선형회귀 (Linear Regression) 모델 - 3

2025 Spring

머신러닝1

이 두 호

개요

- 선형회귀 모델 1
  - 선형회귀 모델 개요
  - 선형회귀 모델 가정
- 선형회귀 모델 2
  - 파라미터 추정 (최소제곱법)
- 선형회귀 모델 3 ✓
  - 결정계수 (R^2)

$$Y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p + \epsilon$$

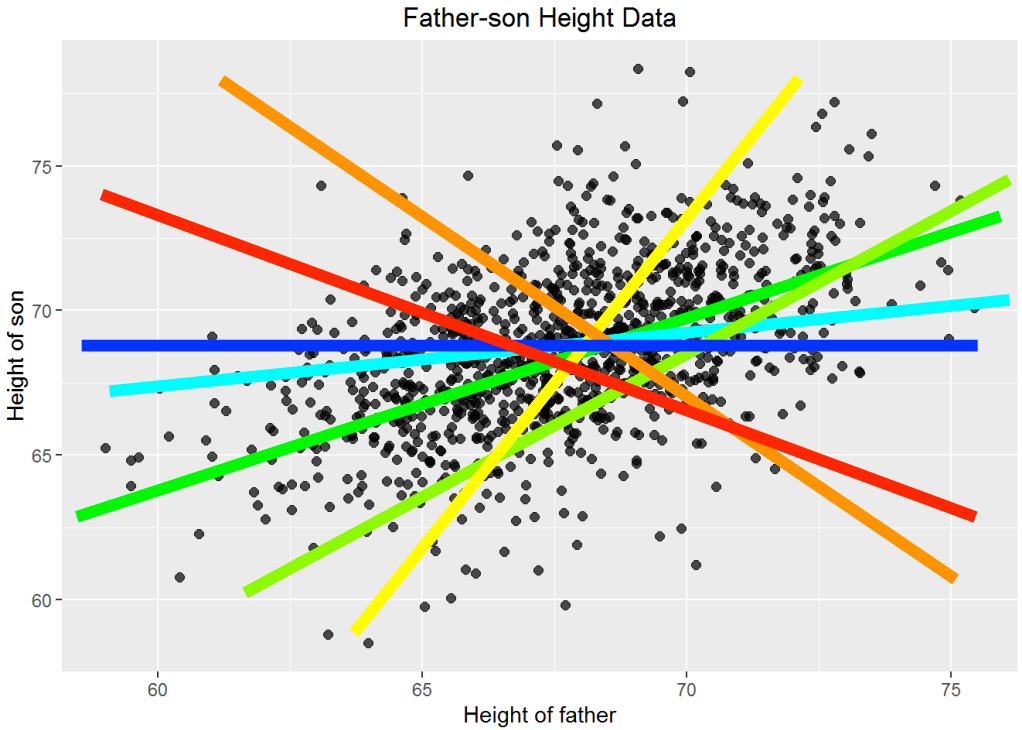
$$E[Y] = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

입력변수(X)와 출력변수(Y) 평균(기대값)

사이의 관계를 정량화하여 선형식으로 표현하기!

선형회귀 모델

$$E[Y] = f(X) = w_0 + w_1 X$$

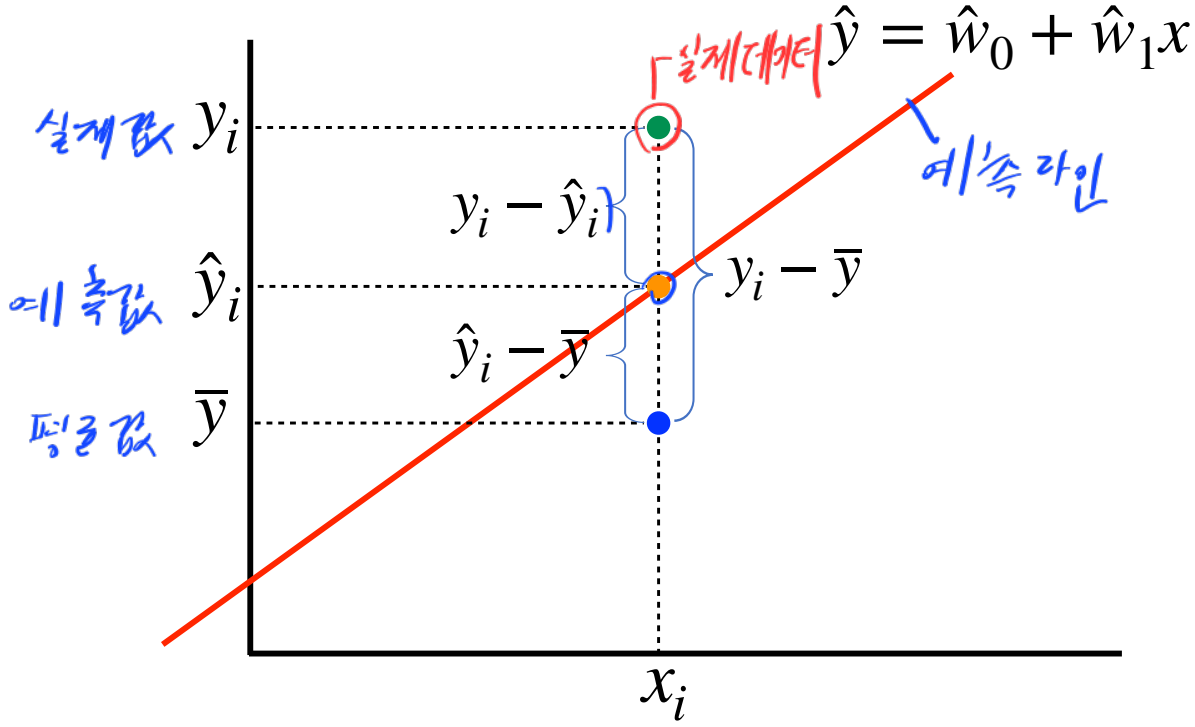


파라미터 (parameter)

파라미터를 찾자 (추정, estimation)

우리가 가지고 있는 데이터들의 함수식으로!

결정계수 (Coefficient of Determination, R^2)

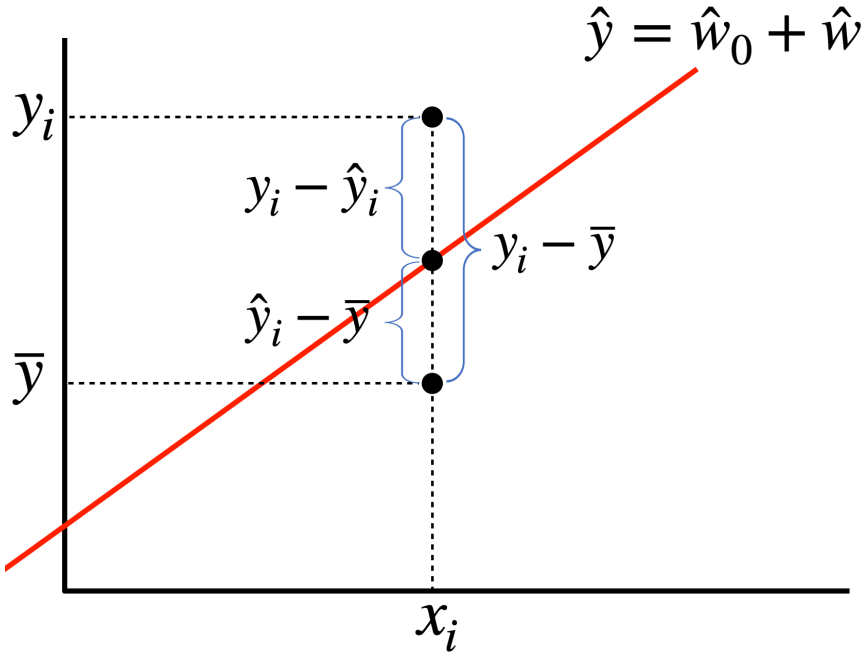


$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

증명법 찾아보기

$$SST = SSE + SSR$$

결정계수 (Coefficient of Determination, R^2)



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

$$SSR = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$$

$$SST = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = SSE + SSR$$

$$\frac{SSR}{SST} = 1 \qquad \frac{SSR}{SST} = 0$$

$$\frac{SSR}{SST} = R^2$$

결정계수 (Coefficient of Determination, R^2)

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

- 결정계수 (R^2)
  - $0 \leq R^2 \leq 1$
  - $R^2 = 1$: 현재 가지고 있는 입력변수 X 들로 출력변수 Y를 100% 설명할 수 있다. 즉, 모든 관측치가 회귀직선(평면) 상에 있다.
  - $R^2 = 0$: 현재 가지고 있는 입력변수 X들은 출력변수 Y를 설명(예측)하는 데에 전혀 도움이 되지 않는다.
  - 사용하고 있는 입력변수 X 들이 출력변수 Y의 분산(변동)을 얼마나 줄였는지 정도
  - 단순히 출력변수 Y의 평균값을 사용했을 때 대비 입력변수 X 들의 정보를 사용함으로써 얻는 학습모델의 성능향상 정도
  - 출력변수 Y를 예측함에 있어서 사용하고 있는 입력변수 X들의 품질

예제 1

	x1	x2	y
1	9	299	18.24
2	12	1036	36.32
3	17	851	43.42
4	25	274	43.10
5	7	1377	34.94
6	23	1375	60.23
7	27	219	49.53
8	16	1217	50.08
9	15	699	33.10
10	3	814	23.52
11	5	818	20.27
12	20	373	36.58
13	13	1110	37.67
14	8	291	20.45
15	18	607	41.03
16	28	1239	65.17
17	11	1411	38.18
18	24	353	44.31
19	6	661	25.93
20	30	141	53.58
21	10	965	31.09
22	21	580	41.91
23	29	1210	66.69
24	2	247	3.86
25	14	522	31.22

```
set.seed(1)
x1 <- sample(2:30, 25)
set.seed(2)
x2 <- sample(36:1460, 25)
set.seed(3)
y <- round(2.341 + 1.6159*x1 + 0.015*x2 + rnorm(25,0,3.25),2)
df1 <- data.frame(x1,x2,y)
View(df1)
```

1. SSE, SSR, SST 를 구하라.

2. R^2 를 구하라.

수정 결정계수 (Adjusted R^2)

$$R^2 = 1 - \frac{SSE}{SST}$$

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

- R^2 은 유의하지 않은 변수가 추가되어도 항상 증가

결정계수의 무의미한 증가 방지

- 수정 결정계수 R^2_{adj} 은 앞에 특정 계수를 곱해 줌으로써 유의하지 않은 변수가 추가 될 경우 증가하지 않게 함
- 입력변수가 서로 다른 회귀모형의 설명력(예측성능)을 비교할 때 사용

$$R^2 \geq R^2_{adj}$$

예제 2

전국적으로 500개의 대리점을 가지고 있는 요플레 제조회사에서 각 대리점의 판매원 수와 광고비 지출이 매출액에 어떤 영향을 미치는가를 알아보기 위해 10개의 대리점에 대한 자료를 수집하였다.

판매원 수 (X_1)	광고비 (X_2)	월간 매출액 (Y)
14	37	850
16	43	970
13	38	730
10	42	940
18	36	920
17	33	830
16	40	940
15	35	900
11	34	760
10	29	710

$$SSE,\ SSR,\ SST,\ R^2,\ R^2_{adj}$$

예제 2

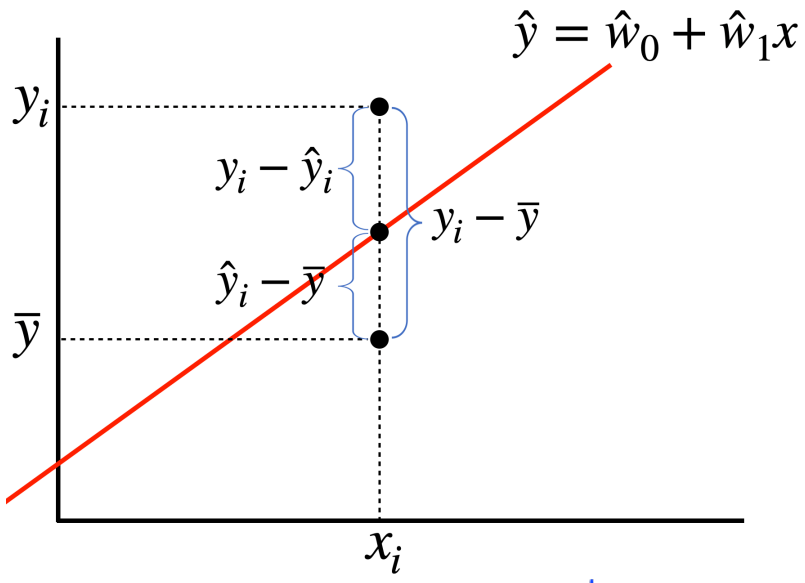
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  141.516    186.220   0.760   0.4721
x1            13.035      7.029   1.854   0.1061
x2            14.469      4.783   3.025   0.0193 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.29 on 7 degrees of freedom
Multiple R-squared:  0.683,     Adjusted R-squared:  0.5924
F-statistic:  7.54 on 2 and 7 DF,  p-value: 0.01794
```

- 판매원 수(X1)와 광고비(X2) 변수에 의해 매출액(Y) 변수의 변동성을 68.3% 감소
- 매출액의 평균 대비 판매원 수와 광고비를 이용하면 설명력이 68.3% 증가
- 현재 분석에 사용하고 있는 판매원 수와 광고비의 “변수 품질” 정도가 100점 만점 기준에 68.3 점

선형회귀모델에서의 분산분석

- 분산분석 : ANalysis Of VAriance (ANOVA)
- 분산(변동량)정보를 이용하여 학습모델의 타당성 분석
- 분산분석은 궁극적으로 가설검정을 행하는 용도로 사용됨



$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$: Y의 총 변동량

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$: X변수에 의해 설명된 양

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$: 오차에 의해 설명된 양

선형회귀모델에서의 분산분석

SST : Y의 총 변동량

SSR : X변수에 의해 설명된 양

SSE : 오차에 의해 설명된 양

귀무가설
대립가설

$$\frac{SSR}{SSE} > 1$$

- X 변수에 의해 설명된 양 > 오차에 의해 설명된 양
- X 변수가 Y를 설명(예측)하는 데에 유의미한 영향
- X 변수의 계수 (가중치)가 0이 아님

$$0 \leq \frac{SSR}{SSE} \leq 1$$

- X 변수에 의해 설명된 양 < 오차에 의해 설명된 양
- X 변수가 Y를 설명(예측)하는 데에 영향을 끼치지 못함
- X 변수의 계수 (가중치)가 0이라 할 수 있음

선형회귀모델에서의 분산분석

$$\frac{SSR}{SSE}$$

$Y_1 \sim \chi^2(n_1)$

$Y_2 \sim \chi^2(n_2)$

$$X = \frac{Y_1/n_1}{Y_2/n_2} = \sim F(n_1, n_2)$$

- 얼마나 커야 큰 값인지? 새로운 사용가능
- 확률분포를 알면 통계적으로 판단할 수 있음
- 안타깝게도 직접적으로 확률분포를 정의할 수 없음
- 하지만, SSR과 SSE가 각각 카이제곱 분포를 따름

Let $Y_1 \sim \chi^2(\nu_1)$ and $Y_2 \sim \chi^2(\nu_2)$, and define $X = \dfrac{Y_1/\nu_1}{Y_2/\nu_2}$. Then, we have $X \sim F(\nu_1, \nu_2)$.

선형회귀모델에서의 분산분석

Let $Y_1 \sim \chi^2(\nu_1)$ and $Y_2 \sim \chi^2(\nu_2)$, and define $X = \dfrac{Y_1/\nu_1}{Y_2/\nu_2}$. Then, we have $X \sim F(\nu_1, \nu_2)$.

$$\frac{SSR}{\sigma^2} \sim \chi^2(p), \quad \frac{SSE}{\sigma^2} \sim \chi^2(n-1-p)$$

$$\frac{SSR/p}{SSE/(n-1-p)} \sim F(p, n-1-p)$$

통계검정 (testing)

가설 (hypothesis)

귀무가설 (null) < 기존의 가설

대립가설 (alternative)

표현식

$H_0 : \mu = 165$

$H_a : \mu \neq 165$ or $\mu > 165$ (H_1)

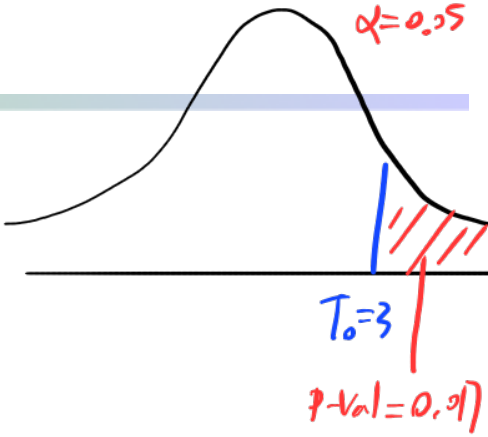
표본 → 검정통계량 (T_0)

유의수준 5% = α (significance level) 일정오류

$\alpha > p\text{-val}$ → 귀무가설을 지지할 근거가 없으므로 기각한다.

유의확률 (p-값) 귀무가설이 참 이라는 가정하에 나의 데이터가 관측될 확률

선형회귀모델에서의 분산분석



Source	DF	SS	MS	F	P
Model	*p*	SSR	MSR	F*	P-value
Error	*n*-1-*p*	SSE	MSE		
Total	*n*−1	SST			

ANOVA Table

$$MSR = SSR/p, \quad MSE = SSE/(n-1-p)$$

$$F^* = \frac{MSR}{MSE}, \quad P\text{-value} = \Pr\{X > F^*\}$$

$$X \sim F(p, n-1-p)$$

예제 3

판매원 수 (X_1)	광고비 (X_2)	월간 매출액 (Y)
14	37	850
16	43	970
13	38	730
10	42	940
18	36	920
17	33	830
16	40	940
15	35	900
11	34	760
10	29	710

Source	DF	SS	MS	F	P-value
Model	2	54809.18	27404.59	7.540	0.01794099
Error	7	25440.82	3634.40		
Total	9	80250.00			

$H_0 : w_1 = w_2 = 0$ vs H_1 : 적어도 하나의 $w \neq 0$

Reject H_0 : 판매원 수 혹은 광고비 혹은 두가지 모두 유의미

예제 2

Source	DF	SS	MS	F	P-value
Model	2	54809.18	27404.59	7.540	0.01794099
Error	7	25440.82	3634.40		
Total	9	80250.00			

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  141.516    186.220   0.760   0.4721
x1            13.035      7.029   1.854   0.1061
x2            14.469      4.783   3.025   0.0193 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.29 on 7 degrees of freedom
Multiple R-squared:  0.683,     Adjusted R-squared:  0.5924
F-statistic:  7.54 on 2 and 7 DF,  p-value: 0.01794
```

HW

state <- as.data.frame(state.x77)

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766
Connecticut	3100	5348	1.1	72.48	3.1	56.0	139	4862
Delaware	579	4809	0.9	70.06	6.2	54.6	103	1982
Florida	8277	4815	1.3	70.66	10.7	52.6	11	54090
Georgia	4931	4091	2.0	68.54	13.9	40.6	60	58073
Hawaii	868	4963	1.9	73.60	6.2	61.9	0	6425
Idaho	813	4119	0.6	71.87	5.3	59.5	126	82677
Illinois	11197	5107	0.9	70.14	10.3	52.6	127	55748
Indiana	5313	4458	0.7	70.88	7.1	52.9	122	36097
Iowa	2861	4628	0.5	72.56	2.3	59.0	140	55941
Kansas	2280	4669	0.6	72.58	4.5	59.9	114	81787
Kentucky	3387	3712	1.6	70.10	10.6	38.5	95	39650
Louisiana	3806	3545	2.8	68.76	13.2	42.2	12	44930
Maine	1058	3694	0.7	70.39	2.7	54.7	161	30920
Maryland	4122	5299	0.9	70.22	8.5	52.3	101	9891
Massachusetts	5814	4755	1.1	71.83	3.3	58.5	103	7826
Michigan	9111	4751	0.9	70.63	11.1	52.8	125	56817
Minnesota	3921	4675	0.6	72.96	2.3	57.6	160	79289
Mississippi	2341	3098	2.4	68.09	12.5	41.0	50	47296

Population, Income, Illiteracy, Life Exp, Frost 를 입력변수로 하고, Murder 를 출력변수로 한다. ANOVA 테이블을 작성하라.