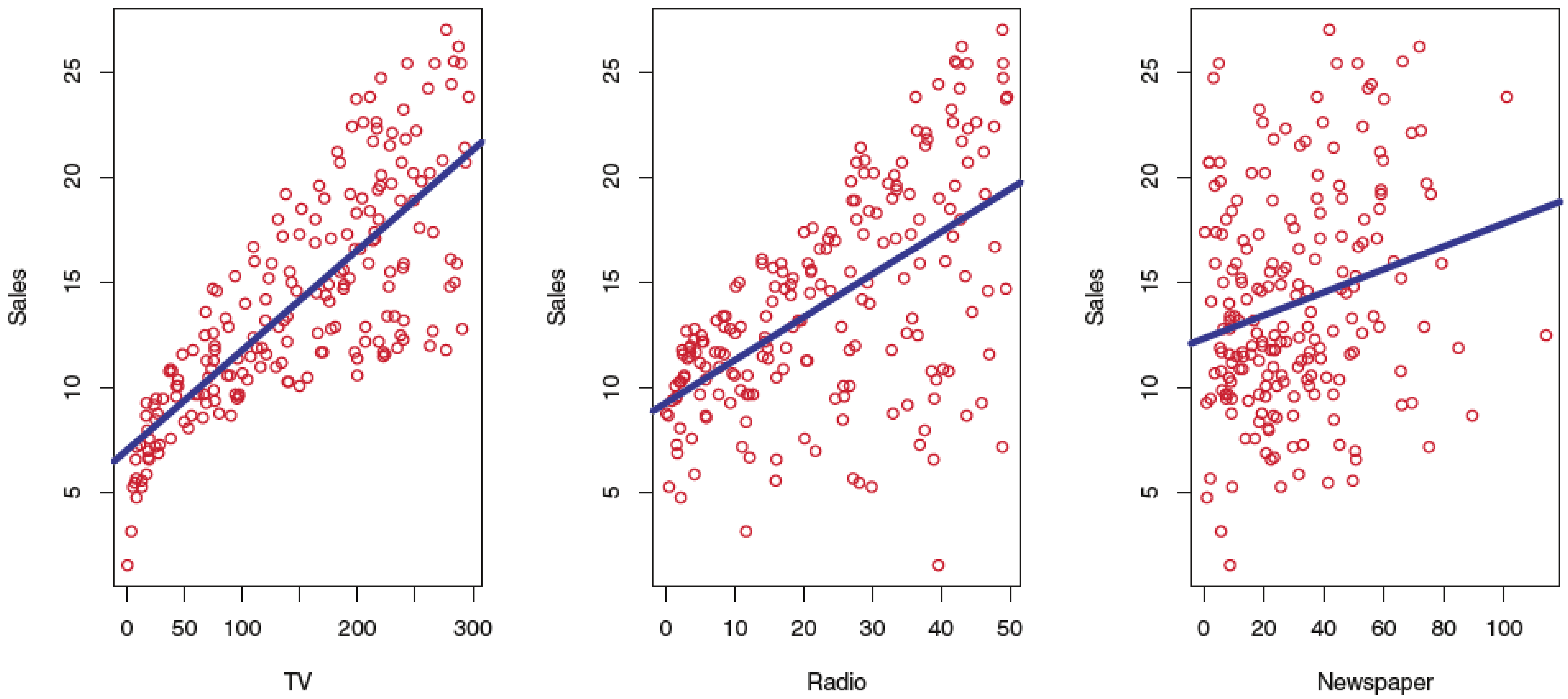
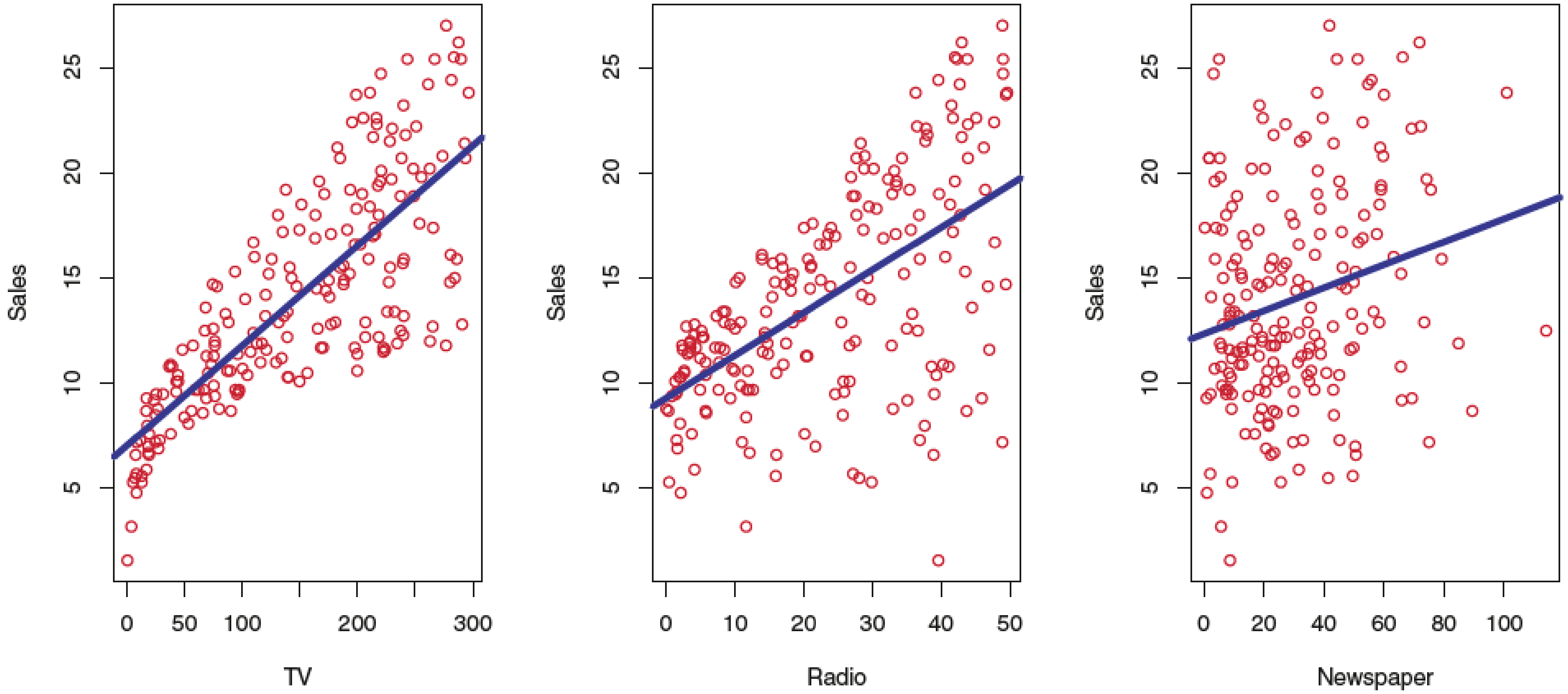
Моделі статистичного навчання: статистичне навчання

Приклад. Нехай нам потрібно порадити як підвищити продажі певного продукту. Нехай маємо дані про продажі цього продукту у 200 різних маркетах та бюджети реклами в кожному маркеті на радіо, ТВ та у газетах.

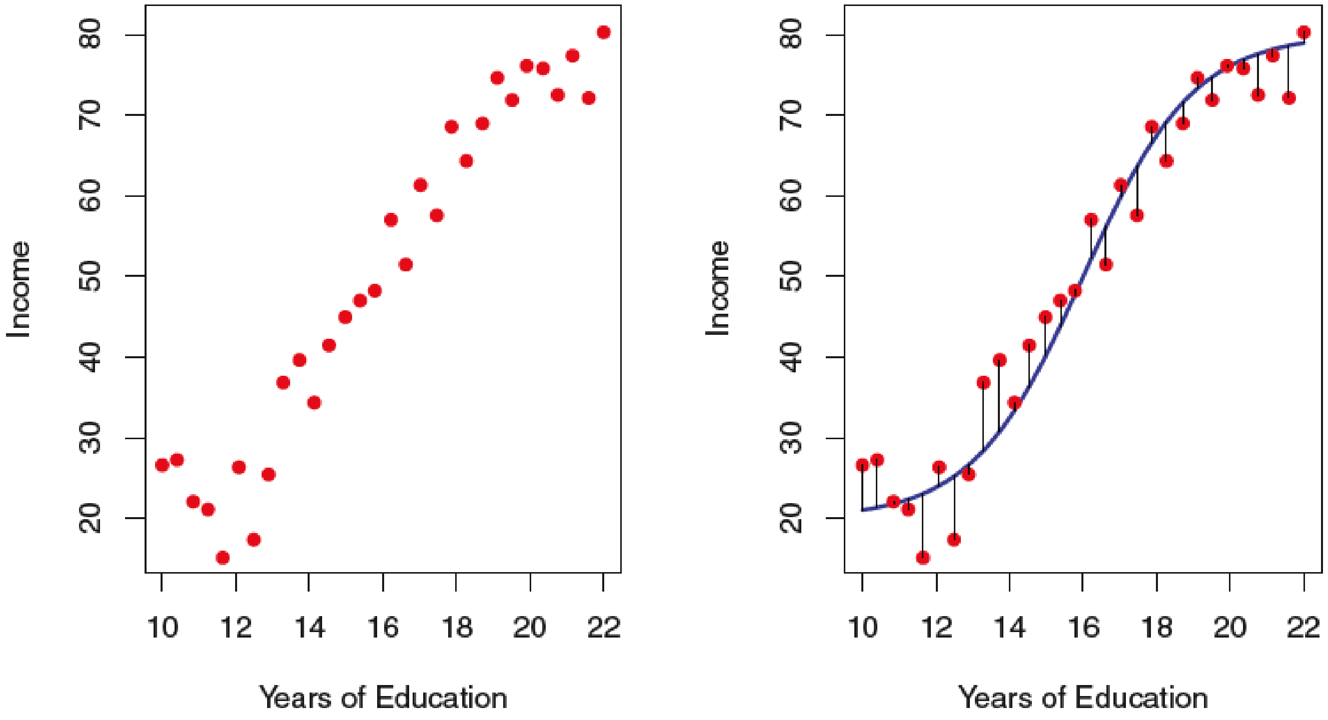
Приклад. Нехай нам потрібно порадити як підвищити продажі певного продукту. Нехай маємо дані про продажі цього продукту у 200 різних маркетах та бюджети реклами в кожному маркеті на радіо, ТВ та у газетах.



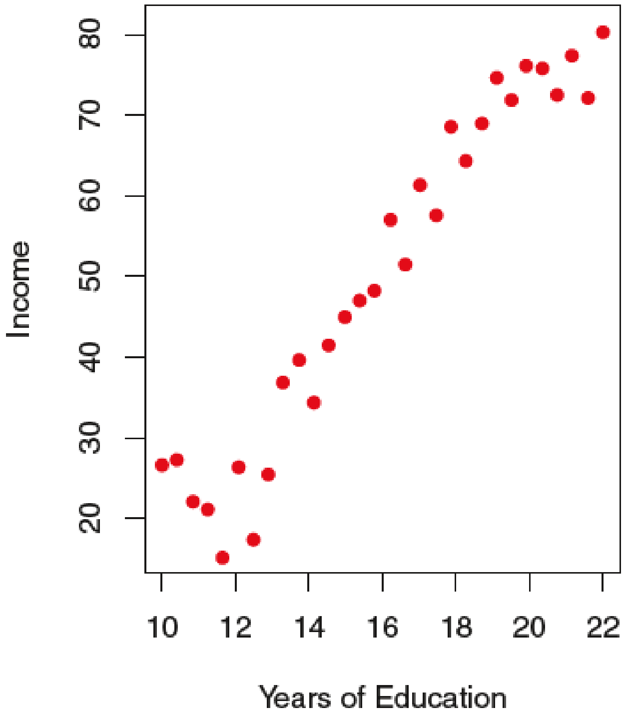
Приклад. Нехай нам потрібно порадити як підвищити продажі певного продукту. Нехай маємо дані про продажі цього продукту у 200 різних маркетах та бюджети реклами в кожному маркеті на радіо, ТВ та у газетах.

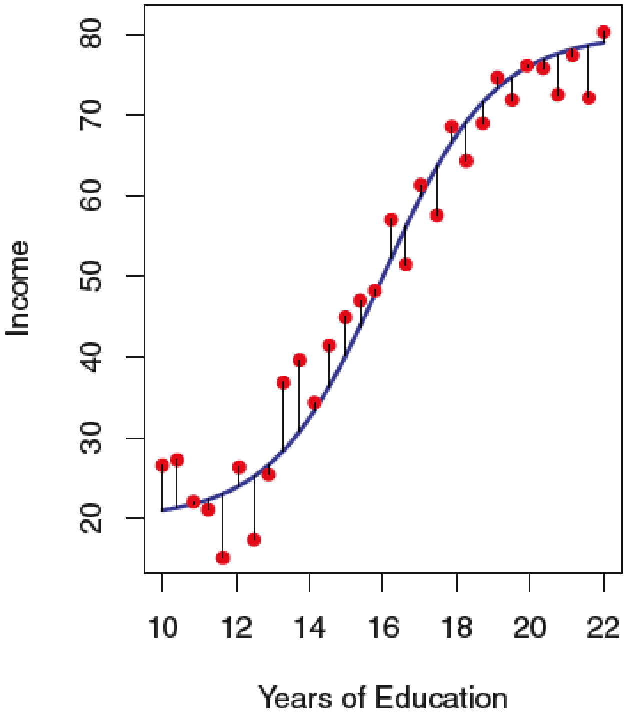


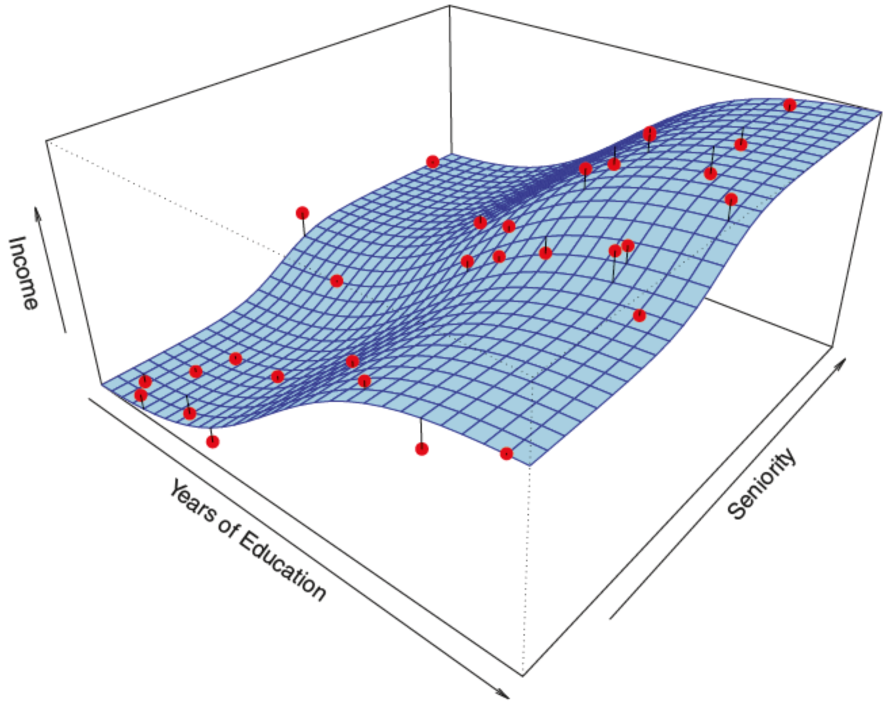
$$Y = f(\mathbf{X}) + \varepsilon, \mathbf{X} = (X_1, X_2, X_3)$$



Income
Years of Education
10 12 14 16 18 20 22
20 30 40 50 60 70 80
Income
Years of Education
10 12 14 16 18 20 22
20 30 40 50 60 70 80


Income
80
70
60
50
40
30
20
10
12
14
16
18
20
22
Years of Education


Income
80
70
60
50
40
30
20
10
12
14
16
18
20
22
Years of Education


Income
Years of Education
Seniority

Для чого оцінювати f: передбачення та висновок.

Для чого оцінювати *f*: передбачення та висновок.

Передбачення. На основі значень вхідних даних визначити результуючі.

$$\hat{Y} = \hat{f}(X)$$

Для чого оцінювати *f*: передбачення та висновок.

Передбачення. На основі значень вхідних даних визначити результуючі.

$$\hat{Y} = \hat{f}(X)$$

Точність оцінки $\hat{Y}$ залежить від двох показників: помилки, що можна зменшити і помилки, що зменшити не можна.

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

Для чого оцінювати *f*: передбачення та висновок.

Передбачення. На основі значень вхідних даних визначити результуючі.

$$\hat{Y} = \hat{f}(X)$$

Точність оцінки $\hat{Y}$ залежить від двох показників: помилки, що можна зменшити і помилки, що зменшити не можна.

$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

Одним з основних завдань є мінімізація помилки (звичайно тієї, яку можна зменшити)

Висновок. Часто на перше місце виходить питання як значення результуючої змінної Y залежить від значень вхідних факторів X_1, X_2, …, X_p.

Висновок. Часто на перше місце виходить питання як значення результуючої змінної Y залежить від значень вхідних факторів X_1, X_2, …, X_p.

Які фактори дійсно пов’язані з результатом?

Висновок. Часто на перше місце виходить питання як значення результуючої змінної Y залежить від значень вхідних факторів X_1, X_2, …, X_p.

Які фактори дійсно пов’язані з результатом?

Яка залежність між результатом та кожним фактором?

Висновок. Часто на перше місце виходить питання як значення результуючої змінної Y залежить від значень вхідних факторів X_1, X_2, …, X_p.

Які фактори дійсно пов’язані з результатом?

Яка залежність між результатом та кожним фактором?

Чи можна коректно описати взаємозв'язок між Y та кожним фактором на основі лінійного рівняння, чи взаємозв'язок є складніший?

Висновок. Часто на перше місце виходить питання як значення результуючої змінної Y залежить від значень вхідних факторів X_1, X_2, …, X_p.

Які фактори дійсно пов’язані з результатом?

Яка залежність між результатом та кожним фактором?

Чи можна коректно описати взаємозв'язок між Y та кожним фактором на основі лінійного рівняння, чи взаємозв'язок є складніший?

Як оцінити *f*?

Висновок. Часто на перше місце виходить питання як значення результуючої змінної Y залежить від значень вхідних факторів X_1, X_2, …, X_p.

Які фактори дійсно пов’язані з результатом?

Яка залежність між результатом та кожним фактором?

Чи можна коректно описати взаємозв'язок між Y та кожним фактором на основі лінійного рівняння, чи взаємозв'язок є складніший?

Як оцінити *f*?

Нехай нам задано n даних, тобто пар $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, де $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

Висновок. Часто на перше місце виходить питання як значення результуючої змінної Y залежить від значень вхідних факторів $X_1, X_2, \ldots, X_p$.

Які фактори дійсно пов'язані з результатом?

Яка залежність між результатом та кожним фактором?

Чи можна коректно описати взаємозв'язок між Y та кожним фактором на основі лінійного рівняння, чи взаємозв'язок є складніший?

Як оцінити *f*?

Нехай нам задано n даних, тобто пар $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, де $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

Параметричні та непараметричні методи.

Параметричні методи.

Параметричні методи. Два етапи:

1. Визначитися з виглядом функції *f*. Наприклад, припустивши, що шукана функція є лінійна, отримаємо

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Параметричні методи. Два етапи:

1. Визначитися з виглядом функції *f*. Наприклад, припустивши, що шукана функція є лінійна, отримаємо

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

2. Далі нам потрібно на основі навчальних даних пристосувати нашу модель. Наприклад, за припущення лінійності функції f нам необхідно оцінити значення параметрів β_0, $\beta_1, \ldots, \beta_p$.

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p$$

Параметричні методи. Два етапи:

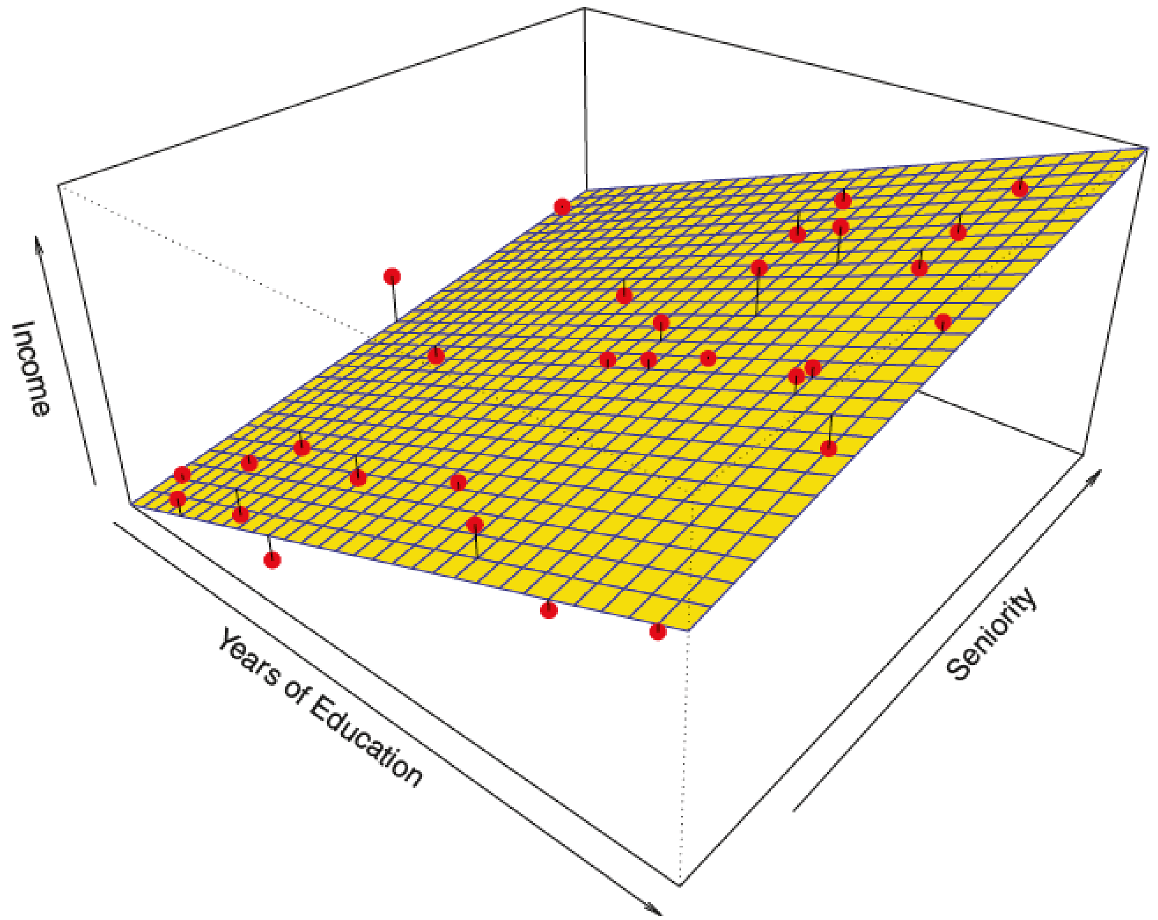
1. Визначитися з виглядом функції *f*. Наприклад, припустивши, що шукана функція є лінійна, отримаємо

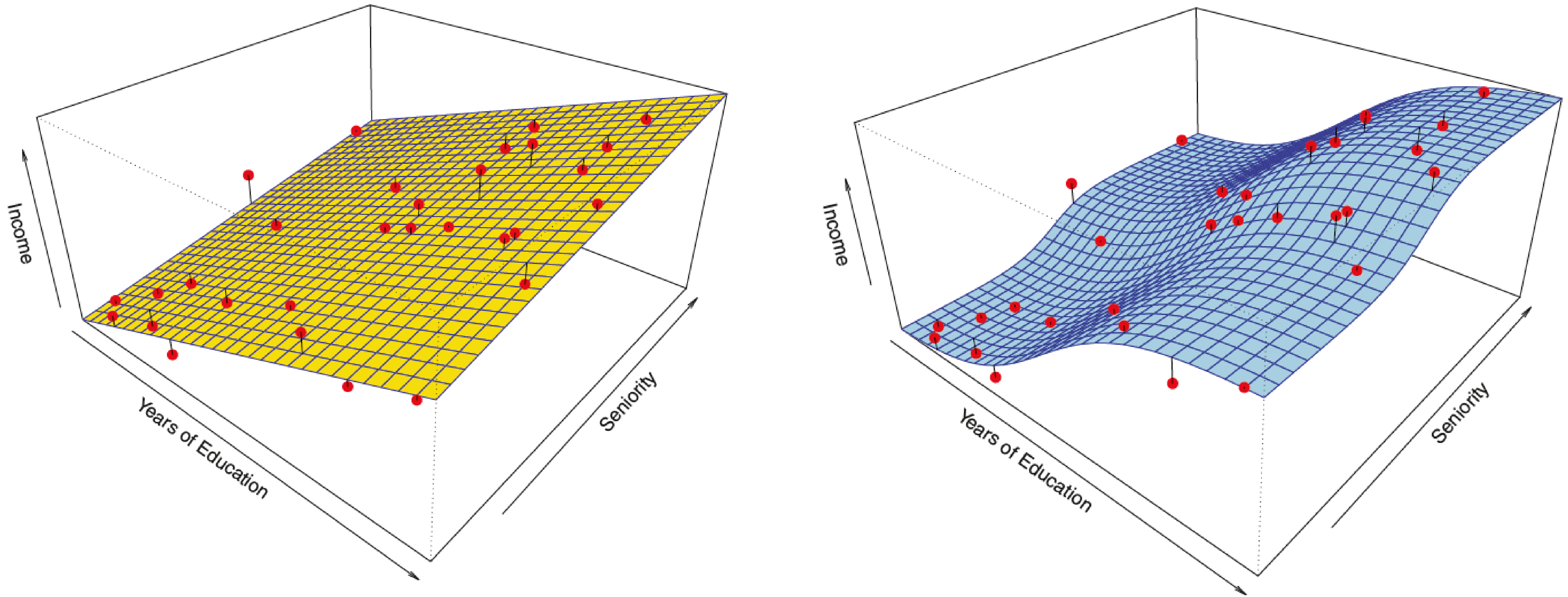
$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

2. Далі нам потрібно на основі навчальних даних пристосувати нашу модель. Наприклад, за припущення лінійності функції *f* нам необхідно оцінити значення параметрів β_0, $\beta_1, \ldots, \beta_p$.

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p$$

$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}.$$


Income
Years of Education
Seniority


Income
Years of Education
Seniority
Income
Years of Education
Seniority

Непараметричні методи.

Непараметричні методи. Не робиться жодного припущення щодо вигляду функції *f*. Основою є побудова функції, яка дає найкраще наближення на всій тренувальній множині.

Непараметричні методи. Не робиться жодного припущення щодо вигляду функції *f*. Основою є побудова функції, яка дає найкраще наближення на всій тренувальній множині.

Основним недоліком є необхідність наявності великої за обсягом тренувальної вибірки.

Непараметричні методи. Не робиться жодного припущення щодо вигляду функції *f*. Основою є побудова функції, яка дає найкраще наближення на всій тренувальній множині.

Основним недоліком є необхідність наявності великої за обсягом тренувальної вибірки.

Для попереднього прикладу, використання тонкопластинчастих сплайнів дає добре наближення шуканої поверхні.

Непараметричні методи. Не робиться жодного припущення щодо вигляду функції f. Основою є побудова функції, яка дає найкраще наближення на всій тренувальній множині.

Основним недоліком є необхідність наявності великої за обсягом тренувальної вибірки.

Для попереднього прикладу, використання тонкопластинчастих сплайнів дає добре наближення шуканої поверхні.

При використанні цього методу потрібно вибрати рівень гладкості поверхні.

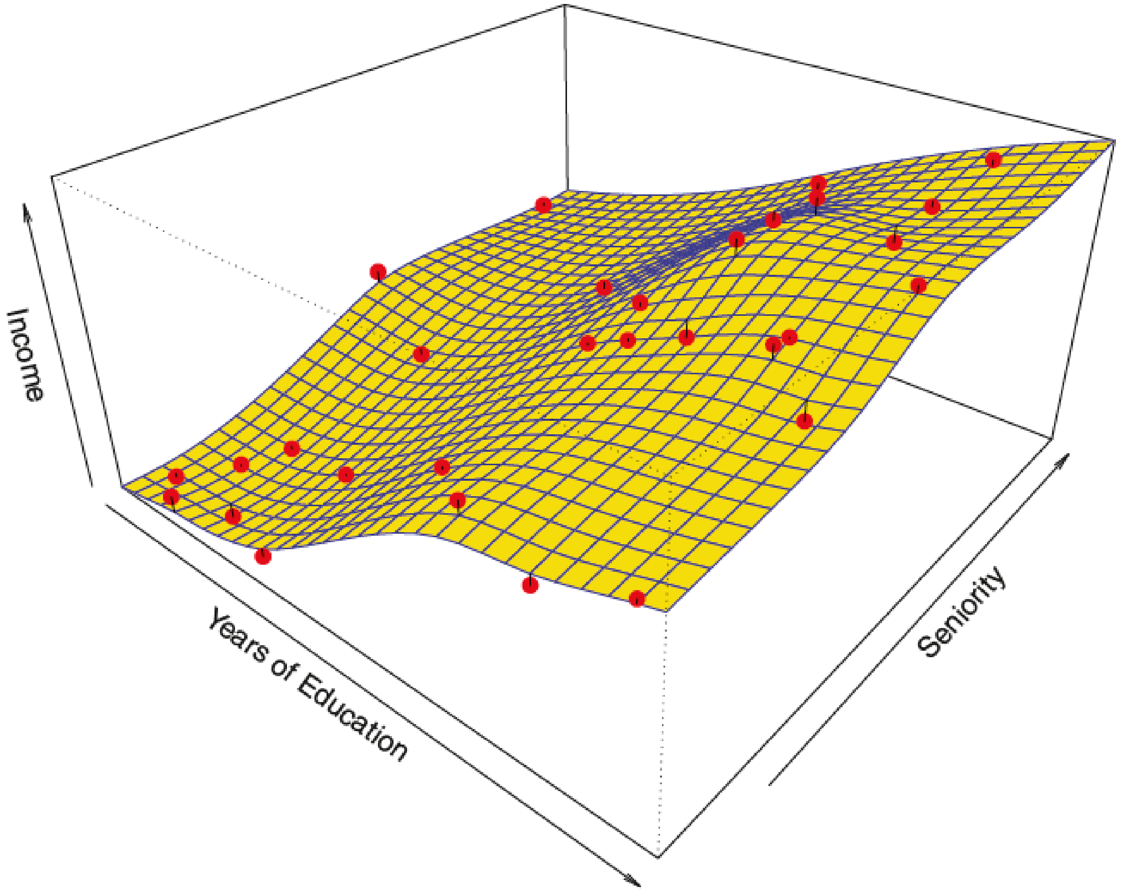
Непараметричні методи. Не робиться жодного припущення щодо вигляду функції *f*. Основою є побудова функції, яка дає найкраще наближення на всій тренувальній множині.

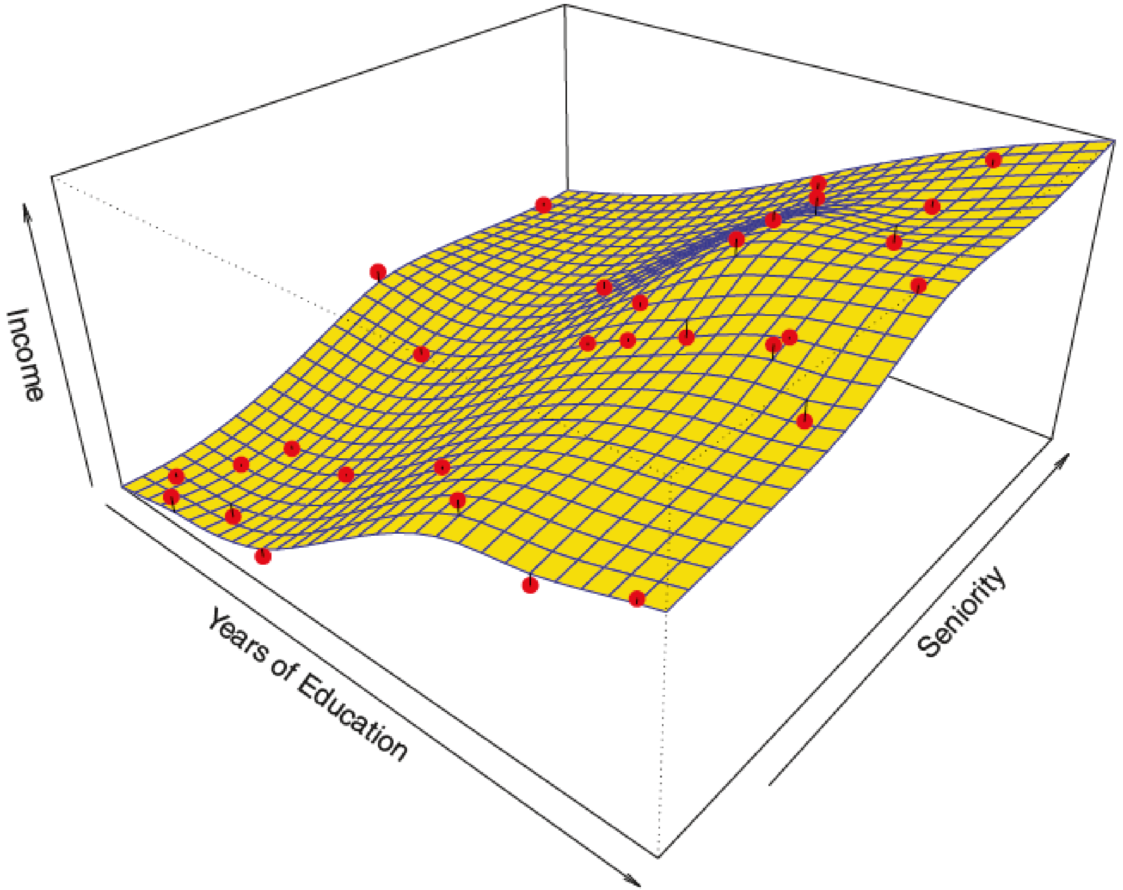
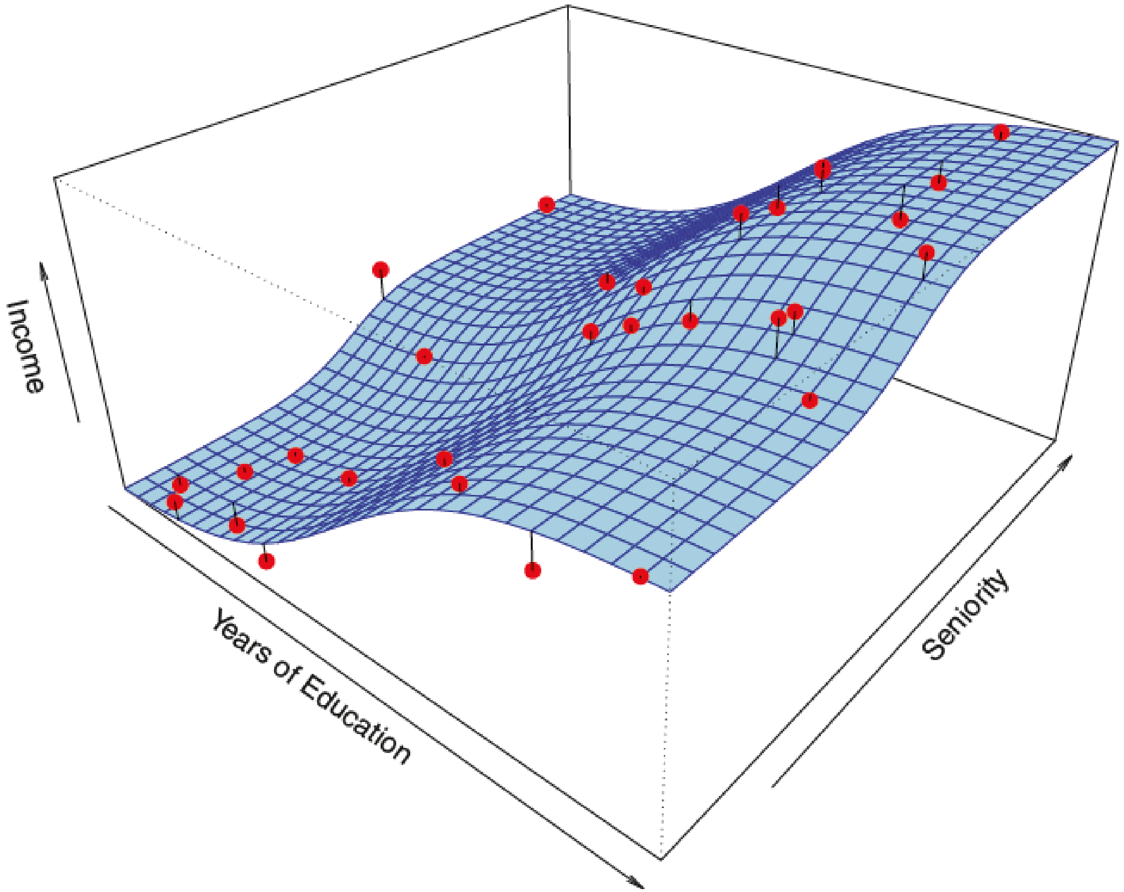
Основним недоліком є необхідність наявності великої за обсягом тренувальної вибірки.

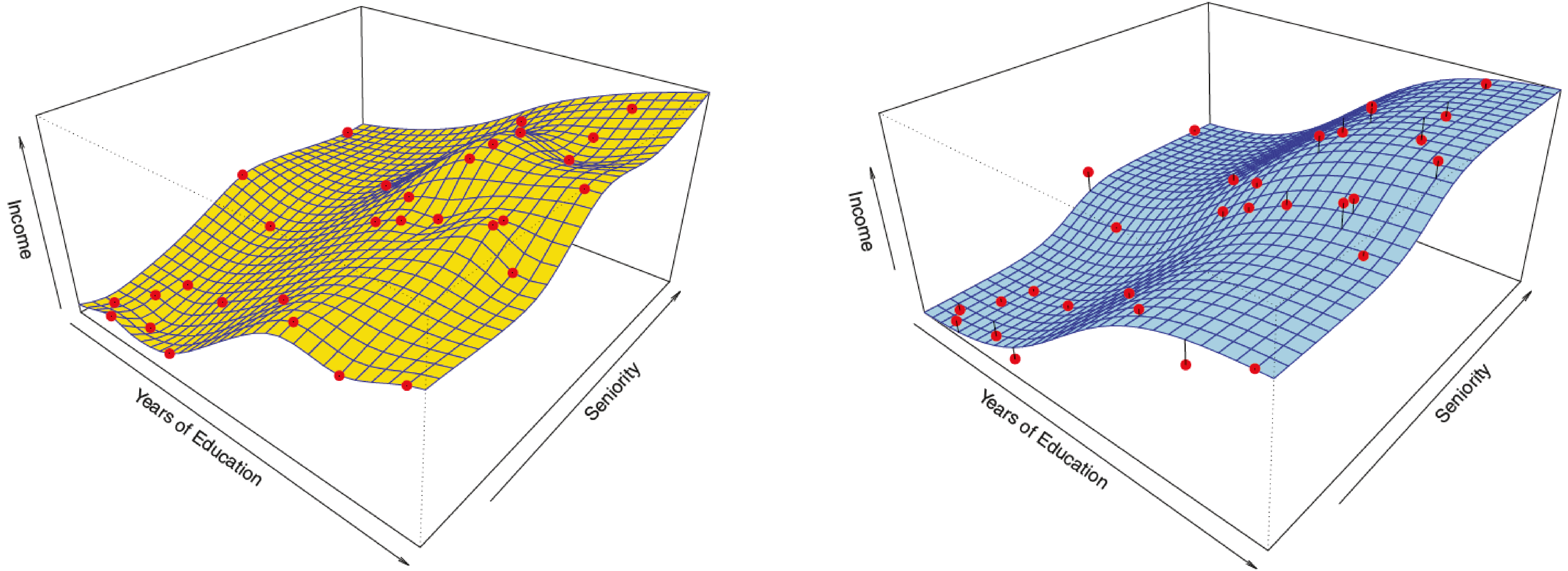
Для попереднього прикладу, використання тонкопластинчастих сплайнів дає добре наближення шуканої поверхні.

При використанні цього методу потрібно вибрати рівень гладкості поверхні.

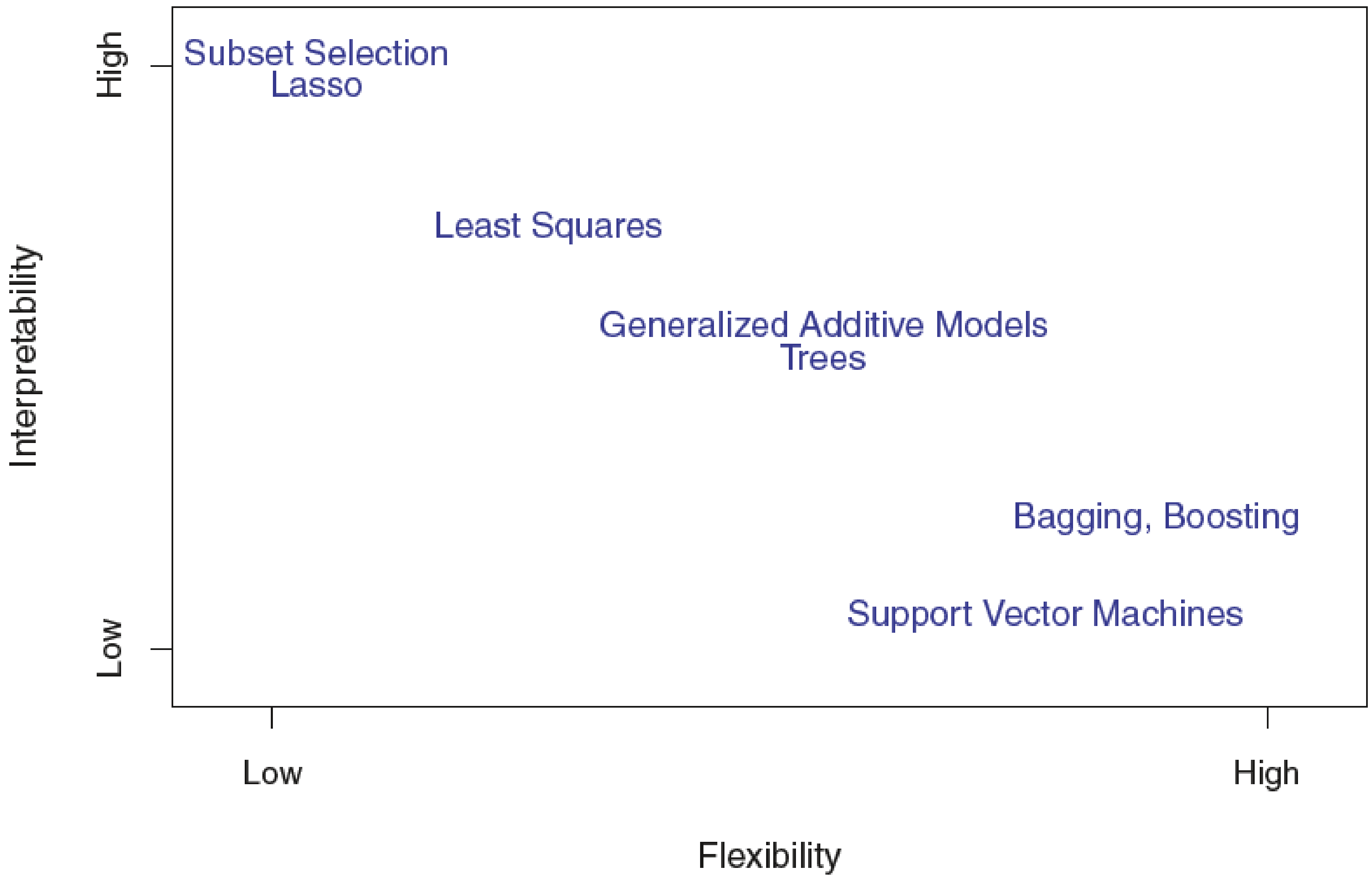
Вибравши менший рівень гладкості для цього методу отримаємо «дуже» добру оцінку у спостережуваних точках, проте поза ними оцінка поверхні є «поганою».


Income
Years of Education
Seniority

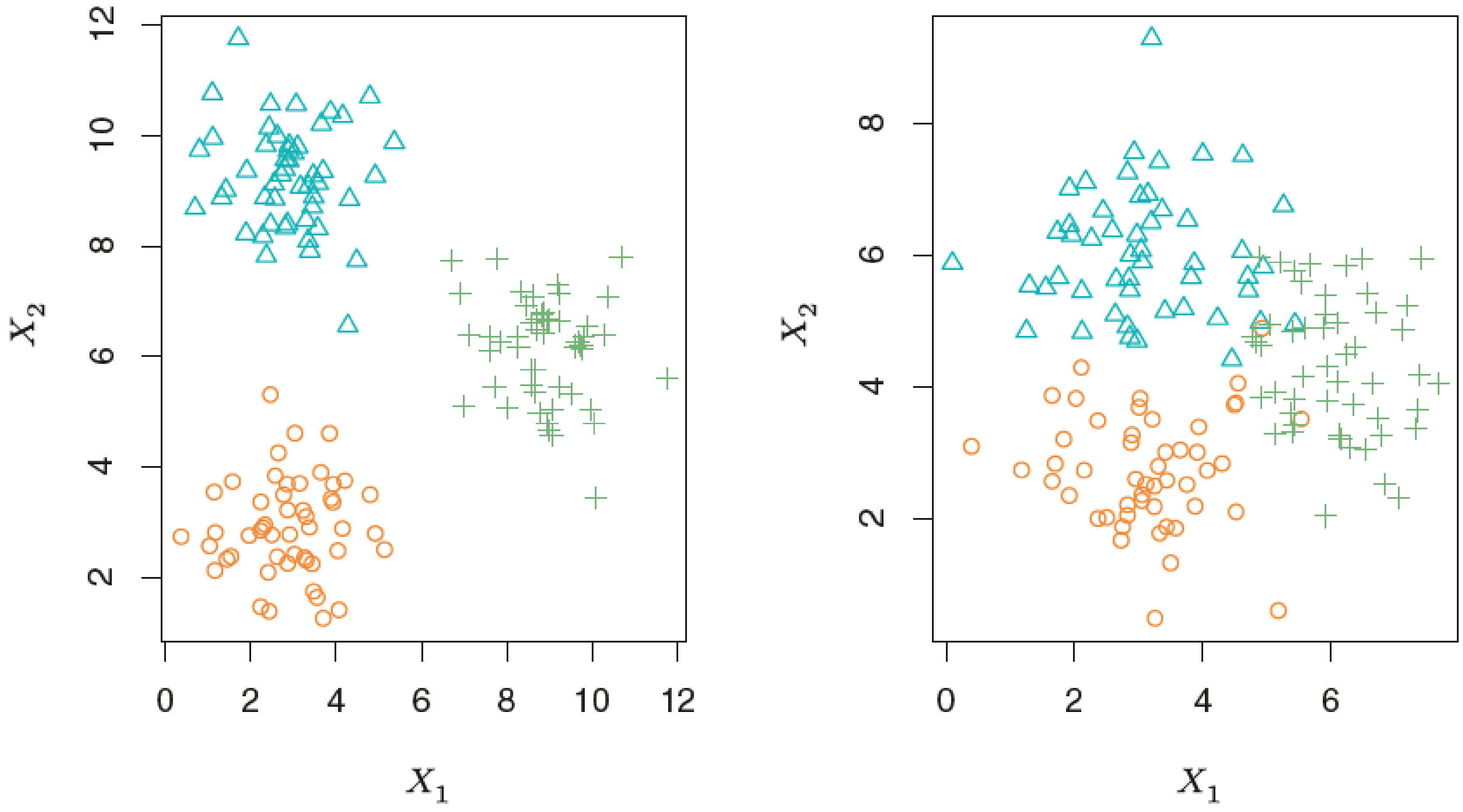

Income
Years of Education
Seniority

Income
Years of Education
Seniority


Income
Years of Education
Seniority
Income
Years of Education
Seniority

Компроміс між гнучкістю та інтерпретованістю моделі.



Навчання без вчителя.



Якість моделі.

Якість моделі.

Якість пристосування моделі.

Якість моделі.

Якість пристосування моделі.

Середня квадратична помилка (*MSE*)

Якість моделі.

Якість пристосування моделі.

Середня квадратична помилка (*MSE*)

Тренувальна помилка

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Якість моделі.

Якість пристосування моделі.

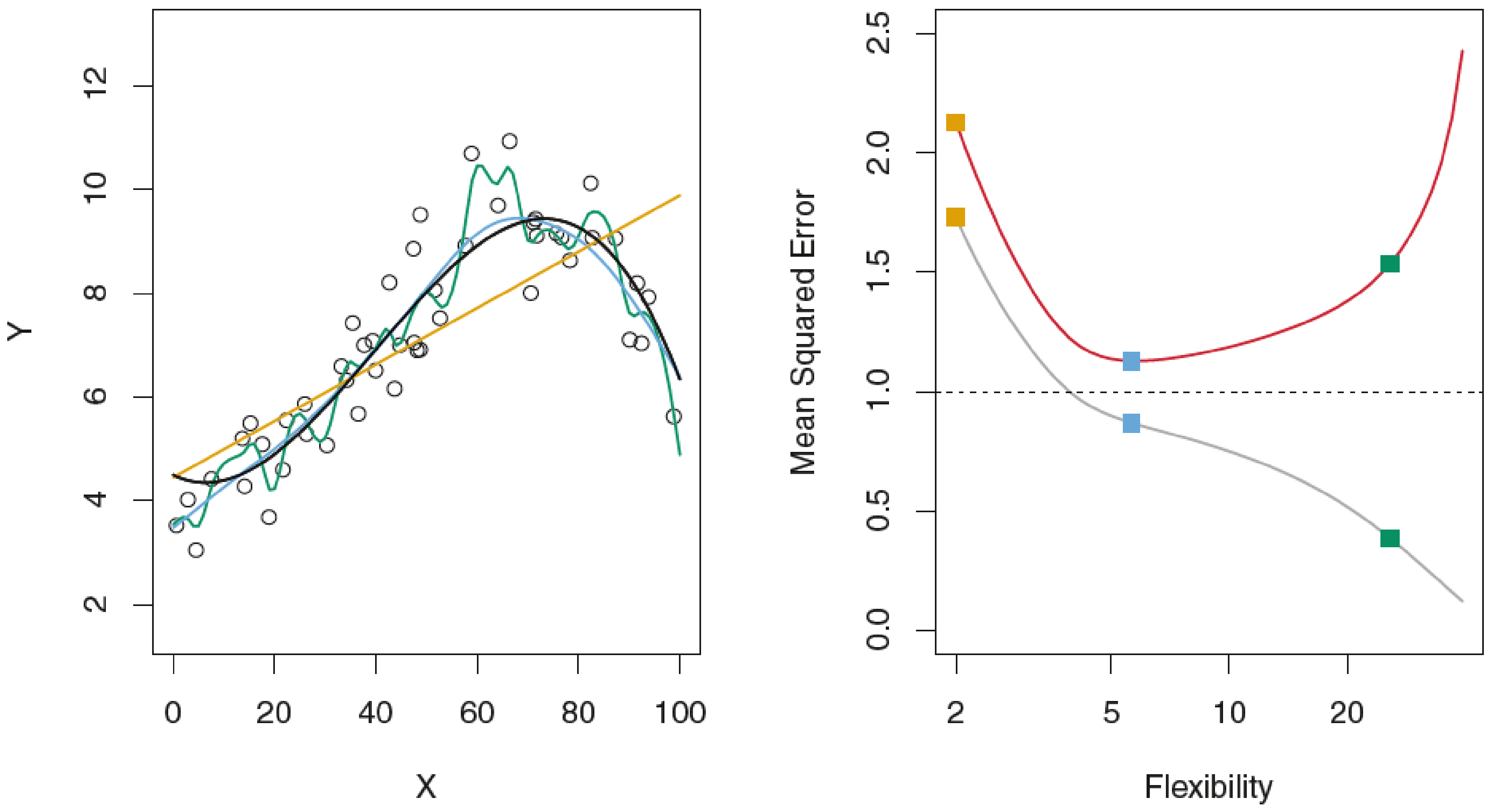
Середня квадратична помилка (*MSE*)

Тренувальна помилка

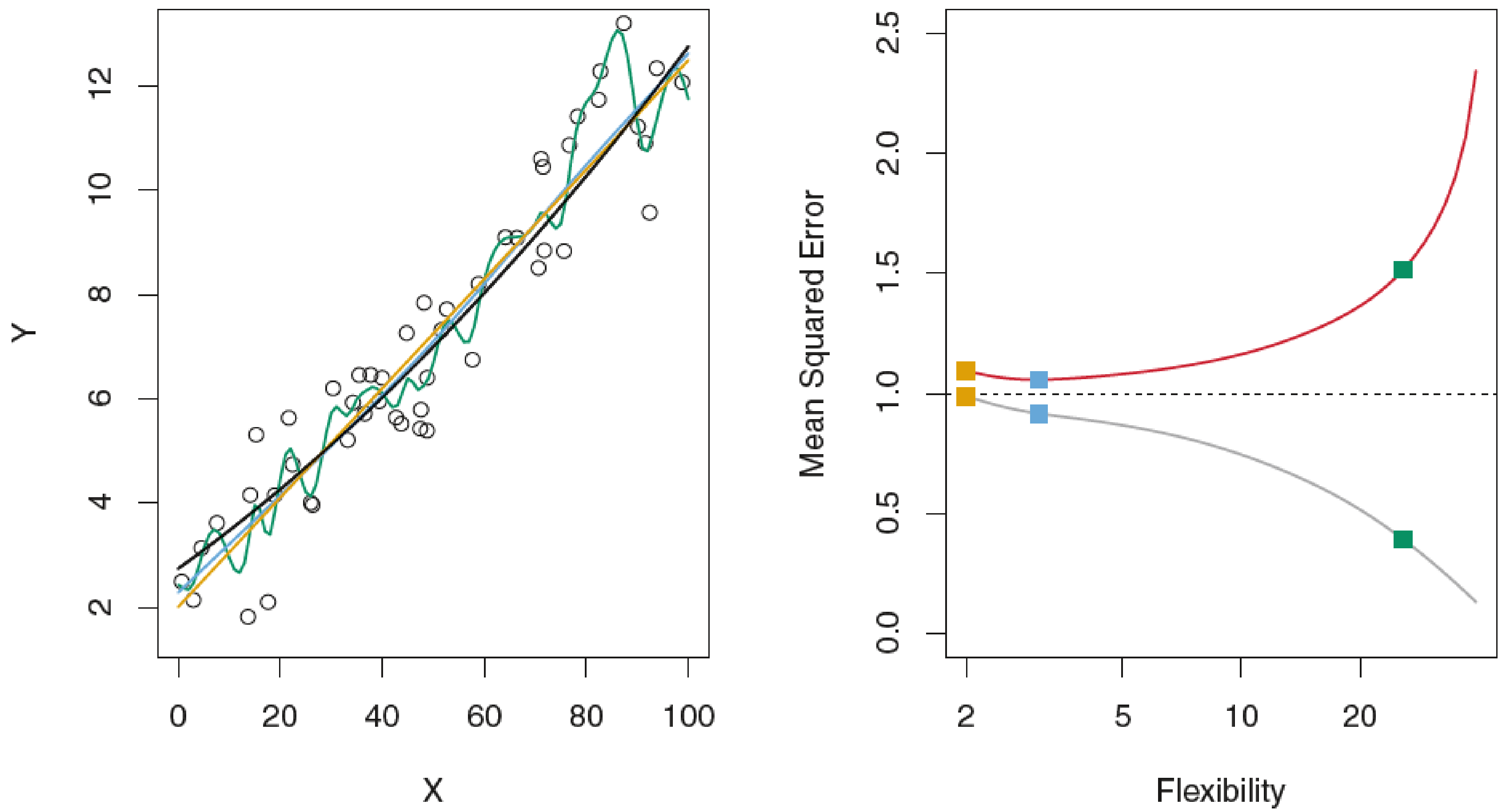
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Тестова помилка

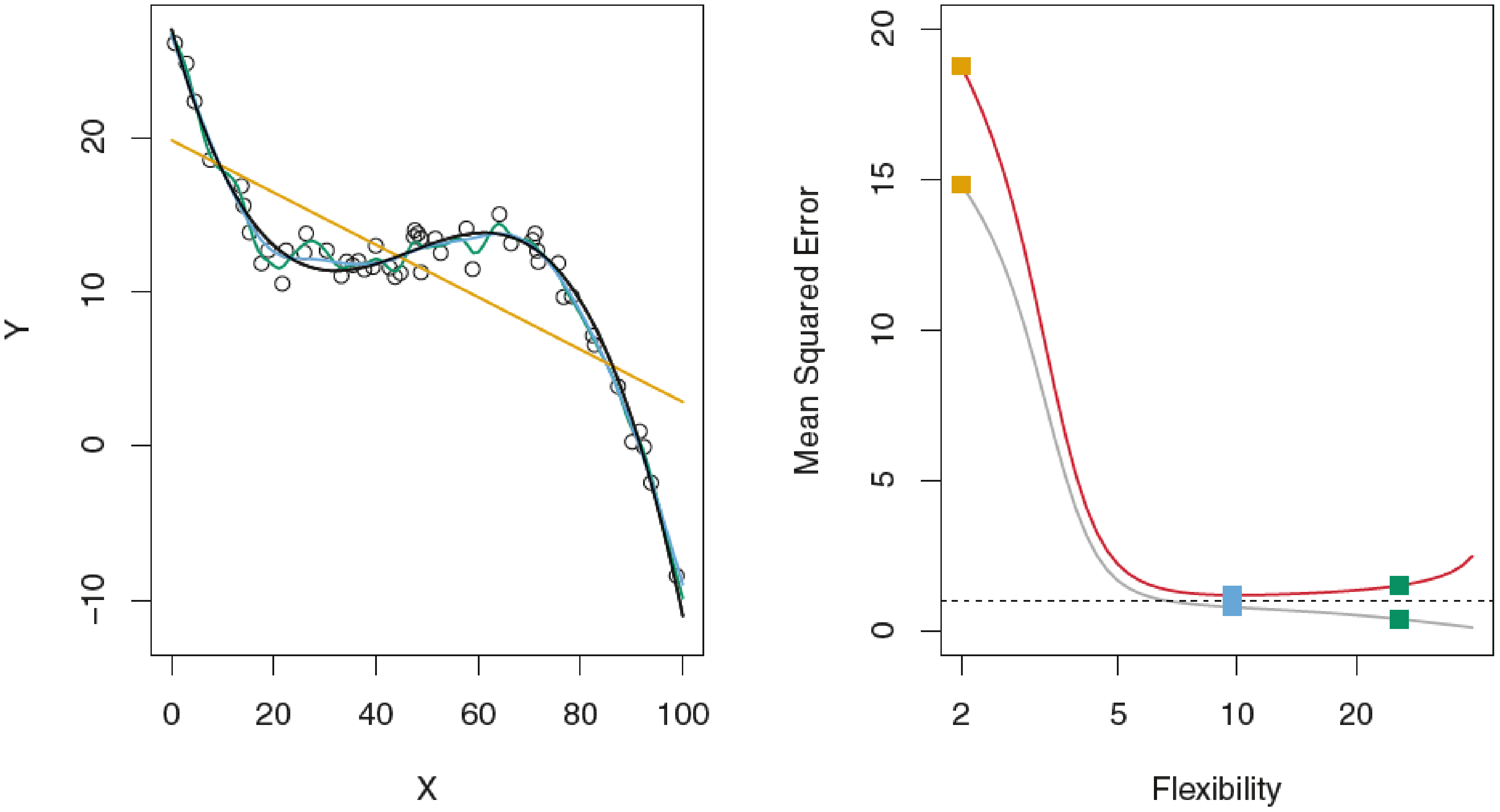
$$\mathrm{Ave}(\hat{f}(x_0) - y_0)^2$$



Y
X
0
20
40
60
80
100
2
4
6
8
10
12
Mean Squared Error
Flexibility
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20



Y
X
Mean Squared Error
Flexibility



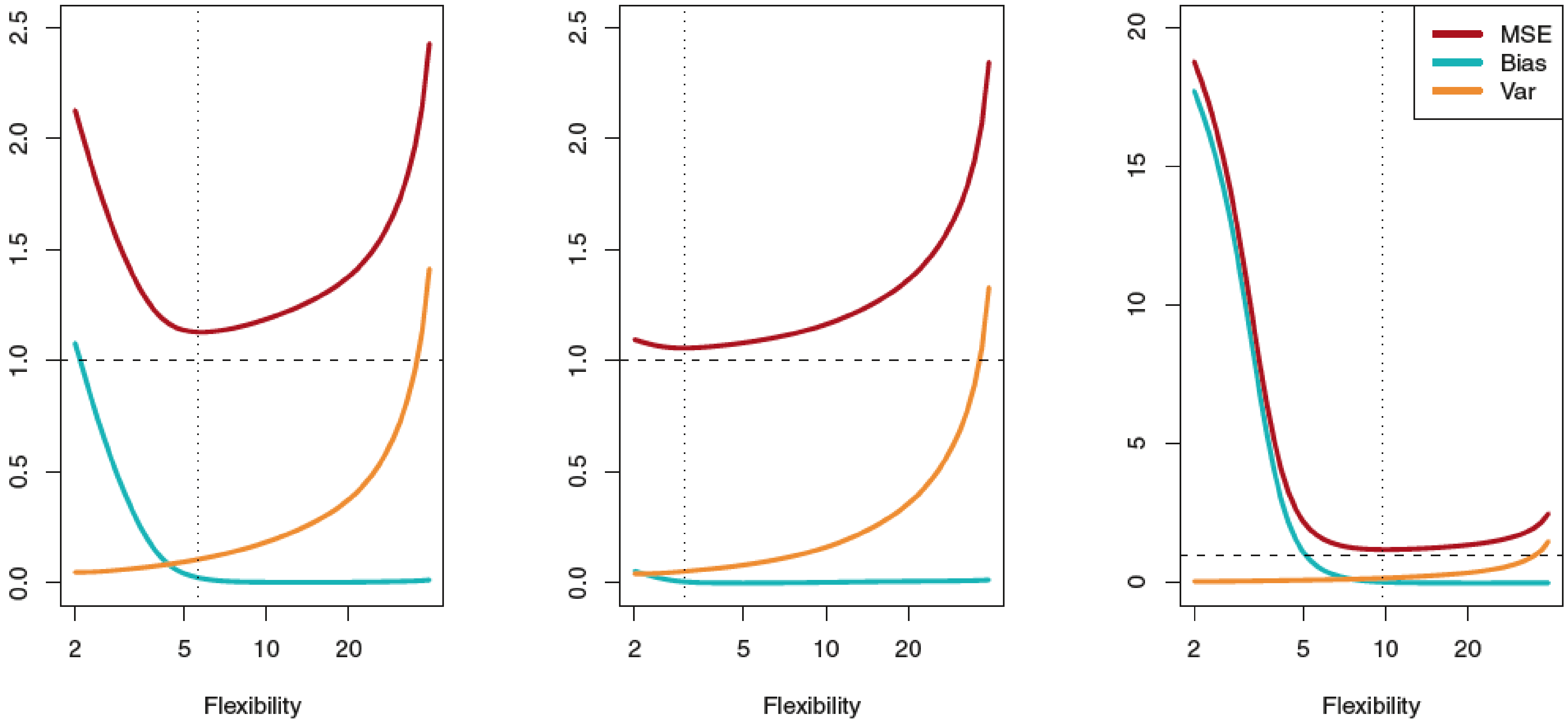
Y
X
0
20
40
60
80
100
−10
0
10
20
Mean Squared Error
Flexibility
2
5
10
20
0
5
10
15
20

Компроміс між зміщенням та дисперсією.

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

Компроміс між зміщенням та дисперсією.

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$



Класифікація.

Класифікація.

Нехай нам задано n даних, тобто пар $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, але тепер $y_1, y_2, \ldots, y_n$ є якісними.

Класифікація.

Нехай нам задано n даних, тобто пар $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, але тепер $y_1, y_2, \ldots, y_n$ є якісними.

Тренувальна помилка

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

Класифікація.

Нехай нам задано n даних, тобто пар $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, але тепер $y_1, y_2, \ldots, y_n$ є якісними.

Тренувальна помилка

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

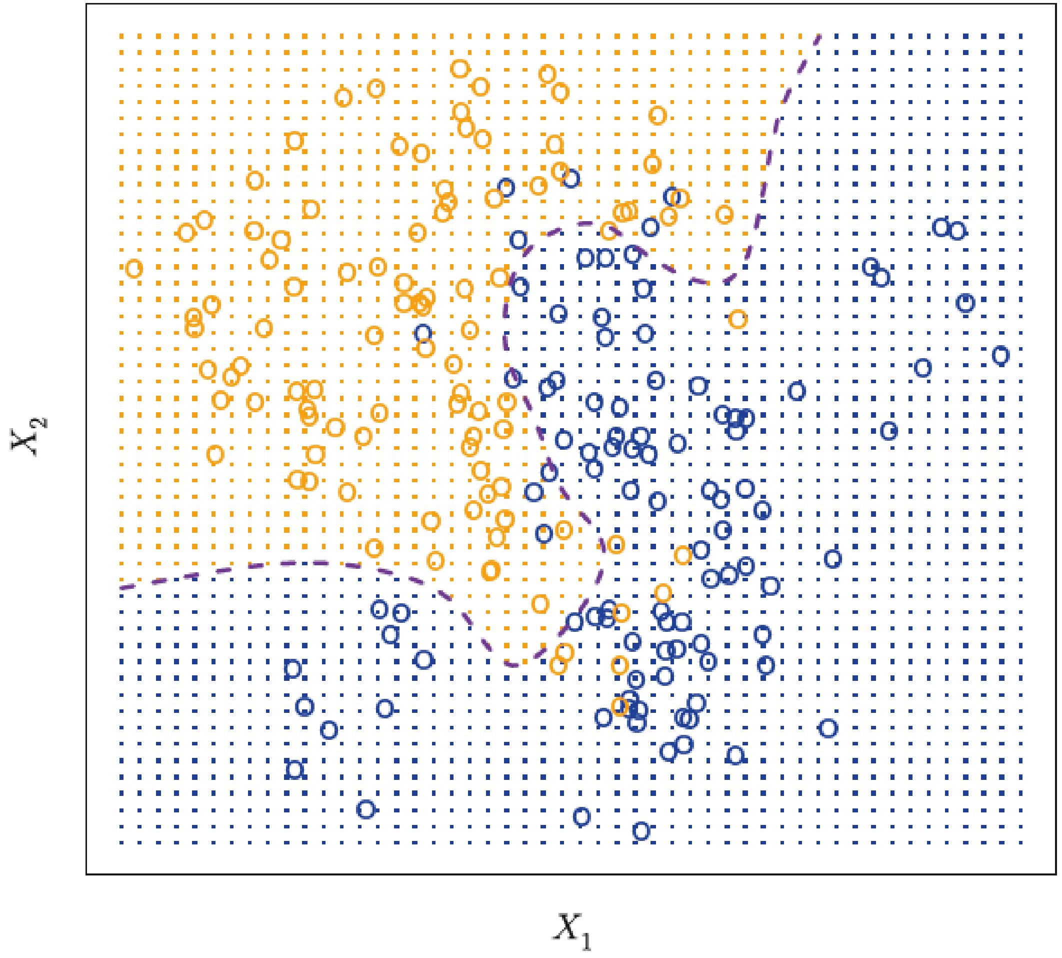
Тестова помилка

$$\operatorname{Ave}\left(I(y_0 \neq \hat{y}_0)\right)$$

Класифікатор Байєса.

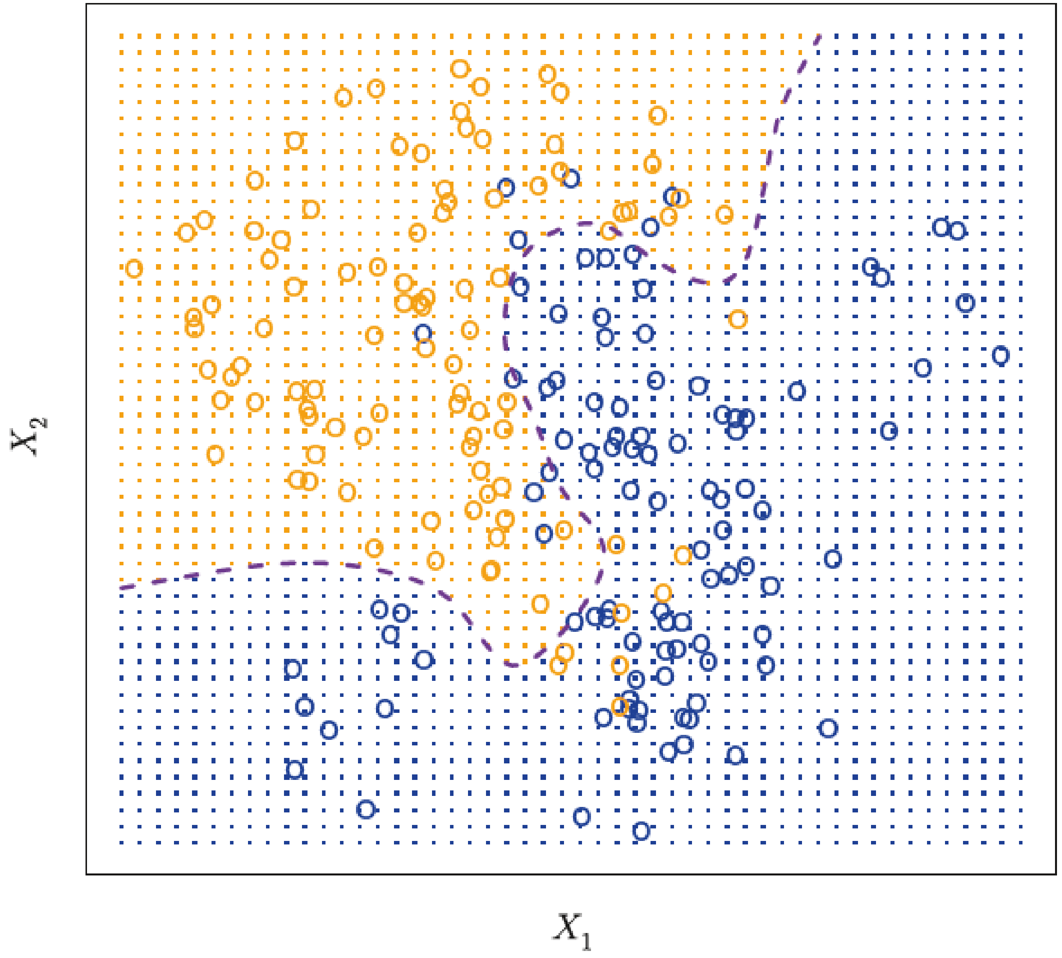
Класифікатор Байєса.

$$\Pr(Y = j | X = x_0)$$



Класифікатор Байєса.

$$\Pr(Y = j | X = x_0)$$



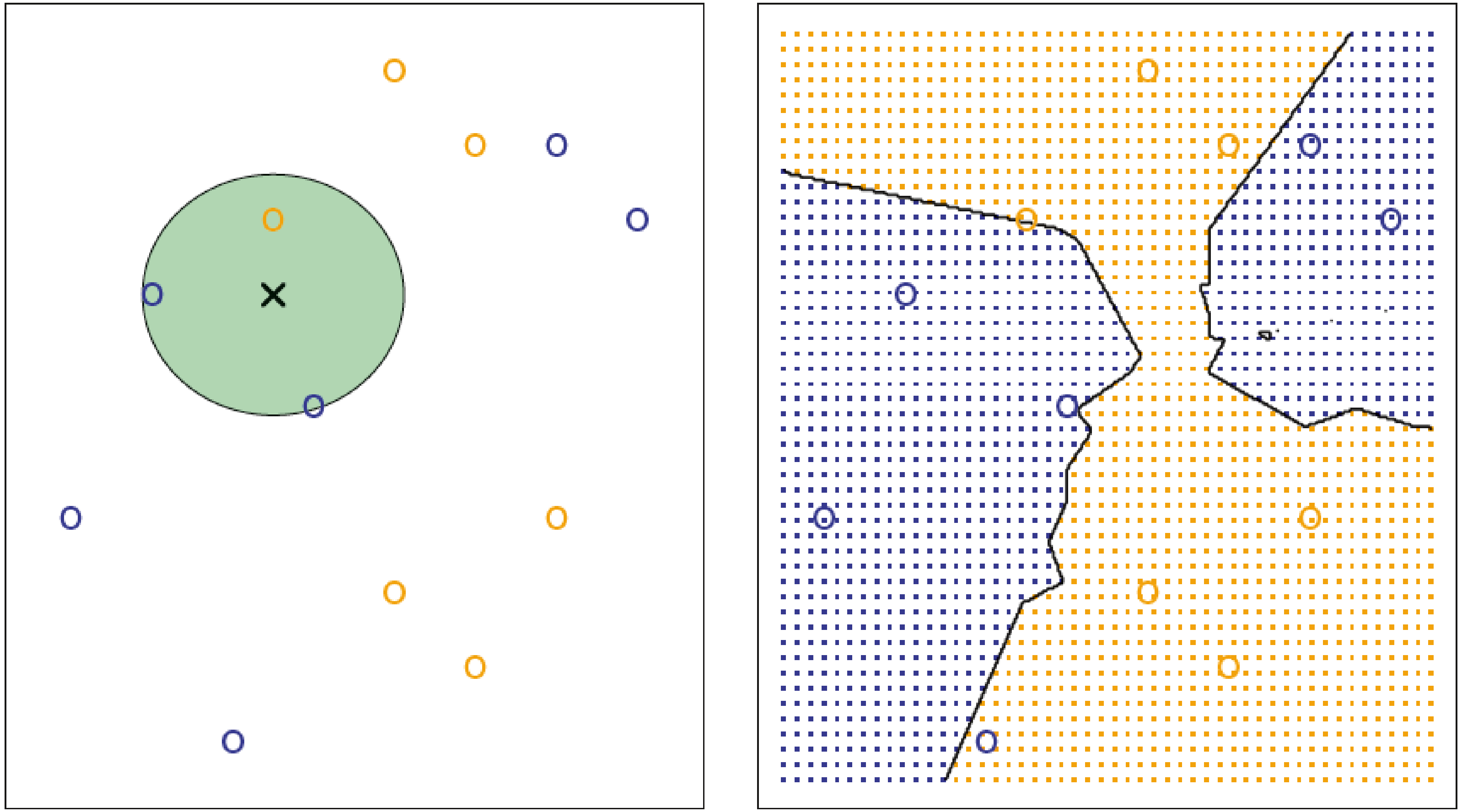
$$1 - E\left(\max_j \Pr(Y = j | X)\right)$$

К-найближчих сусідів.

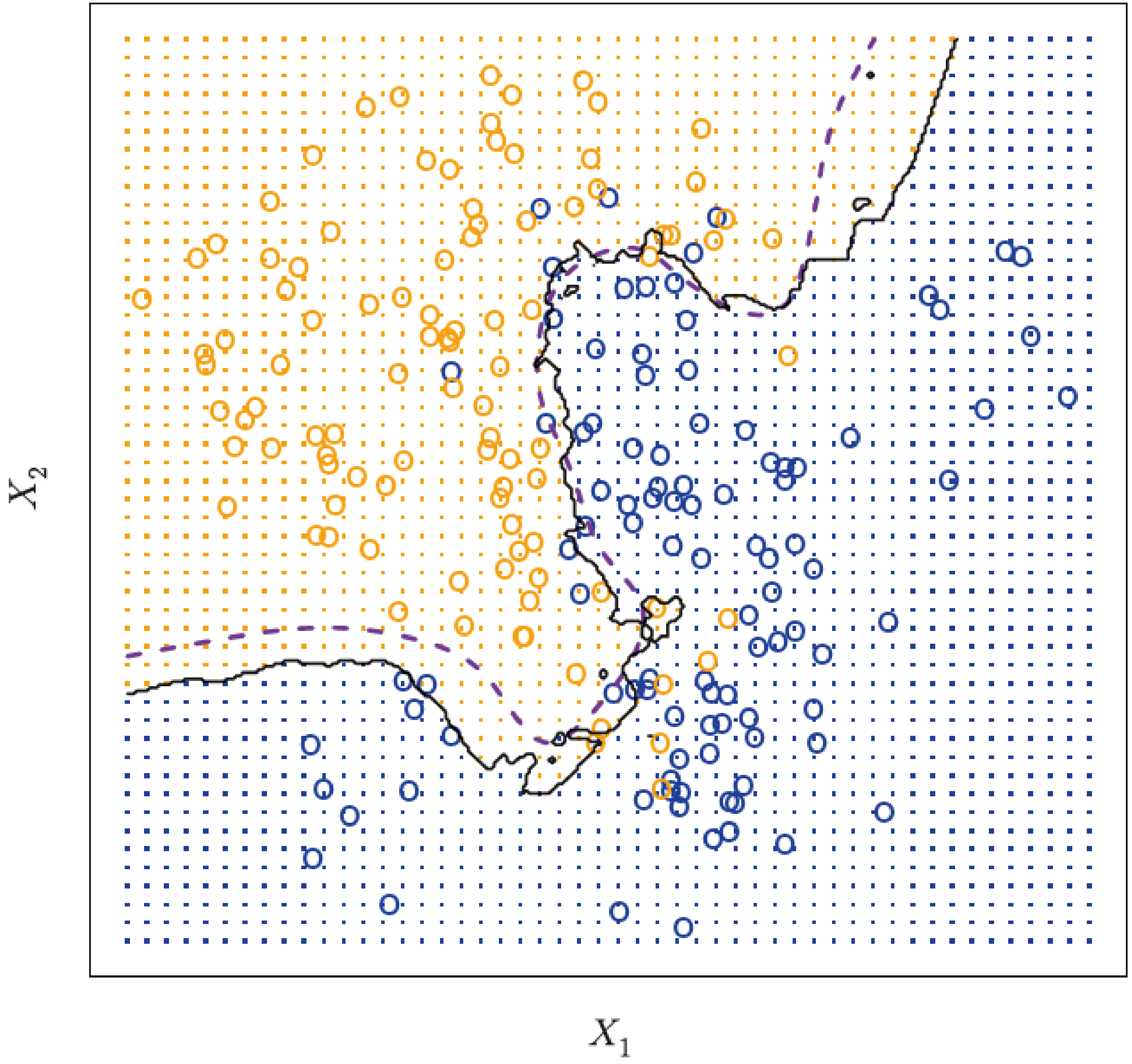
$$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

K-найближчих сусідів.

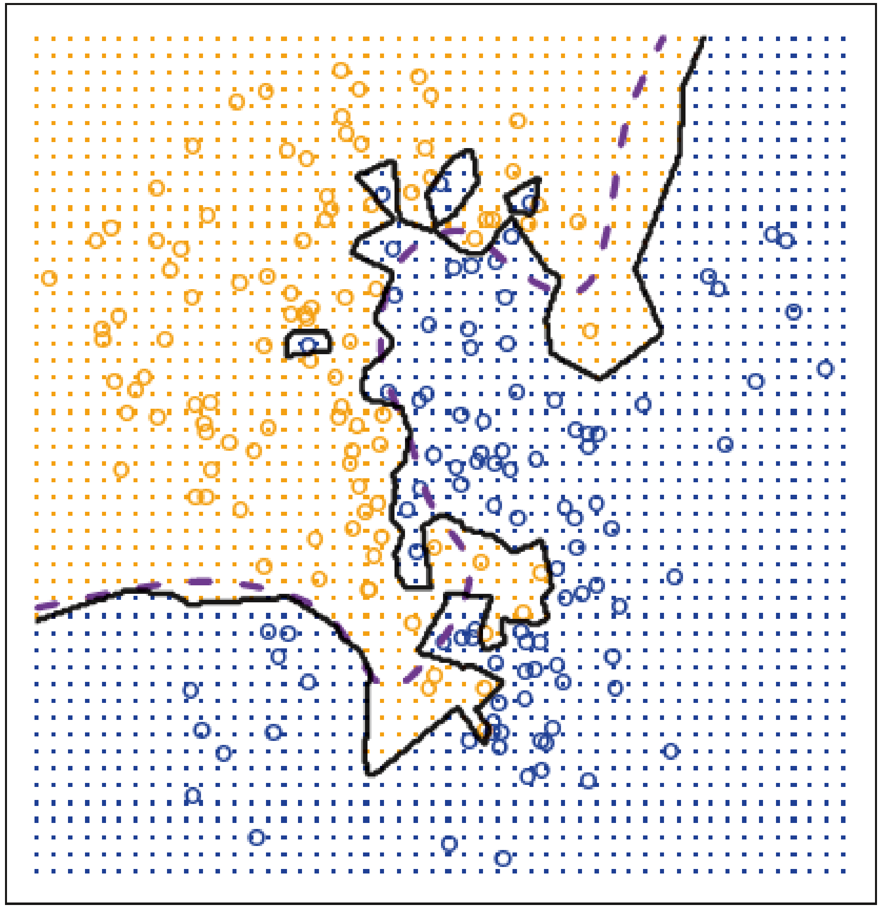
$$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$



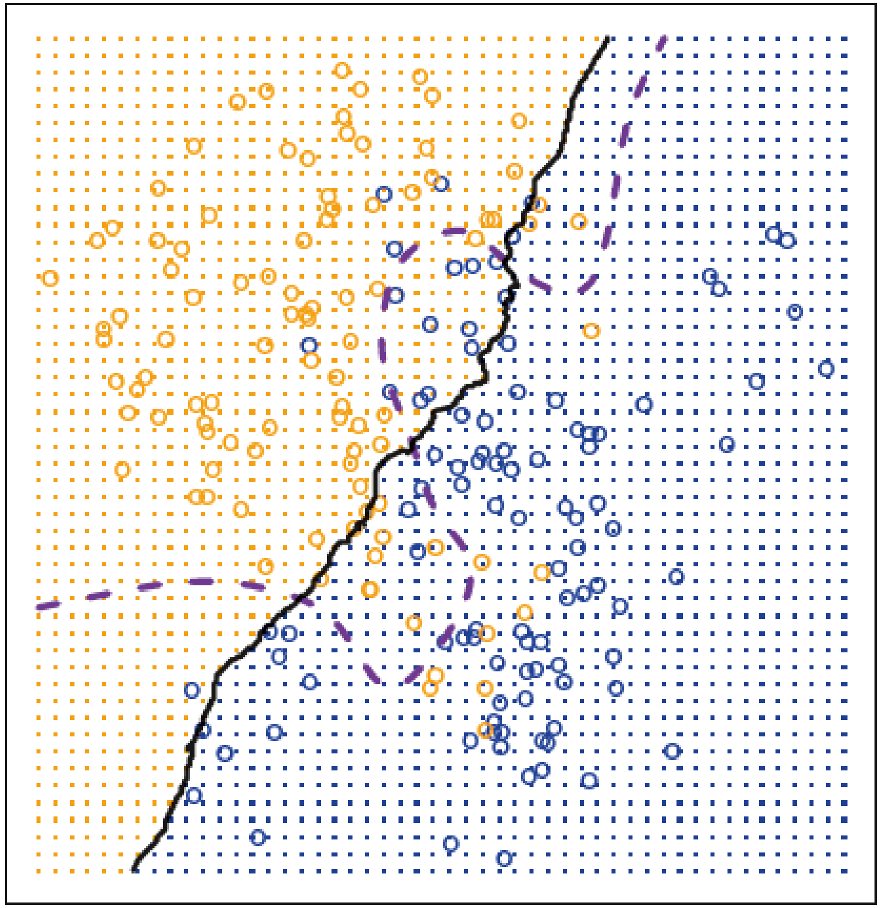
KNN: K=10

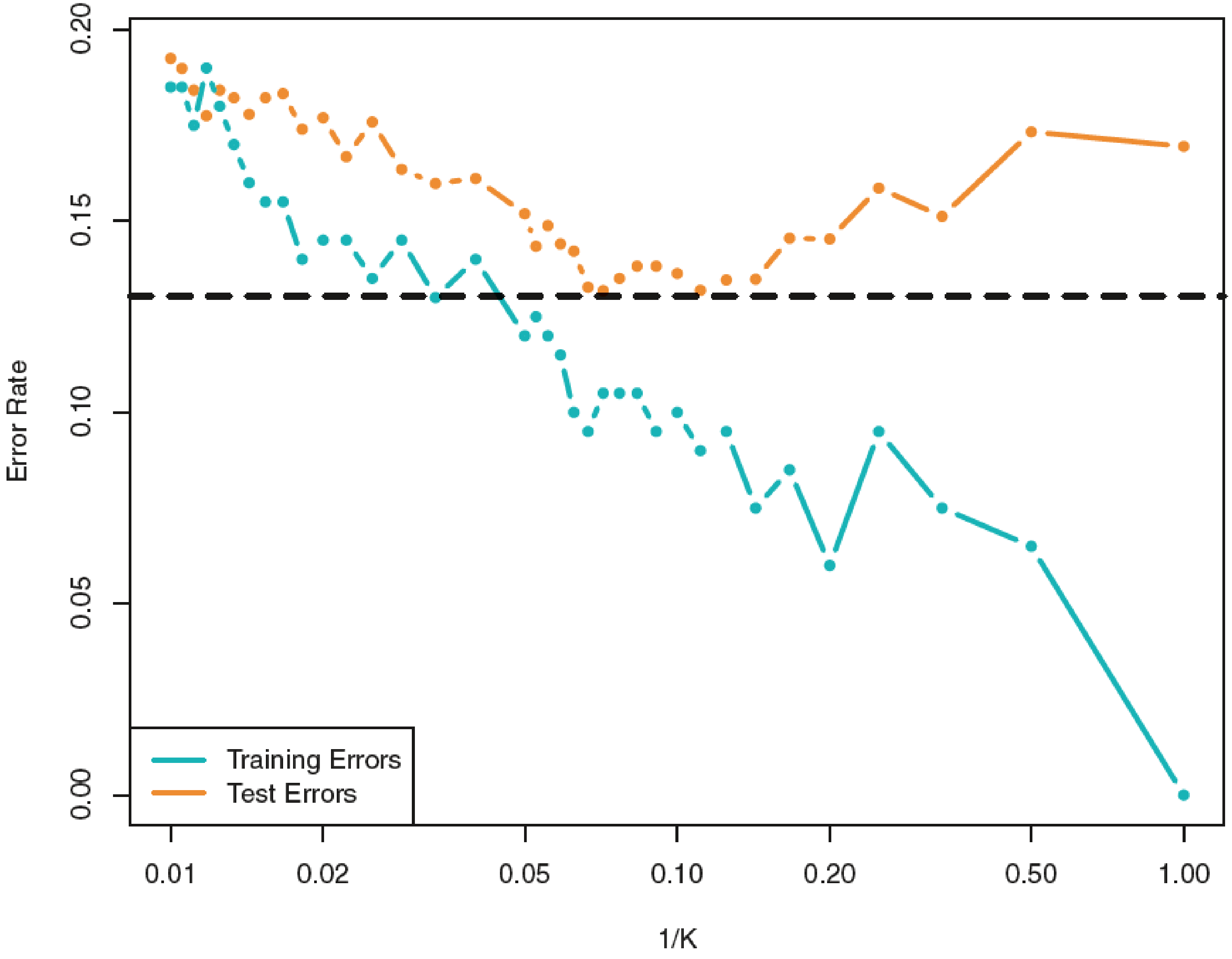


KNN: K=1



KNN: K=100




Error Rate
0.00
0.05
0.10
0.15
0.20
0.01
0.02
0.05
0.10
0.20
0.50
1.00
1/K
Training Errors
Test Errors