Моделі статистичного навчання: нелінійні моделі

Поліноміальна регресія розширює лінійну модель, додаючи додаткові предиктори, отримані шляхом піднесення кожного з початкових предикторів до певного степеня. Наприклад, кубічна регресія використовує три змінні, X, X^2 та X^3, як предиктори. Цей підхід забезпечує простий спосіб використати нелінійні моделі.

Поліноміальна регресія розширює лінійну модель, додаючи додаткові предиктори, отримані шляхом піднесення кожного з початкових предикторів до певного степеня. Наприклад, кубічна регресія використовує три змінні, X, X^2 та X^3, як предиктори. Цей підхід забезпечує простий спосіб використати нелінійні моделі.

Східчасті функції ділять діапазон змінної на K різних областей для отримання якісної змінної. Це має ефект припасування кусково-постійної функції.

Регресійні сплайни є більш гнучкими, ніж поліноми та східчасті функції, і є їхнім узагальненням.

Регресійні сплайни є більш гнучкими, ніж поліноми та східчасті функції, і є їхнім узагальненням. Вони ділять діапазон X на K різних інтервалів. У кожному інтервалі поліноміальна функція припасовується до даних.

Регресійні сплайни є більш гнучкими, ніж поліноми та східчасті функції, і є їхнім узагальненням. Вони ділять діапазон X на K різних інтервалів. У кожному інтервалі поліноміальна функція припасовується до даних. Ці поліноми будуються так, що вони плавно з'єднуються на межах інтервалів, або вузлів. За умови, що вся множина розділена на достатню кількість інтервалів, це може забезпечити надзвичайно точну оцінку.

Регресійні сплайни є більш гнучкими, ніж поліноми та східчасті функції, і є їхнім узагальненням. Вони ділять діапазон X на K різних інтервалів. У кожному інтервалі поліноміальна функція припасовується до даних. Ці поліноми будуються так, що вони плавно з'єднуються на межах інтервалів, або вузлів. За умови, що вся множина розділена на достатню кількість інтервалів, це може забезпечити надзвичайно точну оцінку.

Згладжувальні сплайни схожі на регресійні, але виникають у дещо інших ситуаціях. Згладжування сплайнів є результатом мінімізації суми квадратів залишків.

Локальна регресія подібна до сплайнів, але є суттєва різниця. Інтервали можуть перекриватися, і насправді вони роблять це дуже плавно.

Локальна регресія подібна до сплайнів, але є суттєва різниця. Інтервали можуть перекриватися, і насправді вони роблять це дуже плавно.

Узагальнені адитивні моделі дозволяють розширити вищезазначені методи на випадок з кількома предикторами.

Поліноміальна регресія.

Розширює стандартну лінійну регресію

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

до поліноміальної функції

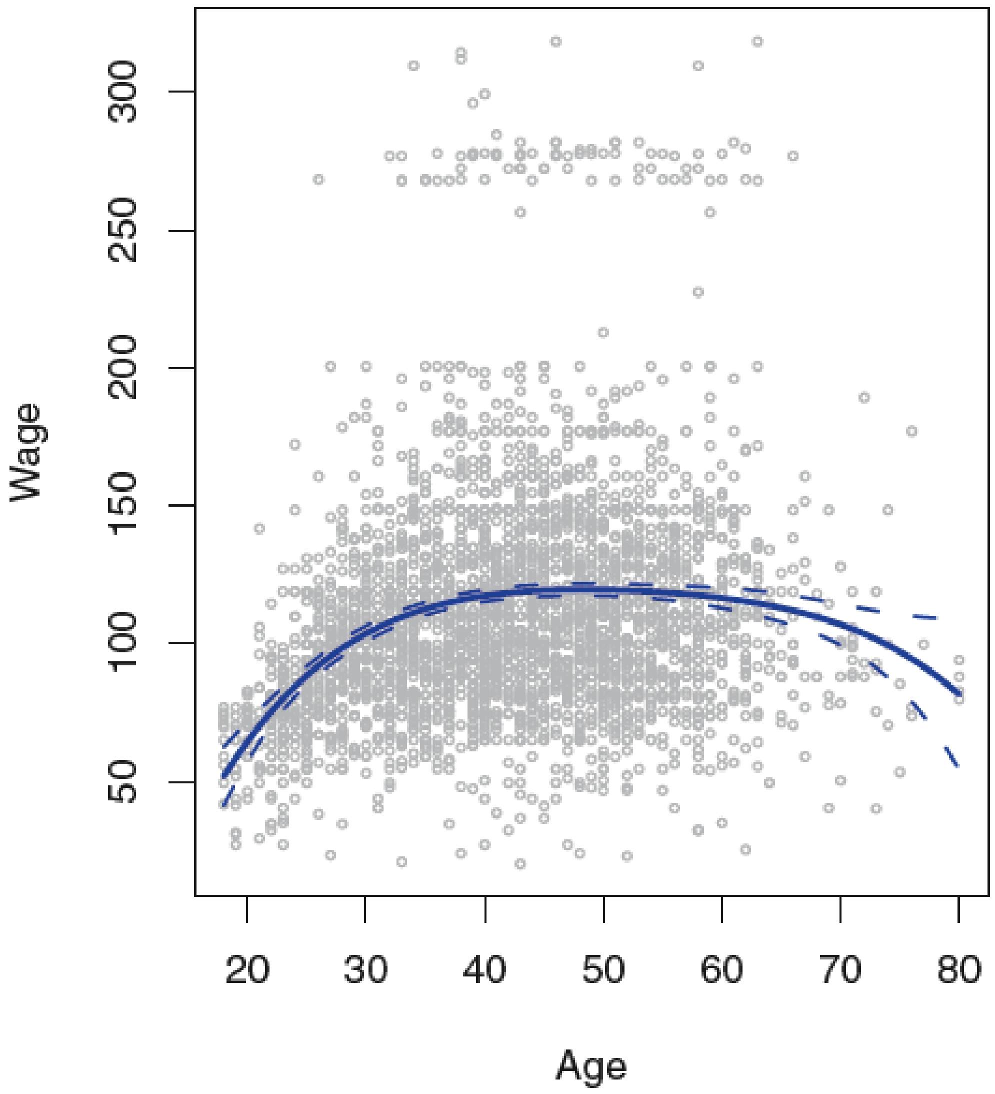
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \varepsilon_i$$

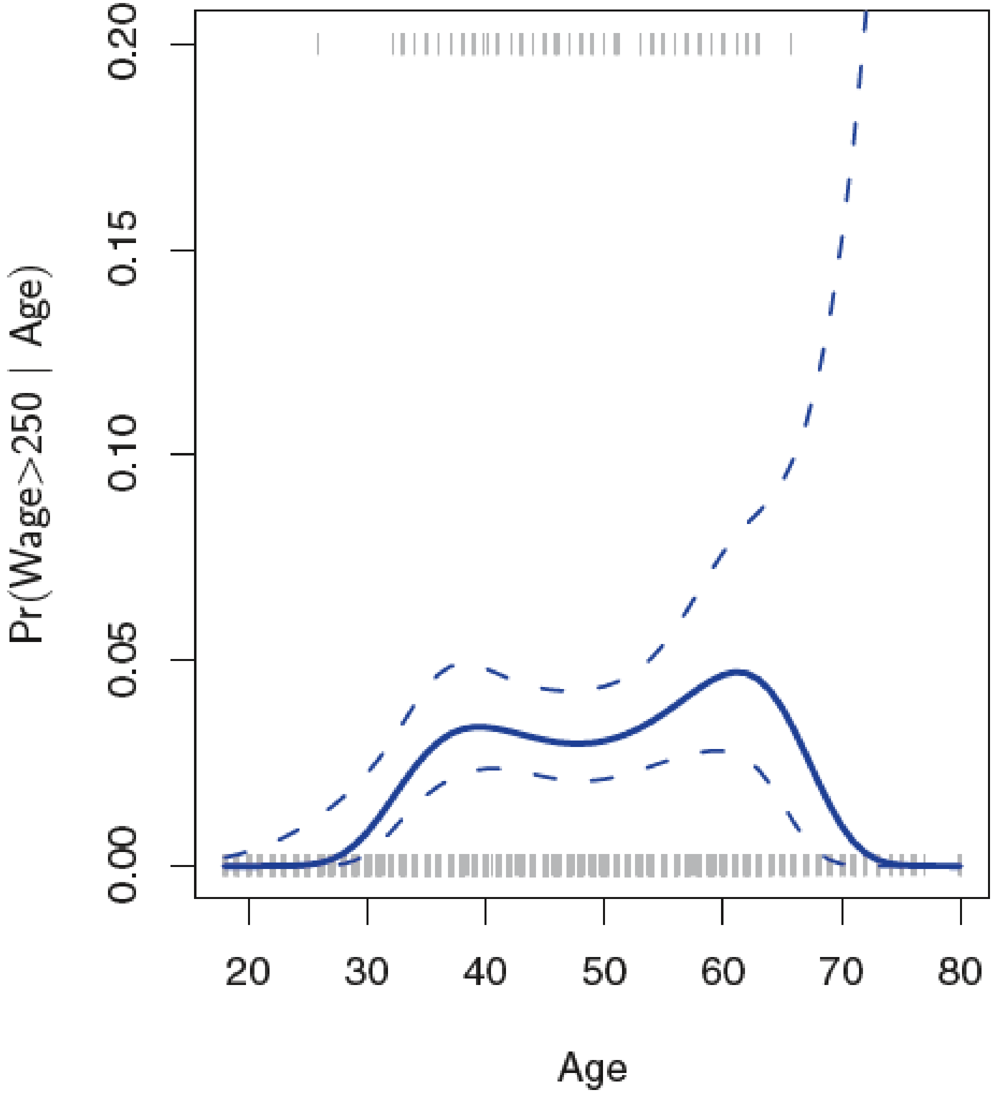
$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4$$

$$\Pr(y_i > 250 | x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}$$

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4$$

$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}$$





Східчасті функції.

$$\begin{array}{lcl} C_0(X) & = & I(X < c_1), \\ C_1(X) & = & I(c_1 \leq X < c_2), \\ C_2(X) & = & I(c_2 \leq X < c_3), \\ & \vdots & \\ C_{K-1}(X) & = & I(c_{K-1} \leq X < c_K), \\ C_K(X) & = & I(c_K \leq X), \end{array}$$

Східчасті функції.

$$
\begin{array}{lcl}
C_0(X) & = & I(X < c_1), \\
C_1(X) & = & I(c_1 \leq X < c_2), \\
C_2(X) & = & I(c_2 \leq X < c_3), \\
 & \vdots & \\
C_{K-1}(X) & = & I(c_{K-1} \leq X < c_K), \\
C_K(X) & = & I(c_K \leq X),
\end{array}
$$

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \epsilon_i$$

$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}{1 + \exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}$$

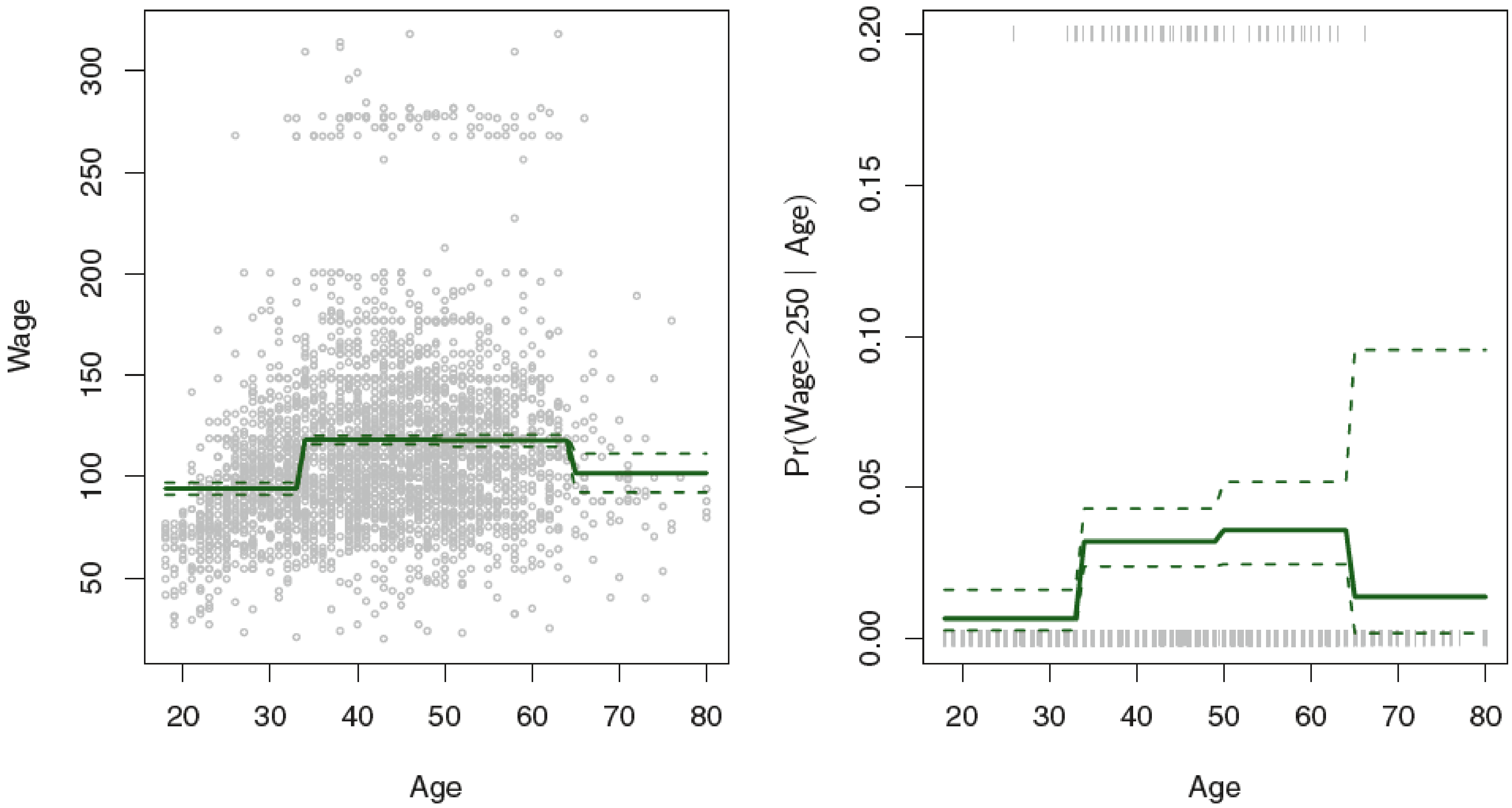
Східчасті функції.

$$
\begin{array}{lcl}
C_0(X) & = & I(X < c_1), \\
C_1(X) & = & I(c_1 \leq X < c_2), \\
C_2(X) & = & I(c_2 \leq X < c_3), \\
 & \vdots & \\
C_{K-1}(X) & = & I(c_{K-1} \leq X < c_K), \\
C_K(X) & = & I(c_K \leq X),
\end{array}
$$

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \epsilon_i$$

$$\Pr(y_i > 250 | x_i) = \frac{\exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}{1 + \exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}$$

Базові функції: $y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \ldots + \beta_K b_K(x_i) + \epsilon_i$



Wage
50
100
150
200
250
300
20
30
40
50
60
70
80
Age
Pr(Wage>250 | Age)
0.00
0.05
0.10
0.15
0.20
20
30
40
50
60
70
80
Age

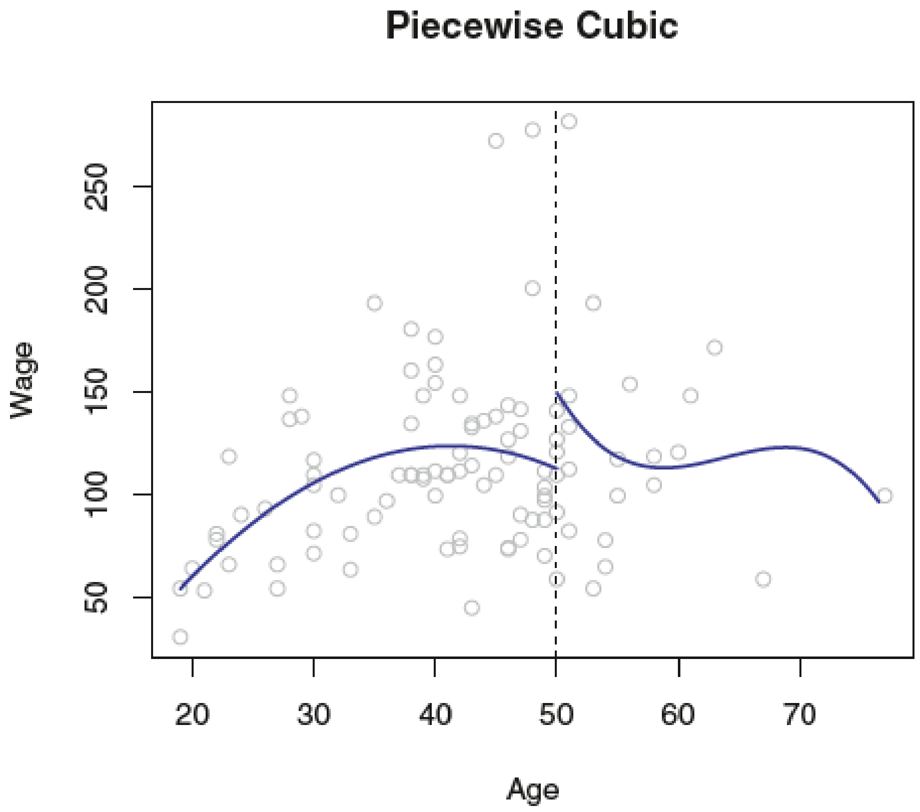
Регресійні сплайни.

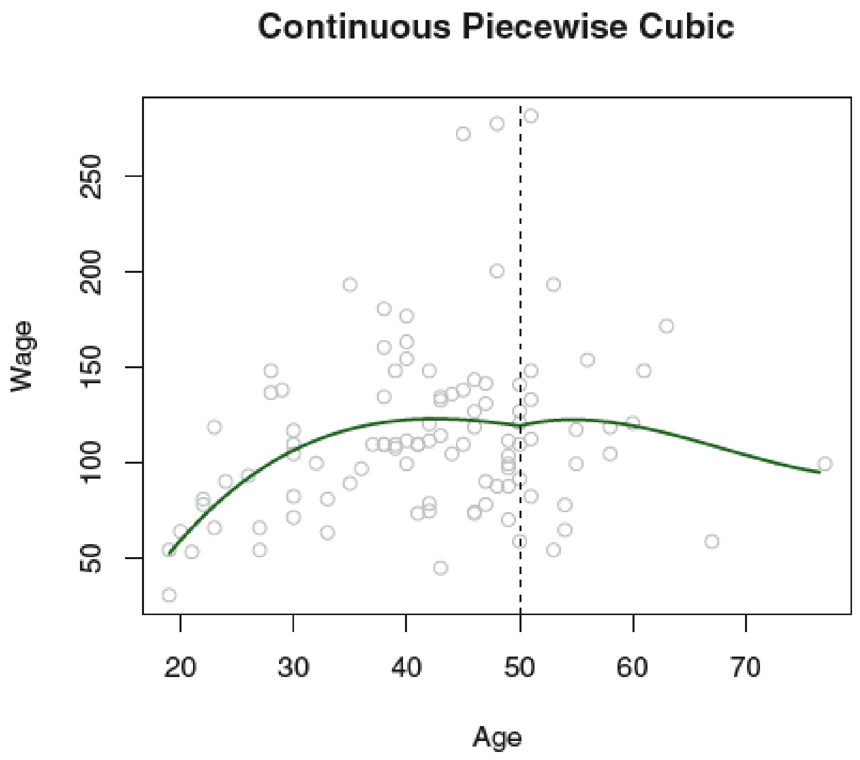
Регресійні сплайни.

$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

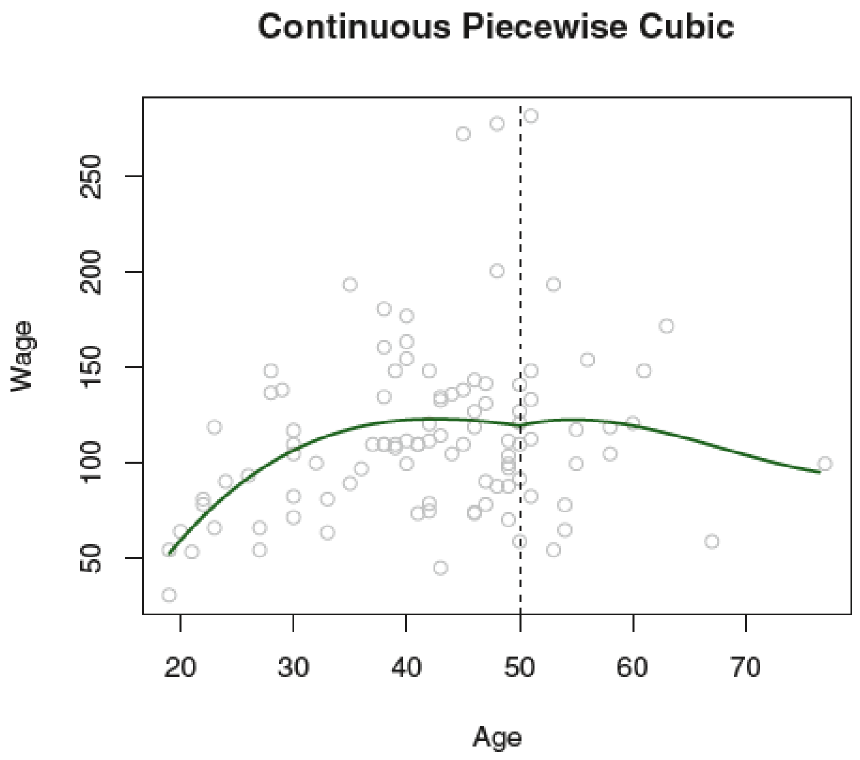
Регресійні сплайни.

$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

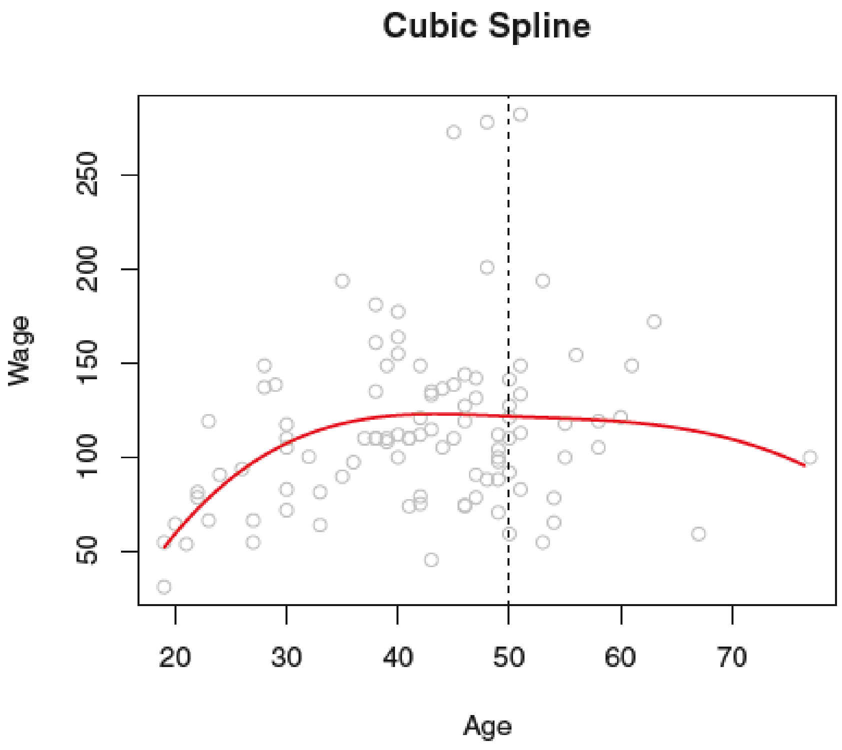




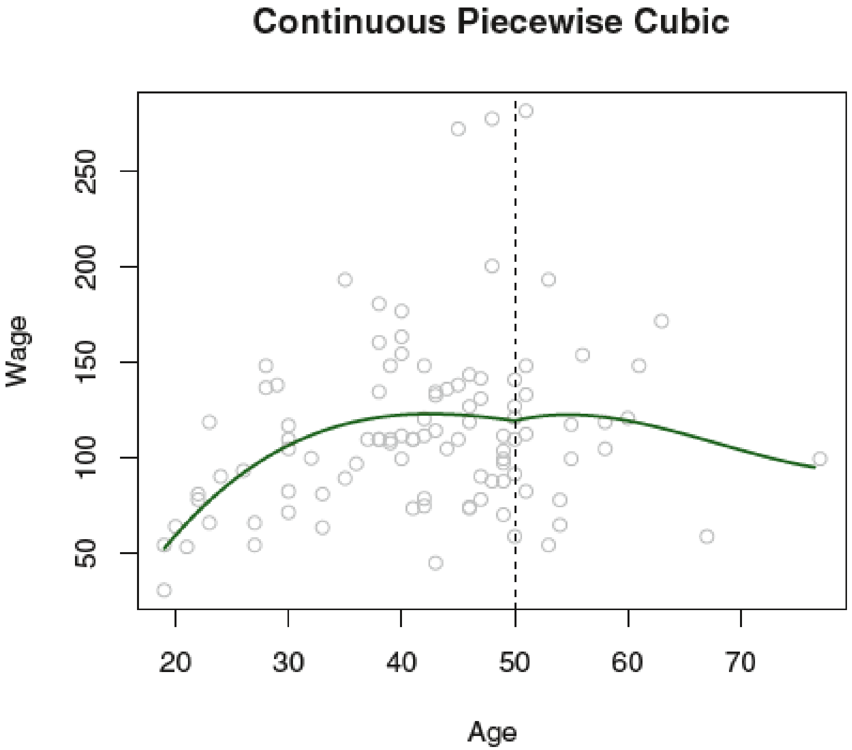
Continuous Piecewise Cubic
Wage
Age
250
200
150
100
50
20
30
40
50
60
70

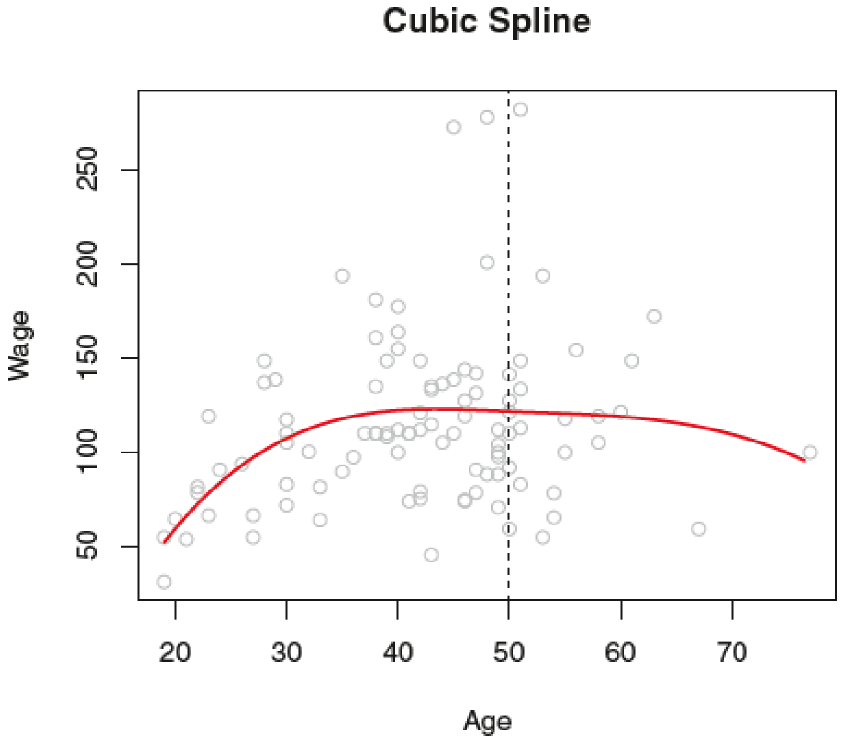


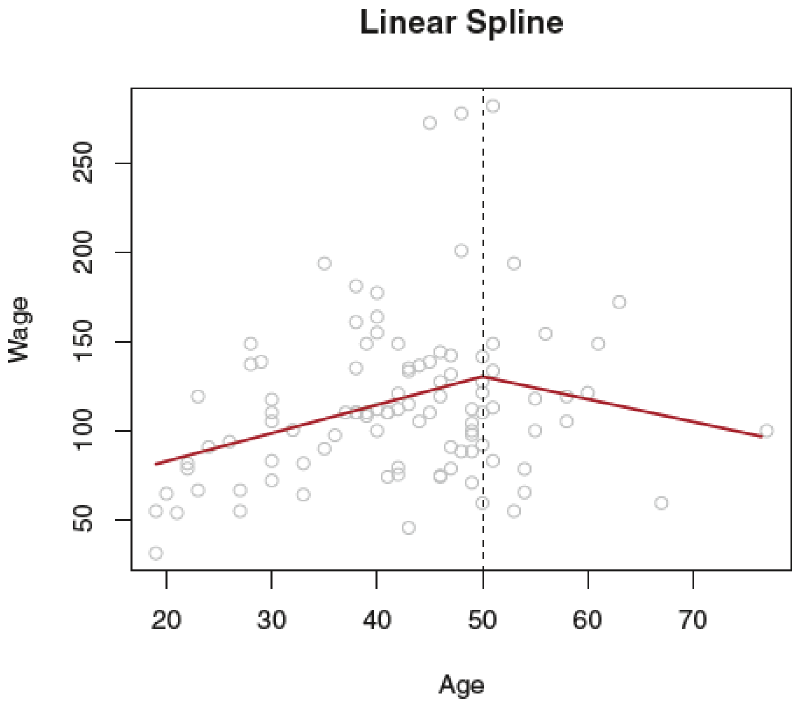
Continuous Piecewise Cubic
Wage
Age



Cubic Spline
Wage
Age


Continuous Piecewise Cubic
Wage
Age
50
100
150
200
250
20
30
40
50
60
70


Cubic Spline
Wage
Age
50
100
150
200
250
20
30
40
50
60
70


Linear Spline
Wage
Age
50
100
150
200
250
20
30
40
50
60
70

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

$$h(x, \xi) = (x - \xi)_+^3 = \begin{cases} (x - \xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise,} \end{cases}$$

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

$$h(x,\xi) = (x-\xi)_+^3 = \begin{cases} (x-\xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise,} \end{cases}$$

$$f(X) = \beta_1 N_1(X) + \beta_2 N_2(X) + \sum_{k=1}^{K-2} (\xi_K - \xi_k)\theta_k N_{k+2}(X)$$

$$N_1(X) = 1, \qquad N_2(X) = X, \qquad N_{k+2}(X) = d_k(X) - d_{K-1}(X),$$

$$d_k(X) = \frac{(X-\xi_k)_+^3 - (X-\xi_K)_+^3}{\xi_K - \xi_k}$$

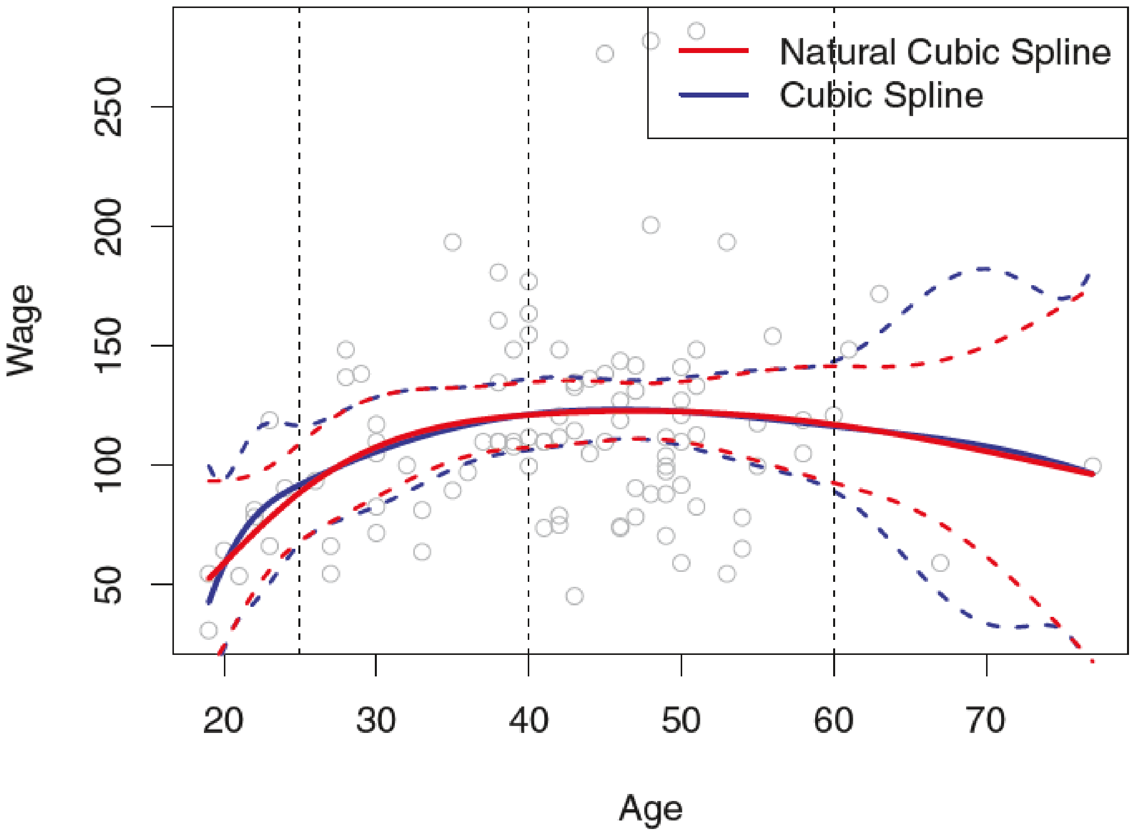
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

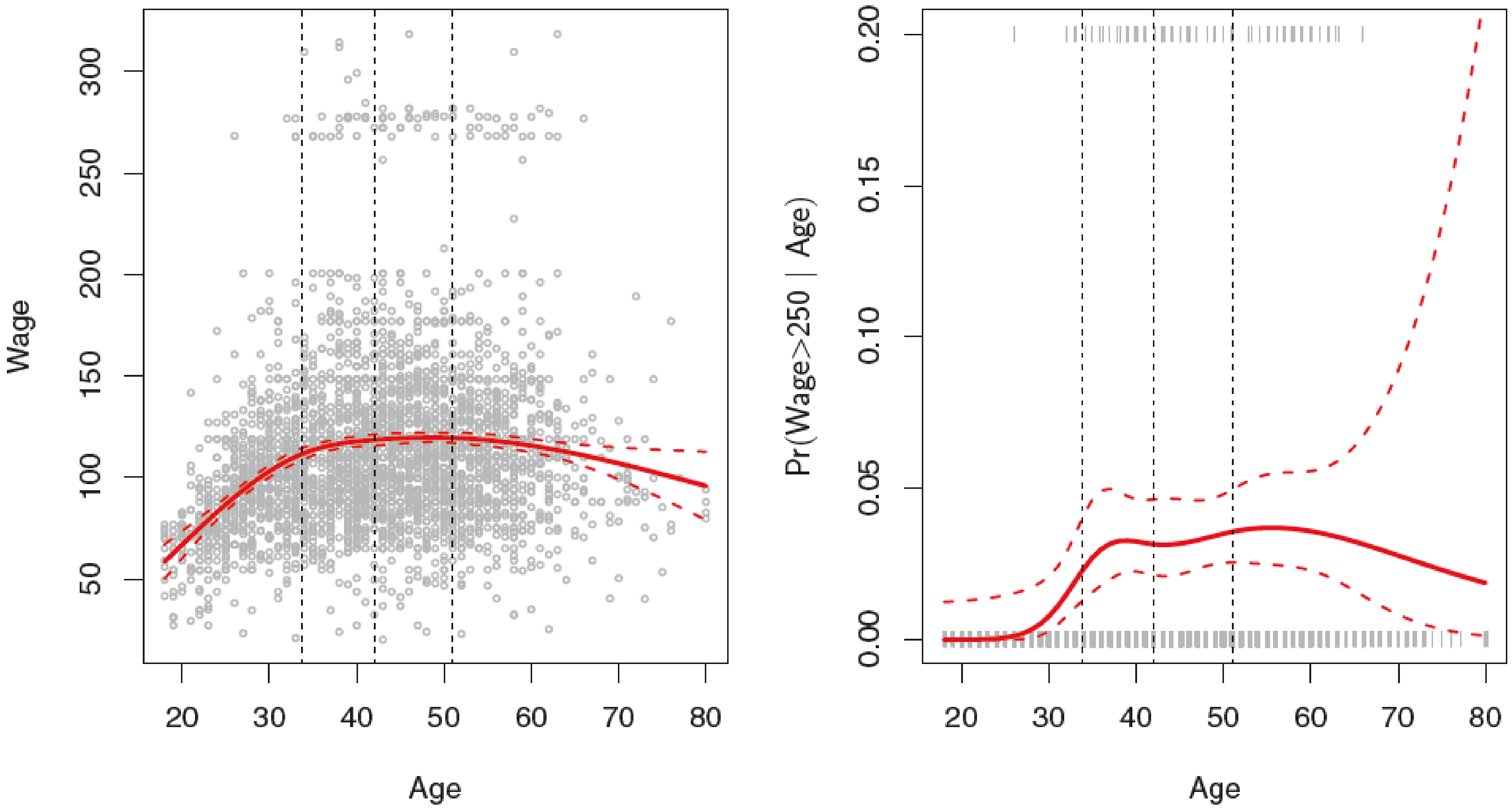
$$h(x, \xi) = (x - \xi)^3_+ = \begin{cases} (x-\xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise,} \end{cases}$$

$$f(X) = \beta_1 N_1(X) + \beta_2 N_2(X) + \sum_{k=1}^{K-2} (\xi_K - \xi_k)\theta_k N_{k+2}(X)$$

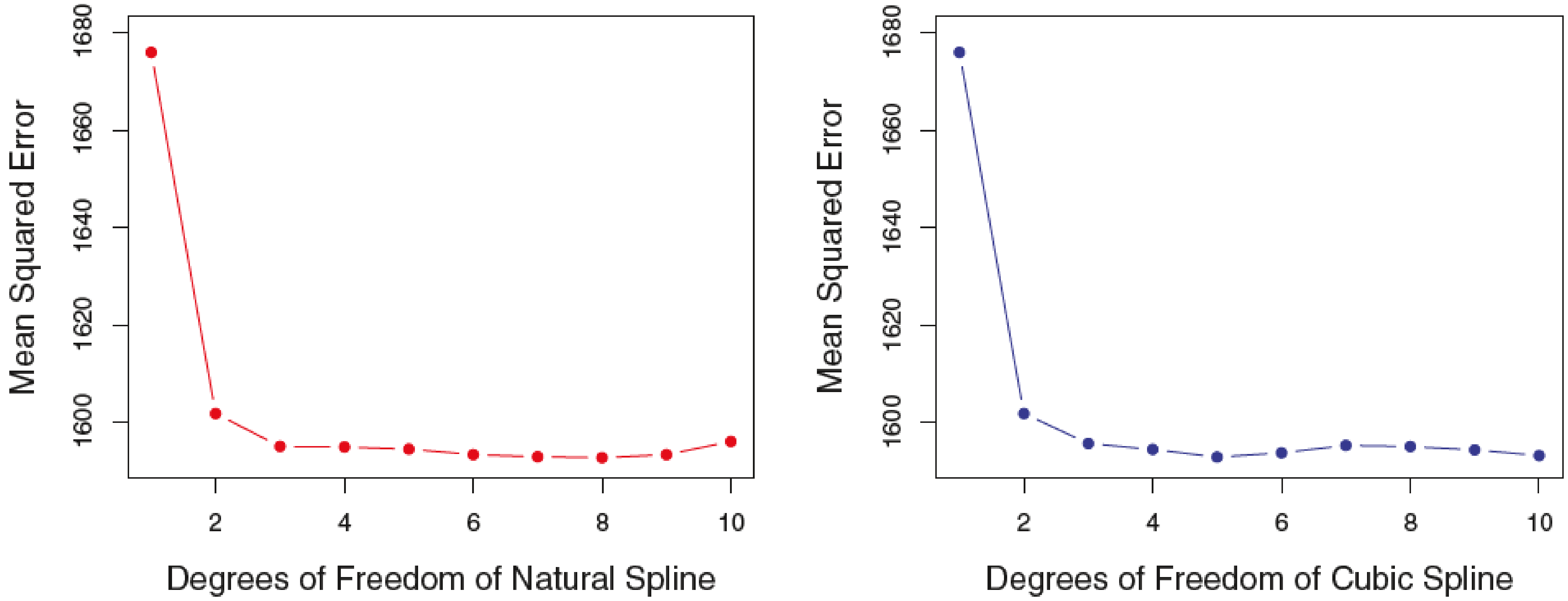
$$N_1(X) = 1, \qquad N_2(X) = X, \qquad N_{k+2}(X) = d_k(X) - d_{K-1}(X),$$

$$d_k(X) = \frac{(X - \xi_k)^3_+ - (X - \xi_K)^3_+}{\xi_K - \xi_k}$$

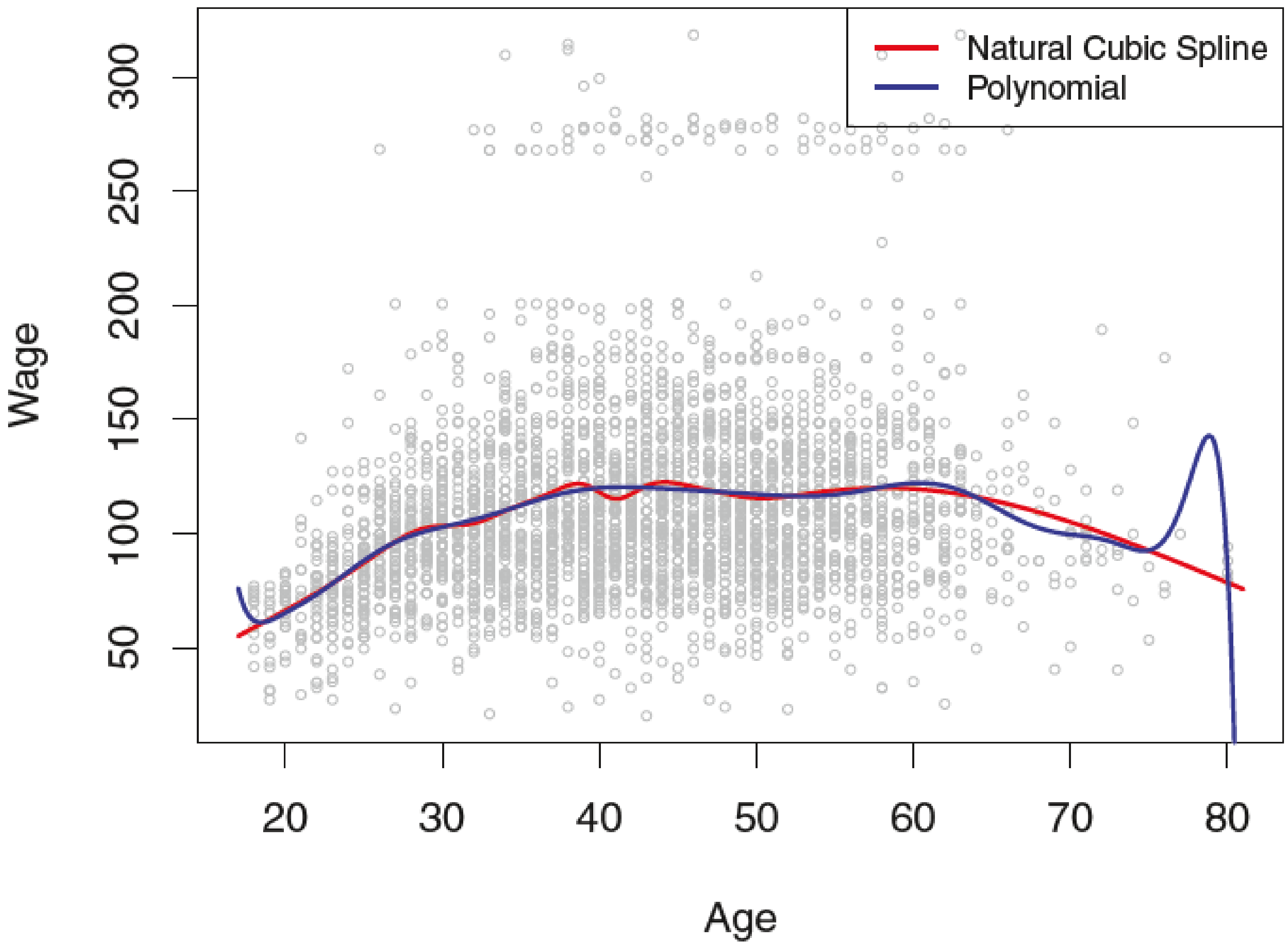




Wage
300
250
200
150
100
50
20
30
40
50
60
70
80
Age
Pr(Wage>250 | Age)
0.20
0.15
0.10
0.05
0.00
20
30
40
50
60
70
80
Age



Mean Squared Error
1600
1620
1640
1660
1680
2
4
6
8
10
Degrees of Freedom of Natural Spline
Mean Squared Error
1600
1620
1640
1660
1680
2
4
6
8
10
Degrees of Freedom of Cubic Spline


Natural Cubic Spline
Polynomial
Wage
Age
50
100
150
200
250
300
20
30
40
50
60
70
80

Згладжувальні сплайни.

Згладжувальні сплайни.

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

Згладжувальні сплайни.

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

Для функції $g(x)$ виконується наступне: вона є кускова кубічна функція з вузлами у всіх значеннях x_1, … x_n; вона є лінійна поза екстремальними вузлами.

Згладжувальні сплайни.

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

Для функції $g(x)$ виконується наступне: вона є кускова кубічна функція з вузлами у всіх значеннях x_1, … x_n; вона є лінійна поза екстремальними вузлами.

Основна проблема – вибір λ.

Маємо $\hat{\mathbf{g}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$, тоді $df_\lambda = \sum_{i=1}^{n}\{\mathbf{S}_\lambda\}_{ii}$

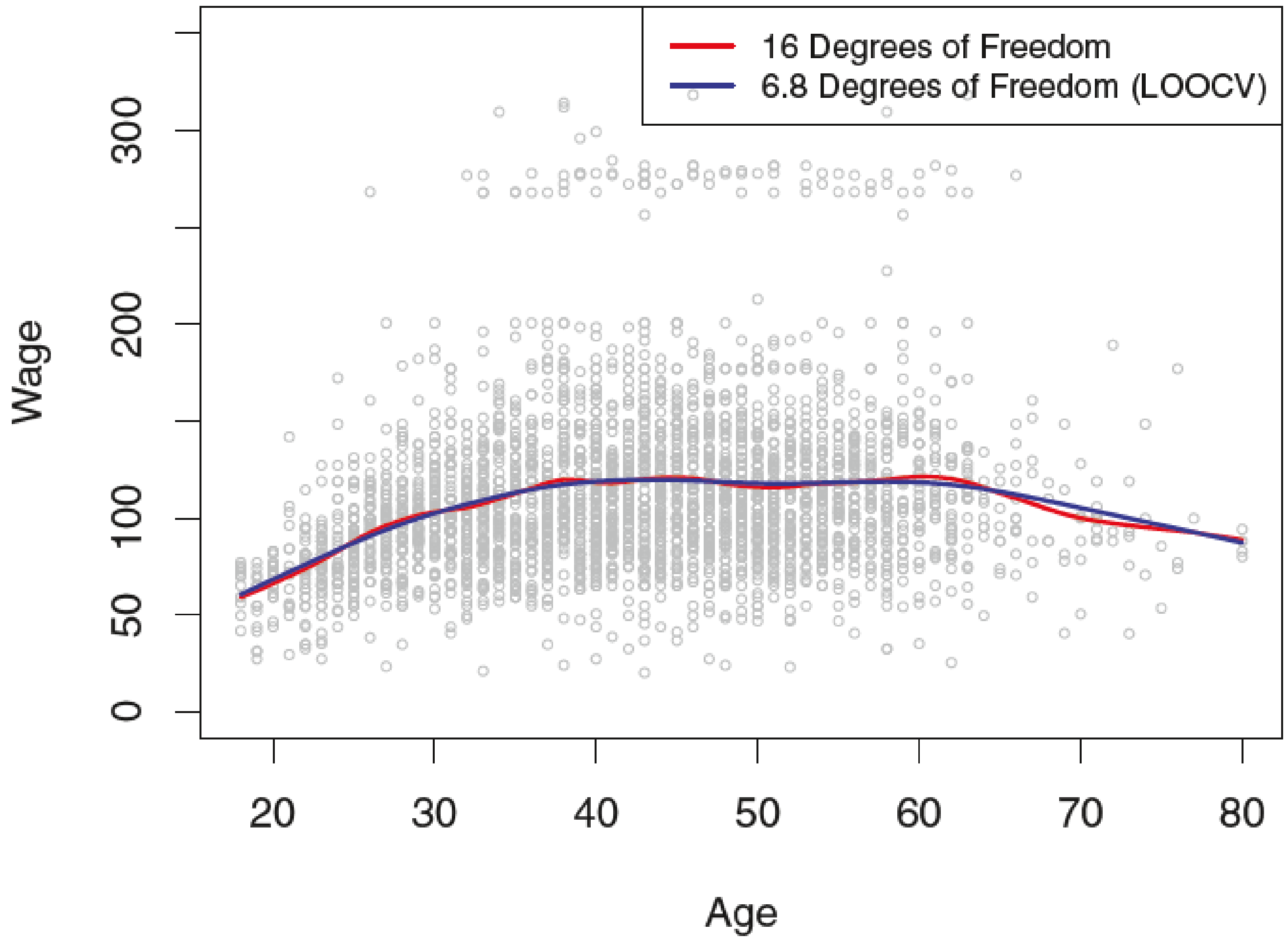
Використовуючи LOOCV, отримаємо

Використовуючи LOOCV, отримаємо

$$\mathrm{RSS}_{cv}(\lambda) = \sum_{i=1}^{n}(y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n}\left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{\mathbf{S}_\lambda\}_{ii}}\right]^2$$

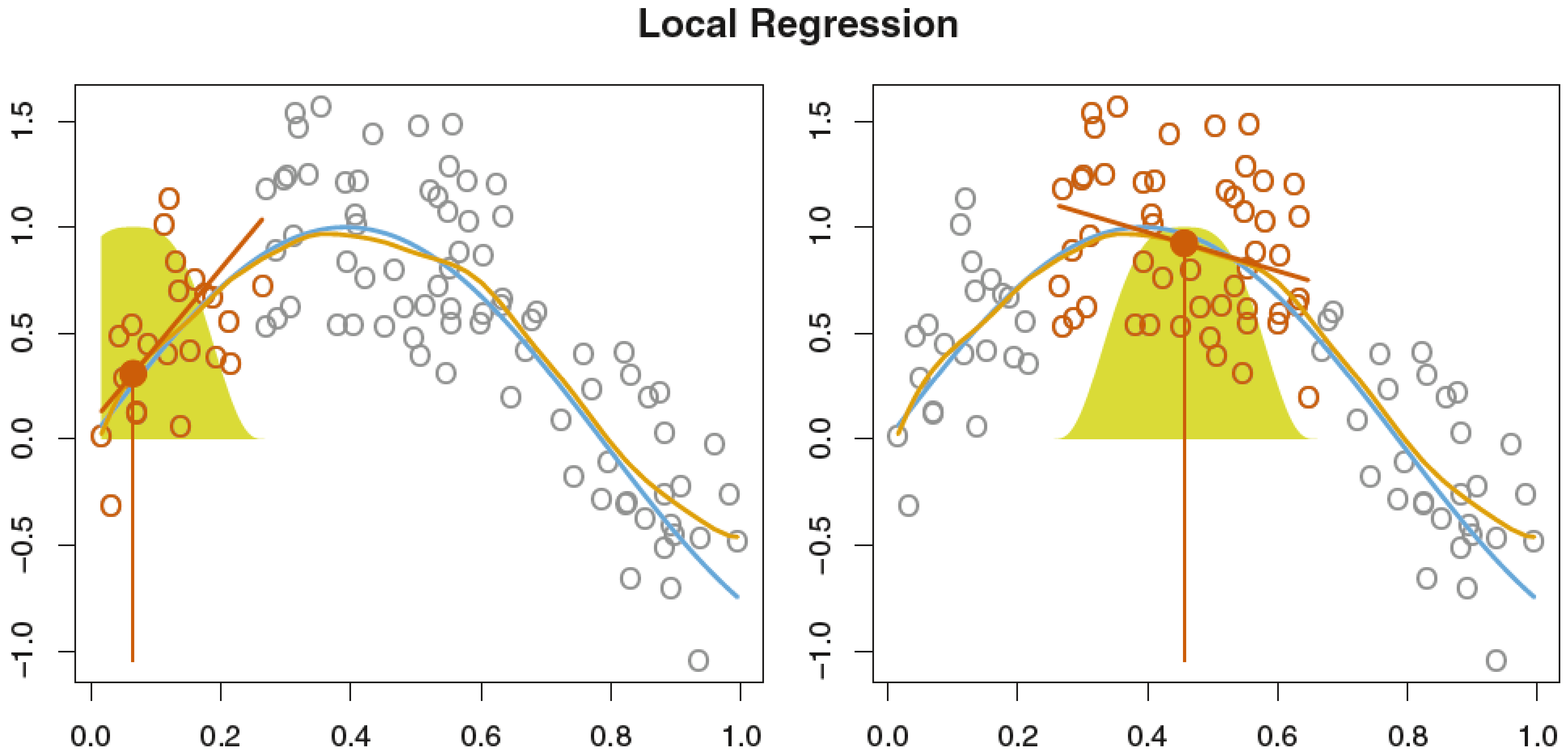
Використовуючи LOOCV, отримаємо

$$\text{RSS}_{cv}(\lambda) = \sum_{i=1}^{n}(y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n}\left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{\mathbf{S}_\lambda\}_{ii}}\right]^2$$



Локальна регресія.

Локальна регресія.



Алгоритм.

Алгоритм.

1. Виберіть частку $s = k / n$ навчальних точок, x_i яких є найближчими до x_0.

Алгоритм.

1. Виберіть частку $s = k / n$ навчальних точок, x_i яких є найближчими до x_0.

2. Призначте вагу $K_{i0} = K(x_i, x_0)$ кожній точці в цьому околі, так що точка, яка найвіддаленіша від x_0, має нульову вагу, а найближча точка має найбільшу вагу. Усі, крім цих найближчих сусідів, мають нульові ваги.

Алгоритм.

1. Виберіть частку $s = k / n$ навчальних точок, x_i яких є найближчими до x_0.

2. Призначте вагу $K_{i0} = K(x_i, x_0)$ кожній точці в цьому околі, так що точка, яка найвіддаленіша від x_0, має нульову вагу, а найближча точка має найбільшу вагу. Усі, крім цих найближчих сусідів, мають нульові ваги.

3. Пристосуйте регресію методом зважених найменших квадратів і обчисліть $\hat{\beta}_0$ і $\hat{\beta}_1$

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2$$

Алгоритм.

1. Виберіть частку $s = k / n$ навчальних точок, x_i яких є найближчими до x_0.

2. Призначте вагу $K_{i0} = K(x_i, x_0)$ кожній точці в цьому околі, так що точка, яка найвіддаленіша від x_0, має нульову вагу, а найближча точка має найбільшу вагу. Усі, крім цих найближчих сусідів, мають нульові ваги.

3. Пристосуйте регресію методом зважених найменших квадратів і обчисліть $\hat{\beta}_0$ і $\hat{\beta}_1$

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2$$

4. Отримаємо в точці x_0

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

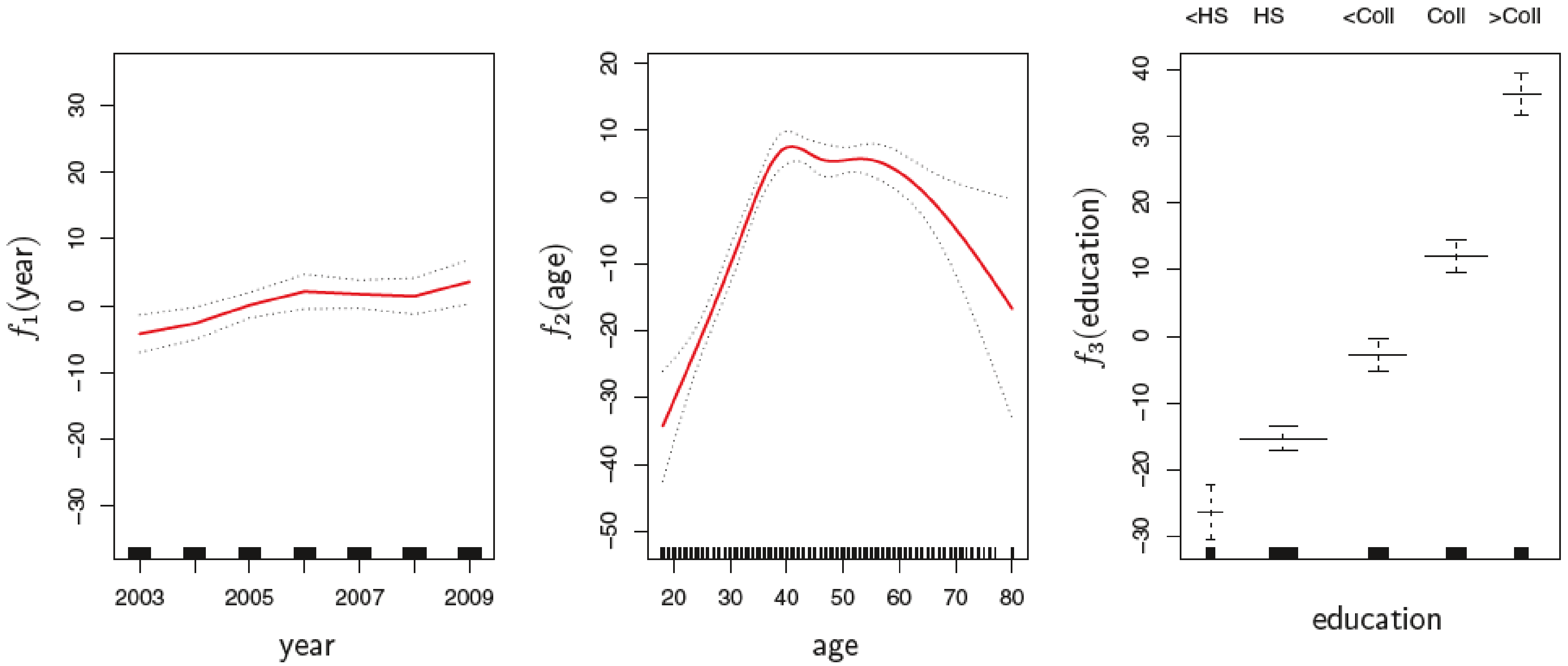
Узагальнені адитивні моделі.

Узагальнені адитивні моделі.

$$
\begin{aligned}
y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\
&= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i
\end{aligned}
$$

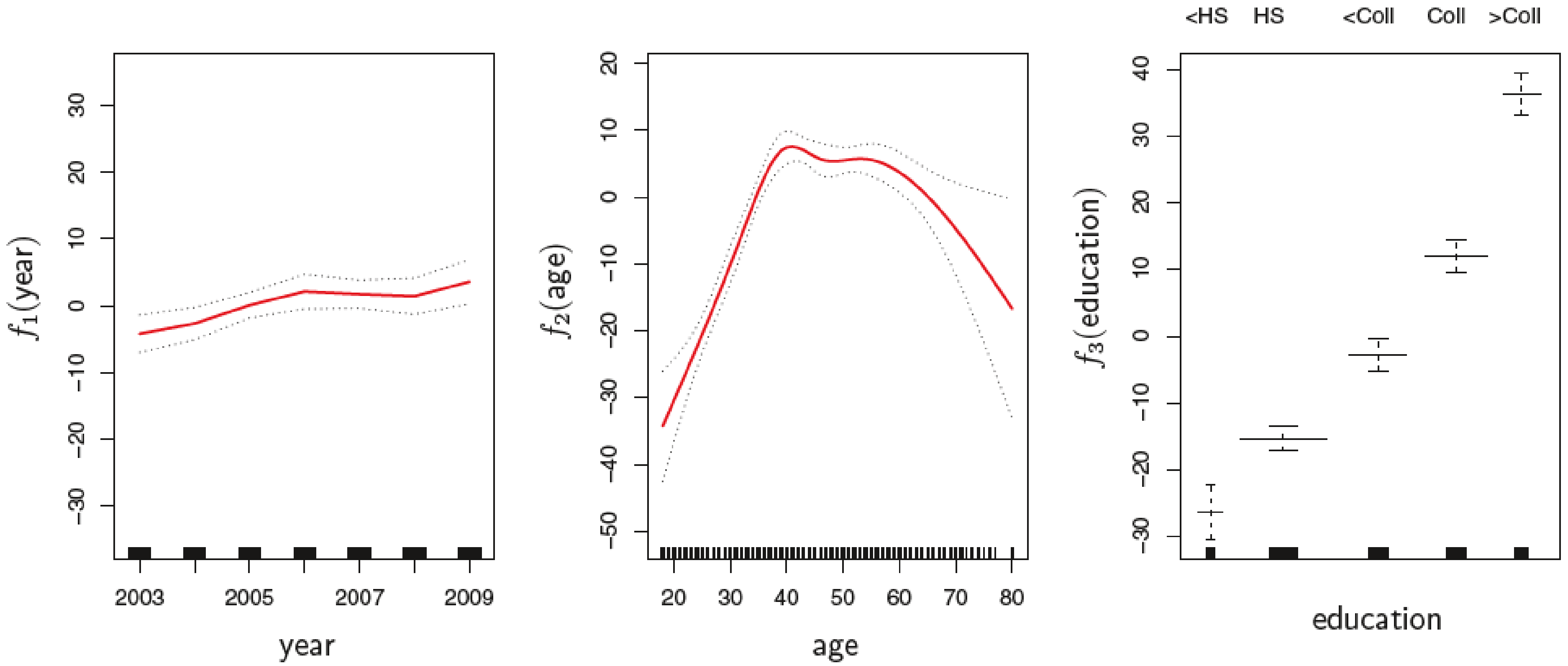
Узагальнені адитивні моделі.

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i \end{aligned}$$



Узагальнені адитивні моделі.

$$
\begin{aligned}
y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\
&= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i
\end{aligned}
$$



$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$

Алгоритм підгонки для згладжувальних сплайнів:

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon$$

Алгоритм підгонки для згладжувальних сплайнів:

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon$$

1. Ініціалізуємо:

$$\hat{\alpha} = \frac{1}{N}\sum_1^N y_i,\ \hat{f}_j \equiv 0, \forall i, j.$$

Алгоритм підгонки для згладжувальних сплайнів:

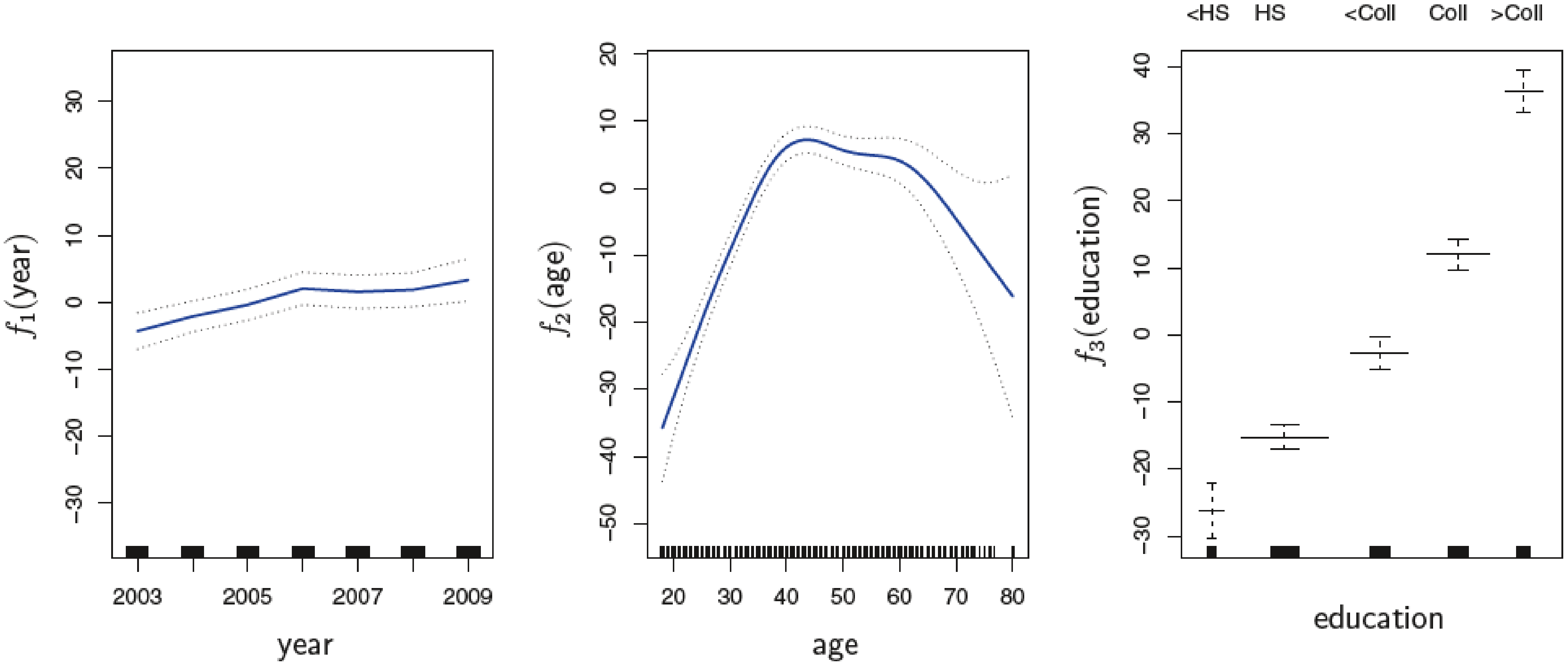
$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon$$

1. Ініціалізуємо:

$$\hat{\alpha} = \frac{1}{N}\sum_1^N y_i,\ \hat{f}_j \equiv 0, \forall i, j.$$

2. Цикл 1, 2, … *p*, 1, 2, …, *p*, …, *p*

$$\begin{aligned}
\hat{f}_j &\leftarrow \mathcal{S}_j\left[\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_1^N\right], \\
\hat{f}_j &\leftarrow \hat{f}_j - \frac{1}{N}\sum_{i=1}^{N} \hat{f}_j(x_{ij}).
\end{aligned}$$



f_1(year)
year
2003
2005
2007
2009
f_2(age)
age
f_3(education)
education
<HS
HS
<Coll
Coll
>Coll

Переваги та обмеження.

УАМ дозволяють нам пристосувати нелінійну функцію f_j до кожного X_j. Це означає, що нам не потрібно намагатися вручну перевіряти багато різних перетворень для кожної змінної окремо.

Переваги та обмеження.

УАМ дозволяють нам пристосувати нелінійну функцію f_j до кожного X_j. Це означає, що нам не потрібно намагатися вручну перевіряти багато різних перетворень для кожної змінної окремо.

Оскільки модель є адитивною, ми все ще можемо вивчити ефект від кожного X_j на Y окремо, утримуючи всі інші змінні сталими.

Переваги та обмеження.

УАМ дозволяють нам пристосувати нелінійну функцію f_j до кожного X_j. Це означає, що нам не потрібно намагатися вручну перевіряти багато різних перетворень для кожної змінної окремо.

Оскільки модель є адитивною, ми все ще можемо вивчити ефект від кожного X_j на Y окремо, утримуючи всі інші змінні сталими.

Гладкість функції f_j для змінної X_j можна підібрати через ступені свободи.

Переваги та обмеження.

УАМ дозволяють нам пристосувати нелінійну функцію f_j до кожного X_j. Це означає, що нам не потрібно намагатися вручну перевіряти багато різних перетворень для кожної змінної окремо.

Оскільки модель є адитивною, ми все ще можемо вивчити ефект від кожного X_j на Y окремо, утримуючи всі інші змінні сталими.

Гладкість функції f_j для змінної X_j можна підібрати через ступені свободи.

Основним обмеженням УАМ є те, що модель є адитивна. За наявності багатьох змінних важливі взаємодії можуть бути пропущені.

Переваги та обмеження.

УАМ дозволяють нам пристосувати нелінійну функцію f_j до кожного X_j. Це означає, що нам не потрібно намагатися вручну перевіряти багато різних перетворень для кожної змінної окремо.

Оскільки модель є адитивною, ми все ще можемо вивчити ефект від кожного X_j на Y окремо, утримуючи всі інші змінні сталими.

Гладкість функції f_j для змінної X_j можна підібрати через ступені свободи.

Основним обмеженням УАМ є те, що модель є адитивна. За наявності багатьох змінних важливі взаємодії можуть бути пропущені. Однак, як і у випадку лінійної регресії, ми можемо вручну додати змінні взаємодії до моделі УАМ, включаючи додаткові предиктори вигляду $X_j \times X_k$. Крім того, ми можемо додати низьковимірну функцію взаємодії вигляду f_{jk} (X_j, X_k); такі функцію можуть бути підігнані за допомогою двовимірних згладжувачів, таких як локальна регресія, або двовимірні сплайни.

Якісна залежна змінна

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

Якісна залежна змінна

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \times \texttt{year} + f_2(\texttt{age}) + f_3(\texttt{education})$$

$$p(X) = \Pr(\texttt{wage} > 250 | \texttt{year}, \texttt{age}, \texttt{education})$$

Якісна залежна змінна

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \times \texttt{year} + f_2(\texttt{age}) + f_3(\texttt{education}) \qquad p(X) = \Pr(\texttt{wage} > 250 | \texttt{year}, \texttt{age}, \texttt{education})$$