Analiza zależności zmiennych ilościowych – regresja

JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska

Plan wykładu

1. Współczynnik korelacji próbkowej
2. Liniowa zależność między dwoma zmiennymi:
   - Prosta regresja
   - Metoda najmniejszych kwadratów
   - Właściwości
3. Zastosowanie różnego oprogramowania
4. Weryfikacja równania regresji
   - Analiza statystyczna
5. Inne zagadnienia
   - Regresja wieloraka
   - Diagnostyka i obserwacje odstające
   - Regresja nieliniowa

Korelacja próbkowa - przypomnienie

- Współczynnik korelacji liniowej Pearsona:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}} = \frac{C(x,y)}{S_x \cdot S_y}$$

gdzie $\bar{x}, \bar{y}$ - średnie art. zmiennych *x* i *y*, a S_x i S_y ich odchylenia standardowe; kowariancja

$$C(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$r_{xy} \in [-1,1]$$

- Zakres stosowalności: zależność dwóch zmiennych ilościowych o charakterze liniowym.

Własności współczynnika korelacji liniowej Pearsona

1. Miara symetryczna
2. Miara niemianowana i unormowaną
   - Można porównywać korelacje dla różnych zestawów zmiennych
3. Pozwala na określenie nie tylko siły, ale i kierunku zależności między zmiennymi
4. Interpretacja wartości współczynnika korelacji:
   - im $|r_{xy}| \to 1$ tym silniejsza korelacja.
5. Ograniczenia
   - Podatny na obserwacje skrajne (ang. *outliers*)

Przykłady obliczania korelacji

Zbadaj zależność dwóch zmiennych opisujących odpowiedzi respondentów w pewnej ankiecie

- *X* - liczba randek w ostatnim tygodniu
- *Y* – ocena satysfakcji z życia na skali punktowej 1,2,3,..,5

X	1	2	…	5
Y	1	2	…	4

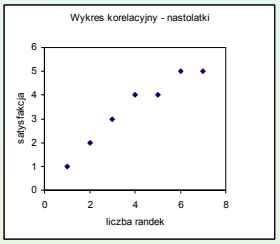
- Dla każdego zestawu odpowiedzi narysuj wykres korelacyjny (każda para wartości odpowiedzi dla jednej osoby przedstawiona jest jako punkt na płaszczyźnie x,y)

Przykłady

Grupa nastolatków

X	1	2	3	4	5	6	7
Y	1	2	3	4	4	5	5

Korelacja = 0.97

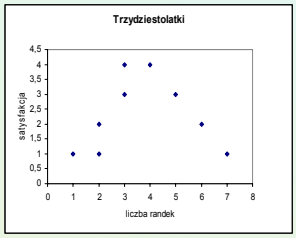


Przykłady 2

Grupa dwudziestoparo-latków

X	1	2	2	3	3	4	5	6	7
Y	1	2	1	3	4	4	3	2	1

Korelacja = ??

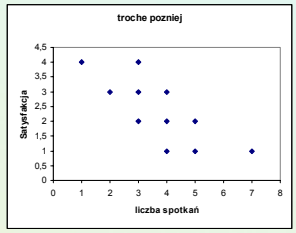


Przykłady 3

Grupa troszkę starszych-latków

X	1	2	3	3	3	4	4	4	5	5	7
Y	4	3	2	3	4	1	2	3	2	1	1

Korelacja = –0.77



Ocena współczynnika korelacji ρ w populacji

- *r* – współczynnik korelacji w próbie → czy może być użyty w odniesieniu do populacji?
- Estymator punktowy?
- Może być także użyty do testowania hipotezy o korelacji zmiennych w populacji.
- Założenia: zmienne (X,Y) populacji mają dwuwymiarowy rozkład normalny o nieznanym współczynniku korelacji ρ. Na podstawie *n* - elementowej próby wyznaczono *r* .
- Testowany układ hipotez:

$$H_0 : \rho = 0 \qquad H_1 : \rho \neq 0$$

- Statystyka testowa:

test $z = \dfrac{r_{xy}}{\sqrt{1-r_{xy}^2}}\sqrt{n}$ lub test $t = \dfrac{r_{xy}}{\sqrt{1-r_{xy}^2}}\sqrt{n-2}$

Przykład testowania istotności współczynnika korelacji

- Współczynnik korelacji między liczbą randek w tygodniu a satysfakcją z życia wynosi *r*=0.493 (N = 16 par pomiarów).
- Czy możemy podjąć decyzję wobec populacji H0: ρ=0.
- Schemat postępowania:
- Testowany układ hipotez:

$$H_0 : \rho = 0 \qquad H_1 : \rho \neq 0$$

- Wybór statystyki testowej

$$t = \frac{r_{xy}}{\sqrt{1-r_{xy}^2}}\sqrt{n-2}$$

- Ma rozkład *t* - Studenta z *n*-2 stopni swobody (14)
- Poziom α=0.05 t_{kryt}=2.145
- *t*=2.11
- Podjęcie decyzji

Problemy w interpretacji współczynnika korelacji

Uwagi po analizie przykładu:

- Należy oglądać dane!
- Współczynnik służy do badania związku liniowego!
- Jeśli związek nie jest liniowy → stosuj regresję krzywoliniową.
- Współczynnik korelacji jest nieistotny → można stwierdzić wyłącznie brak związku liniowego.

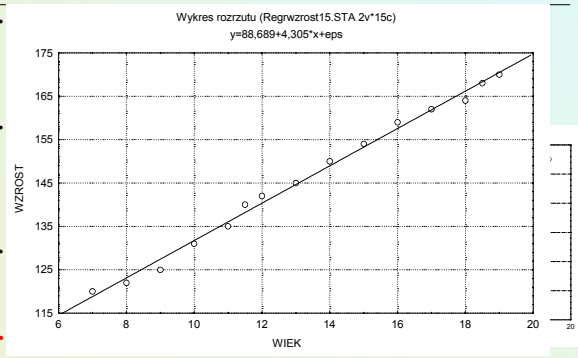
Ponadto pamiętaj:

- Wrażliwość na obserwacje skrajne i ograniczenie zakresy zmienności zmiennej niezależnej.

Gdzie jesteśmy w trakcie wykładu?

1. Współczynnik korelacji próbkowej
2. Liniowa zależność między dwoma zmiennymi:
   - Prosta regresja
   - Metoda najmniejszych kwadratów
   - Właściwości
3. Zastosowanie różnego oprogramowania
4. Weryfikacja równania regresji
   - Analiza statystyczna
5. Inne zagadnienia
   - Regresja wieloraka
   - Diagnostyka i obserwacje odstające
   - Regresja nieliniowa

Analiza regresji



Regresja – model liniowy

- Analityczny sposób przyporządkowania wartości zmiennej zależnej konkretnym wartościom zmiennych niezależnych.
- Liniowa regresja prosta → najprostszy rodzaj regresji, w których zależność zmiennych można opisać za pomocą linii prostej.

$$\hat{y} = a \cdot x + b + \varepsilon$$

gdzie *a* jest *współczynnikiem kierunkowym*, *b* wyraz wolny (punkt przecięcia z osią rzędnych);
x – zmienna niezależna, y – zmienna zależna (objaśniana, przewidywana), ε - błąd losowy.

Założenia modelu regresji

- Związek między *x* i *y* jest liniowy.
- Wartości zmiennej niezależnej nie są losowe. Losowość wartości *y* pochodzi wyłącznie ze składnika losowego.
- Składniki (błędy) losowe mają rozkład normalny o średniej 0 i o stałej wariancji σ^2

- Ciekawa dyskusja założeń w A.Aczel „Statystyka w zarządzaniu".

Liniowa prosta regresji - MNK

- Rzeczywiste dane $(x_1, y_1),\ldots,(x_n, y_n).$
- Wartość teoretyczna funkcji regresji $\hat{y} = f(x)$
- Błąd oszacowania $y_i - \hat{y}_i$ tzw. wartość resztowa lub rezyduum.
- Liniowa regresja prosta → wartości rezyduów powinny być jak najmniejsze dla wszystkich i=1,…,*n*.
- Wskaźnik rozproszenia → suma kwadratów rezyduów.

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Dla liniowego wykresu dużych rezyduów nie ma być zbyt wiele → metoda najmniejszych kwadratów! (F.Gauss) daje ona najlepsze liniowe nieobciążone estymatory parametrów regresji (BLUE)

MNK – jak to się liczy?

- Sprawdź w J.Koronacki, J.Mielniczuk, str. 266.

$$\frac{\partial S}{\partial b} = -2\sum_{i=1}^{n}(y_i - (b + ax_i)) = 0$$

$$\frac{\partial S}{\partial a} = -2\sum_{i=1}^{n} x_i (y_i - (b + ax_i)) = 0$$

Trochę dyskusji właściwości:

- Współczynnik regresji *a* można zapisać jako

$$a = \frac{\mathrm{cov}(x,y)}{S_x^2} = r_{xy} \cdot \frac{S_y}{S_x}$$

- Ponadto

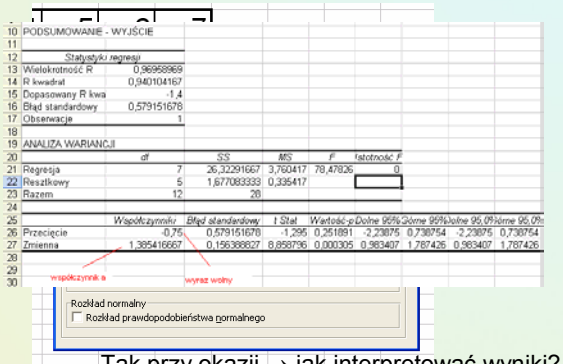
$$r_{xy} = \pm\sqrt{a \cdot b}$$

- Interpretacja wartości współczynnika regresji:
  - Ujemna wartość wskazuję na to, że pod wpływem wzrostu zmiennej niezależnej *x* o jednostkę, zmienna zależna *y* maleje średnio o *a* jednostek
  - Dodatnia wartość → wzrost *y* wraz ze wzrostem *x* o a jedn.
  - *a* = 0 → brak wpływu zmiennej niezależnej na zależną!
- Wyraz wolny rzadko posiada sensowną interpretację.

Zadania

- A teraz trochę popiszemy i policzymy!

Co zrobimy w Excelu? Przykład nastolatki



PODSUMOWANIE - WYJŚCIE

Statystyki regresji

Wielokrotność R	0,96958969
R kwadrat	0,940104167
Dopasowany R kwa	-1,4
Błąd standardowy	0,579151678
Obserwacje	1

ANALIZA WARIANCJI

	df	SS	MS	F	Istotność F
Regresja	7	26,32291667	3,760417	78,47826	0
Resztkowy	5	1,677083333	0,335417		
Razem	12	28			

	Współczynniki	Błąd standardowy	t Stat	Wartość-p	Dolne 95%	Górne 95%	Dolne 95,0%	Górne 95,0%
Przecięcie	0,75	0,579151678	-1,295	0,251891	-2,23875	0,738754	-2,23875	0,738754
Zmienna	1,385416667	0,156388827	8,858796	0,000305	0,983407	1,787426	0,983407	1,787426

współczynnik a

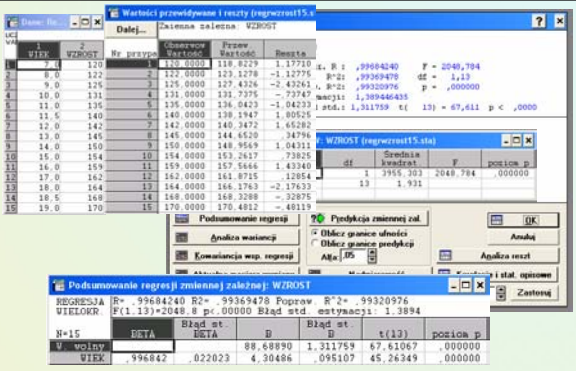
wyraz wolny

Rozkład normalny

Rozkład prawdopodobieństwa normalnego

Tak przy okazji → jak interpretować wyniki?

Przykład wzrost = *f*(wiek) / Statistica (Statsoft)



Weryfikacja modelu regresji

- Ocena dopasowania funkcji regresji do danych empirycznych.
- Składnik resztowy $e_i = y_i - \hat{y}_i$

  tym większy, im większy jest składnik losowy ε,

  może także wynikać z błędnego przyjęcia danej funkcji regresji.

Rozkład całkowitej zmienności zmiennej objaśnianej

- Oceniamy za pomocą wariancji S_y^2 lub całkowitej sumy kwadratów różnic SST

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Ocena modelu regresji

- Całkowitą sumę kwadratów odchyleń (*SST*) w analizie regresji dzieli się na dwie części:

$$SST = SSR + SSE$$

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$

gdzie

- *SSR* – regresyjna suma kwadratów odchyleń (część wyjaśniona przez zbudowany model),
- *SSE* – resztowa suma kwadratów odchyleń (część nie wyjaśniona przez zbudowany model).

Miary dopasowania modelu regresji do danych

- Współczynnik determinacji:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Najważniejsza miara dopasowania funkcji regresji do danych empirycznych; Jest to stosunek zmienności wyjaśnianej przez model do zmienności całkowitej.
- Ponadto błąd standardowy (odchylenie standardowe składnika resztowego)

$$S = \sqrt{\frac{SSE}{n-2}}$$

- Błędy standardowe parametrów *a* i *b*:

$$S(a) = \frac{S}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

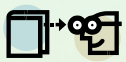
$$S(b) = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Regresja wielokrotna (wielowymiarowa, wieloraka)

- Zmienna objaśniana zależy od więcej niż jednej zmiennej (sytuacja częsta w praktyce).
- Model regresji zmiennej *y* względem zbioru *k* zmiennych niezależnych $x_1, x_2, \ldots, x_k$ jest określony równaniem:

$$y = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots a_k \cdot x_k + \varepsilon$$

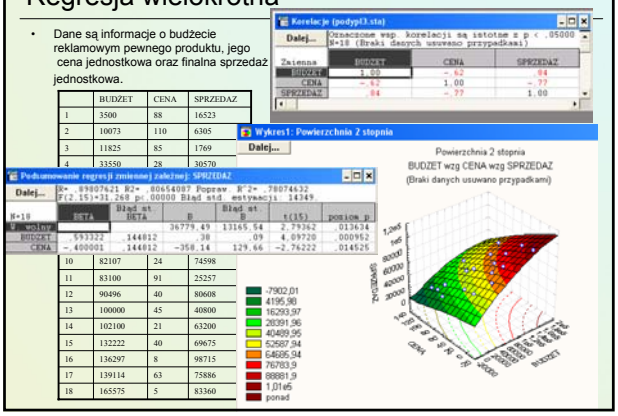
- MNK – rozwiązanie układu równań,
- Zasady oceny i testowania podobne
- Więcej A.Aczel: Statystyka w zarządzaniu (rozdział 11)



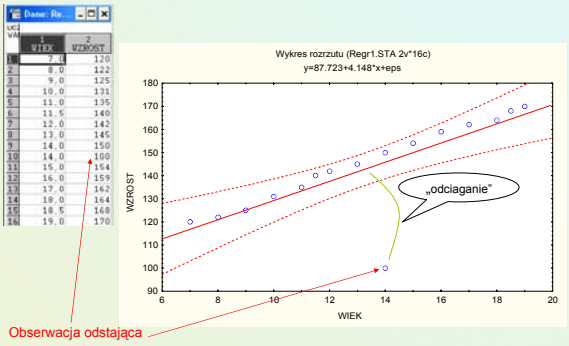
Regresja wielokrotna

- Dane są informacje o budżecie reklamowym pewnego produktu, jego cena jednostkowa oraz finalna sprzedaż jednostkowa.

	BUDŻET	CENA	SPRZEDAŻ
1	3500	88	16523
2	10073	110	6305
3	11825	85	1769
4	33550	28	30570
10	82107	24	74598
11	83100	91	25257
12	90496	40	80608
13	100000	45	40800
14	102100	21	63200
15	132222	40	69675
16	136297	8	98715
17	139114	63	75886
18	165575	5	83360



Regresja – obserwacje odstające / skrajne



Literatura

- Statystyka dla studentów kierunków technicznych i przyrodniczych, Koronacki Jacek, Mielniczuk Jan, WNT, 2001.
- Statystyka w zarządzaniu, A.Aczel, PWN 2000.
- Statystyka praktyczna. W.Starzyńska,
- Statystyka. Wprowadzenie do analizy danych sondażowych i eksperymentalnych. G.Wieczorkowska, Scholar, 2004.
- Przystępny kurs statystyki, Stanisz A., 1997.
- I wiele innych …

Dziękuję za uwagę

Czytaj także podręczniki!

Rodzinnych i radosnych Świąt Bożego Narodzenia oraz szczęśliwego Nowego Roku!