Modele liniowe

Michał Kos

Uniwersytet Wrocławski

Plan wykładu

1. Tabela analizy wariancji dla regresji liniowej prostej
2. Współczynnik determinacji oraz jego modyfikacja
3. Diagnostyka modelu
4. Środki zaradcze

Table of Contents

Analiza wariancji – Analysis of Variance (ANOVA)

Analiza wariancji dla regresji liniowej jest metodą uporządkowania zmienności w wektorze odpowiedzi $Y = (Y_1, ..., Y_n)'$.

Analiza wariancji – Analysis of Variance (ANOVA)

Analiza wariancji dla regresji liniowej jest metodą uporządkowania zmienności w wektorze odpowiedzi $Y = (Y_1, ..., Y_n)'$.
Zgodnie z teoretycznym modelem:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

zmienność w wektorze Y ma dwa źródła:

- zmienność wynikająca z błędów losowych $\epsilon_i \sim N(0, \sigma^2)$ powiązana z wielkością parametru σ^2,
- zmienność, której źródłem są różnice w wartościach oczekiwanych zmiennych objaśnianych $E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i$.

Źródła zmienności w wektorze odpowiedzi (2)

By lepiej zrozumieć dualną naturę zmienności w wektorze Y rozważmy dwie skrajne sytuacje:

1. Przyjmijmy, że $\beta_1 = 0$. Wówczas

$$Y_i = \beta_0 + \epsilon_i \sim N(\beta_0, \sigma^2)$$

Widzimy zatem, że wartość oczekiwana dla dowolnego i jest taka sama $E(Y_i) = \mu_i = \beta_0$, oraz że jedynym źródłem zmienności w wektorze wynikowym Y są losowe błędy ϵ_i (własnosci tego typu danych badali Państwo na PSP).

Źródła zmienności w wektorze odpowiedzi (2)

By lepiej zrozumieć dualną naturę zmienności w wektorze Y rozważmy dwie skrajne sytuacje:

1. Przyjmijmy, że $\beta_1 = 0$. Wówczas

$$Y_i = \beta_0 + \epsilon_i \sim N(\beta_0, \sigma^2)$$

   Widzimy zatem, że wartość oczekiwana dla dowolnego i jest taka sama $E(Y_i) = \mu_i = \beta_0$, oraz że jedynym źródłem zmienności w wektorze wynikowym Y są losowe błędy ϵ_i (własnosci tego typu danych badali Państwo na PSP).

2. Z drugiej strony jeżeli błędy przy pomiarach są pomijalne czyli $\sigma^2 = 0$, wtedy

$$Y_i = E(Y_i) = \beta_0 + \beta_1 X_i$$

   W takiej sytuacji relacja pomiędzy Y_i, a X_i jest ściśle liniowa, a zmienność w wektorze Y nie ma losowego charakteru, lecz jej źródłem jest zmienność wartości regresora X.

Analiza wariancji

Oznaczmy przez $\hat{Y}_i$ model stowarzyszony z estymatorami $\hat{\beta}_0$ oraz $\hat{\beta}_1$, czyli $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ i wprowadźmy następujące oznaczenia:

- $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \|Y - \bar{Y}\mathbb{1}_n\|^2$ – statystyka opisujaca całkowitą zmienność w wektorze Y, $(\mathbb{1}_n = (1, ..., 1)' \in \mathbb{R}^n)$,
- $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \|Y - \hat{Y}\|^2$ – statystyka opisujaca zmienność w wektorze Y, wynikającą z błędów modelu $\hat{Y}_i$ "=" suma kwadratów residuów $(\hat{Y} = (\hat{Y}_1, ..., \hat{Y}_n)')$,
- $SSM = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \|\hat{Y} - \bar{Y}\mathbb{1}_n\|^2$ – statystyka opisujaca zmienność w wektorze Y, wyjaśnioną przy pomocy modelu $\hat{Y}_i$,

"SS" – Sum of squares, "T" – total, "M" – model, "E" – error.

Analiza wariancji

Oznaczmy przez $\hat{Y}_i$ model stowarzyszony z estymatorami $\hat{\beta}_0$ oraz $\hat{\beta}_1$, czyli $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ i wprowadźmy następujące oznaczenia:

- $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \|Y - \bar{Y}\mathbb{1}_n\|^2$ – statystyka opisujaca całkowitą zmienność w wektorze Y, $(\mathbb{1}_n = (1, ..., 1)' \in \mathbb{R}^n)$,
- $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \|Y - \hat{Y}\|^2$ – statystyka opisujaca zmienność w wektorze Y, wynikającą z błędów modelu $\hat{Y}_i$ "=" suma kwadratów residuów $(\hat{Y} = (\hat{Y}_1, ..., \hat{Y}_n)')$,
- $SSM = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \|\hat{Y} - \bar{Y}\mathbb{1}_n\|^2$ – statystyka opisujaca zmienność w wektorze Y, wyjaśnioną przy pomocy modelu $\hat{Y}_i$,

"SS" – Sum of squares, "T" – total, "M" – model, "E" – error.

"Tw. Pitagorasa"

Zachodzi następująca relacja pomiędzy powyższymi statystykami

$$SST = SSM + SSE$$

Zmienność całkowita = Zmienność wyjaśnona przez model + zmienność błędów

Stopnie swobody i średnia suma kwadratów

Do skonstruowania tabeli wariancji potrzebujemy dodatkowo informacji o liczbie stopni swobody (df - degrees of freedom) rozważanych statystyk. Na ich podstawie możemy wyznaczyć statystyki zwane średnimi sumami kwadratów (MS - mean squares).

Source	SS	df	MS
Model	$SSM = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$dfM = 1$	$MSM = SSM/dfM$
Error	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$dfE = n - 2$	$MSE = SSE/dfE = s^2$
Total	$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$dfT = n - 1$	$MST = SST/dfT$
Relations	SST= SSM+SSE	dfT=dfM+dfE	MST≠MSM+MSE

Średnia suma kwadratów - własności

Z naszej perspektywy ważne są własności statystyk MSM oraz MSE. Okazuje się że:

$$E(MSM) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 \qquad \text{oraz} \qquad E(MSE) = E(s^2) = \sigma^2$$

Średnia suma kwadratów - własności

Z naszej perspektywy ważne są własności statystyk MSM oraz MSE. Okazuje się że:

$$E(MSM) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 \qquad \text{oraz} \qquad E(MSE) = E(s^2) = \sigma^2$$

Dodatkowo w problemie testowania istotności współczynnika β_1:

$$H_0 : \ \beta_1 = 0 \qquad vs \qquad H_1 : \ \beta_1 \neq 0$$

gdy zachodzi hipoteza zerowa, statystyka:

$$F = MSM/MSE$$

pochodzi z rozkładu Fishera–Snedecora z *dfM* i *dfE* stopniami swobody ($F \sim F(dfM, dfE) = F(1, n-2)$).
Z drugiej strony, jeżeli zachodzi alternatywa $\beta_1 \neq 0$, to MSM jest średnio większa od MSE i w konsekwencji statystyka F przyjmuje duże wartości.

Test F

Na podstawie powyższych obserwacji skonstruowany jest tzw. test F testujący czy β_1 jest różna od 0:

$$H_0: \quad \beta_1 = 0 \qquad vs \qquad H_1: \quad \beta_1 \neq 0$$

Odrzucamy hipotezę zerową, gdy $F = MSM/MSE > F_c$, gdzie $F_c = F^*(1-\alpha, 1, n-2)$ jest kwantylem rzędu $1-\alpha$ z rozkładu Fishera–Snedecora z *dfM* i *dfE* stopniami swobody.

Test F

Na podstawie powyższych obserwacji skonstruowany jest tzw. test F testujący czy β_1 jest różna od 0:

$$H_0: \quad \beta_1 = 0 \qquad vs \qquad H_1: \quad \beta_1 \neq 0$$

Odrzucamy hipotezę zerową, gdy $F = MSM/MSE > F_c$, gdzie $F_c = F^*(1-\alpha, 1, n-2)$ jest kwantylem rzędu $1-\alpha$ z rozkładu Fishera–Snedecora z *dfM* i *dfE* stopniami swobody.
Zwykle, wnioskowanie dokonywane jest na podstawie p-wartości: $p = P(z > F)$, gdzie $z \sim F(dfM, dfE) = F(1, n-2)$

Alternatywne spojrzenie na test F

Na test F można również spojrzeć w nieco inny sposób. Przyjmijmy że chcemy porównać dwa modele:

H_0 : dane pochodzą z modelu $Y_i = \beta_0 + \epsilon_i$ (tzw. model zredukowany)

H_1 : dane pochodzą z modelu $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ (tzw. model pełny)

W naturalny sposob, powyższe zagadnienie testowe sprowadza się do badania istotnosci parametru β_1.

Alternatywne spojrzenie na test F

Na test F można również spojrzeć w nieco inny sposób. Przyjmijmy że chcemy porównać dwa modele:

H_0 : dane pochodzą z modelu $Y_i = \beta_0 + \epsilon_i$ (tzw. model zredukowany)

H_1 : dane pochodzą z modelu $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ (tzw. model pełny)

W naturalny sposob, powyższe zagadnienie testowe sprowadza się do badania istotnosci parametru β_1.

Porównania modeli możemy dokonać np. przy użyciu statystyk SSE(R) i SSE(F) (SSE odpowiednio dla modelu zredukowanego (R – "reduced") i pełnego (F – "full")).

Alternatywne spojrzenie na test F (2)

Statystyka F wyraża się przy pomocy SSE(R) i SSE(F) w następujący sposób

$$F = \frac{\left(SSE(R) - SSE(F)\right) / \left(dfE(R) - dfE(F)\right)}{MSE(F)}$$

Alternatywne spojrzenie na test F (2)

Statystyka F wyraża się przy pomocy SSE(R) i SSE(F) w następujący sposób

$$F = \frac{(SSE(R) - SSE(F)) / (dfE(R) - dfE(F))}{MSE(F)}$$

Uzasadnienie:

$$SSE(R) = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = SST; \qquad SSE(F) = SSE$$

$$dfE(R) = n - 1; \qquad dfE(F) = dfE = n - 2$$

$$MSE(F) = MSE$$

zatem:

$$SSE(R) - SSE(F) = SST - SSE = SSM; \qquad dfE(R) - dfE(F) = 1 = dfM$$

podstawiając:

$$F = MSM/MSE$$

Tabela analizy wariancji (ANOVA Table)

Dane potrzebne do wykonania testu F zwykle przedstawiane są w tzw. tabeli analizy wariancji, o następującej postaci:

	df	SS	MS	F	p-value
Model	dfM	SSM	MSM	$\mathrm{F} = \mathrm{MSM}/\mathrm{MSE}$	$p = P(z > F)$
Error	dfE	SSE	MSE		

Uwagi do testu F

- gdy zachodzi hipoteza alternatywna statystyka F pochodzi z niecentralnego rozkładu Fishera–Snedecora,
- umożliwia to wyznaczenie funkcji mocy testu,
- do testowania $H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$, wykorzystywaliśmy wcześniej statystykę $T = \hat{\beta}_1 / s(\hat{\beta}_1)$. Można pokazać, że $F = T^2$, i w konsekwencji obie metody są równoważne (np. dają te same p-wartości),
- metody te przestają być równoważne, gdy mamy więcej niż jeden regresor! Wówczas z każdym regresorem stowarzyszona jest inna statystyka T badajaca jego istotność, a test F bada równocześnie istotność wszystkich regresorów (o tym później).

Table of Contents

Współczynnik determinacji R^2

Współczynnik determinacji R^2 jest miarą jakości dopasowania modelu. Mówi on o tym, jaką część całkowitej zmiennosci w wektorze Y (SST) stanowi zmienność wyjaśniona przez model (SSM).

$$R^2 = SSM/SST = 1 - SSE/SST$$

Przyjmuje on wartości od 0 do 1 (czasami wyrażany jest w skali procentowej).

Współczynnik determinacji R^2

Współczynnik determinacji R^2 jest miarą jakości dopasowania modelu. Mówi on o tym, jaką część całkowitej zmiennosci w wektorze Y (SST) stanowi zmienność wyjaśniona przez model (SSM).

$$R^2 = SSM/SST = 1 - SSE/SST$$

Przyjmuje on wartości od 0 do 1 (czasami wyrażany jest w skali procentowej).

- W regresji liniowej prostej R^2 jest tożsamy z kwadratem próbkowej korelacji persona pomiędzy zmiennymi zależną i niezależną,
- Przy użyciu statystyki R^2 możemy porównywać modele o tej samej liczbie regresorów (również w regresji wielorakiej),
- gdy porównujemy modele o różnej liczbie regresorów, zamiast R^2 używa się tzw. modyfikowanego współczynnika determinacji:

$$\tilde{R}^2 = 1 - MSE/MST$$

Table of Contents

Diagnostyka modelu

Bardzo ważnym zagadnieniem jest badanie tego, czy użycie metody regresji liniowej jest adekwatne do analizowanego zbioru danych.

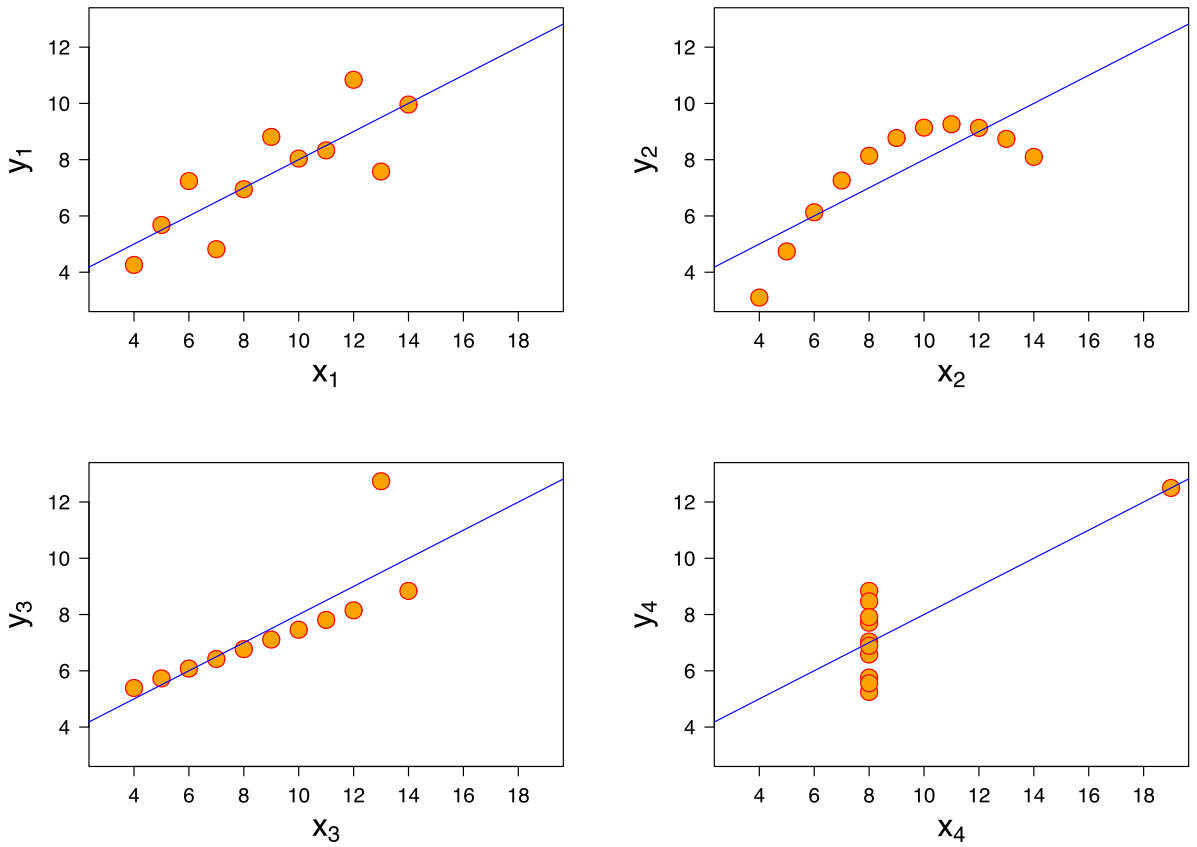
Diagnostyka modelu

Bardzo ważnym zagadnieniem jest badanie tego, czy użycie metody regresji liniowej jest adekwatne do analizowanego zbioru danych.

Zwykle dokonujemy tego poprzez:

1. badanie własności zmiennej niezależnej.
2. badanie własności zmiennej zależnej, uwzględniające wpływ zmiennej niezależnej.

Zbiór danych Anscombe (1973)



Rysunek: Zastosowanie regresji liniowej prostej do powyższych zbiorów danych zwraca modele o identycznych własnościach. Analiza zbioru w R: https://rpubs.com/debosruti007/anscombeQuartet.

Źródło: By Anscombe.svg: Schutz(label using subscripts): Avenue - Anscombe.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=9838454

Badanie własności zmiennej niezależnej X

Formalnie, w teoretycznym modelu liniowym, o zmiennej niezależnej nic nie zakładamy. Wiemy jednak, że ma ona wpływ na własności estymatorów, np. $\bar{X}$ występuje we wzorze na $\hat{\beta}_0$, a suma $\sum(X_i - \bar{X})^2$ występuje w mianowniku wzoru na wariancję estymatora $\hat{\beta}_1$. Dlatego zawsze warto zbadać własności zmiennej niezależnej.

Badanie własności zmiennej niezależnej X

Formalnie, w teoretycznym modelu liniowym, o zmiennej niezależnej nic nie zakładamy. Wiemy jednak, że ma ona wpływ na własności estymatorów, np. $\bar{X}$ występuje we wzorze na $\hat{\beta}_0$, a suma $\sum(X_i - \bar{X})^2$ występuje w mianowniku wzoru na wariancję estymatora $\hat{\beta}_1$. Dlatego zawsze warto zbadać własności zmiennej niezależnej.

Dokonujemy tego przez analizę:

- podstawowych charakterystyk: średnia, mediana, kwartyle, minimum, maksimum, wariancja, rozstęp kwartylowy, itd.
- rysunków: histogram, boxplot, wykres kwantylowo-kwantylowy, wykres X w zależności od kolejności elementów w wektorze, itd.

Badanie własności zmiennej niezależnej X (2)

Powyższa analiza pozwala odpowiedzieć na pytania takie jak:

- czy rozkład X jest symetryczny czy skośny?
- czy rozkład X jest normalny? – jest to czasami korzystne w analizie modelu liniowego,
- czy w zbiorze występują obserwacje odstające? – mogą one (w powiązaniu z odpowiadającą wartością Y_i) silnie wpływać na prostą regresji.
- czy wektor X jest uporządkowany ze względu na kolejność pojawiania się obserwacji w zbiorze? – może dawać to dodatkową informację np. o strukturze eksperymentu.

Badanie własności zmiennej zależnej Y

Badanie własności zmiennej objaśnianej Y w oderwaniu od zmiennej objaśniającej X może być mało informatywne. Wynika to z faktu, iż zgodnie z założeniami teoretycznego modelu, elementami wektora Y są niezależne zmienne losowe z rozkładów normalnych o tej samej wariancji σ^2, ale różnych wartościach oczekiwanych $E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i$.

Badanie własności zmiennej zależnej Y

Badanie własności zmiennej objaśnianej Y w oderwaniu od zmiennej objaśniającej X może być mało informatywne. Wynika to z faktu, iż zgodnie z założeniami teoretycznego modelu, elementami wektora Y są niezależne zmienne losowe z rozkładów normalnych o tej samej wariancji σ^2, ale różnych wartościach oczekiwanych $E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i$.

Uwzględnienie wpływu regresora X umożliwia lepsze i głębsze zrozumienie własności wektora odpowiedzi Y.

Badanie własności zmiennej zależnej Y (2)

Przypomnijmy model teoretyczny:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i; \quad \epsilon_i - \text{niezal. zm. los. z rozkładu } N(0, \sigma^2)$$

Potrzebujemy zbadać kilka własności:

1. liniowość – $E(Y_i) = \beta_0 + \beta_1 X_i$,
2. własności ciągu zm. losowych $\epsilon_1, ..., \epsilon_n$:
   - niezależność
   - stałość wariancji σ^2,
   - normalność,
3. występowanie obserwacji odstających.

W eksploracji liniowości ważna jest prosta regresji:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

a w badaniu własności ciągu błędów losowych $\epsilon_1, ..., \epsilon_n$, istotny jest ciąg residuów:

$$e_i = Y_i - \hat{Y}_i \quad \Rightarrow \quad Y_i = \hat{Y}_i + e_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

Czy relacja jest liniowa? ($E(Y_i) = \beta_0 + \beta_1 X_i$)

Podstawowym narzędziem diagnostycznym są rysunki:

- Y vs X z dodaną prostą regresji $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$,
- e vs X.

Narysowanie wykresu rozrzutu (X_i, Y_i) z dodaną prostą regresji, daje możliwość szybkiej oceny liniowości relacji.

Czy relacja jest liniowa? ($E(Y_i) = \beta_0 + \beta_1 X_i$)

Podstawowym narzędziem diagnostycznym są rysunki:

- Y vs X z dodaną prostą regresji $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$,
- e vs X.

Narysowanie wykresu rozrzutu (X_i, Y_i) z dodaną prostą regresji, daje możliwość szybkiej oceny liniowości relacji.

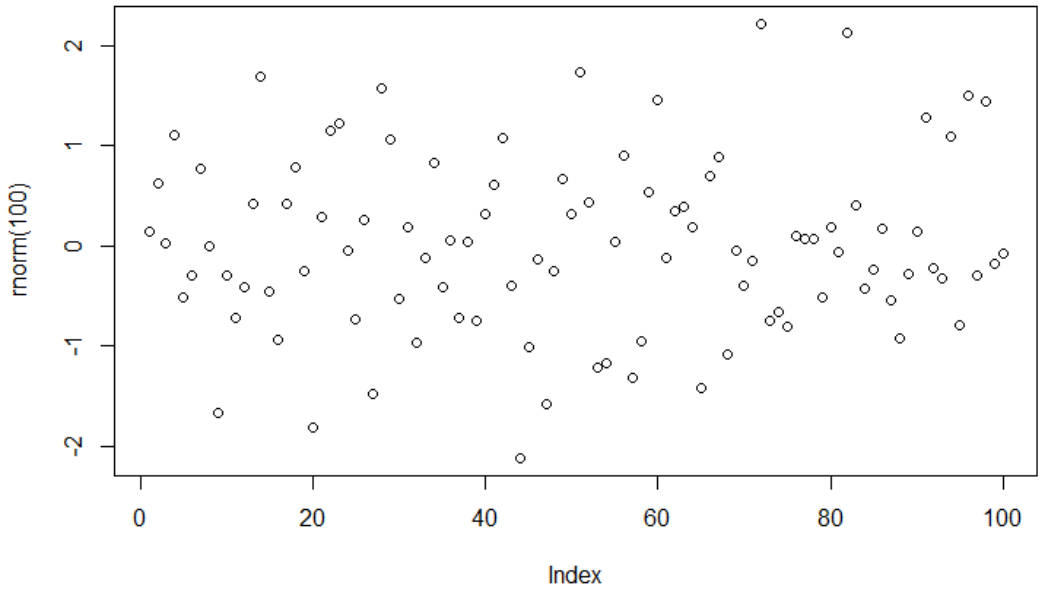
Narysowanie zależności między wartościami zmiennej niezależnej $(X = (X_1, ..., X_n)')$ i odpowiadajacymi im residuami $(e = (e_1, ..., e_n)')$ uwypukla odstępstwa od liniowej struktury.

Własności ciągu zm. losowych $\epsilon_1, ..., \epsilon_n$

Tak jak wspomniano do badania własności błędów losowych $\epsilon_1, ..., \epsilon_n$ przydatna jest analiza własności wektora residuów $e = (e_1, ..., e_n)'$.
Jeżeli dane pochodzą z modelu teoretycznego wówczas błędy losowe są:

- niezależne,
- o wart. oczekiwanej równej 0 i stałej wariancji σ^2,
- z rozkładu normalnego

Oznacza to, że oprócz drgania wartości wokół średniej 0 i stałego rozrzutu, niepowinniśmy obserwować jakiejkolwiek prawidłowości lub struktury w błędach losowych. Istnienie takiej struktury sugeruje łamanie założeń modelu. Podobnego zachowania oczekujemy od residuów.



Niezależność

Łamanie założenia o niezależności błędów losowych możemy zdiagonozować np. na podstawie rysunku e vs kolejność wykonywania pomiarów (kolejność pojawiania się w zbiorze danych).

Jeżeli obserwujemy jakąś strukturę np. trend i/lub cykliczność świadczy to o tym, że błędy losowe są zależne.

Niezależność

Łamanie założenia o niezależności błędów losowych możemy zdiagonozować np. na podstawie rysunku e vs kolejność wykonywania pomiarów (kolejność pojawiania się w zbiorze danych).
Jeżeli obserwujemy jakąś strukturę np. trend i/lub cykliczność świadczy to o tym, że błędy losowe są zależne.

Przykład
Badamy wpływ czasu na południowe temperatury 01.08. w kolejnych latach (np. występowanie efektu cieplarnianego). Jeżeli z każdym rokiem stowarzyszony jest jeden pomiar to raczej nie bedziemy mieli wątpliwości co do niezależności pomiarów (długi odstęp czasu pomiędzy pomiarami). Jeżeli jednak w każdym roku będziemy dokonywali dwóch pomiarów np. o 12:00 i 13:00 to wydaje się naturalne, że będą one ze sobą skorelowane, gdyż zmiany pogodowe rzadko zachodzą tak gwałtownie.

Stałość wariancji (homoskedastyczność)

Wariancja poszczegolnych błędów losowych może zależeć od zmiennej objaśniającej. W takiej sytuacji warto przeanalizować wykres e vs X.

Stałość wariancji (homoskedastyczność)

Wariancja poszczegolnych błędów losowych może zależeć od zmiennej objaśniającej. W takiej sytuacji warto przeanalizować wykres $\underline{e \text{ vs } X}$.

Przykład
Problemy z nadciśnieniem pojawiają się zwykle u osób starszych. Oznacza to, że dla podpopulacji ludzi w wieku powyżej $X = 60$ lat rozrzut ciśnienia (Y), może być większy w zestawieniu z podpopulacją ludzi młodych np. do $X = 20$ lat.

Rozkład normalny

Aby zbadać to czy ciąg zm. losowych $\epsilon_1, ..., \epsilon_n$ pochodzi z rozkładu normalnego analizujemy histogram i wykres kwantylowo–kwantylowy dla wektora residuów e. Możemy również skorzystać z testów np. test Shapiro–Wilka.

Rozkład normalny

Aby zbadać to czy ciąg zm. losowych $\epsilon_1, ..., \epsilon_n$ pochodzi z rozkładu normalnego analizujemy histogram i wykres kwantylowo–kwantylowy dla wektora residuów e. Możemy również skorzystać z testów np. test Shapiro–Wilka.

Jak wcześniej wspomniano istotne jest to, czy dane zbyt mocno nie odbiegają od rozkładu normalnego. Jeżeli ϵ_i nie pochodzą z rozkładu normalnego, tylko z rozkładu o podobnych własnościach (np. symetryczność, stała wariancja) to przedziały ufności i testy zwykle mają dobre własności.

Obserwacje odstające

Do analizy tego czy w danych występują obserwacje odstające służą wykresy:

- Y vs X,
- e vs X.

Obserwacje odstające

Do analizy tego czy w danych występują obserwacje odstające służą wykresy:

- Y vs X,
- e vs X.

Obserwacje odstające mogą (nie muszą) znacząco wpływać na wartości estymatorów!!!

Zwykle zwiększają wartość estymatora s^2. W konsekwencji otrzymujemy szerokie przedziały ufności $(\hat{\beta}_i \pm t_c s(\hat{\beta}_i)$, gdzie $s^2(\hat{\beta}_i) = s^2 f_i(X))$ oraz trudniej jest odrzucić hipotezy statystyczne np. $\beta_0 = 0$, czy $\beta_1 = 0$ $(T = \hat{\beta}_i / s(\hat{\beta}_i))$.

Testy

W diagnostyce modelu oprócz badania własności przy pomocy wykresów, możemy stosować również różne testy, np.:

- test Durbina–Watsona badający niezależność błędów (dwtest lmtest)
- test stałości wariancji Breuscha–Pagana (bptestlmtest)

Jednakże własności tych testów często silnie zależą od wielkości próby.

Dodatkowo, w przeciwieństwie do wykresów, nie dostarczają żadnej informacji na temat aplikowalnych środków zaradczych.

Table of Contents

Metody zaradcze

W poprzedniej część wykładu poznaliśmy metody diagnostyczne modelu. Co jednak zrobić, gdy obserwujemy, że założenia modelu liniowego są łamane?

Metody zaradcze

W poprzedniej część wykładu poznaliśmy metody diagnostyczne modelu. Co jednak zrobić, gdy obserwujemy, że założenia modelu liniowego są łamane?

Niejednokrotnie możemy doprowadzić za pomocą pewnych przekształceń zmiennych X i/lub Y, lub uogólnienia idei zawartej w modelu liniowym do sytuacji, w której możemy zastosować modele liniowe.

Brak "liniowości" $(E(Y_i) = \beta_0 + \beta_1 X_i)$

Przykład 1 (transformacja X)

Załóżmy, że związek pomiędzy X i Y jest następującej postaci:

$$Y_i = \beta_0 + \beta_1 \sqrt{X_i} + \epsilon_i$$

Wówczas $E(Y_i) = \beta_0 + \beta_1 \sqrt{X_i}$. W oczywisty sposób relacja pomiędzy $E(Y)$ i X nie jest liniowa.

Brak "liniowości" ($E(Y_i) = \beta_0 + \beta_1 X_i$)

Przykład 1 (transformacja X)

Załózmy, że związek pomiędzy X i Y jest następującej postaci:

$$Y_i = \beta_0 + \beta_1 \sqrt{X_i} + \epsilon_i$$

Wówczas $E(Y_i) = \beta_0 + \beta_1 \sqrt{X_i}$. W oczywisty sposób relacja pomiędzy $E(Y)$ i X nie jest liniowa.

Łatwo jednak zauważyć że dla nowej zmiennej niezależne postaci $\tilde{X} = \sqrt{X}$ relacja pomiędzy $E(Y)$ i $\tilde{X}$ już jest liniowa:

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \epsilon_i$$

$$E(Y_i) = \beta_0 + \beta_1 \tilde{X}_i$$

Widzimy zatem, że odpowiednie przekształcenie nałożone na zmienną objaśniającą X umożliwia zastosowanie poznanej teorii.

Brak "liniowości" ($E(Y_i) = \beta_0 + \beta_1 X_i$) (regresja liniowa wieloraka)

Przykład 2

Załóżmy, że związek pomiędzy X i Y jest następującej postaci:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

Wówczas zależność pomiędzy $E(Y)$ i X jest relacją kwadratową.

Brak "liniowości" ($E(Y_i) = \beta_0 + \beta_1 X_i$) (regresja liniowa wieloraka)

Przykład 2

Załózmy, że związek pomiędzy X i Y jest następującej postaci:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

Wówczas zależność pomiędzy $E(Y)$ i X jest relacją kwadratową.

Do analizy tego typu danych możemy skorzystać z uogólnienia regresji liniowej prostej do tzw. regresji liniowej wielorakiej (w której mamy więcej niż jedną zmienną objaśniającą). Jeżeli wprowadzimy nową zmienną objaśniającą $\tilde{X} = X^2$, wówczas powyższy model możemy przedstawić w następującej postaci:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \tilde{X}_i + \epsilon_i$$

Wówczas wartość oczekiwana Y zależy liniowo od zmiennych X_i oraz $\tilde{X}_i$:

$$E(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 \tilde{X}_i$$

Brak "liniowości" $(E(Y_i) = \beta_0 + \beta_1 X_i)$

Przykład 3 (przekształcenie Y)

Załóżmy, że związek pomiędzy X i Y jest następującej postaci:

$$Y_i = \beta_0 \exp(\beta_1 X_i) + \epsilon_i$$

Brak "liniowości" $(E(Y_i) = \beta_0 + \beta_1 X_i)$

Przykład 3 (przekształcenie Y)

Załóżmy, że związek pomiędzy X i Y jest następującej postaci:

$$Y_i = \beta_0 \exp(\beta_1 X_i) + \epsilon_i$$

Wówczas $E(Y_i) = \beta_0 \exp(\beta_1 X_i)$ i w konsekwecji $log(E(Y_i)) = log(\beta_0) + \beta_1 X_i$. Możemy w takiej sytuacji rozważyć liniowy model postaci:

$$\log(Y_i) = \tilde{\beta}_0 + \tilde{\beta}_1 X_i + \tilde{\epsilon}_i$$

Brak "liniowości" $(E(Y_i) = \beta_0 + \beta_1 X_i)$

Przykład 3 (przekształcenie Y)

Załóżmy, że związek pomiędzy X i Y jest następującej postaci:

$$Y_i = \beta_0 \exp(\beta_1 X_i) + \epsilon_i$$

Wówczas $E(Y_i) = \beta_0 \exp(\beta_1 X_i)$ i w konsekwecji $log(E(Y_i)) = log(\beta_0) + \beta_1 X_i$. Możemy w takiej sytuacji rozważyć liniowy model postaci:

$$\log(Y_i) = \tilde{\beta}_0 + \tilde{\beta}_1 X_i + \tilde{\epsilon}_i$$

Warto jednak zauważyć, że zmieniliśmy w tym przypadku założenia dotyczące zachowania błędu losowego.

Brak "liniowości" $(E(Y_i) = \beta_0 + \beta_1 X_i)$

Transformacja Boxa–Coxa

W poprzednim przykładzie przed zastosowaniem regresji liniowej, nałożyliśmy przekształcenie na zmienną Y. Ważnym pytaniem jest to jaką funkcję nałożyć na Y aby otrzymać liniową zależność.

Brak "liniowości" $(E(Y_i) = \beta_0 + \beta_1 X_i)$

Transformacja Boxa–Coxa

W poprzednim przykładzie przed zastosowaniem regresji liniowej, nałożyliśmy przekształcenie na zmienną Y. Ważnym pytaniem jest to jaką funkcję nałożyć na Y aby otrzymać liniową zależność.

Tzw. Transformacja Boxa–Coxa umożliwia wybór optymalnego przekształcenia. Dopasowuje ona do danych model postaci:

$$f_\lambda(Y) = \tilde{Y} = \beta_0 + \beta_1 X_i + \epsilon_i$$

gdzie $\tilde{Y} = Y^\lambda$ lub $\tilde{Y} = (Y^\lambda - 1)/\lambda$.
Następnie przy użyciu metody największej wiarogodności estymuje optymalną wartość parametru λ.

Brak "liniowości" $(E(Y_i) = \beta_0 + \beta_1 X_i)$

Transformacja Boxa–Coxa

Ważne szczególne przypadki:

- $\lambda = 1$; $\tilde{Y} = Y$ (brak przekształcenia)
- $\lambda = 0.5$; $\tilde{Y} = \sqrt{Y}$ (pierwiastek)
- $\lambda = -0.5$; $\tilde{Y} = Y^{-0.5}$ ($1/$pierwiastek)
- $\lambda = -1$; $\tilde{Y} = Y^{-1}$ (odwrotność)
- $\lambda \approx 0$; $\tilde{Y} = (Y^{\lambda} - 1)/\lambda$ (granicznym przekształceniem jest logarytm)

Brak "liniowości" $(E(Y_i) = \beta_0 + \beta_1 X_i)$

Transformacja Boxa–Coxa

Ważne szczególne przypadki:

- $\lambda = 1$; $\tilde{Y} = Y$ (brak przekształcenia)
- $\lambda = 0.5$; $\tilde{Y} = \sqrt{Y}$ (pierwiastek)
- $\lambda = -0.5$; $\tilde{Y} = Y^{-0.5}$ ($1/$pierwiastek)
- $\lambda = -1$; $\tilde{Y} = Y^{-1}$ (odwrotność)
- $\lambda \approx 0$; $\tilde{Y} = (Y^{\lambda} - 1)/\lambda$ (granicznym przekształceniem jest logarytm)

W R funkcja za pomocą której możemy wyznaczyć optymalną wartość parametru λ nazywa się boxcox() (z pakietu MASS)

Korelacja między kolejnymi obserwacjami (błędami)

Brak niezależności

W pracy z danymi często możemy się spotkać z sytuacją, w której zakładanie niezależności pomiarów przeczy zdrowemu rozsądkowi ze względu na konstrukcję eksperymentu.

Korelacja między kolejnymi obserwacjami (błędami)

Brak niezależności
W pracy z danymi często możemy się spotkać z sytuacją, w której zakładanie niezależności pomiarów przeczy zdrowemu rozsądkowi ze względu na konstrukcję eksperymentu.

Z taką sytuacją możemy się spotkać gdy wielokrotnie mierzymy pewną wielkość stowarzyszoną z pewnym obiektem oraz chcemy badać zmiany własności tej wielkości w czasie. Taka metoda badania efektów nazywana jest **pomiarami wieloktrotnymi (repeted measures, longitudinal studies)** i ze względu na to, że badamy wielokrotnie ten sam obiekt trudno jest oczekiwać, że obserwacje nie będą ze sobą skorelowane i w konsekwencji zależne.

Korelacja między kolejnymi obserwacjami (błędami)

Przykład – wpływ leku na obniżenie ciśnienia

W eksperymencie badano wpływ nowego leku na obniżenie ciśnienia pacjenta. W tym celu podzielono w losowy sposób grupę pacjentów na dwie podgrupy. Pierwszą podgrupę stanowiły osoby które przyjmowały nowy lek, a druga podgrupa była grupą kontrolną – przyjmowała placebo. Pacjenci przyjmowali lek/placebo przez 4 tygodnie. Ciśnienie było mierzone 4 razy: na początku oraz po 1, 2 i 4 tygodniach.

Korelacja między kolejnymi obserwacjami (błędami)

Przykład – wpływ leku na obniżenie ciśnienia

W eksperymencie badano wpływ nowego leku na obniżenie ciśnienia pacjenta. W tym celu podzielono w losowy sposób grupę pacjentów na dwie podgrupy. Pierwszą podgrupę stanowiły osoby które przyjmowały nowy lek, a druga podgrupa była grupą kontrolną – przyjmowała placebo. Pacjenci przyjmowali lek/placebo przez 4 tygodnie. Ciśnienie było mierzone 4 razy: na początku oraz po 1, 2 i 4 tygodniach.

Metody analizy tego typu danych:

- Ogólny model liniowy (zmiana założeń na relacje miedzy ϵ-ami),
- Modele liniowe z efektami losowymi.

Metody analizy tego typu danych poznają Państwo na wykładach Zaawansowane modele liniowe (wstęp) i Complex data (cały wykład).

Brak stałości wariancji błędów losowych

W sytuacji, w której obserwujemy np. liniową zależność pomiędzy wariancją ϵ_i a zmienną X_i, możemy zastosować tzw. ważoną metodę namniejszych kwadratów do estymacji parametrów modelu. Jest ona wbudowana w funkcję "lm()" w R (parametr "weights" – szczegóły opisano w "pomocy").

Brak normalności błędów losowych

- Często pomaga odpowiednio dobrane przekształcenie,
- Zastosowanie innych metod np. Uogólnione modele liniowe:
  - jeżeli $Y_i \in \{0, 1\}$ – regresja logistyczna,
  - jeżeli $Y_i \in \{0, 1, 2, ...\}$ – regresja Poissona.

Obserwacje odstające

Obesrwacja "odstająca", czy "wpływowa"

Usunąć z danych (?)

vs

zmniejszyć wpływ na estymację
(np. za pomocą parametru "weights")