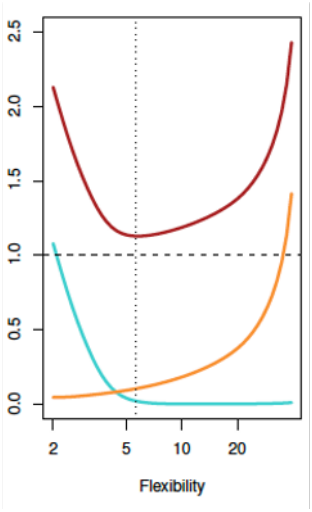
# Kolokwium SAD 2021

Dorota Celińska-Kopczyńska, Krzysztof Gogolewski, Magda Grynkiewicz,
Krzysztof Koras, Błażej Miasojedow, Piotr Pokarowski,
Agnieszka Stępień-Baran, Ewa Szczurek

Kwiecień 2021

**Zadanie 1** [**Autor: ES, punkty: 2, gr 1**] Na wykresie przedstawiono krzywe odpowiadające: - obciążeniu modelu w zależności od jego elastyczności, - wariancji modelu w zależności od jego elastyczności, - błędowi testowemu w zależności od jego elastyczności, - błędowi nieredukowalnemu w zależności od jego elastyczności. Pionową prostą zaznaczono elastyczność modelu M, z którego wygenerowano dane treningowe i testowe.




Rysunek 1: Caption

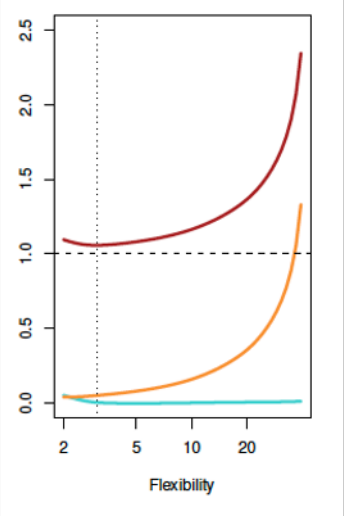
Wybierz właściwą odpowiedź:

- Krzywa czerwona odpowiada wariancji modelu w zależności od jego elastyczności.
- Krzywa pomarańczowa odpowiada obciążeniu modelu w zależności od jego elastyczności.

- Błąd nieredukowalny wzrasta wraz z elastycznością modelu.

**SOL** Błąd testowy zależy od obciążenia, wariancji i błędu nieredukowalnego modelu.

**Zadanie 1 [Autor: ES, punkty: 2, gr 2]** Na wykresie przedstawiono krzywe odpowiadające: - obciążeniu modelu w zależności od jego elastyczności, - wariancji modelu w zależności od jego elastyczności, - błędowi testowemu w zależności od jego elastyczności, - błędowi nieredukowalnemu w zależności od jego elastyczności. Pionową prostą zaznaczono elastyczność modelu M, z którego wygenerowano dane treningowe i testowe.




Rysunek 2: Caption

Która z odpowiedzi jest nieprawdziwa:

**SOL** Krzywa czerwona odpowiada wariancji modelu w zależności od jego elastyczności.

- Błąd nieredukowalny nie zależy od elastyczności modelu.
- Błąd testowy zależy od obciążenia, wariancji i błędu nieredukowalnego modelu.
- Krzywa pomarańczowa odpowiada wariancji modelu w zależności od jego elastyczności.

**Zadanie 2 [Autor: MG, punkty: 3, gr 1]** Załóżmy, że czas rozwiązywania jednego zadania na kolokwium (w minutach) to zmienna losowa $X$ mająca rozkład wykładniczy z nieznanym parametrem $\lambda$. Wiedząc, że dla losowo wybranej

próby 60 zmierzonych czasów rozwiązywania zadań, łączny czas wyniósł 500 minut, podaj (i) wartość estymatora największej wiarygodności parametru $\lambda$ oraz (ii) oczekiwany czas rozwiązywania 10 zadań z dokładnością do minuty.

- (i) 8.33(3), (ii) 83 minuty.
- (i) 0.12, (ii) nie można obliczyć, nie znając odpowiednich wartości funkcji $\Gamma$ i $\gamma$.

**SOL** (i) 0.12, (ii) 83 minuty.

- (i) 8.33(3), (ii) nie można obliczyć, nie znając odpowiednich wartości funkcji $\Gamma$ i $\gamma$.

**Zadanie 2 [Autor: MG, punkty: 3, gr 2]** Załóżmy, że czas rozwiązywania jednego zadania na kolokwium (w minutach) to zmienna losowa $X$ mająca rozkład wykładniczy z nieznanym parametrem $\lambda$. Wiedząc, że dla losowo wybranej próby 120 zmierzonych czasów rozwiązywania zadań, łączny czas wyniósł 960 minut, podaj (i) wartość estymatora największej wiarygodności parametru $\lambda$ oraz (ii) prawdopodobieństwo, że losowo wybrany student rozwiąże wszystkie zadania na kolokwium złożonym z 10 zadań przed upływem 90 minut.

- (i) 0.125, (ii) $\sim 0.675$.

**SOL** (i) 0.125, (ii) nie można obliczyć, nie znając odpowiednich wartości funkcji $\Gamma$ i $\gamma$.

- (i) 8.0, (ii) $\sim 0.675$.
- (i) 8.0, (ii) nie można obliczyć, nie znając odpowiednich wartości funkcji $\Gamma$ i $\gamma$.

**Zadanie 3 [Autor: KK, punkty: 3, gr 1]** Załóżmy, że cena za litr benzyny ma rozkład normalny $N(\mu, \sigma^2)$ o nieznanym $\mu$ i znanym $\sigma$. Na podstawie próby 30 stacji benzynowych wyznaczono przedział ufności na poziomie ufności 0.95 dla średniej ceny za litr benzyny na $(4.56, 5.32)$. (i) Ile wynosi odchylenie standardowe $\sigma$?. (ii) Ile powinien wynosić rozmiar próby (liczba sprawdzonych stacji benzynowych) aby rozważany przedział ufności wynosił $(4.75, 5.13)$ (przy zachowanym tym samym $\sigma$)? Wskazówka: w tym zadaniu może przydać się obliczenie kwantyla w R.

- (i) $\sim 1.062$, (ii) 60
- (i) $\sim 1.270$, (ii) 120

**SOL** (i) $\sim 1.062$, (ii) 120

- (i) $\sim 0.987$, (ii) 120

**Zadanie 3** [**Autor: KK, punkty: 3, gr 2**] Załóżmy, że cena za litr benzyny ma rozkład normalny $N(\mu, \sigma^2)$ o nieznanym $\mu$ i znanym $\sigma$. Na podstawie próby 35 stacji benzynowych wyznaczono przedział ufności na poziomie ufności 0.95 dla średniej ceny za litr benzyny na $(4.28, 4.96)$. (i) Ile wynosi odchylenie standardowe $\sigma$?. (ii) Ile powinien wynosić rozmiar próby (liczba sprawdzonych stacji benzynowych) aby rozważany przedział ufności wynosił $(4.45, 4.79)$ (przy zachowanym tym samym $\sigma$)? Wskazówka: w tym zadaniu może przydać się obliczenie kwantyla w R.

**SOL** (i) $\sim 1.026$, (ii) 140

- (i) $\sim 1.227$, (ii) 140
- (i) $\sim 1.026$, (ii) 70
- (i) $\sim 1.183$, (ii) 140

**Zadanie 4** [**Autor: PP, punkty: 3, gr 1**] Załóżmy, że cena za litr benzyny ma rozkład normalny $N(\mu, 1)$ o nieznanym $\mu$. Na podstawie próby $X_1, \ldots, X_{30}$ z 30 stacji benzynowych otrzymano średnią $\bar{X} = 4.9$, a następnie testem najmocniejszym testowano hipotezę $H_0 \colon \mu = 4.5$ przeciw $H_1 \colon \mu > 4.5$ na poziomie istotności 0.01. (i) Ile wynosi moc testu w $\mu = 5.0$? (ii) Ile wynosi p-wartość statystyki testowej? Wskazówka: tutaj moglibyśmy podać różne kwantyle i wartości dystrybuant, ale zamiast tego zachęcamy do skorzystania z obliczeń w R.

- (i) $\sim 0.34$, (ii) $\sim 0.014$
- (i) $\sim 0.34$ , (ii) $\sim 0.986$
- (i) $\sim 0.905$, (ii) $\sim 0.986$

**SOL** (i) $\sim 0.66$, (ii) $\sim 0.014$

**Zadanie 4** [**Autor: PP, punkty: 3, gr 2**] Załóżmy, że cena za litr benzyny ma rozkład normalny $N(\mu, 1)$ o nieznanym $\mu$. Na podstawie próby $X_1, \ldots, X_{35}$ z 35 stacji benzynowych otrzymano średnią $\bar{X} = 4.8$, a następnie testem najmocniejszym testowano hipotezę $\mu = 4.5$ przeciw $\mu > 4.5$ na poziomie istotności 0.05. (i) Ile wynosi p-wartość statystyki testowej? (ii) Ile wynosi moc testu w $\mu = 5.0$? Wskazówka: tutaj moglibyśmy podać różne kwantyle i wartości dystrybuant, ale zamiast tego zachęcamy do skorzystania z obliczeń w R.

- (i) $\sim 0.96$, (ii) $\sim 0.095$

**SOL** (i) $\sim 0.038$, (ii) $\sim 0.905$

- (i) $\sim 0.038$, (ii) $\sim 0.095$
- (i) $\sim 0.96$, (ii) $\sim 0.905$

**Zadanie 5 [Autor: BM, punkty: 2, gr 1]** Do przewidywania pewnej wielkości $Y$ został zbudowany model liniowy (Model A) z macierzą planu $X$. Następnie znaleziono dodatkowe cechy objaśniające i rozszerzono macierz $X$ o dodatkowe kolumny uzyskując macierz $X'$. Dla nowej macierzy planu zbudowano drugi model liniowy (Model B). Przyjmując, że w modelu liniowym zależność jest postaci

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

gdzie $\mathbb{E}[\varepsilon] = 0$ oraz, że $\mathrm{Var}[\varepsilon] = \sigma^2 I_d$, gdzie $I_d$ to macierz identycznościowa, wskaż zdanie **nieprawdziwe**.

- Może okazać się, że nieobciążony estymator wariancji $\sigma^2$ w modelu A jest mniejszy od nieobciążonego estymatora wariancji w modelu B, albo, że nieobciążony estymator wariancji $\sigma^2$ w modelu B jest mniejszy od nieobciążonego estymatora wariancji w modelu A.

**SOL** Okazało się, że estymator wariancji $\sigma^2$ metodą największej wiarygodności w modelu A jest mniejszy od estymatora w modelu B.

- Okazało się, że $RSS = \|Y - \hat{Y}\|^2$ w modelu A jest większy od RSS w modelu B.
- Może okazać się, że wartość oczekiwana błędu predykcji $\mathbb{E}[\|Y - \hat{Y}\|^2]$ w modelu A była mniejsza niż w modelu B, albo, że wartość oczekiwana błędu predykcji w modelu B była mniejsza niż w modelu A.

**Zadanie 5 [Autor: BM, punkty: 2, gr 2, czas: 5min, Sprawdzający: ES, zadanie OK]** Do przewidywania pewnej wielkości $Y$ został zbudowany model liniowy (Model A) używający macierzy planu $X$. Następnie znaleziono dodatkowe cechy objaśniające i rozszerzono macierz $X$ o dodatkowe kolumny oraz na nowej macierzy planu zbudowano drugi model liniowy (Model B). Przyjmując, że w modelu liniowym zależność jest postaci

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

gdzie $\mathbb{E}[\varepsilon] = 0$ oraz, że $\mathrm{Var}[\varepsilon] = \sigma^2 I_d$, gdzie $I_d$ to macierz identycznościowa, wskaż zdanie **nieprawdziwe**.

**SOL** Wiadomo, że zawsze nieobciążony estymator wariancji $\sigma^2$ w modelu B jest mniejszy od nieobciążonego estymatora wariancji w modelu A.

- Okazało się, że estymator wariancji $\sigma^2$ metodą największej wiarygodności w modelu B jest mniejszy od estymatora w modelu A.
- Okazało się, że $RSS = \|Y - \hat{Y}\|^2$ w modelu A jest większy od RSS w modelu B.
- Może okazać się, że wartość oczekiwana błędu predykcji $\mathbb{E}[\|Y - \hat{Y}\|^2]$ w modelu A była mniejsza niż w modelu B, albo, że wartość oczekiwana błędu predykcji w modelu B była mniejsza niż w modelu A.

**Zadanie 6** [**Autor: DCK, punkty: 2, gr 1**] Firma marketingowa opracowuje kampanię marketingową dla producenta czekolady. Potencjalnym klientom w popularnym serwisie internetowym ma wyświetlać się reklama konkretnego typu czekolady, który z wysokim prawdopodobieństwem może ich zainteresować. Dysponując bazą danych o charakterystykach konsumentów i wybieranych przez nich czekoladach, firma buduje klasyfikator w oparciu o metodę $K$ najbliższych sąsiadów. Tabela poniżej zawiera wyznaczone odległości między punktem odpowiadającym potencjalnemu klientowi a punktem odpowiadającym danej obserwacji trenującej (identyfikowanej id). Na podstawie jedynie informacji z tabeli, reklama której czekolady powinna zostać przedstawiona potencjalnemu klientowi i na podstawie których obserwacji? Przyjmij $K = 3$.

| id | typ | odległość |
|---|---|---|
| 1 | mleczna | 0.01 |
| 2 | gorzka | 0.2 |
| 3 | deserowa | 3.0 |
| 4 | mleczna | 0.5 |
| 5 | deserowa | 9.0 |
| 6 | mleczna | 3.0 |
| 7 | gorzka | 0.1 |
| 8 | mleczna | 0.05 |
| 9 | deserowa | 1.0 |
| 10 | mleczna | 4.0 |

**SOL** mleczna, obserwacje o id: 1, 7, 8

- gorzka, obserwacje o id: 2, 7, 8
- deserowa, obserwacje o id: 1, 3, 5
- Żadne z powyższych

**Zadanie 6** [**Autor: DCK, punkty: 2, gr 2**] Firma marketingowa opracowuje kampanię marketingową dla producenta czekolady. Potencjalnym klientom w popularnym serwisie internetowym ma wyświetlać się reklama konkretnego typu czekolady, który z wysokim prawdopodobieństwem może ich zainteresować. Dysponując bazą danych o charakterystykach konsumentów i wybieranych przez nich czekoladach, firma buduje klasyfikator w oparciu o metodę $K$ najbliższych sąsiadów. Tabela poniżej zawiera wyznaczone odległości między punktem odpowiadającym potencjalnemu klientowi a punktem odpowiadającym danej obserwacji trenującej (identyfikowanej id). Na podstawie jedynie informacji z tabeli, reklama której czekolady powinna zostać przedstawiona potencjalnemu klientowi i na podstawie których obserwacji? Przyjmij $K = 3$.

| id | typ | odległość |
|---|---|---|
| 1 | mleczna | 1.0 |
| 2 | gorzka | 0.2 |
| 3 | deserowa | 3.0 |
| 4 | mleczna | 0.05 |
| 5 | deserowa | 9.0 |
| 6 | mleczna | 0.01 |
| 7 | gorzka | 0.1 |
| 8 | mleczna | 0.05 |
| 9 | deserowa | 1.0 |
| 10 | mleczna | 4.0 |

- mleczna, obserwacje o id: 1, 7, 8
- gorzka, obserwacje o id: 2, 7, 8
- deserowa, obserwacje o id: 1, 3, 5

**SOL** Żadne z powyższych