„Elementy statystyki jako narzędzie sterowania jakością i wynikami badań”

Marek Karwański, Szkoła Główna Gospodarstwa Wiejskiego

Agenda.

1. Wstęp – dwie kultury modelowania danych.
2. Modele efektów stałych i mieszanych.
3. Przykład analizy danych laboratoryjnych.
4. Rozwiązania I, oparte na prostym modelu regresyjnym.
5. Rozwiązanie II, model dla danych powiązanych.
6. Metody estymacji i ich wpływ na wyniki analiz.
7. Ocena dopasowania modeli – kryterium AIC.
8. Podsumowanie.

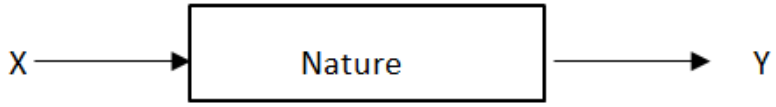
Modelowanie statystyczne: dwie kultury

Statistical Science 2001, Vol. 16, No. 3, 199–231

Wstęp

Statystyka zaczyna się od danych. Pomyśl o danych jako generowanych przez czarną skrzynkę, w której wektor zmiennych wejściowych x (zmienne niezależne) wchodzi z jednej strony, a z drugiej strony wychodzi zmienna odpowiedzi y.

Wewnątrz czarnej skrzynki funkcje „natury” łączą zmienne predykcyjne ze zmiennymi odpowiedzi, więc obraz wygląda tak:



Analiza danych ma dwa cele:

- **Prognoza**. Aby móc przewidzieć, jakie będą reakcje na przyszłe zmienne wejściowe;
- **Informacja**. Aby wydobyć informacje o tym, jak „natura” łączy zmienne odpowiedzi ze zmiennymi wejściowymi.

Istnieją dwa różne podejścia :

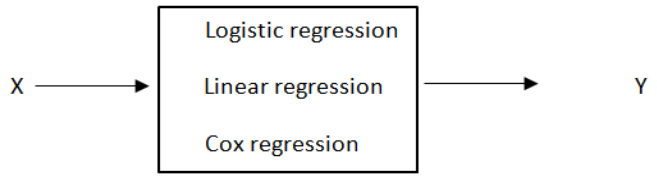
Kultura modelowania statystycznego

Analiza w tej kulturze rozpoczyna się od projektu stochastycznego modelu danych dla wnętrza czarnej skrzynki.

Na przykład powszechnym modelem danych jest to, że dane są generowane przez niezależne losowania:

$$\boldsymbol{odpowiedź} = \boldsymbol{f}(\boldsymbol{zmienne\ predykcyjne}, \boldsymbol{szum\ losowy}, \boldsymbol{parametry})$$

Wartości parametrów są szacowane na podstawie danych i modelu, a następnie wykorzystywanego do informacji i/lub przewidywania. W związku z tym czarne pole jest wypełniane w następujący sposób:



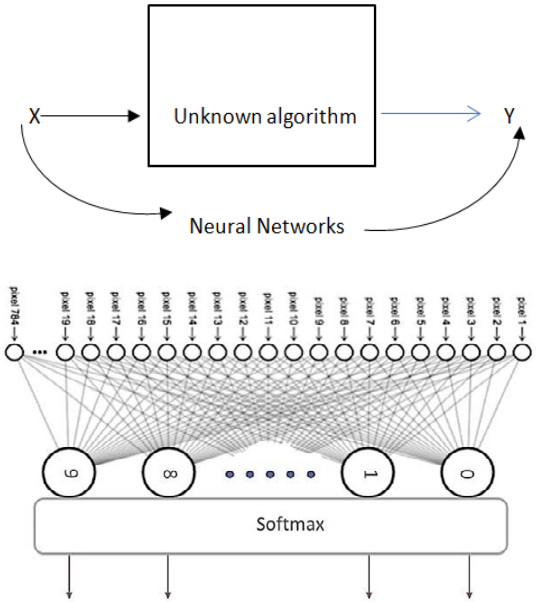
Walidacja modelu.

Tak – nie przy użyciu testów dopasowania i badania resztowego.

Kultura modelowania algorytmicznego

Analiza w tej kulturze rozważa wnętrze pudła złożonego i nieznanego. Podejście polega na znalezieniu funkcji $f(x)$ — algorytmu, który operuje na x, aby przewidzieć odpowiedzi y.

Czarna skrzynka wygląda tak:



Walidacja modelu.

Mierzone przez dokładność predykcyjną.

Model efektów stałych

Możemy zapisać model LM (Linear Model) jako: $Y_i \sim \mathcal{F}_\psi$

gdzie $\mathcal{F}_\psi$ jest rodziną rozkładów normalnych o wartości średniej w postaci funkcji scoringowej:

$$\begin{aligned} y_1 &= \mu + \alpha_1 x_{11} + \alpha_2 x_{12} + \cdots + \alpha_p x_{1p} + e_1 \\ y_2 &= \mu + \alpha_2 x_{21} + \alpha_2 x2_2 + \cdots + \alpha_p x_{2p} + e_2 \\ &\vdots \quad \vdots \qquad\qquad\qquad \vdots \\ y_n &= \mu + \alpha_1 x_{n1} + \alpha_2 x_{n2} + \cdots + \alpha_p x_{np} + e_n \end{aligned}$$

oraz wariancji:

$$\begin{aligned} var(e_1) &= \sigma^2 \\ &\vdots \quad \vdots \\ var(e_n) &= \sigma^2 \end{aligned}$$

W zapisie wektorowym:

$$\boldsymbol{y} = \boldsymbol{X\alpha} + \boldsymbol{e} \qquad , \qquad \boldsymbol{e} \sim \boldsymbol{N(0,V)}$$
$$\boldsymbol{V} = var(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}$$

Model efektów mieszanych

Model mieszany (GLM) rozszerza model efektów stałych o uwzględnienie efektów losowych, czyli losowych współczynników i/lub składników kowariancji w macierzy wariancji.

Rozszerzając nasz model efektów stałych (α) o efekty losowe (β), model mieszany można określić jako

$$y_i = \mu + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_p x_{ip} + \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_p z_{ip} + e_i$$

W zapisie wektorowym (dwa źródła błędów losowych):

$$\boldsymbol{y} = \boldsymbol{X\alpha} + \boldsymbol{Z\beta} + \boldsymbol{e} \qquad \boldsymbol{e} \sim N(0,R)\ , \quad \boldsymbol{\beta} \sim N(0,G)$$

.

Możemy zapisać modele LM/GLM (Linear Model / General Linear Model) jako:

Model efektów stałych (LM)

Efekty stałe tworzą model regresyjny:

$$\mu_i = \mu + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p + e_i$$

$$Y_i|\alpha_i \sim \mathcal{N}(\mu_i, \sigma^2 \mathbb{I}) \quad \text{czyli } (\sim \mathcal{F}_\varphi)$$

gdzie y_i jest odpowiedzią dla i-tej jednostki próbki, a α_i jest wektorem efektów stałych dla tej jednostki

Model efektów mieszanych (GLM)

Mieszane liniowe modele (LMM) mają następującą ogólną reprezentację:

$$\begin{cases} Y_i|\beta_i \sim \mathcal{F}_\psi \\ \beta_i \sim \mathcal{N}(0, \Sigma) \end{cases}$$

ψ jest wektorem rozkładu Y , β_i jest wektorem efektów losowych dla tej jednostki.

Przykład z dziedziny gazometrii medycznej.

Postawowe pojęcia medyczne wprowadzające do problemu:

- Wymiana gazowa w płucach – przenoszenie tlenu O2 z powietrza oddechowego do krwi (utlenowanie krwi) oraz CO2 z krwi do powietrza wydychanego (eliminacja CO2).

- Gazometria krwi tętniczej pozwala ocenić skuteczność wymiany gazowej dzięki pomiarom ciśnienia parcjalnego O2 i CO2.

- Ciśnienie parcjalne (PaO2 i PaCO2) opisuje udział poszczególnych gazów w mieszaninie gazowej. Zmiany ciśnienia opisują dynamikę gdyż gazy przemieszczają się zgodnie z gradientem ciśnień z obszarów o wyższym ciśnieniu parcjalnym do obszarów o niższym ciśnieniu.

- Na poziomie błony pęcherzykowo-włośniczkowej powietrze w pęcherzykach płucnych wykazuje wyższe PaO2 i niższe PaCO2, w porównaniu do krwi włośniczkowej. O2 przemieszcza się z pęcherzyków płucnych dokrwi, a CO2 z krwi do pęcherzyków płucnych. Zaburzenia w tym mechanizmie prowadzą do niewydolności organizmu.

Badanie gazometryczne (ang. blood gases) wykonywane jest na ogół łącznie z badaniem innych parametrów krytycznych koniecznych do ustalenie stopnia niewydolności organizmu

Analiza korelacji będzie narzędziem statystycznym stosowanym w tym kontekście. **Policzymy współczynnik korelacji** między dwiema zmiennymi PaO2 (ciśnienie parcjalne dwutlenku węgla) i HCO3 (aktualne stężenie dwuwęglanów – odpowiadające utlenowaniu krwi).

Przykład z dziedziny gazometrii medycznej

Dane z badania							
Patient	Replicate	PaCO2	HCO3	Patient	Replicate	PaCO2	HCO3
1	1	3.97	4.48	5	3	4.32	6.85
1	2	4.12	4.99	5	4	3.23	6.75
1	3	4.09	5.39	5	5	4.46	6.37
1	4	3.97	6.09	5	6	4.72	7.44
2	1	5.27	4.49	5	7	4.75	7.5
2	2	5.37	4.55	5	8	4.99	9.51
2	3	5.41	5.66	6	1	4.78	4.63
2	4	5.44	7.68	6	2	4.73	5.91
3	1	5.67	5.31	6	3	5.12	6.82
3	2	3.64	6.2	6	4	4.93	5.49
3	3	4.32	6.79	6	5	5.03	7.84
3	4	4.73	6.39	6	6	4.93	8.35
3	5	4.96	7.29	7	1	6.85	3.93
3	6	5.04	6.95	7	2	6.44	4.53
3	7	5.22	9.53	7	3	6.52	5.25
3	8	4.82	8.91	8	1	5.28	6.27
3	9	5.07	9.36	8	2	4.56	6.22
4	1	5.67	6.82	8	3	4.34	6.24
4	2	5.1	6.48	8	4	4.32	6.47
4	3	5.53	7.5	8	5	4.41	6.18
4	4	4.75	5.97	8	6	3.69	7.37
4	5	5.51	7.49	8	7	6.09	6.95
5	1	4.28	6.19	8	8	5.58	8.94
5	2	4.44	6.43				

MODEL STATYSTYCZNY

Interesujące zmienne oznaczmy jako U i W.

Niech (U_{ij}, W_{ij}) będzie j-tą obserwacją u i-tego pacjenta (j=1, ...,m_i) w próbie n pacjentów.

Załóżmy, że para (U_{ij}, W_{ij}) ma dwuwymiarowy rozkład normalny,

$$\begin{bmatrix} U_{ij} \\ W_{ij} \end{bmatrix} \sim Normal\left(\begin{bmatrix} \mu_U \\ \mu_W \end{bmatrix}, \Sigma\right)$$

gdzie μ_U , μ_W są średnimi odpowiednio U i W, a Σ jest macierzą wariancji/kowariancji. Możemy nakładać różne warunki na macierz Σ .

Macierz wariancji/kowariancji możemy zapisać:

$$\Sigma = \begin{bmatrix} \sigma_U^2 & \sigma_U \sigma_W \rho_{UW} \\ \sigma_U \sigma_W \rho_{UW} & \sigma_W^2 \end{bmatrix}$$

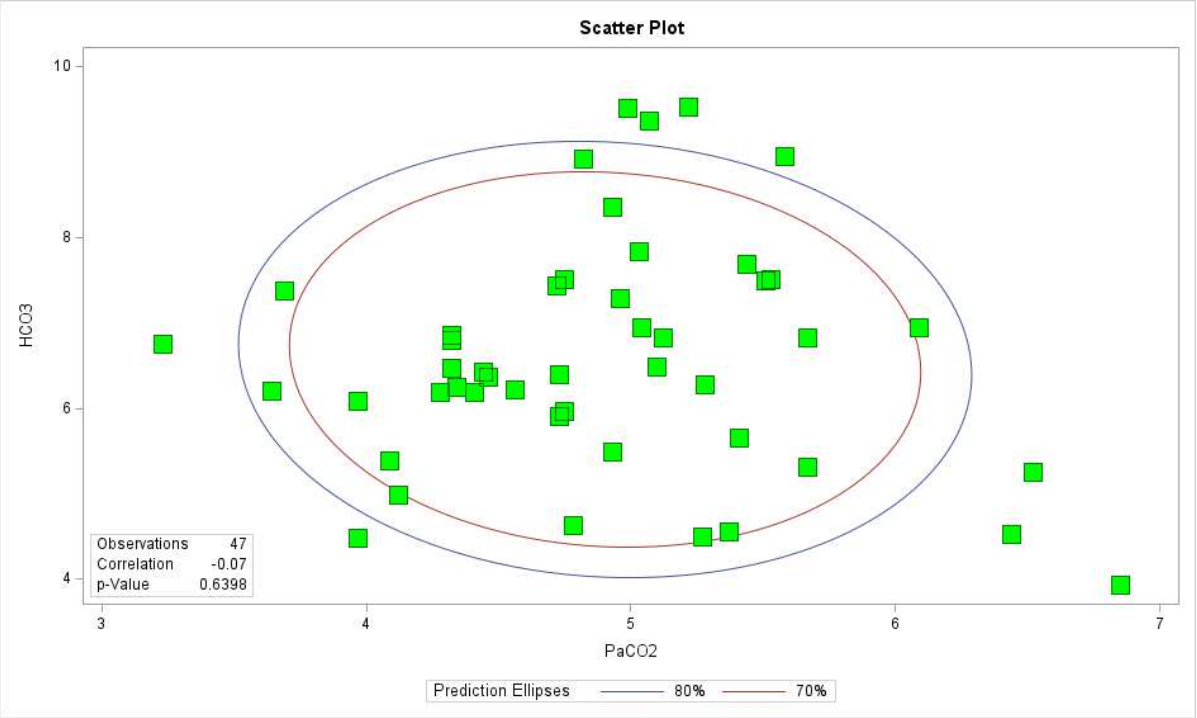
gdzie σ_U^2 i σ_W^2 są wariancjami odpowiednio U i W, a ρ_{UW} ich korelacją .

Dla wygody notacji przydatne będzie używanie terminu kowariancji $\sigma_{UW} \equiv \sigma_U \ \sigma_W \rho_{UW}$.

Współczynnik korelacji liczony bezpośrednio:

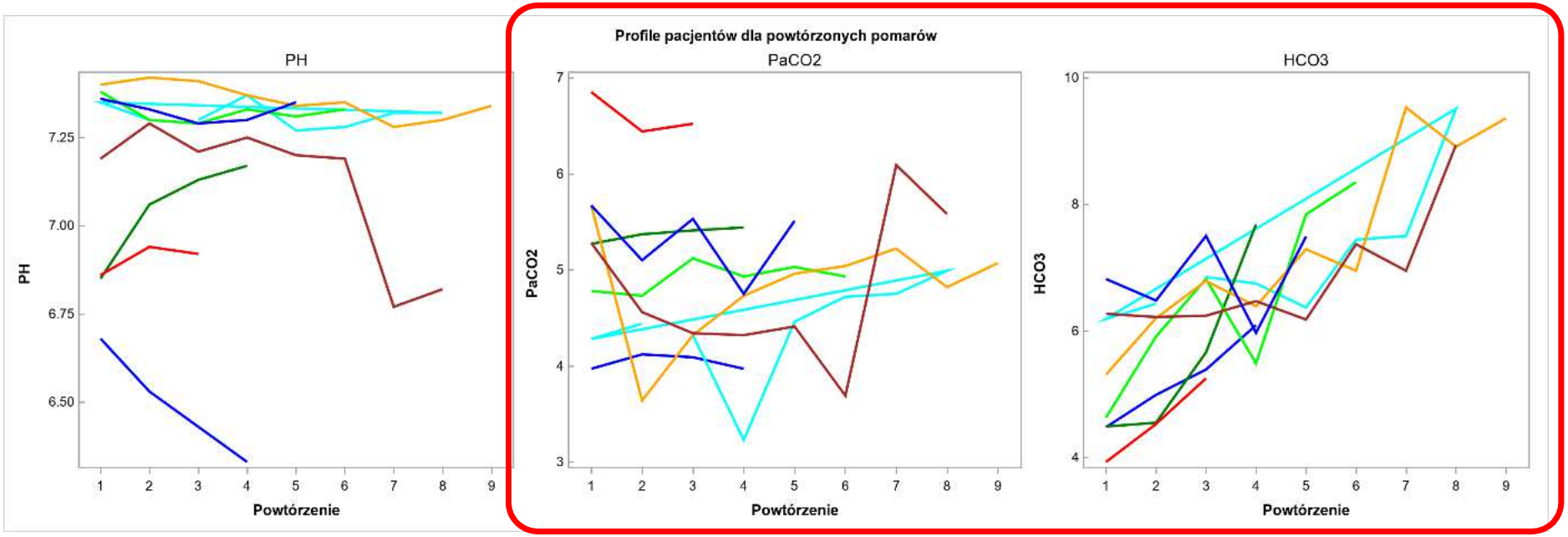
```
Title "Naive Pearson correlation";
proc corr;
 var PaCO2 HCO3;
run;
```

Parametr	Estymator
Naive Pearson correlation	**-0.07006**
Pearson correlation based on means	-0.48794
Weighted correlation based on means	-0.43842



Rysunek poniżej przedstawia wyniki zmiennych PH, PaCO2 i HCO3 dla każdego pacjenta.

Zachowania tych krzywych są różne dla każdej osoby.



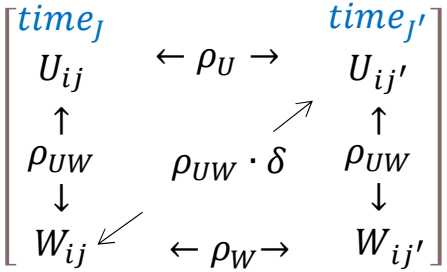
Model dla powiązanych danych (GLM)

Pełna specyfikacja modelu wymaga dalej założeń dotyczących relacji między U i W mierzonych w różnych momentach czasu (replikacjach). Oznaczmy korelacje między pomiarami wykonanymi w dwóch różnych czasach $j \ oraz \ j'$ takich że $j \neq j'$ przez:

$$
\begin{aligned}
corr(U_{ij'}, U_{ij}) &= \rho_U \\
corr(W_{ij'}, W_{ij}) &= \rho_W \\
corr(U_{ij}, W_{ij}) &= \rho_{UW} \\
corr(U_{ij}, W_{ij'}) &= \rho_{UW} \cdot \delta
\end{aligned}
$$

Spodziewamy się, że parametr δ będzie różny od 1, gdyż zazwyczaj korelacje między zmiennymi pomierzonymi w różnych momentach mają inną wartość niż pomierzonymi w tym samym czasie.

Zakładana struktura korelacji może być przedstawiona w następujący sposób:



Aby lepiej pokazać strukturę wariacyjno/kowariancyjną, dla uproszczenia przedstawiony jest zapis pełnej macierzy wariancji/kowariancji dla 2 powtórzonych pomiarów u i-tego pacjenta:

$$V_i = cov\begin{bmatrix} U_{i1} \\ W_{I1} \\ U_{i2} \\ W_{i2} \\ \vdots \\ U_{im_i} \\ W_{im_i} \end{bmatrix} = \begin{bmatrix} \sigma_U^2 & \sigma_{UW} & \sigma_U^2\rho_U & \sigma_{UW}\delta & \dots & \sigma_U^2\rho_U & \sigma_{UW}\delta \\ \sigma_{UW} & \sigma_W^2 & \sigma_{UW}\delta & \sigma_W^2\rho_W & \dots & \sigma_{UW}\delta & \sigma_W^2\rho_W \\ \sigma_U^2\rho_U & \sigma_{UW}\delta & \sigma_U^2 & \sigma_{UW} & \vdots & \sigma_U^2\rho_W & \sigma_{UW}\delta \\ \sigma_{UW}\delta & \sigma_W^2\rho_W & \sigma_{UW} & \sigma_W^2 & \vdots & \sigma_{UW}\delta & \sigma_W^2\rho_W \\ \vdots & \vdots & \dots & \dots & \ddots & \vdots & \vdots \\ \sigma_U^2\rho_U & \sigma_{UW}\delta & \sigma_U^2\rho_U & \sigma_{UW}\delta & \dots & \sigma_U^2 & \sigma_{UW} \\ \sigma_{UW}\delta & \sigma_W^2\rho_w & \sigma_{UW}\delta & \sigma_W^2\rho_W & \dots & \sigma_{UW} & \sigma_W^2 \end{bmatrix}$$

Proszę zwróć uwagę na strukturę blokową tej macierzy, z pod-macierzami odpowiadającymi Σ na głównej przekątnej oraz pod-macierzami (innymi) poza przekątną.

Macierze wariancji/kowariancji dla każdego pacjenta będą miały taką samą strukturę, z wyjątkiem tego, że wymiar może się różnić z powodu różnych liczb powtórzeń pomiarów.

Dla i-tego pacjenta liniowy model efektów mieszanych zadany będzie przez równanie:

$$Y_i = X_i\beta + Z_i\gamma_i + \epsilon_i$$

Gdzie X_i i Z_i są odpowiednio stałą i losową macierzą planu, β jest wektorem stałych efektów, γ_i jest wektorem losowego efektu dla pacjenta i oraz ϵ_i jest błędem losowym dla pacjenta i.

Przyjrzyjmy się teraz dokładniej wpływom obu macierzy na rozkład Y_i:

$$E(Y_i) = \mu_i = X_i\beta \qquad oraz \qquad Var(Y_i) = V_i = Z_i G Z_i^T + R_i$$

Zadanie polega na oszacowaniu β, G, R_i dla zadanej struktury wariancji/kowariancji na podstawie danych obserwowanych.

```
model value = pomiar / ddfm=kr;
  random pomiar /type=un subject=patient;                    #(macierz G)
  repeated pomiar / type=un subject=Replicate(patient);      #(macierz R)
```

	Patient	Replicate	pomiar	value
1	1	1	PaCO2	3.97
2	1	1	HCO3	4.48
3	5	3	PaCO2	4.32
4	5	3	HCO3	6.85
5	1	2	PaCO2	4.12
6	1	2	HCO3	4.99
7	5	4	PaCO2	3.23
8	5	4	HCO3	6.75
9	1	3	PaCO2	4.09
10	1	3	HCO3	5.39
11	5	5	PaCO2	4.46
12	5	5	HCO3	6.37
13	1	4	PaCO2	3.97
14	1	4	HCO3	6.09
15	5	6	PaCO2	4.72
16	5	6	HCO3	7.44
17	2	1	PaCO2	5.27
18	2	1	HCO3	4.49
19	5	7	PaCO2	4.75
20	5	7	HCO3	7.5
21	2	2	PaCO2	5.37

Table 1 Covariance Structure Examples

Description	Structure	Example
Variance components	VC (default)	$\begin{bmatrix} \sigma_B^2 & 0 & 0 & 0 \\ 0 & \sigma_B^2 & 0 & 0 \\ 0 & 0 & \sigma_{AB}^2 & 0 \\ 0 & 0 & 0 & \sigma_{AB}^2 \end{bmatrix}$
Compound symmetry	CS	$\begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{bmatrix}$
Unstructured	UN	$\begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$

ML estymator:

Estimated V Matrix for Patient 1								
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8
1	1.9656	-0.1956	0.5753	-0.3725	0.5753	-0.3725	0.5753	-0.3725
2	-0.1956	0.6816	-0.3725	0.4259	-0.3725	0.4259	-0.3725	0.4259
3	0.5753	-0.3725	1.9656	-0.1956	0.5753	-0.3725	0.5753	-0.3725
4	-0.3725	0.4259	-0.1956	0.6816	-0.3725	0.4259	-0.3725	0.4259
5	0.5753	-0.3725	0.5753	-0.3725	1.9656	-0.1956	0.5753	-0.3725
6	-0.3725	0.4259	-0.3725	0.4259	-0.1956	0.6816	-0.3725	0.4259
7	0.5753	-0.3725	0.5753	-0.3725	0.5753	-0.3725	1.9656	-0.1956
8	-0.3725	0.4259	-0.3725	0.4259	-0.3725	0.4259	-0.1956	0.6816

Estimated V Correlation Matrix for Patient 1								
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8
1	1.0000	-0.1690	0.2927	-0.3218	0.2927	-0.3218	0.2927	-0.3218
2	-0.1690	1.0000	-0.3218	0.6248	-0.3218	0.6248	-0.3218	0.6248
3	0.2927	-0.3218	1.0000	-0.1690	0.2927	-0.3218	0.2927	-0.3218
4	-0.3218	0.6248	-0.1690	1.0000	-0.3218	0.6248	-0.3218	0.6248
5	0.2927	-0.3218	0.2927	-0.3218	1.0000	-0.1690	0.2927	-0.3218
6	-0.3218	0.6248	-0.3218	0.6248	-0.1690	1.0000	-0.3218	0.6248
7	0.2927	-0.3218	0.2927	-0.3218	0.2927	-0.3218	1.0000	-0.1690
8	-0.3218	0.6248	-0.3218	0.6248	-0.3218	0.6248	-0.1690	1.0000

Wyniki z powyższych tabel zebrane w jedno miejsce:

Parametr	Estymator
μ_U	5.0401
μ_W	6.3612
σ_U^2	1.4019
σ_W^2	0.8255
ρ_U	0.2926
ρ_W	0.6248
ρ_{UW}	**-0.1689**
δ	1.9043

Heurystyka: Metoda Największej Wiarogodności (Maximum Likelihood)

$$Y \sim P(X, \theta)$$

Oszacowanie parametrów θ metodą ML polega na znalezieniu:

$$\hat{\theta}_{ML} = \arg\max_{\theta} P(X, \theta)$$

Funkcja prawdopodobieństwa tych modeli ma następującą ogólną postać

$$Lik(\Theta) = \prod_{i=1}^{n} \int p_1(y_i|\gamma_i \psi;) p_2(\gamma_i; \Sigma)$$

$\mathcal{F}_{\psi}$

$\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$

w którym pierwszy składnik jest funkcją prawdopodobieństwa lub funkcji gęstości prawdopodobieństwa $F\psi$ a drugi składnik jest funkcją gęstości prawdopodobieństwa wielowymiarowego rozkładu normalnego dla efektów losowych.

Problem polega na tym, że całka w definicji tej funkcji prawdopodobieństwa nie ma rozwiązania w formie zamkniętej (analitycznej). Dlatego, aby oszacować parametry w tych modelach przy pomocy metody maksymalnego prawdopodobieństwa, musimy w jakiś sposób przybliżyć tę całkę.

Metody estymacji: PQL (Partial Quasi Likelihood), REML (REstricted Maximum Likelihood), ML (Maximum Likelihood), Laplace, Gauss-Hermite są ze sobą powiązane?

Pytanie z bloga.

- Podczas uczenia się o mieszanych uogólnionych liniowych modelach (Generalized Linear Mixed Models - GLMM) często widzę powyższe warunki.

- Czasami wydaje mi się, że są to oddzielne metody oceny (ustalonych? Losowych? Obu?) efektów, ale kiedy czytam literaturę, widzę, że terminy są mieszane. Na przykład PQL i REML. Niektórzy piszą, że PQL działa dobrze w przypadku nienormalnych rozkładów odpowiedzi warunkowej, takiej jak log-normal, ale jest obciążona w klasycznym przypadku dwumianowym lub Poissona, dlatego powinienem użyć wówczas REML lub ML.

- W innych artykułach widzę, że REML lub ML jest używany tylko w modelach liniowych, podczas gdy inne artykuły mówią, że REML jest teraz dostępny również dla GLMM (na przykład pakiet glmmTMB w R-Cran). Rozumiem więc, że PQL jest metodą niezależną od REML? Ale potem widzę książkę, w której porównują różne metody szacowania, w tym PQL przez REML. Czy zatem REML jest specjalnym przypadkiem PQL? A gdzie w grę wchodzi Laplace lub Gauss-Hermitte? Jestem całkowicie zagubiony.

- Czy istnieje sposób na zorganizowanie tych metod?

W literaturze proponowane są dwie główne klasy aproksymacji oraz metoda dwustopniowa REML.

1. Przybliżenie funkcji całkowanej:
   metody te wymagają przybliżenia logarytmu iloczynu dwóch składników $p(y_i|b_i;\psi) \cdot p(b_i;\Sigma)$ przez zmienną o wielowymiarowym rozkładzie normalnym, ponieważ dla tego rozkładu możemy rozwiązać całkę. Do tej kategorii należą metody aproksymacji **PQL i Laplace**.

2. Przybliżone szacowanie całki:
   metody te obejmują przybliżenie całej całki przy pomocy (ważonej) sumy, tj.

$$\int p(y_i|b_i;\psi)p(b_i;\Sigma)\,db_i \approx \sum_k w_k(y_i|b_k;\psi)p(b_k;\Sigma)$$

   Należą do niej niektóre metody należące do kategorii Monte Carlo i adaptacyjne aproksymacje **kwaratury Gaussa np. metoda Gaussa-Hermitta**.

3. **Metoda REML (Restricted Maximum Likelihood)** opierająca się na koncepcji maksymalizacji wiarogodności jako funkcji wariancji Σ_y przy założeniu, że parametr μ jest stały, następnie szacowany jest parametr μ przy założeniu, że wariancja jest stała.

Zalety i wady

- Metody przybliżania funkcji podcałkowej są ogólnie szybsze niż przybliżania całkowania. Nie zapewniają one jednak żadnej kontroli błędu aproksymacji. Z tego powodu metody te działają lepiej, gdy iloczyn dwóch terminów może być dobrze przybliżony przez wielowymiarowy rozkład normalny. Jak to ma miejsce gdy dane są „bardziej" ciągłe. Występuje to, w danych o rozkładach ciągłych a także dwumianowych z dużą liczbą prób i danych Poissona z dużymi oczekiwanymi wartościami.
- Przybliżanie metodą całkowania jest wolniejsze, ale zapewniają kontrolę błędu aproksymacji poprzez użycie większej liczby składników w wyrażeniu sumowania. Można to osiągnąć biorąc pod uwagę większą próbkę Monte Carlo lub więcej punktów kwadraturowych. Dlatego te metody będą działać lepiej w przypadku danych binarnych lub danych Poissona o niskiej oczekiwanej wartości.
- Warto wspomnieć, że istnieją pewne powiązania między dwiema klasami metod. Na przykład przybliżenie Laplace'a jest równoważne adaptacyjnej regule kwadraturowej Gaussa z jednym punktem kwadratury (hiper parametr algorytmu).
- Wreszcie metoda REML jest bardziej przydatna w szacowaniu liniowych modeli mieszanych, dla których całka ma co prawda rozwiązanie w formie zamkniętej, ale głównym celem jest oszacowanie składników wariancji, tj. elementów macierzy kowariancji Σ. Klasyczna procedura największej wiarogodności jest znana z generowania obciążonych wyników do oszacowania tych parametrów, szczególnie w przypadku małych próbek.

  Intuicja REML polega na maksymalizacji zmodyfikowanej funkcji wiarogodności, która jest wolna od komponentu wartości średniej zamiast oryginalnej funkcji wiarogodności, jak w ML.

ML estymator:

Estimated V Matrix for Patient 1								
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8
1	1.9656	-0.1956	0.5753	-0.3725	0.5753	-0.3725	0.5753	-0.3725
2	-0.1956	0.6816	-0.3725	0.4259	-0.3725	0.4259	-0.3725	0.4259
3	0.5753	-0.3725	1.9656	-0.1956	0.5753	-0.3725	0.5753	-0.3725
4	-0.3725	0.4259	-0.1956	0.6816	-0.3725	0.4259	-0.3725	0.4259
5	0.5753	-0.3725	0.5753	-0.3725	1.9656	-0.1956	0.5753	-0.3725
6	-0.3725	0.4259	-0.3725	0.4259	-0.1956	0.6816	-0.3725	0.4259
7	0.5753	-0.3725	0.5753	-0.3725	0.5753	-0.3725	1.9656	-0.1956
8	-0.3725	0.4259	-0.3725	0.4259	-0.3725	0.4259	-0.1956	0.6816

Estimated V Correlation Matrix for Patient 1								
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8
1	1.0000	-0.1690	0.2927	-0.3218	0.2927	-0.3218	0.2927	-0.3218
2	-0.1690	1.0000	-0.3218	0.6248	-0.3218	0.6248	-0.3218	0.6248
3	0.2927	-0.3218	1.0000	-0.1690	0.2927	-0.3218	0.2927	-0.3218
4	-0.3218	0.6248	-0.1690	1.0000	-0.3218	0.6248	-0.3218	0.6248
5	0.2927	-0.3218	0.2927	-0.3218	1.0000	-0.1690	0.2927	-0.3218
6	-0.3218	0.6248	-0.3218	0.6248	-0.1690	1.0000	-0.3218	0.6248
7	0.2927	-0.3218	0.2927	-0.3218	0.2927	-0.3218	1.0000	-0.1690
8	-0.3218	0.6248	-0.3218	0.6248	-0.3218	0.6248	-0.1690	1.0000

Wyniki z powyższych tabel zebrane w jedno miejsce:

Parametr	Estymator
μ_U	5.0401
μ_W	6.3612
σ_U^2	1.4019
σ_W^2	0.8255
ρ_U	0.2926
ρ_W	0.6248
ρ_{UW}	**-0.1689**
δ	1.9043

Metoda estymacji ML (Maximum Likelihood) nie jest najlepszą metodą szacowania komponentów wariacyjnych – zazwyczaj niedoszacowuje ich wartości dlatego lepiej użyć metody REML.

REML estymator:

Estimated V Matrix for Patient 1								
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8
1	2.0892	-0.2434	0.7065	-0.4180	0.7065	-0.4180	0.7065	-0.4180
2	-0.2434	0.7561	-0.4180	0.5015	-0.4180	0.5015	-0.4180	0.5015
3	0.7065	-0.4180	2.0892	-0.2434	0.7065	-0.4180	0.7065	-0.4180
4	-0.4180	0.5015	-0.2434	0.7561	-0.4180	0.5015	-0.4180	0.5015
5	0.7065	-0.4180	0.7065	-0.4180	2.0892	-0.2434	0.7065	-0.4180
6	-0.4180	0.5015	-0.4180	0.5015	-0.2434	0.7561	-0.4180	0.5015
7	0.7065	-0.4180	0.7065	-0.4180	0.7065	-0.4180	2.0892	-0.2434
8	-0.4180	0.5015	-0.4180	0.5015	-0.4180	0.5015	-0.2434	0.7561

Estimated V Correlation Matrix for Patient 1								
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8
1	1.0000	-0.1936	0.3381	-0.3326	0.3381	-0.3326	0.3381	-0.3326
2	-0.1936	1.0000	-0.3326	0.6633	-0.3326	0.6633	-0.3326	0.6633
3	0.3381	-0.3326	1.0000	-0.1936	0.3381	-0.3326	0.3381	-0.3326
4	-0.3326	0.6633	-0.1936	1.0000	-0.3326	0.6633	-0.3326	0.6633
5	0.3381	-0.3326	0.3381	-0.3326	1.0000	-0.1936	0.3381	-0.3326
6	-0.3326	0.6633	-0.3326	0.6633	-0.1936	1.0000	-0.3326	0.6633
7	0.3381	-0.3326	0.3381	-0.3326	0.3381	-0.3326	1.0000	-0.1936
8	-0.3326	0.6633	-0.3326	0.6633	-0.3326	0.6633	-0.1936	1.0000

Metoda estymacji może wpłynąć na wynik oszacowania.

Parametr	Estymatory ML	Estymatory REML
μ_U	5.0401	5.0386
μ_W	6.3612	6.3477
σ_U^2	1.9657	2.0892
σ_W^2	0.6816	0.7561
ρ_U	0.2926	0.3381
ρ_W	0.6248	0.6632
ρ_{UW}	**-0.1689**	**-0.1936**
δ	1.9043	1.7173

Cross Entropy Kullbacka-Leiblera – ocena dopasowania modelu

Informacja Kullbacka-Leiblera (K-L) jest funkcją oznaczoną jako „I" w celach informacyjnych. Funkcja ta ma dwa argumenty: f reprezentuje pełną rzeczywistość lub „prawdę", a g jest modelem.

- Zatem informacja K-L $I(f, g)$ to: „**informacja" tracona** przy zastosowaniu modelu g do przybliżenia pełnej rzeczywistości, f.
- Równoważna i bardzo przydatna interpretacja $I(f, g)$ to: „**odległość" od przybliżającego modelu g** do pełnej rzeczywistości f.

W ramach dowolnej interpretacji. staramy się znaleźć model, który minimalizuje $I(f, g)$ dla hipotezy. reprezentowanej przez modele.

$$I\big(f, g(.|\theta)\big) = \int_{\Omega} f(x) \log\left(\frac{f(x)}{g(x|\theta)}\right) dx$$

Informacja K-L

$$= \int_{\Omega} f(x) \log\big(f(x)\big)\, dx - \int_{\Omega} f(x) \log\big(g(x|\theta)\big)\, dx$$

Hirotagu Akaike podał wzory pozwalające wyliczyć oczekiwaną wartość informacji:

$$E_Y\left[I\left(f, g\left(.\middle|\hat{\theta}(y)\right)\right)\right] \equiv AIC$$

Dla rozkładu normalnego mamy $AIC = n\log\frac{RSS}{n} + 2k$

gdzie RSS – rezydualna suma kwadratów
n - wolumen próbki
k - liczba parametrów modelu

Akaike Information Criterion jest używane do porównywania modeli i wyboru najlepszego

Metoda ML:

Fit Statistics	
-2 Log Likelihood	239.9
AIC (Smaller is Better)	255.9
AICC (Smaller is Better)	257.6
BIC (Smaller is Better)	256.6

Metoda REML:

Fit Statistics	
-2 Res Log Likelihood	241.6
AIC (Smaller is Better)	253.6
AICC (Smaller is Better)	254.6
BIC (Smaller is Better)	254.0

Podsumowanie – wykorzystanie z modeli statytsycznych

Statystycy w badaniach stosowanych uważają modelowanie danych za szablon do analizy statystycznej: w obliczu problemu przede wszystkim myślą o modelu danych. Najważniejszym elementem jest budowa modelu.

Następnie szacowane są parametry i wyciągane wnioski (na podstawie modelu).

Kiedy model jest dopasowany do danych, następuje wyciąganie wniosków ilościowych.

McCullah i Nelder (1989) piszą **„Dane często wskazują z niemal równym naciskiem na kilka możliwych modeli, i ważne jest, aby statystyk uznał to i zaakceptował**". Dobrze powiedziane, ale różne modele, wszystkie jednakowo dobre, mogą dawać różne obrazy relacji między predyktorem a zmiennymi odpowiedzi. Pytanie który najdokładniej odzwierciedla dane, jest trudne do odpowiedzi

Mountain i Hsiao (1989) piszą: „**Trudno sformułować kompleksowy model tak aby: obejmował wszystkie konkurujące modele. Ponadto, przy użyciu skończonych próbek, istnieją wątpliwe implikacje w odniesieniu do trafność i moc różnych testów, które opierają się na teorii asymptotycznej**."

Breiman (2003) ostrzega nas o systematycznych błędach (prowadzących do błędnych wniosków), które zostały popełnione przy zastosowaniu aktualnej praktyki statystycznej w modelowaniu danych. Trzeba zawsze pamiętać, że:

- **kierownictwo poszukuje zysku** - praktycznych odpowiedzi (przewidywań) przydatnych przy podejmowaniu decyzji w krótkim okresie.
- **nauka poszukuje prawdy**, podstawowej wiedzy o przyrodzie, która na dłuższą metę zapewnia zrozumienie i kontrolę.

Notatka historyczna: test t-Studenta ma wiele zastosowań naukowych, ale został wymyślony przez Studenta jako narzędzie zarządzania, dzięki któremu stwierdzono, że piwo Guinness jest lepsze (gorzkie?).

Pytania.