Statistical Learning

https://github.com/ggorr/Machine-Learning/tree/master/ISLR

Data

- http://faculty.marshall.usc.edu/gareth-james/ISL/data.html

- Example: Advertising.csv

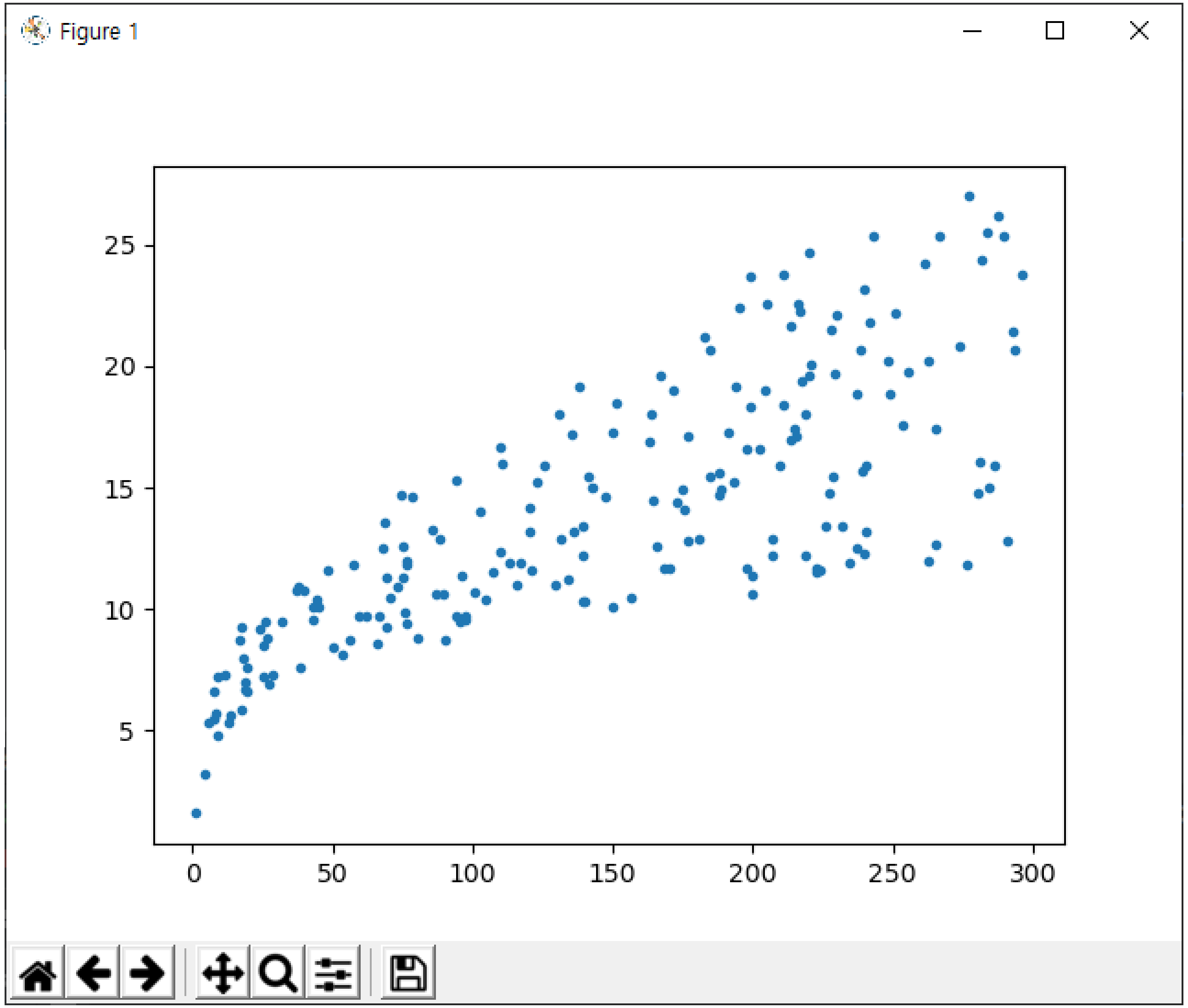
		TV	radio	newspape	sales
1		TV	radio	newspape	sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8

```python
import pandas as pd

df = pd.read_csv('data/Advertising.csv', sep=',')
print(df.keys())
print(df.values.shape)
tv = df.TV.values
print(tv.shape)
```

```
Index(['Unnamed: 0', 'TV', 'radio', 'newspaper', 'sales'], dtype='object')
(200, 5)
(200,)
```

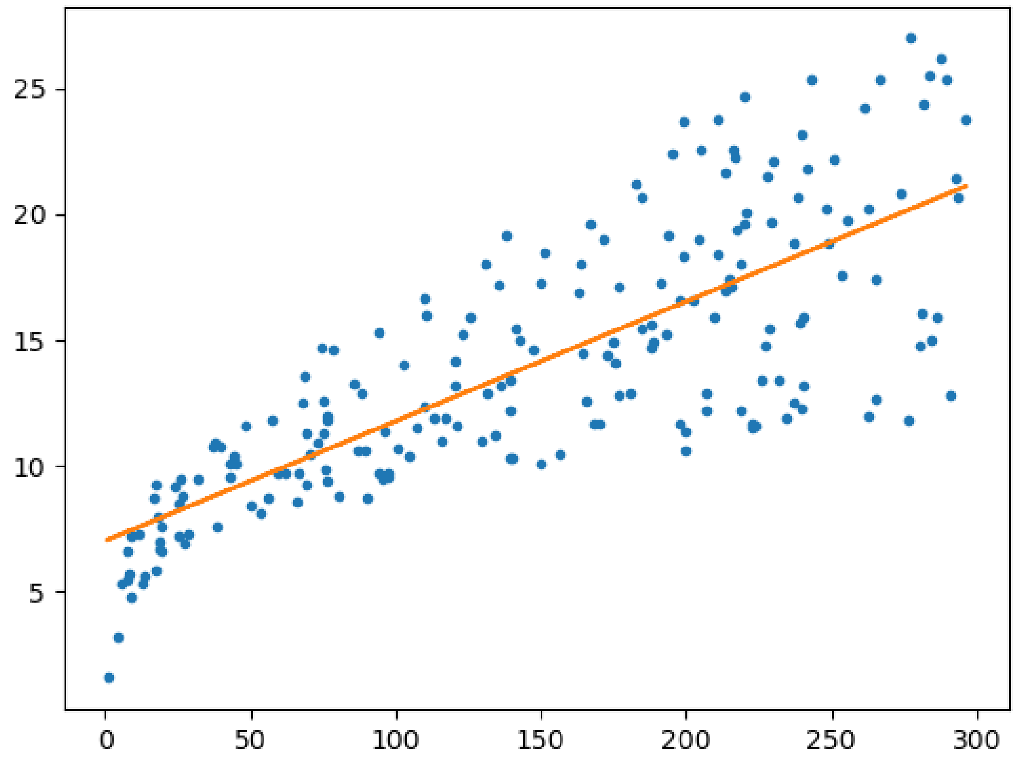
Plot

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('data/Advertising.csv', sep=',')
tv = df.TV.values
sales = df.sales.values
plt.plot(tv, sales, '.')
plt.show()
```



Linear Regression

```python
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression

df = pd.read_csv('data/Advertising.csv', sep=',')
tv = df.TV.values
sales = df.sales.values
lr = LinearRegression()
lr.fit(tv.reshape(-1, 1), sales)
print('beta_0 =', lr.intercept_)
print('beta_1 =', lr.coef_[0])
plt.plot(tv, sales, '.')
plt.plot(tv, lr.predict(tv.reshape(-1,1)))
plt.show()
'''
beta_0 = 7.032593549127695
beta_1 = 0.04753664043301975
'''
```



3 Linear Regression

- 3.1 Simple Linear Regression
- 3.2 Multiple Linear Regression
- 3.3 Other Considerations in the Regression Model
- 3.4 The Marketing Plan
- 3.5 Comparison of Linear Regression with K-Nearest Neighbors
- Lab: Linear Regression
- Exercises

Regression

- Regression
  - Observations
$$Y = f(X) + \epsilon$$
  - Find estimation of f
$$\hat{Y} = \hat{f}(X)$$
- Linear regression
  - Assume that f is linear, i.e.
$$f(x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
  - Estimation of f
$$\hat{f}(x_1, \dots, x_p) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$
  - Find $\hat{\beta}_0, \dots, \hat{\beta}_p$ from a training set

3.1 Simple Linear Regression

- Simple linear regression
  - Single predictor variable
  - Predictor X, response Y

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

  β_0, β_1: coefficients or parameters

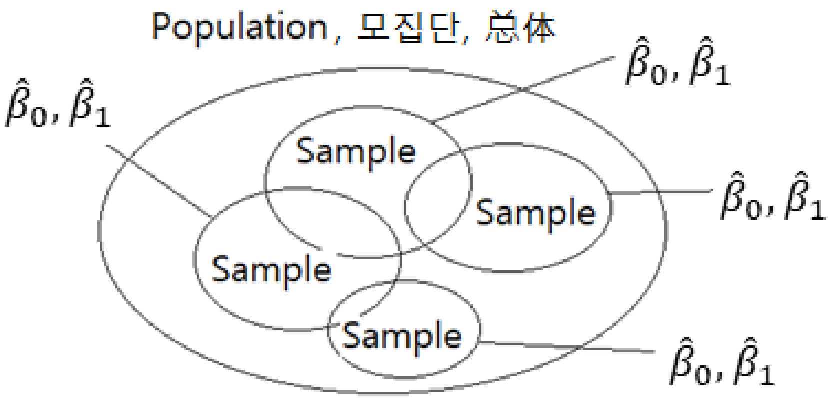
  - Especially, β_0: intercept, β_1: slope
- Example

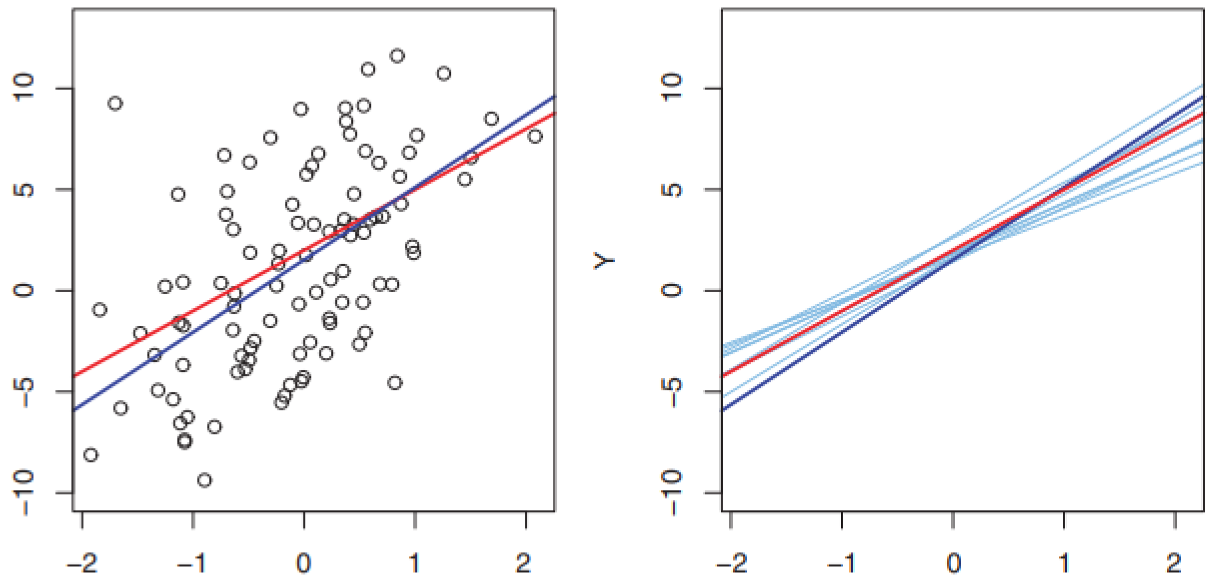
$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

- Estimate from training data

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

A simulated data





3.1.1 Estimating the Coefficients

- Observations

$$(x_1, y_1), \dots, (x_n, y_n)$$

- Prediction

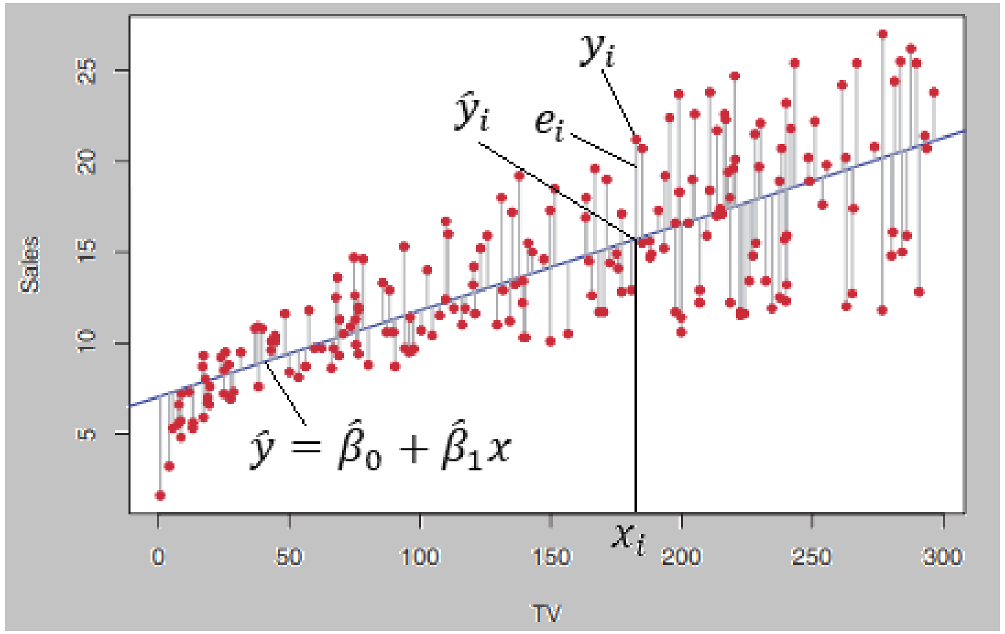
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- i-th residual

$$e_i = y_i - \hat{y}_i$$

- Residual sum of squares(RSS)

$$\begin{aligned} \mathrm{RSS} &= e_1^2 + \cdots + e_n^2 \\ &= \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_n x_n\right)^2 \end{aligned}$$



Least Square Approach

- To minimize the RSS

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

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

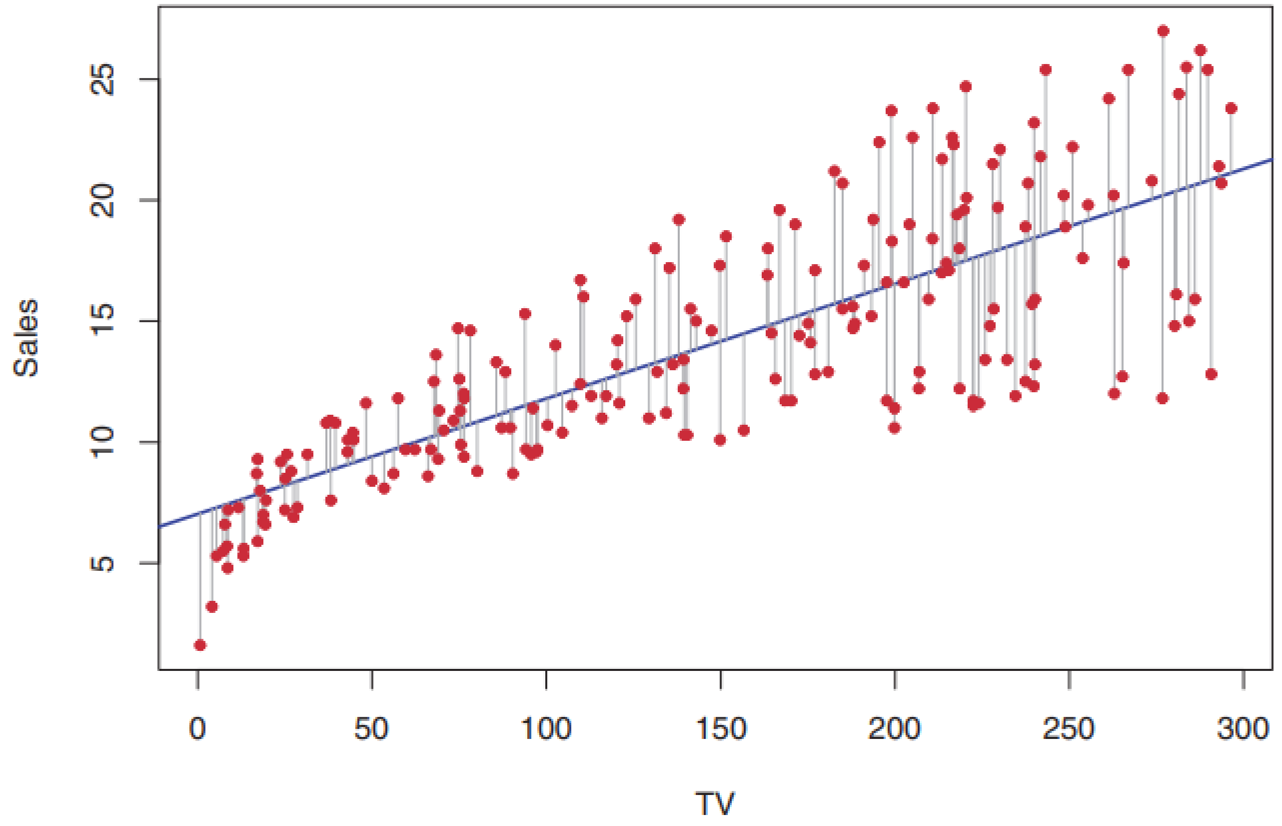
Computation

- To find x and y which minimize
$$f(x,y) = ax^2 + by^2 + cx + dy + e$$
  - Sol 1) quadratic function
$$f(x,y) = a(x - x_0)^2 + b(y - y_0)^2 + m$$
  - Sol 2) solve the system of linear equations
$$\frac{\partial f}{\partial x} = 0, \frac{\partial f}{\partial y} = 0$$
- Our Problem: minimize
$$\mathrm{RSS} = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_n x_n\right)^2$$

Advertising

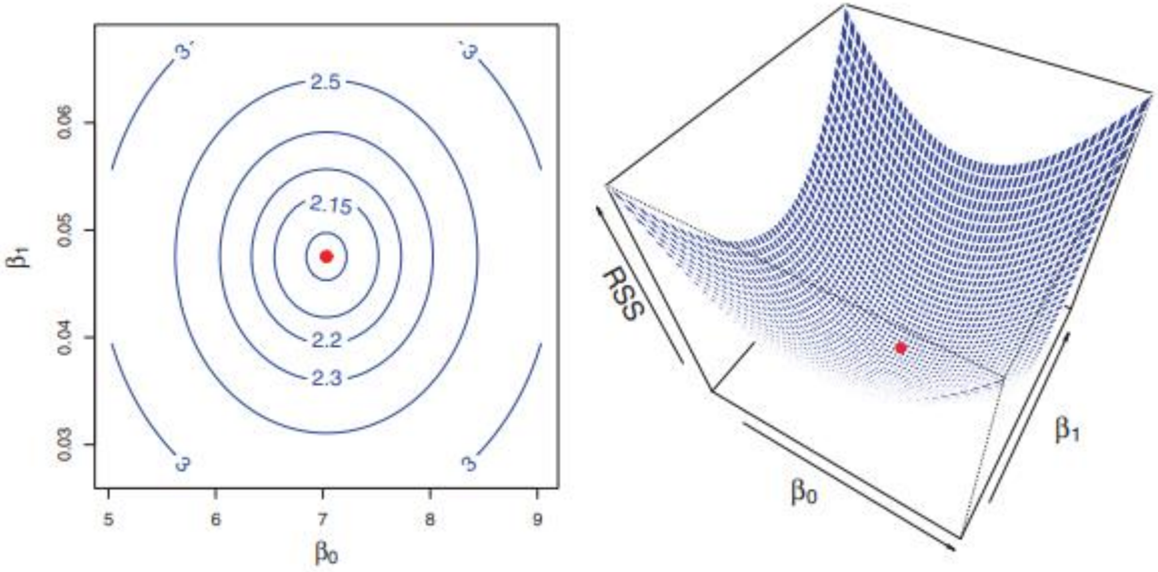
- TV and sales
  - $\hat{\beta}_0 = 7.0325$
  - $\hat{\beta}_1 = 0.0475$

		TV	adio	newspape	sales
1		TV	adio	newspape	sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8



RSS

- RSS is a quadratic function



3.1.2 Assessing the Accuracy of the Coefficient Estimates

- True relationship between X and Y

$$Y = f(X) + \epsilon$$

- Model i.e. assumption

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

  called population regression line
- Estimated regression line

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

Accuracy of the Coefficient Estimates

- Parameters

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2},\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Standard errors (= standard deviations)

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

where $\sigma^2 = \mathrm{Var}(\epsilon)$

  - When σ^2 is not known(it is true in general), estimate it as

$$\sigma = \mathrm{RSE} \text{ or } \sigma^2 = \frac{\mathrm{RSS}}{n-2}$$

- Prove that

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

  - see the page "[Proofs involving ordinary least squares]" in wikipedia

confidence interval

- 95% confidence interval for β_i

$$[\hat{\beta}_i - 2 \cdot \mathrm{SE}(\hat{\beta}_i),\ \hat{\beta}_i + 2 \cdot \mathrm{SE}(\hat{\beta}_i)]$$

Example: advertising data

- $\hat{\beta}_0 = 7.0325$, $\mathrm{SE}\left(\hat{\beta}_0\right) = 0.4578$,
  $\beta_0 \in [6.130, 7.935]$ in 95% confidence
- $\hat{\beta}_1 = 0.0475$, $\mathrm{SE}\left(\hat{\beta}_1\right) = 0.0027$,
  $\beta_1 \in [0.042, 0.053]$ in 95% confidence
- for each $1,000 increase in television advertising, there will be an average increase in sales of between 42 and 53 units.

Hypothesis Test

- Null hypothesis

  H_0: There is no relationship between X and Y

  i.e. $H_0 : \beta_1 = 0$

- Alternative hypothesis

  H_a: There is some relationship between X and Y

  i.e. $H_a : \beta_1 \neq 0$

- Interpretation
  - If

$$0 \notin \left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right), \hat{\beta}_1 + 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right)\right]$$

then, in 95% confidence,

$$\beta_1 \neq 0$$

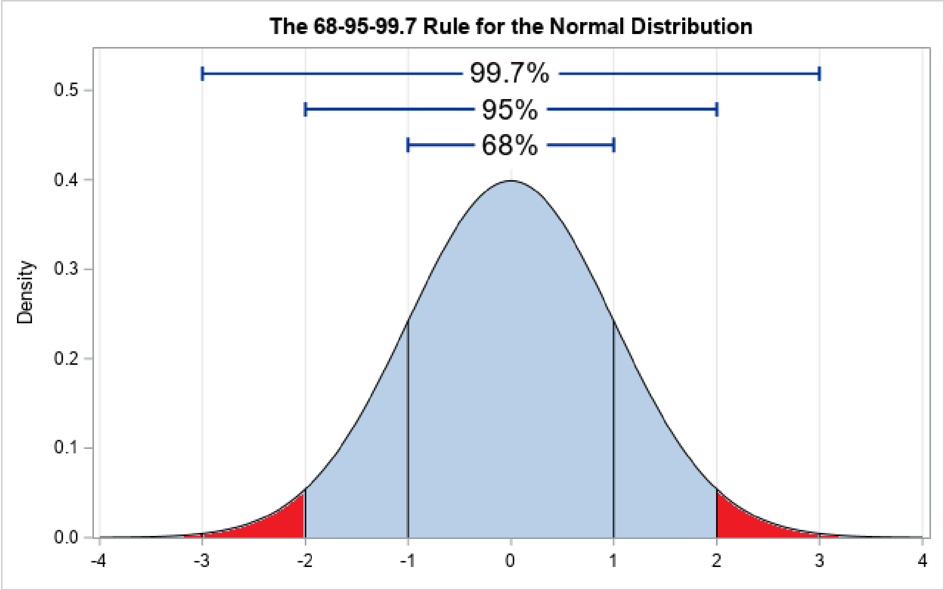
  - Or equivalently, if

$$0 \in \left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right), \hat{\beta}_i + 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right)\right]$$

then we can not say

$$\beta_1 \neq 0$$

in 95% confidence



The equation

$$0 \in \left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right)\right]$$

is equivalent to

$$\hat{\beta}_1 \in \left[0 - 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right),\ \ 0 + 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right)\right]$$

Let

$$\hat{\beta}_1 = 0 + t \cdot \mathrm{SE}(\hat{\beta}_1)$$

Then

$$t = \frac{\widehat{\beta} - 0}{\mathrm{SE}\left(\widehat{\beta}_1\right)}$$

The variable t follows the standard normal distribution, in general. $|t|$ determines whether H_0 is true or not

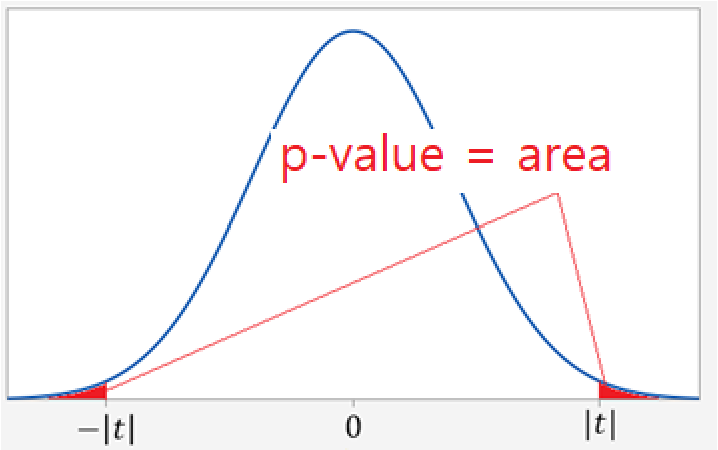
t-statistic

- t-statistic

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

  - If $|t| \geq 2$ then H_a is acceptable in 95% confidence
  - Error rate $< 5\%$

- p-value
  - p-value = error rate
    $=\Pr(|z| < |t|)$
    where $|z|$ is the standard normal distribution



Advertising

	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

3.1.3 Assessing the Accuracy of the Model

- The extent to which the model fits the data
- Methods:
  - Residual Standard Error
  - R^2 statistic

Residual Standard Error

- Residual Standard Error(RSE)

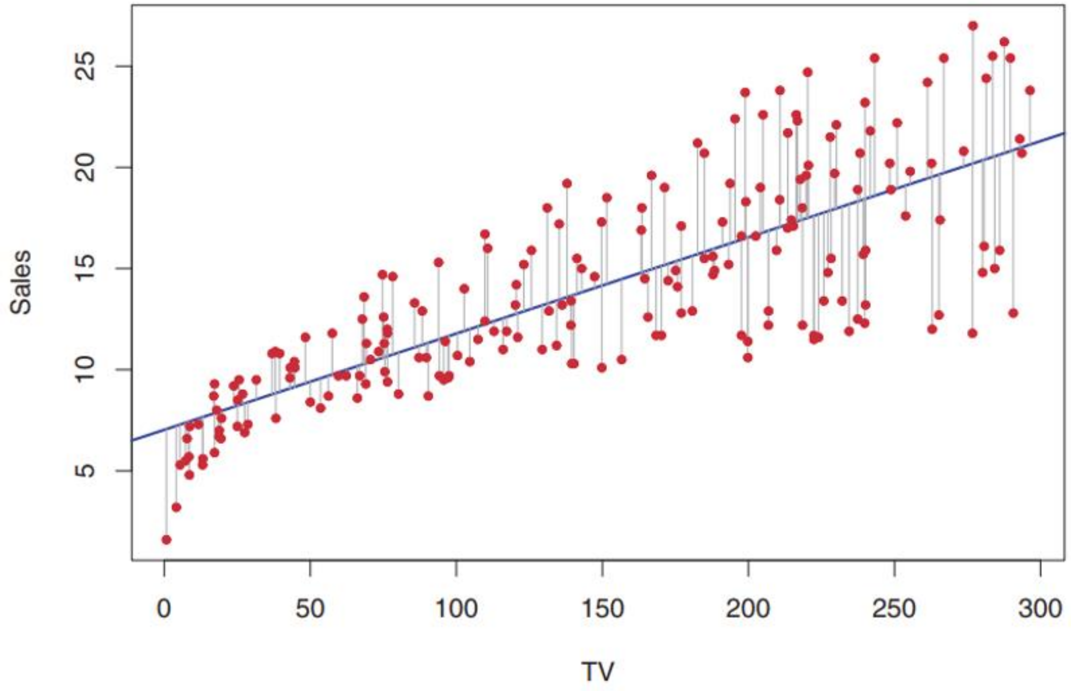
$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2}$$

  - Recall that

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

Advertising

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



- Actual sales in each market deviate from the true regression line by approximately 3,260 units, on average

R^2 Statistic

- R^2 statistic

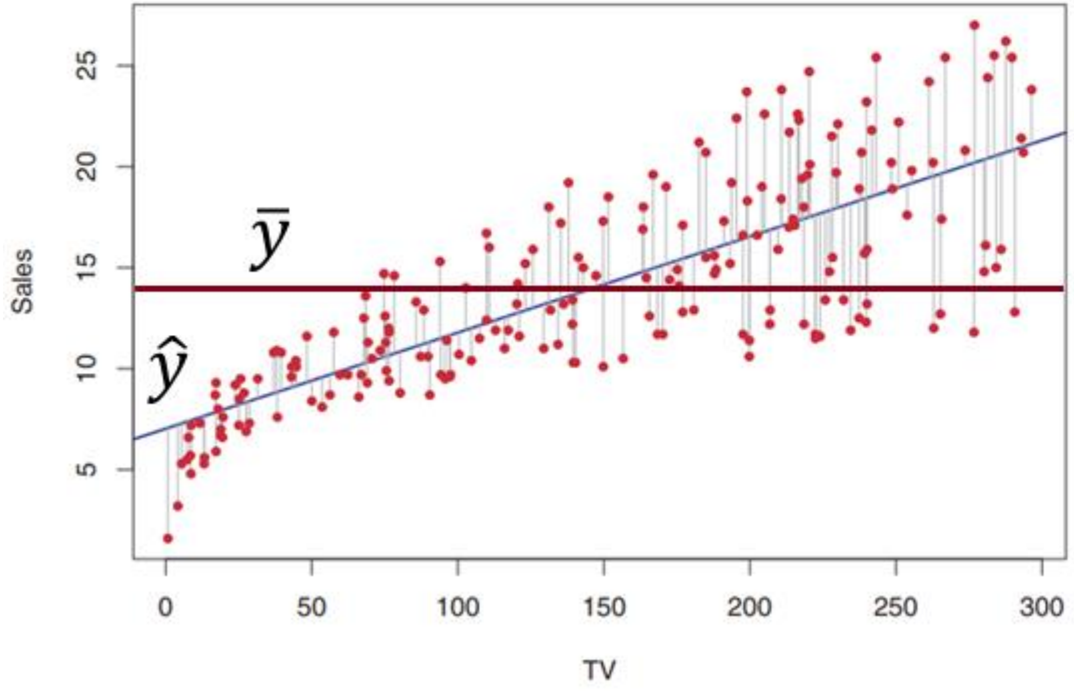
$$R^2 = \frac{\mathrm{TSS}-\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

  - TSS, the total sum of squares

$$\mathrm{TSS} = \sum (y_i - \bar{y})^2$$

  - RSS, the residual sum of squares

$$\mathrm{RSS} = \sum (y_i - \hat{y}_i)^2$$



- R^2 measures the proportion of variability in Y that can be explained using X
  - R^2 is close to 1
    - The linear model is good enough
  - R^2 is close to 0
    - The linear model is wrong, or
    - The inherent error σ^2 is high, or
    - Both
- Advertising

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

Correlation

- Correlation measures the linear relationship between X and Y

$$\mathrm{Cor}(X, Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

  - Cosine similarity
  - $R^2 = \mathrm{Cor}(X, Y)^2$

summary

- $\sigma^2 = \mathrm{Var}(\epsilon)$
- TSS – total sum of square
- RSS – residual sum of square
- RSE – residual standard error
- $\mathrm{SE}(\hat{\beta}_0)$, $\mathrm{SE}(\hat{\beta}_1)$ - standard error
- R^2
- $\mathrm{Cor}(X, Y)$ - correlation

3.2 Multiple Linear Regression

- Is it sufficient?

Simple regression of sales on radio

	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

Simple regression of sales on newspaper

	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

- Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Example: advertising

$$\mathrm{sales} = \beta_0 + \beta_1 \times \mathrm{TV} + \beta_2 \times \mathrm{radio} + \beta_3 \times \mathrm{newspaper} + \epsilon$$

Estimate coefficients

- Find $\hat{\beta}_0, \ldots, \hat{\beta}_p$ to minimize

$$\begin{aligned} RSS &= \sum (y_i - \hat{y}_i)^2 \\ &= \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}\right)^2 \end{aligned}$$

matrix form

- $X = \begin{bmatrix} 230.1 & 37.8 & 69.2 \\ 44.5 & 39.3 & 45.1 \\ \vdots & \vdots & \vdots \end{bmatrix}$
- $y = \begin{bmatrix} 22.1 \\ 10.4 \\ \vdots \end{bmatrix}$
- $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$
- $y = \beta_0 + X\beta + \epsilon$

1		TV	radio	newspape	sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8

matrix form

- $X = \begin{bmatrix} 1 & 230.1 & 37.8 & 69.2 \\ 1 & 44.5 & 39.3 & 45.1 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$
- $y = \begin{bmatrix} 22.1 \\ 10.4 \\ \vdots \end{bmatrix}$
- $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$
- $y = X\beta + \epsilon$
- $\mathrm{RSS} = \| y - X\hat{\beta} \|^2$

1		TV	radio	newspape	sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8

Computation

- Sol 1) solve the system of linear equations

$$\frac{\partial \mathrm{RSS}}{\partial \hat{\beta}_i} = 0$$

- Sol 2) For a training data set (X, y)

$$\begin{aligned} \mathrm{RSS} &= \| X\hat{\beta} - y \|^2 = \left(X\hat{\beta} - y\right)^T \left(X\hat{\beta} - y\right) \\ &= \hat{\beta}^T X^T X \hat{\beta} - 2\hat{\beta}^T X^T y + y^T y \end{aligned}$$

$$0 = \frac{\partial \mathrm{RSS}}{\partial \hat{\beta}} = 2X^T X\hat{\beta} - 2X^T y$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

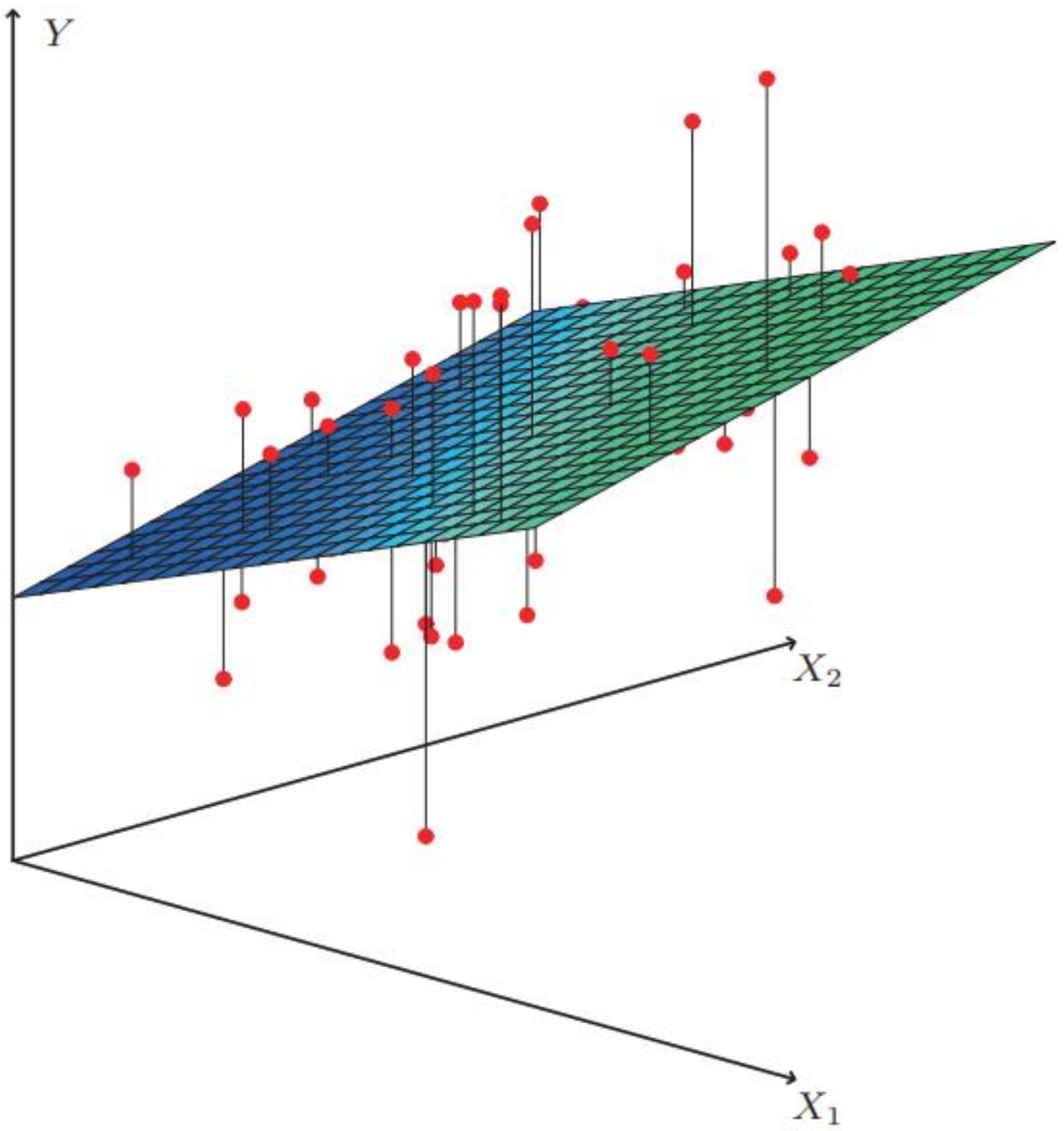
- Sol 3) equation

$$X\hat{\beta} = y$$

  - multiply the Moore-Penrose pseudo inverse $(X^T X)^{-1} X^T$ in both sides

$$(X^T X)^{-1} X^T X \hat{\beta} = (X^T X)^{-1} X^T y$$
$$\hat{\beta} = (X^T X)^{-1} X^T y$$

  - Is $X^T X$ invertiable?


Y
X₂
X₁

Example

- Advertising

	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	−0.001

  - There is no relationship between sales and newspaper

- In simple regression, there is relation between sales and newspaper. Why?

Simple regression of sales on newspaper

	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

TABLE 3.3. *More simple linear regression models for the* Advertising *data. Co-*

- Correlation between radio and newspaper
  - The correlation reveals a tendency to spend more on newspaper advertising in markets where more is spent on radio advertising

	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

TABLE 3.5. *Correlation matrix for* TV, radio, newspaper, *and* sales *for the* Advertising *data.*

  - shark attacks versus ice cream sales for data collected at a given beach community

3.2.2 Some Important Questions

1. Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

One: Is There a Relationship Between the Response and Predictors?

- Null hypothesis

$$H_0: \beta_1 = \cdots = \beta_p = 0$$

- Alternative hypothesis

H_a: at least one β_j is non-zero

F-statistic

- F-statistic

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

  where $\mathrm{TSS} = \sum (y_i - \bar{y})^2$, $\mathrm{RSS} = \sum (y_i - \hat{y}_i)^2$
  - $\mathbb{E}\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2$ if the population model is linear
  - $\mathbb{E}\left[\frac{\mathrm{TSS}-\mathrm{RSS}}{p}\right] \geq \sigma^2$

    In particular, $\mathbb{E}\left[\frac{\mathrm{TSS}-\mathrm{RSS}}{p}\right] = \sigma^2$ if H_0 is true

- $F \geq 1$

- The F-statistic takes on a value close to 1
  - When there is no relationship between the response and predictors
  - Or equivalently, H_0 is true

Advertising

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

Two: Deciding on Important Variables

- Forward selection
  - Start from null model – no variables
  - Add variables which has the lowest RSS, one by one
- Backward selection
  - Start from full model
  - remove variables which has the highest p-value, one by one
- Mixed selection
  - Forward selection
  - Remove variable of which p-value exceeds the threshold

Three: Model Fit

- Numerical measures of model fit: R^2 and RSE
- R^2
  - close to 1 indicates that the model explains a large portion of Y
  - coincides $\mathrm{cor}(Y, \hat{Y})$
- RSE

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-p-1}}$$

Four: Predictions

- Incorrectness of prediction
  - Inaccuracy in the coefficient of estimation
  - Model bias
    - Population is not linear
  - Irreducible error due to ϵ

3.3 Other Considerations in the Regression Model

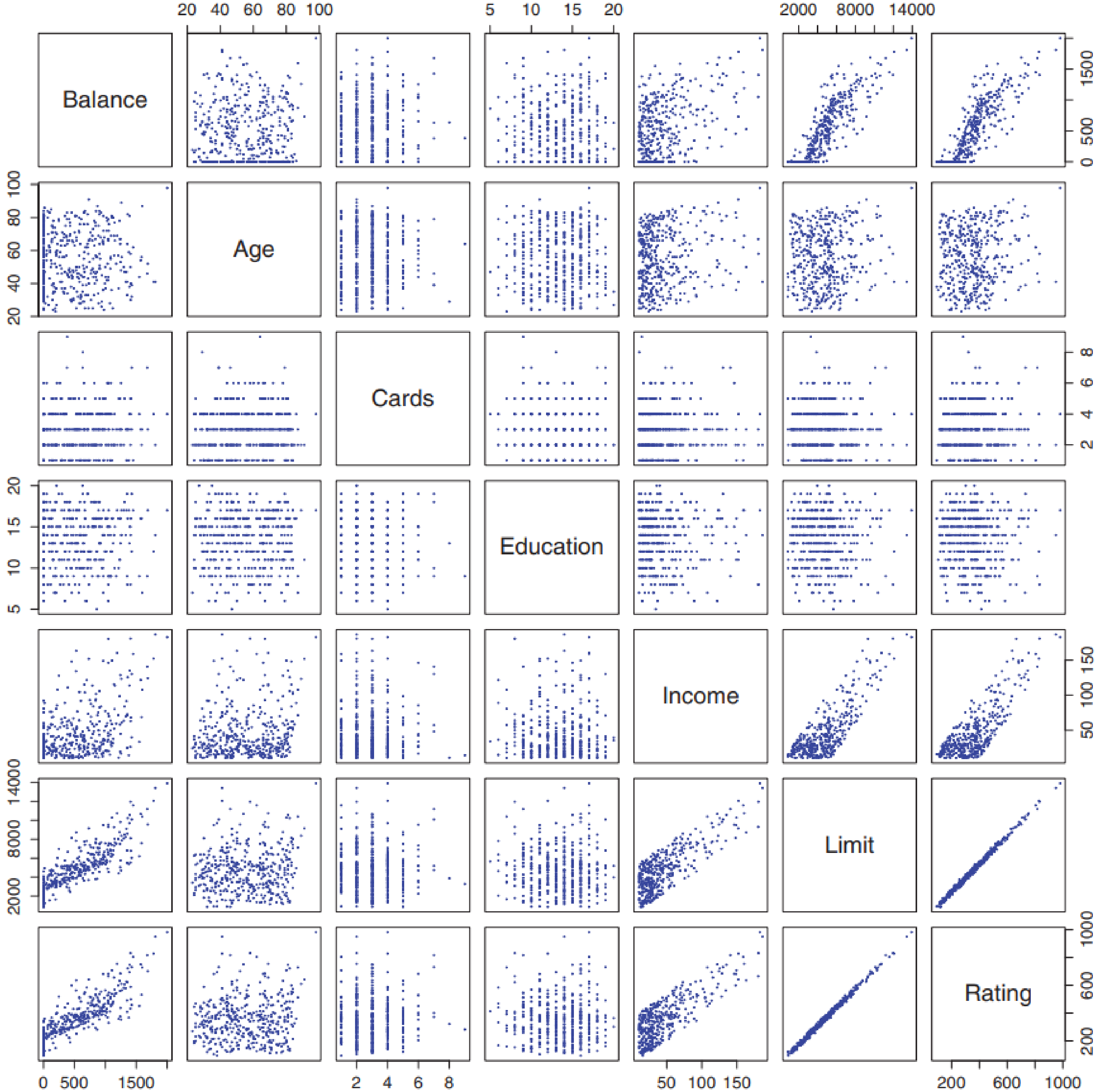
- 3.3.1 Qualitative Predictors
- 3.3.2 Extensions of the Linear Model
- 3.3.3 Potential Problems

3.3.1 Qualitative Predictors

- Variable types
  - Quantitative
  - Qualitative

Example: Credit

- Response
  - Balance
    - Credit card debt
- Predictors
  - age, cards(# of cards), education(years), income, limit(credit limit), rating(credit rating) - quantitative
  - gender, student(student status), status(marital status), ethnicity(Caucasian, African American, Asian) - qualitative



Balance
Age
Cards
Education
Income
Limit
Rating

Predictors with Only Two Levels

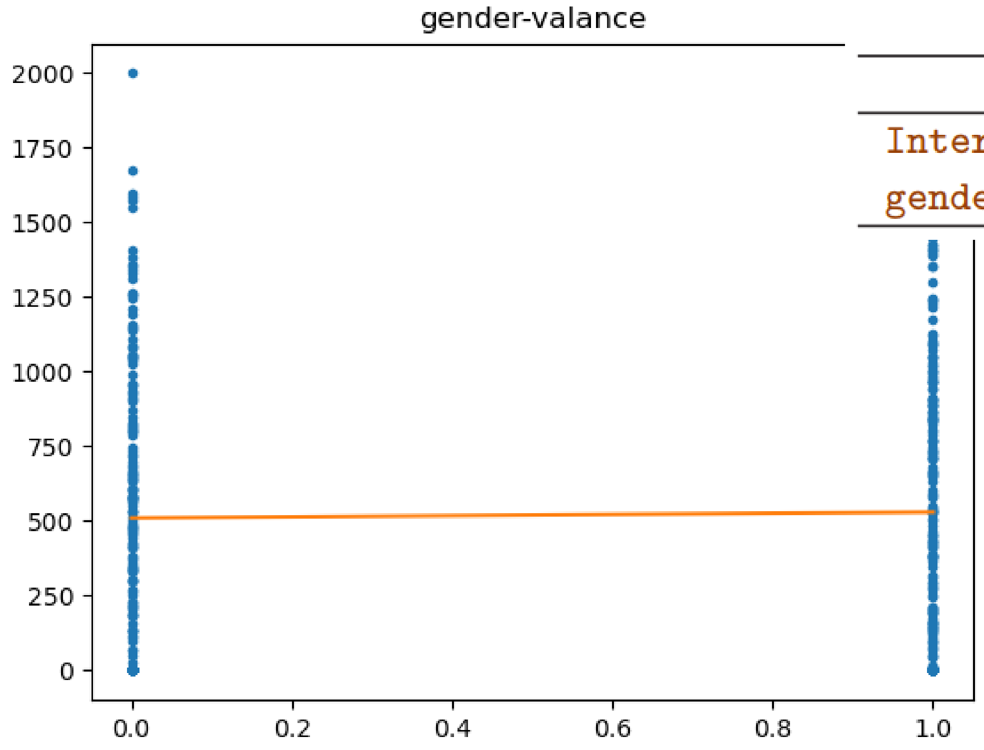
- Qualitative predictor which has only two levels

- Example: credit
  - Regression
  - Predictor – gender
  - Response - balance
  - $(x_1, y_1), \ldots, (x_n, y_n)$

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

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

Linear Regression of balance on gender



	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

- Another labels
  - $x_i = \begin{cases} 1, & \text{if female} \\ 0, & \text{if male} \end{cases}$
  - $x_i = \begin{cases} 1, & \text{if female} \\ -1, & \text{if male} \end{cases}$

Qualitative Predictors with More than Two Levels

- Qualitative predictor which has levels > 2
- Example : credit
  - ethnicity – Caucasian, African American, Asian
  - regression
    - data – ethnicity, balance
    - $(x_{11}, x_{12}, y_1), \ldots, (x_{n1}, x_{n2}, y_n)$

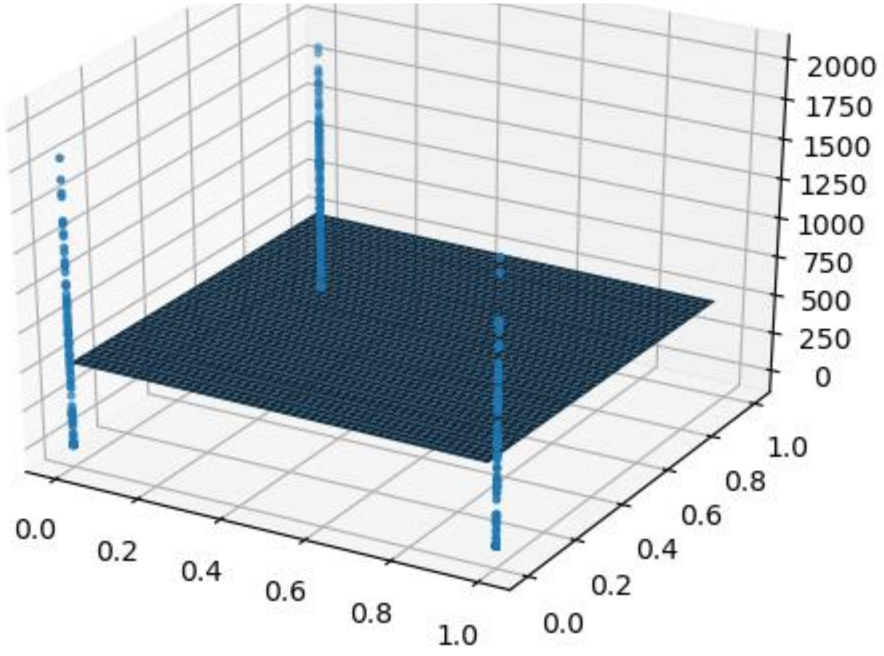
$$x_{i1} = \begin{cases} 1, & \text{if Asian} \\ 0, & \text{if not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & \text{if Caucasian} \\ 0, & \text{if not Caucasian} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Linear regression of balance on ethnicity

	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



3.3.2 Extensions of the Linear Model

- Linear Model
  - Predictors and response are additive and linear

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

Removing the Additive Assumption

- Interaction effect
  - $X_1 X_2$
- Example: advertising
  - $X_1 =$ TV, $X_2 =$ radio
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$

	TV	radio	TV*radio	sales
1	230.1	37.8	8697.78	22.1
2	44.5	39.3	1748.85	10.4
3	17.2	45.9	789.48	9.3
4	151.5	41.3	6256.95	18.5
5	180.8	10.8	1952.64	12.9
6	8.7	48.9	425.43	7.2
7	57.5	32.8	1886	11.8
8	120.2	19.6	2355.92	13.2
9	8.6	2.1	18.06	4.8
10	199.8	2.6	519.48	10.6
11	66.1	5.8	383.38	8.6

Result

	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

TABLE 3.9. *For the* Advertising *data, least squares coefficient estimates associated with the regression of* sales *onto* TV *and* radio*, with an interaction term, as in (3.33).*

- R^2 with interaction 0.968
- R^2 without interaction 0.897

Non-linear Relationships

- polynomial regression

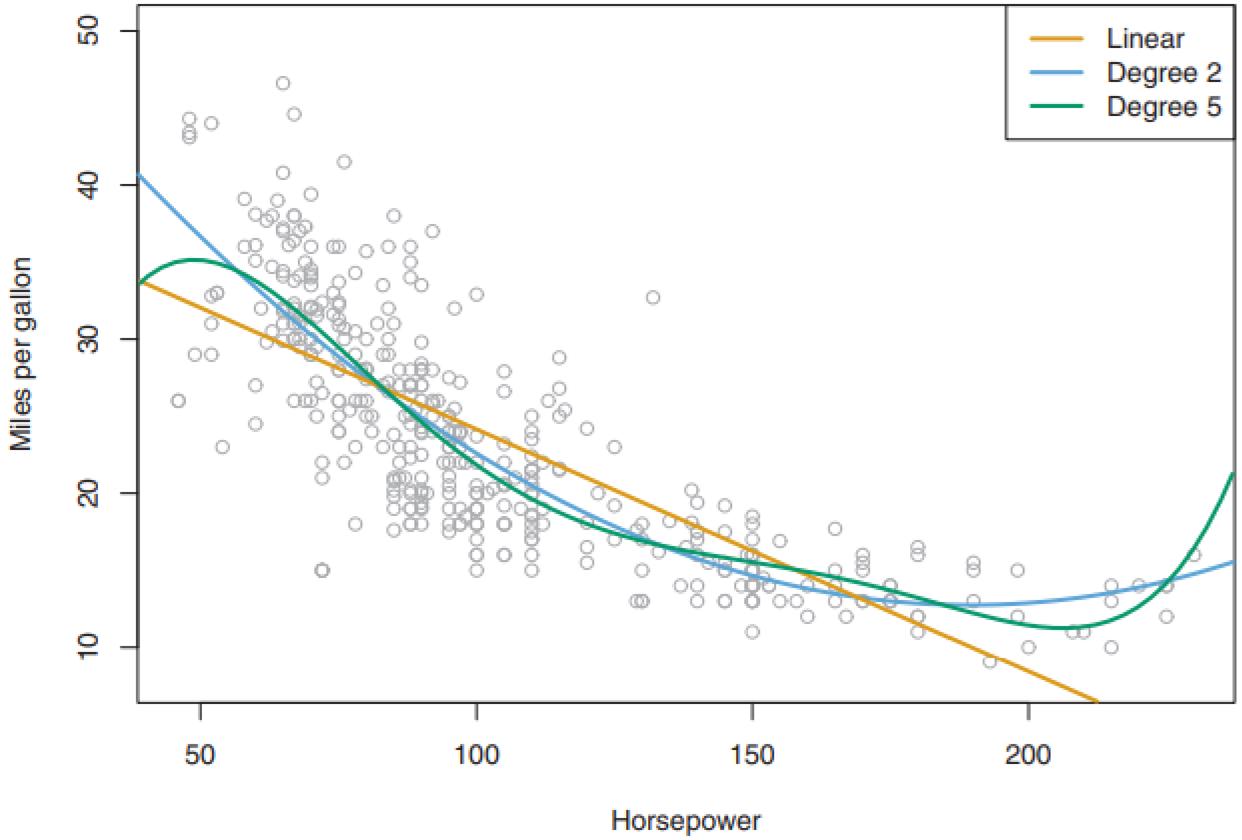
$$Y = \beta_0 + \beta_1 X + \cdots + \beta_p X^p + \epsilon$$

Auto

- Example - auto
  - $\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \epsilon$

mpg	cylinders	displacem	horsepowe	weight
18	8	307	130	350
15	8	350	165	369
18	8	318	150	343
16	8	304	150	343
17	8	302	140	344
15	8	429	198	434
14	8	454	220	435
14	8	440	215	431
14	8	455	225	442
15	8	390	190	385
15	8	383	170	356

mpg	horsepower	horsepower^2	w
18	130	16900	
15	165	27225	
18	150	22500	
16	150	22500	
17	140	19600	
15	198	39204	
14	220	48400	
14	215	46225	
14	225	50625	
15	190	36100	



	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

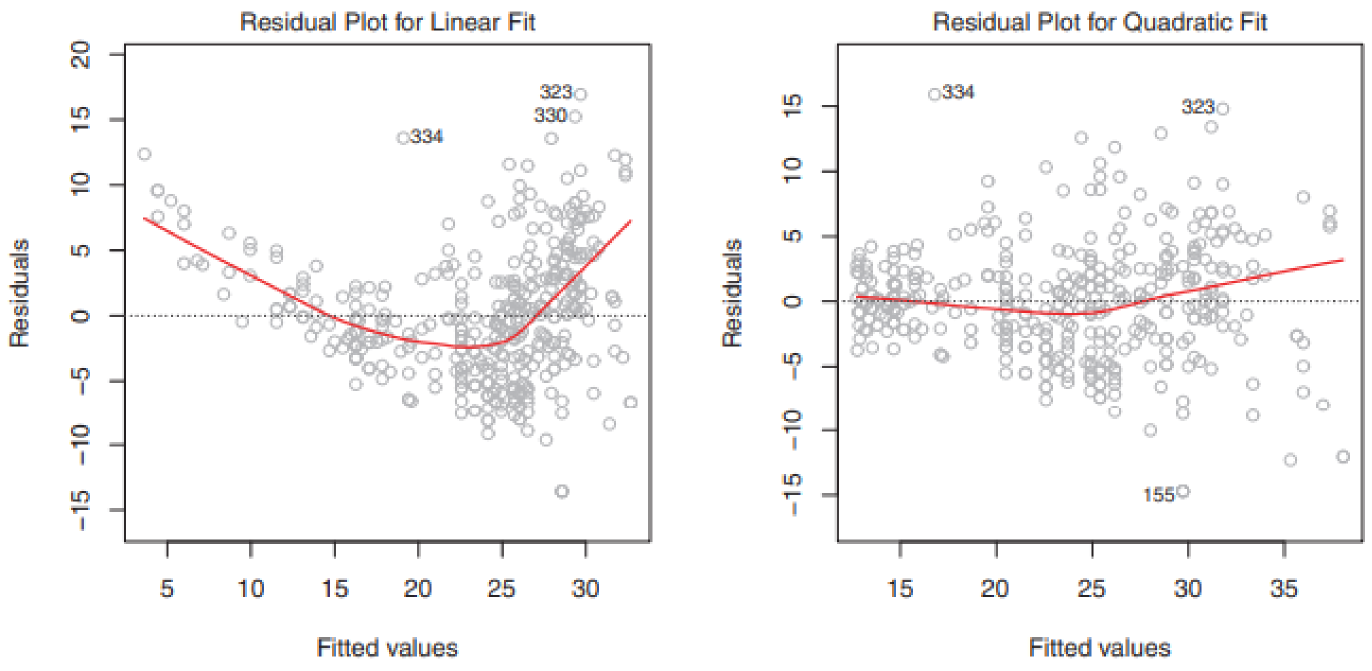
TABLE 3.10. *For the* `Auto` *data set, least squares coefficient estimates associated with the regression of* `mpg` *onto* `horsepower` *and* `horsepower`^{2}.

3.3.3 Potential Problems

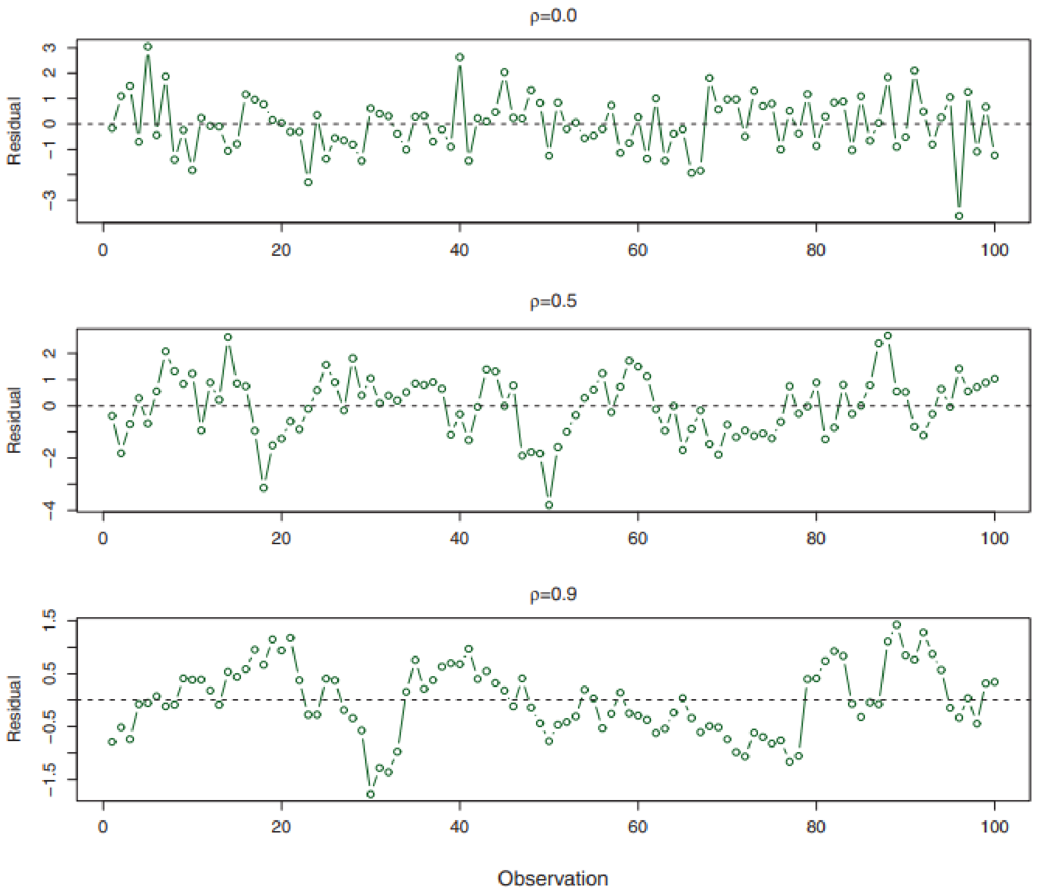
- 1. Non-linearity of the response-predictor relationships.
- 2. Correlation of error terms.
- 3. Non-constant variance of error terms.
- 4. Outliers.
- 5. High-leverage points.
- 6. Collinearity.

1. Non-linearity of the Data

- $\log x$, $\sqrt{x}$, x^2



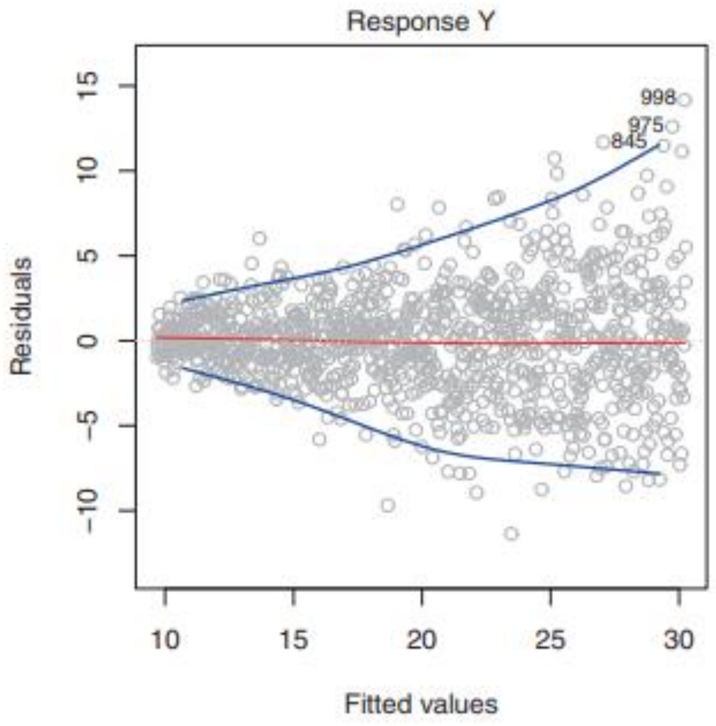
2. Correlation of Error Terms

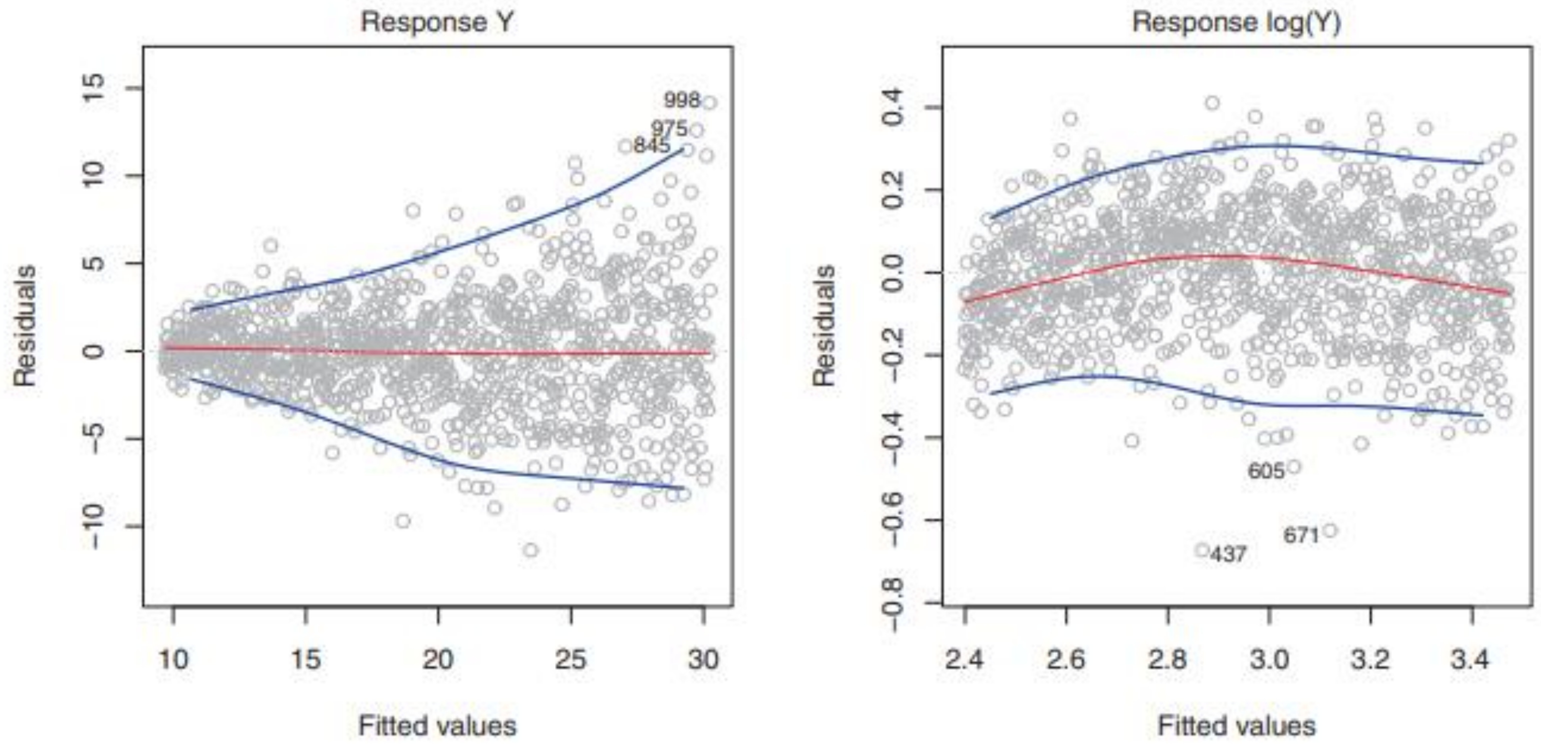


3. Non-constant Variance of Error Terms

- We assume
  - ϵ is independent of predictors
  - $E(\epsilon) = 0$
  - $\mathrm{Var}(\epsilon_i)$ =constant

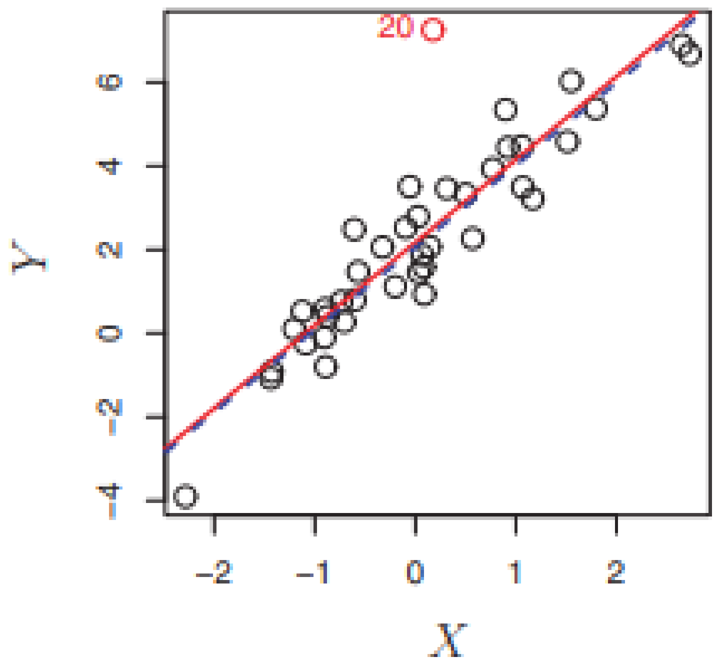
- But $\mathrm{Var}(\epsilon_i)$ may not be constant





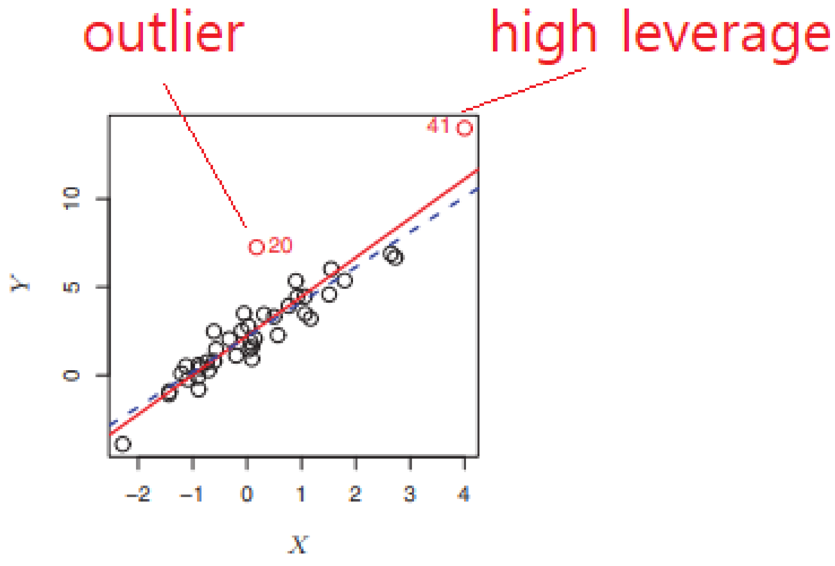
Response Y
Residuals
Fitted values
998
975
845
Response log(Y)
Residuals
Fitted values
605
671
437

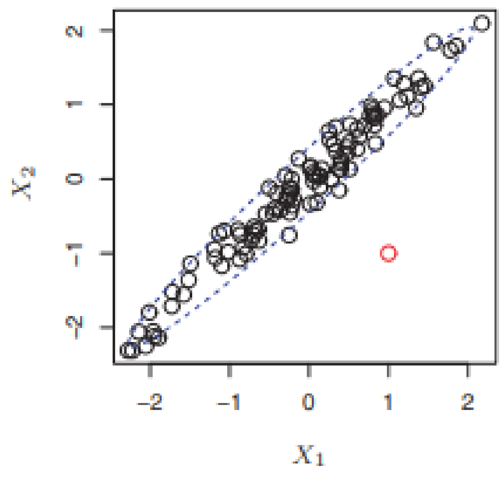
4. Outliers



5. High Leverage Points

- high leverage = an unusual value for x_i





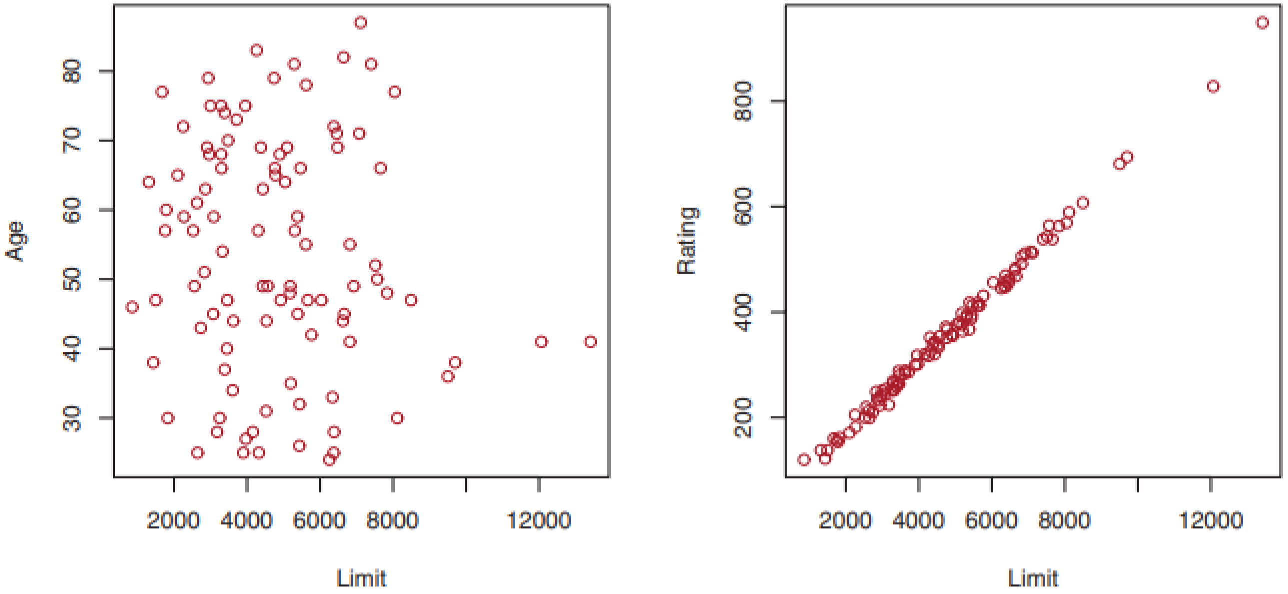
Leverage computation

- Simple linear regression
  - $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$
    - $\frac{1}{n} \leq h_i \leq 1$
    - average of $h_i = \frac{\sum h_i}{n} = \frac{p+1}{n}$
  - large h_i = high leverage

- General case
  - $h_i = i$-th diagonal entry of hat matrix $X(X^T X)^{-1} X^T$

6. Collinearity

- Credit



variance inflation factor (VIF)

- $\mathrm{VIF}\left(\hat{\beta}_j\right) = \frac{1}{1-R^2_{X_j|X_{-j}}}$
  - $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors
- If $R^2_{X_j|X_{-j}}$ is close to one, then collinearity is present, and so the VIF will be large.

3.4 The Marketing Plan

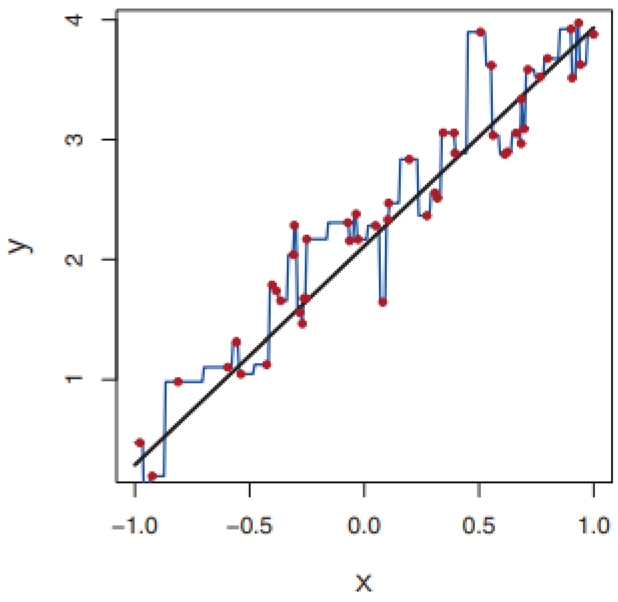
- Read the book

3.5 Comparison of Linear Regression with K-Nearest Neighbors

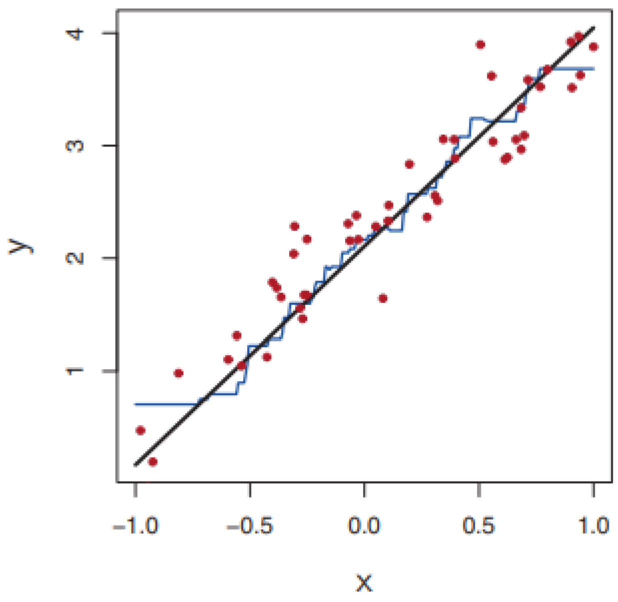
- Parametric regression
  - Linear regression, etc
- Non-parametric regression
  - KNN regression(K-nearest neighbors regression)

KNN

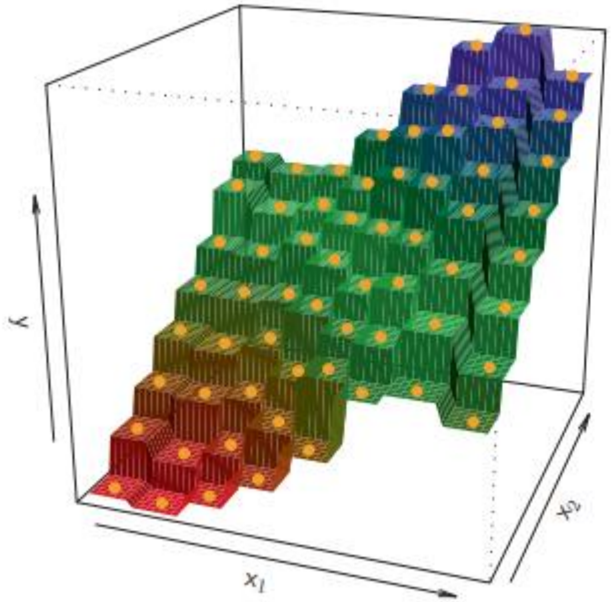
- $\hat{f}(x_0) = \frac{1}{K}\sum_{x_i \in N_0} y_i$
  - $N_0 = K$ training observations that are closest to x_0



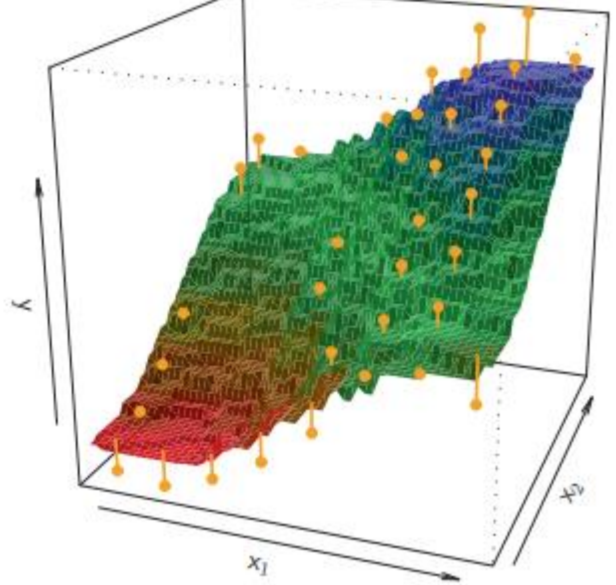
K=1



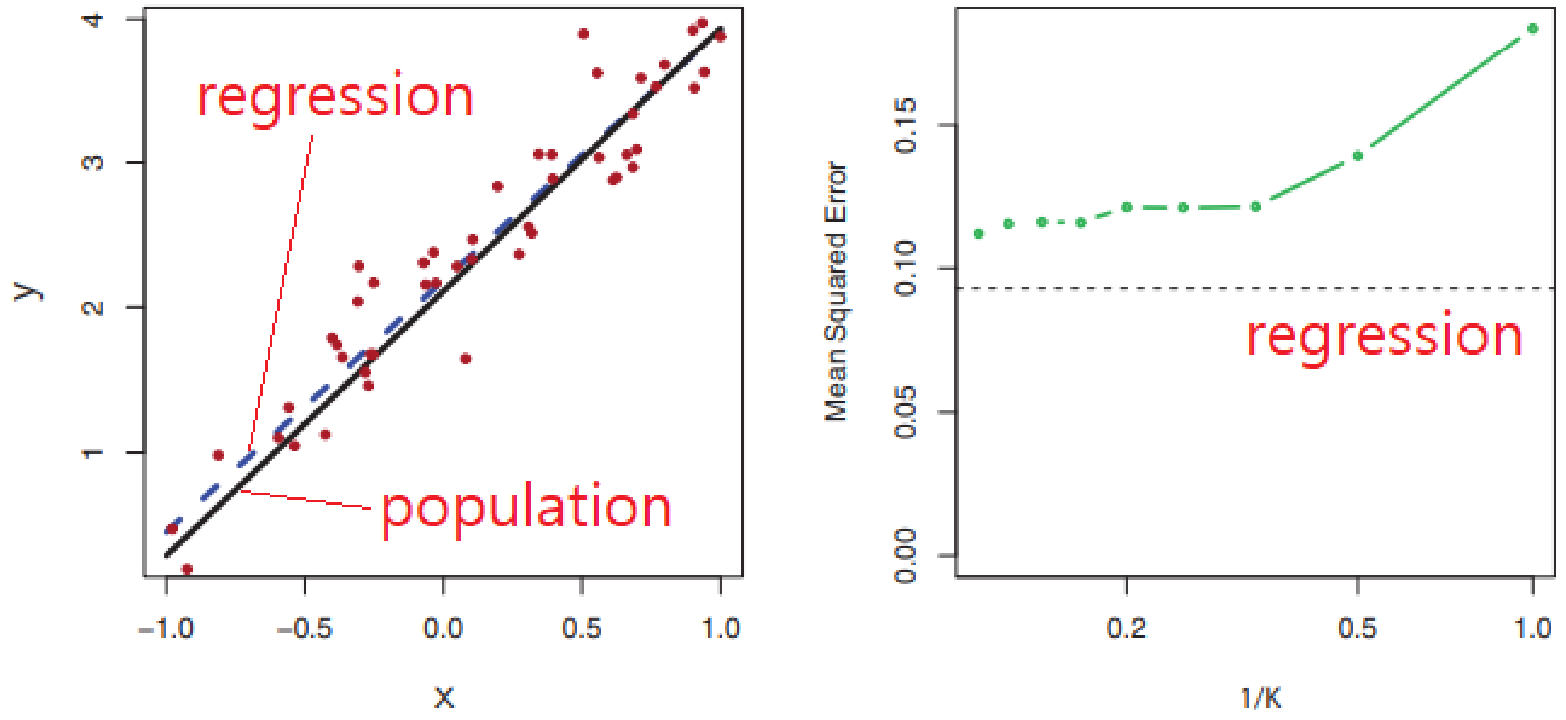
K=9



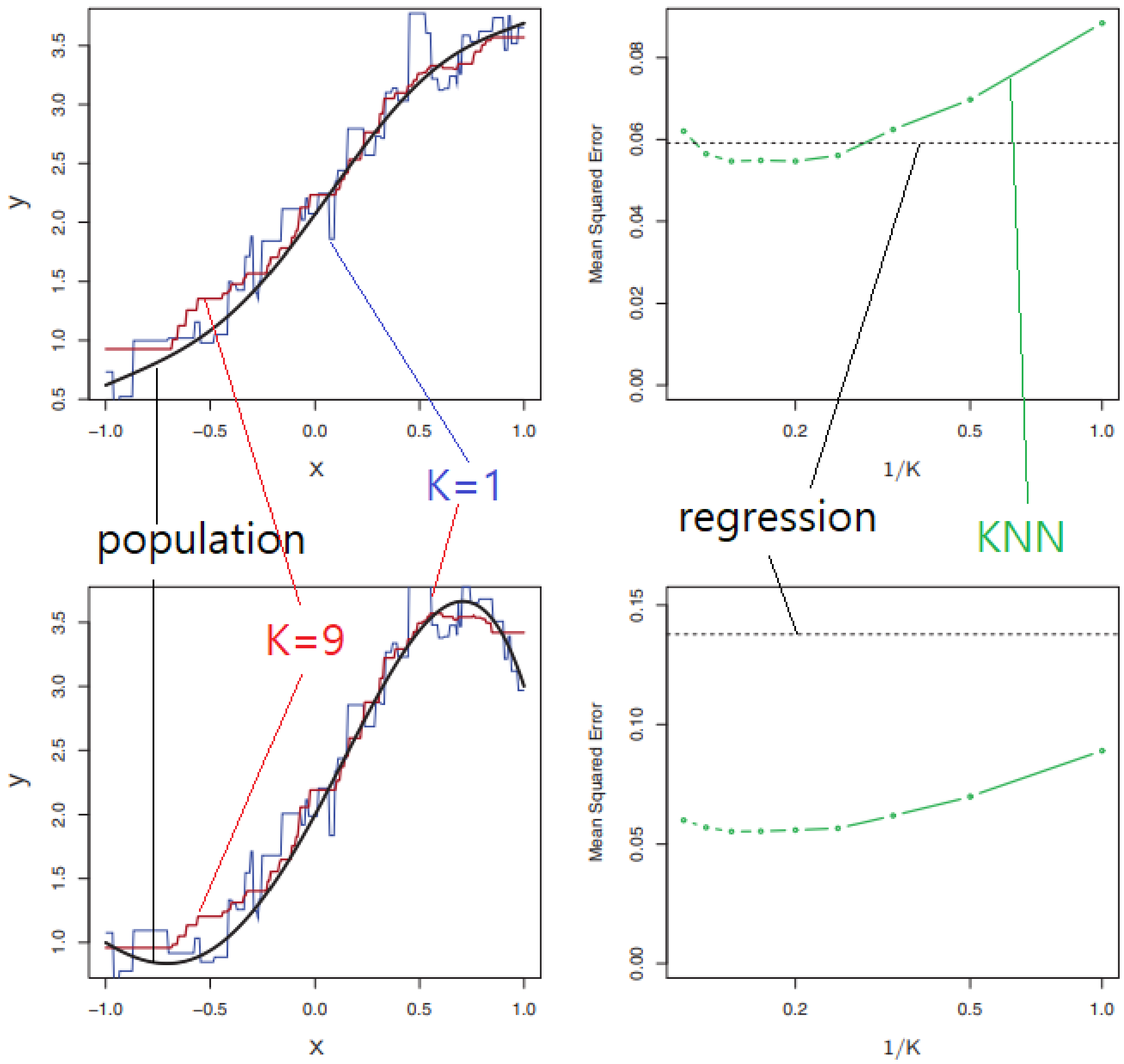
K=1



K=9



regression
population
y
x
−1.0
−0.5
0.0
0.5
1.0
1
2
3
4
Mean Squared Error
0.00
0.05
0.10
0.15
0.2
0.5
1.0
1/K
regression



y
x
Mean Squared Error
1/K
population
K=1
K=9
regression
KNN

curse of dimensionality