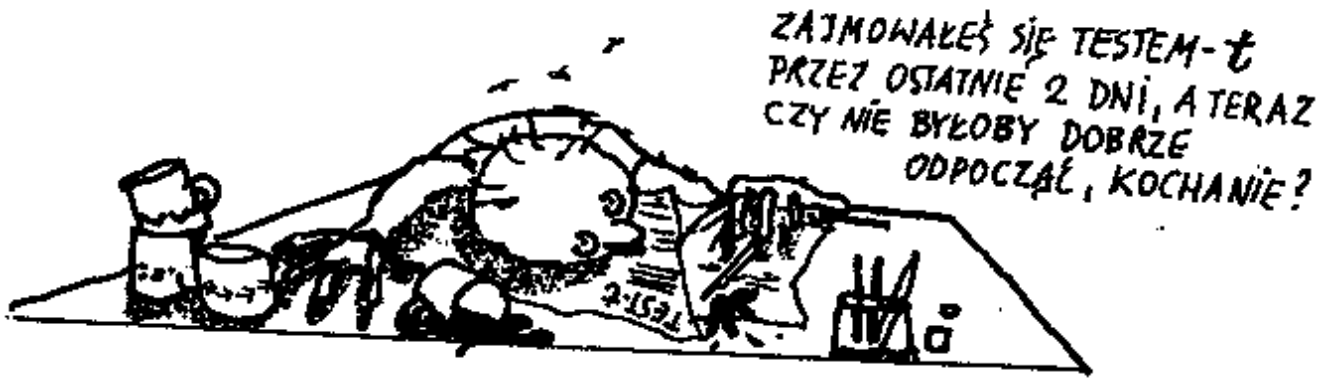
Badanie zależności – trochę uwag na temat poprawności wnioskowania

JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska

Najpierw złe wiadomości

- Czas porozmawiać o kolejnej kartkówce!
- Najpierw kiedy?
  - Z wyprzedzeniem czyli poniedziałek 24 stycznia 2005 ? (identyczna zasada podziału na 2 połówki).
  - Forma podobna jak poprzednio (test z zadaniami)!



Zakres

1. Test *t* - Studenta (porównanie wartości oczekiwanych):
   - Zmienne niezależne,
   - Zmienne powiązane w pary.
2. Testowanie jednorodności wariancji (*F* - test).
3. Współczynnik korelacji próbkowej.
4. Prosta regresja liniowa.
5. Weryfikacja modelu regresji.
6. Test zgodności chi – kwadrat.
7. Test niezależności dwóch zmiennych.
8. Proste testy nieparametryczne (znaków, Wilcoxona), korelacja Spearmana?

Powróćmy do badania zależności między zmiennymi

- W poprzednich wykładach badaliśmy związek między zmiennymi z wykorzystaniem różnych metod: korelacji, regresji, testów chi-kwadrat.
- Teraz chcemy zastanowić się nad zrozumieniem natury związku między zmiennymi, który odkrywaliśmy tymi metodami w dostępnych danych.

Przykłady silnych związków między zmiennymi.

Badanie danych medycznych wskazują między innymi na:

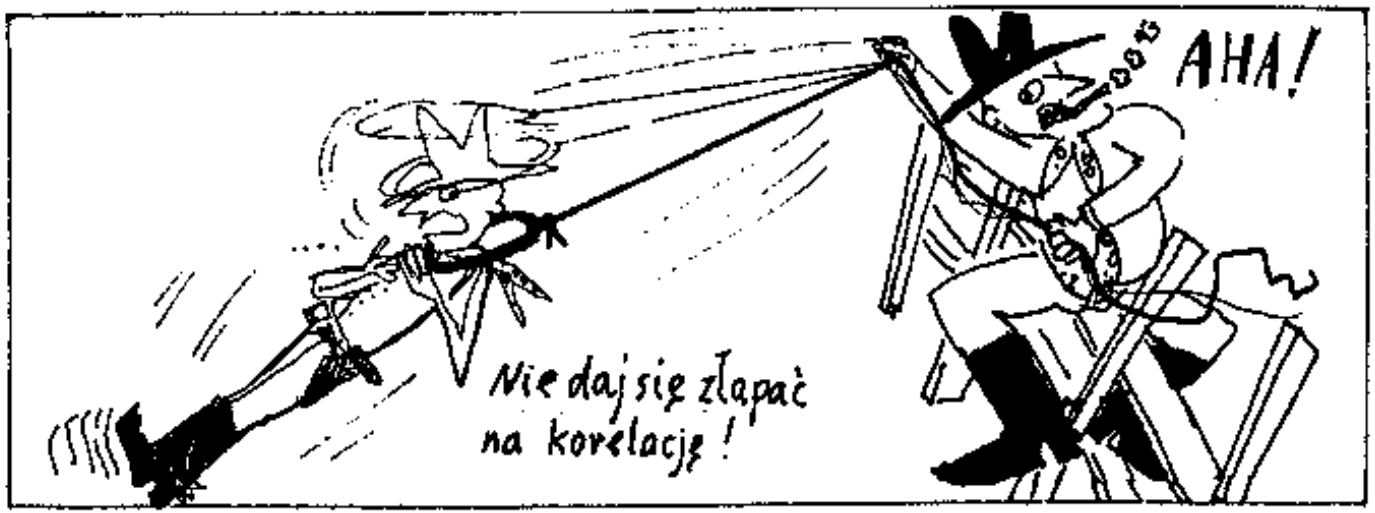
- Silną korelację pomiędzy zachorowaniem na raka a paleniem papierosów.
- Związek pomiędzy ilością cholesterolu we krwi a chorobami serca.

- Co można wywnioskować z takich rezultatów?
- Jakie błędy można popełnić podczas interpretacji rezultatów?

Korelacja a przyczynowość

- Wierzy się potocznie, że gdy istnieje silny związek (korelacja) między dwiema zmiennymi, to jedna ze zmiennych *jest przyczyną* drugiej.
- Uwaga to jest powszechny błąd!

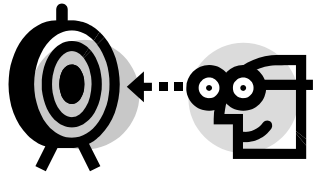
  Istnienie związku między zmiennymi NIE OZNACZA PRZYCZYNOWOŚCI!!
- Nie wpadnij w taką pułapkę



Inne przykłady silnych związków między zmiennymi.

Przykład nr. 1: Długość okresu pobierania nauki i wysokość zarobków są wysoce skorelowane

- Badania ankietowe w Anglii (F.Clegg str. 154).
- Pytanie czy poziom wykształcenia sam w sobie determinuje stanowisko i wysokość zarobków?
- Raczej związek nie jest tak prosty, lecz dość złożony!
  - Inteligencja osoby, cechy osobowościowe, różne umiejętności, no i łut szczęścia, … ☺
- Wysoka korelacja wyłącznie opisuje związek, który istnieje w danych pomiarowych pomiędzy obiema zmiennymi.



Inne przykłady silnych związków między zmiennymi.

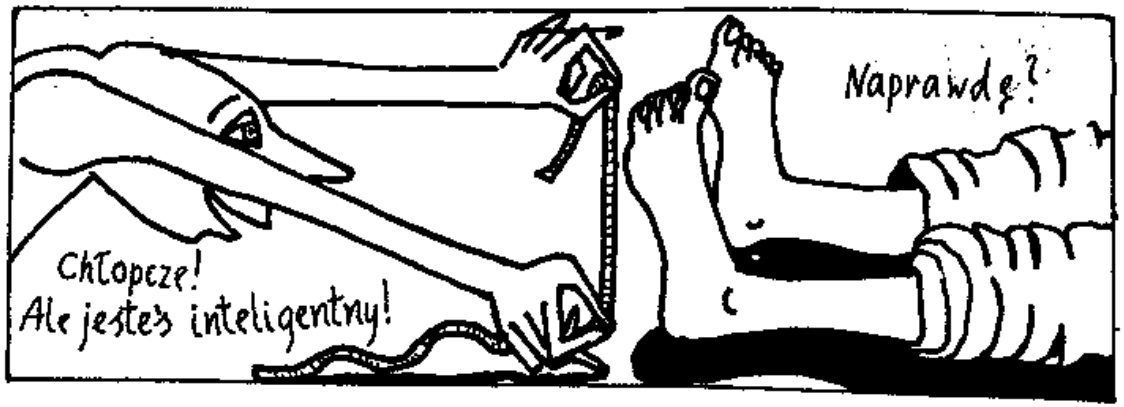
Przykład nr. 2: Oglądalność TV i wskaźnik urodzeń są negatywnie skorelowane

- Badania demograficzne w USA $\rightarrow$ zaobserwowano wysoki wzrost wskaźnika urodzeń, które nastąpiło 9 miesięcy po awarii TV w pewnych rejonach USA.
- Podobnie wiele osób interpretując inne badania wierzy istnieje sprecyzowany związek pomiędzy pokazywaniem przemocy w TV a poziomem agresji!
- Związek przyczynowo-skutkowy nie jest tak prosty i bezpośredni, lecz dość złożony i wymaga uwzględniania innej wiedzy niż wyłącznie korelacja!

Przykłady silnych związków między zmiennymi.

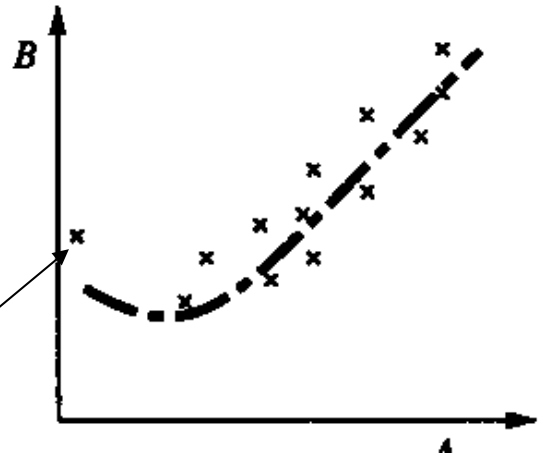
Przykład 3. Silna korelacja pomiędzy długością lewego i prawego ramienia wśród ludzi:

- Zauważona w pewnych badaniach antropometrycznych.
- Uważaj aby oprócz obliczonej korelacji w danych, czegoś nie przeinterpretować?



Czynniki wprowadzające w błąd (korelacja i regresja)

- Obserwacje oddalone (samotniki) / ang. outliers.

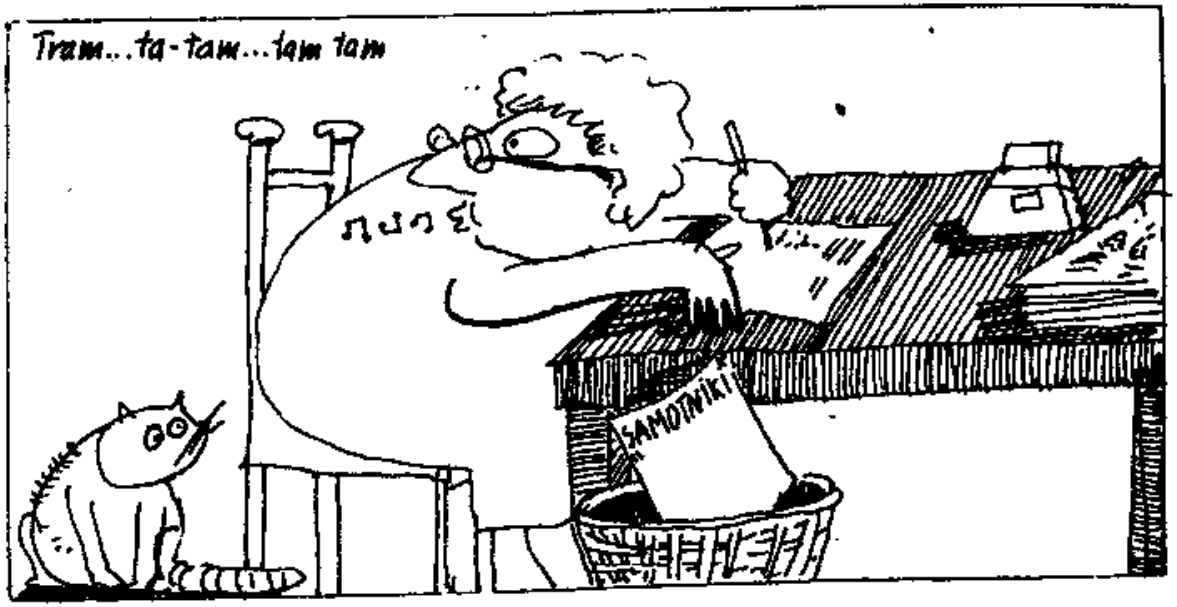


- Dla małych próbek obecność „samotnika” może wywołać mylne konsekwencje w interpretacji danych;

  np. sugerować związek krzywoliniowy,
  gdy w rzeczywistości nie występuje.

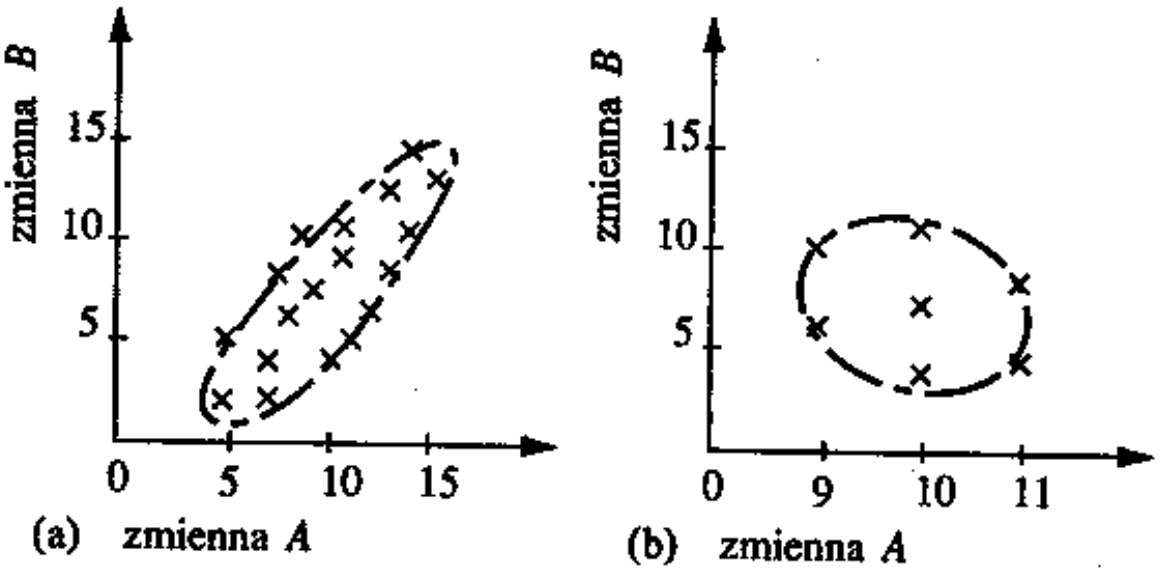
Obserwacje oddalone

- Czy „samotniki” powinny być wykluczane z analizy?
- Nie ma prostej odpowiedzi!
- Zależy od danych (ich natury, wielkości,…) oraz problemu, którego dotyczą.



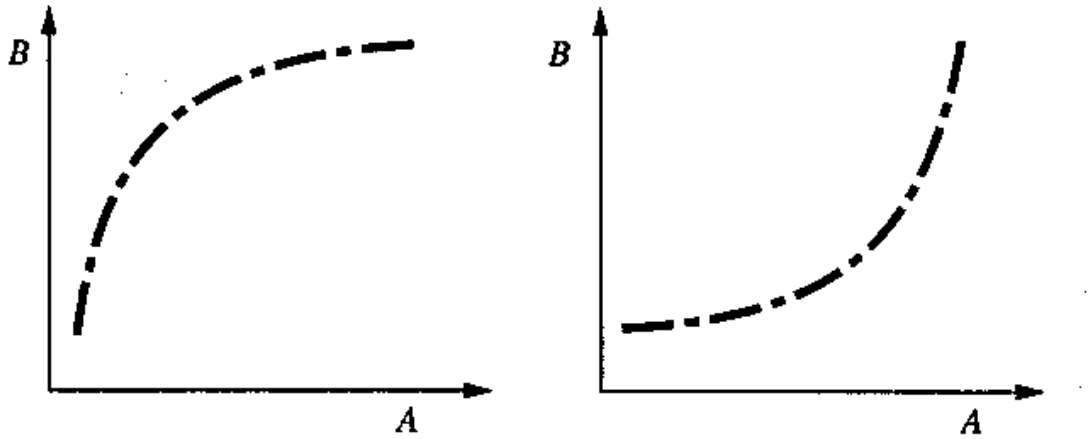
Błędne wybranie zakresu zmiennych.

- Podczas częściowego pobierania zmiennej można błędnie opracowywać wyniki wzięte z jednego krańca rozstępu wartości → ryzykuje się niezauważenie istniejącego związku.



Dobór właściwej miary związku

- W przypadku związków krzywoliniowych $\rightarrow$ nie stosuj współczynnika korelacji liniowej Pearsona!



- Po pierwsze zawsze warto obserwować wykresy!
- W powyższych przypadkach bardziej dogodne byłoby użycie korelacji rho-Spearmana (poczekaj na kolejny wykład).

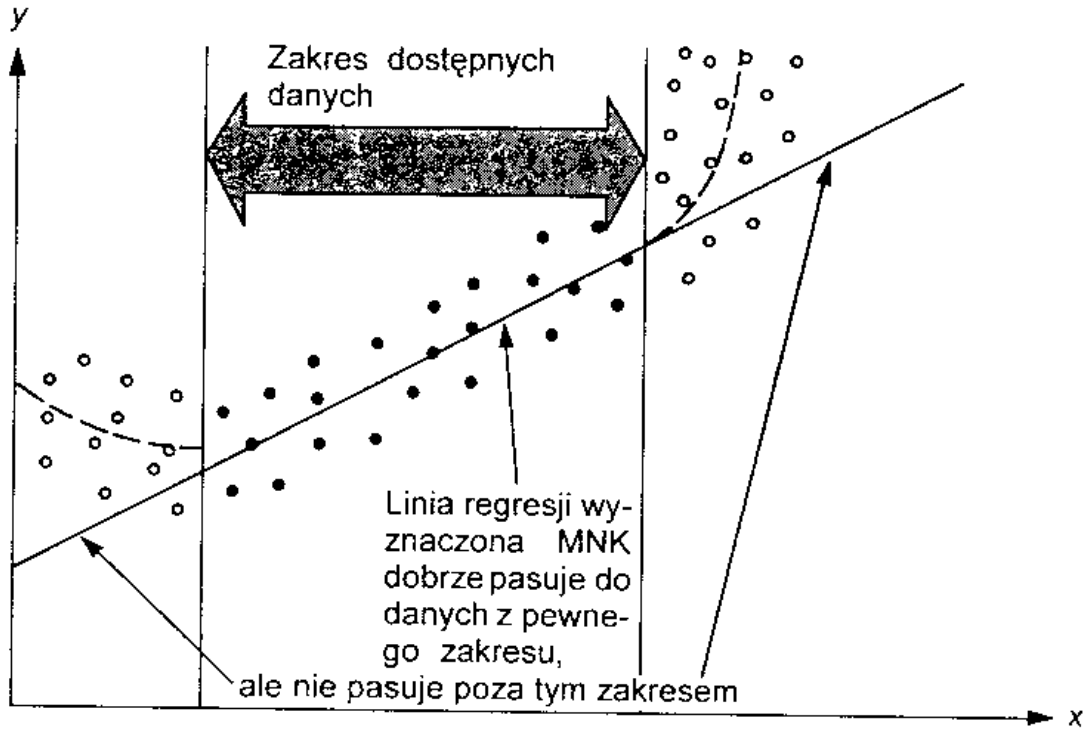
Przewidywanie wartości zmiennej

- Ze stwierdzenia korelacji i związku regresyjnego między zmiennymi nie wynika, że jedna zmienna jest przyczyną drugiej.
- Można jednak wykorzystywać analizę regresji do prognozowania (przewidywania) wartości zmiennej zależnej w oparciu o równanie regresji:

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

Przewidywanie w regresji

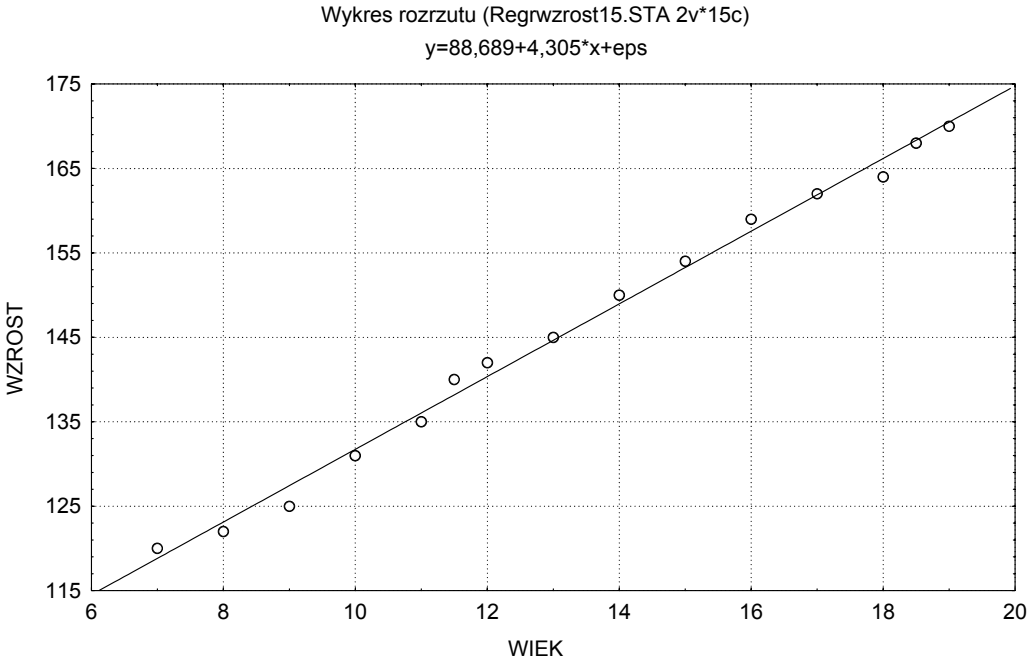
- Wartości prognozowane nie powinny wykraczać poza zakres wartości wykorzystywanych w procedurze szacowania parametrów równania regresji.



Rysunek 10.27. Niebezpieczeństwo ekstrapolacji

Inny przykład

- Z populacji dzieci (w zakresie wieku 7-19 lat) wybrano losowo próbę 15 osobową i określono dla nich dwie cechy: *x* – wiek w latach oraz *y* – wzrost w cm:
- (7,120),(9,125),(18,164), (11.5,140), (8,122),(11,135), (13,145), (17,162), (10,131), (19,170), (14,150), (12,142), (18.5,168), (15,154), (16,159)
- Korelacja **0,99**
- **y=88,689+4,305·x**



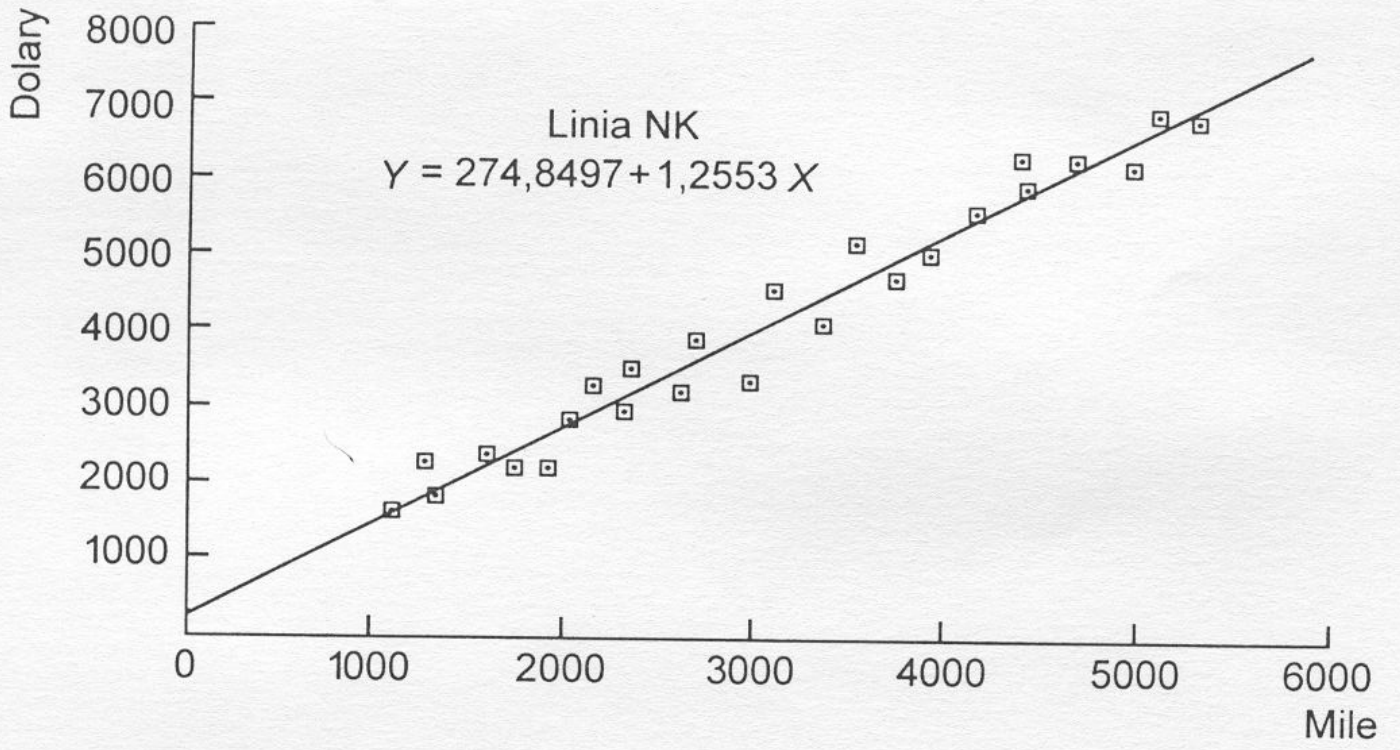
Przykład American Express

- Rozważmy przykład posiadaczy kart kredytowych American Express $\rightarrow$ firma jest przekonana, że posiadacze jej kart podróżują więcej niż inni ludzie.
- W badaniach marketingowych podjęto próbę ustalenie związków między długością tras podróży a obciążeniem karty kredytowej jej posiadacza w danym okresie czasu.
- Więcej w Aczel: Statystyka w zarządzaniu, str. 468.

Tablica 10.1. Dane do badania przeprowadzonego na zlecenie American Express

Długość tras (w milach)	Obciążenie kart (w $)
1 211	1 802
1 345	2 405
1 422	2 005
1 687	2 511
1 849	2 332
2 026	2 305
2 133	3 016
2 253	3 385
2 400	3 090
2 468	3 694
2 699	3 371
2 806	3 998
3 082	3 555
3 209	4 692
3 466	4 244
3 643	5 298
3 852	4 801
4 033	5 147
4 267	5 738
4 498	6 420
4 533	6 059
4 804	6 426
5 090	6 321
5 233	7 026
5 439	6 964

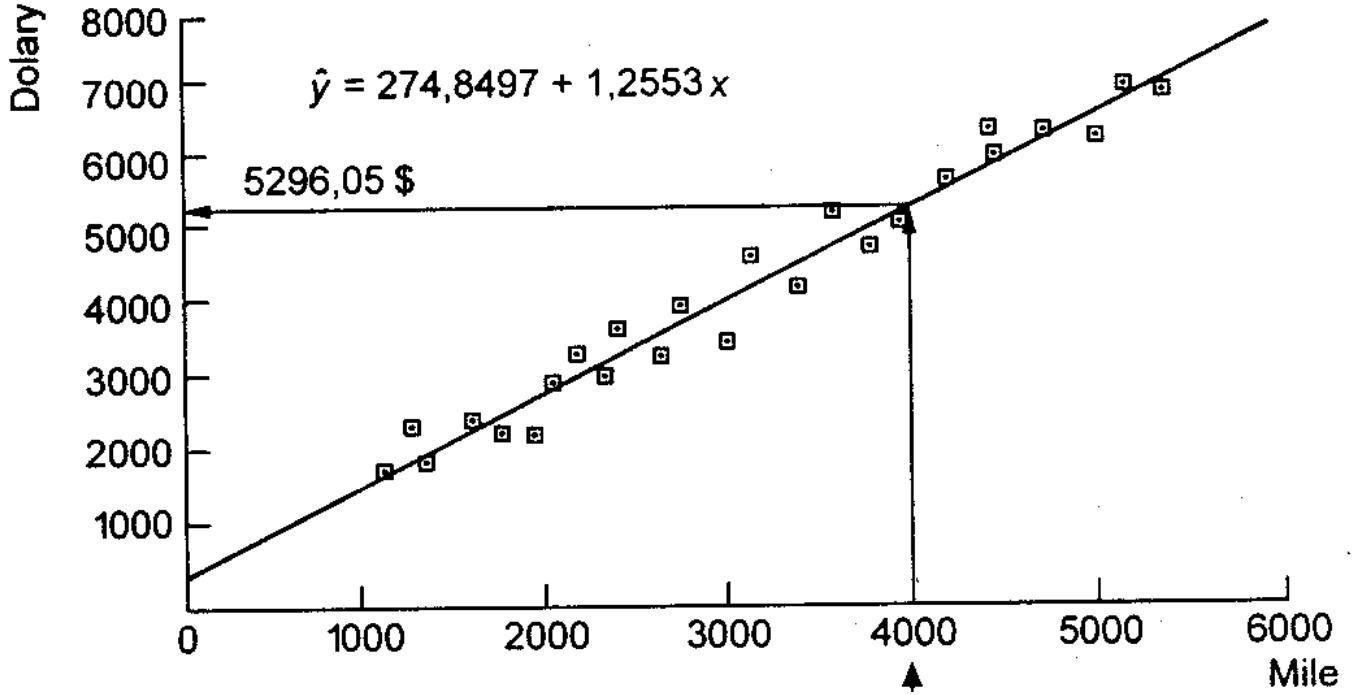
Analiza regresji – American Express



Rysunek 10.12. Linia NK w badaniu zleconym przez American Express

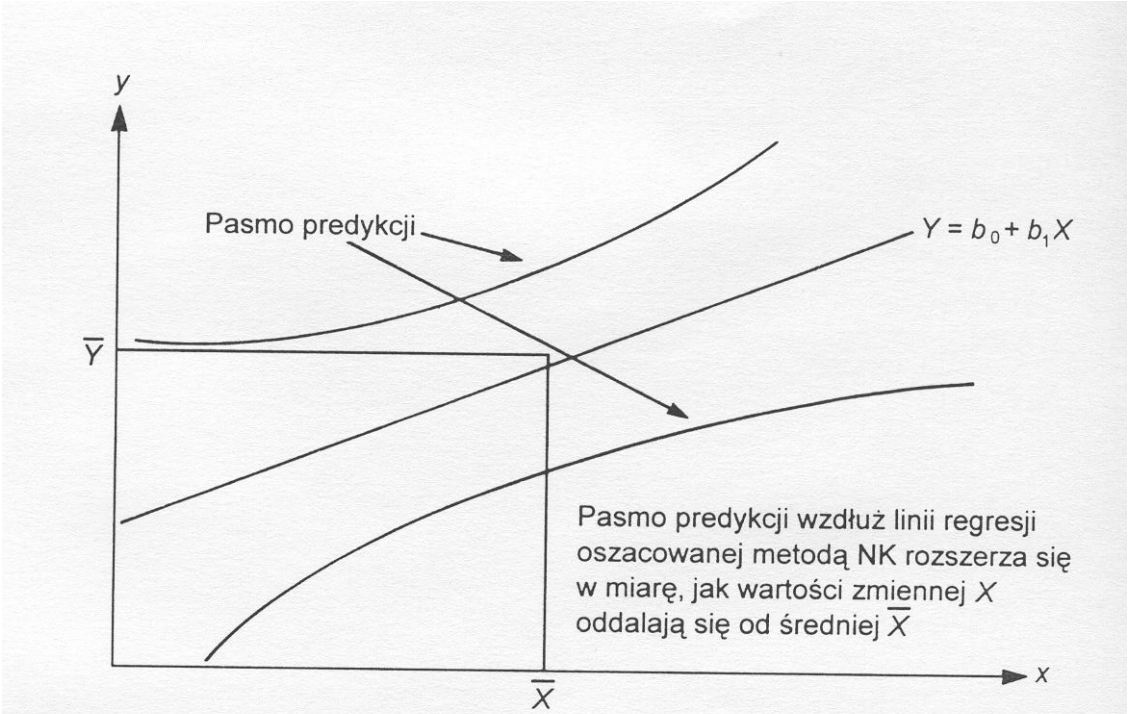
Prognoza punktowa w regresji

- Łatwa na podstawie równania regresji.
- Np. oceń obciążenie kart wśród posiadaczy kart, których trasa podróży osiągnie 4000 mil, w okresie o takiej długości jak okres badany: $\hat{y} = 274{,}85 + 1{,}2663 \cdot x = 274{,}85 + 1{,}2663 \cdot 4000 = 5296{,}05$



Przedziały predykcji

- Ograniczenie prognoz punktowych → błędu pochodzące zarówno z niepewności szacunków, jak i losowej zmienności położenia punktów w stosunku do linii regresji.
- Stosuj wtedy tzw. przedziały predykcji (tzw. prognozy przedziałowe).



Rysunek 10.29. Pasmo predykcji

Przedziały predykcji

- (1-α)·100% przedział predykcji zmiennej Y

$$\hat{y} \pm t_{\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

- Rozpiętość przedziału predykcji zależy od odległości wartości *x* od średniej $\bar{x}$!

Przykład: posiadacz, który przebył 4000 mil i 95% przedział ufności.

- Z analizy danych historycznych:

  $\bar{x}$ = 79448/25=3177,92; SS_x = 40947557,84 a *s* = 318,16
  Ponadto *t* przy 23 stopniach swobody wynosi 2,069

Stąd przedział 5296,05$\pm$676,62 = [4619,43; 5972,67]

- Oznacza to, że w oparciu o wyniki badań można mieć 95% zaufania do prognozy, że posiadacz karty, który przebył trasę 4000 mil w okresie o danej długości obciąży swoją kartę kredytową sumą od 4619.43 do 5972,67$.

Literatura

Wykład jest inspirowany dyskusjami problemów związanych z poprawnym wnioskowaniem przedstawionymi w pozycjach:

- Po prostu statystyka, Frances Clegg, WSiP, 1994. (rozdziały 12 i **15**)
- Statystyka w zarządzaniu, A.Aczel, PWN, 2000. (rozdziały 10 i 15)
- Statystyka dla studentów kierunków technicznych i przyrodniczych, Koronacki Jacek, Mielniczuk Jan, WNT, 2001.

Dziękuję za uwagę

Czytaj także podręczniki!