Regresja

- Regresją nazywamy związek między wartością oczekiwaną zmiennej zależnej i wielkościami zmiennych niezależnych.

- W sensie matematycznym regresja jest warunkową wartością oczekiwaną

$$\mathrm{E}\left(y_i \middle| x_{1i}, \ldots, x_{ki}\right) = m\left(x_{1i}, \ldots, x_{ki}\right)$$

- Regresja liniowa:

$$m\left(x_{1i}, \ldots, x_{ki}\right) = \boldsymbol{\beta}_1 x_{1i} + \ldots + \boldsymbol{\beta}_k x_{ki}$$

- Regresji nieliniowa:

$$m\left(x_{1i},\ldots,x_{ki}\right)=m\left(x_{1i},\ldots,x_{ki},\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_k\right)$$

- W przypadku regresji nieparametrycznej nie zakładamy konkretnej formy $m\left(x_{1i},\ldots,x_{ki}\right)$. Przjmuje się jednak w tym przypadku pewne założenia dotyczące gładkości funkcji $m\left(x_{1i},\ldots,x_{ki}\right)$.

Estymatory nieparametryczne funkcji gęstości

- Zdefiniujmy jądro (*kernel*) jako funkcję $K(u)$ o następujących własnościach:
  - ciągła, ograniczona, symetryczna
  - $\int K(u)\,du = 1$

- Najczęsciej stosowane jądra:

Jednostajne	$\frac{1}{2}\mathbb{I}(\lvert u\rvert \leq 1)$
Trójkątne	$(1-\lvert u\rvert)\,\mathbb{I}(\lvert u \leq 1\rvert)$
Epanechnikov	$\frac{3}{4}\left(1-u^2\right)\mathbb{I}(\lvert u\rvert \leq 1)$
Gaussssian	$\frac{1}{\sqrt{2\boldsymbol{\pi}}}\exp\left(-\frac{1}{2}u^2\right)$

- Estymator jądrowy Rosenblatta-Parzena funkcji gęstości:

$$\widehat{f}_{h_n}(x) = n^{-1} \sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{x - X_i}{h_n}\right)$$

- h_n - jest oknem (*bandwidth*).
  - szerokość okna determinuje stopień wygładzenia funkcji gęstości
  - szersze okno zmniejsza wariancję estymatora ale zwiększa obciążenie
  - okno wybieramy często arbitralnie na podstawie rysunku
  - istnieją także wzory na optymalną szerokość okna

- Uogólnienie na kilka wymiarów

$$\widehat{f}_{h_{1n},\ldots,h_{kn}}(x_1, \ldots, x_K) = n^{-1} \sum_{i=1}^{n} \prod_{k=1}^{K} \frac{1}{h_{kn}} K\left(\frac{x_k - X_{ki}}{h_{kn}}\right)$$

Regresja nieparametryczna, estymator jądrowy

- Dla przypadku regresji dla jednej zmiennej

$$m(x) = \mathrm{E}(y|\, x) = \int y f(y|\, x)\, dy = \frac{1}{f(x)} \int y f(y, x)\, dy$$

- Oszacowanie $f(y, x)$ jest dane wzorem

$$\widehat{f}_{h_n, g_n}(y, x) = n^{-1} \sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{x - X_i}{h_n}\right) \frac{1}{g_n} K\left(\frac{y - Y_i}{g_n}\right)$$

- Przybliżony wzór na

$$\int yf\left(y,x\right)dy \approx \sum_{i=1}^{n} \widehat{f}_{h_n,g_n}\left(Y_i,x\right)Y_i = n^{-1}\sum_{i=1}^{n}\frac{1}{h_n}K\left(\frac{x-X_i}{h_n}\right)Y_i$$

ma tą samą wartość dla wszystkich elementow sumy i możem przyjąć, że $\frac{1}{g_n}K\left(0\right)=1$.

- Estymatorem jądrowy Nadaraya-Watsona funkcji regresji ma więc postać

$$\begin{aligned} m_h\left(x\right) &= \frac{n^{-1}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)Y_i}{n^{-1}\sum_{i=1}^{n}\frac{1}{h_n}K\left(\frac{x-X_i}{h_n}\right)} \\ &= \sum_{i=1}^{n} W_{ih_n}Y_i \end{aligned}$$

gdzie W_{ih_n} są wagami postaci

$$W_{ih} = \frac{\frac{1}{h}K\left(\frac{x-X_i}{h}\right)Y_i}{\sum_{i=1}^{n}\frac{1}{h_n}K\left(\frac{x-X_i}{h_n}\right)}$$

- Uogólnienia wielowymiarowe opiera się na wielowymiarowym opiera się na wielowymiarowym uogólnieniu estymatora funkcji gęstości

Lokalna regresja wielomianowa

- Wiemy, że wystarczajaco gładką funkcję można przybliżyć wokół t za pomocą następującego wzoru Taylora

$$m(t) \approx m(x) + m'(x)(t-x) + \ldots + m^{(p)}(x)(t-x)^p \frac{1}{p!}$$

- Lokalna regresja wilomianowa polega na przeprowadzeniu ważonej regresji następującej postaci:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ Y_i - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_1 (X_i - x) - \ldots - \boldsymbol{\beta}_p (X_i - x)^p \right\} \frac{1}{h} K\left(\frac{x - X_i}{h} \right)$$

- Wagi nadawane poszczególnym obserwacjom związane są z ich odległością od szacowanego punktu

-

Przekleństwo wymiaru

- Zarówno funkcja gęstości jak funkcja regresji przy estymacji metodami nieparametrycznymi jest szacowana na podstawie wielkości sąsiednich obserwacji

- Dla dużej liczby wymiarów precyzja oszacowań szybko spada, ponieważ liczba obserwacji, która sąsiaduje z dowolnie dobranym punktem się szybko zmniejsza wraz ze wzrostem liczby wymiarów

- Dodatkowy problem:

  – jak zdefiniować sąsiedztwo w wielu wymarach w sytuacji, w której poszczególne zmienne mają różne jednostki

- Rozwiązaniem jest przyjęcie mniej ogólnej formy funkcji gęstości
- Jednym z proponowanych rozwiązań jest przyjęcie formy addytywnej

$$m\left(x_1, \ldots, x_k\right) = a + f_1\left(x_1\right) + \ldots + f_K\left(x_k\right)$$

- Można do takiego modelu dodać nieparametryczne interakcje między zmiennymi

$$m\left(x_1, \ldots, x_k\right) = a + f_1\left(x_1\right) + f_2\left(x_k\right) + f_{12}\left(x_1, x_2\right)$$

- Możliwe jest także częściowe sparametryzowanie modelu

$$m\left(x_1, \ldots, x_k\right) = a + \boldsymbol{\beta}_1 x_1 + f_2\left(x_2\right) \ldots + f_K\left(x_k\right)$$