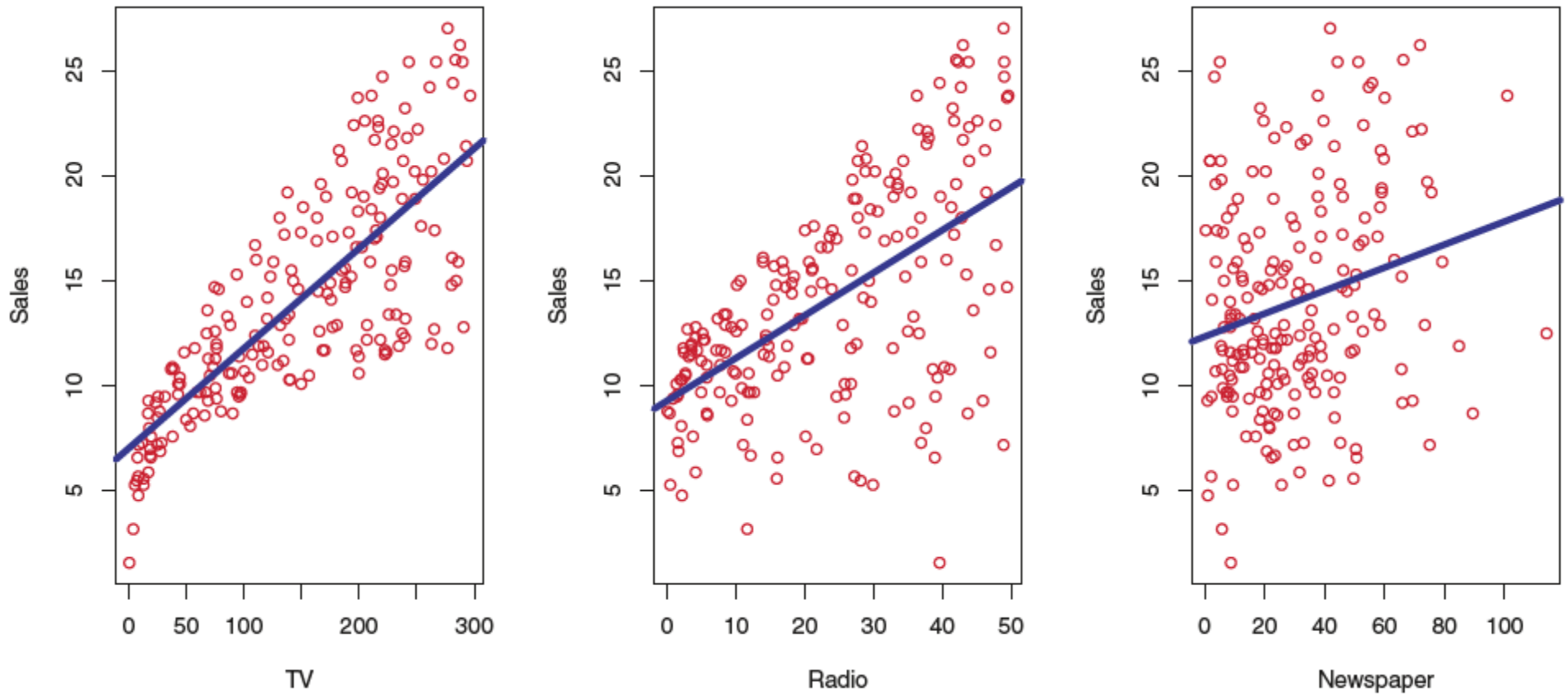
Моделі статистичного навчання: лінійна регресія


Sales
5
10
15
20
25
0
50
100
200
300
TV
Sales
5
10
15
20
25
0
10
20
30
40
50
Radio
Sales
5
10
15
20
25
0
20
40
60
80
100
Newspaper

1. Чи існує взаємозв'язок між рекламним бюджетом та продажами?

2. Наскільки сильний взаємозв'язок між рекламним бюджетом та продажами?

3. Які засоби масової інформації сприяють продажам?

4. Наскільки точно ми можемо оцінити вплив кожного засобу на продажі?

5. Наскільки точно ми можемо передбачити майбутні продажі?

6. Чи взаємозв’язки лінійні?

7. Чи існує взаємодія серед рекламних засобів?

Проста лінійна регресія

Проста лінійна регресія

Маємо

$$Y \approx \beta_0 + \beta_1 X$$

Проста лінійна регресія

Маємо

$$Y \approx \beta_0 + \beta_1 X$$

На практиці β_0 та β_1 невідомі.

Проста лінійна регресія

Маємо

$$Y \approx \beta_0 + \beta_1 X$$

На практиці β_0 та β_1 невідомі.

Нехай $\hat{\beta}_0$ та $\hat{\beta}_1$ оцінки параметрів, тоді

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Проста лінійна регресія

Маємо

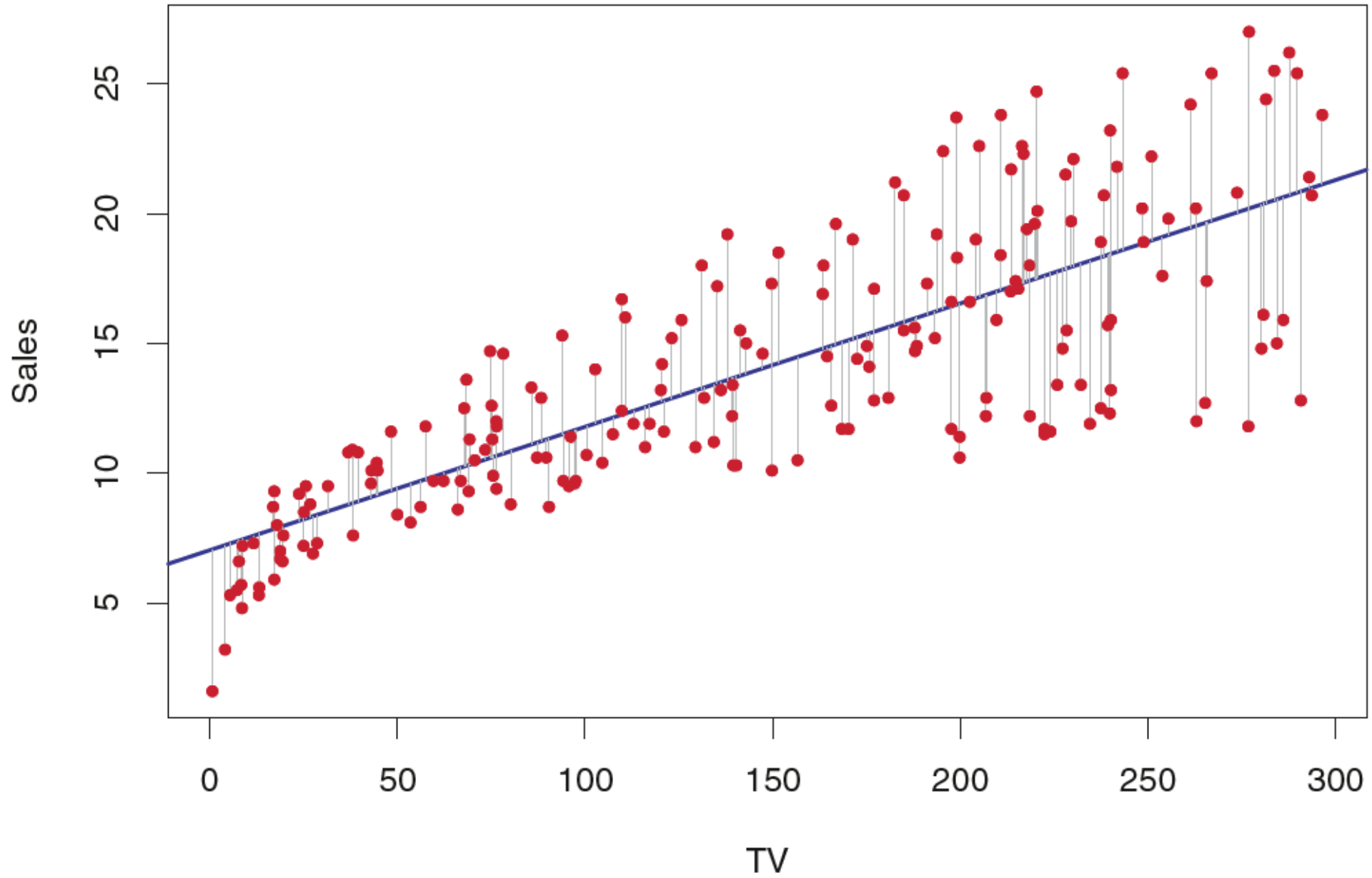
$$Y \approx \beta_0 + \beta_1 X$$

На практиці β_0 та β_1 невідомі.

Нехай $\hat{\beta}_0$ та $\hat{\beta}_1$ оцінки параметрів, тоді

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Нехай нам задано n даних, тобто пар (x_1, y_1), (x_2, y_2), …, (x_n, y_n), ми хочемо отримати такі оцінки, щоб відхилення на множині цих даних було якомога менше.


Sales
5
10
15
20
25
0
50
100
150
200
250
300
TV

Нехай $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, тоді $e_i = y_i - \hat{y}_i$

Нехай $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, тоді $e_i = y_i - \hat{y}_i$

Визначимо величину відхилення як

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

Нехай $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, тоді $e_i = y_i - \hat{y}_i$

Визначимо величину відхилення як

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

Або

$$RSS = \left(\hat{y}_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(\hat{y}_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \ldots + \left(\hat{y}_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2$$

Нехай $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, тоді $e_i = y_i - \hat{y}_i$

Визначимо величину відхилення як

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

Або

$$RSS = \left(\hat{y}_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(\hat{y}_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \ldots + \left(\hat{y}_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2$$

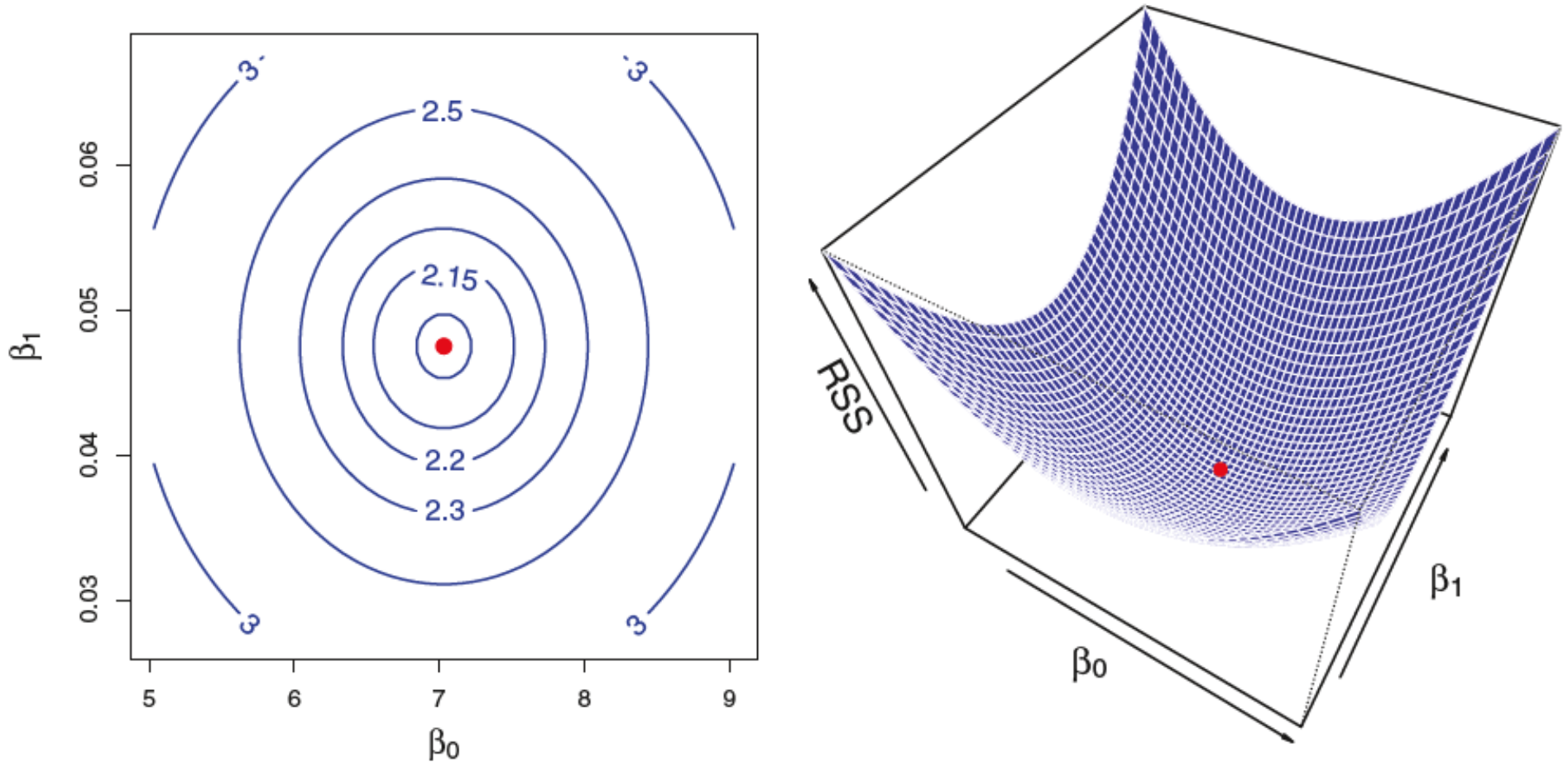
Мінімізуючи *RSS* відносно $\hat{\beta}_0$ та $\hat{\beta}_1$ отримаємо

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Для телебачення отримаємо: $\hat{\beta}_0 = 7.03$ та $\hat{\beta}_1 = 0.0475$

Для телебачення отримаємо: $\hat{\beta}_0 = 7.03$ та $\hat{\beta}_1 = 0.0475$



Оцінка точності оцінок

Оцінка точності оцінок

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Оцінка точності оцінок

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Розглянемо приклад

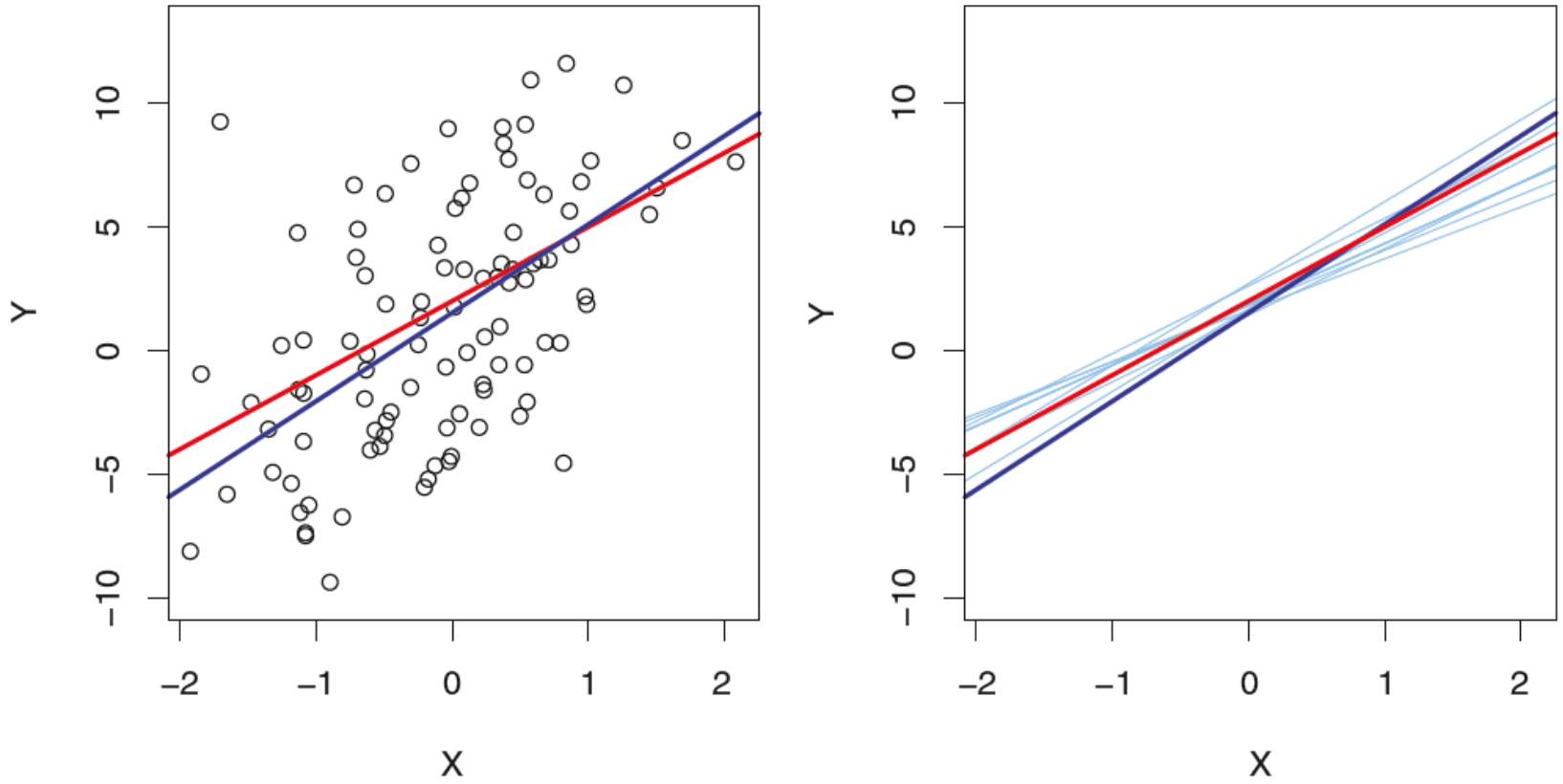
$$Y = 2 + 3X + \varepsilon$$

Оцінка точності оцінок

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Розглянемо приклад

$$Y = 2 + 3X + \varepsilon$$



$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Причому

$$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-2}}$$

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Причому

$$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-2}}$$

(1 – α) інтервали довіри для коефіцієнтів

$$\left[\hat{\beta}_0 - t_{n-2;1-\alpha/2} SE(\hat{\beta}_0); \hat{\beta}_0 + t_{n-2;1-\alpha/2} SE(\hat{\beta}_0)\right]$$

$$\left[\hat{\beta}_1 - t_{n-2;1-\alpha/2} SE(\hat{\beta}_1); \hat{\beta}_1 + t_{n-2;1-\alpha/2} SE(\hat{\beta}_1)\right]$$

Для телебачення отримаємо:

Для телебачення отримаємо:

95% інтервал довіри для:

β_0: [6.130; 7.935]

β_1: [0.042; 0.053]

Для телебачення отримаємо:

95% інтервал довіри для:

β_0: [6.130; 7.935]

β_1: [0.042; 0.053]

Гіпотеза:

H_0: немає ніякого зв’язку між Y і X

vs

H_1: існує певний зв’язок між Y і X

$$H_0: \beta_1 = 0 \text{ vs } H_1: \beta_1 \neq 0$$

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

$$H_0: \beta_1 = 0 \text{ vs } H_1: \beta_1 \neq 0$$

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

p-значення дорівнює імовірності отримати значення тестової статистики більшого за отримане за умови виконання нульової гіпотези. Чим меншим є *p*-значення тим більше підстав відхилити нульову гіпотезу.

$$H_0: \beta_1 = 0 \text{ vs } H_1: \beta_1 \neq 0$$

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

p-значення дорівнює імовірності отримати значення тестової статистики більшого за отримане за умови виконання нульової гіпотези. Чим меншим є *p*-значення тим більше підстав відхилити нульову гіпотезу.

Інший спосіб перевірки гіпотез використання інтервалів довіри!!!

$$H_0: \beta_1 = 0 \text{ vs } H_1: \beta_1 \neq 0$$

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}$$

p-значення дорівнює імовірності отримати значення тестової статистики більшого за отримане за умови виконання нульової гіпотези. Чим меншим є *p*-значення тим більше підстав відхилити нульову гіпотезу.

Інший спосіб перевірки гіпотез використання інтервалів довіри!!!

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Оцінка якості моделі

Оцінка якості моделі

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Оцінка якості моделі

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}\ ,\ \text{де}\quad \mathrm{TSS} = \sum(y_i - \bar{y})^2$$

Багатовимірна регресія

Багатовимірна регресія

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	< 0.0001

Багатовимірна регресія

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	< 0.0001

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

В матричній формі: $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

Оцінка параметрів

Оцінка параметрів

Нехай $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$,

тоді

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip} \right)^2 = \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right)' \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right)$$

Оцінка параметрів

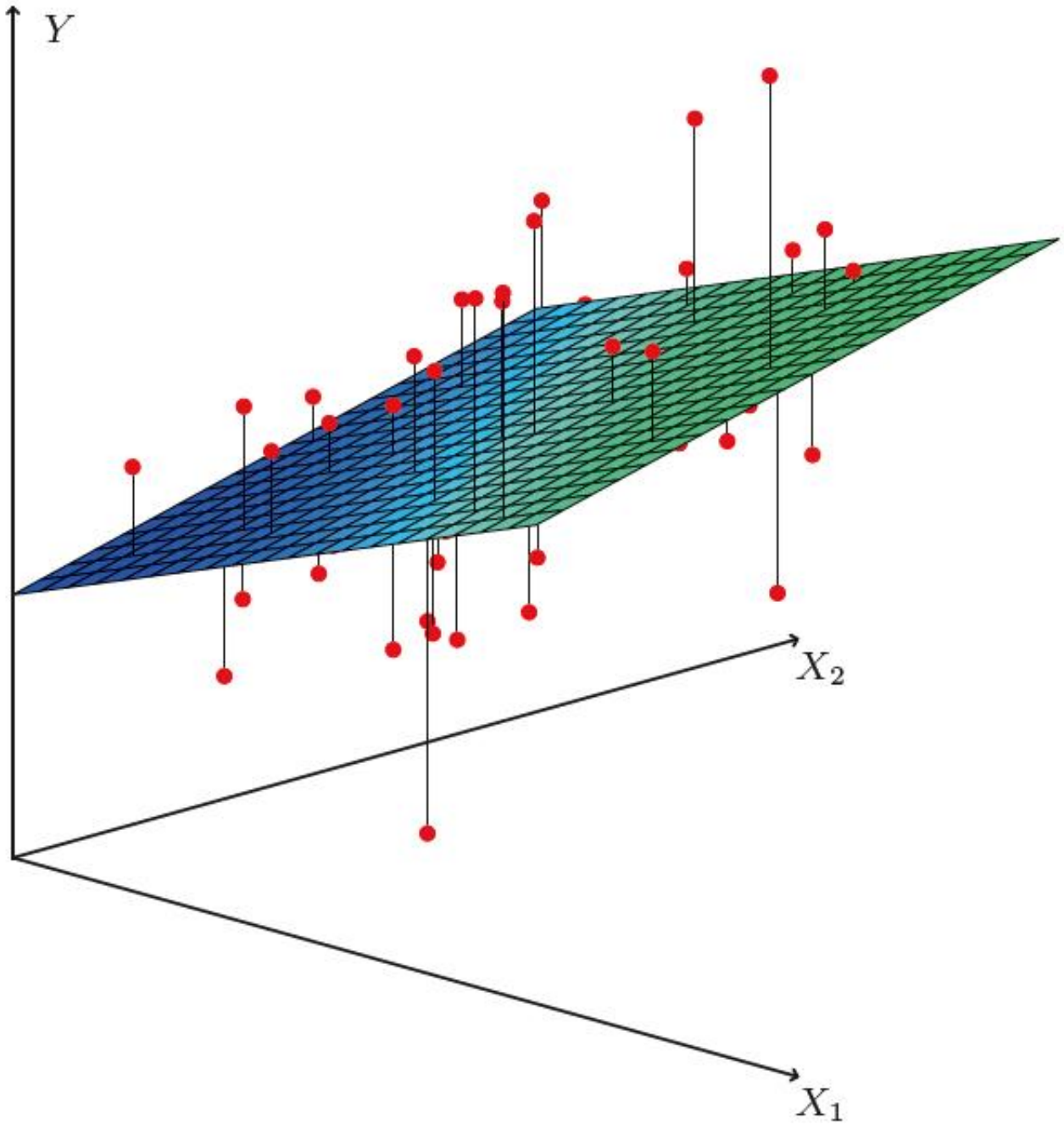
Нехай $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$,

тоді

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip} \right)^2 = \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right)' \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right)$$

Мінімізуючи *RSS* відносно $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Y}$$


Y
X2
X1

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Чи існує взаємозв'язок між **Y** та **X**?

Чи існує взаємозв'язок між **Y** та **X**?

$$H_0: \beta_0 = \beta_1 = \ldots = \beta_p = 0 \quad \text{vs } H_1: \text{існує } j, \text{ що } \beta_j \neq 0$$

Чи існує взаємозв'язок між **Y** та **X**?

$H_0: \beta_0 = \beta_1 = \ldots = \beta_p = 0$ vs H_1: існує j, що $\beta_j \neq 0$

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

$$E\{\mathrm{RSS}/(n-p-1)\} = \sigma^2$$

$$E\{(\mathrm{TSS} - \mathrm{RSS})/p\} = \sigma^2$$

Чи існує взаємозв'язок між **Y** та **X**?

$$H_0: \beta_0 = \beta_1 = \ldots = \beta_p = 0 \quad \text{vs } H_1: \text{існує } j, \text{ що } \beta_j \neq 0$$

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

$$E\{\mathrm{RSS}/(n-p-1)\} = \sigma^2$$

$$E\{(\mathrm{TSS} - \mathrm{RSS})/p\} = \sigma^2$$

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

Чи існує взаємозв'язок між **Y** та підмножиною **X**?

Чи існує взаємозв'язок між **Y** та підмножиною **X**?

$$H_0: \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0 \text{ vs } H_1: \text{існує } j, \text{ що } \beta_j \neq 0$$

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}$$

Чи існує взаємозв'язок між **Y** та підмножиною **X**?

$$H_0: \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0 \text{ vs } H_1: \text{існує } j, \text{ що } \beta_j \neq 0$$

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}$$

Тут RSS_0 RSS моделі без відповідних змінних (у нашому випадку X_{p-q+1}, X_{p-q+2}, X_p

Визначення важливих змінних

Визначення важливих змінних

Вибір вперед (forward selection). Ми починаємо з нульової моделі. Потім ми оцінюємо p простих лінійних регресій і додаємо до нульової моделі змінну, яка призводить до найнижчого RSS. Потім ми додаємо до цієї моделі змінну, яка дає найнижчий RSS для моделі із двома змінними.

Визначення важливих змінних

Вибір вперед (forward selection). Ми починаємо з нульової моделі. Потім ми оцінюємо p простих лінійних регресій і додаємо до нульової моделі змінну, яка призводить до найнижчого RSS. Потім ми додаємо до цієї моделі змінну, яка дає найнижчий RSS для моделі із двома змінними.

• Вибір назад (backward selection). Ми починаємо з усіх змінних у моделі, і видаляємо змінну з найбільшим p-значенням. Аналогічні дії повторяємо для моделей з p - 1, p - 2 ... змінними.

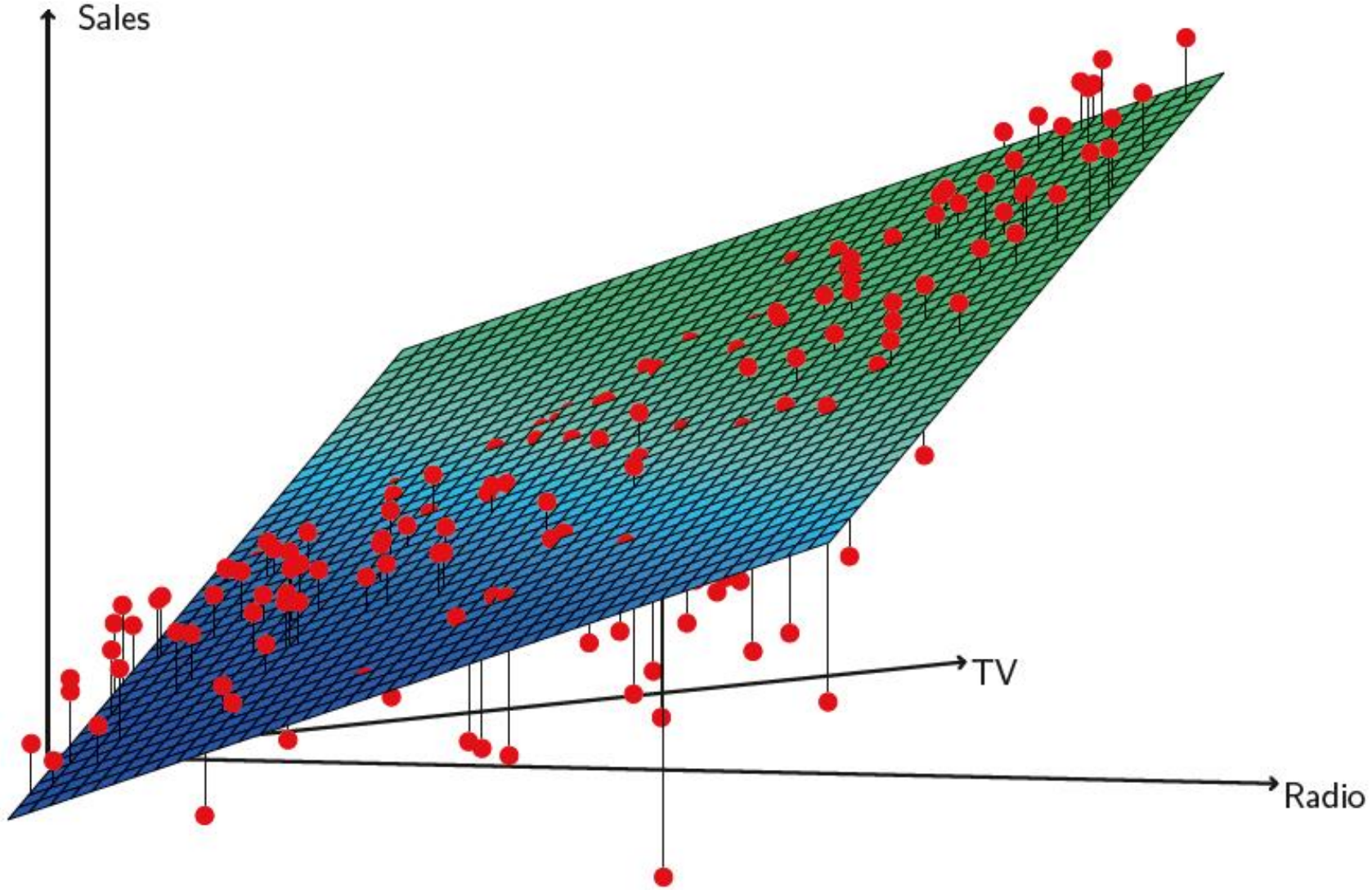
Визначення важливих змінних

Вибір вперед (forward selection). Ми починаємо з нульової моделі. Потім ми оцінюємо p простих лінійних регресій і додаємо до нульової моделі змінну, яка призводить до найнижчого RSS. Потім ми додаємо до цієї моделі змінну, яка дає найнижчий RSS для моделі із двома змінними.

• Вибір назад (backward selection). Ми починаємо з усіх змінних у моделі, і видаляємо змінну з найбільшим p-значенням. Аналогічні дії повторяємо для моделей з p - 1, p - 2 ... змінними.

• Змішаний вибір (mixed selection). Ми починаємо з того, що в моделі немає змінних, і додаємо змінну, яка забезпечує найкращу відповідність. Ми продовжуємо додавати змінні по одній. Якщо в будь-який момент p-значення для однієї зі змінних у моделі перевищує певний поріг, тоді ми видаляємо цю змінну з моделі.

Придатність моделі



Передбачення

Передбачення

Три типи невизначеності:

Передбачення

Три типи невизначеності:

1. Оскільки ми користуємося оцінками, які є випадковими величинами, то і результати, що ми отримуємо є теж випадковими величинами. Дивлячись з точки зору помилки, що можна зменшити, ми маємо можливість побудувати інтервал довіри, щоб оцінити, як близько оцінка $\hat{Y}$ знаходиться від значення

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

Передбачення

Три типи невизначеності:

1. Оскільки ми користуємося оцінками, які є випадковими величинами, то і результати, що ми отримуємо є теж випадковими величинами. Дивлячись з точки зору помилки, що можна зменшити, ми маємо можливість побудувати інтервал довіри, щоб оцінити, як близько оцінка $\hat{Y}$ знаходиться від значення

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

2. Зміщення моделі. Ми використовуємо насправді найкраще лінійне наближення до точної моделі.

Передбачення

Три типи невизначеності:

1. Оскільки ми користуємося оцінками, які є випадковими величинами, то і результати, що ми отримуємо є теж випадковими величинами. Дивлячись з точки зору помилки, що можна зменшити, ми маємо можливість побудувати інтервал довіри, щоб оцінити, як близько оцінка $\hat{Y}$ знаходиться від значення

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

2. Зміщення моделі. Ми використовуємо насправді найкраще лінійне наближення до точної моделі.

3. Врахування залишків моделі. В цьому випадку в нагоді може стати інтервал передбачення.

Інтервал довіри

$$\left[\hat{Y} - t_{1-\alpha/2;n-p-1} RSE \sqrt{\mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_0}; \hat{Y} + t_{1-\alpha/2;n-p-1} RSE \sqrt{\mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_0}\right], \mathbf{x}_0 = \left(1, x_{01}, x_{02}, ..., x_{0p}\right), \mathbf{X} = \begin{pmatrix} 1 & x_{11} & ... & x_{p1} \\ 1 & x_{12} & ... & x_{p2} \\ ... & ... & ... & ... \\ 1 & x_{1p} & ... & x_{pp} \end{pmatrix}$$

Інтервал довіри

$$\left[\hat{Y} - t_{1-\alpha/2;n-p-1} RSE \sqrt{\mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_0}; \hat{Y} + t_{1-\alpha/2;n-p-1} RSE \sqrt{\mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_0}\right], \mathbf{x}_0 = \left(1, x_{01}, x_{02}, ..., x_{0p}\right), \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1p} & \dots & x_{pp} \end{pmatrix}$$

При $p = 1$:

$$\left[\hat{Y} - t_{1-\alpha/2;n-2} RSE \sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}; \hat{Y} + t_{1-\alpha/2;n-2} RSE \sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}\right].$$

Інтервал довіри

$$\left[\hat{Y} - t_{1-\alpha/2;n-p-1} RSE \sqrt{\mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_0}; \hat{Y} + t_{1-\alpha/2;n-p-1} RSE \sqrt{\mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_0}\right], \mathbf{x}_0 = \left(1, x_{01}, x_{02}, ..., x_{0p}\right), \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1p} & \dots & x_{pp} \end{pmatrix}$$

При $p = 1$:

$$\left[\hat{Y} - t_{1-\alpha/2;n-2} RSE \sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}; \hat{Y} + t_{1-\alpha/2;n-2} RSE \sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}\right].$$

Інтервал передбачення

$$\left[\hat{Y} - t_{1-\alpha/2;n-p-1} RSE \sqrt{1 + \mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_0}; \hat{Y} + t_{1-\alpha/2;n-p-1} RSE \sqrt{1 + \mathbf{x}_0^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_0}\right].$$

Інтервал довіри

$$\left[\hat{Y}-t_{1-\alpha/2;n-p-1}RSE\sqrt{\mathbf{x}_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0};\ \hat{Y}+t_{1-\alpha/2;n-p-1}RSE\sqrt{\mathbf{x}_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0}\right],\ \mathbf{x}_0=\left(1,x_{01},x_{02},...,x_{0p}\right),\ \mathbf{X}=\begin{pmatrix}1 & x_{11} & \dots & x_{p1}\\ 1 & x_{12} & \dots & x_{p2}\\ \dots & \dots & \dots & \dots\\ 1 & x_{1p} & \dots & x_{pp}\end{pmatrix}$$

$$\text{При } p=1:\left[\hat{Y}-t_{1-\alpha/2;n-2}RSE\sqrt{\frac{1}{n}+\frac{\left(x_0-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}};\ \hat{Y}+t_{1-\alpha/2;n-2}RSE\sqrt{\frac{1}{n}+\frac{\left(x_0-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}}\right].$$

Інтервал передбачення

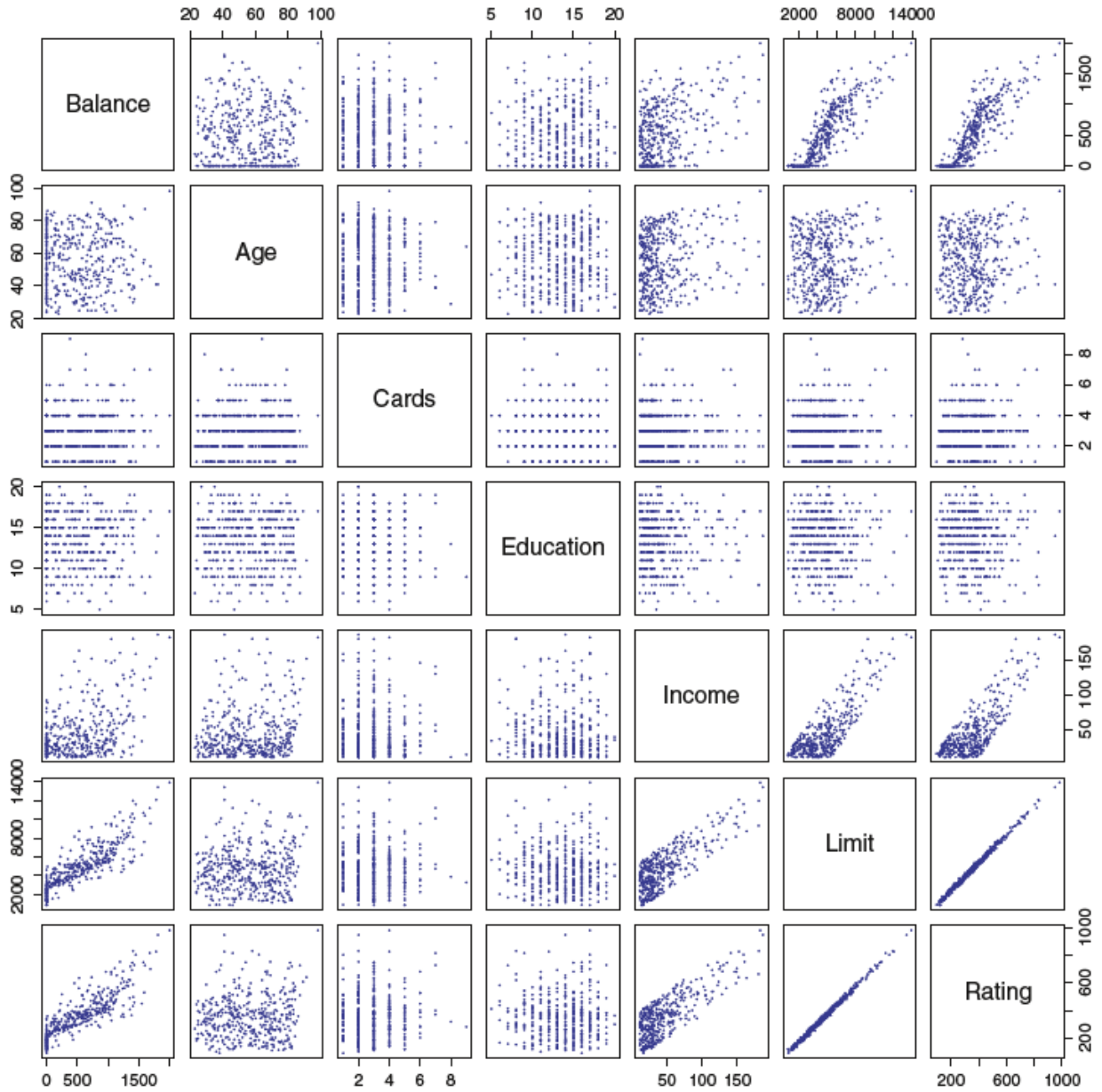
$$\left[\hat{Y}-t_{1-\alpha/2;n-p-1}RSE\sqrt{1+\mathbf{x}_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0};\ \hat{Y}+t_{1-\alpha/2;n-p-1}RSE\sqrt{1+\mathbf{x}_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0}\right].$$

Для нашої моделі:

95% інтервал довіри при витратах на телебачення, що дорівнюють 100000 та витратах на радіо – 20000 матиме вигляд [10,985, 11,528], натомість 95 % інтервал передбачення матиме вигляд [7,930, 14,580].

Якісні змінні

Якісні змінні



Якісні змінні з двома рівнями

Якісні змінні з двома рівнями

Вводимо фіктивні змінні, визначені правилом:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

Якісні змінні з двома рівнями

Вводимо фіктивні змінні, визначені правилом:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

Модель набуде вигляду

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Якісні змінні з двома рівнями

Вводимо фіктивні змінні, визначені правилом:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

Модель набуде вигляду

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

Якісні змінні з більш ніж двома рівнями

Якісні змінні з більш ніж двома рівнями

Розглянемо змінну ethnicity: азіат, кавказець, афроамериканець

Якісні змінні з більш ніж двома рівнями

Розглянемо змінну ethnicity: азіат, кавказець, афроамериканець

Вводимо дві фіктивні змінні, визначені правилом:

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

Якісні змінні з більш ніж двома рівнями

Розглянемо змінну ethnicity: азіат, кавказець, афроамериканець

Вводимо дві фіктивні змінні, визначені правилом:

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

Модель набуде вигляду

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	-18.69	65.02	-0.287	0.7740
`ethnicity[Caucasian]`	-12.50	56.68	-0.221	0.8260

Розширення лінійної моделі.

Розширення лінійної моделі.

Два важливих неявних припущення: адитивність і лінійність.

Розширення лінійної моделі.

Два важливих неявних припущення: адитивність і лінійність.

Усунення адитивності (врахування ефекту взаємодії)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Розширення лінійної моделі.

Два важливих неявних припущення: адитивність і лінійність.

Усунення адитивності (врахування ефекту взаємодії)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

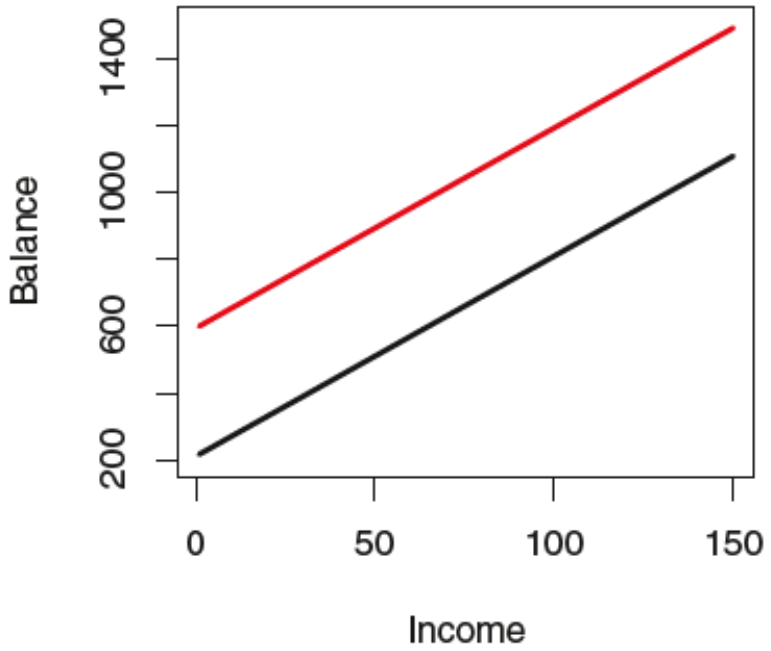
	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Аналогічно врахувати ефект взаємодії можливо і для якісних змінних.

$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}
\end{aligned}$$

Аналогічно врахувати ефект взаємодії можливо і для якісних змінних.

$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}
\end{aligned}$$

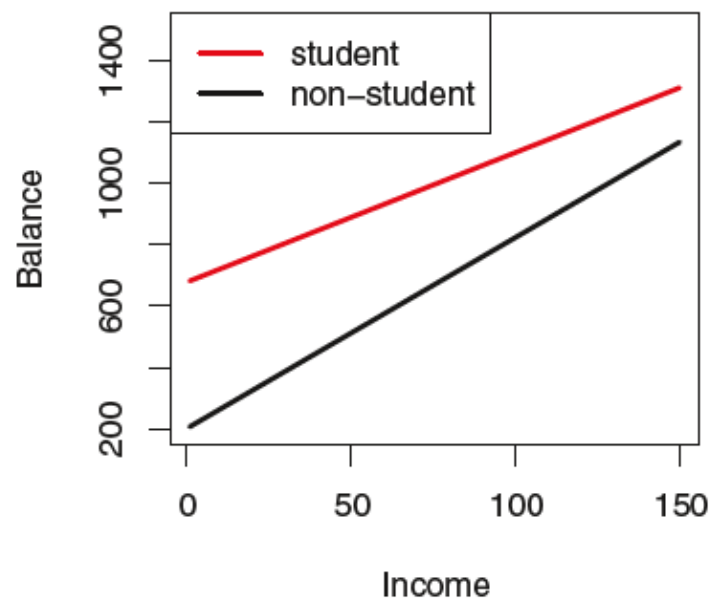


$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}$$

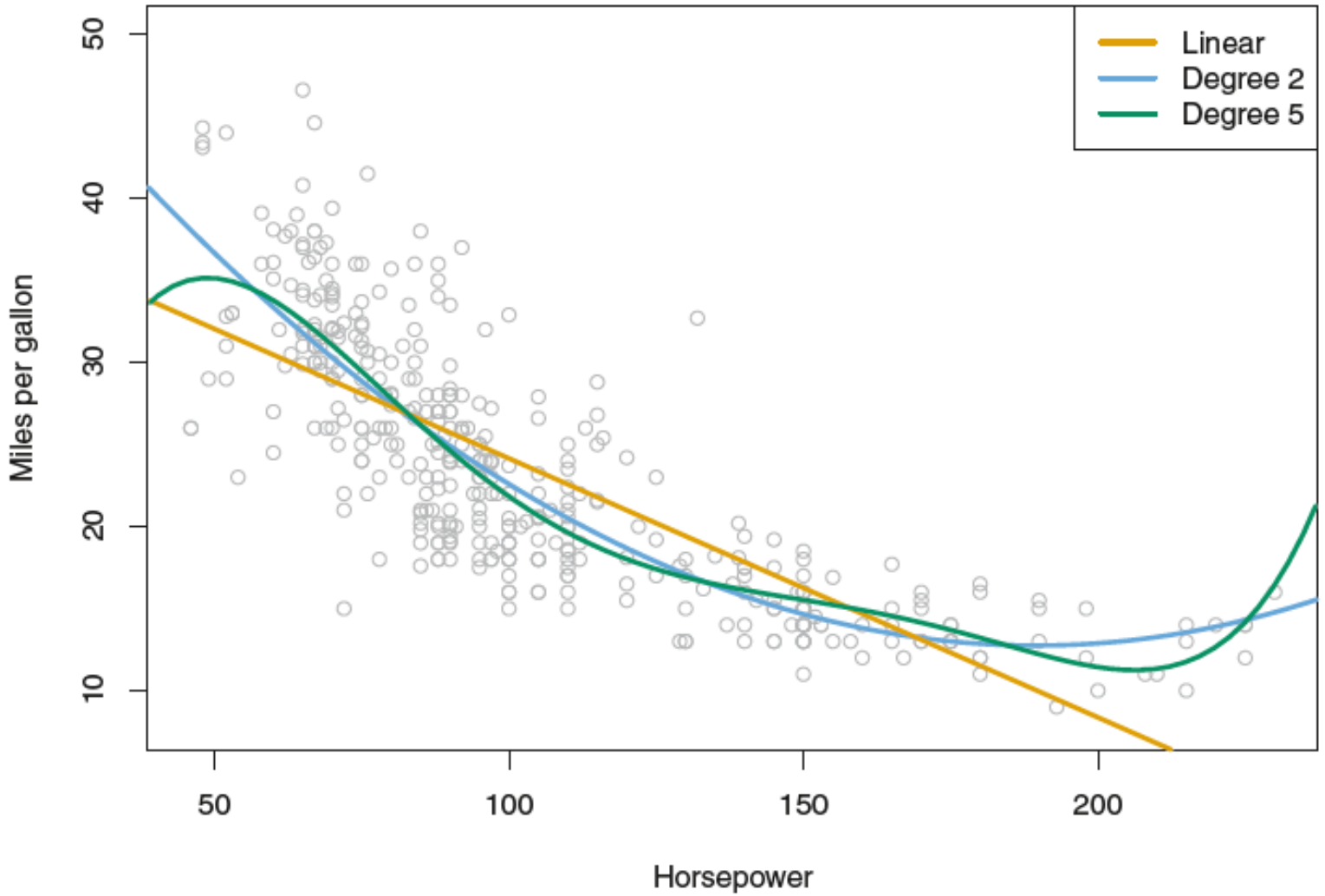
$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}$$



Усунення лінійності.

Усунення лінійності.



$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

Потенційні проблеми.

Потенційні проблеми.

1. Нелінійність взаємозв'язку між Y та X.

2. Кореляція залишків.

3. Непостійна дисперсія залишків.

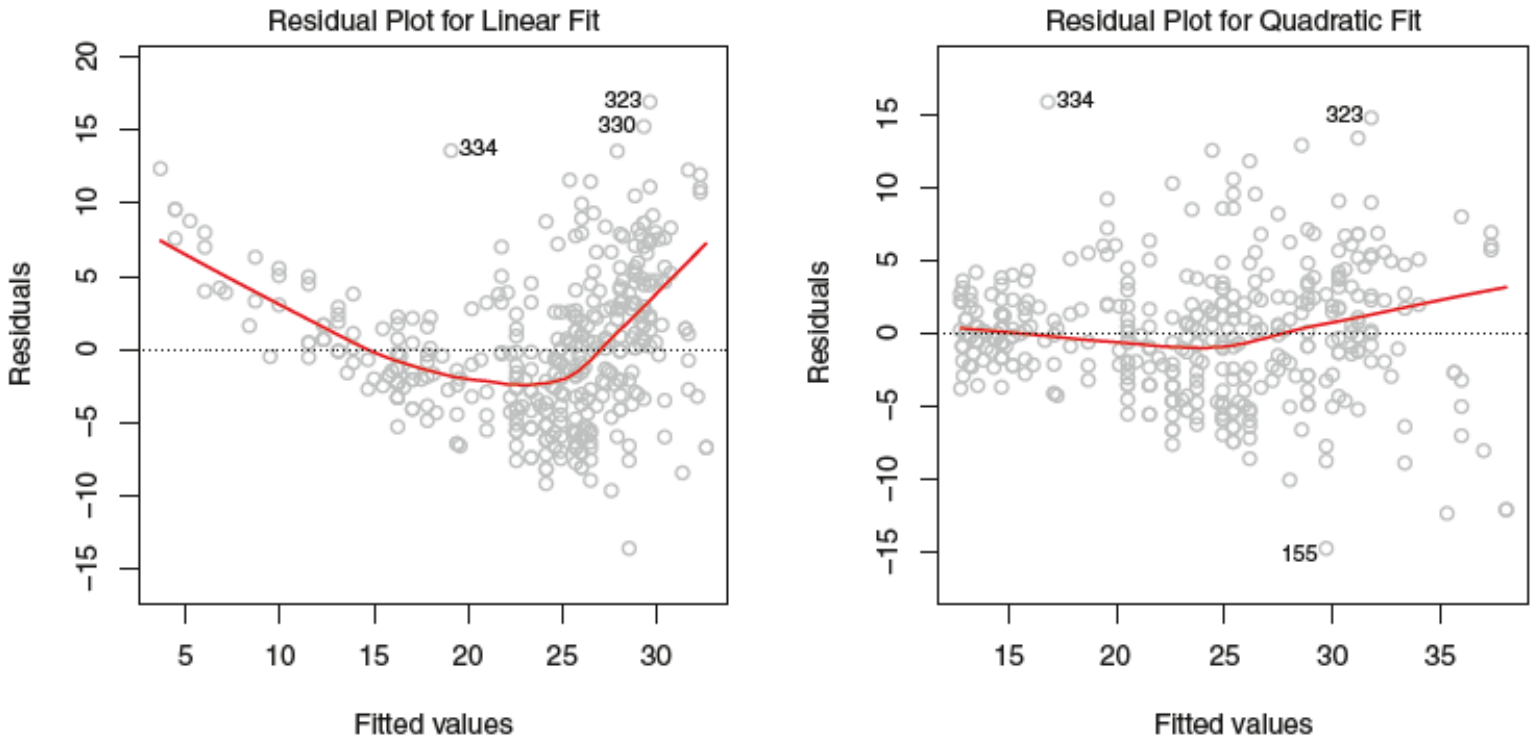
4. Викиди.

5. Точки з великим значенням важелів.

6. Колінеарність.

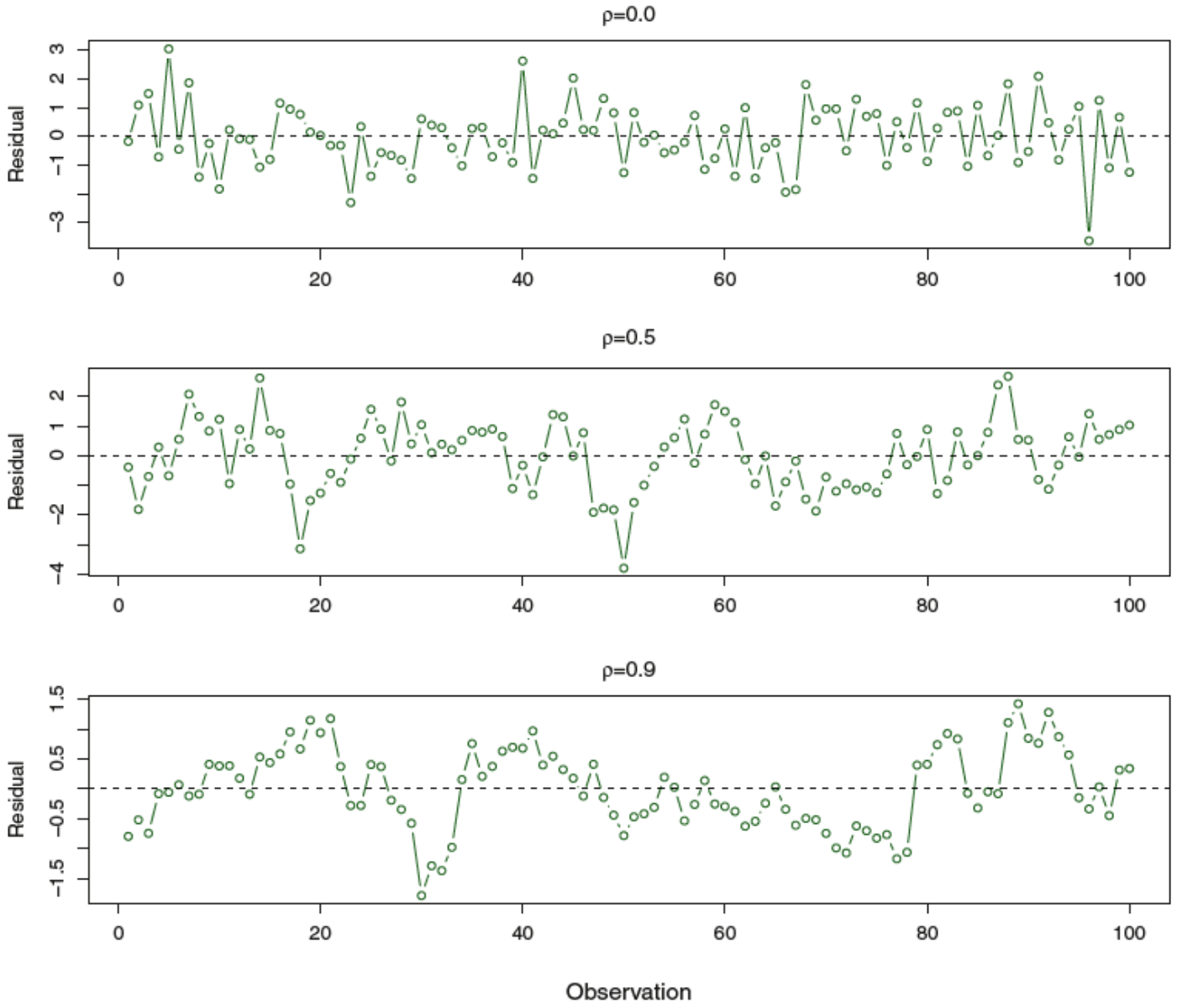
Нелінійність взаємозв'язку між Y та X.

Нелінійність взаємозв'язку між Y та X.



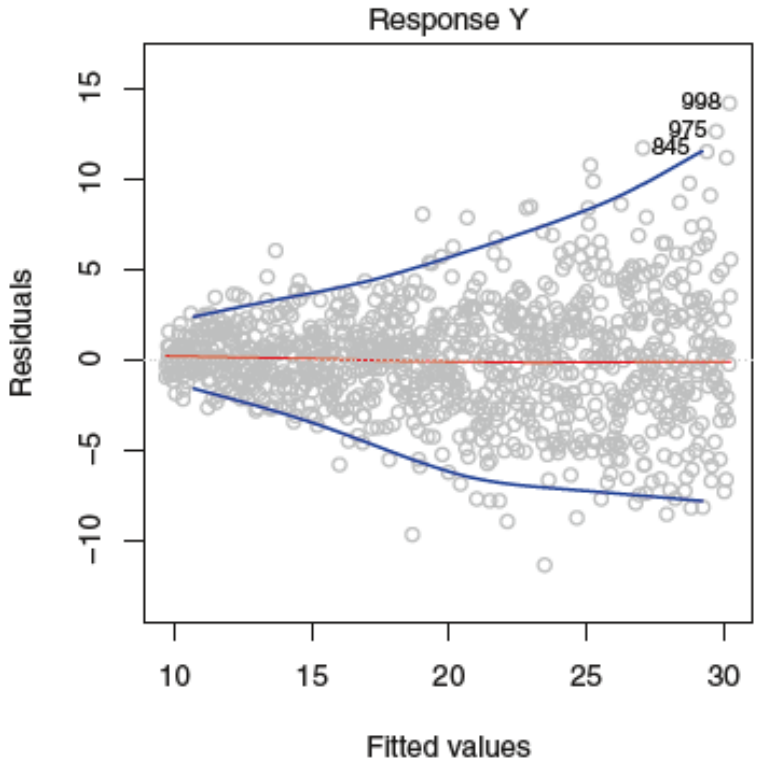
Кореляція залишків.

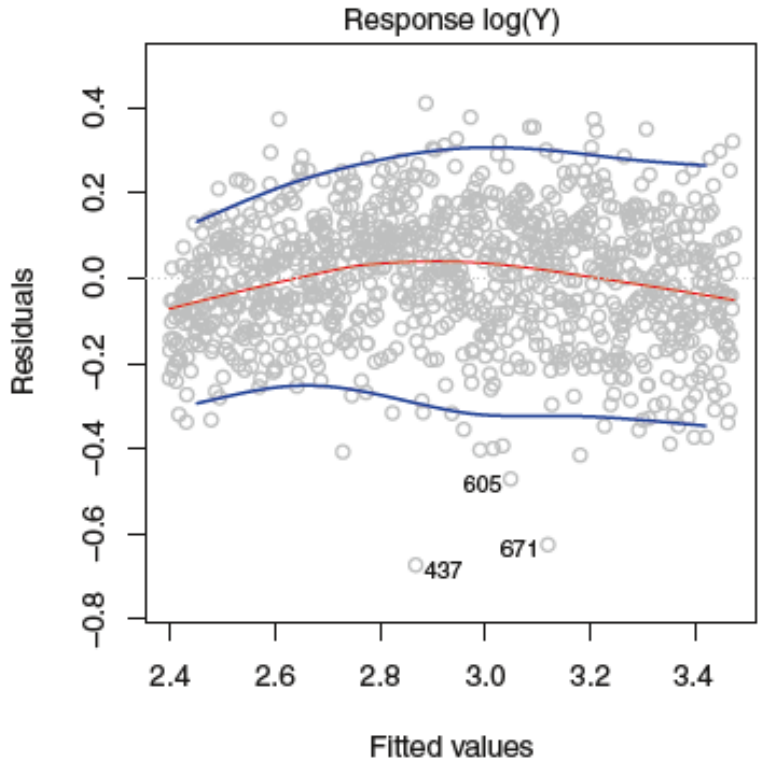
Кореляція залишків.



Непостійна дисперсія залишків.

Непостійна дисперсія залишків.

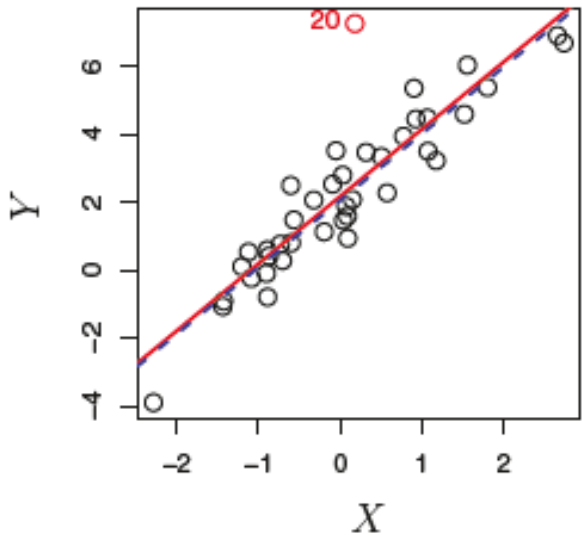


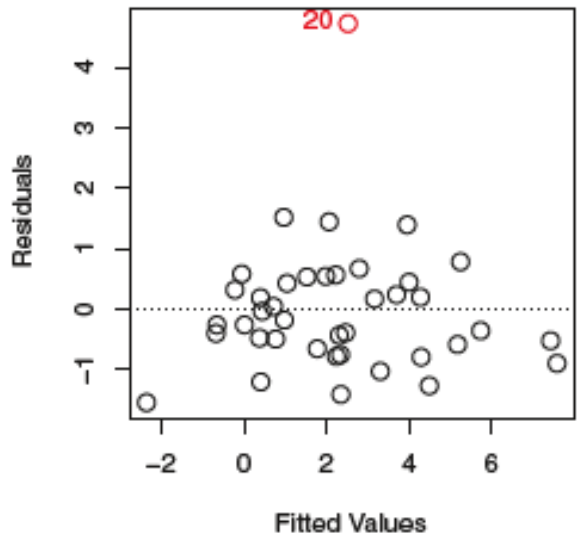


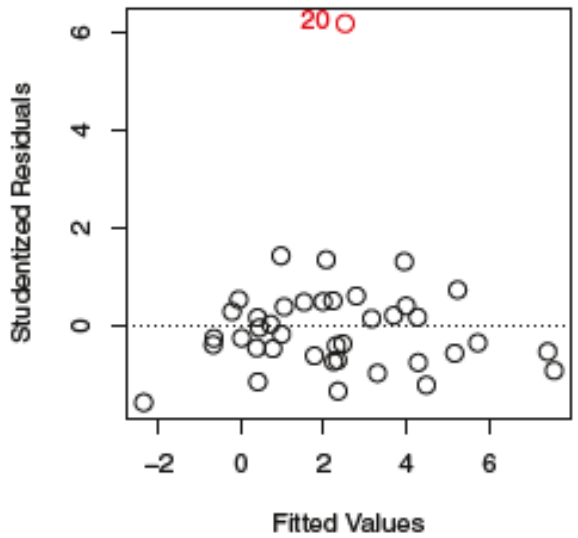
Викиди.

Викиди.

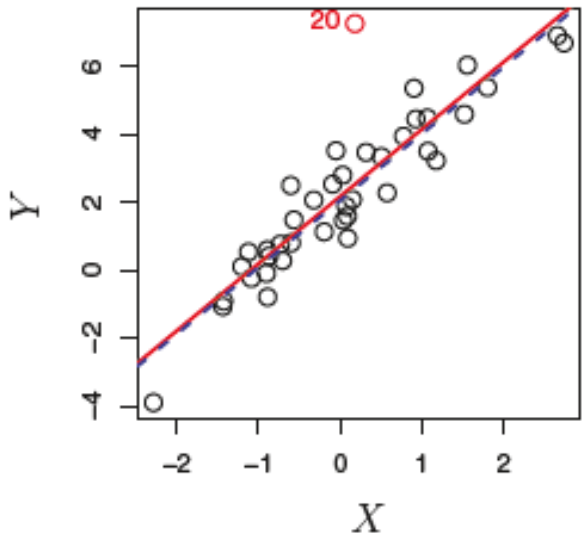
Викиди.

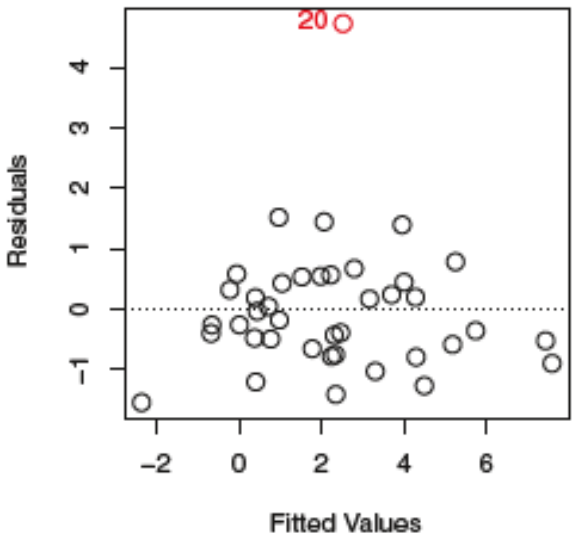

20
Y
X
−4
−2
0
2
4
6
−2
−1
0
1
2

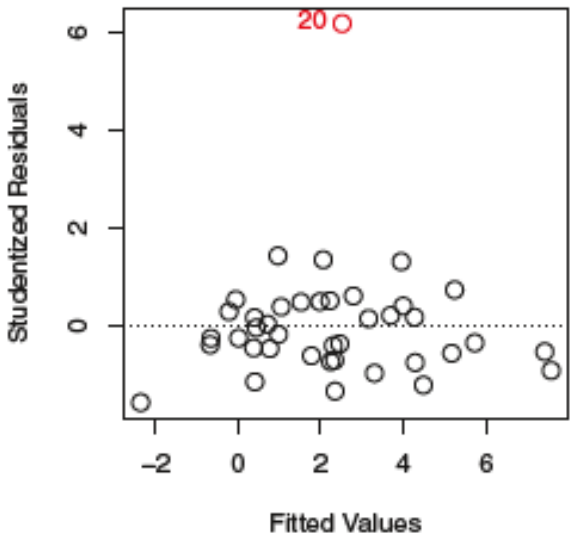

20
Residuals
Fitted Values
−1
0
1
2
3
4
−2
0
2
4
6


20
Studentized Residuals
Fitted Values
0
2
4
6
−2
0
2
4
6

Викиди.



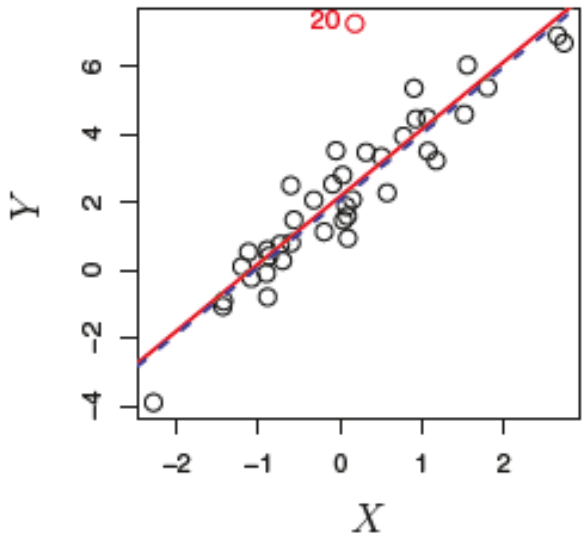


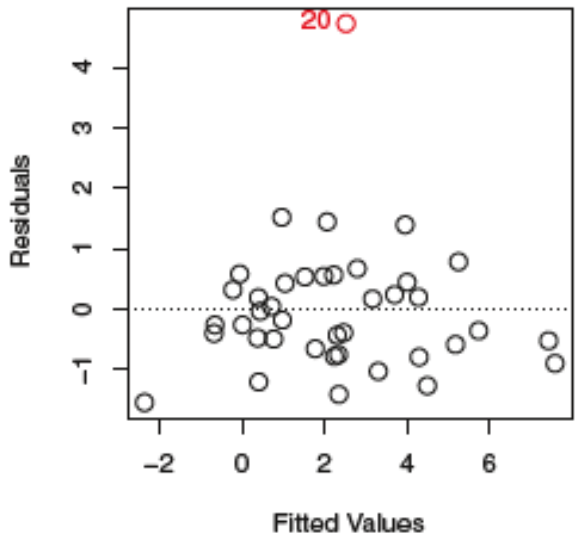


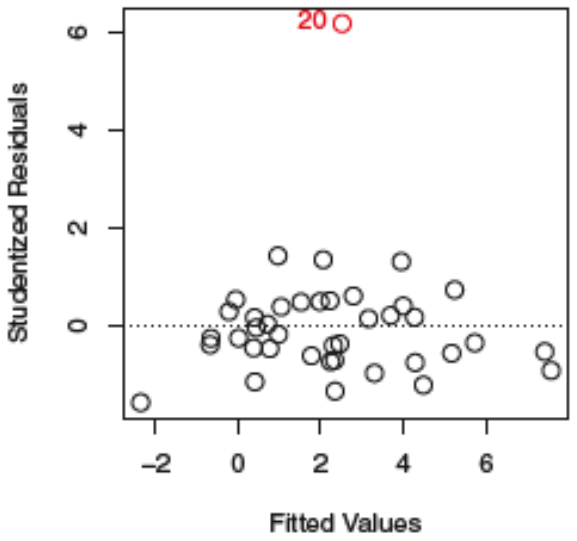
$$r_i = \frac{\varepsilon_i}{RSE\sqrt{1-h_i}},$$ де h_i – i-тий діагональний елемент матриці

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Викиди.







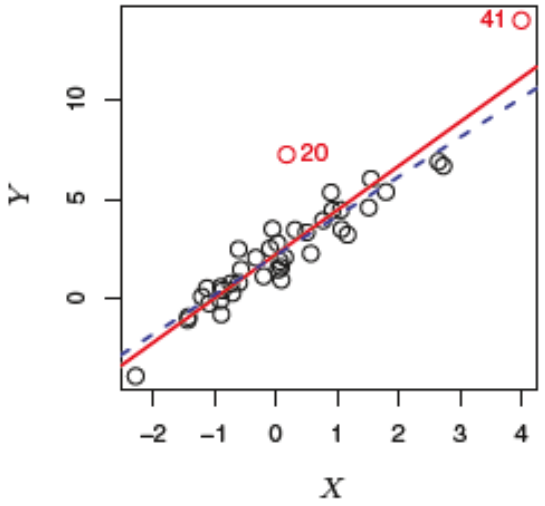
$$r_i = \frac{\varepsilon_i}{RSE\sqrt{1-h_i}},$$ де h_i – i-тий діагональний елемент матриці

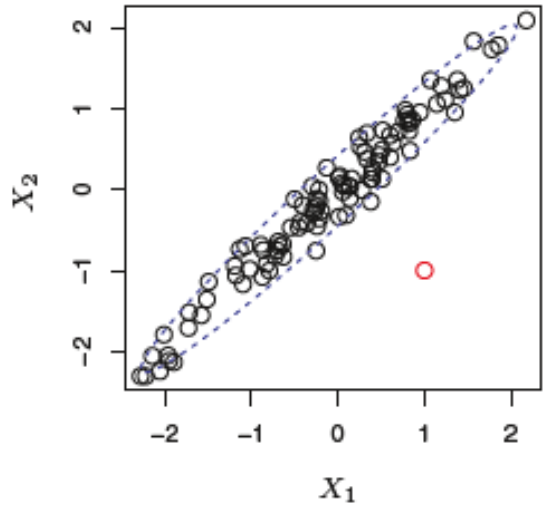
$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

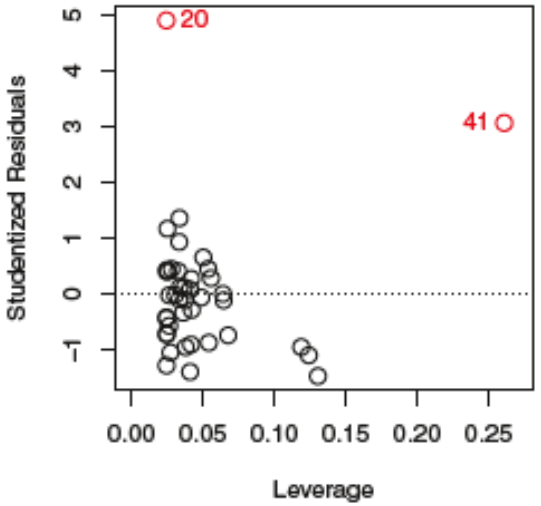
При $p = 1$: $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}.$

Точки з великим значенням важелів.

Точки з великим значенням важелів.





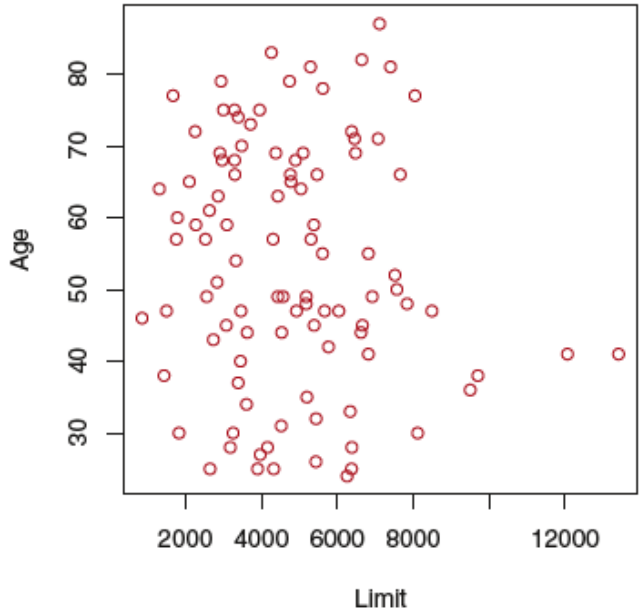


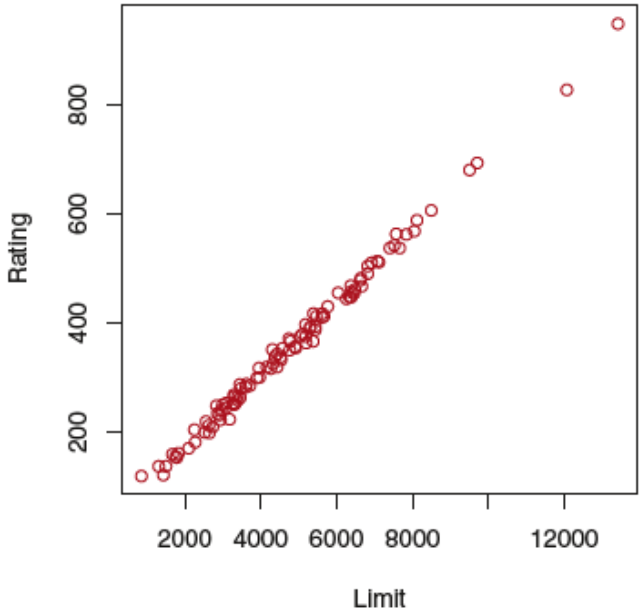
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}.$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

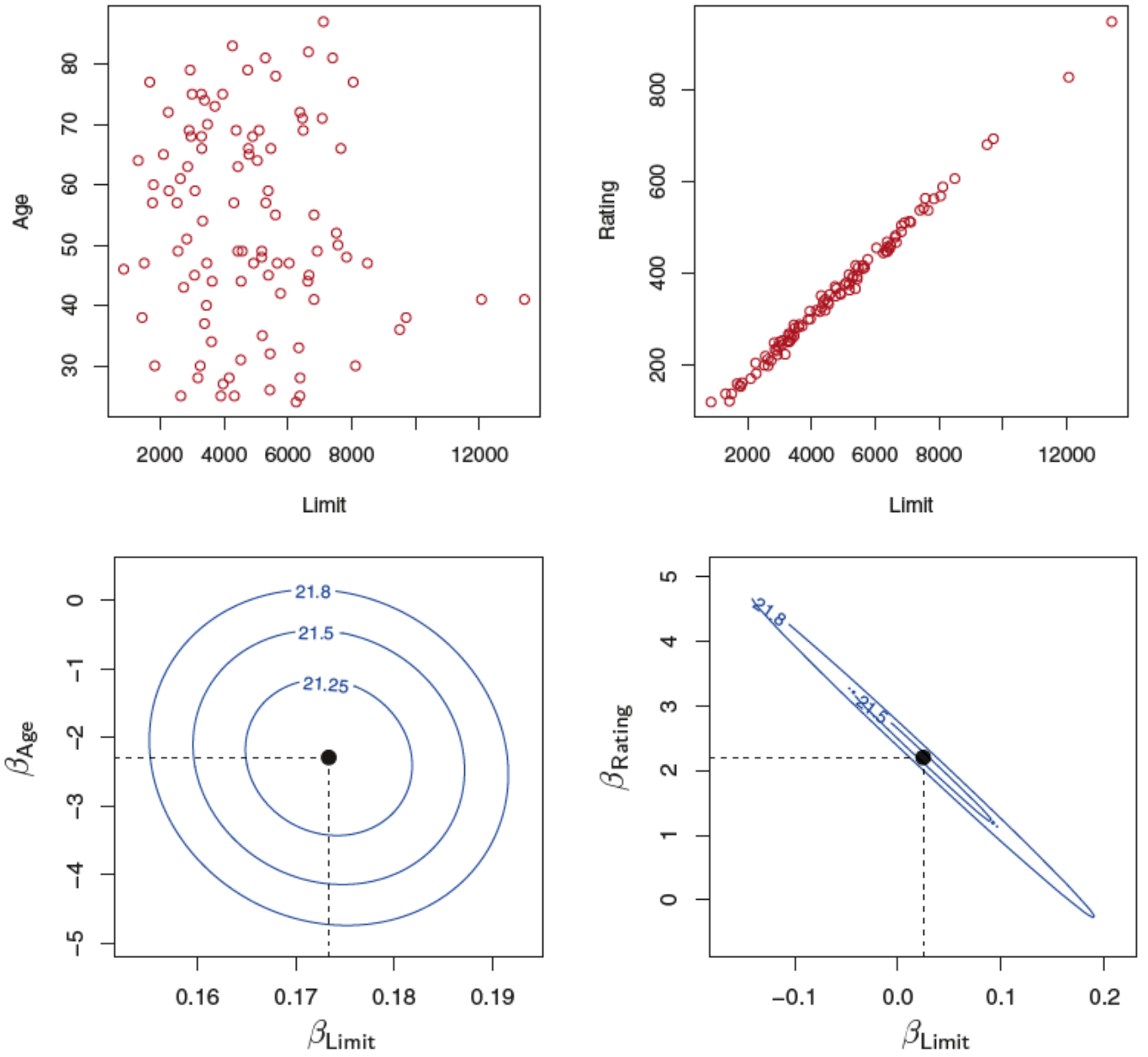
Колінеарність.

Колінеарність.





Колінеарність.



		Coefficient	Std. error	t-statistic	p-value
Model 1	`Intercept`	−173.411	43.828	−3.957	< 0.0001
	`age`	−2.292	0.672	−3.407	0.0007
	`limit`	0.173	0.005	34.496	< 0.0001
Model 2	`Intercept`	−377.537	45.254	−8.343	< 0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

		Coefficient	Std. error	t-statistic	p-value
Model 1	`Intercept`	−173.411	43.828	−3.957	< 0.0001
	`age`	−2.292	0.672	−3.407	0.0007
	`limit`	0.173	0.005	34.496	< 0.0001
Model 2	`Intercept`	−377.537	45.254	−8.343	< 0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

Підсумки

1. Чи існує взаємозв'язок між витратами на рекламу та продажами?

Підсумки

1. Чи існує взаємозв'язок між витратами на рекламу та продажами?

На це питання можна відповісти, побудувавши модель множинної регресії продажів на витрати на рекламу по телевізору, радіо та у газетах та перевіривши гіпотезу

$H_0 : \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0.$

Підсумки

1. Чи існує взаємозв'язок між витратами на рекламу та продажами?

На це питання можна відповісти, побудувавши модель множинної регресії продажів на витрати на рекламу по телевізору, радіо та у газетах та перевіривши гіпотезу

$H_0 : \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0.$

2. Наскільки сильним є зв'язок?

Підсумки

1. Чи існує взаємозв'язок між витратами на рекламу та продажами?

На це питання можна відповісти, побудувавши модель множинної регресії продажів на витрати на рекламу по телевізору, радіо та у газетах та перевіривши гіпотезу

$H_0 : \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0.$

2. Наскільки сильним є зв'язок?

Ми розглянули два показники

RSE (1,681 при середньому значенні продажів 14,022).

R^2, що дорівнювала 90%.

3. Які засоби масової інформації сприяють продажам?

3. Які засоби масової інформації сприяють продажам?

Ми перевірили спочатку *p*-значення кожної змінної зокрема. У випадку множинної регресії виявилося, що *p*-значення для телебачення та радіо є низькі, а для газети - ні.

3. Які засоби масової інформації сприяють продажам?

Ми перевірили спочатку *p*-значення кожної змінної зокрема. У випадку множинної регресії виявилося, що *p*-значення для телебачення та радіо є низькі, а для газети - ні.

4. Наскільки великий вплив кожного засобу на продажі?

3. Які засоби масової інформації сприяють продажам?
Ми перевірили спочатку *p*-значення кожної змінної зокрема. У випадку множинної регресії виявилося, що *p*-значення для телебачення та радіо є низькі, а для газети - ні.

4. Наскільки великий вплив кожного засобу на продажі?

Ми отримали наступні інтервали довіри: (0.043, 0.049) для телебачення, (0.172, 0.206) для радіо, (−0.013, 0.011) для газет. Перевірка на колінеарність дала наступні значення 1.005, 1.145, 1.145.

3. Які засоби масової інформації сприяють продажам?
Ми перевірили спочатку *p*-значення кожної змінної зокрема. У випадку множинної регресії виявилося, що *p*-значення для телебачення та радіо є низькі, а для газети - ні.

4. Наскільки великий вплив кожного засобу на продажі?

Ми отримали наступні інтервали довіри: (0.043, 0.049) для телебачення, (0.172, 0.206) для радіо, (−0.013, 0.011) для газет. Перевірка на колінеарність дала наступні значення 1.005, 1.145, 1.145.

5. Наскільки точно ми можемо передбачити майбутні продажі?

3. Які засоби масової інформації сприяють продажам?
Ми перевірили спочатку *p*-значення кожної змінної зокрема. У випадку множинної регресії виявилося, що *p*-значення для телебачення та радіо є низькі, а для газети - ні.

4. Наскільки великий вплив кожного засобу на продажі?

Ми отримали наступні інтервали довіри: (0.043, 0.049) для телебачення, (0.172, 0.206) для радіо, (−0.013, 0.011) для газет. Перевірка на колінеарність дала наступні значення 1.005, 1.145, 1.145.

5. Наскільки точно ми можемо передбачити майбутні продажі?

Ми можемо використати інтервали довіри чи передбачення.

6. Чи є зв'язок лінійним?

6. Чи є зв'язок лінійним?

Ми зауважили наявність нелінійності. Ми коротко пояснили способи згладжування такої нелінійності.

6. Чи є зв'язок лінійним?

Ми зауважили наявність нелінійності. Ми коротко пояснили способи згладжування такої нелінійності.

7. Чи наявна взаємодія між рекламними носіями?

6. Чи є зв'язок лінійним?

Ми зауважили наявність нелінійності. Ми коротко пояснили способи згладжування такої нелінійності.

7. Чи наявна взаємодія між рекламними носіями?

Ми показали наявність взаємодії. За допомогою введення додаткової змінної нам вдалося досягнути зростання R^2 з 90% до 97%.

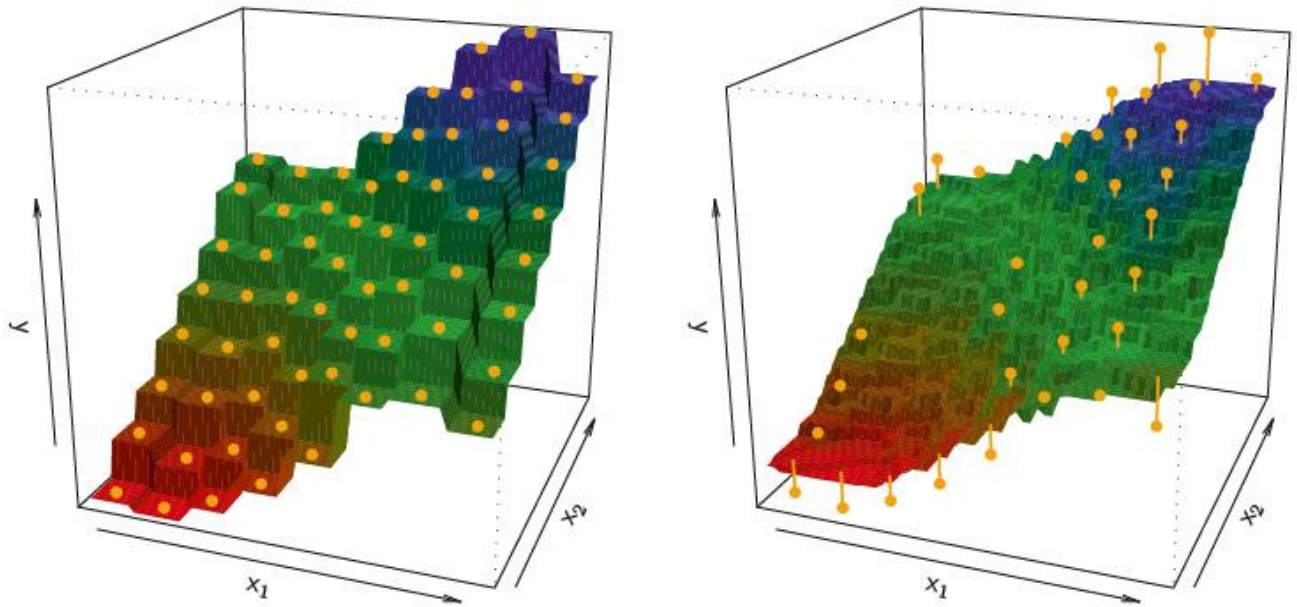
Порівняння методу лінійної регресії з методом K-найближчих сусідів

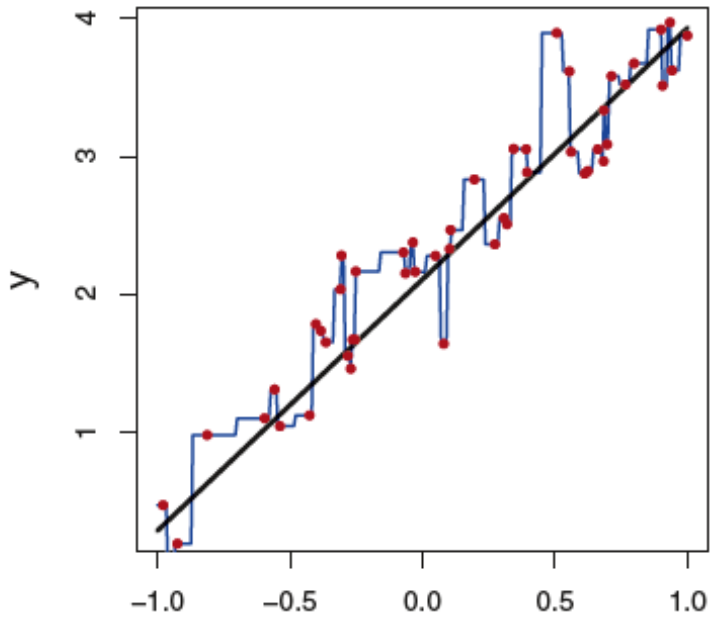
Порівняння методу лінійної регресії з методом K-найближчих сусідів

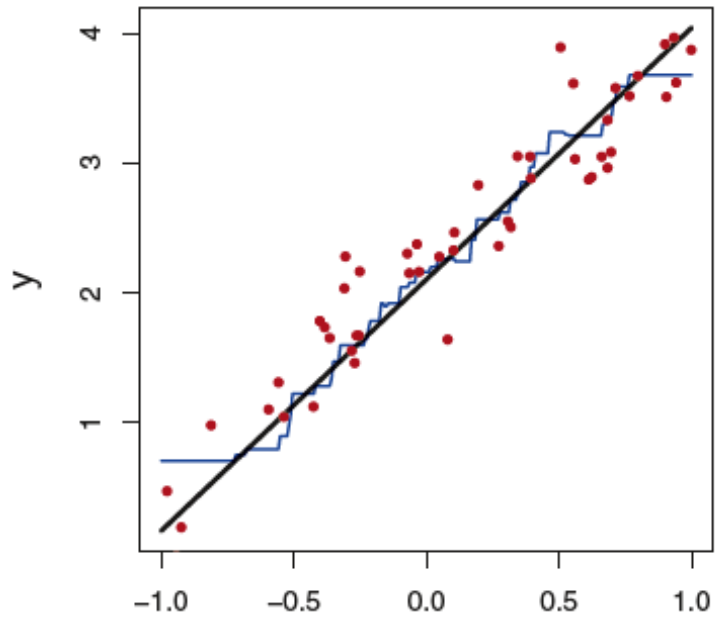
$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

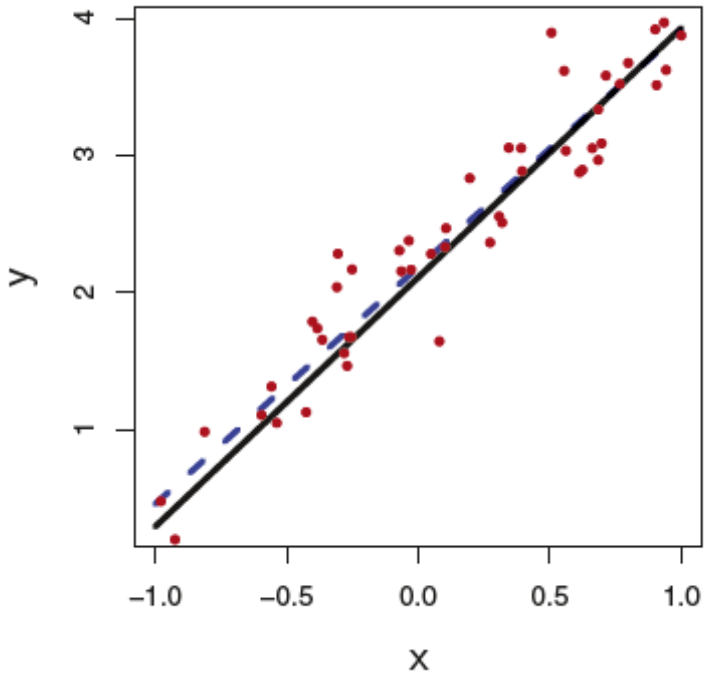
Порівняння методу лінійної регресії з методом K-найближчих сусідів

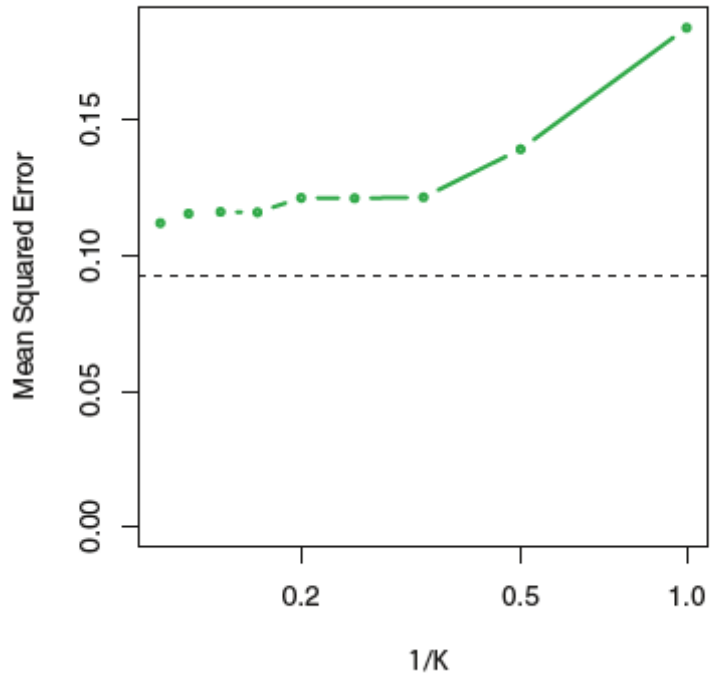
$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

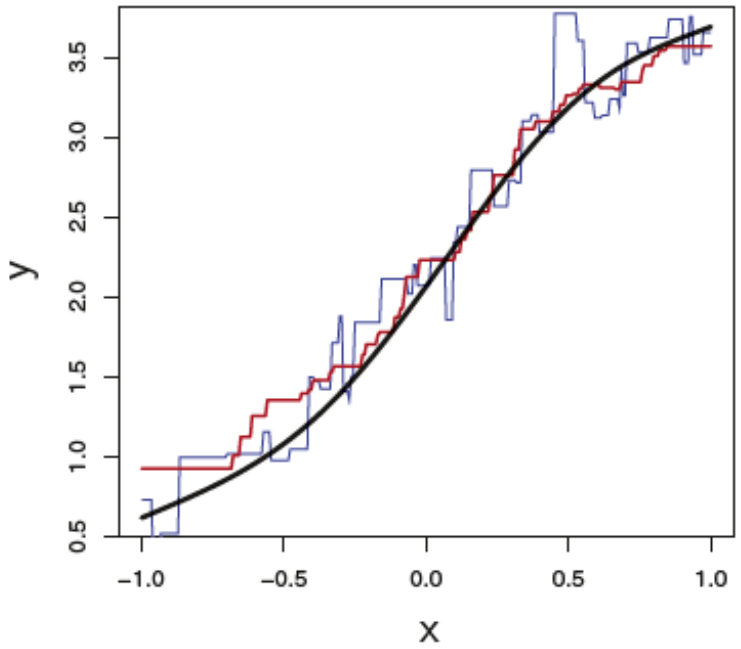


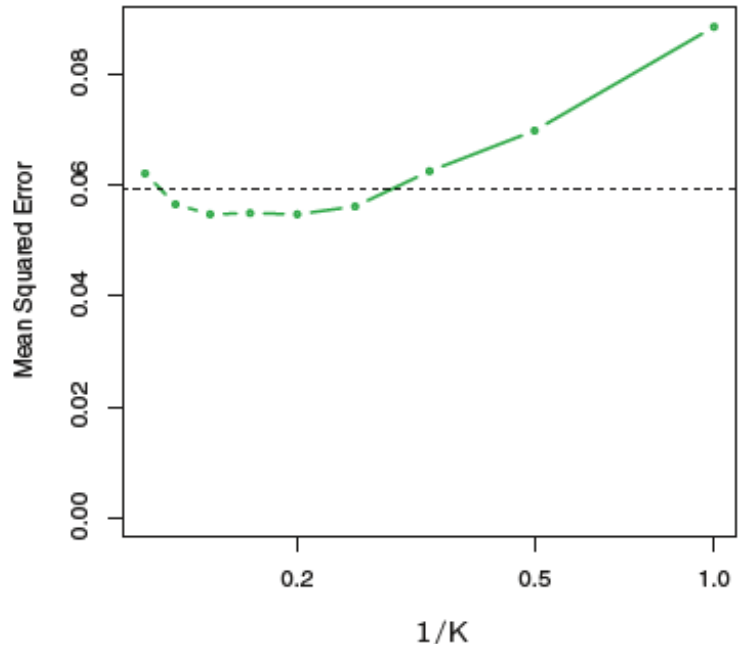

y
4
3
2
1
−1.0
−0.5
0.0
0.5
1.0

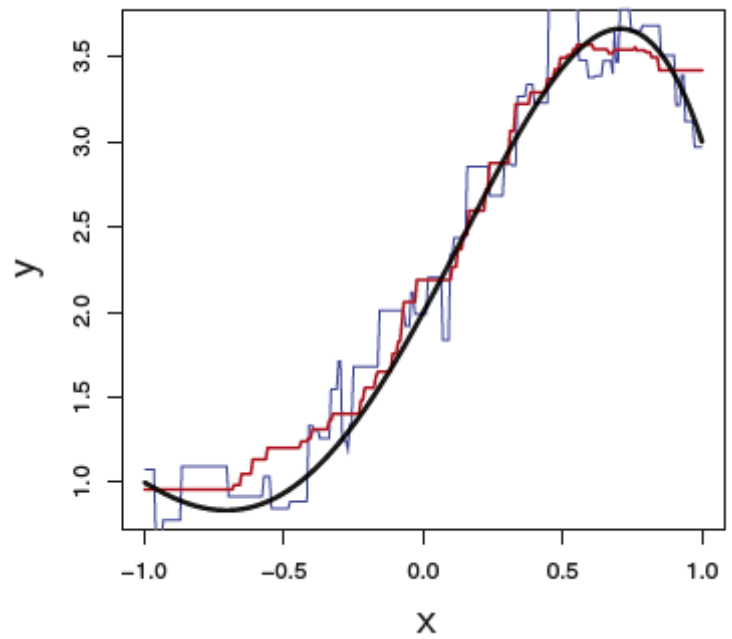

y
4
3
2
1
−1.0
−0.5
0.0
0.5
1.0

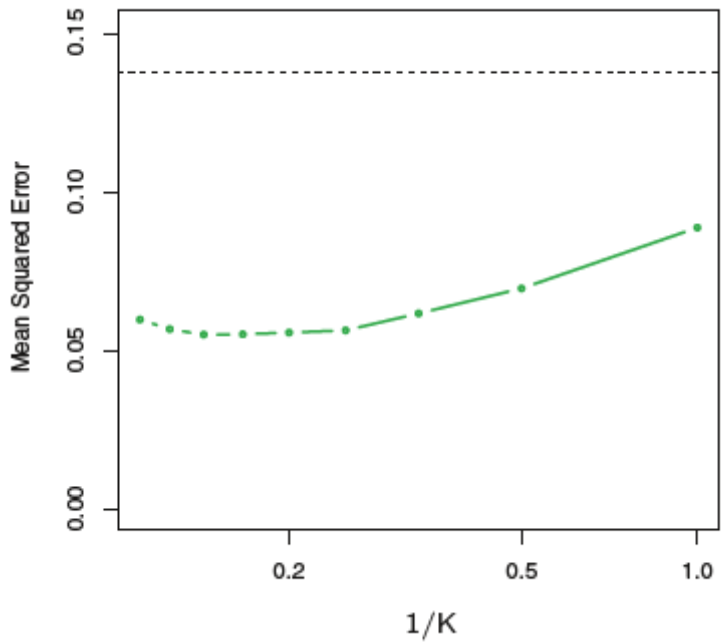

y
X
4
3
2
1
−1.0
−0.5
0.0
0.5
1.0

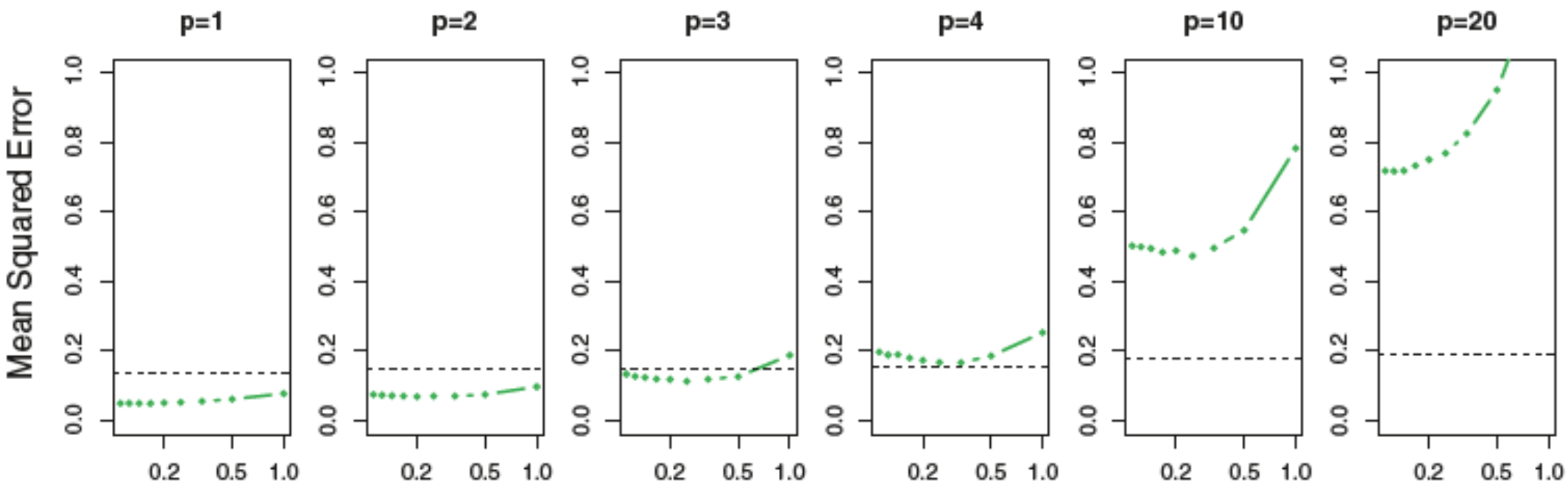

Mean Squared Error
1/K
0.15
0.10
0.05
0.00
0.2
0.5
1.0


y
X
0.5
1.0
1.5
2.0
2.5
3.0
3.5
−1.0
−0.5
0.0
0.5
1.0


Mean Squared Error
1/K
0.00
0.02
0.04
0.06
0.08
0.2
0.5
1.0


y
X
1.0
1.5
2.0
2.5
3.0
3.5
−1.0
−0.5
0.0
0.5
1.0


Mean Squared Error
1/K
0.00
0.05
0.10
0.15
0.2
0.5
1.0


p=1
p=2
p=3
p=4
p=10
p=20
Mean Squared Error
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.5
1.0