Raport 4

Magdalena Potok

2024-01-08

Zadanie 1

(a) Wygeneruję macierz $X_{100\times 2}$ taką, że jej wiersze będą niezależnymi wektorami losowymi z wielowymiarowego rozkłau normalnego $N(0, \Sigma/100)$, gdzie $\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$.

```
Sigma <- matrix(c(1, 0.9, 0.9, 1), nrow = 2, byrow = TRUE)
X <- mvrnorm(n= 100, c(0,0), Sigma = Sigma/100)
```

Pierwsza kolumna macierzy X w tym kodzie, to regresor X_1, natomiast druga kolumna to X_2. Teraz wygeneruję wektor zmiennej odpowiedzi postaci $Y = \beta_1 X_1 + \epsilon$, gdzie $\beta_1 = 3$, $\epsilon \sim N(0, I)$.

```
X1 <- X[,1]
epsilon <- rnorm(100, 0, 1)
Y <- 3*X1 + epsilon
```

(b) Wyznaczę 95% przedział ufności dla wartości β_1 i przeprowadzę t-test na poziomie istotności 0.5 dla hipotezy $\beta_1 = 0$, przy użyciu
- modelu prostej regresji liniowej $Y = \beta_0 + \beta_1 X_1 + \epsilon$,
- modelu z dwiema zmiennymi objaśniającymi $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.

	LPU	PPU	Wartość_krytyczna	T_Statystyka	T_test_wynik
Model 1	1.041	5.639	1.984	2.883	TRUE
Model 2	-0.697	9.305	1.985	1.708	FALSE

Oba przedziały ufności zawierają prawdziwą wartość $\beta_1 = 3$, jednak dla modelu 2 ten przedział jest znacznie szerszy. Wynika to z tego, że drugi mode4l zawiera dodatkową zmienną X_2, co może wpływać na relację między X_1 a zmienną objaśnianą Y. Dodatkowa zmienna wprowadza zmienność, która ma wpływ na szersze prziały ufności. W przypadku 1. modelu na poziomie istotności $\alpha = 0.05$ odrzucamy hipotezę zerową o braku korealcji między Y i X_1, jednak w przypadku 2. modelu nie mamy podstaw aby odrzucić tę hipotezę. Dodanie zmiennej X_2 wpływa na statystykę t-studenta dla współczynnika X_1, przez co w drugim modelu wyszła ona niższa od wartości krytycznej.

(c) Obliczę ręcznie odchylenie standardowe estymatora β_1 i moc identyfikacji X_1 w obu modelach.

$$s^2(\hat{\beta}_1) = \sigma^2(X'X),$$

gdzie σ^2 to odchylenie standardowe błędów losowych, tutaj wynosi 1, a X to macierz, która składa się w pierwszej kolumnie z samych 1, druga kolumna to X_1, a trzecia X_2.

```
MX1 <- cbind(1, X1)
s1 <- ((solve(t(MX1) %*% MX1))^(1/2))[2,2]
MX2 <- cbind(1,X1,X2)
s2 <- ((solve(t(MX2) %*% MX2))^(1/2))[3,3]
```

Mod identyfikacji X_1, to moc testu dla powyżej przeprowadzonego testu t-studenta

$$Moc\ testu = P_{\beta_1=3}(|T| > t_c) = P_{\beta_1=3}(T < -t_c) + P_{\beta_1=3}(T > t_c) = P_{\beta_1=3}(T < -t_c) + 1 - P_{\beta_1=3}(T < t_c)$$

Statystyka T ma niecentralny rozkład studenta z parametrem przesunięcia $ncp = \frac{\beta_1}{s(\hat{\beta}_1)}$.

```
moc1 <- pt(-qt(1-0.05/2, length(X1)-2), length(X1) - 2, 3/s1) + 1 -
  pt(qt(1-0.05/2, length(X1)-2), length(X1) - 2, 3/s1)
moc2 <- pt(-qt(1-0.05/2, length(X1)-3), length(X1) - 3, 3/s2) + 1 -
  pt(qt(1-0.05/2, length(X1)-3), length(X1) - 3, 3/s2)
```

	Odchylenie_stand	Moc_testu
Model 1	1.115	0.759
Model 2	2.416	0.233

Moc testu dla modelu z jedną zmienną objaśniającą jest zadowalający, czyli test jest dobry w wykrywaniu fałszywej hipotezy zerowej. Dla drugiego modelu ten wynik jest znacząco niższy, czyli prawdopodobieństwo odrzucenia H_0, gdy rzeczywiście jest fałszywa jest niskie.
Większe odchylenie dla drugiego modelu oznacza, że precyzja naszej estymacji współczynnika nachylenia X_1 jest niższe niż dla modelu 1. Wyższy wynik oznacza większą zmienność lub rozproszenie estymacji, dodatkowa zmienność może być spowodowana dodaniem drugiej zmiennej.

(d) Wygeneruję 1000 niezależnych kopii wektora błędów losowych ϵ i 1000 odpowiednich kopii wektora zmiennej odpowiedzi. Dla każdego ze zbiorów wyznaczę estymator β_1 i wykonam test istotności dla β_1 w obu modelach. Wyestymuję odchylenie standardowe β_1 oraz moc testu.

	Odchylenie_stand	Moc_testu
Model 1	1.126	0.722
Model 2	2.454	0.237

Wyniki doświadczalne oraz teoretyczne (podpunkt c) są sobie bardzo bliskie, co jest spodziewanym wynikiem przy tak wielu powtórzeniach.

Zadanie 2

(a) Wygeneruję macierz planu $\mathbb{X}_{1000\times 950}$ tak, że jej elementy są niezależnymi zmiennymi losowymi z rozkładu normalnego $N(0, \sigma = 0.1)$. Następnie wygeneruję wektor zmiennej odpowiedzi według modelu

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

gdzie $\beta = (3, 3, 3, 3, 3, 0, ..., 0)^T$.

```
X <- matrix(rnorm(950000, 0, 0.1), nrow = 1000)
beta <- rep(0, 950)
beta[1:5] <- 3
Y <- X%*%beta+rnorm(1000)
```

Oznacza to, że X_1, X_2, X_3, X_4, X_5 będą miały wpływ na model, a pozostałe X_i będą wymnażane przez $\beta_i = 0$.

(b) Wyestymuje wartości współczynników regresji i wykonam t-testy na poziomie istotności 0.05, aby zidentyfikować istotne regresory, gdy model jest zbudowany przy użyciu pierwszych k kolumn macierzy planu dla $k \in \{1, 2, 5, 10, 50, 100, 500, 950\}$. Dla każdego z tych modeli podam:
- sumę kwadratów residów $SSE = \|Y - \hat{Y}\|^2$;

- błąd średniokwadratowy estymatora wartości oczekiwanej Y : $MSE = \|X(\hat{\beta} - \beta)\|^2$;
- wartość kryterium AIC: $AIC = nlog(SSE/n) + 2k$;
- p-wartości dla dwóch pierwszych zmiennych objaśniających;
- liczbę fałszywych odkryć.

	SSE	MSE	AIC	P.value.1	P.value.2	FD
1	1332.552	1.335	287.096	0.000	NA	NA
2	1260.480	1.264	233.492	0.000	0.000	NA
5	992.731	0.999	0.704	0.000	0.000	NA
10	984.746	0.996	2.628	0.000	0.000	1
50	931.052	0.981	26.560	0.000	0.000	5
100	872.806	0.971	61.958	0.000	0.000	9
500	494.289	0.991	293.365	0.000	0.000	23
950	49.487	1.010	-1108.045	0.082	0.004	36

Analiza wyników przedstawia istotne obserwacje dotyczące różnych wskaźników modelu regresji. Suma kwadratów błędów predykcji (SSE) wykazuje tendencję do wzrostu w miarę zmniejszania liczby zmiennych objaśniających (X_i). Model pełny, wykorzystujący wszystkie zmienne, charakteryzuje się najniższym SSE, co sugeruje, że jego predykcje są najbliższe wartościom obserwowanym.
Natomiast średni kwadratowy błąd predykcji (MSE) utrzymuje się na stosunkowo stabilnym poziomie, niezależnie od liczby zmiennych objaśniających w modelu. To wskazuje, że MSE nie zmienia się znacząco wraz ze zmianą X_i.
Wartość kryterium AIC (Akaike's Information Criterion) służy do porównania różnych modeli, a niższa dodatnia wartość AIC wskazuje na lepsze dopasowanie modelu do danych. Model z 5 zmiennymi objaśniającymi osiągnął najniższy AIC, co sugeruje, że najlepiej odwzorowuje zależności w danych. Natomiast model pełny miał wartość ujemną AIC, co może wskazywać na zbyt dobrze dopasowany model.
Analiza P-wartości pokazuje, że w większości przypadków hipotezy zerowe są odrzucane (dla P-wartości_i $<$ α). Warto jednak zauważyć, że nawet dla modelu, w którym nie można odrzucić H_0 (P-wartość_1 $> \alpha$), istnieje istotność drugiej zmiennej (P-wartość_2 $< \alpha$). Wiersze z wartościami 'NA' wskazują, że te konkretne zmienne nie są uwzględnione w danym modelu.
Ostatnia kolumna, dotycząca liczby wyników fałszywie pozytywnych (FD), rośnie wraz ze wzrostem liczby zmiennych objaśniających. Wartości 'NA' w trzech pierwszych wierszach sygnalizują, że te zmienne nie są uwzględnione, co sugeruje brak fałszywych wyników odkryć dla modelu z 5 zmiennymi objaśniającymi, które rzeczywiście mają wpływ na zmienną objaśnianą.

(c) Powtórzę punkt (b), gdy modele są konstruowane przy pomocy zmiennych o największych (niekoniecznie pierwszych) estymowanych współczynnikach regresji.

	SSE	MSE	AIC	P.value.1	P.value.2	FD
1	1308.568	1.311	268.934	0.000	NA	0
2	1216.523	1.220	197.996	0.000	0.000	0
5	1062.577	1.069	68.697	0.000	0.000	0
10	1050.453	1.062	67.221	0.000	0.000	1
50	999.666	1.053	97.666	0.000	0.000	5
100	887.945	0.988	79.155	0.000	0.000	6
500	513.706	1.029	331.896	0.000	0.000	22
950	49.487	1.010	-1108.045	0.003	0.004	36

Analiza porównawcza dwóch zestawów wyników modeli regresji wskazuje na podobne trendy w zależności od liczby zmiennych objaśniających. W obu zestawach obserwujemy spadek zarówno sumy kwadratów błędów predykcji (SSE) jak i średniego kwadratowego błędu (MSE) w miarę zmniejszania liczby zmiennych objaśniających. Zestawienia te pokazują również spadek wartości kryterium AIC wraz z redukcją zmiennych,

gdzie modele z pięcioma zmiennymi objaśniającymi osiągają (prawie)najniższe wartości AIC, sugerując ich najlepsze dopasowanie do danych. Wartości P-wartości również maleją wraz ze zmniejszeniem liczby zmiennych, co wskazuje na istotność statystyczną zmiennych, zwłaszcza dla mniejszych modeli. Obserwujemy wzrost liczby fałszywie pozytywnych wyników (FD) wraz z dodawaniem zmiennych objaśniających, przy czym modele z mniejszą liczbą zmiennych wykazują brak fałszywych wyników odkryć. Oba zestawy danych prezentują spójne wnioski, że modele z pięcioma zmiennymi objaśniającymi wykazują najlepsze dopasowanie do danych, biorąc pod uwagę równowagę między precyzją a złożonością modelu. Gdybyśmy mieli wskazać dla tego podpunktu, który model wybrać jedynie na podstawie AIC, to byłby to model z 10 zmiennymi objaśniającymi, ponieważ to właśnie on ma najmniejszą dodatnią wartość AIC.

(d) Powórzę generowanie ϵ i Y oraz punkty (b) i (c) 50 razy. Dla każdego z zadań obliczę moc identyfikacji X_1, X_2 i średnią liczbę fałszywych odkryć. Dodatkowo oszacuję średni rozmiar modelu wybranego przez AIC dla punktów (b) i (c).

	1	2	5	10	50	100	500	950
AIC(b)	302.194	235.870	3.019	8.036	48.212	92.272	301.962	-1167.166
AIC(c)	339.510	307.912	274.103	245.766	167.902	147.082	304.093	-1167.166
FD(b)	0.000	0.000	0.000	0.200	2.250	4.700	25.850	54.900
FD(c)	0.100	0.150	0.400	0.700	3.950	8.250	25.650	54.900
Moc(b)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.450
Moc(c)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.450

Analiza wyników wskazuje na istotne różnice pomiędzy strategiami wyboru modeli w zależności od liczby zmiennych objaśniających. W przypadku (b), dla 50 powtórzeń, model z 5 zmiennymi (X1, X2, X3, X4, X5) charakteryzuje się najniższym średnim AIC, co jest zgodne z pojedynczym wyliczeniem. Natomiast w (c), model z 100 zmiennymi objaśniającymi wskazuje najniższą średnią wartość AIC, podczas gdy przy pojedynczym obliczeniu najlepszy był model z 10 zmiennymi. Zmiana ta wynika z różnic w uporządkowaniu zmiennych X1, X2, X3, X4, X5 względem wpływu na Y w kolejnych wygenerowanych zestawach danych. Warto zauważyć, że AIC dla poszczególnych liczb zmiennych ma większy rozrzut w (b) niż w (c). Liczba fałszywych odkryć (FD) rośnie wraz z liczbą zmiennych w obu przypadkach, choć nieznacznie więcej występuje ich w (c). Moc, która wskazuje na błędne nieodrzucenie hipotezy zerowej, osiąga maksymalną wartość (1) dla większości modeli z wyjątkiem modelu pełnego, gdzie wynosi 0.450. Średnia moc jest zgodna z informacją z P-wartości, sugerując odrzucenie hipotez zerowych poza modelem pełnym, gdzie nieodrzucenie tych hipotez może wystąpić częściej.