Lista 11

Zadanie 1

Wygeneruj dane:

```
n <- 100
set.seed(123)

x1 <- runif(n, 0, 10)
x2 <- sapply(x1, function(xi) rnorm(1, xi, 1))
x3 <- sapply(x1, function(xi) rnorm(1, xi, 1))
x4 <- rgamma(n, shape=2)
x6 <- rbinom(n, 10, 0.1)
x7 <- rbinom(n, 10, 0.5)
x5 <- sapply(x7, function(xi) rchisq(1, df=xi))
x8 <- sapply(x6, function(xi) rbinom(1, xi, 1))
x9 <- rbinom(n, 2, 0.5)
x10 <- rpois(n, 1/2)

epsilon <- rnorm(n, 0, 2)

X <- data.frame(x1, x2, x3, x4, x5, x6, x7, x8, x9, x10)

y <- 2*sin(x1) + exp(-x4) + 4*x10 - 3*x8 + epsilon
y <- factor(ifelse(y < 0, 0, 1))
```

Wybróbuj różne metody selekcji zmiennych. Postaraj się znaleźć taką metodę, która wybierze istotne zmienne (czyli x1, x4, x8, x10). W przypadku metod tworzących ranking zmiennych, postaraj się, by istotne zmienne zajęły cztery pierwsze miejsca w rankingu.

Zadanie 2

Zadanie 1 (spline)

Wygeneruj zbiór danych zgodnie z rozkładem

$$y = g(x) + \epsilon,$$

gdzie $g(x) = 4.26\left(e^{-x} - 4e^{-2x} + 3e^{-3x}\right)$, ϵ ma rozkład normalny z parametrami $\mu = 0$ i $\sigma = 0.1$, a x są z przedziału $[0, 3]$. Narysuj wykres `y~x`.

1. Do każdego przedziału postaci $[i, i+1]$ dla $i = 0, 1, 2$ osobno dopasuj prostą i zaznacz je na wykresie.
2. Do każdego przedziału postaci $[i, i+1]$ dla $i = 0, 1, 2$ osobno dopasuj prostą, ale tak, by otrzymana łamana była ciągła i zaznacz ją na wykresie. By uzyskać ciągłość, rozszerz macierz eksperymentu o zmienne
$$(x-1)_+ \text{ i } (x-2)_+.$$

Porównaj z krzywą powstałą przez dopasowanie `y~bs(x, degree=1, knots = c(1, 2))` (funkcja `bs` tworzy odpowiednią macierz eksperymentu; pakiet `splines`). i `y~bs(x, degree=1, df=3)`. Jaki jest związek pomiędzy liczbą stopni swobody a liczbą węzłów przy ustalonym stopniu dopasowywanego wielomianu?

3. Do każdego przedziału postaci $[i, i+1]$ dla $i = 0, 1, 2$ dopasuj **spline kubiczny** (czyli stopień dopasowywanego wielomianu ma wynosić 3). Funkcję zapisaną w bazie dla spline'ów kubicznych można przedstawić w poniższy sposób:

$$f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \theta_k (x - \zeta_k)_+^3.$$

4. **Naturalny spline kubiczny**, to taki spline kubiczny, który poza węzłami granicznymi z prawej i lewej strony dopasowuje funkcję liniową (ma to zapobiec niepoprawnemu dopasowywaniu się funkcji na brzegach). Dopasuj do danych naturalny spline kubiczny (funkcja `ns`) z domyślnymi i podanymi węzłami brzegowymi (parametr `Boundary.knots`).

5. **Smooth spline** polega na tym, że używamy jak największej liczby węzłów, ale dodajemy karę za brak gładkości, tzn. szukamy wśród funkcji $f \in C^2$ takiej, która będzie minimalizować błąd

$$RSS(f, \lambda) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt,$$

gdzie λ jest parametrem wygładzającym. Dopasuj smooth spline do danych (funkcja `smooth.spline`). Przetestuj różne wartości parametrów `lambda` i `df`.