawdopodobieństwo
3. m.ciała = m + płeć + wiek + sport + e → prawdopodobieństwo
4. m.ciała = m + płeć + wiek + sport + palenie + e → prawdopodobieństwo

???

Modele Zagnieżdżone

porównanie modeli zagnieżdżonych

Likelihood Ratio Test, Test Ilorazu Wiarygodności → modele liniowe

1. Modele zagnieżdżone:
   - model prosty / model złożony
   - model złożony jest rozwinięciem modelu prostego
   - np. m.ciała = m + płeć + e
     m.ciała = m + płeć + wiek + e

2. Likelihood Ratio Test:
   - LRT = -2 [ln Pr(M0) - ln Pr(M1)]
   - $\sim \chi^2_{m2-m1}$

Likelihood Ratio Test - modele liniowe

1. m.ciała = m + płeć + e → ln Pr(M1)= -6.91
2. m.ciała = m + płeć + wiek + e → ln Pr(M2)= -2.30
3. m.ciała = m + płeć + wiek + sport + e → ln Pr(M3)= -0.22
4. m.ciała = m + płeć + wiek + sport + palenie + e → ln Pr(M4)= -0.11

Likelihood Ratio Test - modele liniowe

1. m.ciała = m + płeć + e
   → ln Pr(M1)= -6.91

2. m.ciała = m + płeć + wiek + e
   → ln Pr(M2)= -2.30 → LRT=-2*(-6.91 + 2.30)=9.21 → α_t=0.0024

3. m.ciała = m + płeć + wiek + sport + e
   → ln Pr(M3)= -0.22 → LRT=-2*(-2.30 + 0.22)=4.16 → α_t=0.0414

4. m.ciała = m + płeć + wiek + sport + palenie + e
   → ln Pr(M4)= -0.11 → LRT=-2*(-0.22 + 0.11)=0.24 → α_t=0.6274

Likelihood Ratio Test - modele liniowe

1. m.ciała = m + płeć + e
   → ln Pr(M1)= -6.91

2. m.ciała = m + płeć + wiek + e
   → ln Pr(M2)= -2.30 → LRT=-2*(-6.91 + 2.30)=9.21 → α_t=0.0024

3. m.ciała = m + płeć + wiek + sport + e
   → ln Pr(M3)= -0.22 → LRT=-2*(-2.30 + 0.22)=4.16 → α_t=0.0414

4. m.ciała = m + płeć + wiek + sport + palenie + e
   → ln Pr(M4)= -0.11 → LRT=-2*(-0.22 + 0.11)=0.24 → α_t=0.6274

porównanie modeli zagnieżdżonych – modele nieliniowe

Analysis of Deviance - modele nieliniowe

nacisk	ilość całkow.	ilość uszkod.
2500	50	10
2700	70	17
	...	
4300	65	51

1. Badanie wytrzymałości złącz w samolotach
2. Zastosowano różne siły nacisku

porównanie modeli zagnieżdżonych

Analysis of Deviance - modele nieliniowe

1. Deviance: $D_M = -2[\ln(\Pr_M) - \ln(dane)]$
2. Modele zagnieżdżone:
   - np. logit (prawdopodobieństwo uszkodzenia) = β_0 + nacisk + e

$$D_{M1} = -2\sum_{i=1}^{K}\left[y_i \ln\left(\frac{y_i}{\hat{y}_{iM1}}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{y}_{iM1}}\right)\right]$$

logit (prawdopodobieństwo uszkodzenia) = β_0 + nacisk + maszyna + e

$$D_{M2} = -2\sum_{i=1}^{K}\left[y_i \ln\left(\frac{y_i}{\hat{y}_{iM2}}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{y}_{iM2}}\right)\right]$$

2. Porównanie modeli:
   - D_{M1}-D_{M2}:

$$-2\sum_{i=1}^{K}\left[y_i \ln\left(\frac{\hat{y}_{iM1}}{\hat{y}_{iM2}}\right) + (n_i - y_i)\ln\left(\frac{n_i - \hat{y}_{iM1}}{n_i - \hat{y}_{iM2}}\right)\right] \sim \chi_{M1-M2}$$

Likelihood Ratio Test - modele nieliniowe

Analysis of Deviance - modele nieliniowe

1. Modele zagnieżdżone:

- np. logit(p) = β_0 + nacisk + e

D_{M1}=151.02 8 st.sw.

logit(p) = β_0 + nacisk + maszyna + e

D_{M2}=105.18 5 st.sw.

→ D_{M1}-D_{M2}=45.84 → α_t<0.0000000001

logit(p) = β_0 + nacisk + maszyna + pracownik + e

D_{M3}=53.44 2 st.sw.

→ D_{M2}-D_{M3}=97.58 → α_t<0.0000000001

Likelihood Ratio Test - modele nieliniowe

Analysis of Deviance - modele nieliniowe

1. Modele zagnieżdżone:

- np. $\text{logit}(p) = \beta_0 + \text{nacisk} + e$

$D_{M1}=151.02$ 8 st.sw.

$\text{logit}(p) = \beta_0 + \text{nacisk} + \text{maszyna} + e$

$D_{M2}=105.18$ 5 st.sw.

→ $D_{M1}-D_{M2}=45.84$ → $\alpha_t<0.0000000001$

$\text{logit}(p) = \beta_0 + \text{nacisk} + \text{maszyna} + \text{pracownik} + e$

$D_{M3}=53.44$ 2 st.sw.

→ $D_{M2}-D_{M3}=97.58$ → $\alpha_t<0.0000000001$

modele niezagnieżdżone

porównanie modeli niezagnieżdżonych

1. Modele niezagnieżdżone:
   - nie można wyróżnić modelu prostego i złożonego
   - np. m.ciała = m + płeć + e

     m.ciała = m + wiek + e

2. Kryteria wyboru modelu
   - Liczba efektów modelu
   - Prawdopodobieństwo dopasowania modelu
   - Procedura nieparametryczna - brak rozkładu

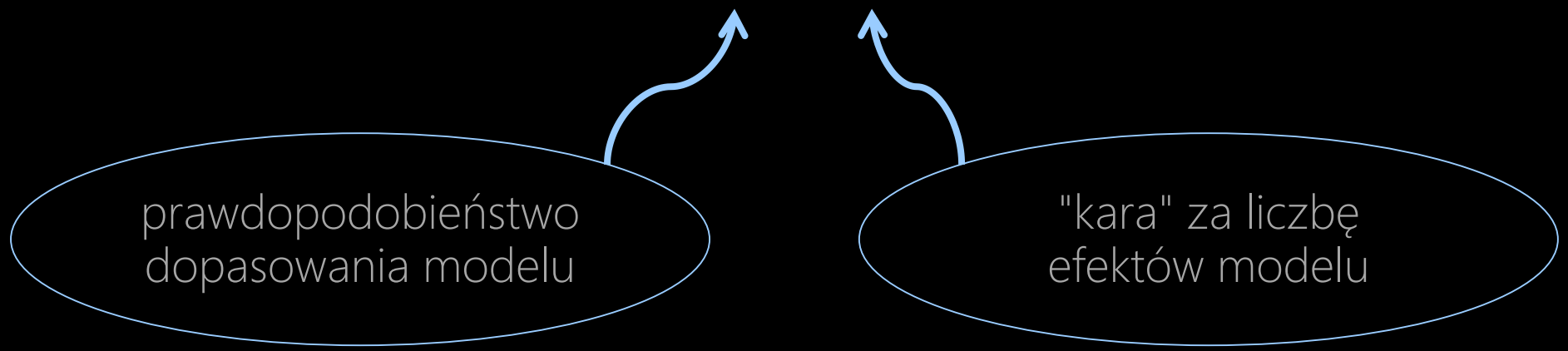
porównanie modeli niezagnieżdżonych

- Akaike Information Criterion

$$AIC = -2\ln(\Pr_M) + 2k$$

- Bayesian Information Criterion

$$BIC = -2\ln(\Pr_M) + k\ln(n)$$



porównanie modeli niezagnieżdżonych

prawdopodobieństwo	ln(p)	k	k2	AIC	k ln(30)	BIC
0.001	-6.91	2	4	17.82	6.80	20.62
0.100	-2.30	7	14	18.61	23.81	28.41
0.230	-1.47	3	6	8.94	10.20	13.14
0.310	-1.17	2	4	6.34	6.80	9.14
0.550	-0.60	5	10	11.20	17.01	18.20
0.670	-0.40	4	8	8.80	13.60	14.41
0.850	-0.16	3	6	6.33	10.20	10.53
0.860	-0.15	6	12	12.30	20.41	20.71
0.900	-0.11	11	22	22.21	37.41	37.62

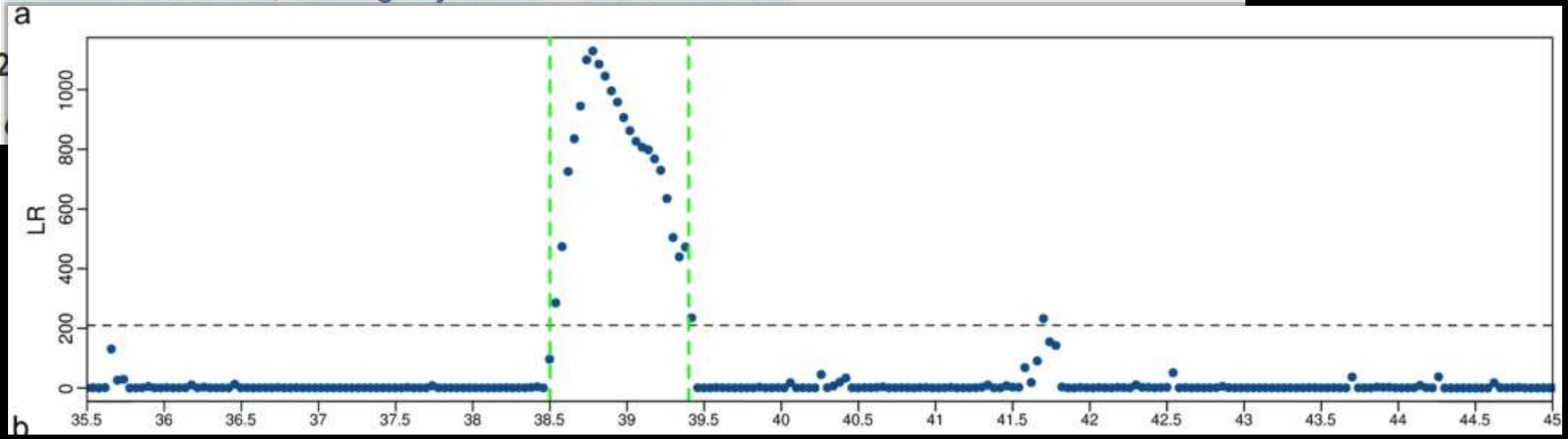
porównanie modeli przykład

Research article | Open Access | Published: 08 January 2020

Assessing genomic diversity and signatures of selection in Original Braunvieh cattle using whole-genome sequencing data

Meenu Bhati ✉, Naveen Kumar Kadri, Danang Crysnanto & Hubert Pausch

BMC Genomics 2



porównanie modeli przykład

Methodology article | Open Access | Published: 19 October 2018

Big data analysis of human mitochondrial DNA substitution models: a regressi

Keren Levinstein Hallak, Shay Tzur & Saharon Rosset

BMC Genomics **19**, Article number: 759 (2018) | Cite this ar

Model #	Codon Position	Direction of Replication	CG pair	Codon	Amino Acid	Right Neighbor	Left Neighbor	Genes	Aggregated Genes	Nucleotide	Response	# of Models	AIC NB	AIC Poisson
1	÷	+	−	÷	−	+	+	−	−	−	All	192	29352	38662
2	÷	+	−	÷	−	+	+	−	+	−	All	192	29407	37931
3	÷	÷	−	+	−	÷	÷	−	+	−	All	18	29411	39079
4	÷	÷	−	+	−	÷	÷	+	−	−	All	18	29430	38413
5	÷	+	−	÷	−	−	−	−	−	−	All	192	29431	39918
6	÷	+	+	÷	−	−	−	−	−	−	All	192	29431	39927
7	÷	÷	÷	+	−	÷	÷	−	+	−	All	36	29447	38739
8	÷	+	÷	÷	−	+	+	−	−	−	All	249	29463	38458
9	÷	÷	÷	+	−	÷	÷	+	−	−	All	36	29469	37853
10	÷	+	−	÷	−	−	−	−	+	−	All	192	29473	39119
11	÷	+	+	÷	−	−	−	−	+	−	All	192	29483	39129
12	÷	÷	−	+	−	−	−	+	−	−	All	6	29491	40374
13	÷	÷	÷	+	−	−	−	+	−	−	All	18	29507	39634
14	÷	÷	−	÷	−	+	+	−	−	−	All	350	29514	38593
15	÷	÷	−	+	−	−	−	−	+	−	All	6	29515	40771
16	÷	÷	÷	+	−	−	−	−	+	−	All	18	29518	40412
17	÷	+	+	+	−	÷	÷	−	+	−	All	9	29540	40060
18	÷	+	÷	÷	−	−	−	−	−	−	All	249	29544	39733
19	÷	+	−	+	−	÷	÷	−	+	−	All	9	29546	40142
20	÷	+	+	+	−	÷	÷	+	−	−	All	9	29547	39360

1. Najlepszy model

2. Porównywanie modeli zagnieżdżonych
   - Likelihood Ratio Test
   - Deviance

3. Porównywanie modeli niezagnieżdżonych
   - Akaike Information Criterion
   - Bayesian Information Criterions