Raport 1

Magdalena Potok

2024-04-09

Celem raportu jest przeprowadzenie analizy na danych wygenerowanych z modelem regresji liniowej, badając wpływ różnych wielkości modelu na estymatory, testy istotności, szerokość przedziałów ufności, oraz liczbę prawdziwych i fałszywych odkryć. Dodatkowo, porównanie wyników zastosowania różnych korekt na wielokrotne testowanie oraz oszacowanie wskaźników FWER i FDR dla różnych procedur testowania.

Wygenerowana została macierz planu $\mathbb{X}_{1000x950}$ tak, że jej elementy są niezależnymi zmiennymi z rozkładu normalnego $N(0, \sigma = \frac{1}{\sqrt{1000}})$. Następnie wygenerowany został wektor zmiennej odpowiedzi zgodnie z modelem

$$Y = \mathbb{X}\beta + \epsilon,$$

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

```
set.seed(567)
X = matrix(rnorm(1000*950, 0, 1/sqrt(1000)), nrow = 1000)
beta = c(5,5,5,5,5,rep(0,945))
eps = rnorm(100)
Y = X %*% beta + eps
```

W niniejszym raporcie została przeprowadzona analiza w oparciu o modele wykorzystjące: (i) pierwszych 5 zmiennych;
(ii) pierwszych 10 zmiennych;
(iii) pierwszych 20 zmiennych;
(iv) pierwszych 100 zmiennych;
(v) pierwszych 500 zmiennych;
(vi) pierwszych 950 zmiennych.

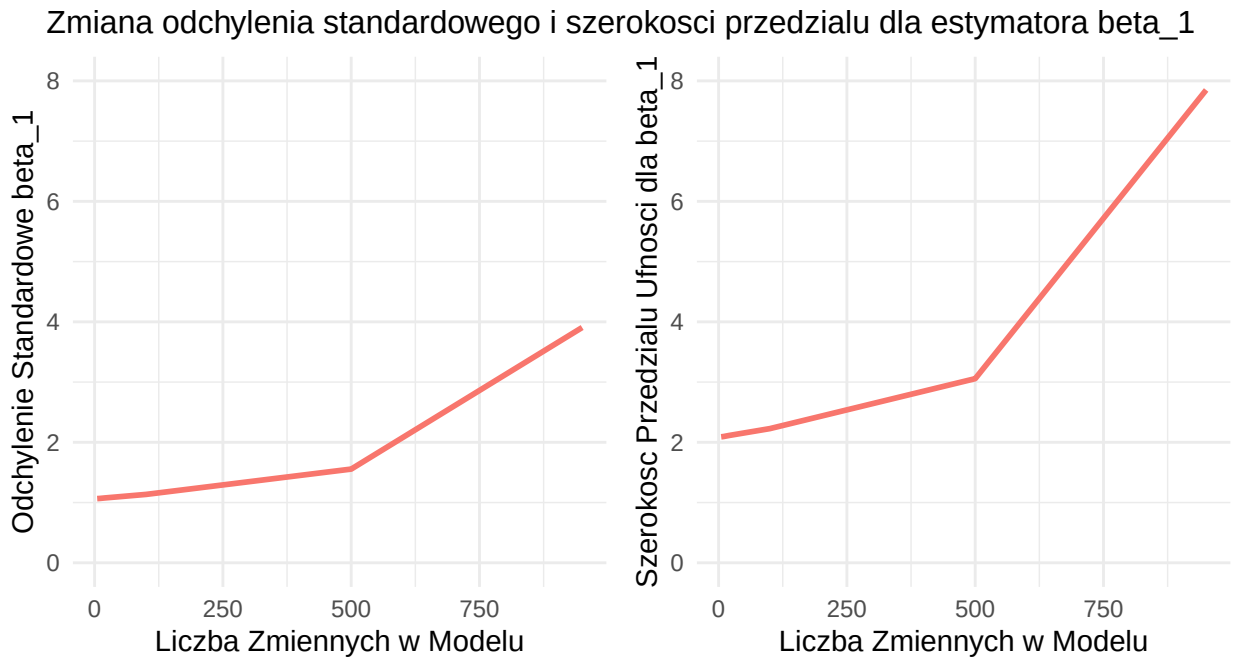
Zadanie 1

(a) Dla każdego z modeli (i)-(vi) wyznaczę estymator najmniejszych kwadratów dla wektora β i wykonam testy istotności jego elementów, wyniki zostały przedstawione w tabeli

Table 1: Liczba istotnych współczynników dla każdego modelu

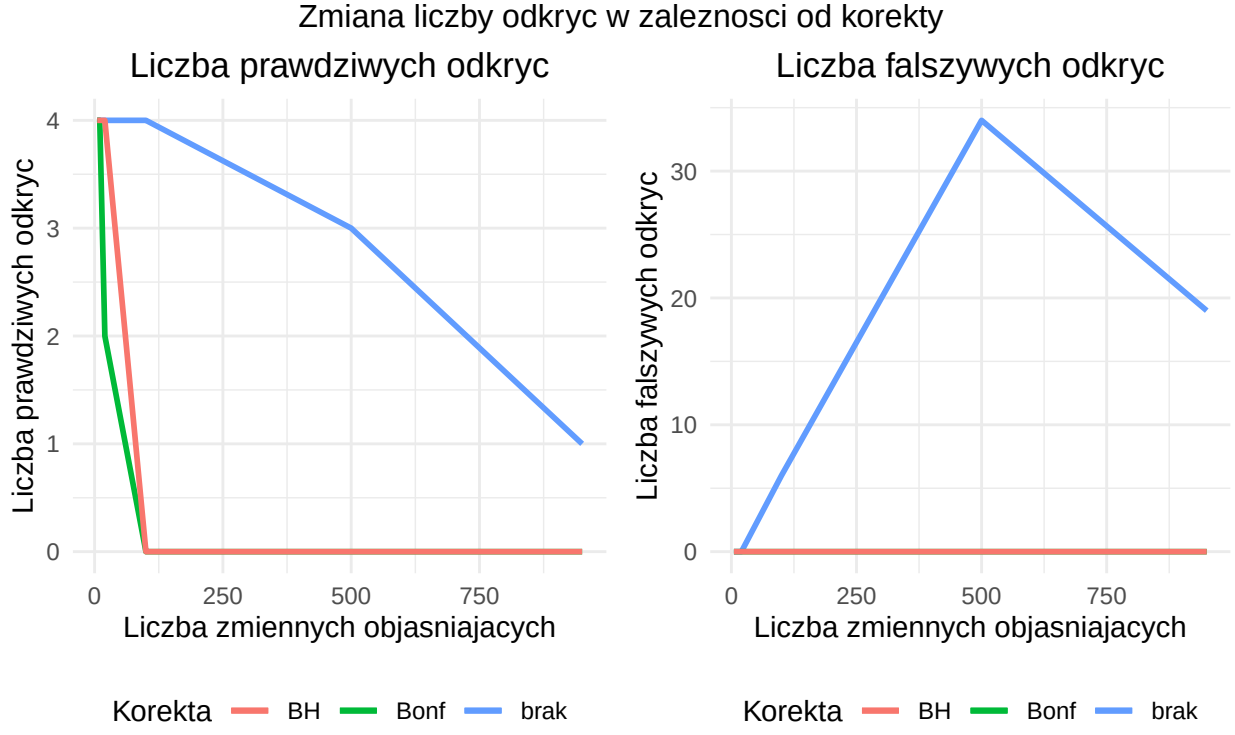
Model	(i)	(ii)	(iii)	(iv)	(v)	(vi)
Istotne_Wspolczynniki	5	6	6	8	28	88

(b) Porównam jak zmienia się odchylenie standardowe estymatora β_1 i szerokość 95% przedziału ufności dla tego parametru w miarę tego jak rośnie rozważany model.



Możemy zauważyć, że wraz ze wzrostem zmiennych objaśniających rośnie odchylenie standardowe oraz szerokość przedziału ufności tego estymatora. Związek między odchyleniem a przedziałem jest naturalny, ponieważ szerokość przedziału ufności estymatora jest wprost związana z jego odchyleniem standardowym. W przypadku rosnącej wartości odchylenia standardowego otaz wzrostu szerokości przedziału, tym trudniej jest odrzucić hipotezę zerową H_0.

(c) i (d) Porównam liczbę prawdziwych i fałszywych odkryć dla różnych modeli nie używając korekty, używając korekty Bonferroniego i używając korekty Benjaminiego-Hochbergana.

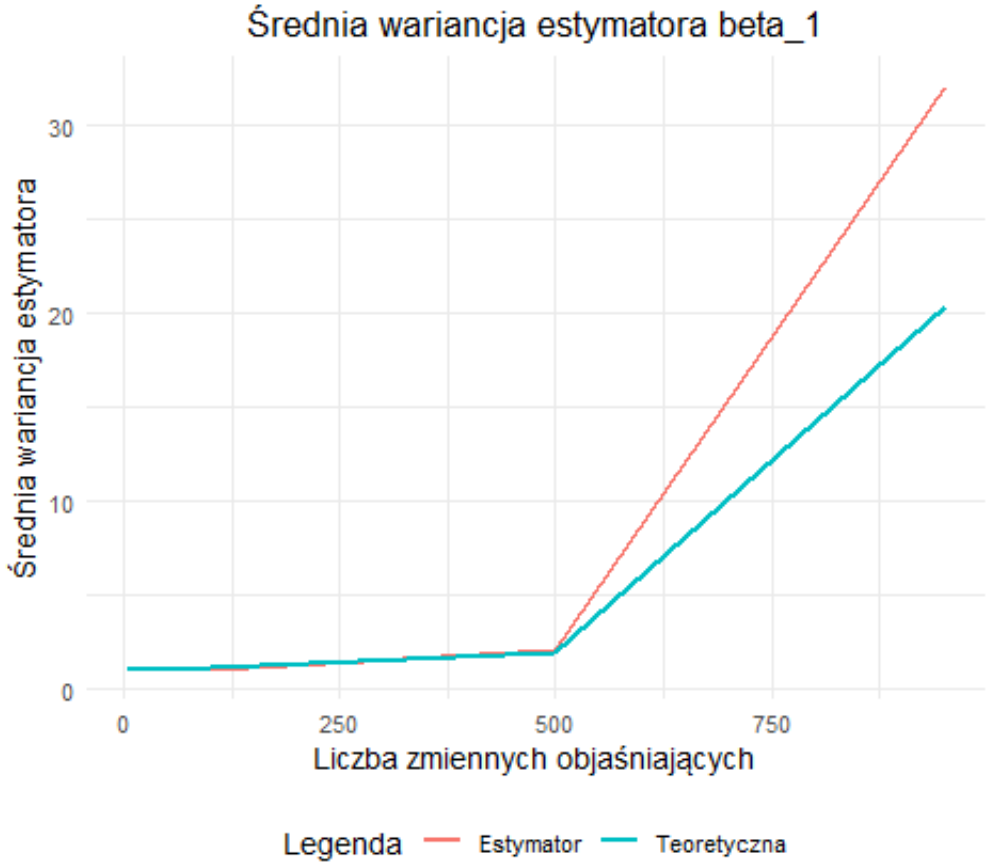


Z wykresu dla prawdziwych odkryć możemy odczytać, że wraz z zwiększającą się liczbą zmiennych objaśniających maleje liczba prawdziwych odkryć w każdej z proponowanych procedur. To co się rzuca w oczy, to fakt, że wybór obu korekt oznacza, że będziemy mieć mniej prawdziwych odkryć, niż w przypadku, gdy nie stosujemy korekty.
Wykres fałszywych odkryć pokazuje, że w przypadku zastosowania obu korekt mamy bardzo małą, bliską zero, ilość fałszywych odkryć - niezależnie od ilości zmiennych objaśniających. W sytuacji, gdy nie mamy korekty na początku ilość fałszywych odkryć rośnie, aż $p = 500$, potem możemy zaobserwować spadek.

Zadanie 2

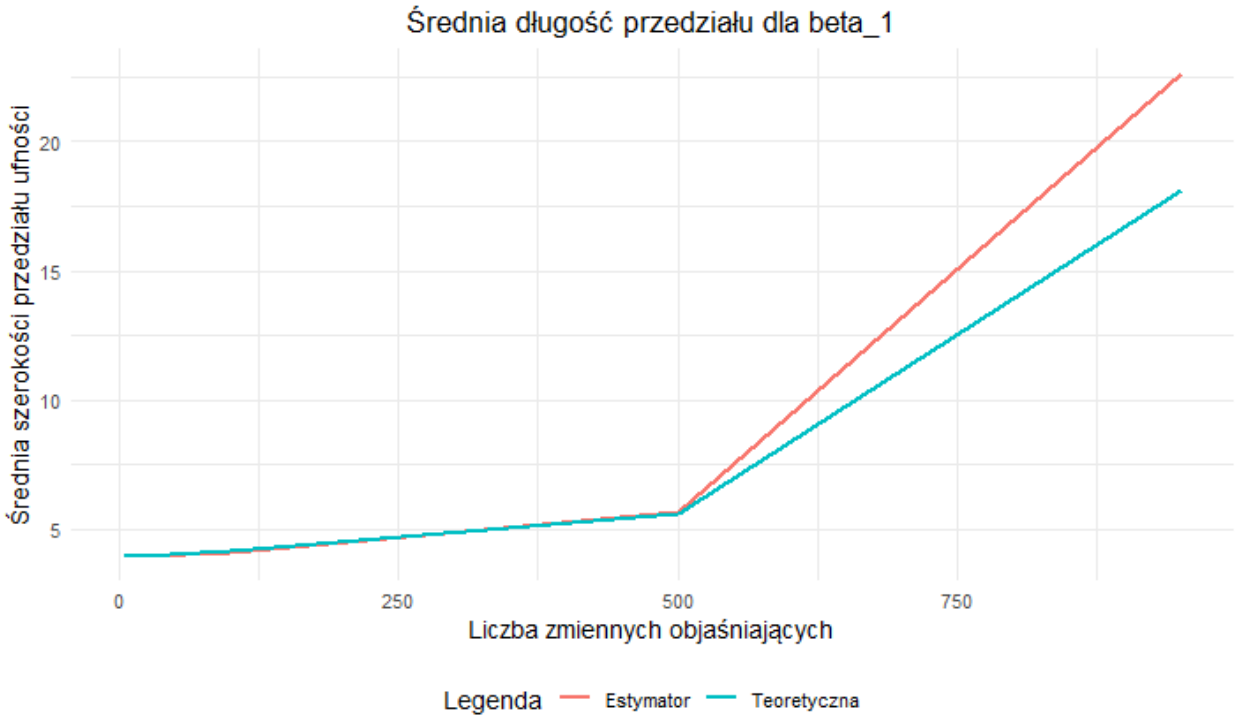
Doświadczenie z 1. zadania powtórzę 1000 razy i wyznaczę dla różnych modeli:

> Średnią wariancję estymatora β_1 i porównam z wartością teoretyczną.



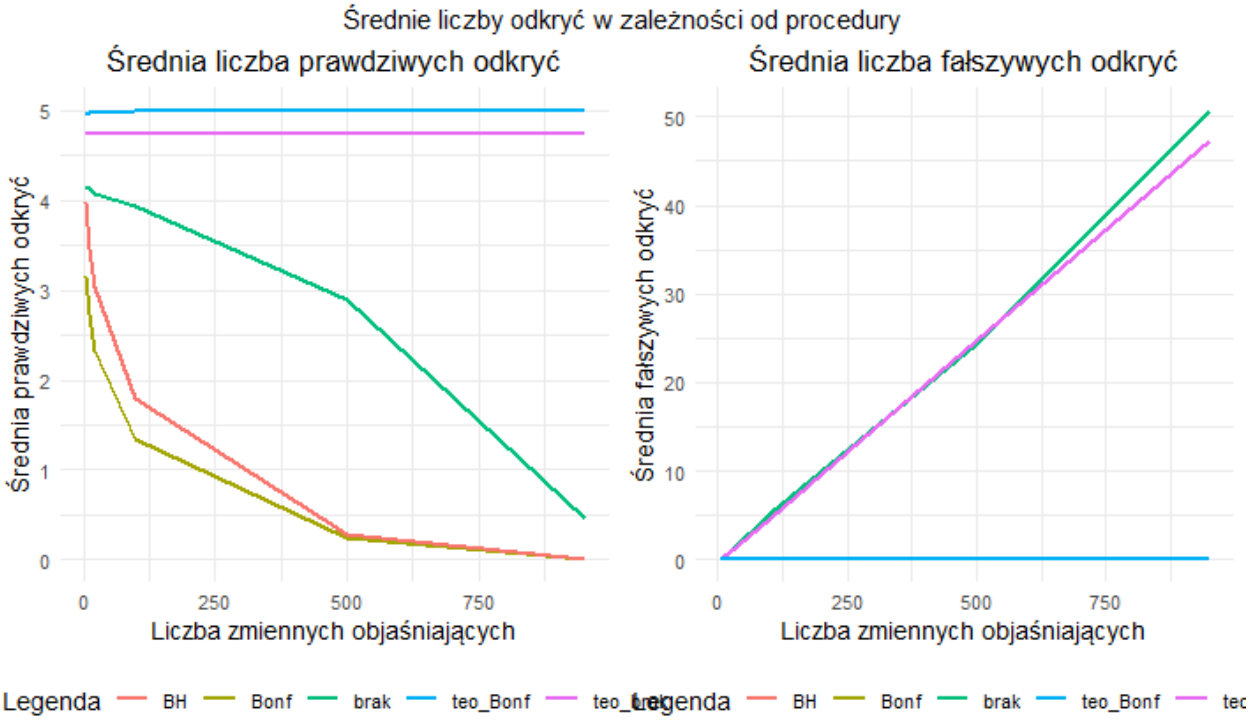
Wartość teoretyczna została wyliczona za pomocą odwrotnego rozkładu Wisharta. Z wykresu możemy odczytać, że wyniki symulacji są całkiem blisko wartości teoretycznej. Przy zwiększaniu się liczby zmiennych objaśniających (od $p = 500$) te wartości zaczynają się rozbiegać, wartość teoretyczna jest mniejsza. Z zwiększającą się liczbą regresorów zwiększa się średnia wariancja estymatora.

> Średnią szerokość 95% przedziału ufności dla β_1 i porównam z teoretycznym oszacowaniem.

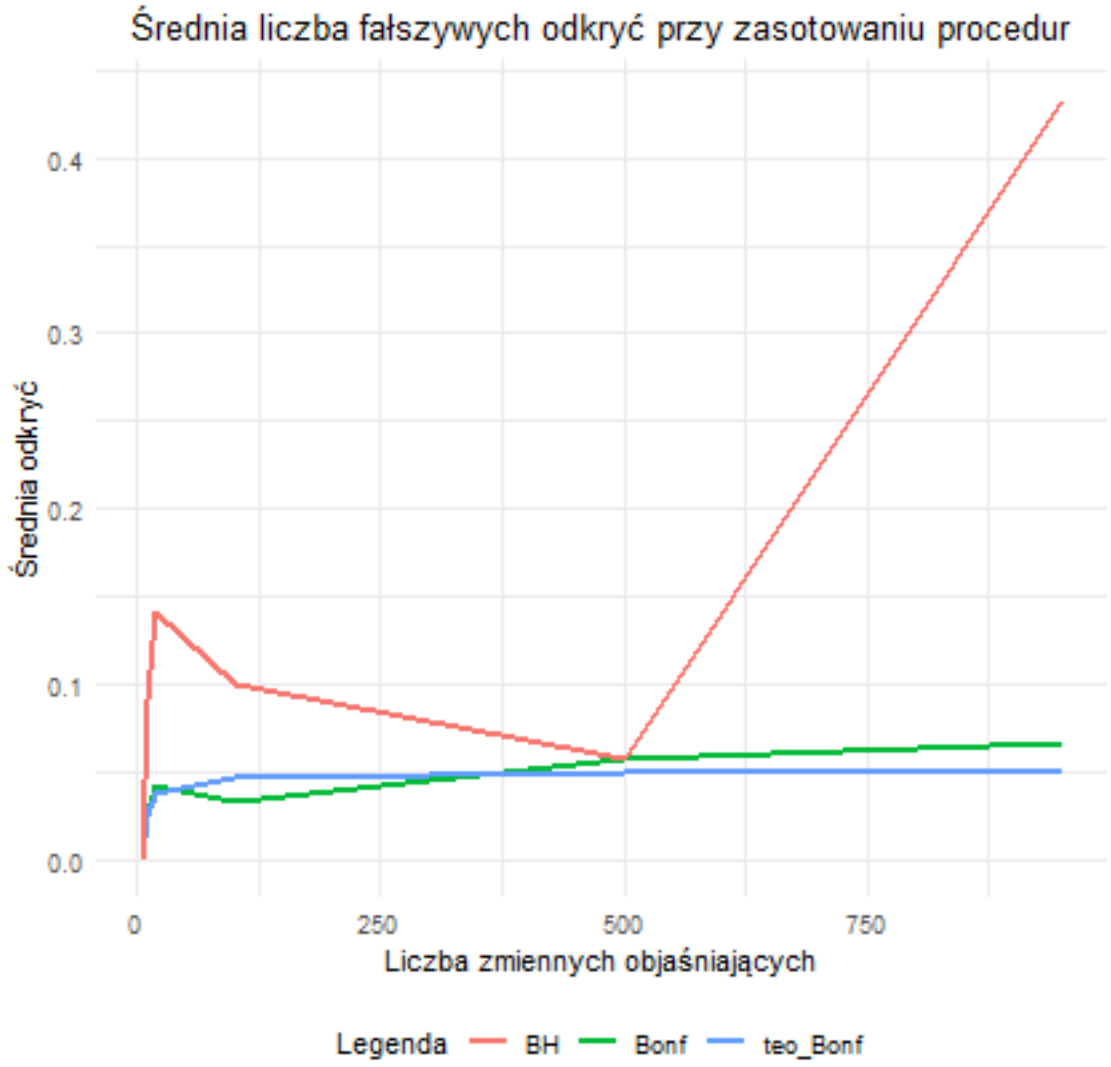


W poprzednim podpunkcie przekonaliśmy się, że ze wzrostem parametru p wzrsata nam średnia wariancja estymatora β_1. Naturalnie z większą wariancją zwiększa nam się szerokość przedziału ufności, co możemy odczytać z wykresu. Tym razem również do wartości $p = 500$ średnia długość teoretyczna pokrywa się z estymowaną, później zaczynają, tak jak w przypadku wariancji, od siebie odbiegać (średnia wariancja wyliczona teoretycznymi wzorami jest mniejsza).

> Średnią liczbę prawdziwych i fałszywych odkryć dla różnych procedur testowania.

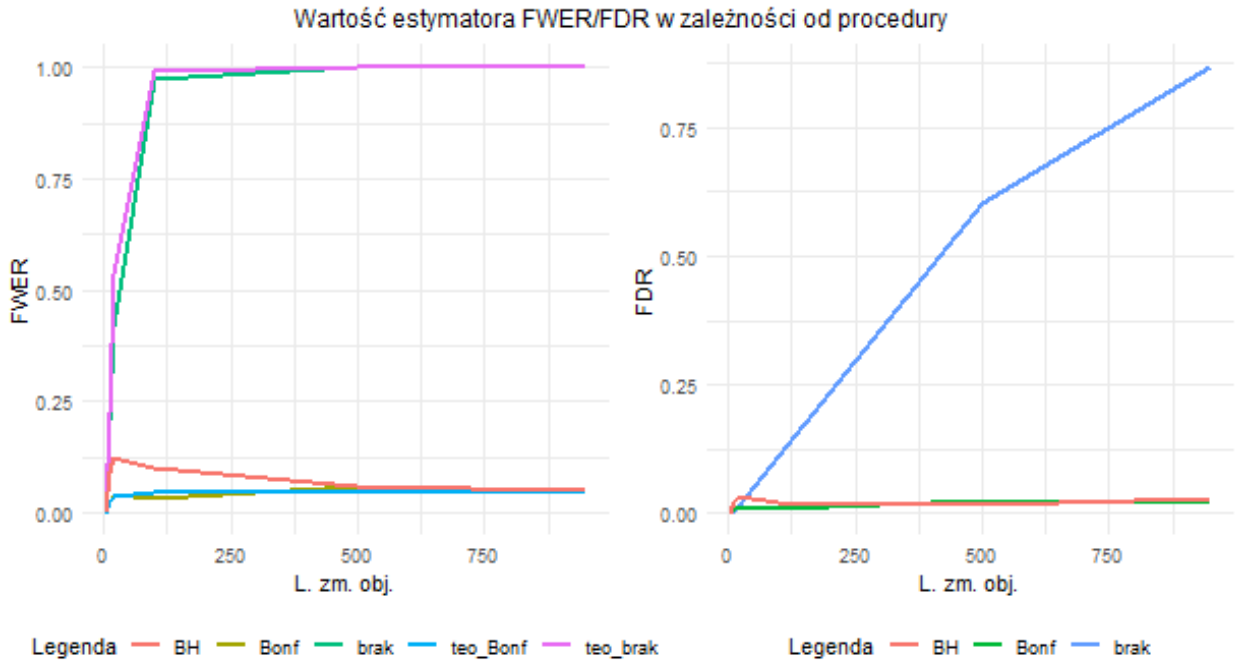


Skupmy się wpierw na wykresie z średnią liczbą prawdziwych odkryć. Możemy zauważyć, że bez używania korekt mamy więcej prawdziwych odkryć niż gdy używamy korekt. Co oznacza, że zwykłe testowanie dobrze sprawdza się, jeśli zależy nam na zidentyfikowaniu istotnych zmiennych. Natomiast wykres średniej fałszywych odkryć tłumaczy nam, dlaczego potrzebne są korekty. Pokazuje on, że zwykłe testowanie wskazuje wiele zmiennych jako istotne, podczas, gdy tak naprawdę są nieznaczące. Jednak przy użyciu korekt Bonferoniego lub Benjaminiego-Hochberga liczba fałszywych odkryć jest znacząco mniejsza, bliska 0. Te metody są preferowane nad zwykłym testowaniem, ponieważ eliminują fałszywie istotne zmienne. Jednakże, kosztem tego jest mniejsza liczba faktycznie istotnych zmiennych, które są identyfikowane. Przyjrzymy się teraz liczbie fałszywych odkryć na osobnym wykresie, ponieważ wspólna skala na jednym wykresie nie pozwala na dokładne przeanalizowanie zależności.



Metoda Bonferroniego charakteryzuje się najmniejszą liczbą fałszywych odkryć, co wynika z jej konserwatywnej natury. Z kolei metoda BH, choć może mieć nieco więcej fałszywych odkryć, to zazwyczaj identyfikuje więcej prawdziwych zmiennych.

> Estymatory FWER i FDR dla procedur testowania bez korekty oraz z korektą Bonferoniego i BH.



Zauważmy, że w przypadku, gdy testujemy bez korekt wartość FWER bardzo szybko zbiega do 1, co jest zdecydowanie więcej niż przyjęty poziom $\alpha = 0.05$, wynik jest bardzo zbliżony do wyliczeń teoretycznych. Gdy popatrzymy na korektę Bonferroniego w przypadku FWER, jak wiemy z wykładu, jest ona kontrolowana na poziomie α, z wykresu widzimy, że ta wartość jest o wiele mniejsza, bliska zero. Ta wartość również zgadza się z teoretycznymi obliczeniami. Korzystając z korekty Bonferroniego, minimalizujemy ryzyko popełnienia błędów typu I, ale kosztem mniejszej czułości testu na wykrycie rzeczywistych efektów. Metoda Benjamini-Hochberga nie kontroluje FWER, co możemy odczytać z wykresu. Jest moment na wykresie, kiedy wartość FWER dla tej metody jest większa niż 0.05, jednak z zwiększającą się liczbą zmiennych objaśniających wartość FWER, tak jak dla Bonf (ale trochę wolniej), zbiega do bardzo małych wartości, bliskich 0.
Wartość FDR, czyli proporcja fałszywych odkryć do ilości odrzuconych hipotez, w przypadku obu korekt jest bardzo mała, wynika to z tego, że obie są kontrolowane na poziomie $\alpha = 0.05$. Na początku ta wartość w przypadku korekty BH jest większksza, ale ze wzrostem zmiennych objaśniających te wartości zaczynają być bardzo bliskie sobie. W przypadku, gdy nie stosujemy korekt wartość FDR rośnie, ta sytuacja jest niepożądana.

Wnioski

- Wzrost liczby zmiennych objaśniających wpływa na wariancję estymatora β_1, a co za tym idzie na jego szerokość przedziału ufności. Zaobserwowaliśmy za równo w przypadku 1 powtórzenia eksperymentu, jak i gdy wykonaliśmy 1000 powtórzeń. Tracimy więc dokładność w estymacji parametru.
- Testowanie bez korekty, z uwagi na ilość fałszywych odkryć, okazuje się mało skutecznym testowaniem i może skutkować nieprawidłowymi wnioskami.
- Korekty, takie jak Bonferroni czy metoda Benjamini-Hochberga, mają istotne znaczenie w ograniczaniu liczby fałszywych odkryć. Po ich zastosowaniu średnia liczba fałszywych odkryć spada praktycznie do zera. Jednakże, stosowanie korekt może również skutkować redukcją liczby faktycznie istotnych zmiennych.
- Wybór odpowiedniej metody korekty zależy od kontekstu badania oraz preferencji badacza, uwzględniając zarówno kontrolę błędów, jak i zachowanie mocy statystycznej testu.