Zaawansowane modele liniowe - raport 1

Łukasz Rębisz

2023-10-10

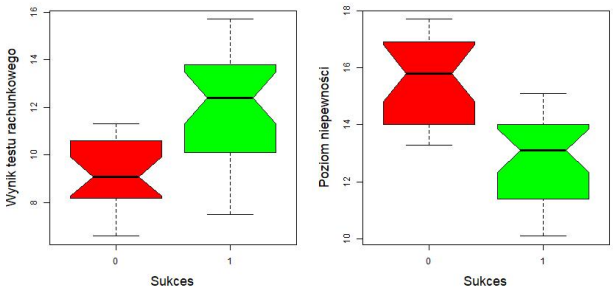
Cel raportu

Celem raportu jest **porównanie** modeli liniowych opartych na **regresji logistycznej** przy zastosowaniu różnych funkcji linkujących. Porównania dokonamy na podstawie wyników otrzymanych dzięki analizie zbioru danych opisującego relacje pomiędzy prawdopodobieństwami przyjęcia na studia a wynikami z testów rachunkowych i poziomem niepewności.

Następnie zbadamy model logistyczny oparty o optymalną w tej sytuacji funkcję linkującą. Przetestujemy dokładność estymatorów, zgodność z wartościami i rozkładami teoretycznymi oraz złożoność obliczeniową ich wyznaczania w zależności od liczby obserwacji, korelacji pomiędzy zmiennymi oraz liczby regresorów. Analizy wyników dokonamy na podstawie symulacji.

Analiza danych

Boxploty dla zmiennych objaśniających



Analiza powyższych wykresów (zwłaszcza porównanie median) pozwala wysunąć następujące wnioski:

- osoby, które nie osiągnęły sukcesu (nie zostały przyjęte na studia) miały średnio niższe wyniki z testów rachunkowych w porównaniu do osób przyjętych na studia,
- z kolei średni poziom niepewności u osób nieprzyjętych na studia był wyższy w porównaniu do osób przyjętych.

Model regresji logistycznej

Skonstruujmy model regresji logistycznej dla powyższych danych, stosując różne funkcje linkujące (*logit, probit, cauchit, cloglog*). Otrzymujemy następujące estymatory parametrów i wyniki testów istotności:

Table 1: Funkcja logit

	Estymator	Wynik z-testu	p-wartość
Wyraz wolny	14.239	2.094	0.036
Wynik testu rachunkowego	0.577	2.327	0.020
Poziom lęku	-1.384	-2.881	0.004

Table 2: Funkcja probit

	Estymator	Wynik z-testu	p-wartość
Wyraz wolny	8.257	2.247	0.025
Wynik testu rachunkowego	0.337	2.464	0.014
Poziom lęku	-0.804	-3.191	0.001

Table 3: Funkcja cauchit

	Estymator	Wynik z-testu	p-wartość
Wyraz wolny	18.383	1.495	0.135
Wynik testu rachunkowego	0.732	1.545	0.122
Poziom lęku	-1.774	-1.789	0.074

Table 4: Funkcja cloglog

	Estymator	Wynik z-testu	p-wartość
Wyraz wolny	9.001	1.935	0.053
Wynik testu rachunkowego	0.402	2.644	0.008
Poziom lęku	-0.939	-2.827	0.005

Powyższe wyniki wskazują na to, że:

- otrzymane wartości estymatorów znacząco różnią się w zależności od wyboru funkcji linkującej (zwłaszcza wyestymowana wartość wyrazu wolnego).
- Wyniki z-testów i w konsekwencji p-wartości także wykazują różnice. Na poziomie istotności $\alpha = 0.05$ nie odrzucamy istotności żadnego z parametrów dla funkcji *logit* i *probit*. W przypadku funkcji *cloglog* odrzucamy jedynie wyraz wolny (jako nieistotny). Natomiast dla funkcji *cauchit* odrzucamy istotność wszystkich parametrów.

Przewidywane prawdodpodobieństwo sukcesu - własna implemetacja

$$logit(\mu) = \eta$$

$$\mu = \frac{\exp(\eta)}{1 + \exp(\eta)} = \frac{1}{1 + \exp(-\eta)},$$

gdzie $\eta_i(\beta) = \beta_0 + x_{i,1}\beta_1 + ... + x_{i,p-1}\beta_{p-1}$.

Przewidywane prawdodpodobieństwo sukcesu u studenta, którego poziom lęku *anxiety* = 13, natomiast wynik z testów rachunkowych *numeracy* = 10, wynosi 0.69296 (własna implementacja) oraz 0.69296 (komenda *predict*).

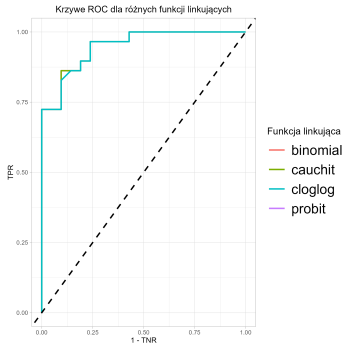
Wyznaczmy tę wartość dla badanych funkcji linkujących (korzystając z funkcji *predict*):

Table 5: Przewidywane prawdodpodobieństwo sukcesu, dla anxiety=13, numeracy=10

	p-stwo
logit	0.8828
probit	0.8806
cauchit	0.8849
cloglog	0.8963

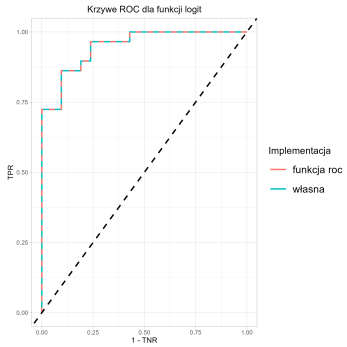
Otrzymaliśmy podobne wartości (z przedziału [0.88, 0.9]) dla wszystkich funkcji linkujących. Najmniejsze prawdopodobieństwo sukcesu zostało osiągnięte dla funkcji *probit*, największe zaś dla funkcji *cloglog*. Różnice są jednak na tyle małe, że możemy stwierdzić, że badane funkcje linkujące (zwłaszcza z pominięciem funkcji *cloglog*) zachowują się w tej sytuacji podobnie.

Krzywe ROC



Analiza powyższego wykresu wskazuje na to, że dla wybranych funkcji linkujących wykresy ROC praktycznie pokrywają się. Z tego powodu wszystkie badane modele dla różnych funkcji w podobny sposób odpowiadają danym. Żadna z funkcji linkujących nie jest wyraźnie lepsza niż pozostałe.

Porównajmy uzyskaną krzywą (dla funkcji *logit*) z krzywą wyznaczoną przy pomocy wbudowanej funkcji *roc*.



Porównanie uzyskanych krzywych pokazuje, że własna implementacja jest zgodna z wynikiem uzyskanym przy pomocy funkcji wbudowanej. Możemy zatem obliczyć pola pod krzywymi ROC wyznaczonymi wbudowaną funkcją *roc* w celu wybrania optymalnej funkcji linkującej. Uzyskujemy następujące wyniki:

Table 6: Pole pod krzywą ROC dla różnych funkcji linkujących

	logit	probit	cauchit	cloglog
Pole pod krzywą	0.949	0.949	0.949	0.947

Analiza powyższych wyników wskazuje na identyczną wartość pola pod wykresem krzywej ROC dla funkcji *logit*, *probit* oraz *cauchit*.

Dla funkcji *cloglog* wartość jest nieznacznie niższa, więc wyniki wskazują na wybór jednej z trzech poprzednich funkcji jako optymalnej. Zbadajmy zatem model oparty na funkcji linkującej *logit*.

Model z funkcją linkującą logit

Estymator macierzy kowariancji

Otrzymaliśmy następującą macierz kowariancji:

$$\begin{array}{c} \\ \text{wyraz wolny} \\ \text{test rach.} \\ \text{lęk} \end{array} \begin{array}{c} \begin{array}{ccc} \text{wyraz wolny} & \text{test rach.} & \text{lęk} \end{array} \\ \begin{pmatrix} 46.23 & -0.248 & -3.062 \\ -0.248 & 0.062 & -0.024 \\ -3.062 & -0.024 & 0.231 \end{pmatrix} \end{array}$$

Porównajmy pierwiastki wartości na przekątnej wyestymowanej macierzy kowariancji z estymatorami odchyleń standardowych zwracanymi przez R:

Table 7: Odchylenia standardowe estymatorów

	estymacja	f. wbudowana	różnica
Wyraz wolny	6.7992310	6.7985192	0.0007118
Wynik testu rachunkowego	0.2480977	0.2480840	0.0000137
Poziom lęku	0.4804650	0.4804027	0.0000623

Powyższe wyniki wskazują na bardzo dużą dokładność otrzymanych wartości odchylenia standardowego. Wskazuje to na dobrą estymację macierzy kowariancji.

Test oparty na statystyce Deviance

Przetestujmy hipotezę, że obie zmienne objaśniające nie mają wpływu na zmienną odpowiedzi (poziom istotności $\alpha = 0.05$).

Wartość statystyki testowej Deviance:

$\chi^2 = D(M_0) - D(M_1) = 39.74 > 2.92 = F^{-1}_{\chi^2}(1-\alpha,\ 2)$.

Powyższa statystyka Deviance ma asymptotyczny rozkład χ^2 z liczbą stopni swobody równą liczbie testowanych zmiennych (czyli 2). Porównanie z odpowiednim kwantylem rozkładu pozwala wysunąć wniosek: NIE odrzucamy istotności obu zmiennych.

Badanie epsilon

Parametr ϵ (domyślna wartość 10^{-8}) jest tolerancją zbieżności dla iteracji

$$\frac{|dev_{i+1} - dev_i|}{|dev_{i+1}| + 0.1} < \epsilon$$

wyznaczającej numerycznie wartości estymatorów.

Porównajmy liczbę iteracji i wartości estymatorów poszczególnych parametrów dla różnych wartości ϵ:

	Wyraz wolny	Test rach.	Lęk	Iteracje
1e-1	12.890	0.538	-1.264	3
1e-2	14.092	0.574	-1.371	4
1e-3	14.237	0.577	-1.384	5
1e-6	14.239	0.577	-1.384	6
1e-8	14.239	0.577	-1.384	6

Powyższe wyniki pokazują, że:

- liczba iteracji rośnie wraz ze wzrostem żądanej dokładności obliczeń,
- wartości estymowanych parametrów są równe z dokładnością do dwóch miejsc po przecinku dla $\epsilon \leq 10^{-3}$.

Symulacje

Wygenerujmy macierz X wymiaru $n = 400,\ p = 3$, której elementy są zmiennymi losowymi z rozkładu $N(0, \sigma^2 = 1/400)$. Załóżmy, że binarny wektor odpowiedzi jest wygenerowany zgodnie z modelem regresji logistycznej z wektorem $\beta = (3, 3, 3)'$. Wówczas macierz Informacji Fishera w puknicie β wynosi

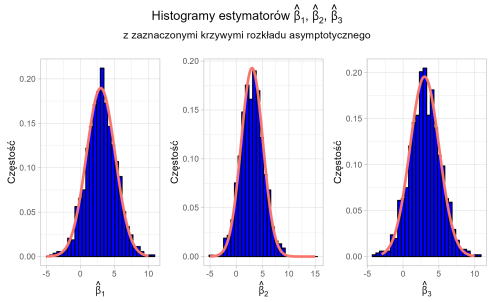
$$J = X'S(\beta)X,$$

gdzie $S(\beta)$ jest macierzą diagonalną z wariancjami parametrów β_k na przekątnej.

Natomiast asymptotyczną macierzą kowariancji estymatorów największej wiarogodności parametrów β jest macierz J^{-1}.

Wyznaczmy 1000-krotnie wartości estymatorów $\hat{\beta}_k$. Otrzymaliśmy następujące wyniki:

Histogramy



Powyższe histogramy wskazują na zgodność rozkładu otrzymanych wartości $\hat{\beta}_k$ z rozkładem asymptotycznym, czyli rozkładem $N(\beta_k, J^{-1}_{k,k})$ dla $k = 1, 2, 3$.

Obciążenie

Obciążenie $Bias(\hat{\beta}_1) = mean(\hat{\beta}_1) - \beta_1 = 0.003$, $Bias(\hat{\beta}_2)$ = -0.024, $Bias(\hat{\beta}_3) = 0.057$. Otrzymane wyniki wskazują na stosunkowo małe obciążenie wyznaczonych estymatorów.

Estymacja macierzy kowariancji

Porównanie asymptotycznej macierzy kowariancji (lewa macierz) z wyestymowaną macierzą kowariancji wektora estymatorów $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ (prawa macierz):

$$\begin{bmatrix} 4.8 & 0.1 & 0.5 \\ 0.1 & 4.1 & -0.1 \\ 0.5 & -0.1 & 4 \end{bmatrix} \begin{bmatrix} 4.6 & 0 & 0.2 \\ 0 & 3.9 & -0.2 \\ 0.2 & -0.2 & 4.2 \end{bmatrix}$$

Wyniki wskazują na zgodność rzędu otrzymanych wielkości. Większość wartości jest zgodna z dokładnością do 0.2. Niemniej jednak nie możemy stwierdzić, że powyższe macierze są w przybliżeniu równe. Być może liczba powtórzeń eksperymentu była zbyt mała.

Wpływ liczby obserwacji

Powtórzmy doświadczenie w przypadku gdy liczba obserwacji $n = 100$.

- Rozkład estymatorów $\hat{\beta}_k$ jest zgodny z odpowiednim rozkładem normalnym (rozkładem asymptotycznym).
- Obciążenia estymatorów wynoszą w tym przypadku $Bias(\hat{\beta}_1) = 0.132$, $Bias(\hat{\beta}_2) = 0.088$, $Bias(\hat{\beta}_3) = 0.018$. Otrzymane wyniki wskazują na większe o rząd wielkości obciążenie estymatorów w porównaniu do modelu opartego na $n = 400$ obserwacjach.

Porównanie asymptotycznej macierzy kowariancji (lewa macierz) z wyestymowaną macierzą kowariancji wektora estymatorów $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ (prawa macierz):

$$\begin{bmatrix} 15.8 & -1.9 & 1.9 \\ -1.9 & 16.5 & -1.6 \\ 1.9 & -1.6 & 21.5 \end{bmatrix} \begin{bmatrix} 17.3 & -2.4 & 3 \\ -2.4 & 17 & -1.9 \\ 3 & -1.9 & 24.1 \end{bmatrix}$$

Otrzymane wyniki obrazują znacząco większe wartości kowariancji pomiędzy zmiennymi w tej sytuacji. W szczególności wariancje estymatorów (wartości na przekątnych) są ok. 4 razy większe niż w przypadku $n = 400$ obserwacji.

Różnice pomiędzy wartościami teoretycznej i wyestymowanej macierzy kowariancji wektora estymatorów $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ wynoszą w tym przypadku:

$$\begin{bmatrix} 1.5 & 0.5 & 1 \\ 0.5 & 0.5 & 0.3 \\ 1 & 0.3 & 2.5 \end{bmatrix}$$

Zgodność pomiędzy macierzami jest więc w tej sytuacji dużo mniejsza niż w przypadku $n = 400$ obserwacji, gdzie wartość różniły się średnio o 0.2.

Wnioski: zmniejszenie liczby obserwacji obniżyło dokładność estymatorów $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, zwiększyło ich wariancje oraz kowariancje. Rozkłady estymatorów pozostały zgodne z rozkładami teoretycznymi.

Wpływ korelacji pomiędzy regresorami

Powtórzmy doświadczenie w przypadku, gdy wiersze macierzy X są niezależnymi wektorami losowymi z rozkładu $N(0, \Sigma)$, gdzie $\Sigma = \frac{1}{n}S$, $S_{ii} = 1, S_{ij} = 0.3$.

- Rozkład estymatorów $\hat{\beta}_k$ jest zgodny z odpowiednim rozkładem normalnym (rozkładem asymptotycznym).

- Obciążenia estymatorów wynoszą w tym przypadku $Bias(\hat{\beta}_1) = 0.183$, $Bias(\hat{\beta}_2)$ = -0.016, $Bias(\hat{\beta}_3) = 0.037$. Otrzymane wyniki wskazują na małe obciążenie estymatorów, porównywalne z obciążeniem uzyskanym w przypadku braku korelacji pomiędzy regresorami.

Porównanie asymptotycznej macierzy kowariancji (lewa macierz) z wyestymowaną macierzą kowariancji wektora estymatorów $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ (prawa macierz):

$$\begin{bmatrix} 4.5 & -0.8 & -0.9 \\ -0.8 & 5.2 & -1.2 \\ -0.9 & -1.2 & 4.7 \end{bmatrix} \begin{bmatrix} 4.8 & -0.9 & -0.9 \\ -0.9 & 5.2 & -1.3 \\ -0.9 & -1.3 & 4.8 \end{bmatrix}$$

Powyższe macierze wskazują na porównywalne wartości wariancji do przypadku, gdy regresory nie były skorelowane. Obserwujemy znaczący wzrost kowariancji pomiędzy regresorami.

Różnice pomiędzy wartościami teoretycznej i wyestymowanej macierzy kowariancji wektora estymatorów $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ wynoszą w tym przypadku:

$$\begin{bmatrix} 0.3 & 0.1 & 0 \\ 0.1 & 0.1 & 0.2 \\ 0 & 0.2 & 0.1 \end{bmatrix}$$

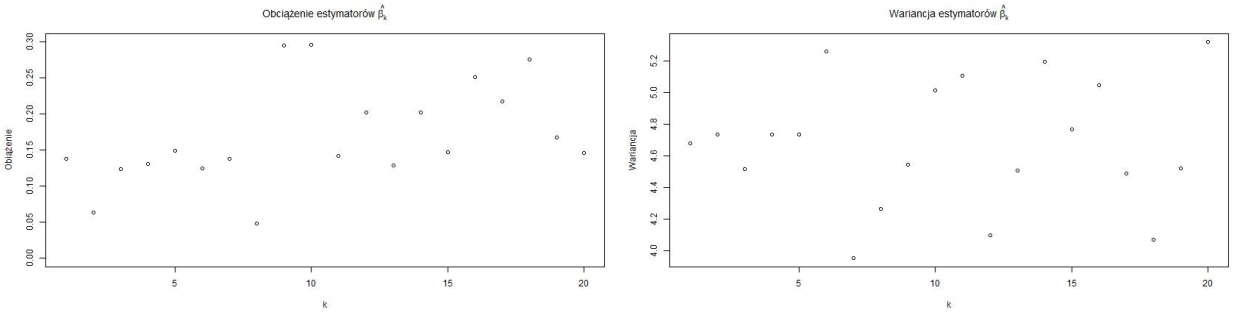
Zgodność pomiędzy macierzami jest więc w tej sytuacji podobna do dokładności uzyskanej w przypadku braku korelacji pomiędzy regresorami.

Wnioski: korelacja pomiędzy regresorami nie wpłynęła na zmianę rozkładów, obciążenia oraz wariancje estymatorów $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$. Zwiększyły się natomiast wartości kowariancji pomiędzy estymatorami.

Wpływ liczby regresorów

Powtórzmy doświadczenie w przypadku, gdy elementy X są niezależne, a $p = 20$.

- Rozkład estymatorów $\hat{\beta}_1$, $\hat{\beta}_2$, ..., $\hat{\beta}_{20}$ jest zgodny z rozkładem asymptotycznym, czyli rozkładem $N(\beta_k, J^{-1}_{k,k})$ dla $k = 1, 2, ..., 20$.



- Estymatory wykazują większe obciążenie niż w przypadku mniejszej liczby regresorów, gdzie obciążenie nie przekraczało wartości 0.1.
- Wariancje estymatorów $\hat{\beta}_k$ należą do przedziału [4, 5.2] - są porównywalne z wariancjami otrzymanymi przy mniejszej liczbie regresorów.
- Kowariancje pomiędzy estymatorami oscylują wokół zera - należą do przedziału [-0.4, 0.6].

Wnioski: zwiększenie liczby regresorów spowodowało zmniejszenie dokładności estymacji (znacząco wzrosło obciążenie estymatorów). Nie zaobserwowaliśmy jednkaże zmiany rozkładu estymatorów oraz ich wariancji i kowariancji.