Analiza dużych zbiorów danych raport nr 1

Dominik Mika

23 marca 2021

Cel raportu

W tym raporcie zajmiemy się problemem wielokrotnego testowania. Dokładnie testowaniem istotności parametrów w modelu liniowym w zależności od ilości rozpatrywanych zmiennych w modelu.

Zadanie 1

W tym zadaniu chcemy wylosować macierz planu $X_{1000\times 950}$, taką że jej elementami są niezależne zmienne losowe z rozkładu $N(0, \sigma = \frac{1}{\sqrt{1000}})$. Następnie wygenerujemy wektor odpowiedzi zgodnie z modelem

$$Y = X\beta + \epsilon,$$

gdzie $\beta = (3, 3, 3, 3, 3, 0, \ldots, 0)^T$, a $\epsilon \sim N(0, I)$.

Następnie wykonamy nasze testy w oparciu odpowiednio o model wykorzystujący 5, 10, 20, 100, 500 i 950 pierwszych zmiennych macierzy planu.

Dla każdego modelu wyznaczymy estymator najmniejszych kwadratów wektora β oraz testy istotności jego elementów. Zbadamy jak zmienia się odchylenie standardowe estymatora β_1 i szerokość 95% przedziału ufności dla tego parametru. Dodatkowo obliczymy liczbę prawdziwych i fałszywych odkryć dla różnych modeli, dla testowania bez korekty, z korektą Bonferroniego oraz korektą Benjaminiego Hochberga na wielokrotne testowanie.

Najpierw chcemy obliczyć odchylenie standardowe estymatora β_1 - $s(\hat{\beta}_1)$, wiemy że zadaje się ono wzorem:

$$s^2(\hat{\beta}_1) = s^2(\mathbb{X}'\mathbb{X})^{-1}_{(1,1)},$$

gdzie $s^2 = MSE$. Natomiast przedział ufności dla tego parametru wygląda następująco:

$$\hat{\beta}_1 \pm t_c s(\hat{\beta}_1),$$

gdzie $t_c = t^*(1 - \frac{\alpha}{2}, n - p)$. Natomiast ilość prawdziwych i fałszywych odkryć wyznaczać będziemy za pomocą p-wartości. Prawdziwe odkrycia to takie, gdy dla istotnych parametrów odrzuciliśmy H_0, czyli p-wartość $\leq \alpha$, gdzie $\alpha = 0.05$. Fałszywe odkrycie to takie, gdy odrzucamy H_0, dla nieistotnego parametru (p-wartość $\leq \alpha$). W przypadku korekty Bonferroniego parametr α zastępujemy parametrem $\frac{\alpha}{n}$, gdzie n to liczba wykonywanych testów. Korekta Benjaminiego Hochberga jest bardziej złożona. Na początku ustawiamy rosnąco p-wartości naszych testów. Następnie szukamy maksymalnego $i = i_0$, takiego, że zachodzi $p_{(i)} \leq \frac{i}{n}\alpha$. Wtedy odrzucamy wszystkie $H_{0(i)}$, dla wszystkich $i \leq i_0$.

Po wylosowaniu macierzy planu i zastosowaniu naszych metod wyniki prezentujemy w poniższej tabeli.

lz	$s(\hat{\beta}_1)$	IW	TD	FD	TD-B	FD-B	TD-BH	FD-BH
5	0.966	3.79	4	0	4	0	4	0
10	0.970	3.805	4	0	3	0	4	0
20	0.977	3.834	4	0	3	0	3	0
100	1.023	3.834	4	8	1	0	1	0
500	1.340	4.015	4	26	0	0	0	0
950	4.428	17.790	0	43	0	0	0	0

Legenda:

lz- liczba zmiennych

IW- szerokość przedziału

TD- liczba prawdziwych odkryć

FD- liczba fałszywych odkryć

-B- po korekcie Bonferrioniego

-BH- po korekcie Benjaminiego Hochberga

Widzimy, że wszystkie wartości zmieniają się w zależności od rozpatrywanego modelu. Odchylenie standardowe i szerokość przedziału rosną wraz z ilością zmiennych, wynika to z ich teoretycznych własności, które opiszę w następnym zadaniu. Natomiast widzimy, że gdy zwiększamy liczbę naszych testowanych hipotez to nasze wyniki dotyczące odkryć są gorsze, ale wydaje się to być naturalne. Możemy zauważyć, że po zastosowaniu korekt Bonferroniego i Benjaminiego Hochberga nasze wyniki są takie same oraz zdecydowanie lepsze w porównaniu do tych bez korekty. Niestety po zastosowaniu tych korekt, gdy rozważaliśmy model z 500 i 950 zmiennych nie odrzuciliśmy żadnej hipotezy zerowej.

Zadanie 2

W tym zadaniu powtórzymy eksperyment z poprzedniego zadania 1000 razy i dla różnych modeli wyznaczymy:

- średnią wariancję β_1,
- średnią szerokość 95% przedziału ufności β_1,
- średnią liczbę prawdziwych i fałszywych odkryć oraz estymatory FWER i FDR dla procedur testowania bez korekty oraz z korektą Bonferoniego i BH.

Wykonujemy nasz eksperyment 1000 razy, a następnie liczymy średnią dla naszych wyników. Przedstawiają się one w poniższej tabeli.

lz	$s^2(\hat{\beta}_1)$	IW	TD	FD	TD-B	FD-B	TD-BH	FD-BH
5	1.006	3.934	4.255	0	3.294	0	4.173	0
10	1.011	3.943	4.256	0.266	2.854	0.029	3.758	0.137
20	1.021	3.964	4.255	0.722	2.395	0.037	3.197	0.169
100	1.112	4.138	4.058	4.744	1.355	0.049	1.77	0.168
500	2.01	5.565	2.79	24.636	0.19	0.045	0.215	0.077
950	20.784	18.128	0.485	46.77	0.003	0.035	0.004	0.126

Widzimy, że wyniki zgadzają się z tymi z poprzedniego zadania. Zauważamy też te same własności.

Następnie chcielibyśmy porównać wartości średnich wariancji β_1 z teoretycznymi wartościami. W tym celu skorzystamy z odwrotnego rozkładu Wisharta. Do policzenia wariancji chcemy znaleźć rozkład macierzy $(\mathbb{X}'\mathbb{X})^{-1}$. Kolumnami macierzy $\mathbb{X}$ są wektory z rozkładu normalnego. Stąd z definicji rozkładu Wisharta macierz $\mathbb{X}'\mathbb{X}$ ma rozkład Wisharta. Dlatego z definicji odwrotnego rozkładu Wisharta macierz $(\mathbb{X}'\mathbb{X})^{-1}$ ma właśnie ten rozkład. Chcielibyśmy znaleźć wartość oczekiwaną takiej macierzy. Wynosi ona:

$$\frac{\Psi}{\nu - p + 1}.$$

W naszym przypadku ma ona postać:

$$\frac{1000I}{1000 - p + 1}.$$

Potrzebujemy wartość dla β_1, więc weźmiemy współrzędną $(1, 1)$ tej macierzy, czyli teoretyczna wartość wynosi

$$\frac{1000}{1000 - p + 1}.$$

Natomiast teoretyczna szerokość 95% przedziału ufności dla β_1, gdy mamy już wariancję tego estymatora jest dość prosta do wyliczenia. Szerokość przedziału wynosi:

$$\beta_1 + z_{0.975}\sigma(\beta_1) - \beta_1 + z_{0.975}\sigma(\beta_1) = 2z_{0.975}\sigma(\beta_1).$$

Wyniki tych teoretycznych wartości przedstawiamy w poniższej tabeli:

lz	5	10	20	100	500	950
$\sigma^2(\beta_1)$	1.006	1.011	1.021	1.112	2.004	20.408
IW	3.932	3.942	3.962	4.134	5.549	17.708

Zauważmy, że wyniki empiryczne pokrywają się z tymi teoretycznymi.

Dodatkowo chcielbyśmy policzyć statystyki FWER i FDR. Zadają się one wzorami:

$$\text{FWER} = \mathbb{E}[\mathbb{P}(V \geq 1)],$$
$$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{E}\left[\frac{V}{\max(R, 1)}\right]$$

Te wyniki znajdują się w poniższej tabeli.

lz	FWER	FDR	FWER-B	FDR-B	FWER-BH	FDR-BH
5	0	0	0	0	0	0
10	0.237	0.05	0.029	0.009	0.128	0.028
20	0.504	0.125	0.036	0.014	0.152	0.039
100	0.991	0.513	0.045	0.024	0.141	0.051
500	1	0.893	0.043	0.038	0.062	0.047
950	1	0.987	0.029	0.028	0.032	0.031

Widzimy, że statystyki liczone bez korekt wraz ze wzrostem rozpatrywanych zmiennych w modelu zblizają się do 1. Natomiast dla tych po korektach nasze wyniki znacznie się poprawiły, czyli prawdopodobieństwo popełnienia błędnych odkryć znacznie się zmniejszyło.

Chcielibyśmy jeszcze znaleźć na jakim poziomie kontrolujemy nasze statystyki FWER i FDR. Z teoretycznych własności wynika, że FWER bez korekt jest ograniczone w następujący sposób:

$$\text{FWER} \leq \alpha n_0,$$

gdzie $\alpha = 0.05$, n_0 to liczba prawdziwych hipotez zerowych, czyli nieistotnych zmiennych w modelu.

Natomiast oszacowanie dla FWER z korektą Bonforroniego wygląda następująco:

$$\text{FWER} \leq \alpha \frac{n_0}{n},$$

gdzie n to liczba wszytkich testowanych hipotez.

Mamy jeszcze ograniczyć FDR, ale wiemy, że FDR jest kontrolowane przez FWER, więc mamy zawsze:

$$\text{FDR} \leq \text{FWER}.$$

Stąd, nasze oszacowania dla FWER są też oszacowaniami dla FDR.

Przedstawiają się one w poniższej tabeli:

lz	5	10	20	100	500	950
FWER	0	0.25	0.75	1	1	1
FWER-B	0	0.025	0.0375	0.0475	0.0495	0.0497

Tutaj widzimy, że faktycznie nasze teoretyczne oszacowania ograniczają FWER i FDR.

Podsumowanie

Po zaobserwowaniu wszystkich wyników naszych zadań mogliśmy zauważyć znaczące różnice między innymi metodami na wielokrotne testowanie. W przypadku testowania bez żadnych korekt popełnialiśmy bardzo dużo błędnych odkryć dla nieistotnych zmiennych, ale z drugiej strony mieliśmy więcej prawdziwych odkryć.

Natomiast po zastosowaniu którejkolwiek z korekt liczba fałszywych odkryć znacznie się zmniejszyła, ale stało się to kosztem liczby prawdziwych odkryć. Właśnie w przypadku prawdziwych odkryć lepiej radziła sobie korekta Benjaminiego Hochberga.