Học có giám sát

Nguyễn Thanh Tùng
Khoa Công nghệ thông tin – Đại học Thủy Lợi
tungnt@tlu.edu.vn

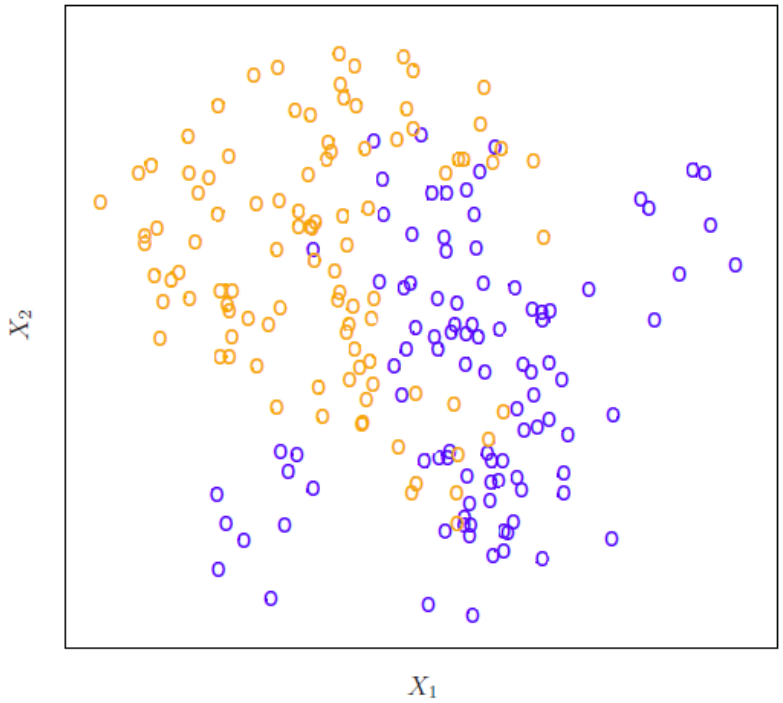
Website môn học: https://sites.google.com/a/wru.vn/cse445fall2017

Bài giảng có sử dụng hình vẽ trong cuốn sách "An Introduction to Statistical Learning with Applications in R" với sự cho phép của tác giả, có sử dụng slides các khóa học CME250 của ĐH Stanford và IOM530 của ĐH Southern California


Giải thuật phân lớp đơn giản
(nhắc lại Bài 1)

K-Nearest Neighbor classifier (KNN)

- Ý tưởng: phân lớp các mẫu dựa trên "hàng xóm" các mẫu đã biết nhãn



K-Nearest Neighbor classifier (KNN)

- Bộ phân lớp: Chia không gian thuộc tính thành nhiều vùng
  - Mỗi vùng được gắn với 1 nhãn lớp (class label)
  - *Ranh giới quyết định* chia tách các vùng quyết định
- Các phương pháp phân lớp xây dựng mô hình có dạng:

$$Pr\left(Y \mid X\right)$$

K-Nearest Neighbor classifier (KNN)

- Bộ phân lớp KNN
  - Việc dự đoán lớp cho mẫu X là *lớp phổ biến nhất giữa K láng giềng gần nhất* (trong tập học)
  - Mô hình phân lớp:

$$Pr(X \text{ belongs to class } Y) \approx \frac{\#\,(\text{neighbors of } X \text{ in class } Y)}{K}$$

K-Nearest Neighbor classifier (KNN)

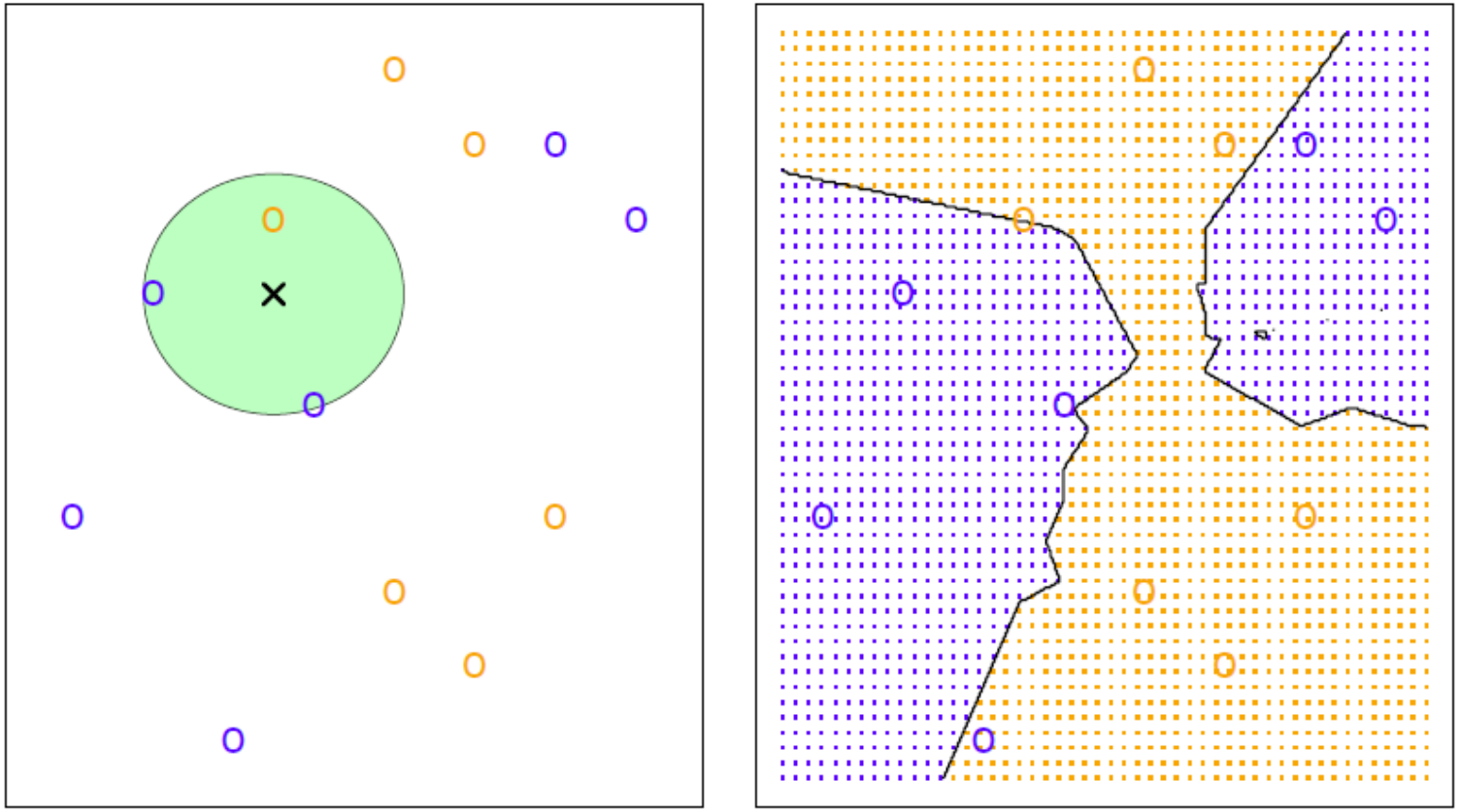


Figure 2.14, ISL 2013

Lựa chọn K (bộ phân lớp KNN)

- K nhỏ
  - Ranh giới quyết định linh hoạt hơn, tuy nhiên dễ bị *overfit*
- K lớn
  - Ranh giới quyết định ít linh hoạt nhưng ít bị *overfit*
- *Overfitting*: Cho kết quả tốt trên tập học nhưng kém trên tập thử nghiệm

Lựa chọn K (bộ phân lớp KNN)

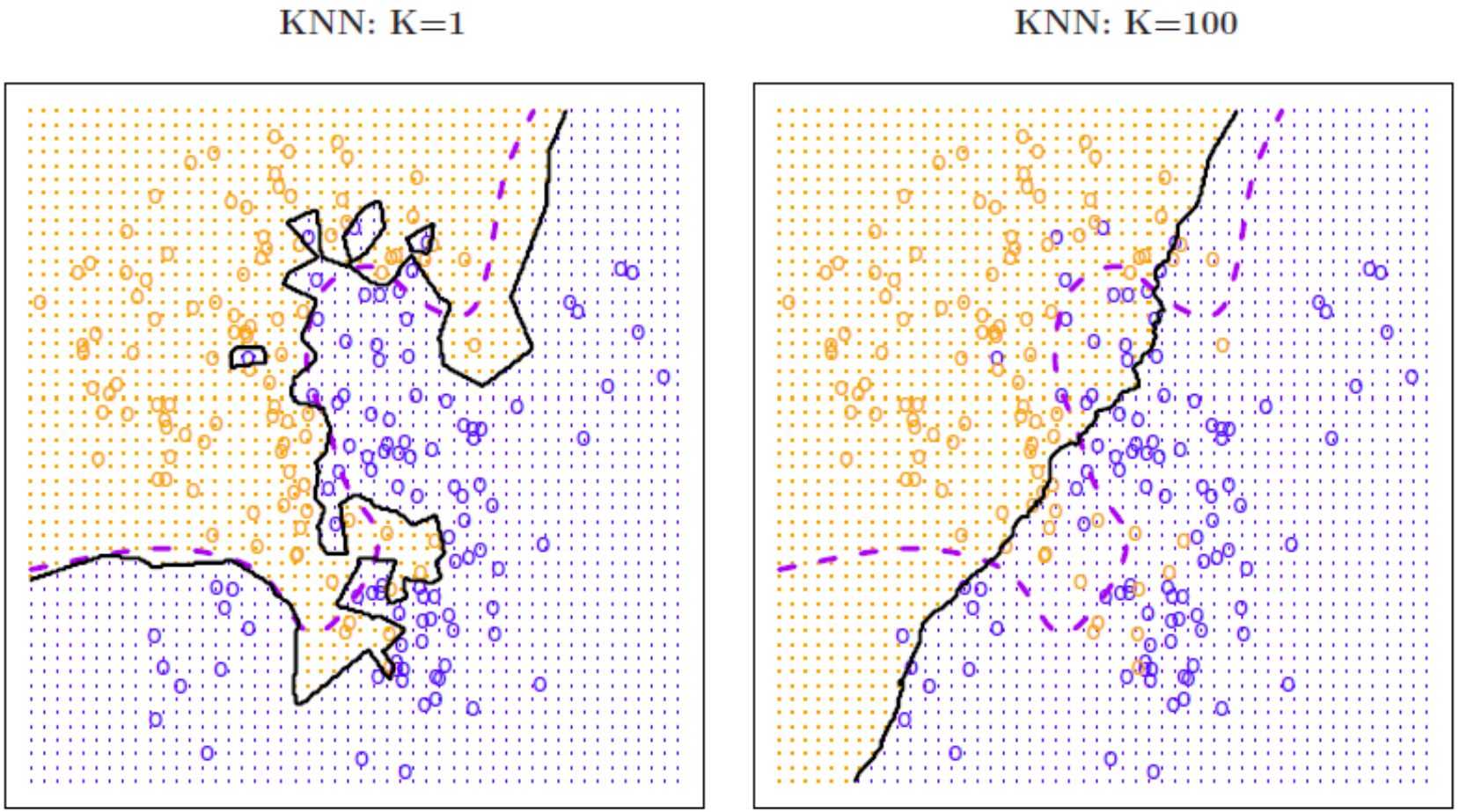


Figure 2.16, ISL 2013

Lựa chọn K (bộ phân lớp KNN)

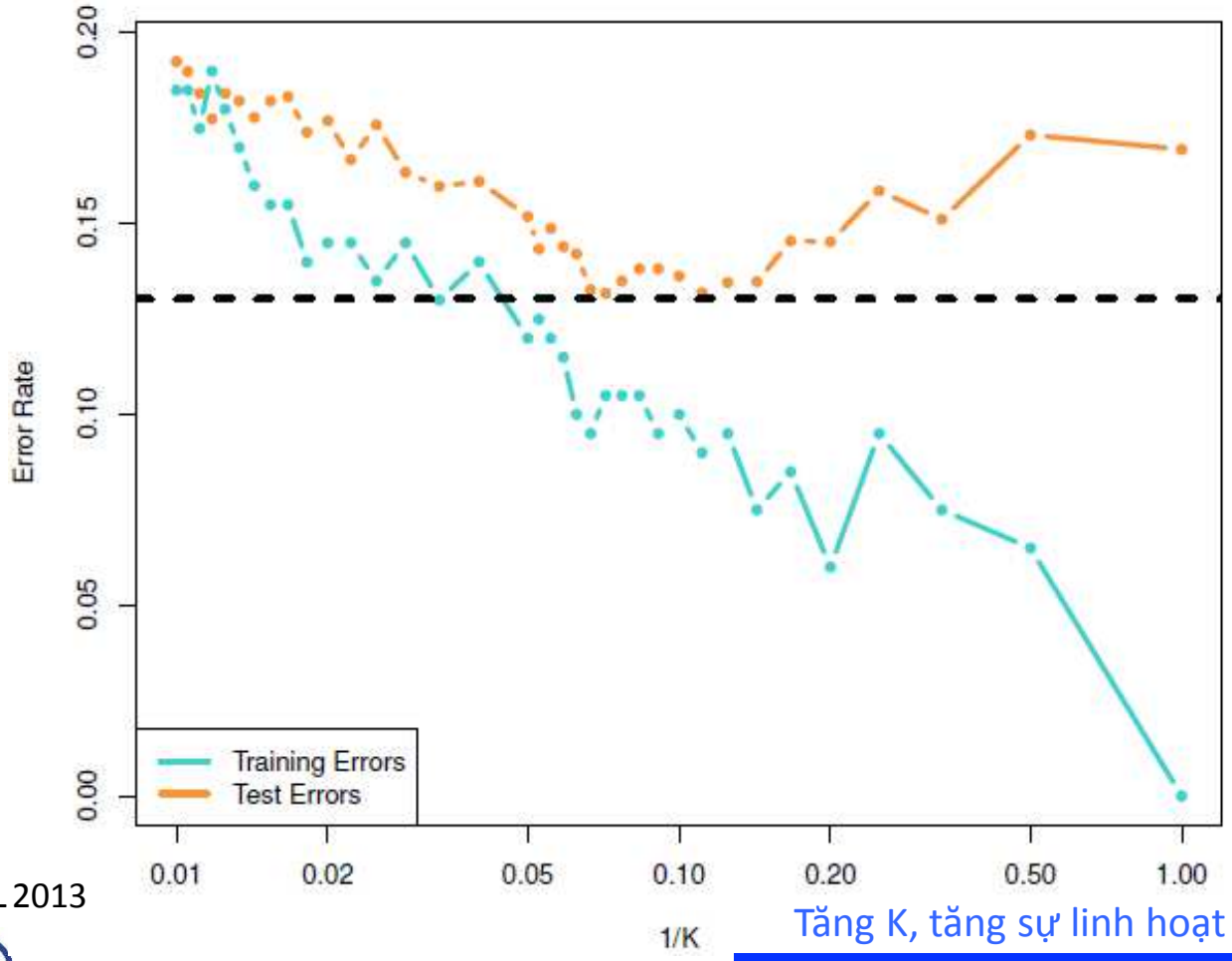


Figure 2.17, ISL 2013

Tại sao lỗi huấn luyện (trên dữ liệu học) tăng cùng K?

Tại sao lỗi kiểm thử lại khác?

Tăng K, tăng sự linh hoạt →

Lựa chọn K (bộ phân lớp KNN)

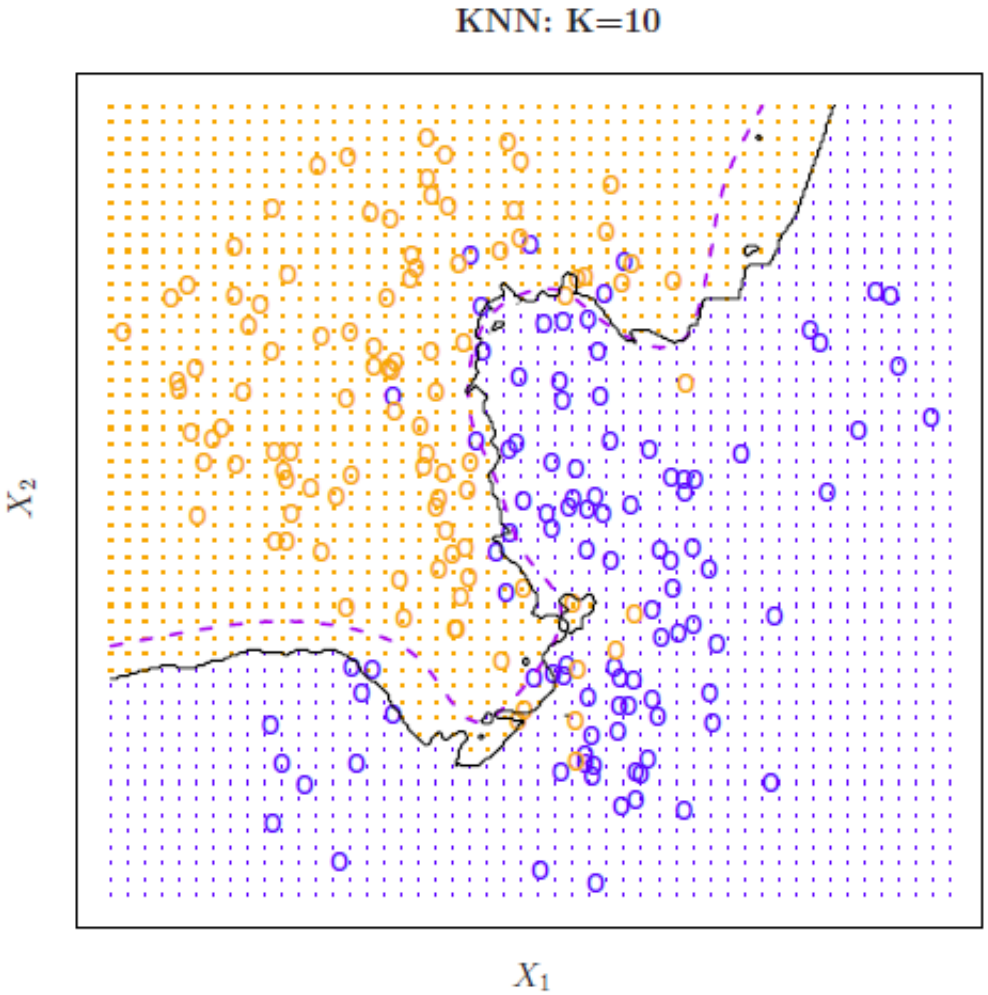


Figure 2.15, ISL 2013

Câu đố:

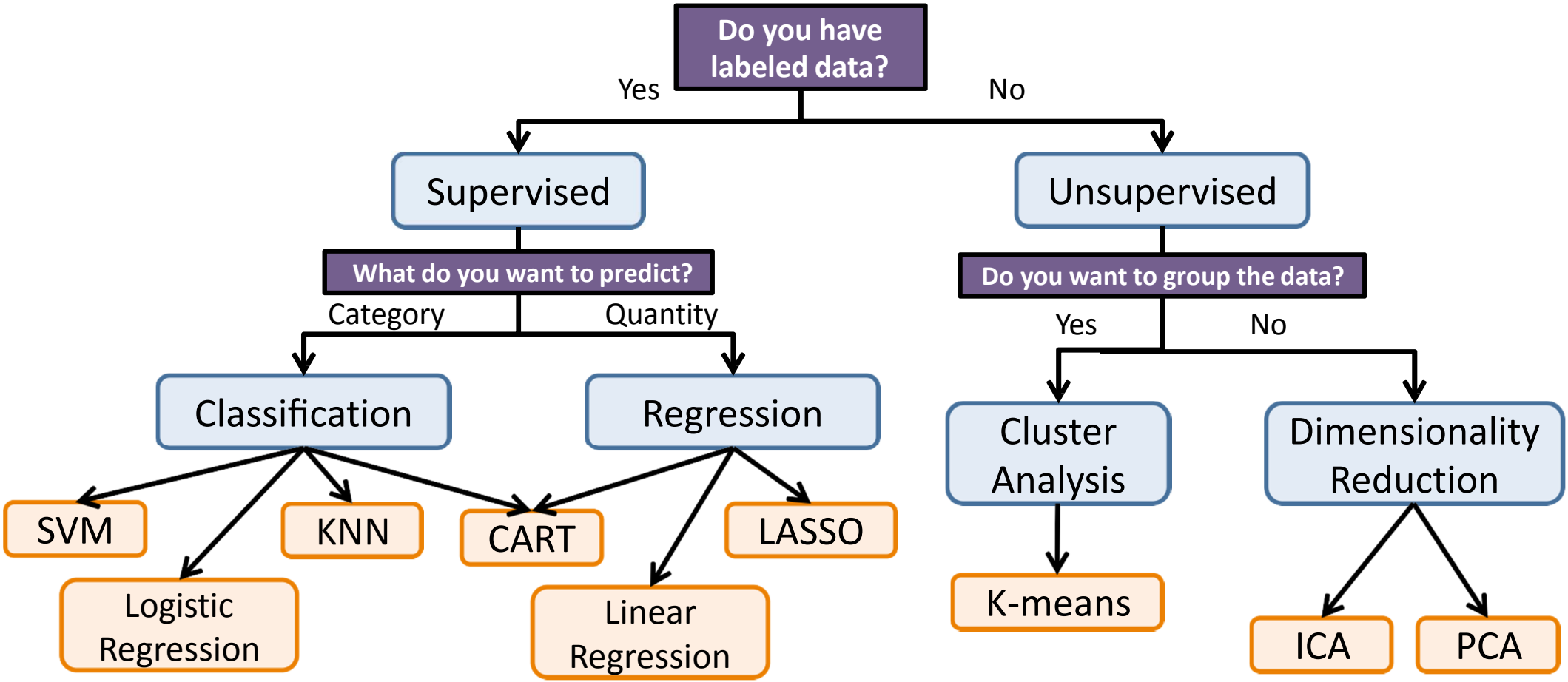
- Bộ phân lớp KNN là tham số hay phi tham số?
  - ***Nhắc lại:***

    Kỹ thuật tham số phải đặt các giả định của mô hình về dữ liệu

    (chẳng hạn, dữ liệu theo xu hướng tuyến tính; dữ liệu tuân theo phân bố chuẩn)

- Liệu ta có thể dùng kỹ thuật KNN dự đoán một giá trị số thay cho giá trị định danh (i.e. “KNN hồi quy”)?

Các dạng giải thuật học máy



Giải thuật Học máy "Tốt nhất"

- Tin tồi: Không có giải thuật nào tốt nhất
  - Không có giải thuật học máy nào thực hiện tốt cho mọi bài toán
- Tin tốt: Tất cả các giải thuật học máy đều tốt
  - Mỗi giải thuật học máy thực hiện tốt cho một số bài toán
- Định lý "No free lunch"
  - Wolpert (1996): các giải thuật thực hiện như nhau khi ta lấy trung bình kết quả chúng thực hiện trên tất cả các bài toán

Trade-offs (đánh đổi) trong Học máy

- Độ lệch vs. Phương sai
- Độ chính xác vs. Khả năng diễn giải (một tính chất của mô hình về khả năng thấy được mối quan hệ giữa các biến)
- Độ chính xác vs. Khả năng mở rộng giải thuật
- Phạm vi kiến thức vs. Hướng dữ liệu
- Nhiều dữ liệu vs. Giải thuật tốt hơn

Chuẩn bị dữ liệu

- Các giải thuật học máy cần phải có dữ liệu!
- Tiền xử lý dữ liệu để chuyển đổi dữ liệu trước khi áp dụng vào giải thuật học máy
  - Lấy mẫu: chọn tập con các quan sát/mẫu
  - *Trích chọn thuộc tính*: Chọn các biến đầu vào
  - *Chuẩn hóa dữ liệu (Normalization)* (standardization, scaling, binarization)
  - Xử lý dữ liệu thiếu và phần tử ngoại lai (missing data and outliers)
- Ngoài ra, còn phụ thuộc vào giải thuật học máy
  - Cây quyết định có thể xử lý dữ liệu thiếu/phần tử ngoại lai
  - PCA yêu cầu dữ liệu đã được chuẩn hóa


Các câu hỏi?

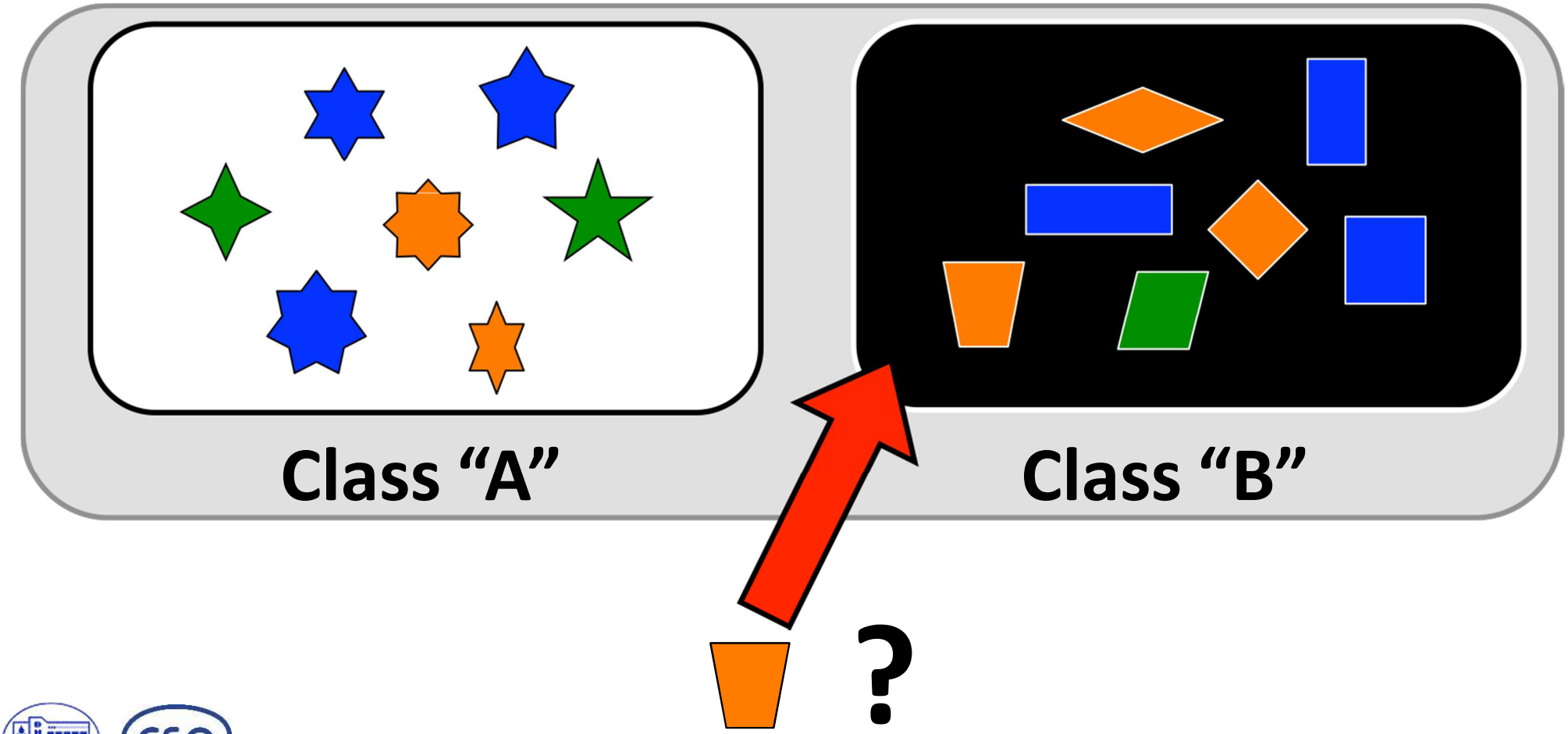
Giới thiệu về Học có giám sát

Học giám sát

- Xét: $$Y = f(X) + \epsilon$$
- Các phương pháp học giám sát:
  - Học bởi các ví dụ (quan sát)-“Learn by example”
  - Xây dựng mô hình $\hat{f}$ sử dụng tập các quan sát đã được gắn nhãn

$$\left(X^{(1)}, Y^{(1)}\right), \ldots, \left(X^{(n)}, Y^{(n)}\right)$$

Dữ liệu học



Dữ liệu học

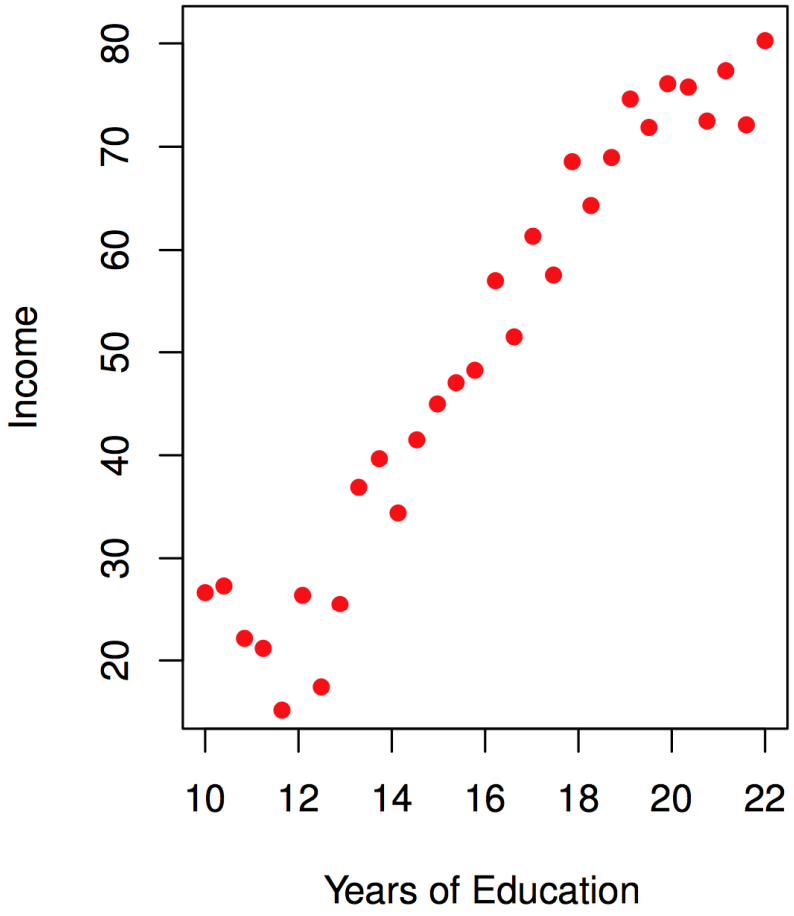


Figure 2.2 , ISL 2013

Học có giám sát

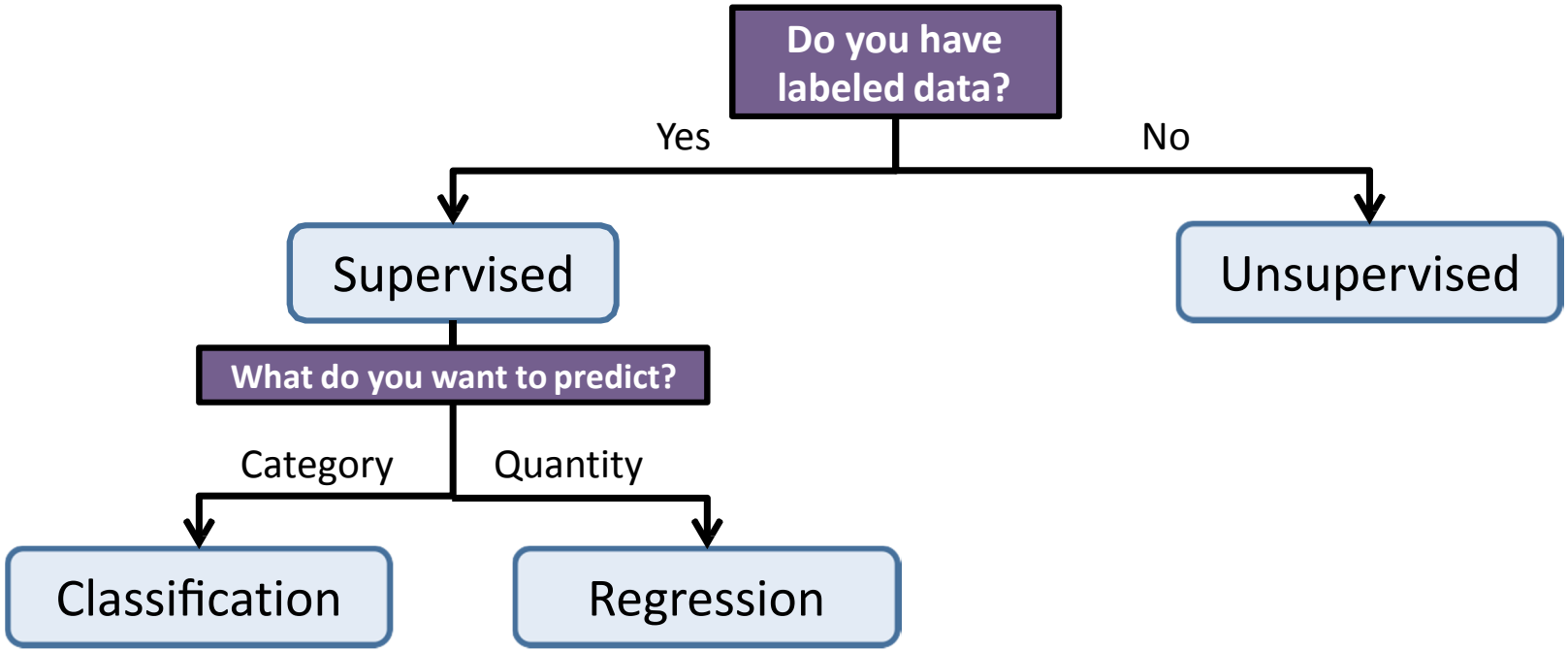
- Giải thuật học có giám sát
  - Lấy hàm ước lượng "tốt nhất" $\hat{f}$ trong tập các hàm
- Ví dụ: Hồi quy tuyến tính
  - Chọn 1 ước lượng tốt nhất từ *dữ liệu học* trong tập các hàm tuyến tính

$$f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d$$

Phân lớp và Hồi quy

- Bài toán học có giám sát gồm 2 dạng:
  - Hồi quy: biến đầu ra Y là định lượng (quantitative)
  - Phân lớp: biến đầu ra Y là định tính/hạng mục/rời rạc

Các dạng giải thuật học máy



Độ chính xác của mô hình

Đo hiệu năng bài toán hồi quy

- Hàm tổn thất (Loss function): loại hàm dùng để đo lường sai số của mô hình
- Vd: Sai số bình phương trung bình (Mean squared error - MSE)
  - Độ đo thông dụng dùng để tính độ chính xác bài toán hồi quy

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

  - Tập trung đo các sai số lớn hơn là các sai số nhỏ


Đo hiệu năng bài toán hồi quy

- Mục tiêu: xây dựng mô hình khái quát hóa (*generalizes)*
  - Ta muốn cực tiểu hóa lỗi trên dữ liệu chưa biết, không phải trên dữ liệu học.
  - Vd: Dự đoán giá cổ phiếu *trong tương lai* vs. giá cổ phiếu trong quá khứ

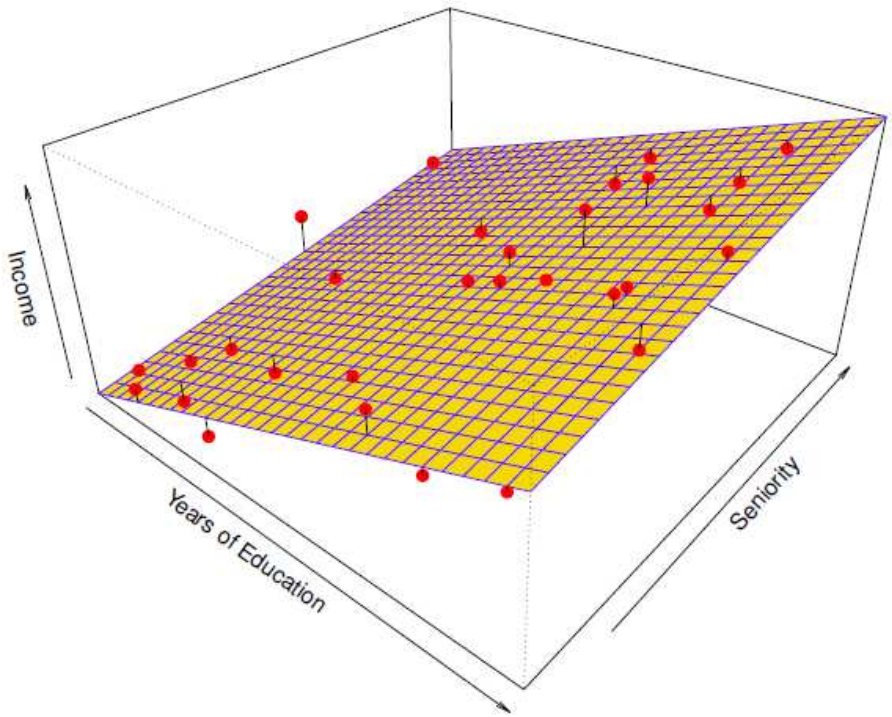
- Chúng ta muốn cực tiểu tổn thất kỳ vọng (*expected* loss)
  - Vấn đề: Ta không thể cực tiểu lỗi trên dữ liệu huấn luyện.

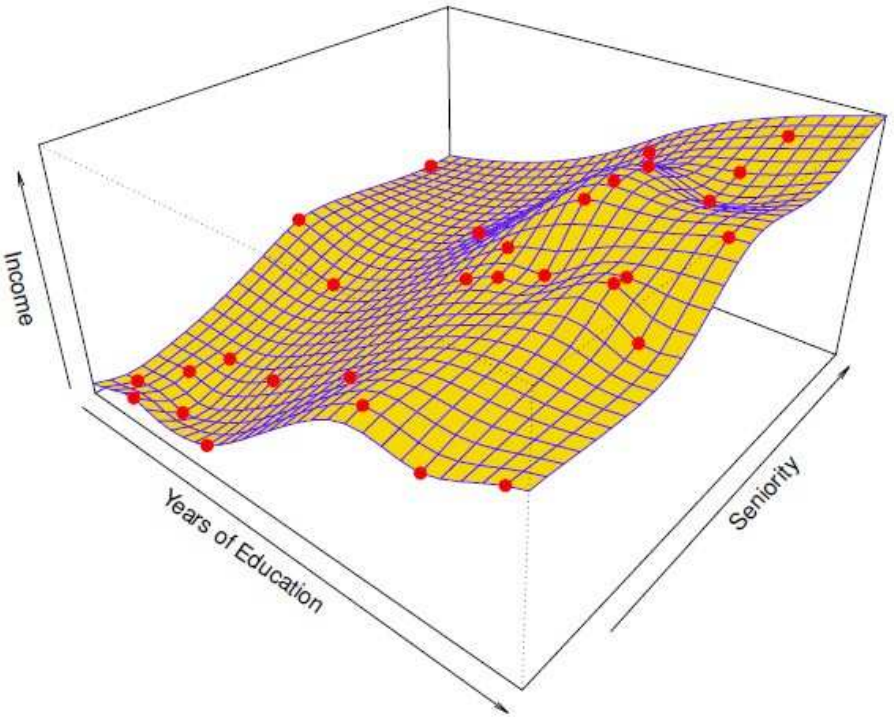
Vấn đề: Overfitting

- *Quá khớp (Overfitting)*: Học sự biến thiên ngẫu nhiên trong dữ liệu hơn là xu hướng cơ bản

- Đặc điểm của overfitting:
  - Mô hình có hiệu năng cao trên dữ liệu học nhưng kém trên tập dữ liệu thử nghiệm.

Vấn đề: Overfitting





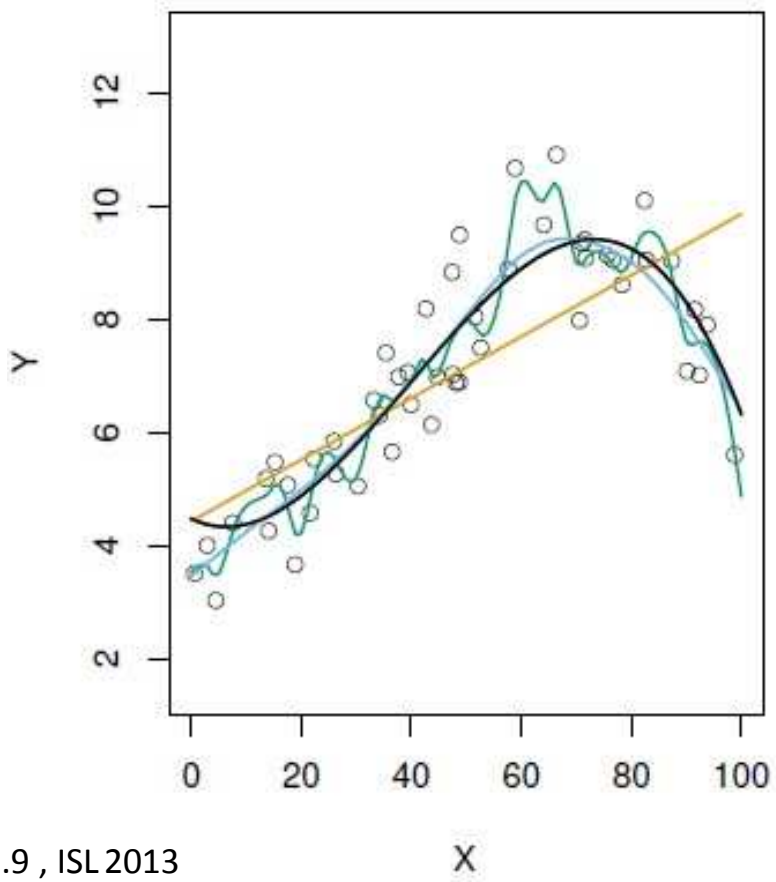
Figures 2.4 and 2.6 , ISL 2013

Đánh giá hiệu năng

- Lỗi huấn luyện và lỗi kiểm thử thể hiện khác nhau
  - Tính linh hoạt của mô hình tăng lên…
  - *Lỗi huấn luyện* giảm
  - *Lỗi kiểm thử ban đầu* giảm,

    Nhưng sau đó tăng lên vì overfitting →"U-shaped" lỗi kiểm thử dạng chữ U.

Đánh giá hiệu năng



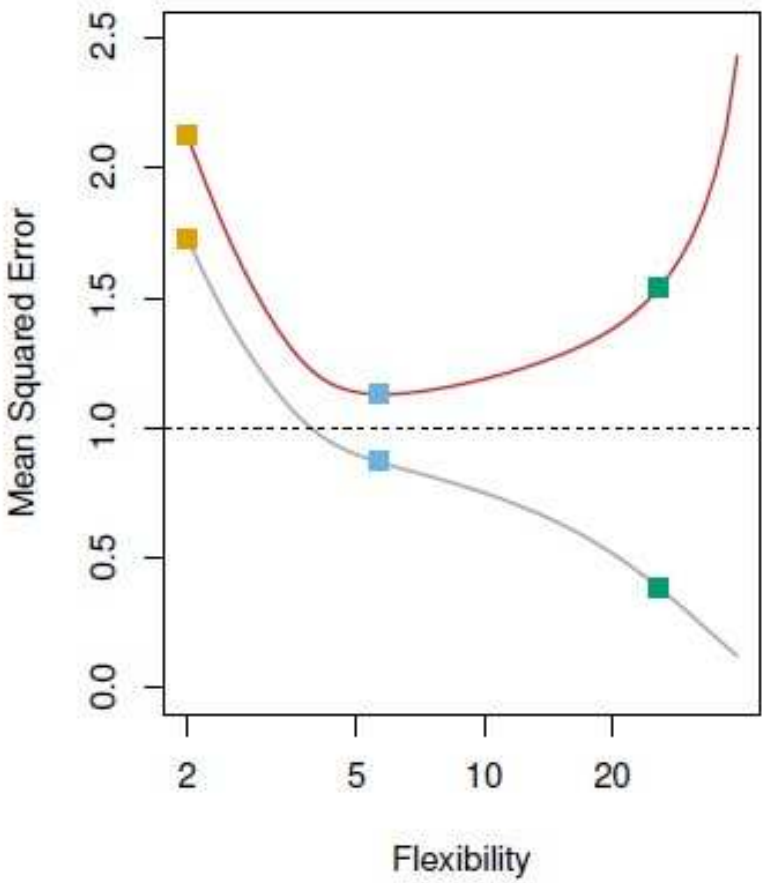


Figure 2.9 , ISL 2013

Đánh giá hiệu năng

- Làm sao để ước lượng lỗi kiểm thử để tìm một mô hình tốt?

- *Kỹ thuật kiểm tra chéo (Cross-validation):*

  một tập các kỹ thuật nhằm sử dụng dữ liệu huấn luyện để ước lượng lỗi tổng quát (generalization error)

Dữ liệu

- *Dữ liệu huấn luyện (Training data)*
  - Tập các quan sát (bản ghi) được sử dụng để xây dựng (học) mô hình.
- *Dữ liệu kiểm chứng (Validation data)*
  - Tập các quan sát dùng để ước lượng lỗi nhằm tìm tham số hoặc lựa chọn mô hình.
- *Dữ liệu kiểm thử (Test data)*
  - Tập các quan sát dùng để đánh giá hiệu năng trên dữ liệu chưa biết (unseen) trong tương lai.
  - Dữ liệu này không sử dụng cho giải thuật học máy trong quá trình xây dựng mô hình.


Trade-off: Độ lệch vs. Phương sai

- Lỗi kiểm thử đường cong hình chữ U (U-shaped) xảy ra dựa trên 2 đặc điểm của mô hình học máy:

$$\mathbb{E}\left[\text{test error}\right] = var(\hat{f}) + bias(\hat{f})^2 + var(\epsilon)$$

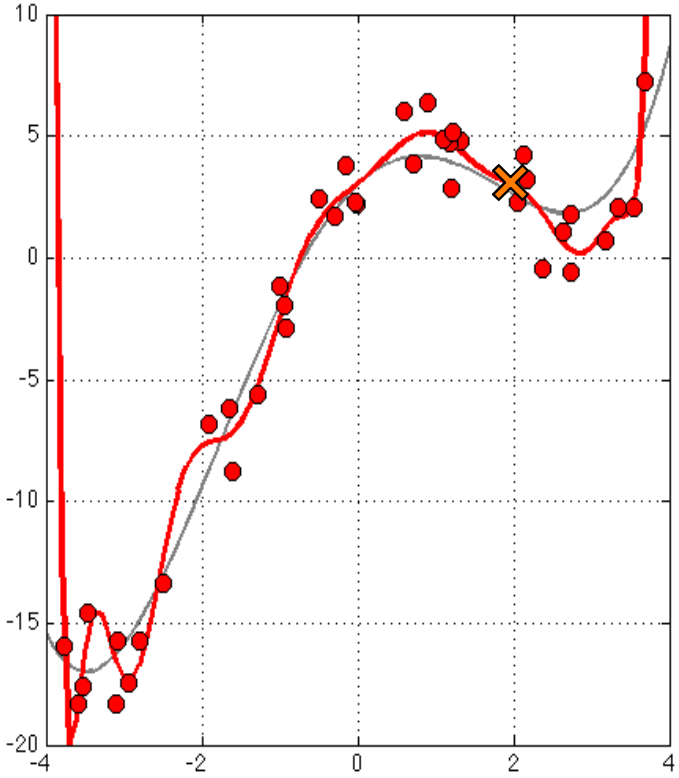
  - $var(\hat{f})$: *Phương sai (variance)* của hàm ước lượng
  - $bias(\hat{f})$: *Độ chệch/sai lệch (bias)* của hàm ước lượng

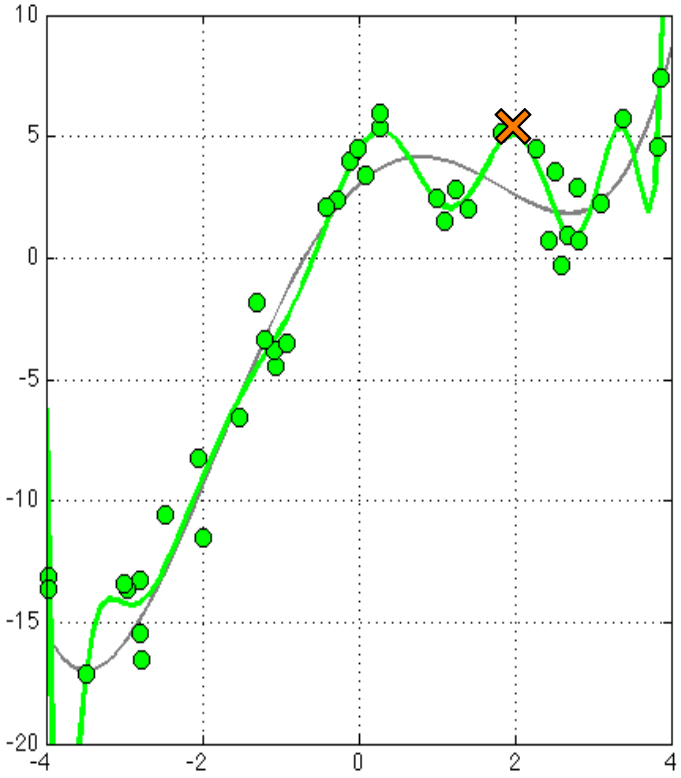
Trade-off: Độ lệch vs. Phương sai

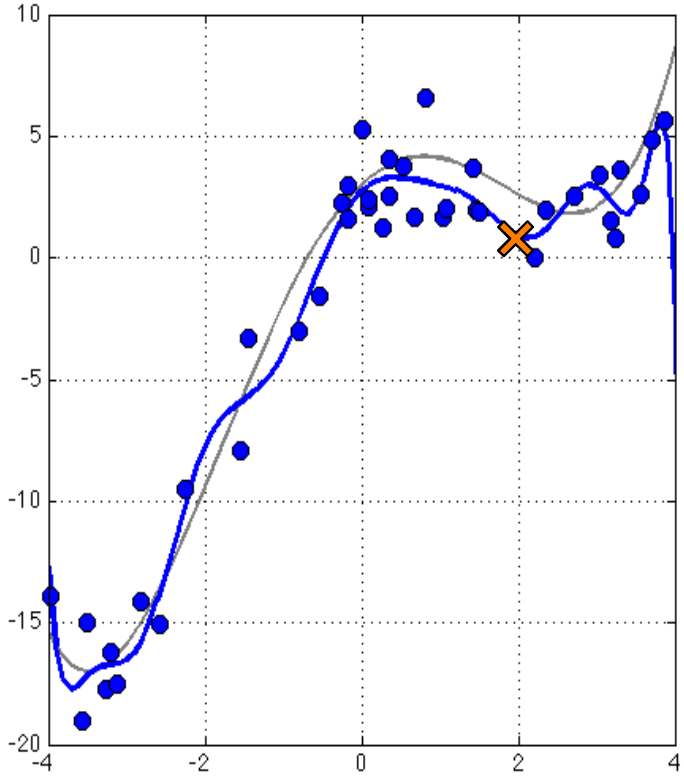
- *Phương sai của hàm ước lượng*
  - Chênh lệch giữa kết quả áp dụng mô hình với các quan sát đầu vào khác nhau.

- Phương sai cao: các thay đổi nhỏ trong tập huấn luyện
  →Các thay đổi lớn trong hàm ước lượng thống kê.
  - Các phương pháp càng linh hoạt → Phương sai càng lớn.

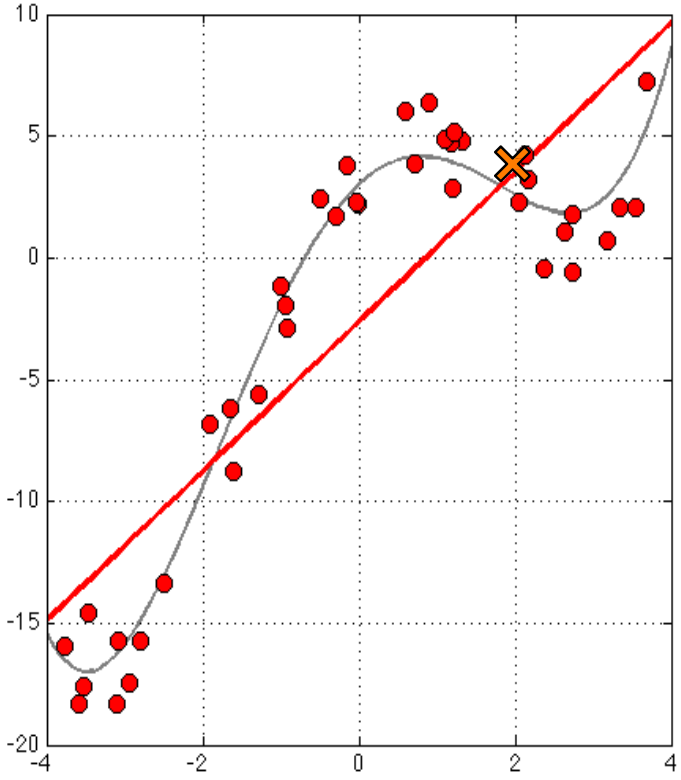
Trade-off: Độ lệch vs. Phương sai

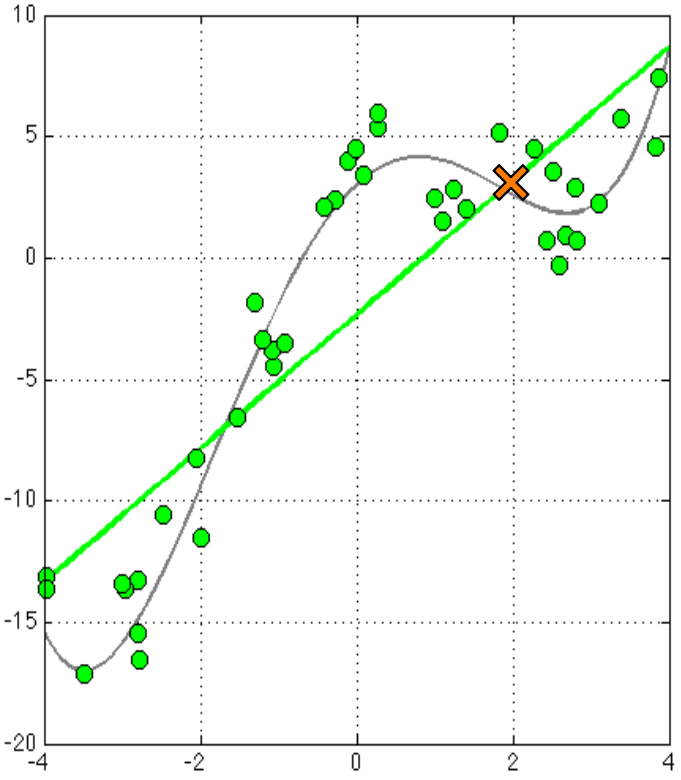


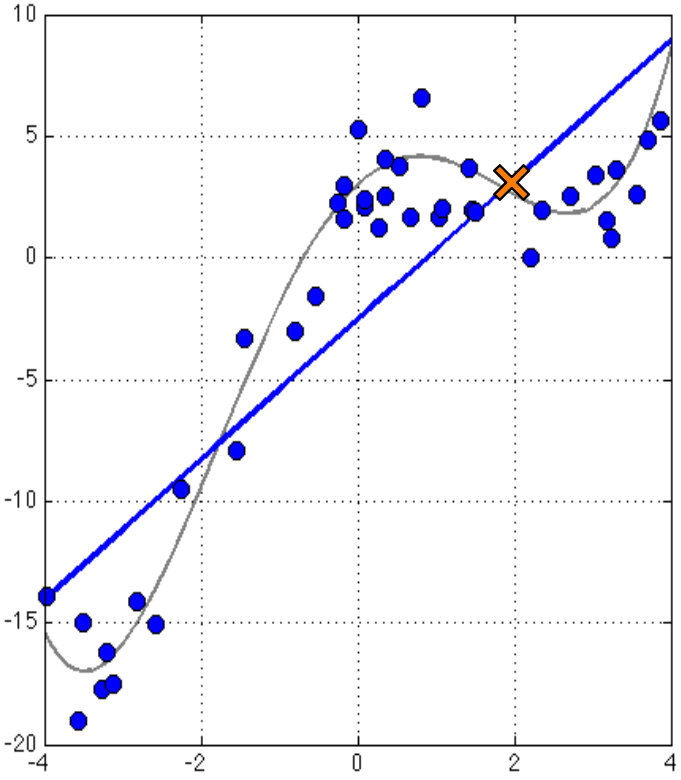




Trade-off: Độ lệch vs. Phương sai







Trade-off: Độ lệch vs. Phương sai

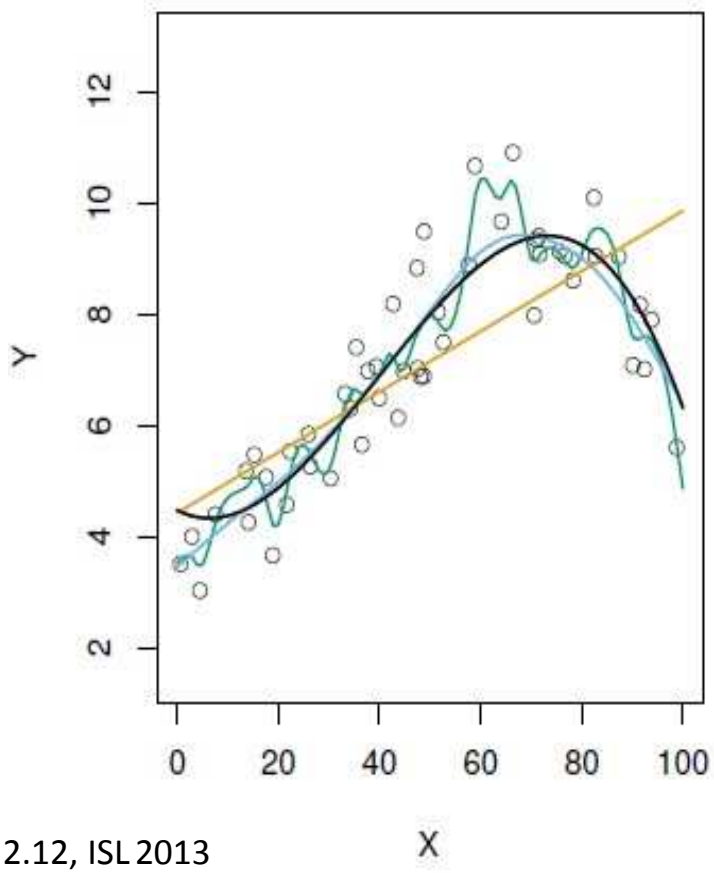
- *Độ lệch (bias) của hàm ước lượng*
  - Bias là độ sai lệch giữa kết quả dự đoán của mô hình và thực tế, sai số xấp xỉ một hàm khi áp dụng một mô hình đơn giản.
  - Vd: Hồi quy tuyến tính giả định các biến phải quan hệ tuyến tính.

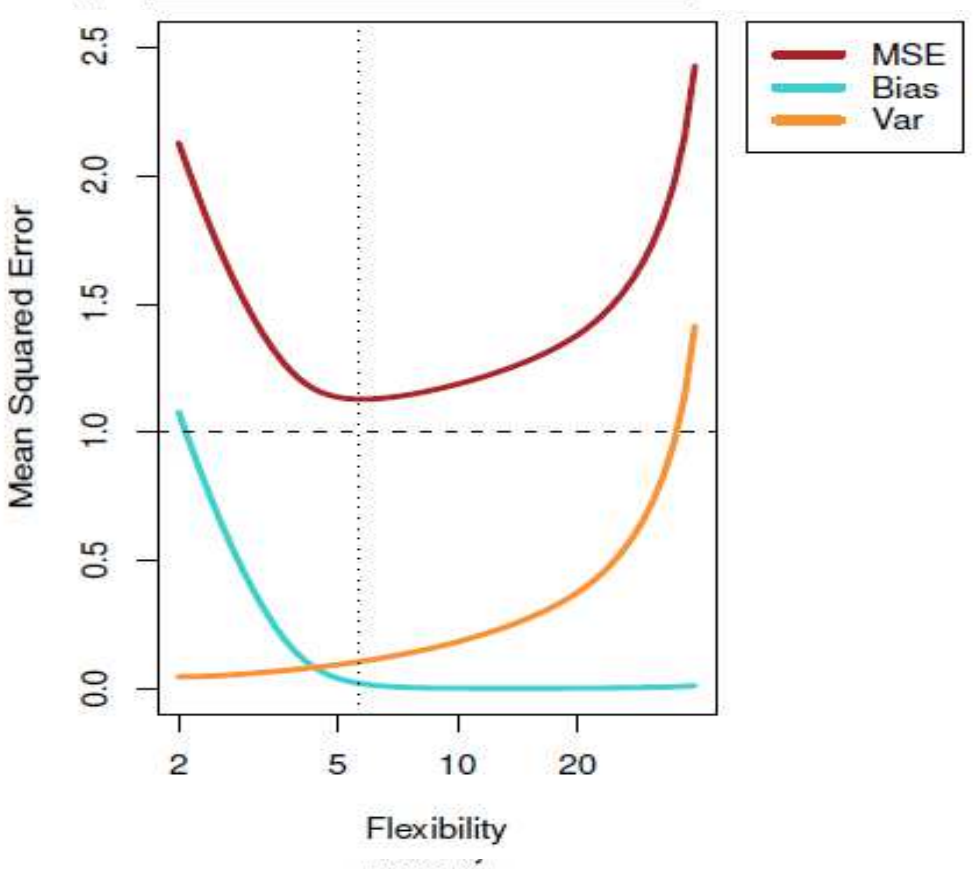
    → lỗi bias xuất hiện khi hệ thống là phi tuyến.
  - Các phương pháp càng linh hoạt → bias nhỏ.


Trade-off: Độ lệch vs. Phương sai

- Phương sai thấp *và* bias thấp → Lỗi kiểm thử cũng thấp.
- Càng linh hoạt (phức tạp) → Phương sai tăng, bias giảm.
- Lỗi kiểm thử đường cong hình chữ U (U-shaped):
  - Ban đầu độ linh hoạt mô hình tăng, ta thấy bias giảm nhanh hơn tăng phương sai → lỗi kiểm thử MSE giảm.
  - Độ linh hoạt của mô hình có ảnh hưởng nhỏ hơn đến việc giảm bias, tuy nhiên khi tăng độ linh hoạt nó ảnh hưởng lớn đến phương sai → lỗi kiểm thử MSE tăng.

Trade-off: Độ lệch vs. Phương sai





Figures 2.9, 2.12, ISL 2013

Trade-off: Độ lệch vs. Phương sai

- Phương pháp linh hoạt (phức tạp)
  - Có thể xấp xỉ sát hàm ước lượng thống kê (bias thấp),
  - Tuy nhiên các lỗi/rủi ro của mô hình học lại quá phụ thuộc vào dữ liệu huấn luyện (phương sai cao)
- Phương pháp đơn giản hơn
  - Có thể xấp xỉ hàm ước lượng với độ chính xác không cao (bias cao),
  - Tuy nhiên chúng ít phụ thuộc vào dữ liệu huấn luyện (phương sai thấp)
- Tradeloff
  - Dễ đạt được phương sai thấp/bias cao *hoặc* phương sai cao/bias thấp,
  - Tuy nhiên rất khó để đạt được cả phương sai *và* bias cùng thấp


Hồi quy:
Hồi quy tuyến tính

Hồi quy tuyến tính

- *Hồi quy tuyến tính*: là phương pháp học máy có giám sát đơn giản, được sử dụng để dự đoán giá trị biến đầu ra dạng số (định lượng)
  - Nhiều phương pháp học máy là dạng tổng quát hóa của hồi quy tuyến tính
  - Là ví dụ để minh họa các khái niệm quan trọng trong bài toán học máy có giám sát

Hồi quy tuyến tính

- Tại sao dùng hồi quy tuyến tính?
  - Mối quan hệ tuyến tính: là sự biến đổi tuân theo quy luật hàm bậc nhất
  - Tìm một mô hình (phương trình) để mô tả một mối liên quan giữa X và Y
  - Ta có thể biến đổi các biến đầu vào để tạo ra mối quan hệ tuyến tính
  - Diễn giải các mối quan hệ giữa biến đầu vào và đầu ra - sử dụng cho bài toán suy diễn


Hồi quy tuyến tính đơn giản

- Biến đầu ra Y và biến đầu vào X có mối quan hệ tuyến tính giữa X và Y như sau:

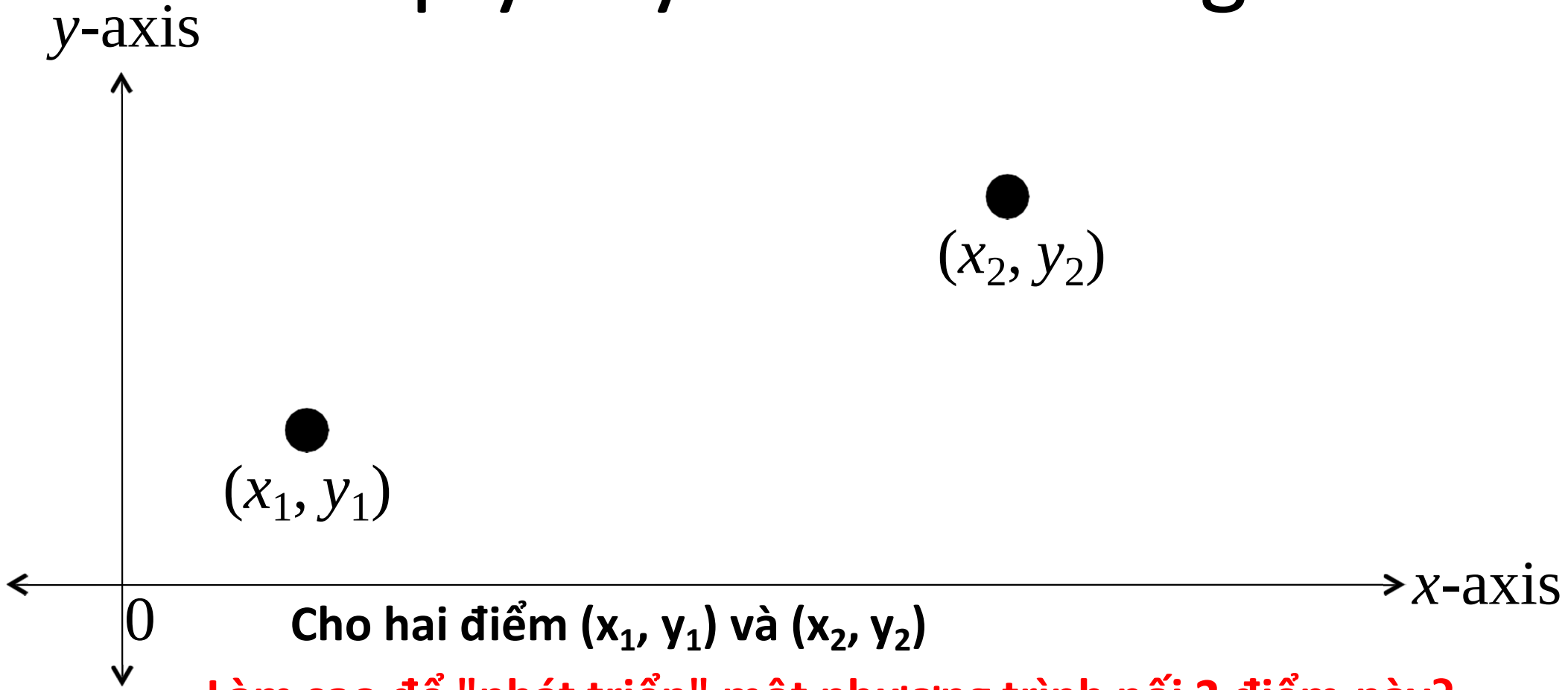
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Các tham số của mô hình:

β_0 intercept hệ số chặn (khi các x_i=0)

β_1 slope độ dốc

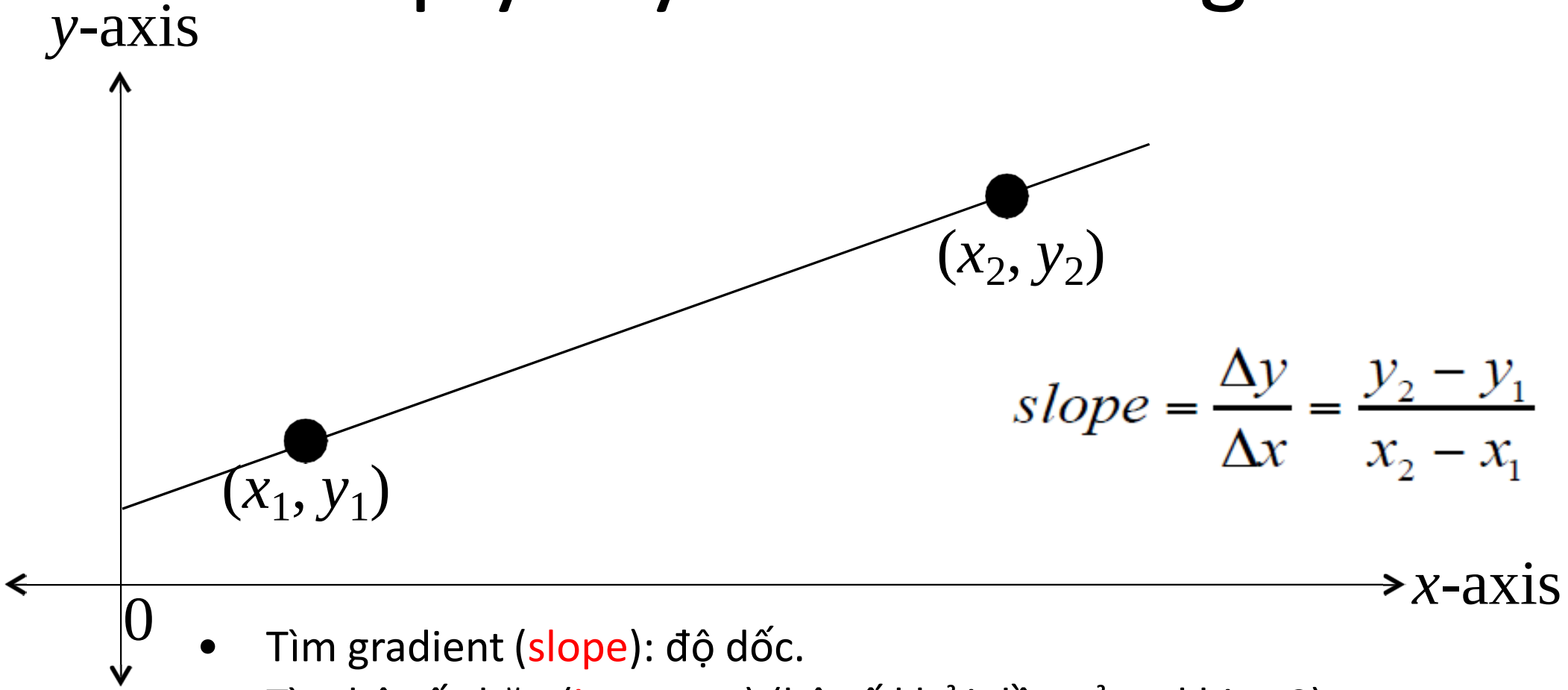
Hồi quy tuyến tính đơn giản



Cho hai điểm (x_1, y_1) và (x_2, y_2)

Làm sao để "phát triển" một phương trình nối 2 điểm này?

Hồi quy tuyến tính đơn giản



$$slope = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

- Tìm gradient (slope): độ dốc.
- Tìm hệ số chặn (intercept) (hệ số khởi đầu của y khi x=0)

Hồi quy tuyến tính đơn giản

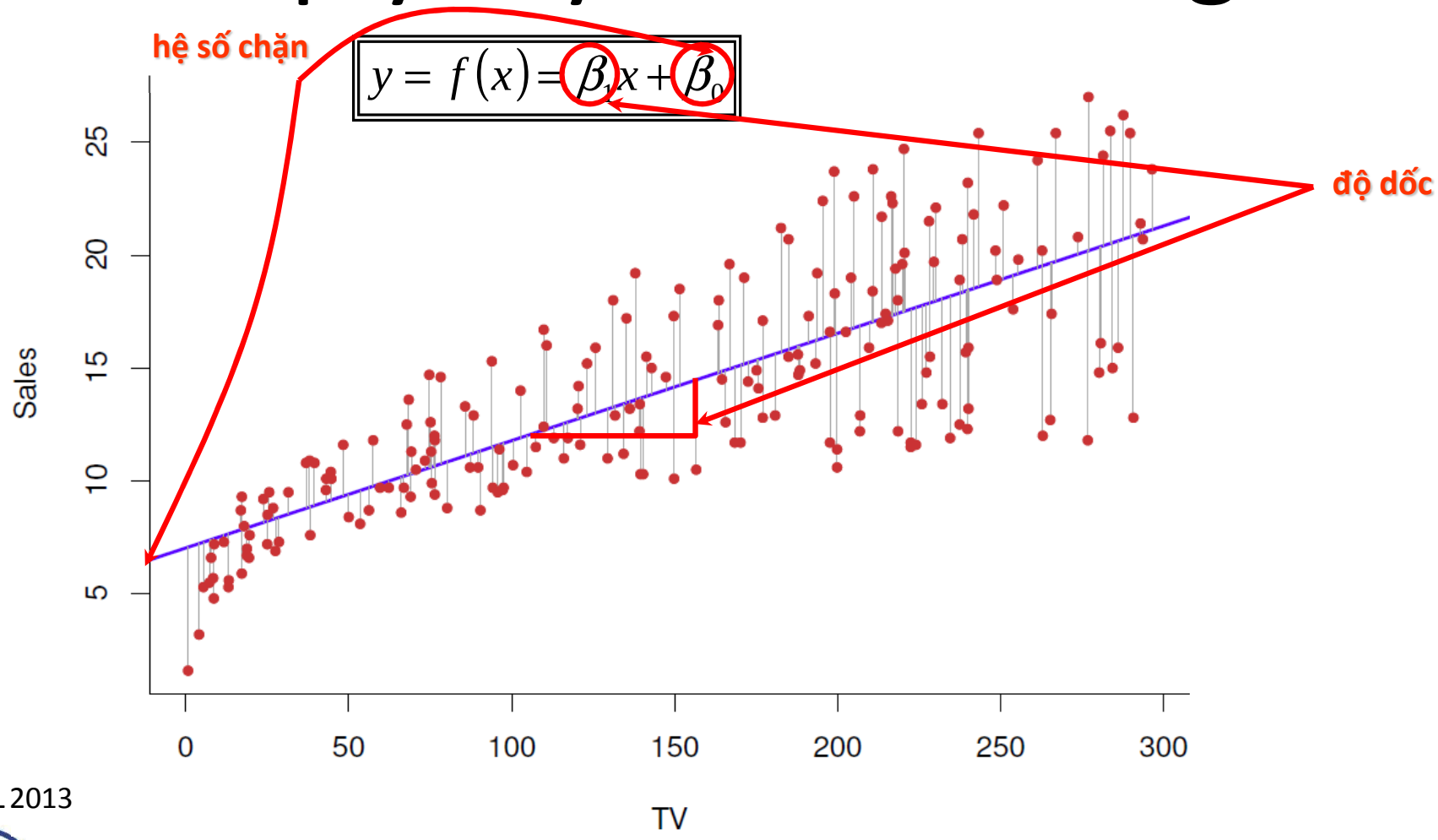


Figure 3.1 , ISL 2013

Hồi quy tuyến tính đơn giản

- β_0 và β_1 chưa biết → Ta ước tính giá trị của chúng từ dữ liệu đầu vào

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- Lấy $\hat{\beta}_0, \hat{\beta}_1$ sao cho mô hình đạt "xấp xỉ tốt nhất" ("good fit") đối với tập huấn luyện

$$Y^{(i)} \approx \hat{\beta}_0 + \hat{\beta}_1 X^{(i)}, \quad i = 1, \ldots, n$$

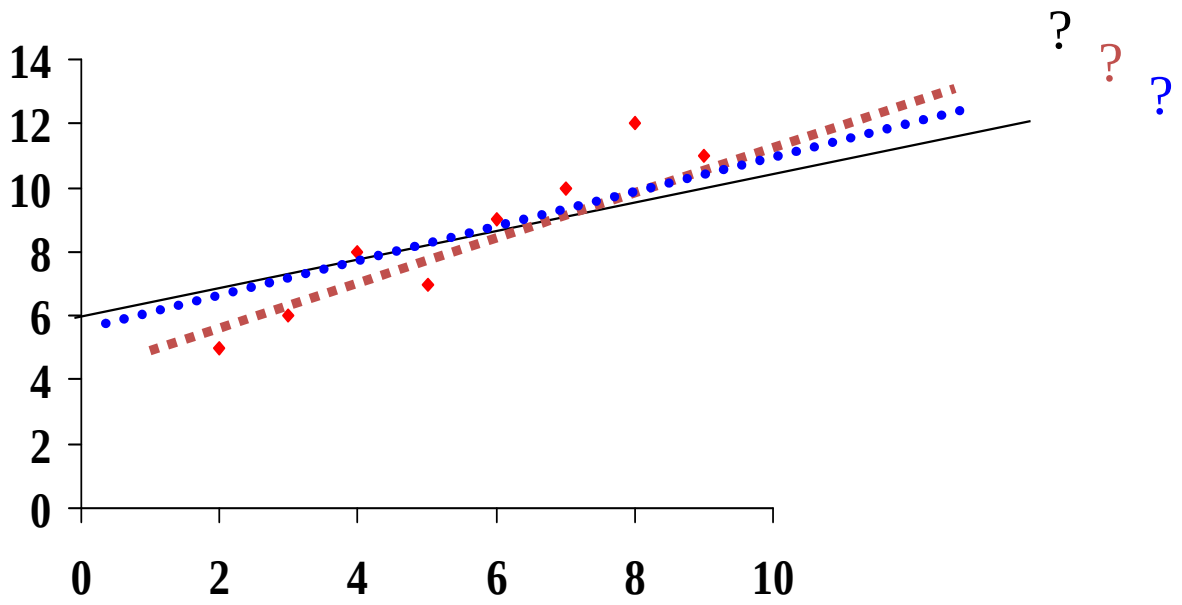
Các giả định

- Mối liên quan giữa X và Y là tuyến tinh (linear) về *tham số*
- *X* không có sai số ngẫu nhiên
- Giá trị của *Y* độc lập với nhau (vd, Y_1 không liên quan với Y_2) ;
- Sai số ngẫu nhiên (ε): phân bố chuẩn, trung bình 0, phương sai bất biến

$$\varepsilon \sim N(0, \sigma^2)$$

Đường thẳng phù hợp nhất

Cho tập dữ liệu đầu vào, ta cần tìm cách tính toán các tham số của phương trình đường thẳng



Bình phương nhỏ nhất

- Thông thường, để đánh giá độ phù hợp của mô hình từ dữ liệu quan sát ta sử dụng phương pháp *bình phương nhỏ nhất* (*least squares)*

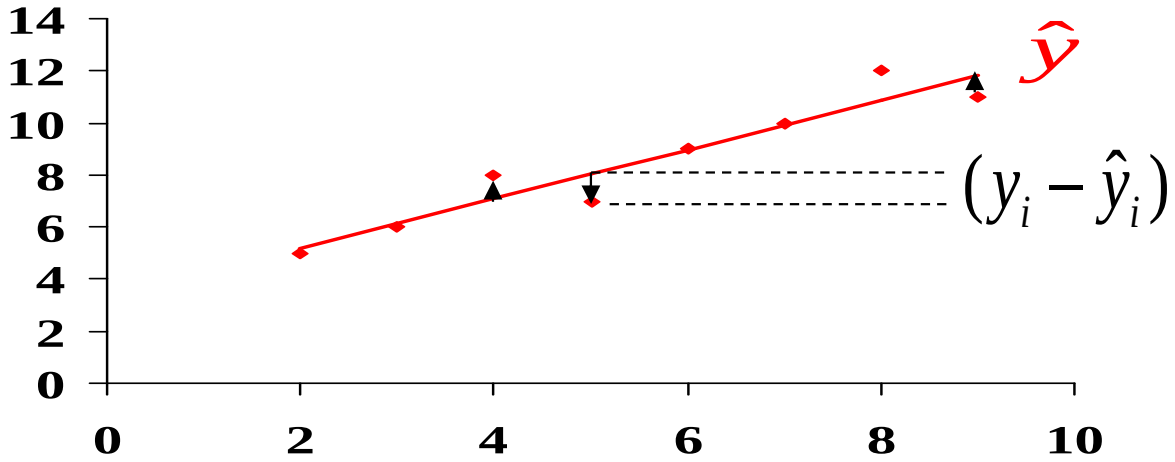
- Lỗi bình phương trung bình (Mean squared error):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y^{(i)} - \hat{Y}^{(i)}\right)^2$$

Đường thẳng phù hợp nhất

Rất hiếm để có 1 đường thẳng khớp chính xác với dữ liệu, do vậy luôn tồn tại lỗi gắn liền với đường thẳng

Đường thẳng phù hợp nhất là đường giảm thiểu độ dao động của các lỗi này



Phần dư (lỗi)

Biểu thức (y_i - $\hat{y}$) được gọi là lỗi hoặc *phần dư*

$$\varepsilon_i = (y_i - \hat{y})$$

Đường thẳng phù hợp nhất tìm thấy khi tổng bình phương lỗi là nhỏ nhất

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

Ước lượng tham số

- Các ước số $\hat{\beta}_0$, $\hat{\beta}_1$ tính được bằng cách cực tiểu hóa MSE

$$\min_{(\hat{\beta}_0, \hat{\beta}_1)} \left[\frac{1}{n} \sum_{i=1}^{n} \left(Y^{(i)} - \left(\hat{\beta}_0 + \hat{\beta}_1 X^{(i)} \right) \right)^2 \right]$$

  - Hệ số chặn của đường thẳng $\hat{\beta}_1 = \frac{SS_{xy}}{SS_x}$

trong đó: $SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ và $SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2$

Ước lượng tham số

Hệ số chặn của đường thẳng

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

trong đó

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Hồi quy tuyến tính đơn giản

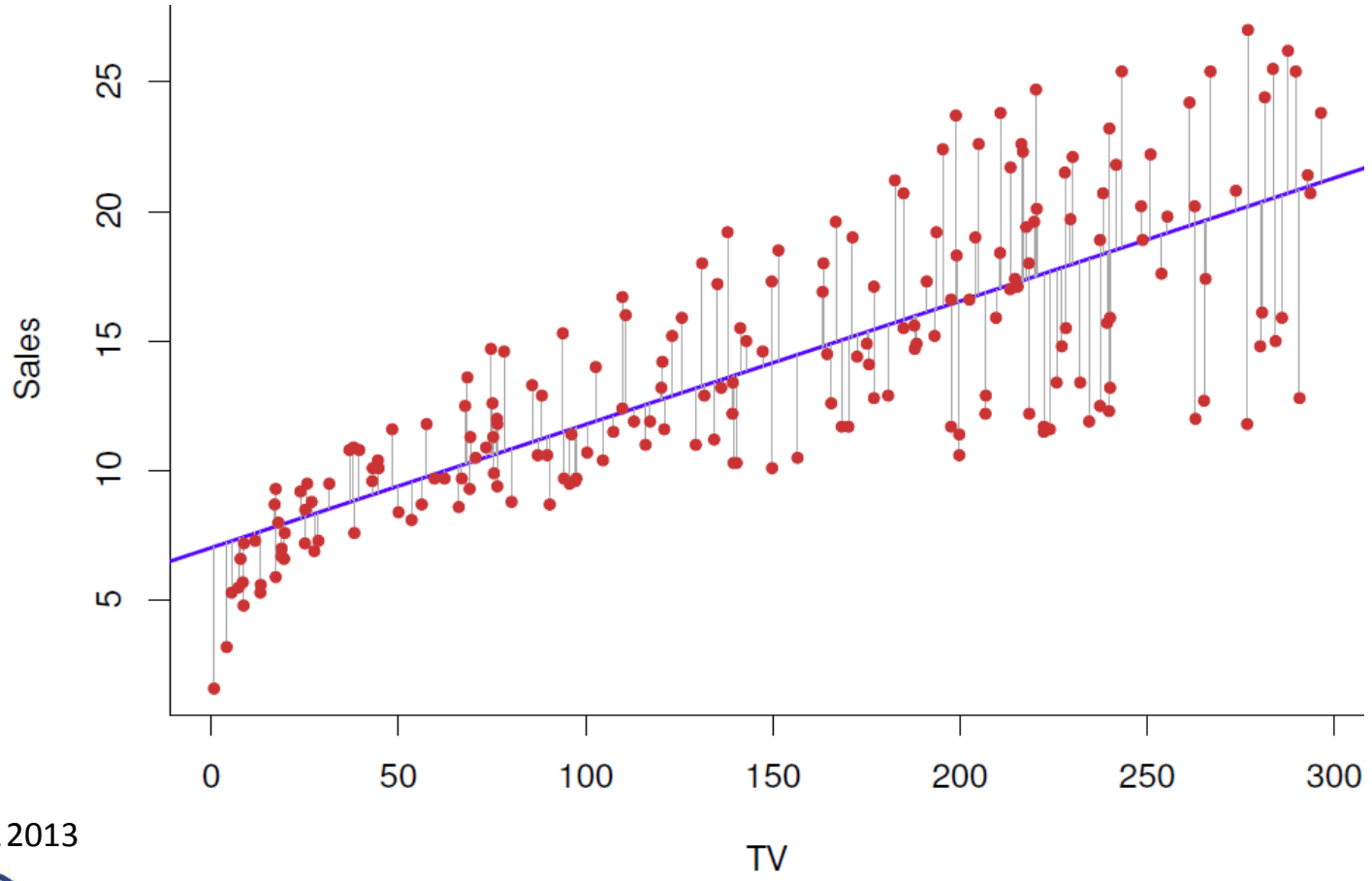
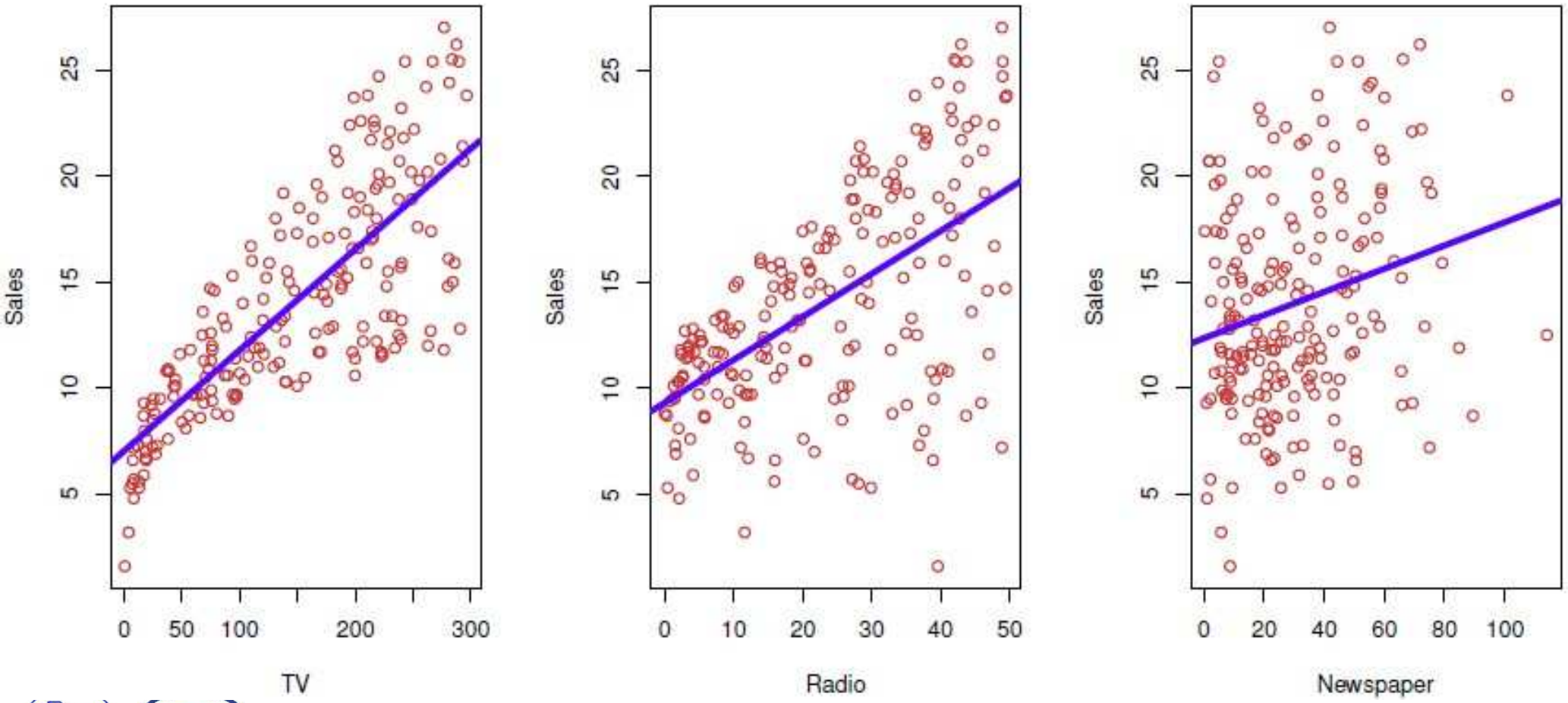


Figure 3.1 , ISL 2013

Hồi quy tuyến tính đơn giản



Ví dụ

X kilos	Y giá $
17	132
21	150
35	160
39	162
50	149
65	170

$$\bar{x} = 37.83$$

$$\bar{y} = 153.83$$

$$SS_{xy} = 891.83$$

$$SS_x = 1612.83$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{891.83}{1612.83} = 0.533$$

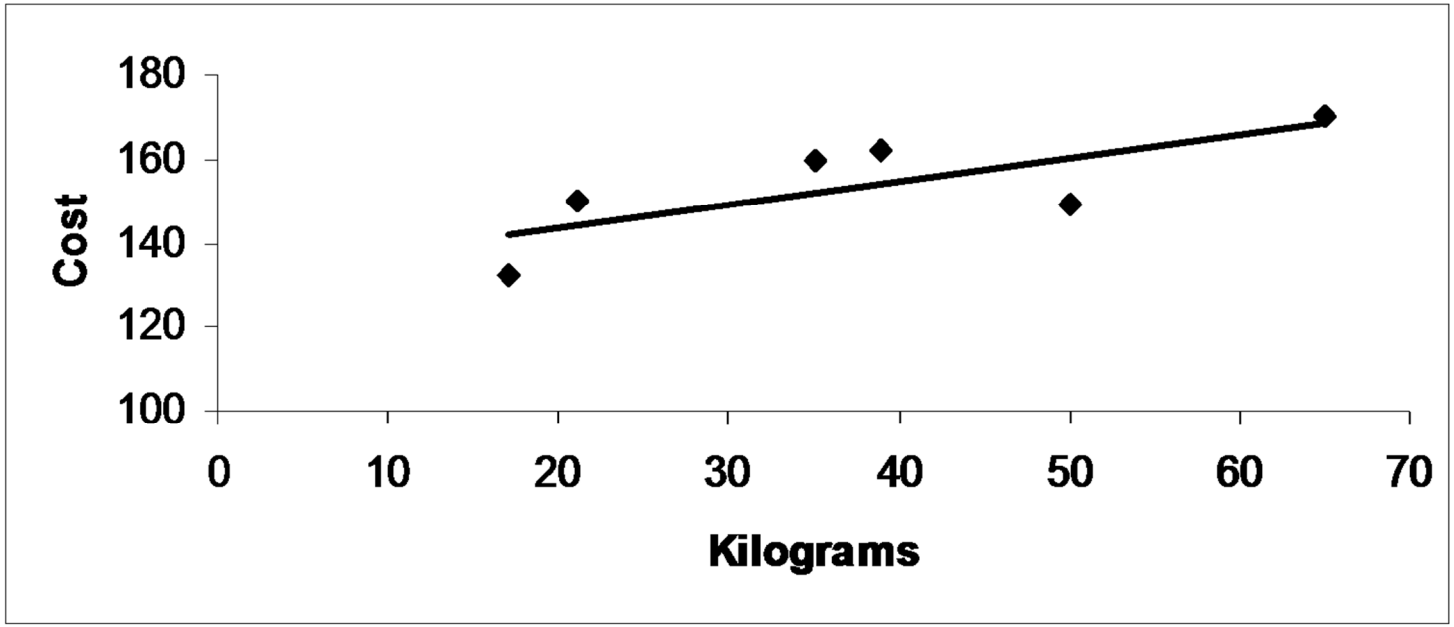
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 153.83 - 0.553 \times 37.83 = 132.91$$

phương trình tìm được là

$$Y = 132.91 + 0.553 * X$$

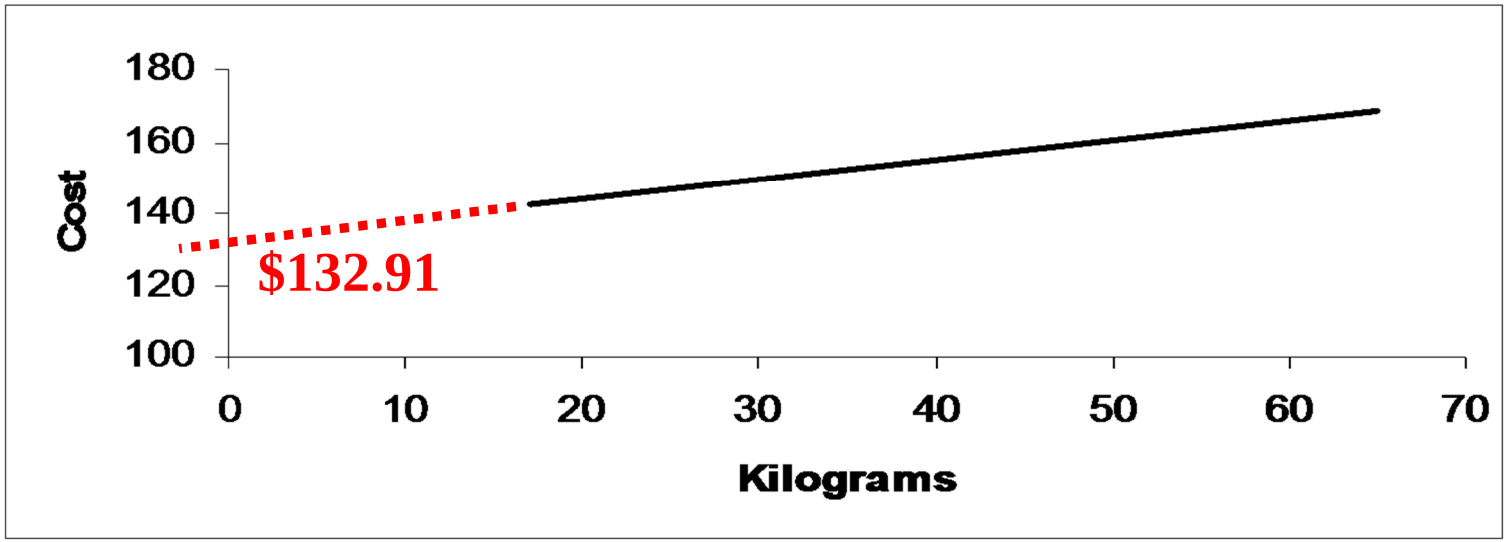
Diễn giải tham số

Trong ví dụ trước, tham số ước lượng $\hat{\beta}_1$ của độ dốc là 0.553. Điều này có nghĩa là khi thay đổi 1 kg của X, giá của Y thay đổi 0.553 $



Diễn giải tham số

$\hat{\beta}_0$ là hệ số chặn của Y. Nghĩa là, điểm mà đường thẳng cắt trục tung Y. Trong ví dụ này là $132.91



Đây là giá trị của Y khi X = 0

Dữ liệu phân tích: Boston

- **Boston data:** liên quan đến giá nhà đất
- Các biến số
  - **crim**: tỉ lệ tội phạm của thị trấn
  - **zn**: tỉ lệ khu đất có diện tich trên 25,000 feet vuông
  - **indus**: tỉ lệ doanh nghiệm tương đối lớn
  - **chas**: gần sông Charles (1=yes, 0=no)
  - **nos**: nồng độ nitric oxides (parts/10 triệu)
  - **rm**: số phòng trung bình mỗi nhà
  - **age**: tỉ lệ căn hộ (unit) xây trước 1940
  - **dis**: khoảng cách đến các trung tâm kĩ nghệ (tìm việc làm)

Dữ liệu phân tích: Boston

- **Boston data:** liên quan đến giá nhà đất
- Các biến số
  - **rad**: chỉ số gần xa lộ radial
  - **tax**: tỉ suất thuế tinh trên $10,000
  - **ptratio:** tỉ số học trò trên giáo viên của thị trấn
  - **black**: chỉ số về số người da đen trong thị trấn (Bk – 0.63)^2
  - **lstat**: tỉ lệ dân số thành phần kinh tế thấp
  - **medv:** trị giá nhà ($1000)

Ước tính bằng R

- Chúng ta muốn ước tinh mối liên quan giữa số phòng (rm) và giá căn nhà
- Mô hình hồi qui tuyến tinh:

$$medv = \alpha + \beta * rm + \varepsilon$$

- R

```
lm(medv ~ rm, data=Boston)
```

Phân tích bằng R

```
attach(Boston)
# Phân tích hồi qui tuyến tính
m1 = lm(medv ~ rm, data= Boston)
summary(m1)

# vẽ biểu đồ
plot(medv ~ rm, pch=16)
abline(m1, col="red")
```

Phân tích bằng R

```
Residuals:
    Min      1Q  Median      3Q     Max
-23.346  -2.547   0.090   2.986  39.433

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -34.671      2.650  -13.08   <2e-16 ***
rm             9.102      0.419   21.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 6.616 on 504 degrees of freedom
Multiple R-squared:  0.4835,    Adjusted R-squared:  0.4825
F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

Diễn giải kết quả

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -34.671      2.650  -13.08   <2e-16 ***
rm              9.102      0.419   21.72   <2e-16 ***
```

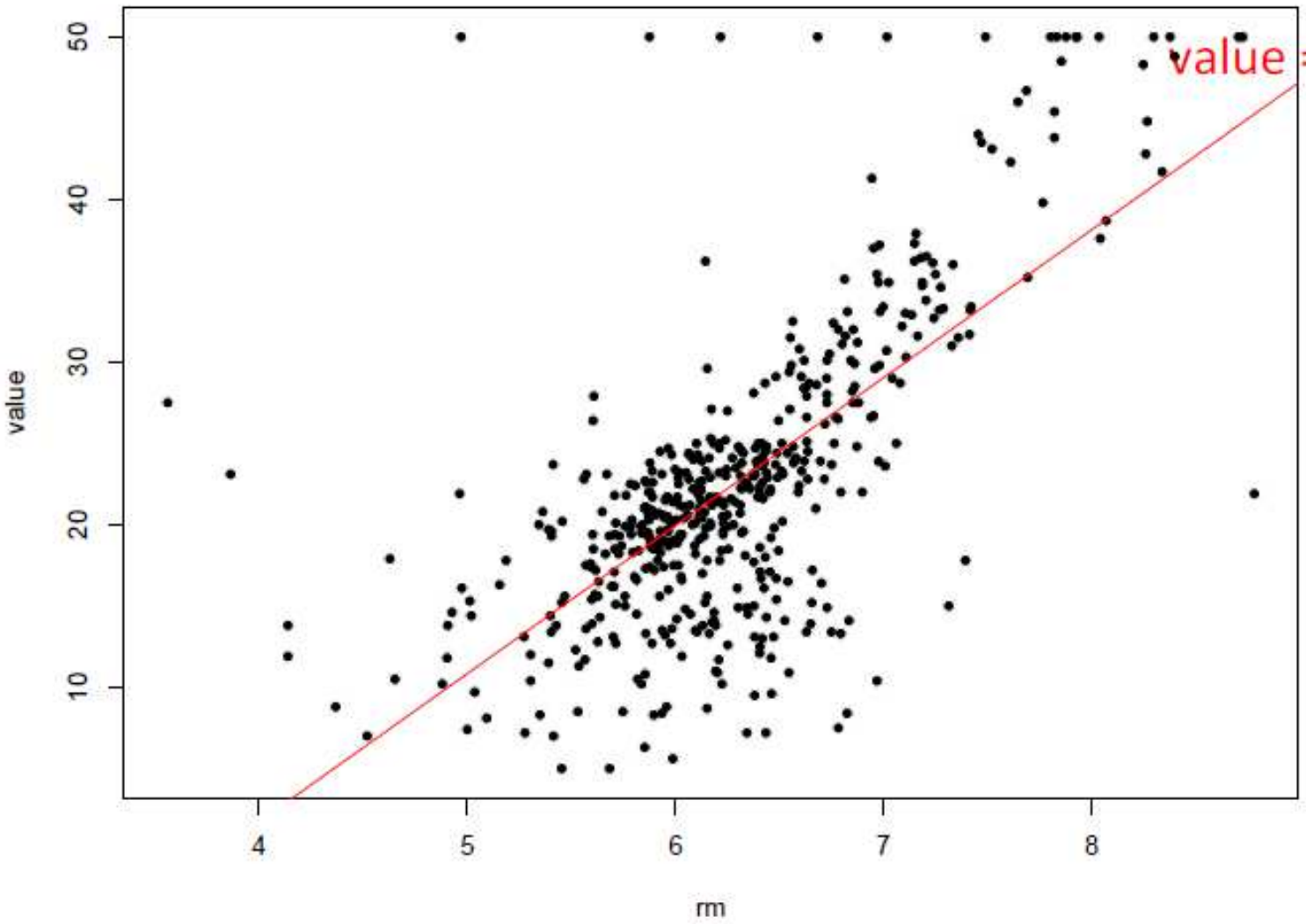
- Nhớ rằng mô hình là:

medv= a + b*rm

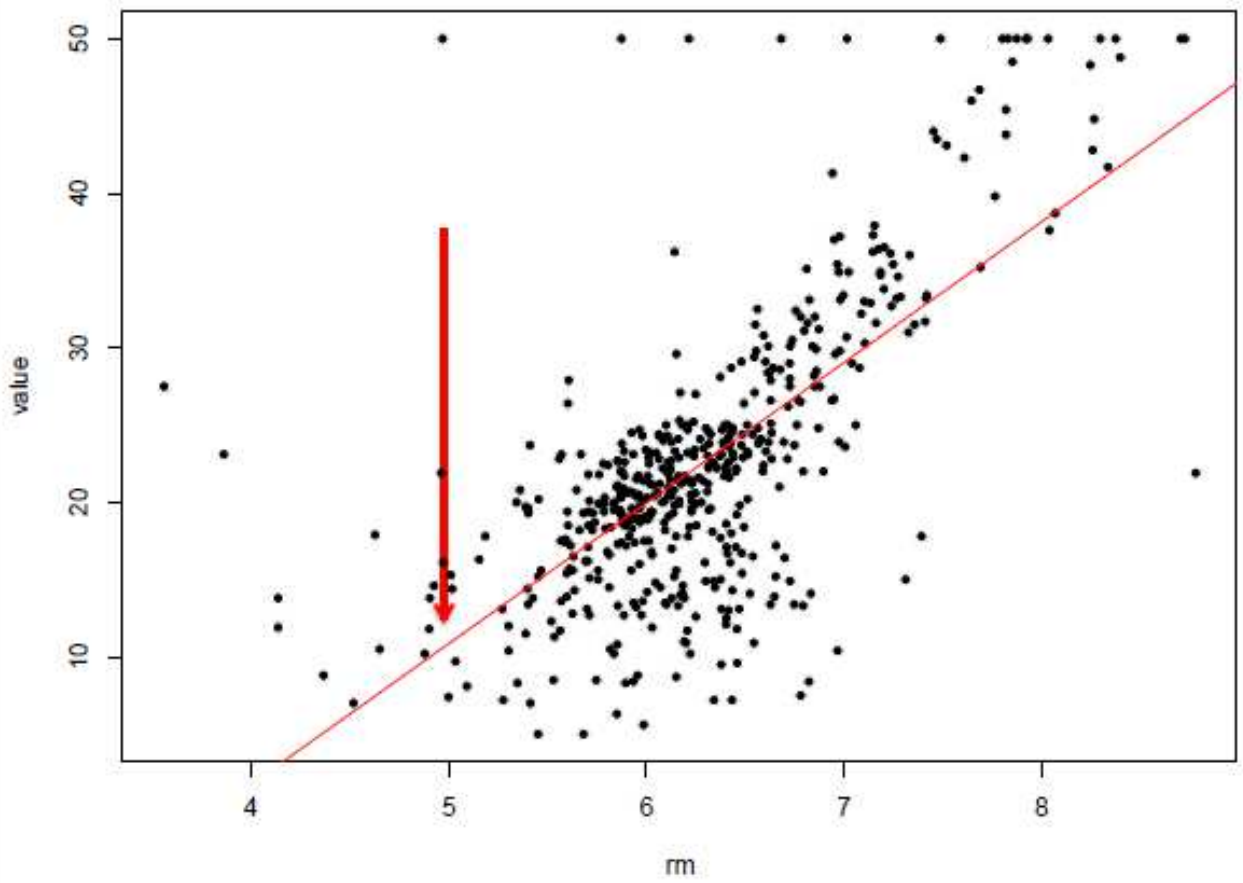
- Phương trình:

medv= -34.67 + 9.10*rooms

- Ý nghĩa: nhà có thêm 1 phòng tăng 9100 USD cho giá trị căn nhà. Mối tương quan này có **ý nghĩa thống kê (P < 0.0001)**


value = -34.67 + 9.10*rooms
value
rm
10
20
30
40
50
4
5
6
7
8

Ý nghĩa của đường biểu diễn



Giá trị trung bình (kì vọng)

medv= -34.67 + 9.10*rooms

Khi room = 5,
medv= -34.67 + 9.10*5 = **10.83**

Khi room **= 6**
medv= -34.67 + 9.10*6 = **19.93**

Khi room **= 8**
medv= -34.67 + 9.10*8 = **38.13**

Hồi quy tuyến tính đa biến

- Hồi quy tuyến tính đa biến: mô hình có nhiều hơn 1 biến dùng để dự đoán biến đích

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d + \epsilon$$

Hồi quy tuyến tính đa biến

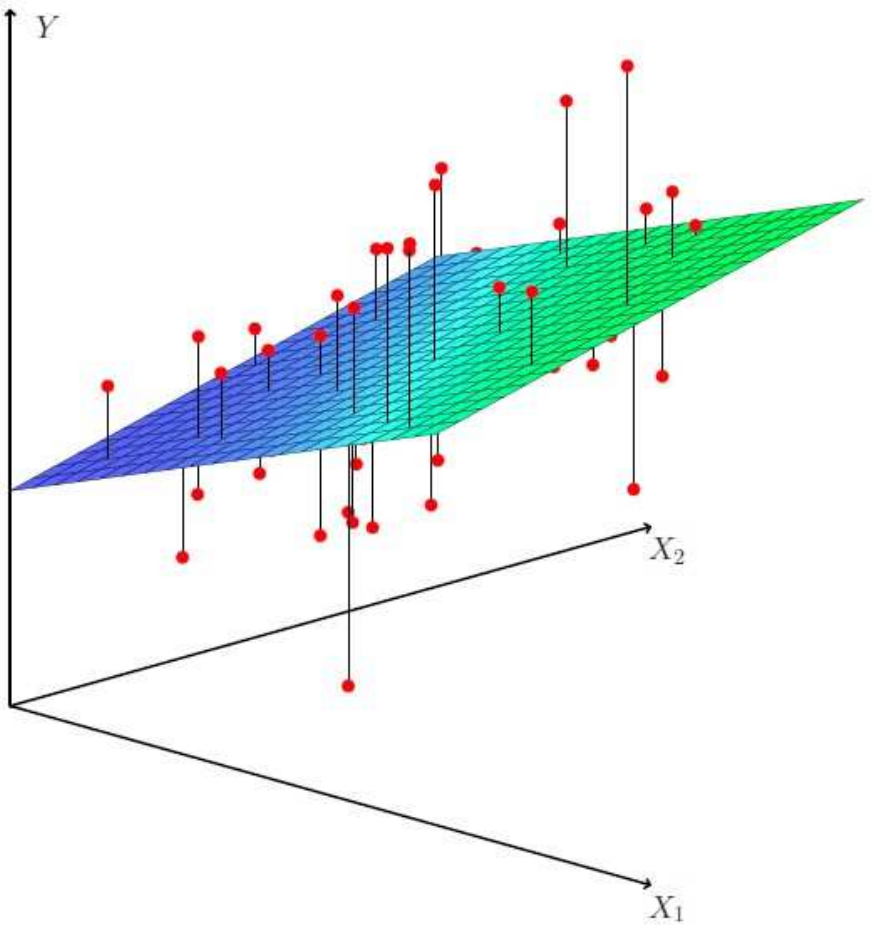


Figure 3.4 , ISL 2013

Hồi quy tuyến tính đa biến

- Diễn giải hệ số β_j:

khi tăng X_j lên một đơn vị → Y sẽ tăng trung bình một lượng là β_j

	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	−0.001

Bình phương nhỏ nhất

- Tìm các ước số bằng phương pháp bình phương nhỏ nhất

$$\hat{\beta} = \arg\min_{\beta} \|Y - X^T\beta\|^2 \qquad X = \begin{bmatrix} 1 & X^{(1)T} \\ & \cdots \\ 1 & X^{(n)T} \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdots \\ \hat{\beta}_d \end{bmatrix}$$

$$Y = \begin{bmatrix} Y^{(1)} \\ \cdots \\ Y^{(n)} \end{bmatrix}$$

- Giải phương trình để tìm $\hat{\beta}$:

$$X^T X\hat{\beta} = X^T Y \quad \rightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T Y$$

Hồi quy tuyến tính đa biến

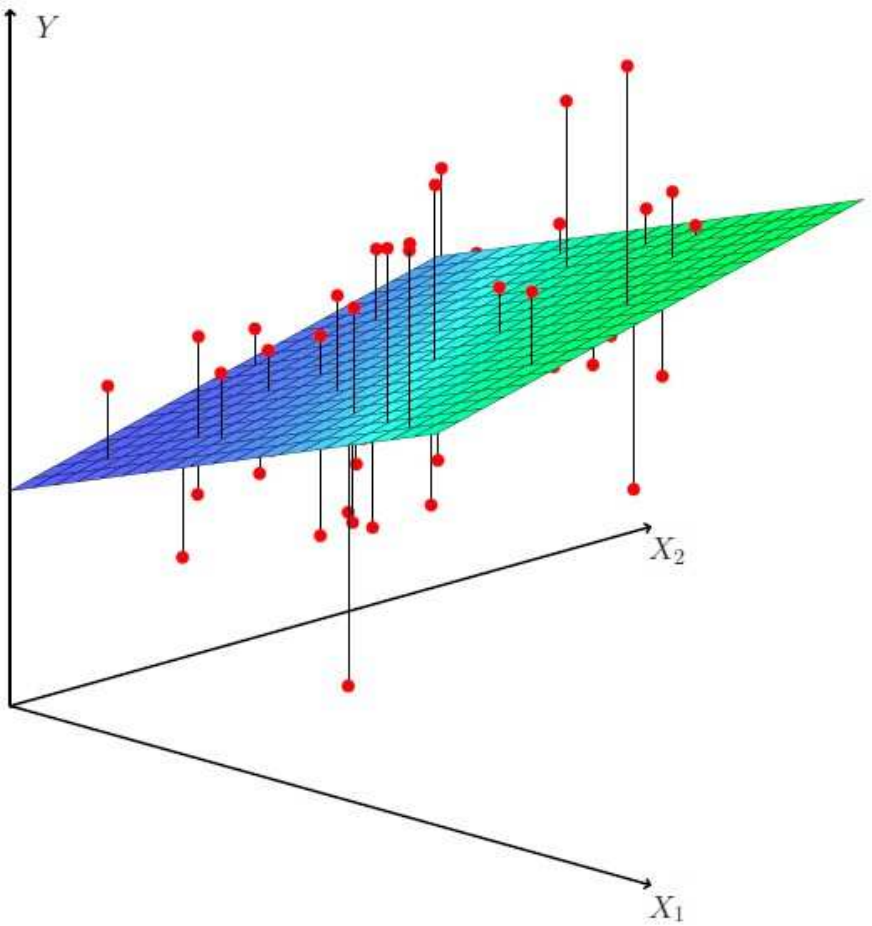


Figure 3.4 , ISL 2013

Ví dụ

Cho

$$y = \begin{bmatrix} 6 \\ 9 \\ 12 \\ 5 \\ 13 \\ 2 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 3 & 9 & 16 \\ 1 & 6 & 13 & 13 \\ 1 & 4 & 3 & 17 \\ 1 & 8 & 2 & 10 \\ 1 & 3 & 4 & 9 \\ 1 & 2 & 4 & 7 \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

Ví dụ

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 6 & 4 & 8 & 3 & 2 \\ 9 & 13 & 3 & 2 & 4 & 4 \\ 16 & 13 & 17 & 10 & 9 & 7 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 6 & 26 & 35 & 72 \\ 26 & 138 & 153 & 315 \\ 35 & 153 & 295 & 448 \\ 72 & 315 & 448 & 944 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 47 \\ 203 \\ 277 \\ 598 \end{bmatrix}$$

Ví dụ

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 2.59578 & -0.15375 & -0.01962 & -0.13737 \\ -0.15375 & 0.03965 & -0.00014 & -0.00144 \\ -0.01962 & -0.00014 & 0.01234 & -0.00431 \\ -0.13737 & -0.00144 & -0.00431 & 0.01406 \end{bmatrix} \begin{bmatrix} 47 \\ 203 \\ 277 \\ 598 \end{bmatrix}$$

$$= \begin{bmatrix} 3.20975 \\ -0.07573 \\ -0.11162 \\ 0.46691 \end{bmatrix}$$

$$\hat{\beta}_0 = 3.20975 \quad \hat{\beta}_1 = -0.07573 \quad \hat{\beta}_2 = -0.11162 \quad \hat{\beta}_3 = 0.46691$$

$$\hat{y} = 3.20975 - 0.07573x_1 - 0.11162x_2 + 0.46691x_3$$

Dữ liệu định tính

- Xử lý dữ liệu dạng định tính (định danh, hạng mục) trong mô hình hồi quy tuyến tính
  - vd: biến "giới tính": "male" hoặc "female"

- Nếu chỉ có 2 khả năng trên, ta tạo *biến giả* (*dummy variable)*

$$X_j = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

Dữ liệu định tính

- Nếu có nhiều hơn 2 giá trị, ta biểu diễn biến chúng dùng nhiều biến giả
  - vd: biến "màu mắt": "blue", "green" or "brown"

$$X_j = \begin{cases} 1 & \text{if blue} \\ 0 & \text{if not blue} \end{cases}$$

$$X_{j+1} = \begin{cases} 1 & \text{if brown} \\ 0 & \text{if not brown} \end{cases}$$

Hồi quy tuyến tính

- Ưu điểm:
  - Mô hình đơn giản, dễ hiểu
  - Dễ diễn giải hệ số hồi quy
  - Nhận được kết quả tốt khi dữ liệu quan sát nhỏ
  - Nhiều cải tiến/mở rộng

- Nhược điểm:
  - Mô hình hơi đơn giản nên khó dự đoán chính xác với dữ liệu có miền giá trị rộng
  - Khả năng ngoại suy (extrapolation) kém
  - Nhạy cảm với dữ liệu ngoại lai (outliers) – do dung phương pháp bình phương nhỏ nhất


Câu hỏi?