Modele liniowe

Michał Kos

Uniwersytet Wrocławski

Plan wykładu

Table of Contents

Diagnostyka modelu

Diagnostyka modelu w regresji liniowej wielorakiej jest analogiczna do diagnostyki dla regresji liniowej prostej.
W pierwszym kroku zwykle dokonujemy badania własności każdej ze zmiennych niezależnych:

- analiza podstawowych charakterystyk (średnia, rozrzut itp.) oraz wykresów (histogram, boxplot, wykres kwantylowo-kwantylowy, zależność od czasu (porzadku w danych))
- odpowiedzi na pytania o rozkład zm. objaśniającej, obs. odstające, zależność w czasie,
- badanie relacji pomiedzy zmiennymi niezależnymi: korelacje (cor()), wykresy rozrzutu (pairs()),

Diagnostyka modelu

W drugim kroku dokonujemy badania własności zmiennej zależnej, uwzględniające wpływ zmiennych niezależnych.

- analiza podstawowych charakterystyk wektora residuów e (średnia, rozrzut itp.) oraz wykresów (histogram, wykres kwantylowo-kwantylowy)
- wykresy wektora residuów e vs każda zmienna niezależna lub czas
- odpowiedź na pytania:
  - Czy wektor residuów jest w przybliżeniu normalny?
  - Czy wektor residuów niezależy od zmiennych niezależnych?
  - Czy wariancja jest stała? (np. przez wyrysowanie kwadratów residuów vs cokolwiek od czego może wariancja zależeć np. predykcje, zm. niezależne, czas)

Remedia

Jeżeli analiza wskazuje na łamanie założeń modelu liniowego możemy zastosować środki zaradcze np.:

- transformacja danych np. transformacja Boxa–Coxa,
- analiza danych z wyłączeniem obserwacji odstających.

Table of Contents

Analiza wariancji

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

- 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 $E(Y_i) = \mu_i = \beta_0 + X_{i1}\beta_1 + ... + X_{ip-1}\beta_{p-1}$.

Analiza wariancji

Do analizy zmienności w wektorze odpowiedzi Y wykorzystywaliśmy w regresji liniowej prostej, poniższe statystyki (dla wartości $p = 2$):

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

Całość rozumowania w przypadku regresji liniowej wielorakiej jest taka sama z tą różnicą, że w tym przypadku liczba stopni swobody statystyk SSM i SSE zależy od liczby kolumn p.

Liczba stopni swobody

Uzasadnimy obecnie skąd bierze się kolumna z liczbą stopni swobody. Jest ona równa śladowi macierzy, na którą wykonywany jest rzut ortogonalny wektora Y związany z odpowiednimi sumami kwadratów SSM, SSE oraz SST.

Liczba stopni swobody

Uzasadnimy obecnie skąd bierze się kolumna z liczbą stopni swobody. Jest ona równa śladowi macierzy, na którą wykonywany jest rzut ortogonalny wektora Y związany z odpowiednimi sumami kwadratów SSM, SSE oraz SST.

Zacznijmy od statystyki SSE, gdyż dla niej do pewnego stopnia analiza już została wykonana. Zauważmy, że:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \|Y - \hat{Y}\|_2^2 = \|(\mathbb{I} - H)Y\|_2^2 = \|e\|_2^2$$

Z wcześniejszej analizy wiemy dodatkowo, że ślad macierzy rzutu ortogonalnego $Tr(\mathbb{I} - H) = n - p$.

Liczba stopni swobody

W analogiczny sposób możemy potraktować statystykę SST. Wystarczy zauważyć, że jest ona tożsama ze statystyką SSE dla szczególnej macierzy rzutu $\tilde{H}$:

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \|(\mathbb{I} - \tilde{H})Y\|_2^2 = \|\tilde{e}\|_2^2$$

gdzie $\tilde{H}$ jest rzutem ortogonalnym na przestrzeń rozpiętą na kolumnie jedynek związanej z interceptem $Lin(\mathbb{1} = (1, .., 1)')$ (brak zmiennych objaśniających).

$$\tilde{H} = \mathbb{1}(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1}' = \frac{1}{n}\mathbb{1}\mathbb{1}' = \begin{pmatrix} 1/n & \cdots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \cdots & 1/n \end{pmatrix}; \quad \tilde{H}Y = \begin{pmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix}$$

Łatwo pokazać, że $Tr(\tilde{H}) = 1$, dlatego $dfT = Tr(\mathbb{I} - \tilde{H}) = n - 1$.

Liczba stopni swobody

W przypadku statystyki SSM wykorzystamy wyniki dla poprzednich punktów

$$SSM = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \|(H - \tilde{H})Y\|_2^2$$

Stąd $dfM = Tr(H - \tilde{H}) = Tr(H) - Tr(\tilde{H}) = p - 1$.
Wypada wykazać że $H - \tilde{H}$ jest macierzą rzutu ortogonalnego. Symetryczność jest oczywista, także pozostaje wykazać idempotentność. Jednakże w pierwszej kolejności należy zauważyć że:

$$H\tilde{H} = \tilde{H} \quad \text{oraz} \quad \tilde{H}H = \tilde{H}$$

Powyższe równości wynikają bezpośrednio z faktu, iż przestrzeń $Lin(\mathbb{1})$ (na którą rzutuje $\tilde{H}$) jest podprzestrzenią $Lin(\mathbb{X})$ (na którą rzutuje H). Wykorzystując powyższe fakty uzyskujemy idempotentność macierzy $H - \tilde{H}$:

$$(H - \tilde{H})^2 = H^2 - H\tilde{H} - \tilde{H}H + \tilde{H}^2 = H - \tilde{H}$$

Test F

Na podstawie statystyk *MSM* i *MSE* skonstruowany jest tzw. test F testujący hipotezę o istotności wpływu wszystkich regresorów (równocześnie):

$$H_0: \quad \beta_1 = ... = \beta_{p-1} = 0 \qquad vs \qquad H_1: \quad \beta_i \neq 0 \text{ przynajmniej dla jednego i}$$

Statystyka $F = MSM/MSE$ przy H_0 ma rozkład Fishera–Snedecora z *dfM* i *dfE* stopniami swobody. Odrzucamy hipotezę zerową, gdy $F > F_c$, gdzie $F_c = F^*(1-\alpha, dfM, dfE)$ jest kwantylem rzędu $1-\alpha$ z rozkładu Fishera–Snedecora z *dfM* i *dfE* st. sw.

Test F

Na podstawie statystyk *MSM* i *MSE* skonstruowany jest tzw. test F testujący hipotezę o istotności wpływu wszystkich regresorów (równocześnie):

$$H_0: \quad \beta_1 = ... = \beta_{p-1} = 0 \qquad vs \qquad H_1: \quad \beta_i \neq 0 \text{ przynajmniej dla jednego i}$$

Statystyka $F = MSM/MSE$ przy H_0 ma rozkład Fishera–Snedecora z *dfM* i *dfE* stopniami swobody. Odrzucamy hipotezę zerową, gdy $F > F_c$, gdzie $F_c = F^*(1-\alpha, dfM, dfE)$ jest kwantylem rzędu $1-\alpha$ z rozkładu Fishera–Snedecora z *dfM* i *dfE* st. sw.

Zwykle, wnioskowanie dokonywane jest na podstawie p-wartości: $p = P(z > F) = 1 - \tilde{F}_z(F)$, gdzie $z \sim F(dfM, dfE)$ a $\tilde{F}_z$ jest dystrybuantą zm. z.

Test F

Na podstawie statystyk *MSM* i *MSE* skonstruowany jest tzw. test F testujący hipotezę o istotności wpływu wszystkich regresorów (równocześnie):

$$H_0: \quad \beta_1 = ... = \beta_{p-1} = 0 \qquad vs \qquad H_1: \quad \beta_i \neq 0 \text{ przynajmniej dla jednego } i$$

Statystyka $F = MSM/MSE$ przy H_0 ma rozkład Fishera–Snedecora z *dfM* i *dfE* stopniami swobody. Odrzucamy hipotezę zerową, gdy $F > F_c$, gdzie $F_c = F^*(1-\alpha, dfM, dfE)$ jest kwantylem rzędu $1-\alpha$ z rozkładu Fishera–Snedecora z *dfM* i *dfE* st. sw.

Zwykle, wnioskowanie dokonywane jest na podstawie p-wartości: $p = P(z > F) = 1 - \tilde{F}_z(F)$, gdzie $z \sim F(dfM, dfE)$ a $\tilde{F}_z$ jest dystrybuantą zm. z.

Na hipotezy w tescie F można również spojrzeć jak na porównanie dwóch modeli:

$$H_0: \text{ dane pochodzą z modelu } Y_i = \beta_0 + \epsilon_i$$

$$H_1: \text{ dane pochodzą z modelu } Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{ip-1} + \epsilon_i$$

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	$F = \text{MSM}/\text{MSE}$	$p = P(z > F)$
Error	dfE	SSE	MSE		

Uwagi o teście F:

- gdy zachodzi hipoteza alternatywna statystyka F pochodzi z niecentralnego rozkładu Fishera–Snedecora,
- umożliwia to wyznaczenie funkcji mocy testu,
- gdy $p = 2$ (mamy jedeną zm. objaśniającą) test F i test oparty na statystyce T dla parametru β_1 są równoważne.

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).

- Przy użyciu statystyki R^2 możemy porównywać modele o tej samej liczbie regresorów. Wynika to z faktu, iż statystyka SSM (licznik w R^2) z definicji jest funkcją rosnącą ze względu na dodawanie nowych kolumn do macierzy planu. Oznacza to, że jeżeli do macierzy $\mathbb{X}$ będziemy dodawać kolejne kolumny (bez względu na ich związek z wektorem odpowiedz) to R^2 będzie rosnąć.
- czasami gdy porównujemy modele o różnej liczbie regresorów, zamiast R^2 używa się tzw. modyfikowanego współczynnika determinacji:

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

istnieją jednak lepsze metody.

Table of Contents

Ogólny test F

Przedstawiony 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

Przedstawiony 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

Przedstawiony 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:

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

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:

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

Ogólny test F – przykłady

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:

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

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 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 $dfE(R) - dfE(F) = \#I = 1$.

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.