Metody matematyczne w transporcie

Modelowanie czasu reakcji kierującego w sytuacji zagrożenia na drodze. Diagnoza modelu wielokrotnej regresji liniowej

Marzena Nowakowska

Katedra Technologii Informatycznych

Wydział Zarządzania i Modelowania Komputerowego

Sprawdzenie założeń modelu liniowego

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

k – liczba zmiennych niezależnych

Normalność rozkładu błędów (sprawdza się reszty)

Błędy mają średnią zerową i stałą wariancję (sprawdza się reszty)

Obserwacje są niezależne (sprawdza się losowość reszt)

Liczebność próby jest większa od liczby parametrów

W modelu wielokrotnej regresji liniowej:

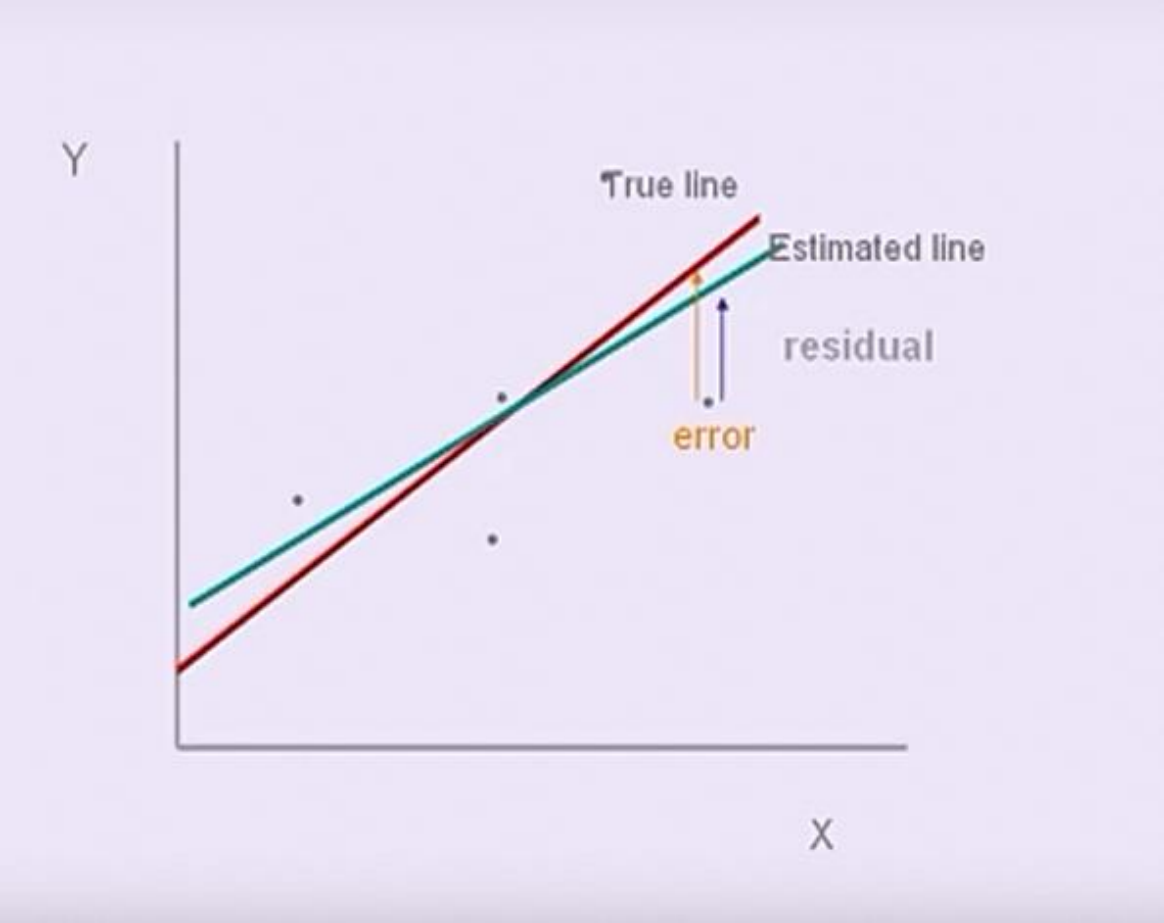
u_i, (i = 1, ..., n) są wartościami błędu (estymowane przez reszty)

r_i (i = 1, …, n) są resztami: $r_i = y_i - \hat{y}_i = y_i - (B_0 + B_1 x_{i1} + \ldots + B_k x_{ik})$

Błąd = Składnik losowy ≠ Reszta = Rezyduum

Założenia sprawdza się wykonując właściwe testy oraz za pomocą wykresów diagnostycznych dostarczanych przez komputerowe systemy analityczne.

Model liniowy: błąd vs. reszta



Źródło: https://ppt-online.org/215018

Normalność rozkładu błędów

Wykres histogramu, test normalności, wykres kwantylowy.

Kwantyl rzędu q ($0 < q < 1$) w populacji jest taką liczbą x_q, że q*100% elementów tej populacji ma wartość badanej cechy mniejszą lub równą x_q.
Do najbardziej popularnych kwantyli zalicza się:

- kwartyle (dzielące populację na cztery części – pierwszy kwartyl, mediana i trzeci kwartyl),
- kwintyle (czyli kwantyle rzędu 1/5, 2/5, 3/5, 4/5),
- decyle (dzielące na dziesięć części),
- percentyle (dzielące populację na 100 części).

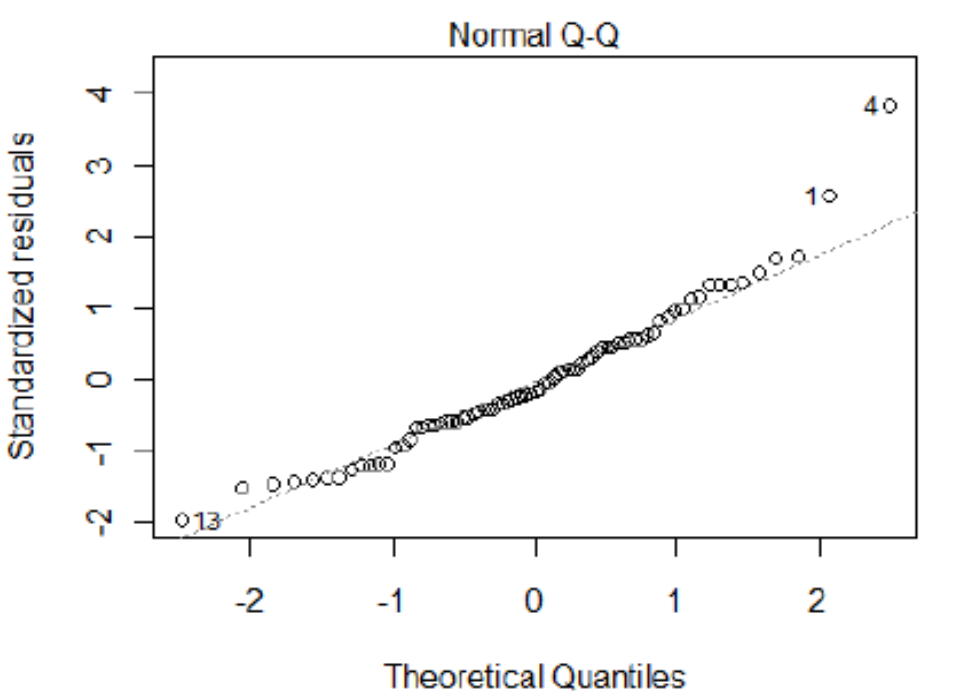
Aby otrzymać test normalności reszt należy te reszty wygenerować.
Przykładowo:

Reszty_gaz_srednia = gaz_srednia - predicted_gaz_srednia

Wykres kwantylowy

Wykres kwantyli próbkowych rozkładu reszt względem kwantyli rozkładu normalnego odpowiedniego rzędu.

Wykres kwantylowy odkłada na osi X kwantyle rozkładu N(0,1) a na osi Y kwanyle standaryzowanego rozkładu obserwowanego rezyduuów (reszt).

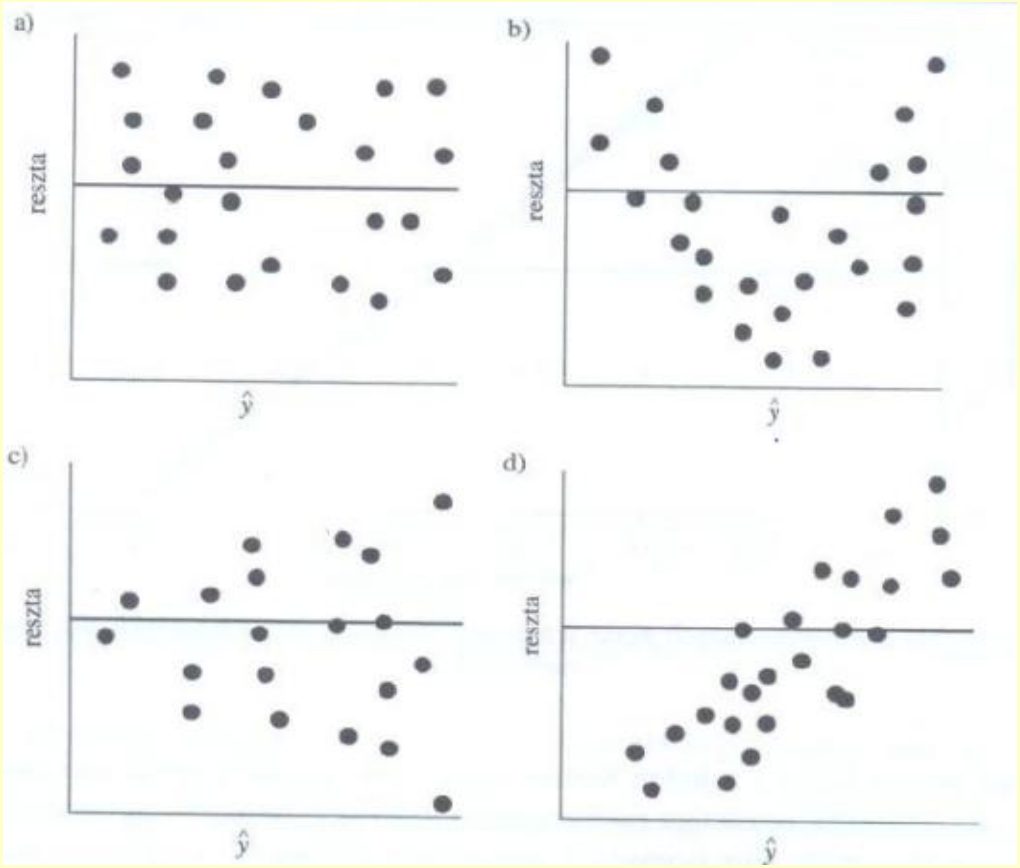


Jeżeli rozkład jest normalny to punkty na wykresie powinny tworzyć linię prostą.
Systematyczne odchylenia od linii prostej mogą wskazywać na nieliniowość.

Wartość standaryzowana r_i: $(r_i - r_{śr})/odstd$

Zerowa średnia, stała wariancja błędów

Wykres standaryzowanych rezyduów (reszt) względem wartości prognozowanych



Źródło: Interent

A . Poprawny wykres, brak widocznych wzorców. Punkty zajmują określony (w pasie o pewnej szerokości) obszar od lewej do prawej.

B. Występuje krzywizna, co świadczy o naruszonym założeniu niezależności (zakrzywiony wzorzec – reszty następne zależą od poprzednich i rozbiegają się ponad ograniczony pas).

C. Wzór lejka świadczy o naruszonym założeniu o stałej wariancji (dla małych wartości prognoz wartości reszt są mniejsze, dla dużych większe) – heteroskedastyczność

D. Obraz wzorca rosnącego od lewej strony do prawej - naruszone założenie o zerowej średniej (dla małych wartości prognoz reszty są mniejsze od zera, podczas, gdy dla dużych wartości reszty są większe od 0).

Weryfikacja założeń. Dodatkowe informacje

Stała wariancja - test heteroskedastyczności.
Opcja ta testuje, czy pierwszy i drugi moment modelu zostały poprawnie podane.
Hipoteza zerowa:

H0: wariancja błędów = constans

Test specyfikacji pierwszego i drugiego momentu. Brak podstaw do odrzucenia hipotezy można traktować jako potwierdzenie homoskedastyczności reszt.

Jeżeli **wykres kwantylowy** nie pokazuje systematycznego odchylenia od linii prostej a **wykres reszty względem wartości przewidywanych** nie pokazuje widocznych wzorców, to możemy wyciągnąć wniosek, że nie ma graficznych dowodów na naruszenie założeń regresji i możemy kontynuować analizę regresji.

Weryfikacja modelu liniowego

- **Merytoryczna**

  Nie może budzić zastrzeżeń merytorycznych

- **Statystyczna**

  Powinien być bardzo dobrze dopasowany do danych

  Wszystkie zmienne objaśniające w modelu muszą być istotne statystycznie

Merytoryczna weryfikacja modelu liniowego

Celem jest stwierdzenie, czy model jest zgodny z wiedzą na temat badanego zjawiska, teorią z dziedziny, której dotyczy oraz zdrowym rozsądkiem.

Bada się:

- Zgodność znaków parametrów strukturalnych ze znakami z analizy korelacji (koincydencja),
- Skale parametrów, czyli wartości bezwzględne parametrów (ocenia się, czy są do przyjęcia),

Ocena merytoryczna wymaga wiedzy o badanych związkach (doświadczenia).

Statystyczna weryfikacja modelu liniowego

Celem jest stwierdzenie, czy model spełnia postulaty sformułowane w teorii ekonometrii i statystyki.

- Suma kwadratów: regresyjna i resztowa
- Stopień dopasowania modelu do danych empirycznych
- Statystyczna istotność modelu (Global test of significance *BETA* = 0) – test F
- Statystyczna istotność wejść i parametrów strukturalnych modelu – testy t

W przypadku modeli nieliniowych weryfikacja sprowadza się do wyznaczenia i weryfikacji pomocniczego modelu liniowego (linearyzacja modeli).

Model liniowy badania reakcji kierującego na zagrożenie na drodze - przykład

Gaz_średnia = f(TTC, nr_próby, hamulec_średnia, skręt_średnia)

= b0 + b1*TTC + b2* nr_próby + b3*hamulec_średnia + b4* skręt_średnia

b0=0,54

b1 = 0,13 (TTC)

b2 = -0,003 (nr_próby) - weryfikacja merytoryczna

b3 = 30,70 (hamulec_średnia)

b4 = -0,64 (skręt_średnia) - weryfikacja merytoryczna

Gaz_srednia = 0,542 + 0,128 * TTC + (-0,0025 nr_proby) + 0,702 * hamulec_srednia – 0,644 * skret_srednia

Model po reestymacji

Gaz_srednia = 0,399 + 0,046 * TTC + 0,377 * hamulec_srednia

Źródło zmienności

Źródło zmienności (*Źródło*) :

- dla modelu (*Model*) dopasowanej regresji liniowej jest przedstawiona regresyjna suma kwadratów SSR (*Sum of Squared Regression)* - określa tę część zmienności *Y*, która jest wyjaśniona przez model:

$$\text{RegresyjnaSK} = SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- dla błędu (*Błąd*) jest przedstawiona resztowa suma kwadratów - suma kwadratów błędów (składnika resztowego) SSE - *Sum of Squared Errors* (inaczej: RSS – Residual Sum of Squares), określa ona wpływ czynnika losowego na zmienność *Y* (zmienność resztowa):

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

Całkowita suma kwadratów (Total Sum of Squares) – zmienność całkowita:
SST = SSR + SSE

Jeżeli regresyjna suma kwadratów jest większa od resztowej sumy kwadratów, to świadczy o dobrym dopasowaniu modelu.

Dopasowanie modelu do danych

W zależnościach poniżej: n – liczba obserwacji, k – liczba estymowanych współczynników kierunkowych modelu.

Współczynnik determinacji R^2 oznaczony przez *R-kwadrat* określa w ilu procentach zmienność *Y* jest wyjaśniona przez model:
$R^2 = SSR/SST = 1 - SSE/SST$ ($R^2 \in [0, 100]$ %)

W przypadku niekorzystnych proporcji szerokości (liczba zmiennych) do długości (liczba obserwacji) analizowanego zbioru danych stosuje się skorygowany współczynnik determinacji (Skor. *R-kwadrat*): $R^2_{adj} = 1 - SSE/SST \cdot (n-1)/(n-k-1)$

Średni błąd oszacowania S_e zmiennej zależnej *Y* wyznacza się ze statystyki Mean Square Error MSE (SAS: *Error* → *Mean Square*) – Pierw. z MSE. Określa on o ile średnio wartości rzeczywiste zmiennej objaśnianej (czyli *Y*) różnią się od wartości przewidywanych przez model.

$$MSE = SSE / (n - k - 1), \quad S_e = \sqrt{MSE}$$

Statystyczna istotność postaci modelu

Global test of significance *BETA* = 0

H0: $\beta_0 = \beta_1 = \ldots = \beta_k = 0$

H1: przynajmniej jeden z parametrów modelu jest istotnie różny od zera

W hipotezie zerowej zakłada się brak związku liniowego określając, że wszystkie parametry strukturalne modelu są równe zero. Weryfikacji H0 dokonuje się w oparciu o statystykę testową *F*.

Wartość statystyki testowej ma rozkład *F* (*F Value*) z liczbą stopni swobody *k* oraz (*n*-*k*-1) :

$$F = [SSR/k] / [SSE/(n-k-1)]$$

Statystyka oraz wyznaczone dla tej statystyki prawdopodobieństwo testowe *p* (*Pr>P*) dostarczają informacji niezbędnych do przeprowadzenia testu istotności modelu liniowego.

Jeśli wartość statystyki testowej przekroczy wartość krytyczną, wyznaczając małe prawdopodobieństwo testowe (mniejsze od założonego poziomu istotności α), model liniowy należy uznać za istotny statystycznie.

Wnioskowanie dla efektów oraz odpowiadających im parametrów strukturalnych modelu

Każda zmienna (parametr) jest testowana indywidualnie.
Hipotezy mają postać:

H0: $\beta_i = 0$
H1: $\beta_i \neq 0$

Hipoteza alternatywna wyznacza dwustronny obszar krytyczny.
Do weryfikacji hipotez służy statystyka *t* z liczbą stopni swobody równą (*n*-*k*-1).

Współczynnik modelu regresji liniowej jest istotny statystycznie (H0 odrzucone), jeżeli prawdopodobieństwo testowe wyznaczonej statystyki testowej jest mniejsze niż poziom istotności.

Interpretacja matematyczna esytmatorów parametrów strukturalnych modelu

Wartość współczynnika informuje jak zmienia się wartość zmiennej objaśnianej (celu), jeżeli wartość analizowanej zmiennej objaśniającej (wejściowej) zmienia się o jeden, przy ustalonych (niezmieniających się) wartościach pozostałych wejść.

$$Y = 13,6 + 22,7 X_1 + (-4,6 X_2)$$

Gdy wartość X_1 wzrośnie o jeden, to przy niezmieniającej się wartości X_2 wartość zmiennej Y zwiększy się (znak +) o 22,7.

Jeżeli wartość X_2 wzrośnie o jeden, to przy niezmieniającej się wartości X_1 wartość zmiennej Y zmniejszy się (znak-) o 4,6.

W interpretacji należy uwzględnić dopuszczalny zakres wartości i interpretację fizyczną (ocena merytoryczna) cechy oraz zdrowy rozsądek.

Metody wyznaczania modelu

Parametry modelu są wyznaczane metodą maksimum wiarygodności.

Może być stosowana metoda selekcji, która pomaga w doborze „najlepszych zmiennych”

- **Backward** – „do tyłu”; zaczyna od pełnego zbioru zmiennych niezależnych (efektów) i eliminuje mało istotne
- **Forward** – „do przodu”; zaczyna od stałej i dodaje istotne zmienne niezależne
- **Stepwise** – „mądre kroki”, zaczyna od stałej i dodaje istotne zmienne zależne, przy czym możliwe jest usunięcie z modelu dodanego wcześniej efektu, jeżeli nie ma on dużego wpływu na zmienną prognozowaną

Uwagi

- Jeżeli jest wiele zmiennych wejściowych kandydujących do objaśniania modelu, można zastosować iteracyjne metody selekcji (*forward, backward, stepwise selection*).
- Zaleca się zbadać analizowany zbiór cech na okoliczność nadmiernej korelacji pomiędzy parami cech (macierz korelacji) oraz współliniowości (indeks uwarunkowania – wskaźnik warunku).
- Związki między zmienną zależną i zbiorem zmiennych niezależnych niekoniecznie musza być liniowe. Metoda regresji liniowej może być stosowana, jeżeli zmienne „powodujące problemy” można przekształcić tak, aby w ostateczności otrzymać model liniowy.
- Zaleca się dokonywać normalizacji (standaryzacja, unitaryzacja, przekształcenie ilorazowe) zmiennych w przypadku dużych różnic w zakresach wartości i różnych mian (nieporównywalność wartości).