Modele liniowe

Michał Kos

Uniwersytet Wrocławski

Plan wykładu

1. Ogólny test F
2. Zagadnienie standaryzacji danych
3. Problem multikolinearności

Table of Contents

Ogólny test F

Klasyczny test F umożliwia badanie następującego problemu:

$$H_0 : \quad \beta_1 = ... = \beta_{p-1} = 0 \qquad vs \qquad H_1 : \ (\exists i)\ \beta_i \neq 0$$

Odpowiada on zatem na fundamentalne pytanie o to, czy którakolwiek ze zmiennych objaśniających ma istotny wpływ na zmienną wynikową.

Brak podstaw do odrzucenenia H_0 oznacza, że **żadna** zmienna objaśniająca nie wpływa na zmienną odpowiedzi w istotny sposób.
Odrzucenie H_0 oznacza, że **przynajmniej jedna** zmienna objaśniająca wpływa na zmienną odpowiedzi (nie mówi jednak które!)

Ogólny test F

Klasyczny test F umożliwia badanie następującego problemu:

$$H_0: \quad \beta_1 = ... = \beta_{p-1} = 0 \qquad vs \qquad H_1: \ (\exists i) \ \beta_i \neq 0$$

Odpowiada on zatem na fundamentalne pytanie o to, czy którakolwiek ze zmiennych objaśniających ma istotny wpływ na zmienną wynikową.

Brak podstaw do odrzucenenia H_0 oznacza, że **żadna** zmienna objaśniająca nie wpływa na zmienną odpowiedzi w istotny sposób. Odrzucenie H_0 oznacza, że **przynajmniej jedna** zmienna objaśniająca wpływa na zmienną odpowiedzi (nie mówi jednak które!)

Jeżeli odrzucimy H_0 to pojawia się kolejne fundamentalne pytanie:

Które zmienne objaśniające w istotny sposób wpływają na zmienną objaśnianą, a dla których ów wpływ jest pomijalny?

Powyższe zagadnienie określane jest **"problemem wyboru istotnych zmiennych do modelu"** lub (krócej) **"problemem wyboru modelu"**.

Ogólny test F

Klasyczny test F umożliwia badanie następującego problemu:

$$H_0: \quad \beta_1 = ... = \beta_{p-1} = 0 \qquad vs \qquad H_1: \ (\exists i)\ \beta_i \neq 0$$

Odpowiada on zatem na fundamentalne pytanie o to, czy którakolwiek ze zmiennych objaśniających ma istotny wpływ na zmienną wynikową.

Brak podstaw do odrzucenenia H_0 oznacza, że **żadna** zmienna objaśniająca nie wpływa na zmienną odpowiedzi w istotny sposób.
Odrzucenie H_0 oznacza, że **przynajmniej jedna** zmienna objaśniająca wpływa na zmienną odpowiedzi (nie mówi jednak które!)

Jeżeli odrzucimy H_0 to pojawia się kolejne fundamentalne pytanie:

Które zmienne objaśniające w istotny sposób wpływają na zmienną objaśnianą, a dla których ów wpływ jest pomijalny?

Powyższe zagadnienie określane jest **"problemem wyboru istotnych zmiennych do modelu"** lub (krócej) **"problemem wyboru modelu"**.

Odpowiedź można wyrazić w języku **nośnika wektora parametrów** β, czyli zbioru indeksów, dla których parametry β_i są różne od zera
$S = Supp(\beta) = \{i : \beta_i \neq 0\}$ oraz jego dopełnienia: $S^c = \{i : \beta_i = 0\}$.

Ogólny test F

Okazuje się, że test F można uogólnić w taki sposób, by badał równocześnie istotność kilku ale nie wszystkich zmiennych objaśniających, np. dla zm. objaśniających o indeksach w zbiorze $I \subset \{1, 2, ..., p-1\}$:

$$H_0 : \ (\forall i \in I)\ \beta_i = 0 \qquad vs \qquad H_1 : \ (\exists i \in I)\ \beta_i \neq 0$$

lub (równoważnie) porównywał modele:

H_0 : dane pochodzą z modelu $Y_i = \beta_0 + \sum_{j \in I^c} \beta_j X_{ij} + \epsilon_i$ model zredukowany

H_1 : dane pochodzą z modelu $Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \epsilon_i$ model pełny

Łatwo zauważyć, że powyższe modele różnią się wyborem zmiennych objaśnianych włączanych do modelu. W modelu pełnym występuja wszystkie zmienne, a model zredukowany zawiera wyłącznie zmienne o indeksach z dopełnienia zbioru I.

Ogólny test F

Statystyka testowa dla tak postawionego problemu ma postać:

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

gdzie w nawiasach podano, dla którego modelu wyznaczana jest odpowiednia statystyka (SSE, MSE,dfE). Litera F oznacza model pełny (full model), a R model zredukowany (reduced model).

Ogólny test F

Statystyka testowa dla tak postawionego problemu ma postać:

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

gdzie w nawiasach podano, dla którego modelu wyznaczana jest odpowiednia statystyka (SSE, MSE,dfE). Litera F oznacza model pełny (full model), a R model zredukowany (reduced model).

Statystyka F przy założeniu prawdziwości H_0 pochodzi z rozkładu Fishera–Snedecora o $dfE(R) - dfE(F)$ i $dfE(F)$ stopniach swobody.

Łatwo pokazać, że $dfE(R) - dfE(F)$ jest równe liczbie zerowanych parametrów β_i przy H_0, czyli $dfE(R) - dfE(F) = \#I$

Ogólny test F

Statystyka testowa dla tak postawionego problemu ma postać:

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

gdzie w nawiasach podano, dla którego modelu wyznaczana jest odpowiednia statystyka (SSE, MSE,dfE). Litera F oznacza model pełny (full model), a R model zredukowany (reduced model).

Statystyka F przy założeniu prawdziwości H_0 pochodzi z rozkładu Fishera–Snedecora o $dfE(R) - dfE(F)$ i $dfE(F)$ stopniach swobody.

Łatwo pokazać, że $dfE(R) - dfE(F)$ jest równe liczbie zerowanych parametrów β_i przy H_0, czyli $dfE(R) - dfE(F) = \#I$

Test odrzuca hipotezę zerową na poziomie istotnosci α dla dużych wartości statystyki F, tzn. dla $F > F^*$ gdzie $F^* = F^*(1-\alpha, dfE(R) - dfE(F), dfE(F))$ jest kwantylem rzędu $1-\alpha$ z rozkładu Fishera-Snedecora o $dfE(R) - dfE(F)$ i $dfE(F)$ stopniach swobody.
Zwykle, wnioskowanie dokonywane jest na podstawie p-wartości:
$p = P(z > F) = 1 - \tilde{F}_z(F)$, gdzie $z \sim F(dfE(R) - dfE(F), dfE)$ a $\tilde{F}_z$ jest dystrybuantą zmiennej losowej z.

Ogólny test F – przykłady

Przykład 1

Przyjmijmy, że $n = 100$, oraz że mamy 5 zmiennych objaśniających: $X_1, ..., X_5$.
Podejrzewamy, że zmienne X_4, X_5 nie wpływają w istotny sposób na zm. objaśnianą.
Za pomocą ogólnego testu F możemy zbadać powyższe podejrzenia:

$$H_0 : \quad \beta_4 = \beta_5 = 0 \qquad vs \qquad H_1 : \; \beta_4 \neq 0 \text{ i/lub } \beta_5 \neq 0$$

Kluczowym zagadnieniem jest wyznaczenie liczby stopni swobody statystyki F:

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

Na ich podstawie możemy wyznaczyć wartość krytyczną F^* dla testu na poziomie istotności α.

Ogólny test F – przykłady

Przykład 1

Przyjmijmy, że $n = 100$, oraz że mamy 5 zmiennych objaśniających: $X_1, ..., X_5$. Podejrzewamy, że zmienne X_4, X_5 nie wpływają w istotny sposób na zm. objaśnianą.
Za pomocą ogólnego testu F możemy zbadać powyższe podejrzenia:

$$H_0: \ \ \beta_4 = \beta_5 = 0 \qquad vs \qquad H_1: \ \ \beta_4 \neq 0 \text{ i/lub } \beta_5 \neq 0$$

Kluczowym zagadnieniem jest wyznaczenie liczby stopni swobody statystyki F:

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

Na ich podstawie możemy wyznaczyć wartość krytyczną F^* dla testu na poziomie istotności α.

Liczba stopni swobody mianownika statystyki F wynosi $dfE(F) = n - 6$, a licznika $dfE(R) - dfE(F) = n - 4 - (n - 6) = 2 = \#I$. Zatem statystyka F przy H_0 pochodzi z rozkładu Fishera–Snedecora z 2 i $n - 6$ stopniami swobody.

Ogólny test F – przykłady

Przykład 1 c.d. (dodatkowe sumy kwadratów)

Często w przypadku modeli linowych stosuje się notację opisującą w jednoznaczny sposób, które zmienne są włączone do modelu np.: $SSE(X_{i_1}, ..., X_{i_m})$ opisuje statystykę SSE dla modelu w którym występują zmienne objaśniane $X_{i_1}, ..., X_{i_m}$.

Ogólny test F – przykłady

Przykład 1 c.d. (dodatkowe sumy kwadratów)

Często w przypadku modeli linowych stosuje się notację opisującą w jednoznaczny sposób, które zmienne są włączone do modelu np.: $SSE(X_{i_1}, ..., X_{i_m})$ opisuje statystykę SSE dla modelu w którym występują zmienne objaśniane $X_{i_1}, ..., X_{i_m}$.

W rozpatrywanym przykładzie statystyki $SSE(R)$ i $SSE(F)$ możemy wyrazić we wprowadzonej notacji jako:

$$SSE(F) = SSE(X_1, X_2, X_3, X_4, X_5); \quad SSE(R) = SSE(X_1, X_2, X_3);$$

Dodatkowo różnicę między statystykami oznacza się w następujacy sposób:

$$SSE(X_4, X_5|X_1, X_2, X_3) = SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4, X_5)$$

Różnica $SSE(X_4, X_5|X_1, X_2, X_3)$ występuje w liczniku statystyki F, stąd:

$$F = \frac{(SSE(R) - SSE(F))/(dfE(R) - dfE(F))}{MSE(F)} = \frac{SSE(X_4, X_5|X_1, X_2, X_3)/2}{SSE(X_1, X_2, X_3, X_4, X_5)/(n-6)}$$

Przykład 1 c.d. (dodatkowe sumy kwadratów)

W analogiczny sposób oznacza się statystyki $SSM(X_{i_1}, ..., X_{i_m})$. Jednakże, różnica zdefiniowana jest wzorem

$$SSM(X_4, X_5|X_1, X_2, X_3) = SSM(X_1, X_2, X_3, X_4, X_5) - SSM(X_1, X_2, X_3)$$

Obie różnice są zdefiniowane w taki sposób by przyjmowały wartości nieujemne. Można ponadto pokazać, że są sobie równe:

$$SSE(X_4, X_5|X_1, X_2, X_3) = SSE(R) - SSE(F) =$$

$$SST - SSM(R) - (SST - SSM(F)) = SSM(F) - SSM(R) = SSM(X_4, X_5|X_1, X_2, X_3)$$

Drugą ważną własnością różnicy jest następująca relacja:

$$SST = SSM(X_1, X_2, X_3) + SSM(X_4, X_5|X_1, X_2, X_3) + SSE(X_1, X_2, X_3, X_4, X_5)$$

Wynika z niej, że różnica $SSM(X_4, X_5|X_1, X_2, X_3)$ opisuje zmienność w wektorze Y objaśnianą przez dodatkowe zmienne X_4 i X_5 ponad to co objaśniają zmienne (X_1, X_2, X_3).

Ogólny test F – przykłady

Przykład 2 (standardowy test F)
Badamy następujący problem:

$$H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \qquad vs \qquad H_1: \ (\exists i)\ \beta_i \neq 0$$

Zgodnie z ogólnym testem F statystyka testowa ma postać:

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

Pokażemy, że powyższa formula sprowadza się do $F = MSM/MSE$ (postaci statystyki dla stand. test F):
1. Model pełny zawiera wszystkie zmienne objaśniające, zatem:

$$SSE(F) = SSE; \quad dfE(F) = dfE = n - p; \quad MSE(F) = MSE$$

2. Model zredukowany zawiera wyłacznie Intercept, zatem:

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

zatem:

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

podstawiając:

$$F = MSM/MSE$$

Ogólny test F – przykłady

Przykład 3 ("alternatywa" dla testu T_i dla β_i)
Za pomocą ogólnego testu F możemy badać następujący problem:

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

Jest to "alternatywny" test do testu opartego na statystyce T_i badajacego istotność parametru β_i. Zgodnie z ogólnym testem F statystyka testowa dla powższego problemu ma postać:

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

$$= \frac{SSM(X_i | X_1, ..., X_{i-1}, X_{i+1}, ..., X_{p-1})}{MSE(F)}$$

ostatnia równość wynika z faktu, iż liczba zerowanych elementów wektora β jest równa $dfE(R) - dfE(F) = \#I = 1$.

Ogólny test F – przykłady

Przykład 3 ("alternatywa" dla testu T_i dla β_i)
Za pomocą ogólnego testu F możemy badać następujący problem:

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

Jest to "alternatywny" test do testu opartego na statystyce T_i badajacego istotność parametru β_i. Zgodnie z ogólnym testem F statystyka testowa dla powyższego problemu ma postać:

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

$$= \frac{SSM(X_i | X_1, ..., X_{i-1}, X_{i+1}, ..., X_{p-1})}{MSE(F)}$$

ostatnia równość wynika z faktu, iż liczba zerowanych elementów wektora β jest równa $dfE(R) - dfE(F) = \#I = 1$.

Można pokazać, że statystka $F = T_i^2$. W konsekwencji oba testy są równoważne!

Dwa typy dodatkowych sum w R

Suma z poprzedniego przykładu postaci $SSM(X_i|X_1, ..., X_{i-1}, X_{i+1}, ..., X_{p-1})$, w R nazywana jest **sumą II typu** (type II SS). Opisuje ona to, ile zmienności w Y, objaśnia zmienna X_i po uwzględnieniu wpływu pozostałych zmiennych $X_1, ..., X_{i-1}, X_{i+1}, ..., X_{p-1}$.

Z kolei tzw. **sumy I typu** (domyślne w R) dodają po jednej zmiennej w każdym kroku:

1. Wpływ pierwszej zmiennej $= SSM(X_1)$
2. Wpływ drugiej zmiennej (po uwzględnieniu pierwszej) $= SSM(X_2|X_1)$
3. Wpływ trzeciej zmiennej (po uwzględnieniu pierwszej i drugiej) $=$ $SSM(X_3|X_1, X_2)$
4. $SSM(X_4|X_1, X_2, X_3)$ ... itd.

Łatwo pokazać że sumy I typu stanowią rozkład statystyki $SSM(X_1, ..., X_{p-1})$:

$$SSM(X_1, ..., X_{p-1}) = SSM(X_1) + SSM(X_2|X_1) + ... + SSM(X_{p-1}|X_1, ..., X_{p-2})$$

Dwa typy dodatkowych sum w R

Ogólne testy F dla ciągu sum typu pierwszego porównują modele:

$$H_0: \text{ dane pochodzą z modelu } Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \text{ model zredukowany}$$

$$H_1: \text{ dane pochodzą z modelu } Y_i = \beta_0 + \sum_{j=1}^{k+1} \beta_j X_{ij} + \epsilon_i \text{ model pełny}$$

Statystki mają postać:

$$F_k = \frac{SSM(X_k|X_1, ..., X_{k-1})}{MSE(F)} \quad k = 0, ..., p-1$$

Każda ze statystyk F_k pochodzi z rozkładu Fishera-Snedecora z 1 i $n-p$ stopniami swobody.

Dwa typy dodatkowych sum w R

Uwagi:

- Oba typy dodatkowych sum mocno się różnią
- zmiana porządku zmiennych zależnych w modelu wpływa znacząco na sumy typu I, ale nie zmienia sum typu II.

Table of Contents

Standaryzacja danych

W regresji liniowej często pojawia się sytuacja, w której regresory są wyrażone w różnych jednostkach. Taka sytuacja powoduje niejednokrotnie brak możliwości wzajemnego porównania wpływu regresorów.

Przykład

Załóżmy, że chcemy przewidzić obwód w pasie pacjenta (Y) na podstawie wzrostu wyrażonego w centymetrach (X_1) i wagi wyrażonej w kilogramach (X_2). Załóżmy, że uzyskany model jest postaci:

$$\hat{Y} = 1 + 2X_1 + 3X_2$$

Na podstawie modelu można by wysnuć wniosek, że większy wpływ na zmienną Y ma waga ($2 < 3$). Jednakże gdybyśmy wyrazili wzrost w metrach $\tilde{X}_1 = X_1/100$, wówczas ten sam model by miał postać:

$$\hat{Y} = 1 + 200(X_1/100) + 3X_2 = 1 + 200\tilde{X}_1 + 3X_2$$

i wniosek byłby odwrotny ($200 > 2$). Widzimy zatem, że w takiej sytuacji, aby móc porównywać wpływ regresorów potrzebna jest standaryzacja danych.

Standaryzacja danych

W przypadku regresji liniowej standaryzacji dokonuje się wprowadzając następujące przekształcenie zmiennych niezależnych:

$$\tilde{X}_i = s(Y)X_i/s(X_i) \quad i = 1, ..., p-1$$

W wyniku powyższej standaryzacji wariancja każdej z nowych zmiennych $\tilde{X}_i$ jest taka sama. W konsekwencji wzrost o dowolną ustaloną wartość K dla każdej zmiennej $\tilde{X}_i$ jest tak samo prawdopodobny.
Jak wygląda model wyrażony w języku nowych zmiennych:

$$\hat{Y} = ... + \hat{\beta}_i X_i + ... = ... + \left(\hat{\beta}_i s(X_i)/s(Y)\right)(s(Y)X_i/s(X_i)) + ... =$$

$$= ... + \hat{\tilde{\beta}}_i \tilde{X}_i + ...$$

gdzie $\hat{\tilde{\beta}}_i = \hat{\beta}_i s(X_i)/s(Y)$.
Jak interpretować wpływ nowych zmiennych:
Wzrost o wartość 1 dowolnej zmiennej $\tilde{X}_i$ powoduje wzrost $\hat{Y}$ o $\hat{\tilde{\beta}}_i$. Ponieważ wzrost te są tak samo prawdopodobne, zatem $\hat{\tilde{\beta}}_i$ są porównywalne i mówią o wzajemnej sile wpływu poszczególnych zmiennych $\tilde{X}_i$ (X_i).

Table of Contents

Multikolinearność

W kontekście modeli liniowych o zjawisku multikolinearności mówimy, gdy macierz $\mathbb{X}'\mathbb{X}$ jest bliska macierzy singularnej. Miarą tej bliskości może być np. najmniejsza wartość własna macierzy $\mathbb{X}'\mathbb{X}$ (λ_{min}) (dla macierzy singularnej $\lambda_{min} = 0$). Pojawiają się wówczas 2 problemy:

- Obliczeniowy: pomimo tego, że macierz $\mathbb{X}'\mathbb{X}$ formalnie jest odwracalna to ze względu na skończoną dokładność obliczeniową komputerów, wyznaczenie macierzy $(\mathbb{X}'\mathbb{X})^{-1}$ stanowi duże wyzwanie, co wpływa na dokładność np. $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$
- Statystyczny: ponieważ wartości własne z macierzy A^{-1} są odwrotnościami wartości własnych macierzy A, zatem jeżeli $\lambda_{min}(\mathbb{X}'\mathbb{X})$ jest bliska 0 to $\lambda_{max}((\mathbb{X}'\mathbb{X})^{-1}) = \lambda_{min}^{-1}(\mathbb{X}'\mathbb{X})$ jest bardzo duża. Przekład się to na duże wartości wariancji estymatorów $\hat{\beta}_i$ ($\hat{\beta} \sim N(\beta, \sigma^2(\mathbb{X}'\mathbb{X})^{-1})$) (ślad macierzy $tr((\mathbb{X}'\mathbb{X})^{-1}) = \sum_{i=1}^{p} \lambda_i((\mathbb{X}'\mathbb{X})^{-1})$).

Okazuje się zatem, że zjawisko multikolinearności wpływa negatywnie zarówno na własności statystyczne estymatora $\hat{\beta}$ jak i dokładność jego wyznaczenia.

Multikolinearność

Problem multikolinearności jest ściśle związany z własnościami macierzy eksperymentu $\mathbb{X}$. Można pokazać, że zjawisko to pojawia się, gdy kolumny macierzy planu są prawie liniowo zależne.

Multikolinearność

Problem multikolinearności jest ściśle związany z własnościami macierzy eksperymentu $\mathbb{X}$. Można pokazać, że zjawisko to pojawia się, gdy kolumny macierzy planu są prawie liniowo zależne.
Jak wiemy z algebry liniowej zbiór wektorów $\{v_1, v_2, ..., v_{p-1}\}$ (kolumn macierzy $\mathbb{X}$) jest liniowo zależny gdy:

$$\exists i \ : \ v_i = \sum_{j \neq i} \gamma_j v_j$$

gdzie $\gamma_j \in \mathbb{R}$ to waga wektora v_j. Innymi słowy wektor v_i jest kombinacją liniową pozostałych wektorów.

Multikolinearność

Problem multikolinearności jest ściśle związany z własnościami macierzy eksperymentu $\mathbb{X}$. Można pokazać, że zjawisko to pojawia się, gdy kolumny macierzy planu są prawie liniowo zależne.
Jak wiemy z algebry liniowej zbiór wektorów $\{v_1, v_2, ..., v_{p-1}\}$ (kolumn macierzy $\mathbb{X}$) jest liniowo zależny gdy:

$$\exists i \ : \ v_i = \sum_{j \neq i} \gamma_j v_j$$

gdzie $\gamma_j \in \mathbb{R}$ to waga wektora v_j. Innymi słowy wektor v_i jest kombinacją liniową pozostałych wektorów.
Powyższa zależność implikuje fakt, iż:

$$cor(v_i, \sum_{j \neq i} \gamma_j v_j) = 1$$

W powyższym języku wyrażana jest prawie liniowa zależność:
Problem multikolinearności występuje, gdy:

$$\exists i \ cor(v_i, \sum_{j \neq i} \gamma_j v_j) \approx 1$$

Multikolinearność

Przykład
Wpływ korelacji pomiędzy dwoma kolumnami w macierzy $\mathbb{X}'\mathbb{X}$ na wielkość maksymalnej wartości własnej macierzy $(\mathbb{X}'\mathbb{X})^{-1}$: