Metody statystyki medycznej stosowane w badaniach klinicznych

Statistics for clinical research & post-marketing surveillance

<u>część III</u>

Program szkolenia część III

Model regresji liniowej

- Współczynnik korelacji Pearsona a regresja liniowa dwóch zmiennych.
- Zmienne w analizie regresji.
- Założenia modelu regresji.
- Równanie regresji wielorakiej (wieloczynnikowej).
- Dobór zmiennych do modelu regresji.
- Interpretacja parametrów modelu liniowego.
- Testowanie hipotez o współczynnikach regresji.
- Analiza dopasowania modelu regresji.
- Analiza własności rozkładu reszt.
- Graficzna prezentacja równania regresji oraz zależności korelacyjnej.
- Analiza modeli regresji nieliniowej.
- Graficzna prezentacja równania regresji bazującego na postaci nieliniowej.

Program szkolenia część III cd.

Analiza dynamiki zjawisk

- Podstawowe zagadnienia związane z analizą dynamiki.
- Indeksy dynamiki:
  - Indeksy o podstawie stałej.
  - Indeksy o podstawie zmiennej.
- Średnie tempo wzrostu.

Modele trendu oparte na analizie regresji

- Analiza trendu liniowego (regresja liniowa względem czasu).

Model przeżywalności Kaplan-Meier

Model regresji liniowej
(linear regression model)

$$y = a_1 x + a_0 + \xi$$

Gdzie:
x - zmienna niezależna;
y – zmienna zależna;
α_1 – parametr kierunkowy;
α_0 - wyraz wolny;
ξ - składnik losowy (teoretyczny twór w modelu)

Modelowanie oznacza wyznaczenie odpowiedniego opisu matematycznego (*modelu*) będącego odzwierciedleniem badanego procesu.

Model tworzy się za pomocą równań różnego typu: liniowych lub nieliniowych, statycznych lub dynamicznych, ciągłych lub dyskretnych.

Model regresji liniowej

Wyznaczenie modelu oznacza:

1. Wybór typu równania (należy ocenić typ zależności)
2. Wybór jego struktury (*rzędu równania*)
3. Obliczenie wartości współczynników równania (*parametrów modelu* α_i)
4. Przeprowadzenie weryfikacji modelu (*sprawdzenie jego poprawności*).

Praktyczna funkcja modelowania: na podstawie modelu możliwe jest oszacowanie wartości zmiennej **y** na podstawie znanych wartości zmiennej **x**.

Model regresji liniowej

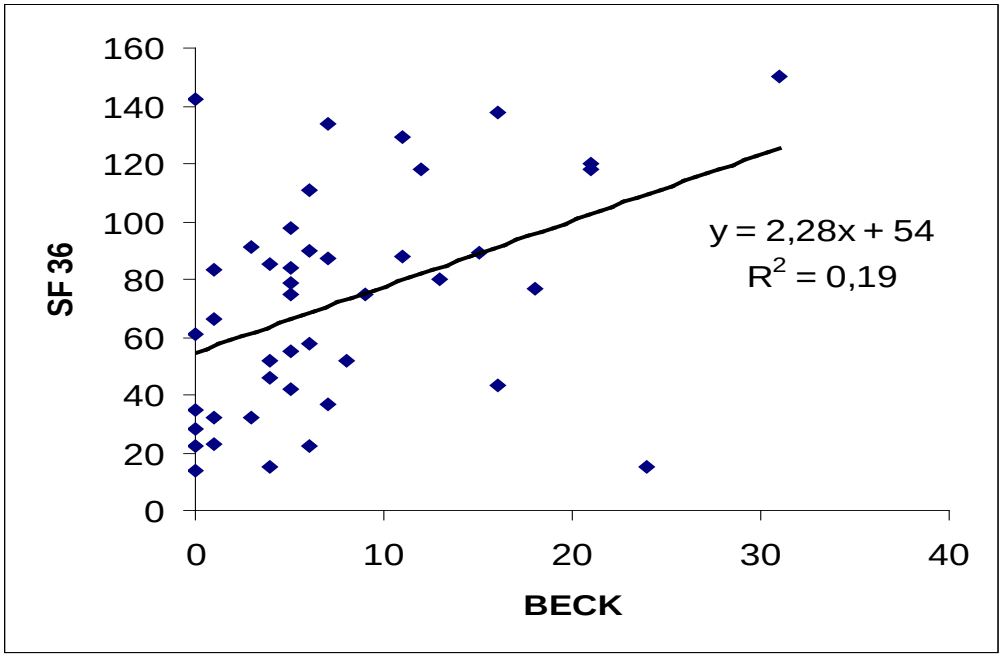
Wykres korelacyjny (wykres rozrzutu punktów empirycznych)

Przykład

Zbadano zależność pomiędzy wiekiem respondentów a osiąganymi przez nich dochodami.

y : skala SF 36 służąca do oceny jakości życia-zmienna zależna (dependent variable)

x : skala BECK do oceny stopnia depresji - zmienna niezależna (independent variable)



Model regresji liniowej

Konieczne jest wyznaczenie parametrów $\mathbf{a_1}$ oraz $\mathbf{a_o}$ tej funkcji.

Wzory:

$$a_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad a_0 = \bar{y} - a_1 \bar{x}$$

Aby obliczyć $\mathbf{a_1}$ z funkcji statystycznych użyj funkcji statystycznych „nachylenie”
Aby obliczyć $\mathbf{a_o}$ z funkcji statystycznych użyj funkcji statystycznych „odcięta”

Powyższe formuły dotyczą metody najmniejszych kwadratów do szacowania funkcji liniowych.
Uwagi: współczynnik korelacji r oraz parametr kierunkowy są zawsze tych samych znaków.

Model regresji liniowej

Przykład

Model:

$$y = 2{,}28x + 54$$

Gdzie:
y : skala SF 36 służąca do oceny jakości życia-zmienna zależna (dependent variable)
x : skala BECK do oceny stopnia depresji - zmienna niezależna (independent variable)

Interpretacja parametrów:

Wzrost skali BECK o jednostkę powoduje, średnio rzecz biorąc wzrost skali SF o 2,28.

Weryfikacja modelu (regresji liniowej)

Miary dopasowania modelu:

- Współczynnik determinacji (R2) {R Square}
  Wartość współczynnika determinacji R2 zawiera się w przedziale <0,1> i informuje jaka część zaobserwowanej, całkowitej zmienności y została wyjaśniona przez model.

- Współczynnik zbieżności ($\phi 2 = 1-R2$)
  Wartość współczynnika zbieżności zawiera się w przedziale <0,1> i informuje jaka część zaobserwowanej, całkowitej zmienności y NIE została wyjaśniona przez model.

- Odchylenie standardowe reszt (SE)
  Wielkość odchylenia standardowego reszt interpretujemy jako przeciętne odchylenie zaobserwowanych wartości zmiennej y od odpowiadających im wartości funkcji (wartości teoretycznych)

Model regresji liniowej

Testy weryfikujące istotność modelu (parametrów modelu)

1. Test istotności modelu

2. Test istotności parametrów

Pierwszy z nich zakłada, że wszystkie parametry modelu, poza wyrazem wolnym są równe zero. Hipoteza H_0 zakłada więc, że model nie jest istotny. Sprawdzianem tej hipotezy jest statystyka F. Dla tej statystyki testowej z tablic rozkładu Fishera odczytuje się poziom istotności (p-value). Zakłada się, że p-value niższe od wartości α (zwykle 0,05 lub 0,01) oznacza odrzucenie hipotezy H_0 (model jest istotny).

Model regresji liniowej

Test istotności parametrów modelu

W badaniu istotności modelu możliwe jest także inne podejście. Zakłada ono osobne badanie poszczególnych parametrów. Badanie istotności w tym przypadku sprowadza się do weryfikacji hipotez:

$$Ho: \alpha_i = 0$$

Czyli hipoteza, że i–ta zmienna nie wpływa w istotny sposób na zmienną endogeniczną. (w niniejszym opracowaniu badano głównie istotność parametru kierunkowego zmiennej czasowej). Hipoteza alternatywna jest sformułowana następująco:

$$H1: \alpha_i \neq 0$$

Model regresji liniowej

Test istotności parametrów modelu

Sprawdzianem testu jest statystyka t-Studenta o n - k stopniach swobody wyznaczana jako:

$$t_i = \frac{a_i - \alpha_i}{D(a_i)}$$

Gdzie:
a_i - ocena i-tego parametru,
α_i - prawdziwa wartość parametru (zgodnie z hipotezą zerową $\alpha i=0$),
$D(a_i)$ – błąd średni szacunku parametru.

W obu typach testów istotności modelu trendu zakłada się że wartość poziomu istotności wyznaczonej dla statystyki testowej niższa od α(np. 0,05 lub 0,01) świadczy o istotności parametru (lub inaczej mówiąc o istotnym wpływie zmiennej X na zmienną Y).

Model regresji liniowej

Przykład

Model prognozujący przeżycie po operacji kardiologicznej w zależności od

Zmienna zależna:

Y – przeżywalność (0-wypis; 1-zgon)

Zmienne niezależne:

X_1 – chirurgia aorty

X_2 powikłania neurologiczne

X3 - pozawałowe VSD

	Współczynniki
Przecięcie	0,019
X1 - chirurgia aorty	0,116
x2 – powikłania neurologiczne	0,262
x3 - pozawałowe VSD	0,613

Interpretacja parametrów:

- Częstsza chirurgia aorty powoduje wyższe ryzyko zgonu pooperacyjnego (dodatni parametr 0,116);
- Powikłania neurologiczne – im częstsze powikłania neurologiczne tym wyższe ryzyko zgonu pooperacyjnego.

Model regresji liniowej

R2 = 0,13 ; 13% zmienności zgonów pooperacyjnych jest wyjaśnione przez model

Su = 0,16
Odchylenie standardowe reszt: wartości teoretyczne (modelowe) odchylają się od wartości empirycznych średnio o +/-0,16.

Badanie istotności parametrów modelu liniowego

Test istotności modelu:

H_0 : R=0 (lub, α_i =0, dla i=1,2)
H_1: R > 0 (lub przynajmniej jedno $\alpha_i \neq 0$)

Statystyka testowa: F=205
Poziom istotności statystyki : p-value < 0,001

Wniosek: Należy odrzucić hipotezę zerową. Model jest statystycznie istotny.

Model regresji liniowej

Test istotności poszczególnych parametrów modelu

Hipotezy

H_0: $\alpha_1 = 0$

H_1: $\alpha_1 \neq 0$

	Współczynnik i	*t Stat*	*Wartość-p*
Przecięcie	0,019	7,153	p<0,001
x1-chirurgia_aorty	0,116	8,925	p<0,001
x2 - powikłania_neurologiczne	0,262	14,554	p<0,001
x3 - pozawałowe VSD	0,613	13,790	p<0,001

Dla wyniku statystyki testowej t_1=8,92 wyznaczany jest poziom istotności p-value. Jest on w tym przypadku niższy niż 0,001. Wynik ten pozwala na odrzucenie hipotezy H_0. można zatem twierdzić, że parametr α_1 jest istotny (zmienna x_1 istotnie wpływa na zmienną y. Można uznać zmienną X1 za niezależny czynnik ryzyka zgonu.

Model regresji liniowej

Przykład

Na podstawie danych oszacuj stosując opcję Regresja model zależności zmiennej Y od zmiennych x1, x2, x3

Narzędzia → Analiza danych → Regresja

Statystyki regresji	
Wielokrotność R	0,37
R kwadrat	0,13
Dopasowany R kwadrat	0,13
Błąd standardowy	0,16
Obserwacje	3999

Model regresji liniowej

Należy zwrócić uwagę, na przedziały ufności dla parametrów (dolne i górne 95%).
Interpretacja : z prawdopodobieństwem 95% przedział ufności o końcach 0,09 oaz 0,141 pokrywa rzeczywistą nieznaną wartość parametru α_1.

	Współczynniki	*Dolne 95%*	*Górne 95%*
Przecięcie	0,019	0,014	0,024
x1 - chirurgia_aorty	0,116	0,090	0,141
x2 - powikłania_neurologiczne	0,262	0,227	0,298
x3 - pozawałowe VSD	0,613	0,526	0,700

Istnieje możliwość zdefiniowania również przedziałów i innym prawdopodobieństwie (wybór w opcji „regresja" innej wartości „poziom ufności")
Dane tego modelu doskonale nadają się do analizy modelem **regresji logistycznej**

Model regresji liniowej

Zadanie 1.

Dysponując danymi w arkuszu CRF zbuduj model regresji zależności poziomu Hb od wieku pacjenta.

- Czy model jest statystycznie istotny? (poziom istotności 0,05).

Zadanie 2.

Dysponując danymi w arkuszu CRF zbuduj model regresji zależności poziomu Hb od wagi pacjenta.

- Czy model jest statystycznie istotny? (poziom istotności 0,05).
- Czy zastosowana dawka leku na wizycie 3 wpływa liniowo na poziomy Hb i Ferrytyny na wizycie 4?

Model regresji liniowej

W przypadku, gdy model regresji budowany jest na podstawie danych szeregu czasowego należy dodatkowo zbadać własności rozkładu reszt modelu.

Własności:

- Rozkład reszt powinien być niezautokorelowany
- Rozkład reszt powinien być losowy
- Rozkład reszt powinien być symetryczny
- Rozkład reszt powinien mieć rozkład normalny

Na podstawie rzetelnie opracowanego modelu regresji możliwe jest symulowanie zachowań. Czyli np.:

Jaka będzie teoretyczna wartość zmiennej Y jeżeli zmienna X osiągnie wartość 15?

Po podstawieniu do modelu regresji za X wartości 15 uzyskamy odpowiedź na tak zadane pytanie badawcze.

Analiza dynamiki zjawisk

Podstawowe zagadnienia związane z analizą dynamiki.

Indeksy dynamiki:

- Indeksy o podstawie stałej (informują o % przyroście w stosunku do jednego okresu bazowego)
- Indeksy o podstawie zmiennej, indeksy łańcuchowe (informują o % przyroście w stosunku do okresu poprzedzającego).

Średnie tempo wzrostu – jest to średnia geometryczna, obliczana zgodnie z formułą:

$$\bar{i} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

Prognozowanie na przyszłe okresy na podstawie średniego tempa zmian:

$$\bar{y}_{t+n} = y_t \cdot i^{n-1}$$

Analiza dynamiki zjawisk

Zadanie 3.

Dysponujemy danymi dotyczącymi poziomu Hb u pacjenta (pomiary wykonywane co tydzień przez okres 2 miesięcy).

- Wyznacz indeks o podstawie stałej (podstawa z tygodnia 1)
- Wyznacz indeks o podstawie stałej (podstawa z tygodnia 5)
- Wyznacz indeks łańcuchowy
- Wyznacz średnie tempo zmian
- Wyznacz przewidywany poziom Hb w tygodniach: 11 oraz 14.

Zadanie 4.

Zbadaj za pomocą indeksów dynamiki zmiany poziomu Ferrytyny u pacjenta oznaczonego jako lp. 5.

Modele trendu

Time series analysis

Model szeregu czasowego, w którym występuje tendencja rozwojowa, wahania sezonowe oraz wahania przypadkowe, a rolę zmiennej objaśniającej odgrywa zmienna czasowa nazywamy **modelem tendencji rozwojowej** (modele trendu).

Zapis modelu jest następujący:

$$y_t = f(t) + g(t) + \varepsilon_t$$

Gdzie:

t = 1, 2, ... ,n

f(t) - funkcja czasu, charakteryzująca tendencję rozwojową szeregu, nazywana funkcją trendu,

g(t) - funkcja opisująca wahania sezonowe,

ε_t - zmienna losowa, charakteryzująca efekty oddziaływania wahań przypadkowych.

Modele trendu

Najczęściej spotykanymi postaciami analitycznymi funkcji trendu są:

- **trend liniowy**

$$y_t = a+b*t$$

Jest to najczęściej wykorzystywana w praktyce postać funkcji trendu, stosowana zawsze w przypadku, gdy można przyjąć założenie o stałych przyrostach wartości zmiennej y w jednostce czasu. Parametry funkcji zostały oszacowane metodą najmniejszych kwadratów.

- **trend logarytmiczny**

$$y_t = a + b * \log t$$

Ten typ trendu wybieramy gdy wzrost badanej zmiennej jest coraz wolniejszy.

Można ten typ modelu oszacować podstawieniem

$t' = \log t$

otrzymujemy równanie liniowe:

$y_t = a + b\, t'$

- **trend wielomianowy**

$$yt = at^2 + bt + c$$

Modele trendu

Zadanie 5.

Oszacuj i zbadaj dopasowanie poszczególnych modeli trendu na podstawie danych liczby ludności Polski

Bazując na tych samych danych oszacuj model liniowy i zbadaj istotność parametrów modelu.

Wyznacz średnie tempo zmian.

Model przeżywalności Kaplan-Meier

Kaplan Meier Survival Analysis

Przykład.

Założenia metody:

- 100 pacjentów obserwowano poprzez okres 5 lat.
- obserwowano zgony pacjentów.
- w trakcie badania część pacjentów „odchodzi” (Became Unavailable (Censored)).

Time	At Risk	Became_Unavailable	Died	Survived
Year 1	100	3	5	
Year 2		3	10	
Year 3		3	15	
Year 4		3	20	
Year 5		3	25	

Time	At Risk	Became Unavailable	Died	Survived
Year 1	100	3	5	95
Year 2	92	3	10	82
Year 3	79	3	15	64
Year 4	61	3	20	41
Year 5	38	3	25	13

Model przeżywalności Kaplan-Meier

Time	At Risk	Became Unavailable	Died	Survived
Year 1	100	3	5	95
Year 2	92	3	10	82
Year 3	79	3	15	64
Year 4	61	3	20	41
Year 5	38	3	25	13

Model przeżywalności Kaplan-Meier

Time Period	At Risk	Became Unavailable (Censored)	Died	Survived	Kaplan-Meier Survival Probability Estimate
Year 1	100	3	5	95	(95/100)=0.95
Year 2	92	3	10	82	(95/100)x(82/92)=0.8467
Year 3	79	3	15	64	(95/100)x(82/92)x(64/79)= 0.70
Year 4	61	3	20	41	(95/100)x(82/92)x(64/79)x (41/61)=0.4611
Year 5	38	3	25	13	(95/100)x(82/92)x(64/79)x (41/61)x(13/38)=0.1577

Model przeżywalności Kaplan-Meier

Model Kaplan – Meier jest powszechnie stosowany również do porównania przeżywalności dwóch lub więcej metod leczenia.

Testy stosowane w modelu Kaplana Meiera do porównania przeżywalności:

- Log Rank (Mantel-Cox)
- Breslow (Generalized Wilcoxon)
- Tarone-Ware

Model przeżywalności Kaplan-Meier

Wyniki testu:

	Oszacowana średnia przeżycia	Błąd Standardowy	95% przedział ufności	
Wielkość guza			Dolna Granica	Górna granica
<= 2 cm	**126,7**	1,3	124,2	129,2
2-5 cm	**108,5**	3,2	102,1	114,8
> 5 cm	**63,2**	9,7	44,1	82,3
Całkowite	123,0	1,3	120,4	125,6

Model przeżywalności Kaplan-Meier

Wyniki testów porównujących przeżywalność w grupach

	Chi-kwadrat	df	Istotność
Log Rank (Mantel-Cox)	32,995	2	0,000 (p<0,001)
Breslow (Generalized Wilcoxon)	31,763	2	0,000 (p<0,001)
Tarone-Ware	33,575	2	0,000 (p<0,001)

Wniosek: testy wskazują na istotność różnic w przeżywalności pomiędzy grupami pacjentów z małym (do 2cm), średnim (2-5cm) oraz dużym rozmiarem nowotworu (powyżej 5cm).

Model przeżywalności Kaplan-Meier

Funkcja hazardu (ryzyka) - przeciwnie do funkcji przeżycia skupia się na pojawieniu niekorzystnego zdarzenia, na przykład śmierci. Przedstawia jakby "negatywne" uzupełnienie informacji niesionej przez funkcję przeżycia. Wartość funkcji hazardu w momencie *t* traktujemy jako chwilowy potencjał pojawiającego się zdarzenia (np. śmierci lub choroby), pod warunkiem że osoba dożyła czasu *t*.