



Multivariate Linear Regression

강필성

고려대학교 산업경영공학부

pilsung_kang@korea.ac.kr

회귀 분석: Regression Analysis

- 도요타 코롤라 자동차의 중고차 가격 예측



종속 변수
(target)

설명 변수
(attributes, features)


Variable	Description
Price	Offer Price in EUROS
Age_08_04	Age in months as in August 2004
KM	Accumulated Kilometers on odometer
Fuel_Type	Fuel Type (Petrol, Diesel, CNG)
HP	Horse Power
Met_Color	Metallic Color? (Yes=1, No=0)
Automatic	Automatic (Yes=1, No=0)
CC	Cylinder Volume in cubic centimeters
Doors	Number of doors
Quarterly_Tax	Quarterly road tax in EUROS
Weight	Weight in Kilograms

다중 선형 회귀 모형: Multivariate Linear Regression

- 목적

- ✓ 종속변수 Y 와 설명변수 집합 X_1, X_2, \dots, X_p 사이의 관계를 **선형으로 가정**하고 이를 가장 잘 설명할 수 있는 회귀 계수 **regression coefficients**를 추정
- ✓ 예: 웨이퍼의 수율(Y)은 FDC 파라미터들(X)의 선형 결합으로 표현될 수 있음을 가정

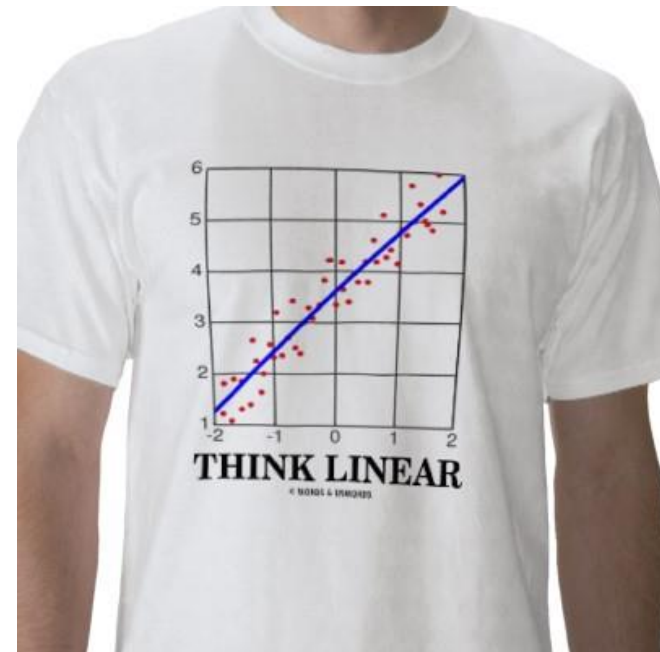
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_d x_d + \epsilon$$



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$

coefficients

unexplained

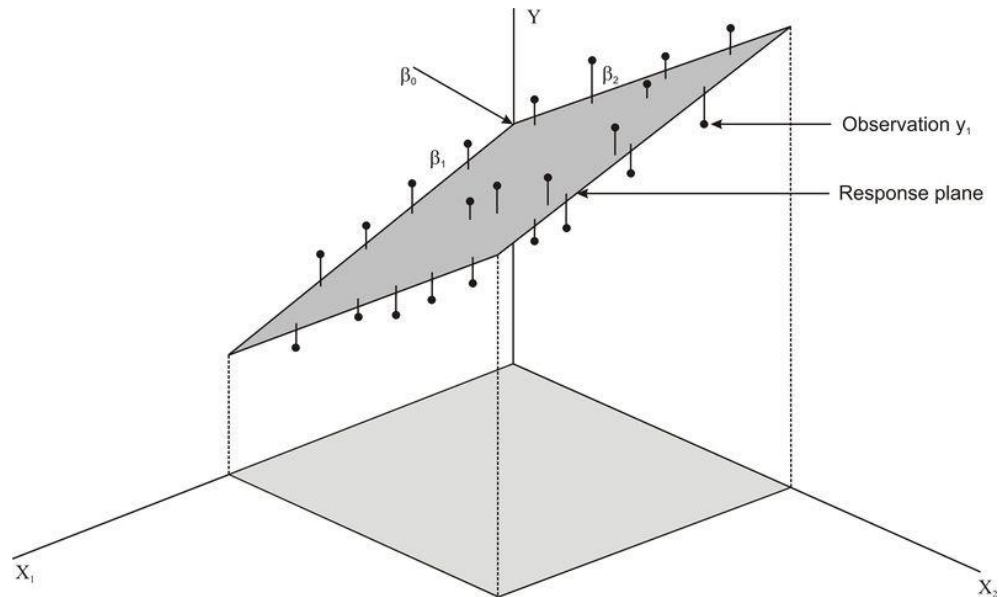
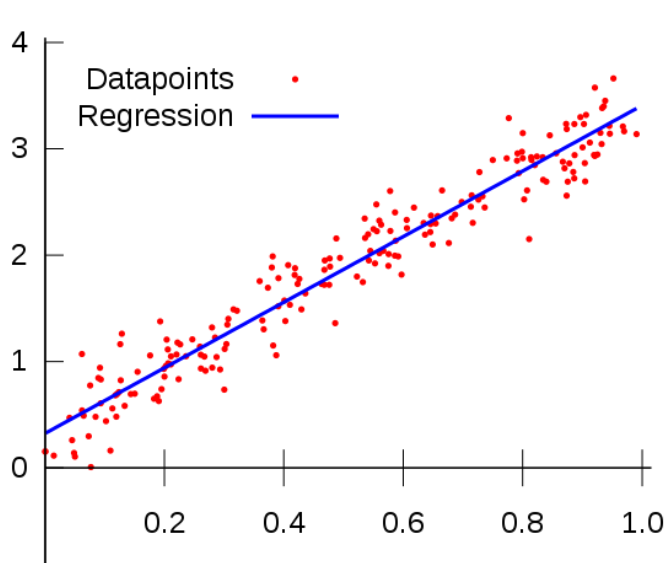


다중 선형 회귀 모형: Multivariate Linear Regression

- 다중 선형 회귀 모형: Multivariate Linear Regression

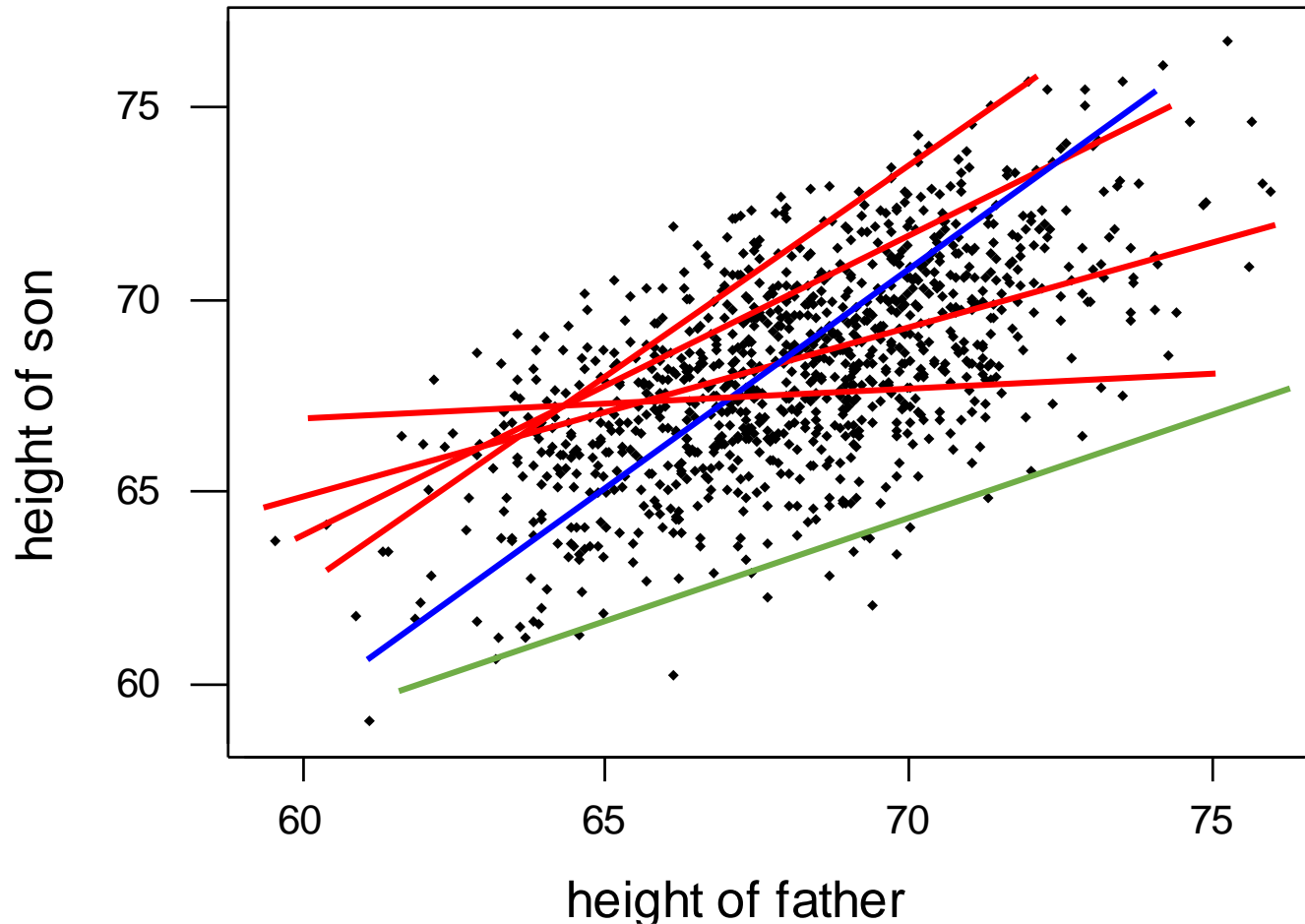
✓ 반응변수들과 설명변수 사이의 관계를 선형으로 표현

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$



다중 선형 회귀 모형: Multivariate Linear Regression

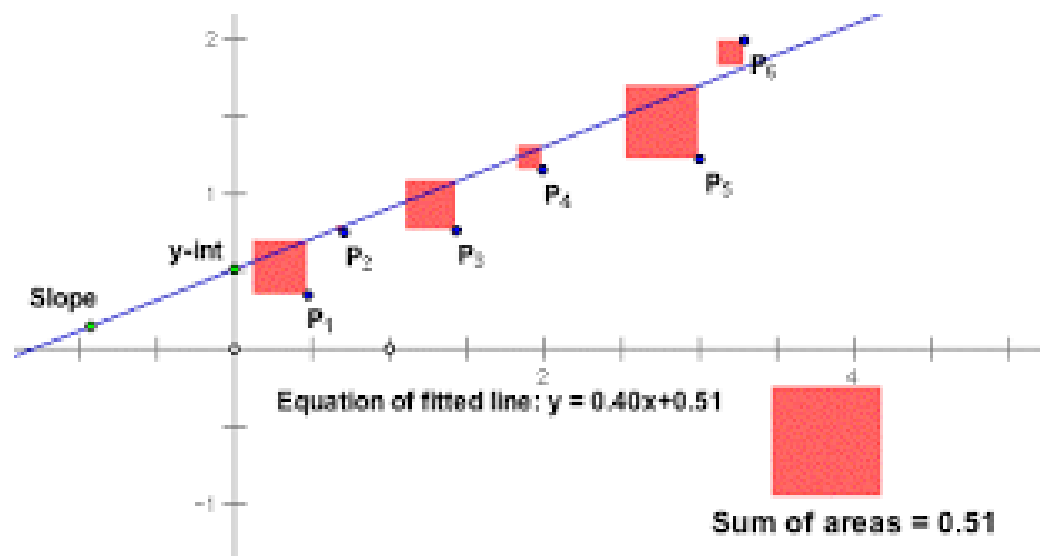
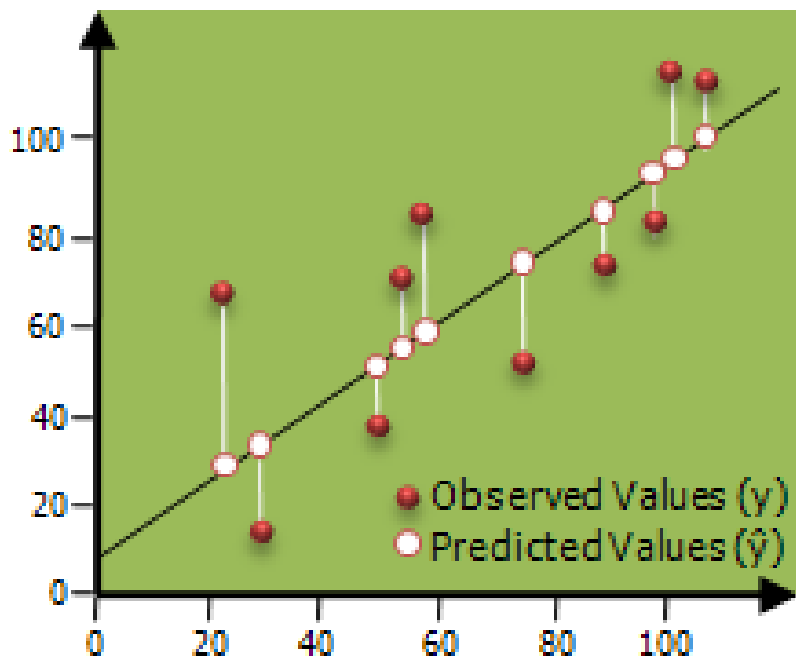
- 어떤 직선이 설명변수와 종속변수를 가장 잘 표현하는가?



다중 선형 회귀 모형: 회귀 계수의 추정

- 최소자승법: Ordinary Least Squares (OLS)

✓ 추정된 회귀식에 의해 결정된 값과 실제 종속변수 값의 차이를 최소한으로 줄이는 것을 목적으로 함



다중 선형 회귀 모형: 회귀 계수의 추정

- 회귀 계수의 추정

- ✓ 최소자승법: Ordinary least square (OLS)

- Actual target: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_d x_d + \epsilon$

- Predicted target: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$

- **목적:** 실제 종속변수 값과 예측된 종속변수 값 사이의 오차 제곱합을 최소화

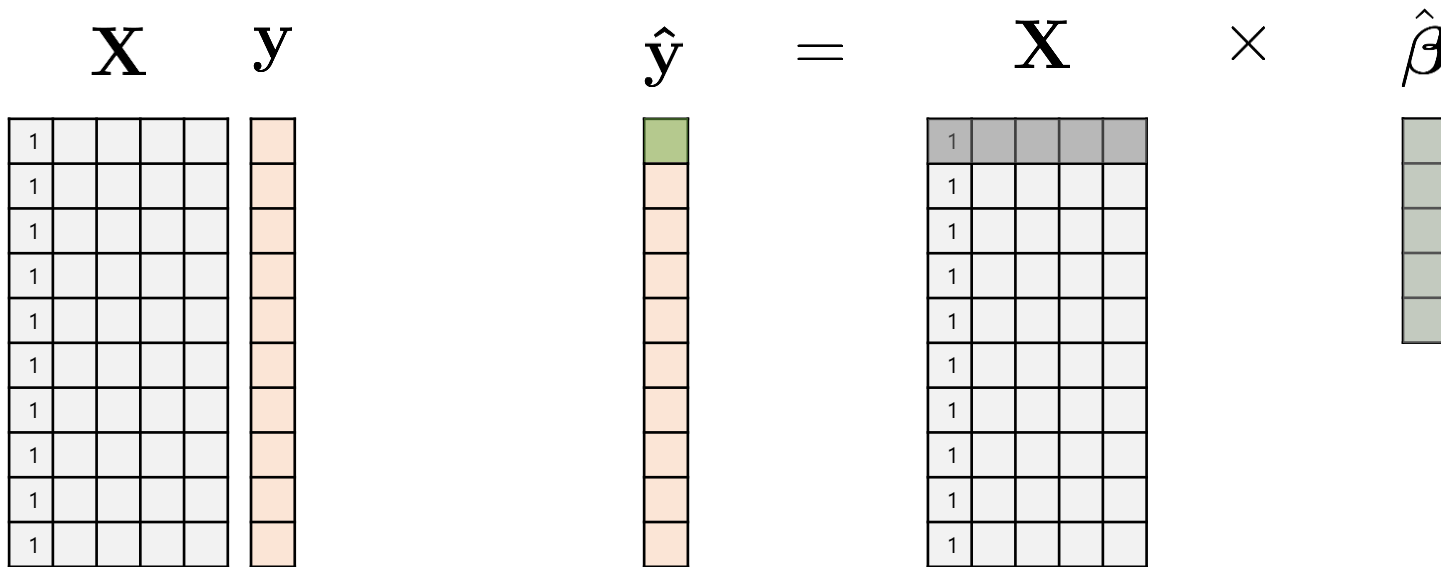
$$\begin{aligned} \min \frac{1}{2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ = \frac{1}{2} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \cdots + \hat{\beta}_d x_{id}) \right)^2 \end{aligned}$$

다중 선형 회귀 모형: 회귀 계수의 추정

- 최소 자승법: 행렬을 이용한 해 구하기

$\mathbf{X} : n \times (d + 1)$ matrix, $\mathbf{y} : n \times 1$ vector

$\hat{\boldsymbol{\beta}} : (d + 1) \times 1$ vector



상수항을 취급하기 위한 장치

다중 선형 회귀 모형: 회귀 계수의 추정

- 최소 자승법: 행렬을 이용한 해 구하기

$\mathbf{X} : n \times (d + 1)$ matrix, $\mathbf{y} : n \times 1$ vector

$\hat{\boldsymbol{\beta}} : (d + 1) \times 1$ vector

$$\min E(\mathbf{X}) = \frac{1}{2} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right)^T \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right)$$

$$\Rightarrow \frac{\partial E(\mathbf{X})}{\partial \hat{\boldsymbol{\beta}}} = -\mathbf{X}^T \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right) = 0$$

$$\Rightarrow -\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = 0$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y} \longrightarrow \text{학습 데이터에 대해 유일하고 명시적인 해(unique and deterministic solution)가 존재!}$$

다중 선형 회귀 모형: 회귀 계수의 추정

- 최소 자승법: 행렬을 이용한 해 구하기

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

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

The diagram illustrates the matrix equation $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ using grid representations for the matrices and vectors. The components are as follows:

- $\hat{\beta}$: A 5x1 green grid.
- \mathbf{X}^T : A 5x10 grid with the first row containing 1s and the rest empty.
- \mathbf{X} : A 10x5 grid with the first column containing 1s and the rest empty.
- \mathbf{X}^T : A 10x5 grid with the first row containing 1s and the rest empty.
- \mathbf{y} : A 10x1 orange grid.

The entire equation is enclosed in large parentheses with a -1 superscript on the \mathbf{X} matrix.

회귀 계수에 대한 Closed form solution이 존재

다중 선형 회귀 모형: 회귀 계수의 추정

- 최소자승법

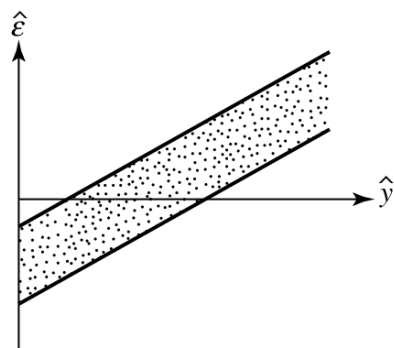
✓ 아래 조건을 만족할 경우 최소자승법으로 구한 회귀계수 β 는 최적해임

- 오차항 ε 이 정규분포를 따름
- 설명변수와 종속변수 사이에 선형관계가 성립함
- 각 관측치들은 서로 독립
- 종속변수 Y 에 대한 오차항^{residual}은 설명변수 값의 범위에 관계없이 일정함
(homoskedasticity)

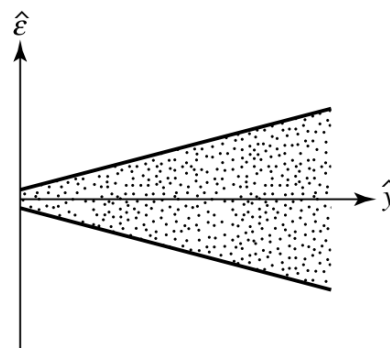
다중 선형 회귀 모형: 회귀 모형의 적합도

- 잔차도: Residual Plot

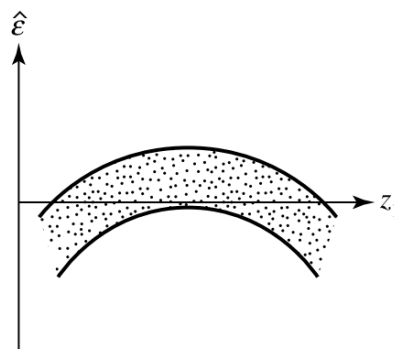
✓ 특정 설명변수 및 종속변수의 크기에 잔차가 영향을 받지 않아야 함



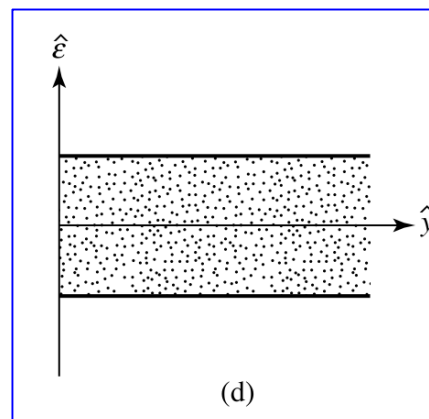
(a)



(b)



(c)



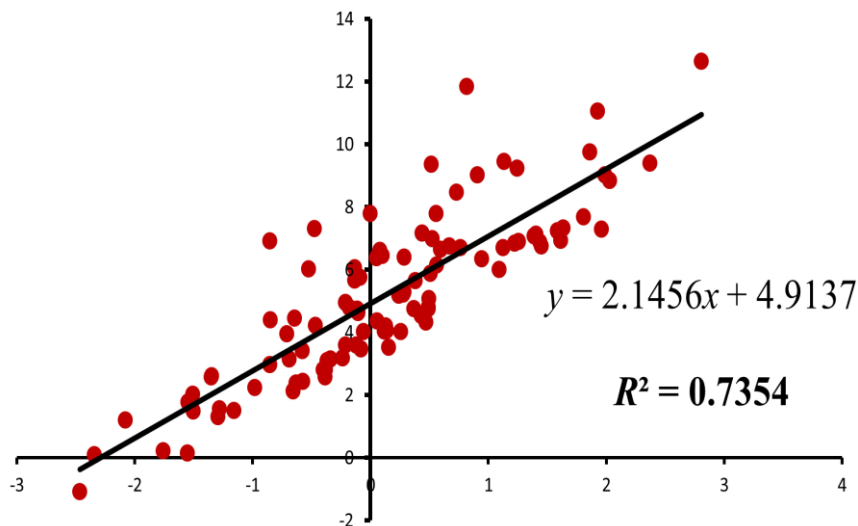
(d)

다중 선형 회귀 모형: 회귀 모형의 적합도

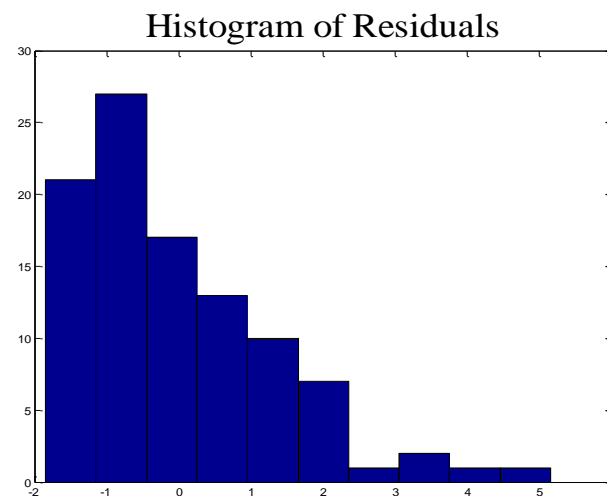
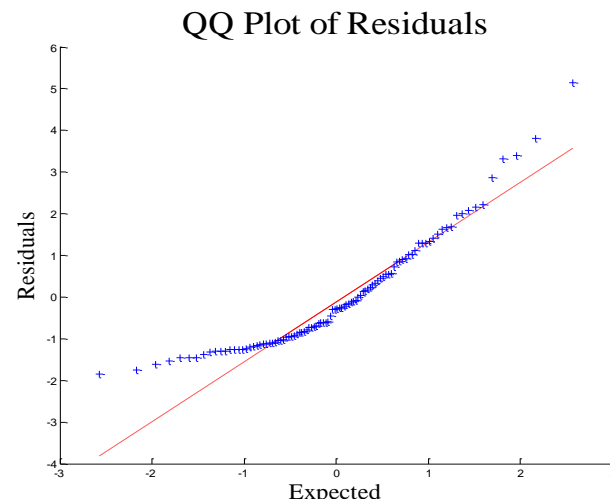
- 잔차의 정규성

✓ 산점도와 회귀직선을 보면 그럴듯 하지만...

$$y = 2x + \varepsilon, \quad \varepsilon \sim \text{Gamma}(2,1)$$



Regression model



다중 선형 회귀 모형: 회귀 모형의 적합도

- 종속변수의 전체 변동성(분산) = 회귀식이 설명할 수 있는 변동성 + 회귀식이 설명할 수 없는 변동성

$$\sum_{j=1}^n (y_j - \bar{y})^2 = \sum_{j=1}^n (\hat{y}_j - \bar{y})^2 + \sum_{j=1}^n \hat{\varepsilon}_j^2 .$$

(total sum of squares
about mean)

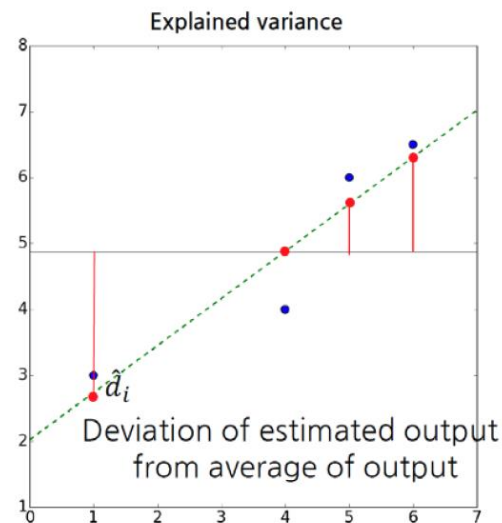
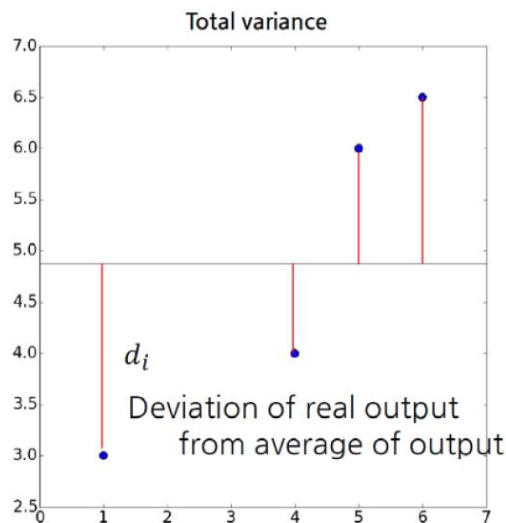
SST

(regression
sum of squares)

SSR

(residual (error)
sum of squares)

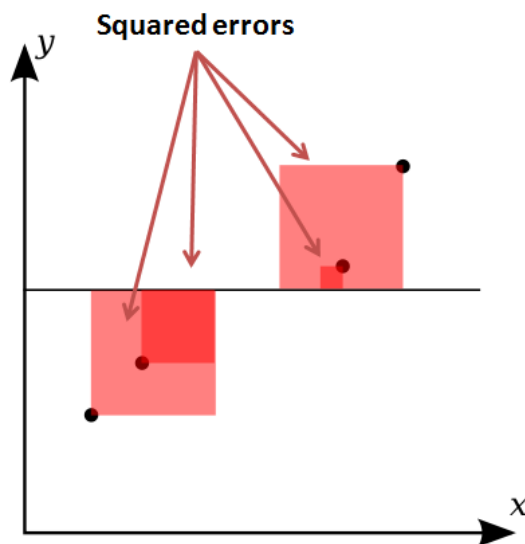
SSE



다중 선형 회귀 모형: 회귀 모형의 적합도

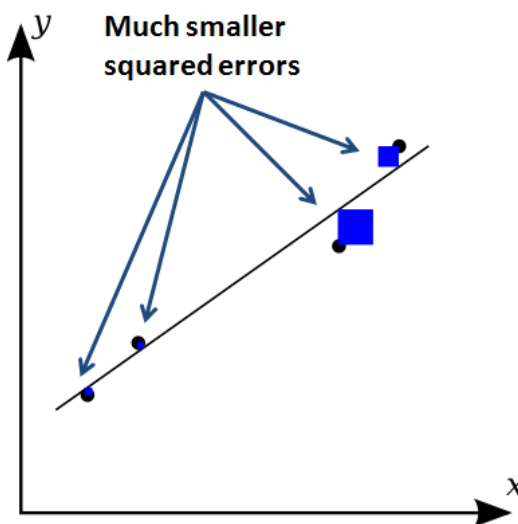
- 회귀 모형의 적합도 (결정계수, R^2) = 전체 변동성 중 회귀식이 설명할 수 있는 변동성의 비율

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \quad 0 \leq R^2 \leq 1$$



Computationally:

$$R\text{-squared} = 1 - \left[\frac{SS_{\text{error}}}{SS_{\text{total}}} \right]$$



Conceptually:

Force x and y to be independent, calculate **the squared error**.

Allow for a relationship between x and y, does this reduce your **error**?

다중 선형 회귀 모형: 회귀 모형의 적합도

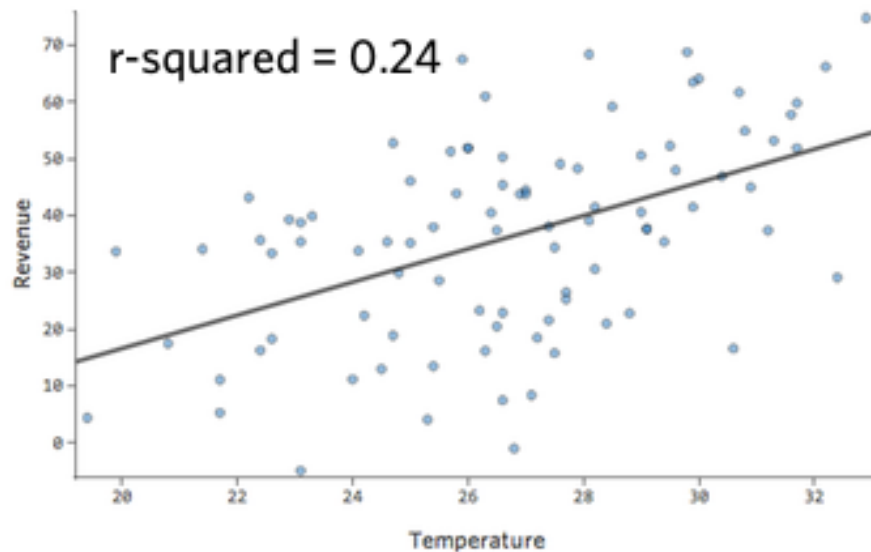
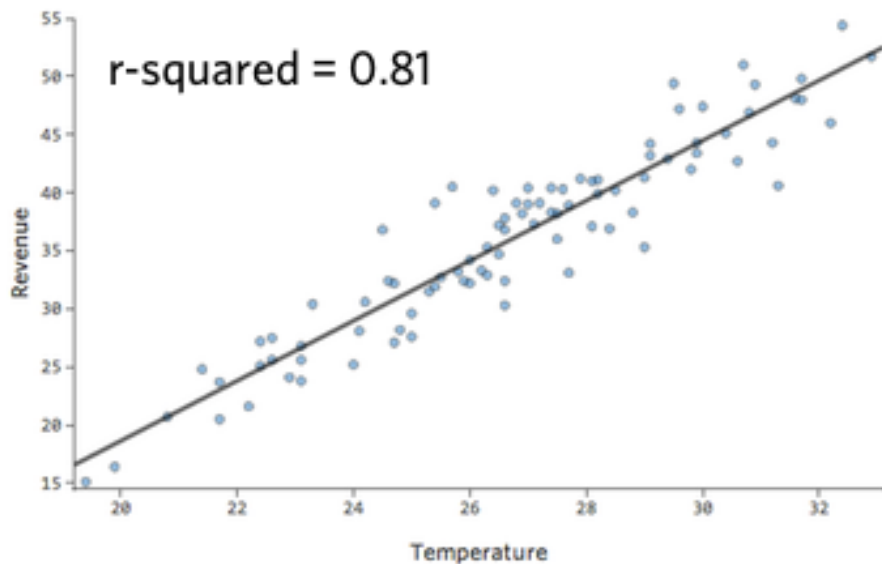
- 회귀 모형의 적합도 (결정계수, R^2) = 전체 변동성 중 회귀식이 설명할 수 있는 변동성의 비율

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \quad 0 \leq R^2 \leq 1$$

- ✓ 반응변수 (Y)의 전체 변동 중 예측변수(X)가 차지하는 변동의 비율
- ✓ R^2 는 0과 1 사이에 존재
- ✓ $R^2=1$: 회귀직선으로 Y의 총변동이 완전히 설명됨 (모든 측정값들이 회귀직선 위에 있는 경우)
- ✓ $R^2=0$: 추정된 회귀직선은 X와 Y의 관계를 전혀 설명하지 못함

다중 선형 회귀 모형: 회귀 모형의 적합도

- 회귀 모형의 적합도



✓ 선형회귀분석을 해봤더니 R^2 가 높게 나왔다

- 내가 분석을 잘했구나! (NO)
- 데이터의 입력변수와 출력변수 사이에 강한 선형 관계가 있구나 (YES)
- 왜? 동일한 데이터에 대해서는 누가 해도 같은 결과가 나오니까...

다중 선형 회귀 모형: 회귀 모형의 적합도

- 회귀 모형의 적합도

✓ 수정 결정계수 Adjusted R²:

$$R_{adj}^2 = 1 - \left[\frac{n-1}{n-(p+1)} \right] \frac{SSE}{SST} \leq 1 - \frac{SSE}{SST} = R^2$$

- R²는 유의하지 않은 변수가 추가되어도 항상 증가
- 수정 R²는 이러한 단점을 앞에 계수를 곱해줌으로써 보정
- 유의하지 않은 변수가 추가될 경우 수정 결정계수는 증가하지 않음

- 모형의 검토

✓ 추정된 모형이 다음 가정을 만족하는지 확인

- 예측변수와 반응변수 간 관계가 선형
- 오차항들이 서로 독립
- 오차항은 평균이 0이며 분산이 일정한 정규분포를 따름

다중 선형 회귀 모형: 예시

- 예시: 도요타 코롤라 중고차 가격 예측

Y **X**

Price	Age_08_04	KM	Fuel_Type	HP	Met_Color	Automatic	cc	Doors	Quarterly_Tax	Weight
13500	23	46986	Diesel	90	1	0	2000	3	210	1165
13750	23	72937	Diesel	90	1	0	2000	3	210	1165
13950	24	41711	Diesel	90	1	0	2000	3	210	1165
14950	26	48000	Diesel	90	0	0	2000	3	210	1165
13750	30	38500	Diesel	90	0	0	2000	3	210	1170
12950	32	61000	Diesel	90	0	0	2000	3	210	1170
16900	27	94612	Diesel	90	1	0	2000	3	210	1245
18600	30	75889	Diesel	90	1	0	2000	3	210	1245
21500	27	19700	Petrol	192	0	0	1800	3	100	1185
12950	23	71138	Diesel	69	0	0	1900	3	185	1105
20950	25	31461	Petrol	192	0	0	1800	3	100	1185
19950	22	43610	Petrol	192	0	0	1800	3	100	1185
19600	25	32189	Petrol	192	0	0	1800	3	100	1185
21500	31	23000	Petrol	192	1	0	1800	3	100	1185
22500	32	34131	Petrol	192	1	0	1800	3	100	1185
22000	28	18739	Petrol	192	0	0	1800	3	100	1185
22750	30	34000	Petrol	192	1	0	1800	3	100	1185
17950	24	21716	Petrol	110	1	0	1600	3	85	1105
16750	24	25563	Petrol	110	0	0	1600	3	19	1065

다중 선형 회귀 모형: 예시

- 데이터 전처리

- ✓ 원칙: Fuel type 변수에 대한 1-of-C coding 변환
- ✓ 선형회귀분석에서는 다중공선성 문제로 1-of-(C-1) coding을 사용

	Fuel_type = Diesel	Fuel_type = Petrol	Fuel_type = CNG
Diesel	1	0	0
Petrol	0	1	0
CNG	0	0	1

- 데이터 구분

- ✓ 가용한 모든 데이터를 전부 학습에 사용하면 과적합 (지금 가지고 있는 데이터는 잘 맞추지만 새로운 데이터를 잘 맞추지 못하는 문제) 위험이 있음
- ✓ 문제집을 답 보고 전부 풀었다고 해서 시험에서 100점을 맞는 것은 아니니까...

다중 선형 회귀 모형: 예시

- 다중회귀분석 결과물 해석

✓ 다중회귀분석을 수행하고 나면 다음과 같은 표를 결과로 얻을 수 있음

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.00000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000

다중 선형 회귀 모형: 예시

- 다중회귀분석 결과물 해석

- ✓ 회귀계수 Coefficient

- 선형회귀분석에서 각 변수에 대응하는 베타값임
- 해당 변수가 1단위 증가할 때 종속변수의 변화량을 의미
- 양수이면 해당 설명변수와 종속변수는 양의 상관관계, 음수이면 음의 상관관계

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.00000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000

다중 선형 회귀 모형: 예시

• 다중회귀분석 결과물 해석

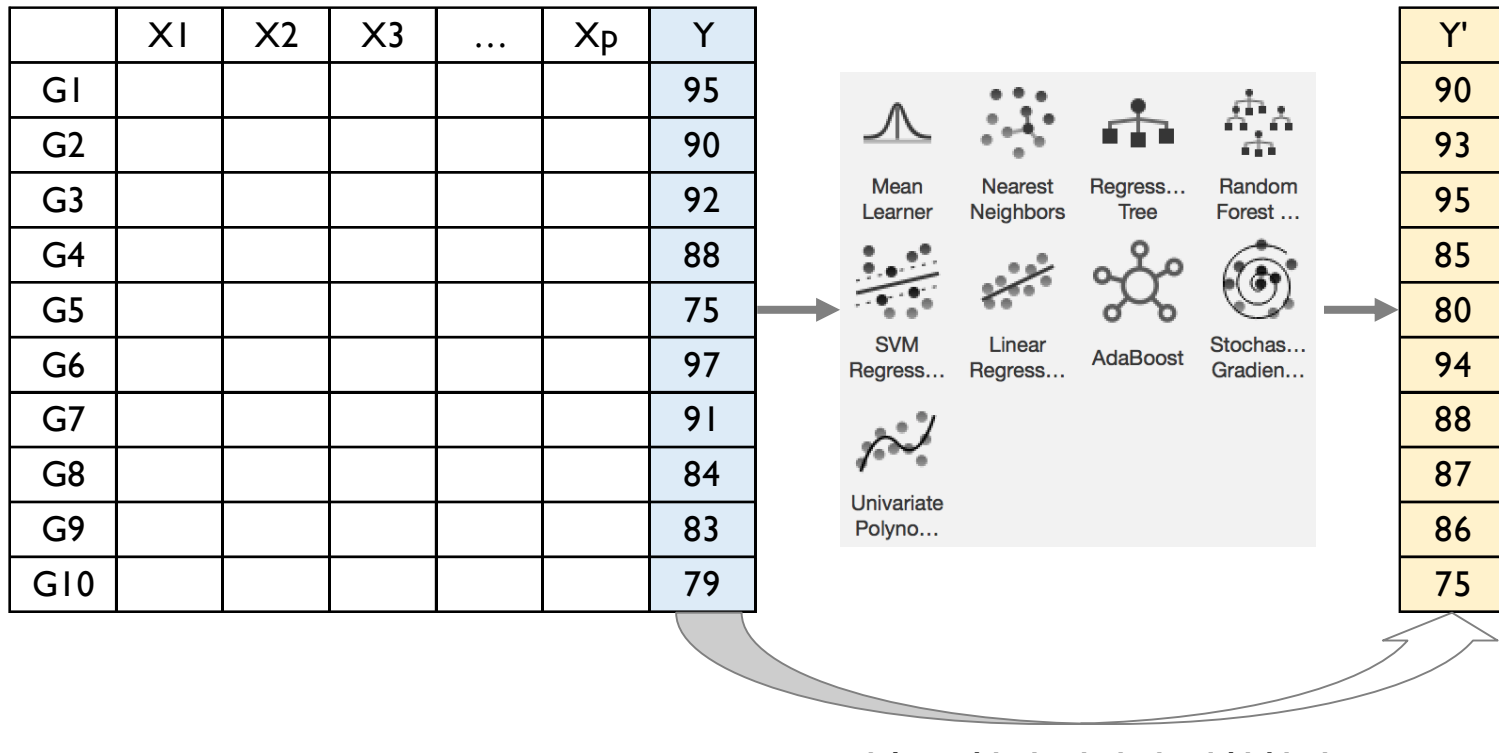
✓ 유의확률 p-value

- 선형회귀분석에서 해당 변수가 통계적으로 유의미한지 알려주는 지표
- 0에 가까울수록 모델링에 중요한 변수이며, 1에 가까울수록 유의미하지 않은 변수임
- 특정 유의수준(α)을 설정하여 해당 값 미만의 변수만을 사용하여 다시 선형회귀분석을 구축하는 것도 가능함 (주로 $\alpha = 0.05$ 사용)

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.00000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000

회귀모형 성능평가

- 예시: 설비 파라미터 X에 대한 제품의 수율(y) 예측

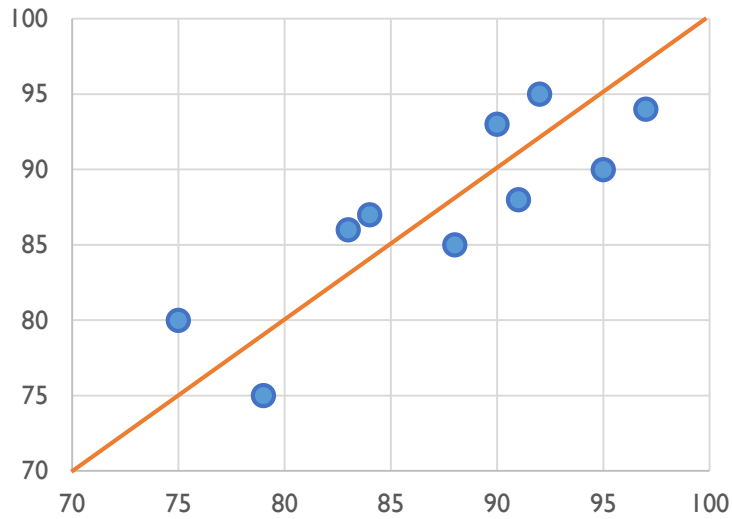


회귀모형 성능평가

- 성능지표 I: 평균오차 (Average Error)

- ✓ 실제 값에 비해 과대/과소 추정 여부를 판단
- ✓ 부호로 인해 잘못된 결론을 내릴 위험이 있음

$$\text{Average Error} = \frac{1}{n} \sum_{i=1}^n (y_i - y'_i)$$



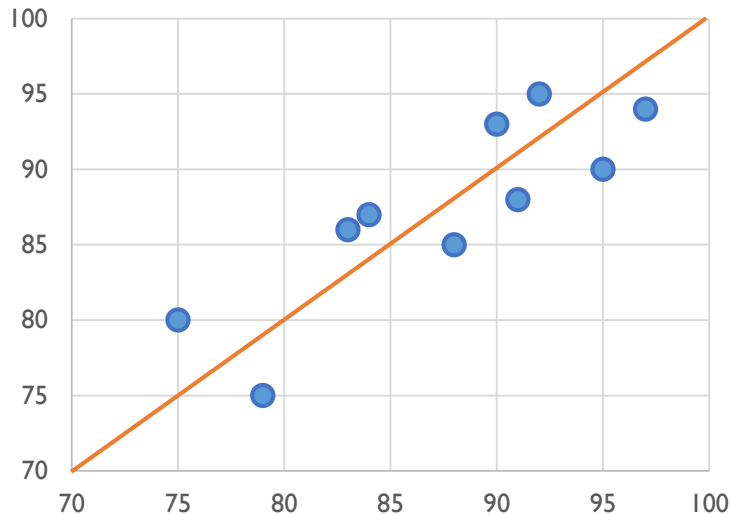
Y	Y'	Y-Y'
95	90	5
90	93	-3
92	95	-3
88	85	3
75	80	-5
97	94	3
91	88	3
84	87	-3
83	86	-3
79	75	4
Average Error		0.1

회귀모형 성능평가

- 성능지표 2: 평균 절대 오차(Mean absolute error; MAE)

✓ 실제 값과 예측 값 사이의 절대적인 오차의 평균을 이용

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - y'_i|$$



Y	Y'	Y-Y'
95	90	5
90	93	3
92	95	3
88	85	3
75	80	5
97	94	3
91	88	3
84	87	3
83	86	3
79	75	4
MAE		3.5

회귀모형 성능평가

- 성능지표 3: Mean absolute percentage error (MAPE)

- ✓ MAE의 단점: 실제 값과 절대적인 차이에 대한 정보만 제공하고, 상대적인 차이에 대한 정보를 제공하지 못함

- ✓ 아래 두 예시의 MAE는 모두 1임

Y	Y'	Y-Y'
1	0	1
1	2	1
1	0	1
1	2	1
1	0	1
1	2	1
1	0	1
1	2	1
1	0	1
1	2	1
MAE		1

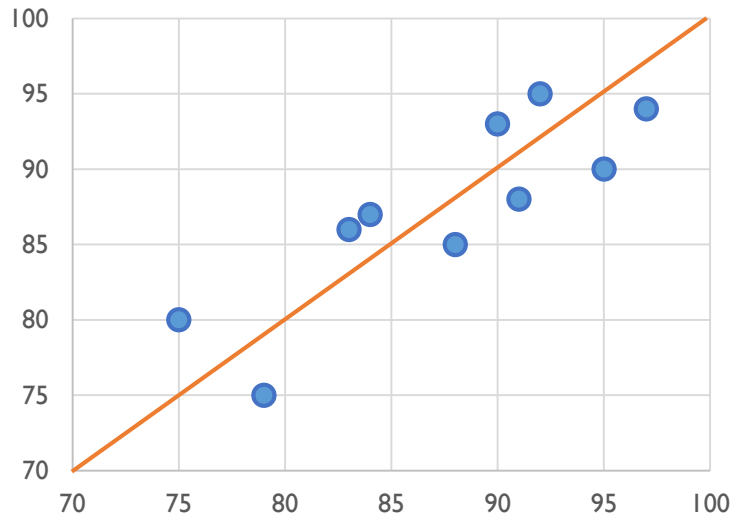
Y	Y'	Y-Y'
100	99	1
100	101	1
100	99	1
100	101	1
100	99	1
100	101	1
100	99	1
100	101	1
100	99	1
100	101	1
MAE		1

회귀모형 성능평가

- 성능지표 3: Mean absolute percentage error (MAPE)

- ✓ 실제값 대비 얼마나 예측 값이 차이가 있는지를 %로 표현
- ✓ 상대적인 오차를 추정하는데 주로 사용

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - y'_i}{y_i} \right|$$



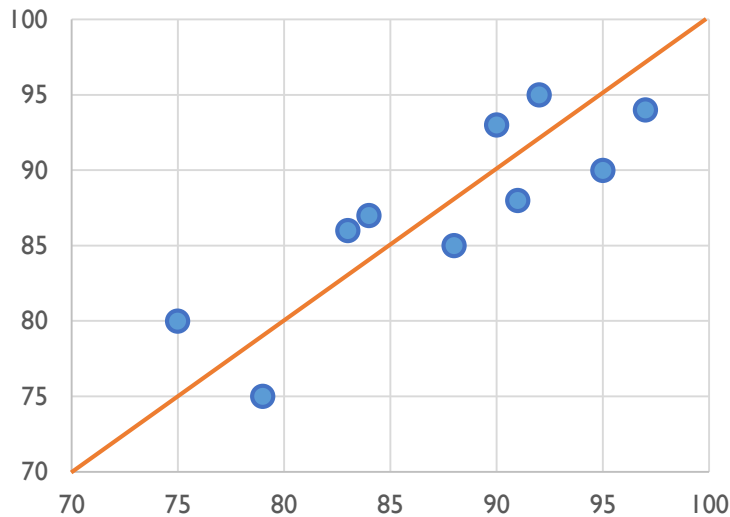
Y	Y'	Y-Y'	Y-Y' / Y
95	90	5	5.26%
90	93	3	3.33%
92	95	3	3.26%
88	85	3	3.41%
75	80	5	6.67%
97	94	3	3.09%
91	88	3	3.30%
84	87	3	3.57%
83	86	3	3.61%
79	75	4	5.06%
MAE		3.5	4.06%

회귀모형 성능평가

- 성능지표 4 & 5: (Root) Mean squared error ((R)MSE)

✓ 부호의 영향을 제거하기 위해 절대값이 아닌 제곱을 취한 지표

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2, \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2}$$



Y	Y'	(Y-Y') ²
95	90	25
90	93	9
92	95	9
88	85	9
75	80	25
97	94	9
91	88	9
84	87	9
83	86	9
79	75	16
MSE		12.9

$$\text{RMSE} = \sqrt{12.9} = 3.59$$

