Metody matematyczne w transporcie

Wielokrotna regresja liniowa

Marzena Nowakowska
Wydział Zarządzania i Modelowania Komputerowego

Pojęcia podstawowe

Modelowanie regresyjne jest to tworzenie modelu abstrakcyjnego (matematycznego), za pomocą którego wartość konkretnej zmiennej wyznacza się na podstawie znajomości wartości innych zmiennych. Modelowanie takie może być wyrażone w postaci zależności: $Y = f(X_1, ..., X_k)$.

$$\boldsymbol{Y} = f(X_1, X_2, ..., X_k)$$

Zmienna objaśniania
Zmienna zależna
Zmienna wyjściowa
Zmienna odpowiedzi, odpowiedź
Zmienna endogeniczna

Zmienne objaśniające
Zmienne niezależne
Zmienne wejściowe
Atrybuty (wejściowe)
Predyktory
Zmienne egzogeniczne

Analiza korelacji i regresji

Dział statystyki obejmujący ustalanie i mierzenie związku cech nosi nazwę teorii korelacji i regresji. Zagadnienia, które ma ta teoria rozstrzygać nie są proste, gdyż związek cech ustalany zazwyczaj na podstawie pewnego badania częściowego, czyli próby, może mieć charakter przypadkowy. Może być więc wynikiem działania pewnych przyczyn ubocznych, a nie rezultatem istnienia rzeczywistej zależności między badanymi cechami.

Liniowa współzależność cech
Kowariancja

Kowariancja jako miara współzależności cech *X* i *Y*:

$$C(X,Y) = E((X-E(X))*(Y-E(Y))$$

E(Z) - wartość oczekiwana zmiennej losowej *Z*

Estymator (kowariancja z próby)

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})$$

Wada: posiadanie miana

Liniowa współzależność cech Współczynnik korelacji liniowej – współczynnik Pearsona

Korelacja Pearsona jako miara współzależności **liniowej** cech *X* i *Y*:

$$\varrho(X,Y) = C(X,Y) / (\sigma_X \sigma_Y)$$

σ_X, σ_Y – odchylenie standardowe zmiennych losowych *X* i *Y*

Oznaczenie estymatora: r

Estymator:

$$r_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Zaleta: brak wymiaru, łatwość w interpretacji.

Współczynnik korelacji Pearsona ma wartości z przedziału [-1,1].

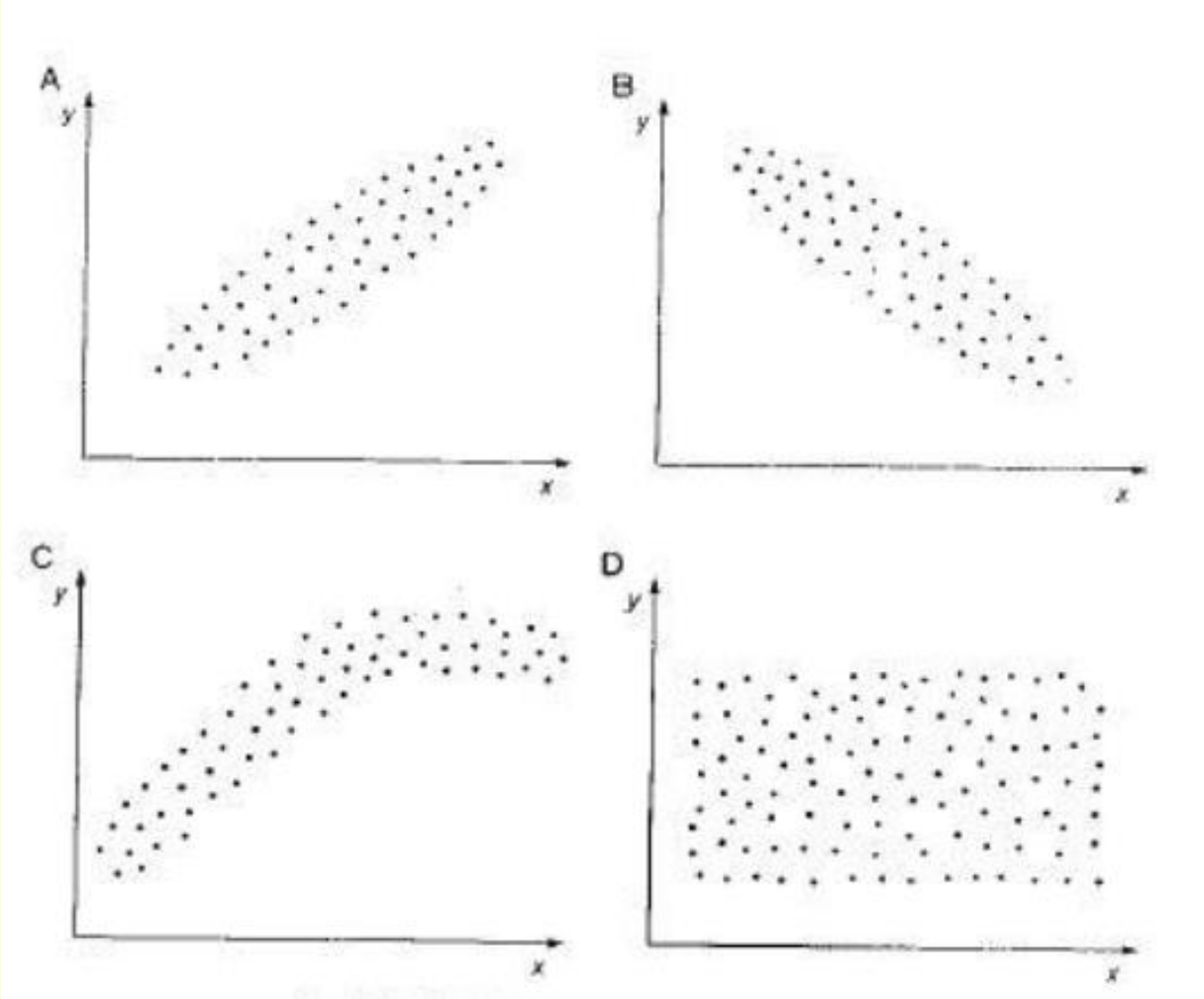
Inne miary współzależność dwóch cech

Korelacja rang Spearmana (lub: korelacja rangowa Spearmana, rho Spearmana) – nieparametryczna miara monotonicznej zależności miedzy zmiennymi losowymi. Mierzy także zależność nieliniową. Wartości: [-1, 1]

Kendall's tau-*b* – miara zgodności na podstawie danych na skali porządkowej. Miara bazuje na liczbie par obserwacji zgodnych (concordant) i niezgodnych (discordant) oraz związanych (tie). Jest miarą monotonicznej zależności między zmiennymi.
Wartości: [-1, 1]

Hoeffding's D – miara do wykrywania odstępstw od niezależności. Jeżeli nie ma par związanych, miara ma wartości [-0.5, 1], z wartością 1 wskazująca całkowitą zależność (dla par związanych mniejsze niż 1). Im większa wartość miary, tym silniejszy związek między zmiennymi.

Wykresy rozrzutu ilustrujące potencjalne związki między dwiema zmiennymi X i Y



Źródło: Internet

Testowanie istotności współczynnika korelacji liniowej

Procedura wnioskowania statystycznego opiera się o parametryczny test istotności, a treść hipotez o braku związku liniowego między cechami jest następująca:

H0: $\rho = 0$

wobec hipotezy alternatywnej

H1: $\rho > 0$ albo H1: $\rho < 0$ albo H1: $\rho \neq 0$

Statystyka testowa: $$t = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n - 2}$$

ma rozkład t-Studenta o n-2 stopniach swobody

Jeśli prawdopodobieństwo testowe odpowiadające wartości statystyki testowej jest mniejsze od przyjętego poziomu istotności, to hipotezę zerową o braku korelacji należy odrzucić.

Cechy współczynnika korelacji Pearsona

- Związek może być silny (na podstawie r – z próby) a mimo to nieistotny i odwrotnie, związek może być słaby (na podstawie próby) ale istotny.
- Kluczowa jest wielkość próby.
- Dla małych zbiorów jest stosunkowo łatwo uzyskać silną korelację przez przypadek i trzeba zwrócić uwagę na poziom istotności zanim wyciągnie się ostateczne wnioski, by nie odrzucić prawdziwej hipotezy zerowej, czyli nie popełnić błędu I rodzaju.
- Dla większych zbiorów, jest bardzo łatwo osiągnąć istotność, ale trzeba zwrócić uwagę na siłę korelacji (wartość bezwzględna współczynnika), żeby mieć pewność, że występuje rzeczywisty związek.

Interpretacja wartości r

Współczynnik korelacji ma wartości z przedziału [-1,1].

Im korelacja jest bliższa +/-1, tym bliższa jest idealnemu liniowemu związkowi. Przykładowa interpretacja korelacji dla wartości bezwzględnej współczynnika:

- <0,2, 0,4> bardzo słaby związek lub jego brak
- <0,4, 0,7> związek umiarkowany
- <0,7, 1> związek silny lub bardzo silny

Nie są to sztywne kryteria klasyfikacji podziałów.

W niektórych sytuacjach można zmienić poziom słabej korelacji np. rozszerzyć przedział (od 0,2 do 0,6) a w innych przesunąć przedział w górę (od 0,5 do 0,8).

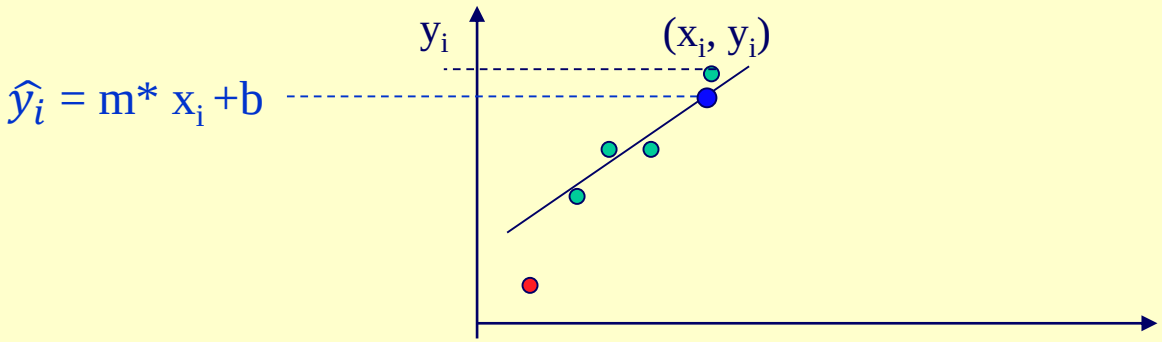
Opinia jest subiektywna, w której należy uwzględnić liczebność próby.

(Bardzo) prosta regresja liniowa

Jeżeli smuga punktów na wykresie układa się wzdłuż linii prostej, to dopasowuje się do niej funkcję liniową, którą można opisać zależnością: $Y = m*X+b$.

Położenie prostej na płaszczyźnie jednoznacznie określają obie stałe *m* i *b* przytoczonej funkcji, zwanej liniową funkcją regresji.

Funkcja ta określa trend liniowy zależności *Y* od *X*. Analiza takiego związku nosi nazwę analizy regresji liniowej.



Wielokrotna regresja liniowa

Ogólnym celem jest ilościowe ujęcie związk**ów między wieloma zmiennymi niezależnymi a zmienną zależną.**

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i \qquad i = 1, \ldots, n$$

gdzie:

- u_i, $i = 1, \ldots, n$ są wartościami błędu, niezależnymi od siebie dla kolejnych obserwacji i o takim samym rozkładzie (normalnym): $E(u_i) = 0$ and $V(u_i) = \sigma^2_u$,
- rozkład błędu jest niezależny od łącznego rozkładu zmiennych objaśniających $X_1, \ldots, X_k$
- funkcja regresji zależności zmiennej Y od $(X_1, \ldots, X_k)$ definiuje warunkową wartość oczekiwaną: $E(Y| X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \quad + \beta_k X_k$ przy czym $V(Y| X_1, \ldots, X_k) = \sigma^2_u$,
- nieznane parametry β_0 , $\beta_1, \ldots, \beta_k$ są stałymi

Estymowany (oszacowany) model liniowej regresji wielokrotnej ma postać:

$$Y = B_0 + B_1 X_1 + \ldots + B_k X_k$$

gdzie $B_0, B_1, \ldots B_k$ są estymatorami prawdziwych wartości parametrów z próby.

Oszacowanie modelu regresji liniowej

Oszacować (estymować) model oznacza znaleźć wartości parametrów strukturalnych na podstawie konkretnej próby.

Oszacowany (estymowany) model liniowej regresji wielokrotnej ma postać:

$$Y = B_0 + B_1 X_1 + \ldots + B_k X_k$$

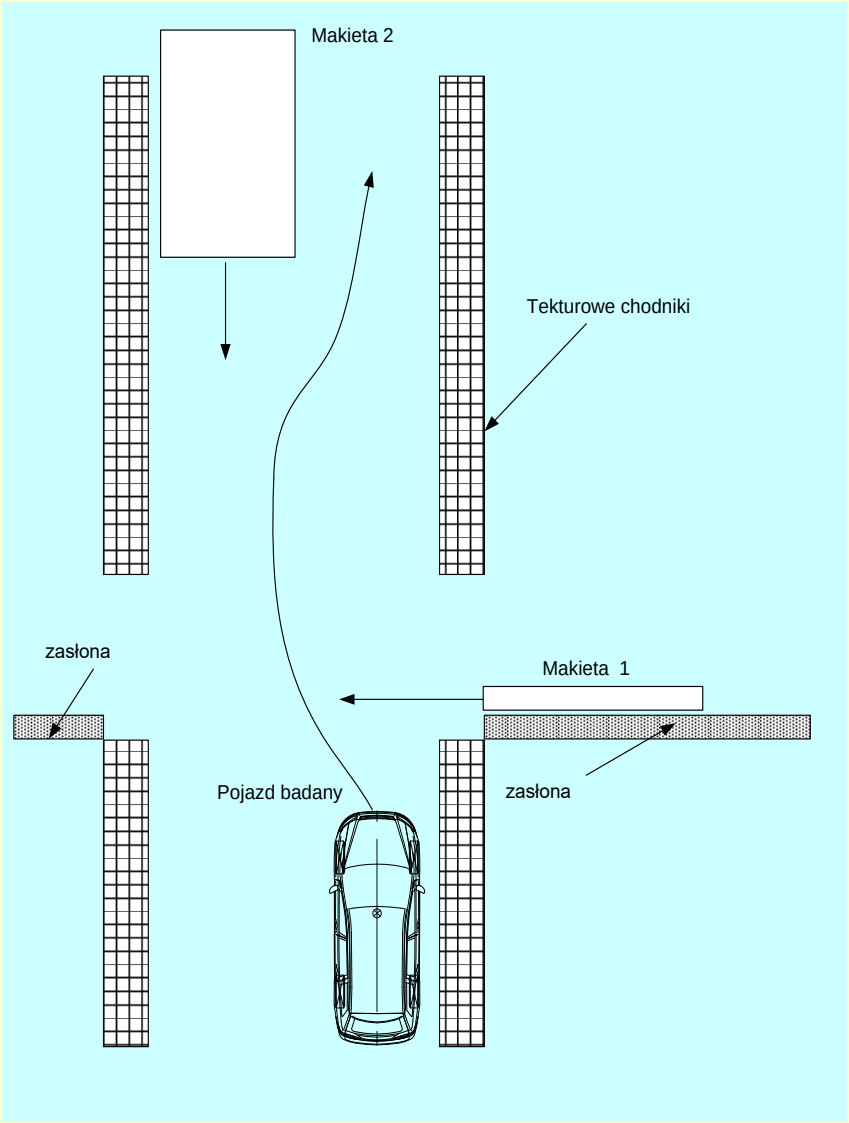
gdzie:

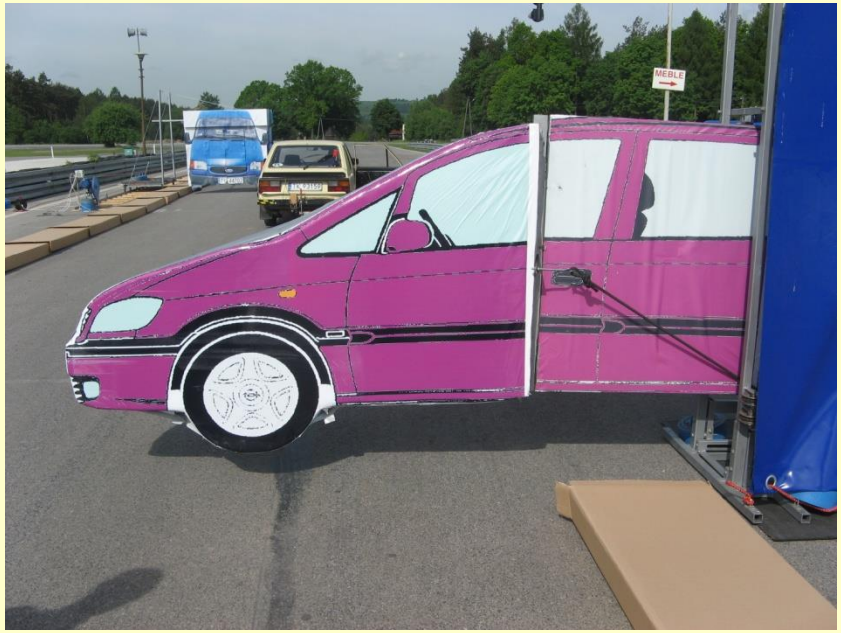
B_0, B_1, ...B_k są estymatorami prawdziwych wartości parametrów policzonymi z próby.

Metody szacowania parametrów:

- **MNK – Metoda Najmniejszych Kwadratów** (własności estymatorów: nieobciążone, zgodne, efektywne)
- MM – Metoda Momentów
- MNW – Metoda Największej Wiarygodności
- inne

Eksperyment badawczy





Dane: udostępnione przez Katedrę Pojazdów Samochodowych i Transportu, Politechniki Świętokrzyskiej
Miejsce: tor samochodowy na Miedzianej Górze
Rejestrowane: wielkości charakteryzujące działania podjęte przez kierowcę oraz parametry ruchu pojazdu badawczego

Zmienne rozważane w analizie

Nr_proby: numer określający kolejność realizowanych prób

Czas_ryzyka (TTC – Time To Colission): czas, jaki upłynąłby od momentu zauważenia przeszkody do chwili zderzenia z nią, gdyby kierowca nie zmodyfikował swojego zachowania

Gaz_srednia: średnia czasu reakcji psychicznej kierowcy, tzn. czasu od chwili zauważenia zagrożenia (przeszkody) do chwili rozpoczęcia zdejmowania nogi z pedału gazu

Gaz_odchylenie: odchylenie standardowe dla czasu reakcji psychicznej kierowcy

Hamulec_srednia: średnia czasu reakcji psychomotorycznej kierowcy, tzn. czasu od chwili zauważenia zagrożenia (przeszkody) do chwili rozpoczęcia naciskania pedału hamulca

Hamulec_odchylenie: odchylenie standardowe dla czasu reakcji psychomotorycznej kierowcy

Skret_srednia: średnia czasu reakcji psychomotorycznej kierowcy przy skręcie, tzn. czasu od chwili zauważenia zagrożenia (przeszkody) do chwili rozpoczęcia manewru omijania

Skret_odchylenie: odchylenie standardowe dla czasu reakcji psychomotorycznej kierowcy przy skręcie

S: odległość od przeszkody

V: prędkość jazdy samochodu (prędkość najazdu)

Poszukiwanie związków

Wykorzystując metodę wielokrotnej regresji liniowej, opracować modele zgodnie z podanymi niżej zależnościami:

Gaz_srednia = f(Nr_proby, TTC, Hamulec_srednia, Skret_srednia)

Hamulec_srednia = f(Nr_proby, TTC, Gaz_srednia, Skret_srednia)

Skret_srednia = f(Nr_proby, TTC, Gaz_srednia, Hamulec_srednia)

Gaz_srednia=f(TTC, Nr_proby, Hamulec_srednia, Skret_srednia)

Hamulec_srednia = f(Nr_proby, S, V, TTC, Gaz_srednia, Skret_srednia)

Skret_srednia = f(Nr_proby, S, V, TTC, Gaz_srednia, Hamulec_srednia)

itd.

Przykładowa postać wielokrotnej regresji liniowej:

Skret_srednia = B0 + B1* Nr_proby + B2*TTC +B3*Gaz_srednia + B4*Hamulec_srednia

Gaz_srednia = B0 + B1*TTC + B2* Nr_proby + B3*Hamulec_srednia + B4*Skret_srednia

Współliniowość zmiennych objaśniających

Gdy zmienne objaśniające są wysoko skorelowane, wyniki analizy regresji mogą być niestabilne, obciążone lub nieprawidłowe. Właściwość, w której istnieje zależność liniowa między zmiennymi wejściowymi nosi nazwę **współliniowości** (*collinearity*) lub **wielowspółliniowości** (*multicollinearity*).

Możliwości sprawdzenia:

- Współczynnik inflacji (podbicia) wariancji WIW dla danej zmiennej X_j wskazuje, o ile wariancja współczynnika przy zmiennej X_j jest zawyżona z powodu zależności liniowej tej zmiennej od innych zmiennych wejściowych w modelu.
  - $WIW_j >= 5$ ($R_j^2 \approx 0.80$) – wskazanie umiarkowanej współliniowości
  - $WIW_j >= 10$ ($R_j^2 \approx 0,9$) – wskazanie silnej współliniowości
- Macierz korelacji – wykrycie silnie skorelowanych zmiennych
- Wskaźnik warunku (uwarunkowania) macierzy $\mathbf{X} \leftarrow (n^x k)$ - wymiarowa macierz wartości zmiennych objaśniających. Dla każdej zmiennej objaśniającej określa się, jaka część zmienności estymatora jest przypisana każdej wartości własnej macierzy $\mathbf{X}^T\mathbf{X}$. Problem współliniowości jest sygnalizowany wysoką wartością składowej związanej z dużą wartością wskaźnika warunku jednej lub większej liczby zmiennych wejściowych.
  - Wskaźnik warunku $\approx$ 10 – sygnalizacja problemu
  - Wskaźnik warunku $\approx$ 100 – być może konieczność zmiany postaci modelu

Współliniowość zmiennych objaśniających metody postępowania

- Sprawdzić wartość **współczynnika inflacji wariancji**. Jeśli ta wartość jest mniejsza niż 10 dla wszystkich zmiennych niezależnych, temat jest zamknięty i nie ma potrzeby wykonywać żadnych operacji na zmiennych.
- Jeśli są tylko dwie wartości WIW powyżej 10, można założyć, że istnieje problem kolinearności między tymi dwiema wartościami. W takiej sytuacji należy wyeliminować jedną zmienną albo utworzyć kompozyt.
- Jeśli jest więcej niż dwie zmienne objaśniające z WIW powyżej 10, należy przeanalizować diagnostykę kolinearności (współczynnik uwarunkowania).
- Zidentyfikować w macierzy **współczynników uwarunkowania** wiersze z wartością tego wskaźnika powyżej 15.
- Jeżeli w tych wierszach jest więcej niż jedna kolumna (więcej niż jedna zmienna niezależna) o wartościach powyżej 0,90 w proporcjach wariancji, zakłada się problem kolinearności między tymi zmiennymi, które mają te wysokie wartości. Obecność tylko jednej takiej zmiennej w wierszu o wysokiej wartości (powyżej 0,90), nie ma znaczenia.
- Jeśli nie można zidentyfikować źródła wielowspółliniowości, ponieważ nie ma wierszy o kilku wartościach wariancji powyżej 0,9, to kryterium można zmniejszyć i rozważyć na przykład pary zmiennych niezależnych (lub ich grup) z wartości powyżej 0,8 lub 0,7.