Лабораторна робота №5
з непараметричної статистики

Горбунов Даніел Денисович
1 курс магістратури
група "Прикладна та теоретична статистика"
Варіант №4

7 травня 2022 р.

Вступ.

У даній роботі розглядалася задача оцінювання невідомої функції зв'язку в структурній моделі регресії:

$$Y_j = g(X_j) + \varepsilon_j,\ j = \overline{1,n},\ n = 300$$

де $\{X_j\}_{j=1}^n$ – н.о.р., $X_1 \sim N(0,1)$, $\{\varepsilon_j\}_{j=1}^n$ – н.о.р., $\varepsilon_1 \sim N(0,0.5)$, $\{X_j\}_{j=1}^n$ та $\{\varepsilon_j\}_{j=1}^n$ є незалежними в сукупності. Невідомою функцією зв'язку є $g(x) = 2|x|$. Для оцінювання $g(x)$ використано три оцінки: оцінки ковзаючого середнього та медіани та локально-лінійна регресія з ядром Єпанєчнікова. Досліджено поведінку оцінок у разі додавання до вибірки викидів. Параметри згладжування у запропонованих оцінках підібрані на око.

Хід роботи.

Моделювання даних.

Надалі будемо працювати лише з однією вибіркою, згенерованою з зернини 0:

```
# Фіксуємо зернину
set.seed(0)
# Генеруємо вибірку, користуючих початковими відомостями
n <- 300
# Регресор
x <- rnorm(n, 0, 1)
# Похибка
e <- rnorm(n, 0, sqrt(0.5))
# Невідома функція зв'язку
g <- function(t) { 2 * abs(t) }
# Відгук
y <- g(x) + e
```

Зобразимо дані на діаграмі розсіювання:

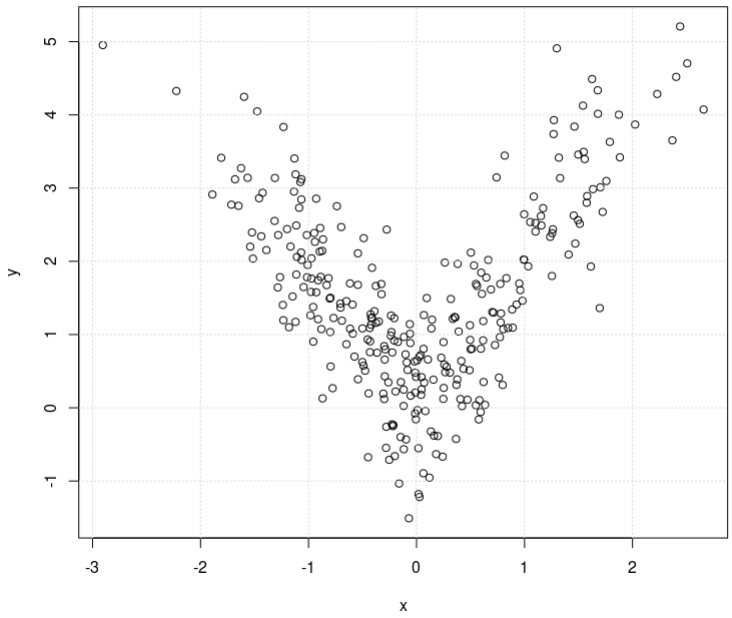

Рис. 1: Діаграма розсіювання початкових даних.

Можливо будуть деякі проблеми на кінцях та всередині тренду: перша пояснюється тим, що на кінцях густина спостережень зовсім мала (може призвести до крайового ефекту для ковзаючих оцінок), а щодо другого, то імітувати гострий кут наврядчи добре вийде.

Ковзаюче середнє: формула, реалізація, застосування.

Оцінка ковзаючого середнього із заданою шириною вікна $h > 0$ у точці x дорівнює

$$\hat{g}^{ma}(x) = \frac{\sum_{j=1}^{n} Y_j \mathbb{1}\{|x - X_j| < h/2\}}{\sum_{j=1}^{n} \mathbb{1}\{|x - X_j| < h/2\}}$$

Програмна реалізація підрахунку оцінки така:

```
# Повертає функцію, що рахує значення оцінки КС у заданій точці t.
# Аргументи: x - регресор, y - відгук, h - ширина вікна
moving.average <- function(x, y, h)
{
  # Опишемо функцію для однієї точки t
  ma.univar <- function(t)
  {
    ones <- (abs(t - x) < h / 2)
    sum(y * ones) / sum(ones)
  }
  # Векторизуємо попередню функцію
  function(t) sapply(t, ma.univar)
}
```

Тут і для інших оцінок функції зв'язку, параметр згладжування $h > 0$ обираємо на око з інтервалу $(0, 1)$.

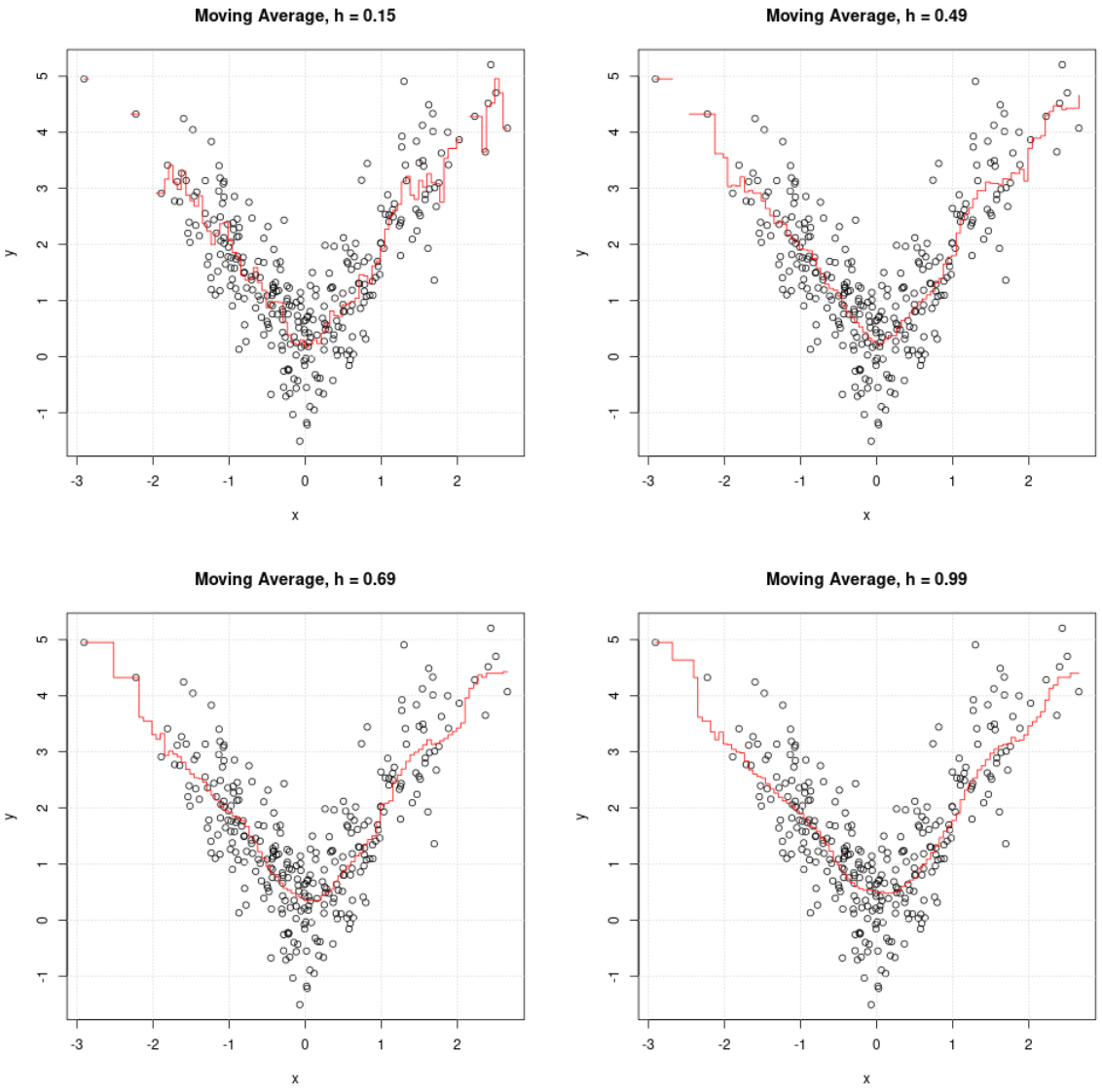

Рис. 2: Різні графіки ковзаючого середнього, в залежності від h.

Ефект перезгладжування (коли графік майже проходить через кожне спостереження) спостерігається для h приблизно менших за 0.4. Добре наближення вже починається при h приблизно більших за 0.4 і приблизно менших за 0.7. Для достатньо малих значень h спостерігається раніше згаданий крайових ефект на кінцях тренду: внаслідок браку потрібних спостережень, зліва оцінка стає невизначеною, а справа спостерігаємо сильні коливання. Для того, щоб оцінка була визначеню всюди (що буде стосуватися й аналогічних оцінок), ширина вікна має буде більшою за $\min\limits_{1\le j\le n-1}(X_{(j+1)} - X_{(j)}) \approx 0.6809988$. Розв'язавши одну проблему, з'являється інша: для таких h вже відчувається, що оцінка стає занадто гладкою: "нижній кут" стає менш гострим, ніж яким має бути насправді. Зокрема можна спостерігати як при збільшенні ширини вікна графік вироджується до вибіркового середнього по всій вибірці. Отже на роль "хороших" значень, що визначають ширину вікна h, можна взяти два наступні:

1. Непогана імітація функції зв'язку, але не визначена всюди: $h_1 := 0.49$,
2. Трохи перезгладжена оцінка, але повністю визначена: $h_2 := 0.69$.

Надалі будемо використовувати h_2.

Ковзаюча медіана: формула, реалізація, застосування.

Оцінка ковзаючої медіани із заданою шириною вікна $h > 0$ у точці x дорівнює

$$\hat{g}^{mm}(x) = \text{med}\{Y_j \mid j = \overline{1,n} : \ |x - X_j| < h/2\}$$

Програмна реалізація підрахунку оцінки така:

```
moving.median <- function(x, y, h)
{
  mm.univar <- function(t) { median(y[abs(t - x) < h / 2]) }
  function(t) sapply(t, mm.univar)
}
```

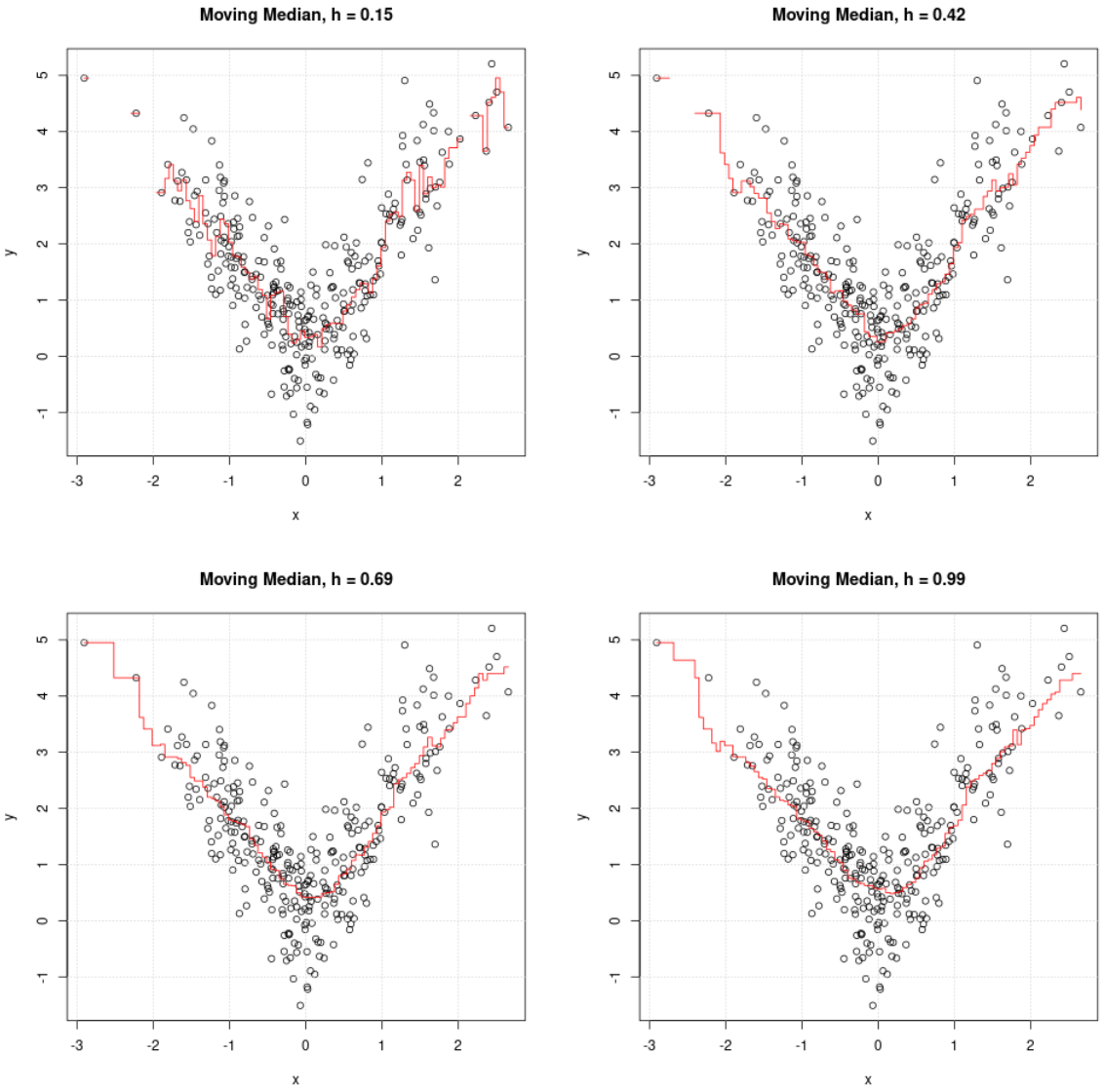

Рис. 3: Різні графіки ковзаючої медіани, в залежності від h.

Розподіл похибок є симетричним, а тому можна очікувати, що ковзаюча медіана буде давати майже такі ж результати, які можна було б отримати для попередньої оцінки (поки не закинути викиди). Можна взяти в якості оптимальних на око значення параметрів з попереднього випадку, однак тут в якості h_1 покладемо 0.42.

Локально-лінійна регресія: формула, реалізація, застосування.

Ідея полягає в тому, щоб застосувати лінійне набляження у заданій точці x_0 (у певному сенсі схоже на лінеаризацію Тейлора):

$$g(x_0) \approx b_0 + b_1(x - x_0)$$

Підгонку параметрів $b_j = b_j(x_0)$ робимо за допомогою зваженого МНК із ваговими коефіцієнтами вигляду:

$$w_j(x_0) = K\left(\frac{x_0 - X_j}{h}\right),\ j = \overline{1,n}$$

Де K будемо називати локалізуючим ядром, яке є парним та затухає за межами $[-1, 1]$. Тут $h > 0$ відіграє роль половини ширини інтервалу, який і задає локальність. Оскільки підхід дещо нагадує лінеаризацію Тейлора для гладких функцій, то із нашою функцією зв'язку, здається, можуть виникнути проблеми в нулі.

Програмна реалізація підрахунку оцінки така:

```
loc.lin.regr <- function(x, y, h, K, corr.w = 1)
{
  llr.univar <- function(x0)
  {
    # Обчислюємо вагові коефіцієнти для кожного спостереження
    w <- K((x0 - x) / h)
    # Робимо підгонку параметрів за допомогою зваженого МНК
    loc.lm <- lm(y ~ I(x - x0), weights = corr.w * w)
    # Прогнозом функції зв'язку буде оцінка зсуву у лінійній моделі
    coef(loc.lm)[1]
  }
  function(t) sapply(t, llr.univar)
}
```

Пропонується використовувати ядро Єпанєчнікова: $K(t) = 0.75 \cdot (1 - t^2) \cdot \mathbb{1}\{|t| < 1\}$

```
K.e <- function(t)
{
  0.75 * (1 - t^2) * (abs(t) < 1)
}
```

Нижче покажемо графіки прогнозу для деяких h.

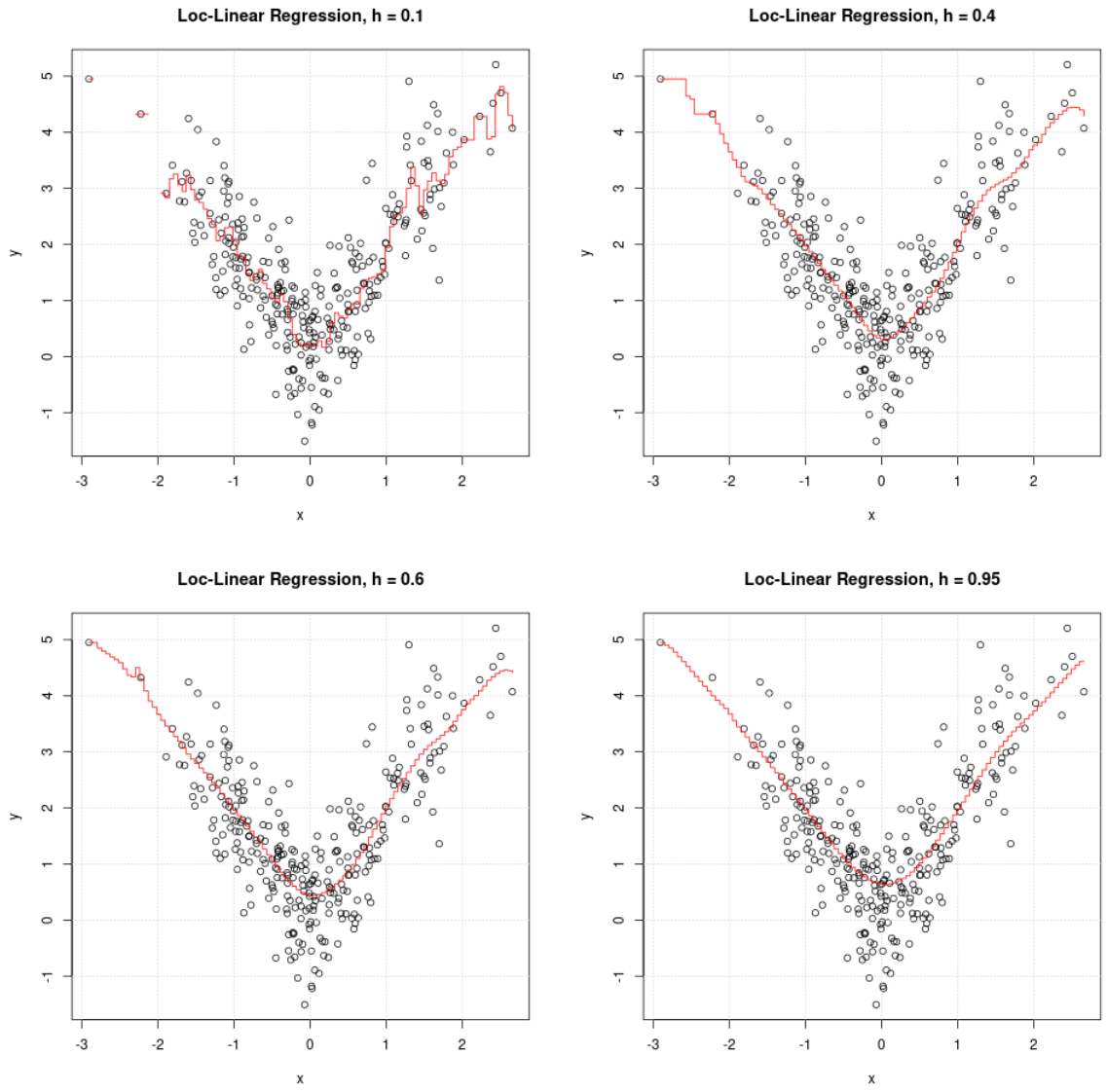

Рис. 4: Різні графіки прогнозу локально-лінійної регресії, в залежності від h.

Видно, що для $0.4 \leq h \leq 0.6$ (приблизні межі) маємо відносно непогані наближення невідомої функції зв'язку. Якщо при $h = 0.4$ ширина виявилася замалою, то для $h = 0.6$ поведінка стає більш-менш злагодженою. Для більших або менших значеннях h маємо перезгладужвання або недозгладжування відповідно. Тому оптимальним на око в цьому випадку зафіксуємо $h^* = 0.6$.

Залучення викидів.

До початкових даних додамо по п'ять пар спостережень, які зовсім вибиваються з тренду:

```
x <- c(x, rep(c(-2, 0, 2), 5))
y <- c(y, rep(c(10, -5, 10), 5))
```

На діаграмі розсіювання це виглядатиме так:

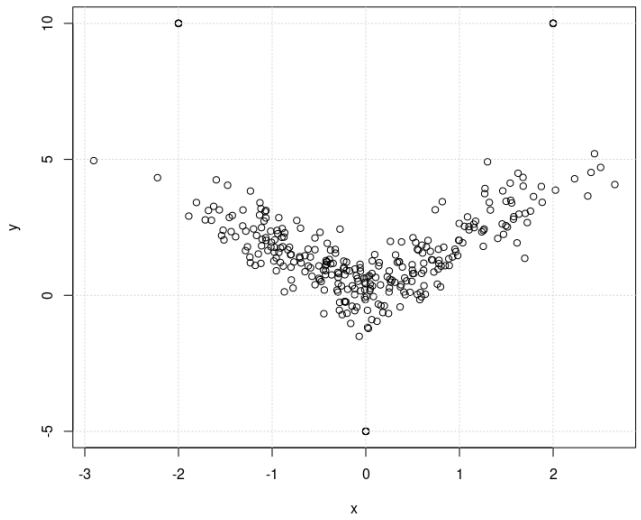

Рис. 5: Діаграма розсіювання початкових даних з викидами.

Лівий верхній викид хороший тим, що навіть ковзаюча медіана його не зможе оминути внаслідок низької густини значень по X_j зліва. Для оцінок з підібраними на око параметрами згладжування покажемо як сильно змінилися графіки:

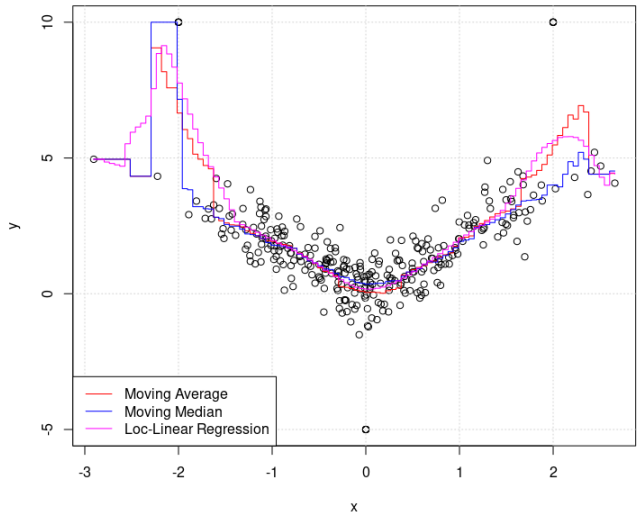

Рис. 6: Поведінка різних оцінок на даних з викидами.

Коротко кажучи, графіки змінилися в гіршу сторону. Викиди на кінцях витягують всі оцінки догори. Центральний викид в околі нуля трохи зсунув значення ковзаючого середнього і прогнозу локально-лінійної регресії вниз. Але в центрі ковзаюча медіана не зсунулася (дякувати за це значному скупченню різноманітних значень X_j поблизу).

Поправка на викиди у локально-лінійній регресії.

Застосуємо підхід, який використовується для Lowess, для виправлення неадекватної поведінки прогнозу локально-лінійної регресії внаслідок викидів.

Процесс описується у декілька кроків:

1. Обчислюємо залишки прогнозу $U_j = Y_j - \hat{g}_j^{ll}\text{m}\ j = \overline{1,n}$.

2. Обчислюємо коефіцієнт "нехтування" спостереженням, використовуючи біквадратне ядро:

$$\delta_j = B(U_j/(6 \cdot \mu)),\ \mu = \text{med}\{|U_j|\},\ B(t) = (1-t^2)^2 \cdot \mathbb{1}\{|t|<1\},\ j = \overline{1,n}$$

3. Робимо повторну підгонку параметрів у локальній регресії з новими ваговими коефіцієнтами: $w_j'(x_0) = \delta_j \cdot w_j(x_0)$.

4. Повторюємо кроки (1)-(3), поки не стано ясно, що аномалії усунуті.

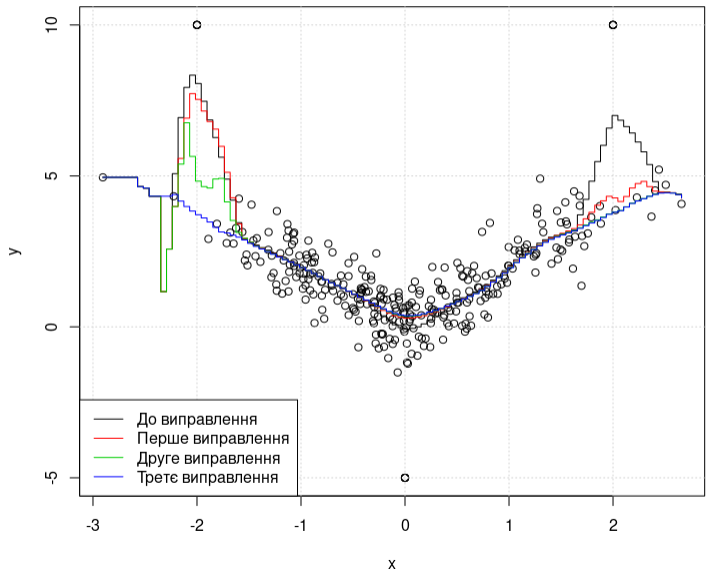

Рис. 7: Поведінка прогнозу локальної регресії від кількості поправок.

З рисунку видно, що вже на третьому кроці поведінка прогнозу стає адекватною. Гнучкість локально-лінійної моделі врятувала ситуацію з викидами, чого не дозволяє ковзаюча медіана.

Висновки.

У цьому "іграшковому" прикладі краще не морочити голову над застосуванням непараметричної регресії і правильно застосувати класичну лінійну регресії, бо з діаграми розсіювання неважко побачити, що невідома функція зв'язку має лінійний характер та можна прикинути точку зміни тренду. З іншого боку, цей штучний приклад дає зрозуміти з якими проблемами доводиться спіткатися при використанні непараметричної регресії та способами їх можливого усунення. Досить непогано відтворює функцію зв'язку оцінка ковзацючої медіани: на відміну від ковзаючого середнього та локально-лінійної регресії, в околі нуля краще імітується "гострий кут" модуля. Незважаючи на те, що виходить в околі нуля для прогнозу локально-лінійної регресії, оцінка хороша тим, що не допускає нестабільної поведінки прогнозу на кінцях. Ковзаюча медіана також хороша стійкістю до забруднень даних, хоча для локально-лінійної регресії можна зробити адаптацію вагових коефіцієнтів для усунення збурень прогнозу.

Слід зазначити, що застосування технік непараметричної регресії відчувається виснажливим в часовому та обчислювальному розумінні, поки в параметричній регресії все обчислюється досить швидко.