

Monitorowanie stanu przeszczepionej nerki i problemy statystyczne związane z tym zagadnieniem

Przemek Biecek

MIMUW



Matematyczne Metody Biologii, 22 X 2008

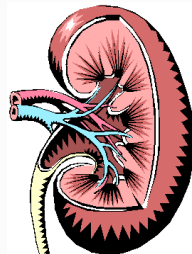
Plan

- Kilka słów o przeszczepie nerki,
- Jak wyglądają nasze dane,
- Próby „ataku” tych danych
 - Standardowe techniki w analizie przeżycia,
 - Konsekwencje korelacji statystyk testowych,
 - Kontrola współczynników FDR lub FWER,
 - Regresja PCR,
- Podsumowanie.

Słów kilka o przeszczepie nerki

Pierwszy przeszczep nerki w Polsce został wykonany w roku 1966. Po 40 latach w Polsce wykonuje się około 1000 przeszczepów rocznie.

Nerkę do przeszczepu wybiera się tak, by zminimalizować ryzyko pojawienia się zjawiska odrzucania przeszczepu. Odrzucanie zachodzi gdy organizm gospodarza uzna przeszczep za obce ciało. Rzadko pojawia się atak przeszczepu na gospodarza (Graft-Versus-Host Disease).



Odrzucanie może pojawić się po kilku dniach lub po kilku miesiącach. Aby zminimalizować ryzyko odrzucenia pacjentom podaje się leki immunosupresyjne (blokujące produkcję limfocytów T i hamujące produkcję przeciwciał).

Kreatynina jest bezwodnikiem kreatyny, występującym w krwi oraz moczu, z którym jest wydalana stanowiąc oprócz mocznika jeden z głównych związków azotowych. Kreatynina to produkt uboczny dla mięśnia, który powstaje w wyniku niechcianej konwersacji kreatyny dawkowanej z zewnątrz w postaci preparatów kreatynowych.

Kreatynina, która znajduje się w krwiobiegu musi zostać usunięta przez nerki. Nadmierne obciążenie nerek na skutek zbyt dużej konwersacji kreatyny do kreatyniny może prowadzić do upośledzenia możliwości filtracyjnych nerek. Produkt uboczny, jakim jest kreatynina nie ma żadnego pozytywnego znaczenia dla mięśni.

Współczynnik przesączania kłębuszkowego (ang. glomerular filtration rate, GFR) - pozwala na ocenę stopnia wydolności nerek.

Współczynnik przesączania kłębuszkowego (ang. glomerular filtration rate, GFR) - pozwala na ocenę stopnia wydolności nerek.

Wzór Cockcroft'a-Gault'a

$$C_{Cr}(ml/min) = \frac{(140 - \text{wiek}) \times m.c.}{(72 \times S_{Cr})} \times W$$

- C_{Cr} - klirens kreatyniny,
- S_{Cr} - stężenie kreatyniny w mg/dl,
- wiek - w latach,
- m.c. - masa ciała w kg,
- W - wskaźnik płci - 0,85 (dla kobiet) lub 1,0 (dla mężczyzn).

Przjrzyjmy się teraz danym

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	plec biorec	wiek biorec	plec dawcy	wiek dawcy	czas obse	rok utraty prze	niezgdn	niezgdn	wagastart	wagaend	MDRD7	MDRD30	MDRD6m	MDRD12m	MDRD24m	MDRD36m	MDRD60m
2	K	56 M		44	5		2	2	67	76	40	49	40	48	44	59	49
3	M	43 M		51	5	2008	3	2	57	64	39	51	50	51	56	65	60
4	M	28 K		36	5		2	2	66	77	66	78	78	78	69	76	17
5	M	31 M		42	5		2	1	68	75	29	50	48	62	53	56	47
6	K	25 K		39	5		3	1	49	56	36	40	47	44	48	56	48
7	M	54 K		62	5		3	1	84	98.3	36	57	40	44	52	44	43
8	M	52 M		44	5		4	0	83	84.5	16	30	48	48	51	52	53
9	K	43 M		29	5.4		2	0	84	98	43	62	60	60	60	68	75
10	M	56 M		31	5.5		2	1	80	92	16	26	54	52	65	69	75
11	M	51 M		55	5.5		1	1	82	97	32	48	48	51	48	48	27
12	K	25 M		47	5.5		3	1	40	48	43	56	68	51	45	56	60
13	M	61 M		53	5.5		3	1	87	88	17	28	40	36	36	35	34
14	K	52 K		52	5.5		0	1	53.9	56.5	26	36	37	38	36	36	33
15	M	36 M		47	5.5		3	0	77	93	16	36	45	44	39	42	38
16	M	49 M		61	5.5		2	0	86	95	27	48	50	44	52	56	46
17	M	32 M		36	5.5	2005	2	0	76	100	48	48	52	52	48	45	48
18	M	43 M		30	5.5		2	0	78	95	15	57	80	80	66	55	26
19	M	38 M		20	5.5		2	0	78	85	16	35	45	41	45	45	42
20	M	61 K		61	5.5		3	0	66	69	17	42	33	35	36	37	37
21	K	32 K		55	5.5		2	0	52	61	16	32	32	29	35	32	39
22	M	53 M		51	5.6		3	0	85.6	84	16	42	44	44	48	44	41
23	M	58 M		65	5.7		3	1	68	75	36	44	41	43	51	51	52
24	K	43 K		48	5.7		2	1	62	65	15	33	48	56	64	65	75
25	M	58 M		44	5.8	2007	2	1	96	120	19	40	48	46	51	54	48
26	M	43 M		24	5.8		2	1	71	80	45	66	72	72	68	69	71
27	M	30 K		53	5.8		1	1	55	53	20	28	42	40	41	45	40
28	K	51 M		36	5.8		2	1	60	64	48	52	48	53	60	60	59
29	M	45 K		57	5.8		4	0	79	79	28	44	47	47	44	44	47
30	K	52 M		46	6		2	2	57	73	21	31	39	39	39	42	53
31	M	65 M		58	6		2	1	79	83	16	42	64	59	59	64	56
32	K	30 K		47	6		1	1	49	51	29	55	44	40	40	37	34
33	M	47 K		52	6		3	1	77	77	16	29	32	32	34	32	33
34	M	50 K		47	6	2007	3	1	58	77	16	39	56	64	53	54	64
35	K	47 K		52	6		1	1	62	52	29	36	43	46	44	44	54
36	M	38 K		30	6	2005	2	1	64	63	46	50	45	56	62	44	29
37	M	58 K		52	6		1	1	61	58	17	20	20	22	21	22	26

K < > H \ dane/

Gotowy

Standardowe techniki w analizie przeżycia

Przedstawimy dwie standardowe techniki:

- krzywe przeżycia,
- Model Coxa.

Przedstawimy te techniki na przykładzie danych nefrologicznych, stosować można je w wielu innych sytuacjach

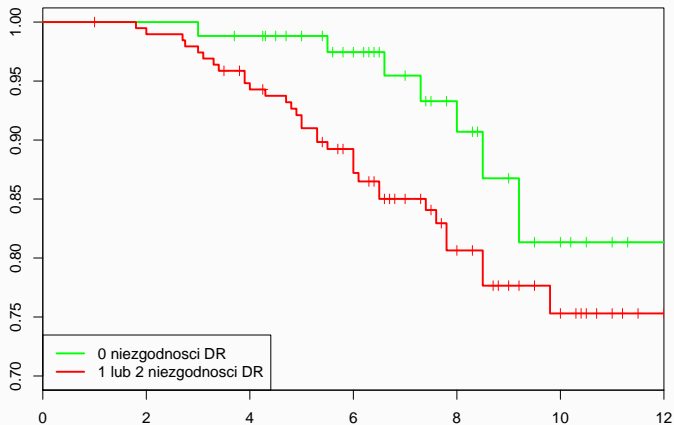
- ekonomia (wypłacalność kredytobiorcy, bankrutowanie firm),
- genetyka (ekspresja genów jako szereg czasowy),
- socjologia (rozpady małżeństw),
- wiele, wiele więcej...

Dane cenzorowane

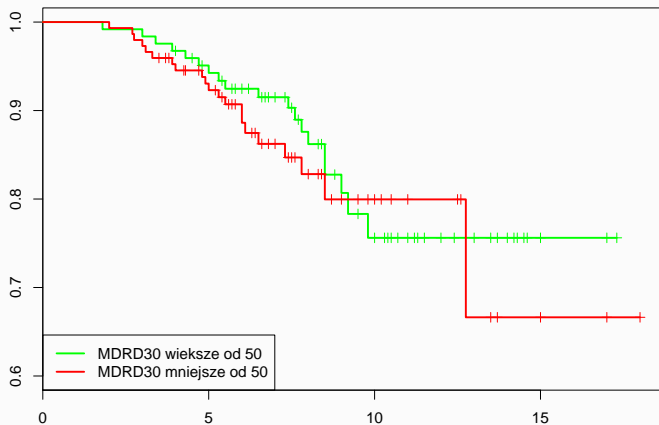
Opisując i porównując czasy przeżycia (czasy utrzymania przeszczepu) nie możemy korzystać ze technik typu regresja liniowa, ANOVA, itp. ponieważ:

- rozkład czasów przeżycia nie ma (zazwyczaj) nic wspólnego z rozkładem normalnym,
- o pewnych obserwacjach mamy jedynie częściową informację (dane cenzorowane), wiemy, że czas był dłuższy od X ale nie wiemy ile dokładnie wynosił.
- standardowe techniki nie pozwalają na operowanie na funkcji hazardu, funkcji skumulowanego hazardu itp.

Funkcja przeżycia, $S(t) = Pr(T > t)$



Funkcja przeżycia, $S(t) = Pr(T > t)$



Test Logrank - porównanie krzywych przeżycia

Nazywany też testem Mantela-Haenszela, wartość statystyki testowej wyznacza się ze wzoru

$$Z = \frac{\sum_{j=1}^J (O_{1j} - E_j)}{\sqrt{\sum_{j=1}^J V_j}}$$

gdzie

$$E_j = O_j \frac{N_{1j}}{N_j},$$

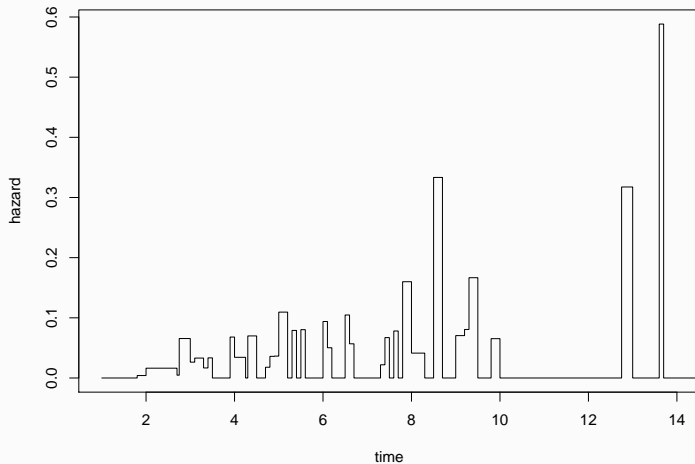
$$V_j = \frac{O_j(N_{1j}/N_j)(1 - N_{1j}/N_j)(N_j - O_j)}{N_j - 1},$$

N_j - liczba obiektów „at risk” w czasie t ,

O_j - liczba „zdarzeń” w czasie t ,

Przy prawdziwej H_0 statystyka Z ma rozkład asympt. $\mathcal{N}(0, 1)$.

Funkcja hazardu, $\lambda(t) = \Pr(t \leq T < t + dt \mid T < t) = -\frac{S'(t) dt}{S(t)}$



Model Coxa przy założeniu proporcjonalnego hazardu

Model dla funkcji hazardu jest następujący

$$h(t) = h_0(t) \exp(X\beta).$$

Dzieląc i logarytmując otrzymujemy

$$\ln \frac{h(t, X)}{h_0(t)} = X\beta,$$

czyli (w miarę) standardowy model regresji. Dla takiego modelu wiemy jak estymować współczynniki β i testować hipotezy dotyczące β .

Jak zaaplikować te metody do naszych danych?

Interesuje nas wpływ MDRD na czas życia nerki, możemy więc:

- Podzielić pacjentów na grupy o wysokim MDRD i o niskim MDRD a następnie testować różnice pomiędzy krzywymi przeżycia.
- Zastosować model regresji Coxa dla każdego MDRD osobno i wybrać model z MDRD lepiej dopasowany do danych.
- Zastosować model regresji Coxa z wszystkimi MDRD w jednym modelu.

Albo:

- Dla różnych MDRD mamy dostępne różne liczby obserwacji, jakiego modelu możemy porównać?
- MDRD są silnie skorelowane, co powoduje problemy, jeżeli uwzględnimy je razem w regresji
- Czy kontrolujemy błąd I rodzaju??

Czy to rzeczywisty problem?

Jak zaaplikować te metody do naszych danych?

Interesuje nas wpływ MDRD na czas życia nerki, możemy więc:

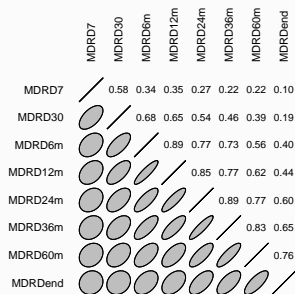
- Podzielić pacjentów na grupy o wysokim MDRD i o niskim MDRD a następnie testować różnice pomiędzy krzywymi przeżycia.
- Zastosować model regresji Coxa dla każdego MDRD osobno i wybrać model z MDRD lepiej dopasowany do danych.
- Zastosować model regresji Coxa z wszystkimi MDRD w jednym modelu.

Ale...

- Dla różnych MDRD mamy dostępne różne liczby obserwacji, takie modele ciężko porównać,
- MDRD są silnie skorelowane co powoduje problemy, jeżeli uwzględnimy je razem w regresji.
- Brak kontroli błędu I rodzaju!!!

Czy to rzeczywisty problem?

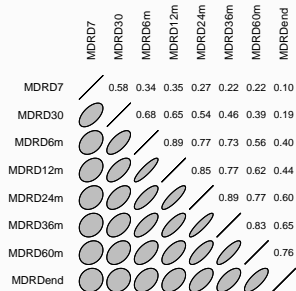
Korelacja MDRD



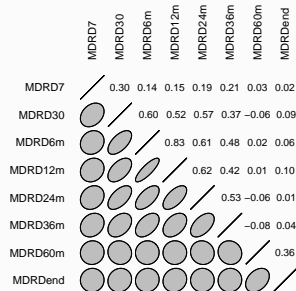
Wartości własne

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
MDRD	5.06	1.33	0.71	0.31	0.26	0.15	0.11	0.08

Korelacja MDRD



Korelacja s. testowych



Wartości własne

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
MDRD	5.06	1.33	0.71	0.31	0.26	0.15	0.11	0.08
S. testowa	3.33	1.37	0.97	0.68	0.64	0.47	0.38	0.15

Standardowa notacja

Wyniki testowania zbioru hipotez można przedstawić następującą tabelką

	#przyjęte H_0	#odrzucone H_0	
#prawdziwe H_0	U	V	m_0
#fałszywe H_0	T	S	m_1
suma	m - R	R	m.

Współczynniki błędów

Per-family error rate (PFER)

$$PFER = E(V).$$

Family-wise error rate (FWER)

$$FWER = Pr(V > 0).$$

False discovery rate (FDR)

$$FDR = E(Q),$$

where

$$Q = \begin{cases} V/R & R > 0, \\ 0 & R = 0. \end{cases}$$

Zmodyfikowana korekta Bonferroniego

W korekcie Bonferroniego zamieniamy liczbę testów na „efektywną liczbę testów”

$$\alpha_0 = \alpha / N^{\text{eff}},$$

α to poziom istotności dla współczynnika FWER a efektywną liczbę testów wyznacza się ze wzoru

$$N^{\text{eff}} = \min_n \max_{\substack{I \subseteq \{1..N\} \\ \#I=n}} \sum_{i \in I} \lambda_i \geq \delta N$$

gdzie λ_i oznacza i tą wartość własną macierzy korelacji statystyk testowych (oczywiście $\sum_{i=1}^N \lambda_i = N$) a $\delta \in [0, 1]$ jest parametrem. Innymi słowy N^{eff} jest minimalną liczbą wartości własnych sumujących się do δN lub liczbą zmiennych wyjaśniających $\delta * 100\%$ zmienności.

Wyniki symulacyjne

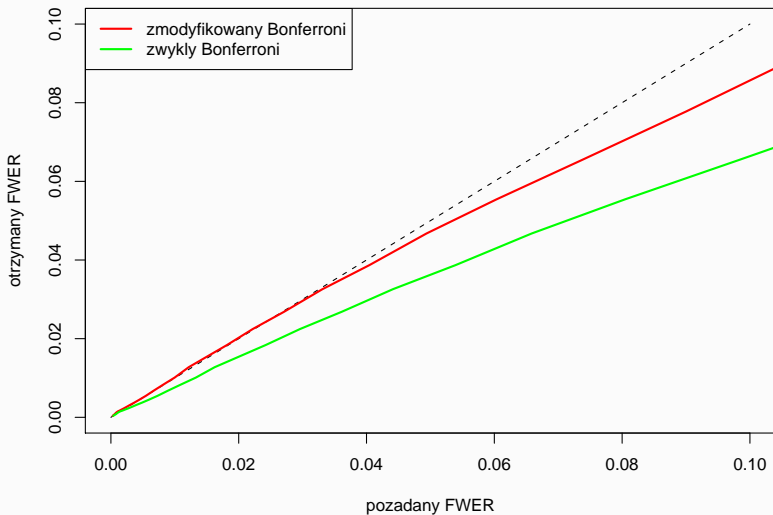
Sprawdźmy jaki wpływ na błąd FWER ma obserwowana korelacja pomiędzy zmiennymi.

Permutując oryginalne dane (bootstrap blokowy) wyznaczymy rozkład brzegowy i łączny statystyk testowych przy prawdziwej H_0 . Wartości MDRD pozostawiamy niezmiennione, permutujemy czasy przeżycia i informacje o utracie przeszczepu.

Znając łączny rozkład p-wartości można ocenić rzeczywisty współczynnik FWER.

Przy badaniu współczynnika FWER wystarczy badać rozkład minimum p-wartości.

Wyniki symulacyjne



Wyniki symulacyjne

Wyniki dla $\alpha = 0.05$.

Stosując korektę Bonferroniego $\alpha_0 = 0.05/8$, otrzymujemy

$$\widehat{FWER} = 0.0363$$

Stosując zmodyfikowaną korektę Bonferroniego $\alpha_0 = 0.05/6$, otrzymujemy

$$\widehat{FWER} = 0.0473$$

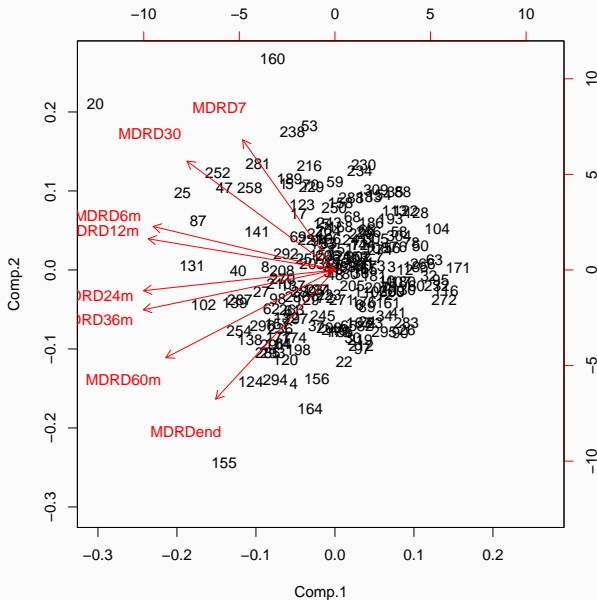
Różnica w błędzie I rodzaju 20%, różnica w mocy jest wyższa (jeżeli test jest nieobciążony to moc jest wypukłą funkcją poziomu istotności).

Składowe główne PCA

```
> loadings(princomp(daneMDRD))
```

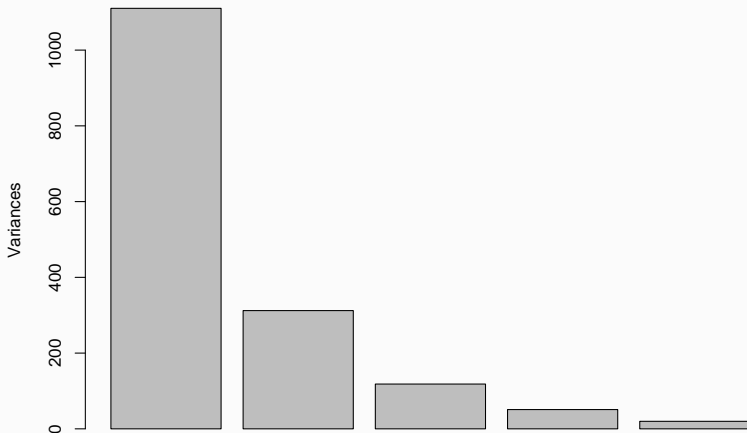
Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
MDRD7	-0.265	0.828	-0.492			
MDRD30	-0.493	0.298	0.776	-0.231		-0.101
MDRD6m	-0.427	-0.123		0.510	0.503	0.537
MDRD12m	-0.452	-0.195	-0.147	0.539	-0.402	-0.532
MDRD24m	-0.390	-0.272	-0.224	-0.384	-0.552	0.521
MDRD36m	-0.384	-0.313	-0.289	-0.497	0.527	-0.382



Regresja PCR i model Cox

Wariancja wyjaśniona przez kolejne składowe



Regresja PCR i model Cox

p-wartości wyznaczone z modelu regresyjnego Coxa

p-wartość	pojedyncza	wspólny model	składowe główne
Zmienna 1	0.0788	0.19	0.0072
Zmienna 2	0.0175	0.69	0.7307
Zmienna 3	0.0086	0.13	0.8798
Zmienna 4	0.0316	0.44	0.5555
Zmienna 5	0.0590	0.26	0.5671
Zmienna 6	0.2505	0.20	0.1559
...

Poziomy istotności dla $\alpha = 0.05$

$$\alpha_0^{standard} = 0.05/8 = 0.00625; \quad \alpha_0^{nowe} = 0.05/6 = 0.00833$$

Pełny model z uwzględnieniem MDRD

Redukcja liczby zmiennych opisujących MDRD ułatwia uwzględnienie w modelu innych zmiennych

	coef	IQR	coef**IQR	p-value
13. wiek.donor	1.01	20	1.30	0.31
20. LN.DR	1.89	1	1.89	0.1
12. tHDterapii	1.01	22	1.23	0.05
57. RR1	1.04	30	3.47	1.1e-05
58. RR2	1.06	10	1.73	0.0027
73. Hb	0.865	3.2	0.628	0.082

Podsumowanie

Wyniki

- Czas do utraty przeszczepu jest w istotny sposób zależny od MDRD,
- Używając PCA można uwzględnić wpływ MDRD wybierając tylko pierwszą składową główną (dzięki temu można budować większe modele z innymi czynnikami takimi jak płeć itp.).

Metodologia

- Dane cenzorowane wymagają specjalnego traktowania,
- Korelacja pomiędzy zmiennymi powinna być uwzględniona przy kontroli współczynników FWER i FDR,
- Technika PCR pozwala na redukcję liczby zmiennych w modelu.

Literatura

- Natasa Rajcic, D. Finkelstein, D. Schoenfeld (2006) „Survival analysis of longitudinal microarrays” Bioinformatics
- E. L. Kaplan, Paul Meier (1959) „Nonparametric Estimation from Incomplete Observations” JASA,
- Mark Stevenson, (2007) „An Introduction to Survival Analysis”.
- H. Martens, (1989) Multivariate calibration. Chichester: Wiley.
- Małgorzata Zynek-Litwin, ..., Przemysław Biecek, M. Klinger, (2008) „Neutrophils degranulation marker - leucocyte elastase-A1 protease inhibitor ... kidney graft injury markers” Transplantation.
- Przemysław Biecek, (2006) „A modified Bonferroni correction for strongly correlated test statistics” Proceedings of the XI National Conference Application of Mathematics to Biology and Medicine.