

5장 데이터 분석기법(회귀분석)

- 회귀분석

- 선형회귀분석
 - 단순선형회귀분석
 - 다중선형회귀분석
- 결과 분석 방법
- 이상값 (outlier)의 영향
- 역 회귀 분석
- 주성분 분석에 의한 설명변수의 합성
- 로지스틱 회귀분석

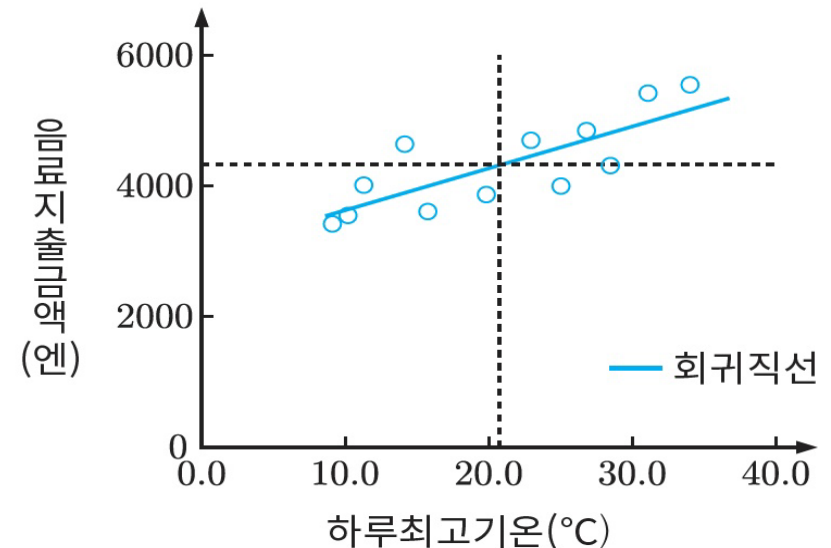
01-1 회귀분석과 상관관계

• 회귀분석과 상관분석

- 변수들 간의 관계를 분석한다.
- 회귀분석은 하나의 변수값으로부터 다른 변수값을 예측하는것이 목적이 되며, 상관분석은 두 변수 간의 관계가 얼마나 강한지를 나타낸다.
- 즉, 회귀는 변수들 간에 원인과 결과의 관계를 나타내고, 상관은 변수들이 함께 변화하는 정도를 알려 준다.

(ex) 하루최고기온으로부터 음료지출을 예측:

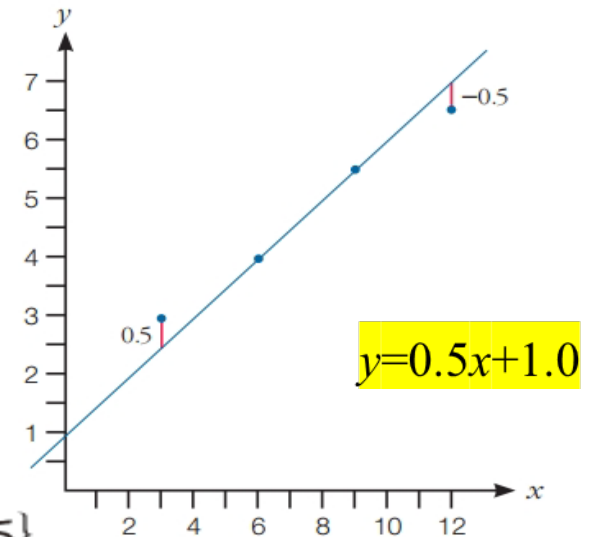
- 상관계수(r): 0.8
- 결정계수(R^2): 0.64
- 회귀직선의 실제 데이터에 대한 적합성은 나쁘지 않음



01-2 독립변수와 종속변수

- 회귀분석은 독립변수와 종속변수의 관계를 함수식으로 나타내고, 독립변수를 이용하여 종속변수의 값을 설명하거나 예측한다.

- X: 독립변수(independent variable),
설명변수(explanatory variable)
- Y: 종속변수(dependent variable),
반응변수(response variable)



$$X = \{3.0, 6.0, 9.0, 12.0\}, Y = \{3.0, 4.0, 5.5, 6.5\}$$

- 두 변수 간에 상관관계가 높으면 독립변수는 종속변수를 더 잘 설명할 수 있고, 또한 독립변수의 값으로부터 종속변수의 값을 보다 정확히 예측할 수 있다. 따라서 회귀분석은 상관관계 분석을 바탕으로 이루어진다.

01-3 회귀분석의 종류

- **선형 회귀분석(Linear Regression)**: 일반적으로 사용하는 모델링 기술 중 하나이며 종속 변수(y)는 연속적이며, 독립 변수(x)는 연속적이거나 이산적일 수 있다. 회귀선은 선형을 가진다.
 - 단순선형 회귀(simple linear regression): 독립변수가 1개 존재
 - 다중선형회귀(multiple linear regression): 독립변수가 2개 이상 존재

Simple
Linear
Regression

$$y = b_0 + b_1 * x_1$$

Multiple
Linear
Regression

Dependent variable (DV) Independent variables (IVs)

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$$

Constant Coefficients

- **로지스틱 회귀분석(Logistic Regression)**: 종속변수(y)가 이진(1/0, 참/거짓, 예/아니요)일 때 주로 사용되는 회귀 분석이다.
- **비선형 회귀분석(Nonlinear Regression)**

01-4 모델 구축 (설명 모델과 예측모델)

- 회귀 분석의 목적

- 통계학 분야: 독립변수와 종속변수 사이의 관계를 설명하고자 함.
- 데이터마이닝: 새로운 사례에 대한 결과값을 예측하고자 함.

- 설명모델 (Explanatory model)

- 데이터가 작은 경우의 모델
- 모집단에서 가정하는 가설관계에 대한 정보를 잘 반영할 수 있도록 전체 데이터를 사용하여 최상의 적합모델을 추정하고자 함

- 예측모델 (Predictive model)

- 데이터가 많은 경우의 모델 (데이터마이닝 분야)
- 적합한 모델을 이용하여 알려지지 않은 데이터에 대한 예측을 목적으로 함.
- 학습용 데이터(모델 구축)와 평가용 데이터(성능평가)로 나누어 사용함.

02-1 선형회귀분석: 단순선형회귀 분석의 예

- 하루최고기온으로부터 음료지불금액을 예측

월	1	2	3	4	5	6
하루최고기온(℃)	9.1	10.2	14.1	19.8	25.0	26.8
음료지출금액(엔)	3416	3549	4639	3857	3989	4837
월	7	8	9	10	11	12
하루최고기온(℃)	31.1	34.0	28.5	22.9	15.7	11.3
음료지출금액(엔)	5419	5548	4311	4692	3607	4002

$$y = b_0 + b_1 \times x_1$$

- 최소제곱법을 사용해 회귀직선을 구함

$$\hat{y} = 2947.8 + 66.4 \times x$$

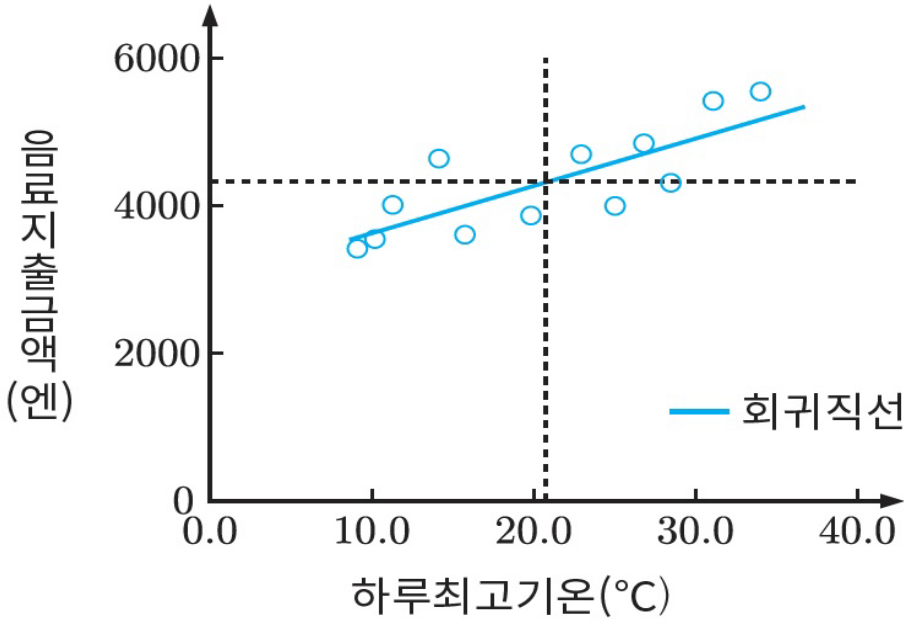
하루최고기온: 9.1~34.0℃

하루최고기온이 10℃인 경우 음료지불금액

: $2947.8 + 66.4 \times 10 = 3611.8$

기온이 1℃ 상승하면 평균적으로 음료지불금액이 66.4엔씩 증가

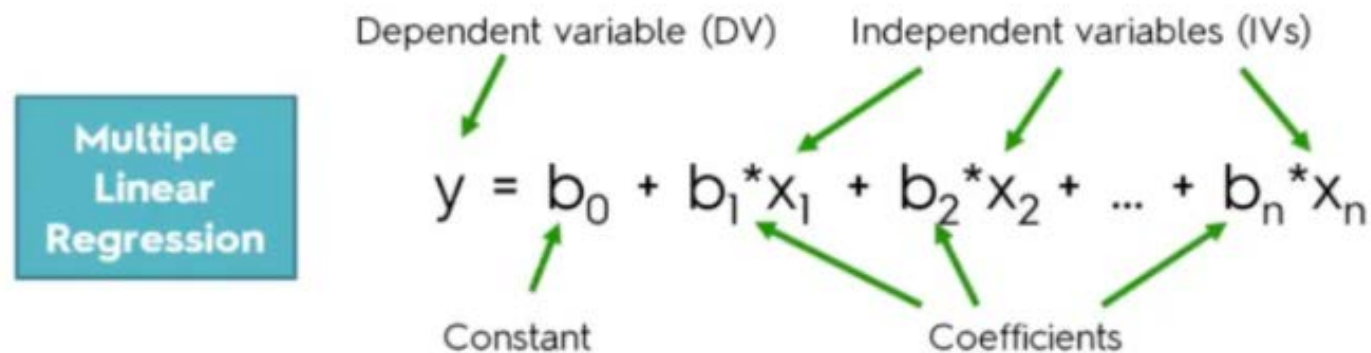
회귀직선은 각 변수의 평균값을 좌표로 갖는 점(20.7, 4322.2)를 통과



$$b_0 = \bar{y} - b_1 \bar{x}$$

02-2 선형회귀분석: 다중 선형회귀 분석

- 종속 변수의 변화가 하나의 독립변수만으로 충분히 설명할 수 없는 경우가 많음. 따라서 독립변수를 적절히 여러 개 선택하여 이들의 함수로서 종속변수를 설명하는 것이 더 정확할 수 있음.
- 이 경우의 모델을 다중선형회귀 모델이라 한다.



The diagram illustrates the Multiple Linear Regression equation: $y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$. A blue box on the left contains the text "Multiple Linear Regression". Green arrows point from labels to parts of the equation: "Dependent variable (DV)" points to y ; "Independent variables (IVs)" points to x_1, x_2, \dots, x_n ; "Constant" points to b_0 ; and "Coefficients" points to b_1, b_2, \dots, b_n .

- 여기에서, $b_0, b_1, b_2, \dots, b_n$ 은 구해야 할 회귀계수이다.
- 최소제곱법(least square method)에 의해 회귀계수를 구할 수 있다.
- 각 회귀계수에 대한 검정도 단순회귀분석의 경우와 동일하게 수행한다. (예: 결정계수)

02-2 선형회귀분석: 다중 선형회귀 분석 (예제)

- "더운 날에는 아이스크림이 많이 팔릴 것이다"

→ 아이스크림 판매수량과 최고기온 데이터를 조사하여 회귀분석을 수행

$$\hat{y} = 210.8 + 134.2x$$

→ 예상최고기온이 30 °C 라면 예상 판매수량은 4236.8개

최고기온이 1°C 상승하면,
아이스크림의 판매수량은
134.2개씩 증가

- "가격을 저렴하게 하면 많이 판매된다"

→ 회귀식의 오른쪽 항에 변수를 추가

- "평일보다는 휴일에 많이 팔린다"

→ 더미변수(dummy variable): 휴일인 경우에는 1, 평일인 경우에는 0

- 회귀직선: $\hat{y} = 195.4 + 118.1x - 5.8p + 30.4D$

p: 아이스크림 1개의 가격(엔)

D: 더미변수(휴일이면 1, 평일이면 0)

"아이스크림의 가격을 1엔 올리면 매상이 5.8개 줄어든다"

"휴일은 평일에 비해서 매상이 30.4개 올라간다"

02-2 선형회귀분석: 다중 선형회귀 분석

The diagram shows the equation $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$. Green arrows point from labels to parts of the equation: 'Dependent variable (DV)' points to y , 'Independent variables (IVs)' points to x_1, x_2, \dots, x_n , 'Constant' points to b_0 , and 'Coefficients' points to b_1, b_2, \dots, b_n .

- 회귀모델에 포함시키는 독립변수의 선정 기준

- 종속변수와 높은 상관관계를 갖는다.
- 선택된 독립변수들은 서로 상호간에 낮은 상관관계를 갖는다. (다중공선성 문제 회피)
- 독립변수의 개수는 적을수록 좋다

❖ **다중공선성(multicollinearity):** 독립변수들 간에 밀접한 상관관계가 존재하는 것을 말하며, 이와 같은 경우에는 독립변수의 계수가 정확히 추정되지 못하는 문제가 발생함.

- 독립변수의 선택 방법

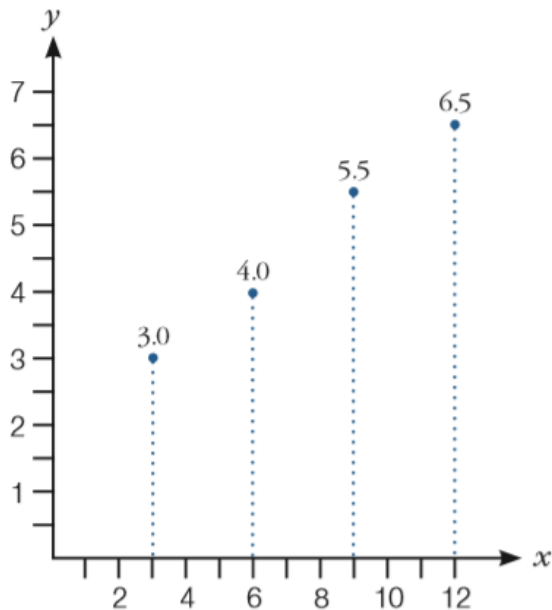
- All possible regression: 변수들의 모든 가능한 조합으로부터 최적의 모델을 찾아낸다. 탐색시간이 많이 드는 단점이 있음.
- Forward stepwise selection: 기여도가 높은 유의한 변수부터 하나씩 추가하는 방법. 탐색시간이 빠르다.
- Backward stepwise selection: 모든 변수를 포함한 상태에서 불필요한 변수를 제거해 나가는 방법. 중요변수가 제외될 가능성이 적음.

02-3 회귀모형의 결과분석 방법 (단순선형회귀 분석 예)

`lm(formula, data, ...)`: 선형회귀 모델을 생성하기 위한 함수

- `formula` : 반응변수 ~ 설명변수의 형태로 지정한 식
- `data` : 변수가 포함된 데이터 프레임

$X = \{3.0, 6.0, 9.0, 12.0\}$, $Y = \{3.0, 4.0, 5.5, 6.5\}$



```
> x = c(3.0, 6.0, 9.0, 12.0)
> y = c(3.0, 4.0, 5.5, 6.5)
> m = lm(y ~ x)
> m
```

Call:

```
lm(formula = y ~ x)
```

Coefficients:

(Intercept)	x
1.75	0.40

$$y = 0.4x + 1.75$$

lm(formula, data=)로 구한 모델을 summary()로 요약하면

```
> summary(m)                # 모델의 상세 분석

Call:
lm(formula = y ~ x)

Residuals:
    1     2     3     4 
0.05 -0.15  0.15 -0.05 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.75000    0.19365   9.037  0.01202 *
x            0.40000    0.02357  16.971  0.00345 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1581 on 2 degrees of freedom
Multiple R-squared:  0.9931,    Adjusted R-squared:  0.9897 
F-statistic: 288 on 1 and 2 DF,  p-value: 0.003454
```

- **Residual**: 모델로 예측한 y 값과 실제 데이터의 y 값과의 차이를 의미한다. 패턴/추세가 보이면 안되며, residual plot으로 관측 가능하다.
- **Coefficient**: Estimate 컬럼: 절편과 각 x 의 기울기 값이 출력된다. 또한 예측한 각 회귀계수에 대한 유의성을 나타낸다.
 - t 값(t 검정에 대한 통계량)은 독립변수(x)와 종속변수(y)간에 선형관계(관련성)가 존재하는 정도를 나타낸다. t 값은 회귀계수 나누기 표준오차(표준편차)가 된다. 유의미한 결과가 나오려면 t 값이 커야 한다 (절대값이 2보다 커야 한다).
(ex) $1.75/0.19365=9.037$ $0.4/0.02357=16.971$
 - **유의수준(significance level)**: p -value로 표기되며, 관찰된 데이터의 검정 통계량이 귀무가설(영가설)을 지지하는 정도를 확률로 표현한 것. 유의수준 0.05로 설정한다면, p -값이 유의수준보다 작으므로 **귀무가설** (독립변수 x 는 종속변수 y 는 아무 관련이 없다)은 기각된다. 즉 두 변수는 관련이 있다라는 **대립 가설**이 받아들여진다.

Residuals:

1	2	3	4
0.05	-0.15	0.15	-0.05

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.75000	0.19365	9.037	0.01202 *
x	0.40000	0.02357	16.971	0.00345 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> plot(x, y)
> abline(m, col = 'red')
```

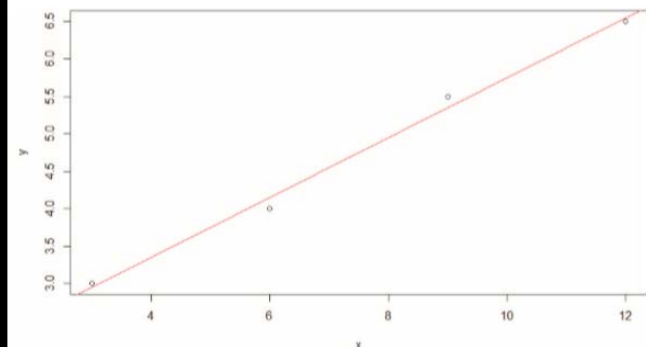


그림 7-4 lm 함수로 찾은 최적 모델

귀무가설/대립가설

- 통계학의 궁극적인 목표는 기존 주장이 맞는지 아니면 새로운 연구 또는 실험으로 발견된 주장이 맞는지 검토하는 것이다. 그래서 최종적으로 귀무가설을 채택하거나 기각, 또는 대립가설을 채택하거나 기각하는 선택을 하는 것이다.
- **귀무가설 (null hypothesis, H_0)**은 우리가 증명하고자 하는 가설의 반대되는 가설을 의미하며 우리가 증명 또는 입증하고자 하는 가설을 **대립가설(alternative hypothesis, H_1)**이라고 한다.
- 유의성 검정(Null Hypothesis Significance Testing(NHST))은 두 개의 가설 중 어느 쪽이 참인지를 판단하기 위해 진행하는 검증 과정이다.
- 유의확률은 보통 p-value로 표현하며, **p-value가 0.05이하이면 이 귀무가설을 옳지 않은 것으로 본다**(‘귀무가설을 기각’한다고 표현함). p-value가 0.05 미만이라면, 이 통계치에서 귀무가설을 참으로 봤을 때 표본에서 실제로 해당되는 통계치가 나올 가능성은 5% 미만이라는 의미가 된다. 즉 해당 통계치는 95%의 확률로 대립가설이 참이 될 가능성이 훨씬 더 높은 것이라고 할 수 있다.

가설 검정

- 가설: 어떤 사실을 설명하거나 증명하기 위해서 설정한 가정을 의미한다.
- (예제)
 - 가설: A 유전자가 위암을 유발한다. (대립가설, H_1)
 - ✓ 귀무가설(H_0): A 유전자는 위암을 유발하지 않는다.
 - ✓ 'A 유전자는 위암을 유발하지 않는다.' 라는 귀무가설이 틀렸다는 사실을 증명하고자 함
 - ✓ 방법: 귀무가설의 유의성 검증을 통해서 p-value가 0.05이하이면 이 귀무가설이 옳지 않은 것으로 판단.
 - ✓ 즉, 'A 유전자가 위암을 유발한다.' 는 대립가설이 성립되는 것을 증명하는 것이다.

- **Residual**: 모델로 예측한 y값과 실제 데이터의 y값과의 차이를 의미한다. 패턴/추세가 보이면 안되며, residual plot으로 관측 가능하다.
- **Coefficient**: Estimate 컬럼: 절편과 각 x의 기울기 값이 출력된다. 또한 예측한 각 회귀계수에 대한 유의성을 나타낸다.
 - t값(t검정에 대한 통계량)은 독립변수(x)와 종속변수(y)간에 선형관계(관련성)가 존재하는 정도를 나타낸다. t 값은 회귀계수 나누기 표준오차(표준편차)가 된다. 유의미한 결과가 나오려면 t 값이 커야 한다 (절대값이 2보다 커야 한다).
(ex) $1.75/1.19365=9.037$ $0.4/0.02357=16.971$
 - **유의수준(significance level)**: p-value로 표기되며, 관찰된 데이터의 검정 통계량이 귀무가설(영가설)을 지지하는 정도를 확률로 표현한 것. 유의수준 0.05로 설정한다면, p-값이 유의수준보다 작으므로 **귀무가설** (독립변수 x는 종속변수 y는 아무 관련이 없다)은 기각된다. 즉 두 변수는 관련이 있다라는 **대립 가설**이 받아들여진다.

Residuals:

1	2	3	4
0.05	-0.15	0.15	-0.05

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.75000	0.19365	9.037	0.01202 *
x	0.40000	0.02357	16.971	0.00345 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> plot(x, y)
> abline(m, col = 'red')
```

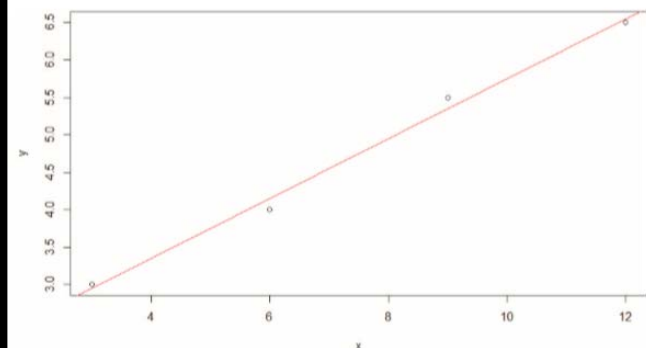


그림 7-4 lm 함수로 찾은 최적 모델

- **Multiple R-Squared**: 회귀모델의 설명력을 나타내는 결정계수로서 값이 1이면 실제 관측값들이 회귀선상에 정확히 일치함을 나타낸다. 만약 0.65이면 35%는 회귀식으로 설명할 수 없음을 의미한다. 단, 독립변수의 개수가 증가할수록 값이 증가하는 특징을 가지고 있다.
- **Adjusted R-Squared** : 독립 변수의 개수가 고려되어 보정된 R-Squared이다.
- **F-statistic**: 회귀식 전체의 유의성을 검정하는 값으로 가정 먼저 확인하여야 하는 값이다. “모든 회귀 계수가 0이다”라는 귀무가설(H_0)의 기각여부를 검증하는 것이다. 즉 해당 p-value가 작은 값이면 최소 한 변수가 유의하다는 것을 알 수 있다. 어떤 변수가 유의한지는 coefficient table을 보면 된다.

Multiple R-squared: 0.9931, Adjusted R-squared: 0.9897
F-statistic: 288 on 1 and 2 DF, p-value: 0.003454

02-4 회귀모형의 결과 분석 (다중선형회귀 분석 예)

```
Call:
lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-10.5932	-1.0690	0.2902	1.4272	3.3951

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.526667	0.374290	9.422	<2e-16 ***
youtube	0.045765	0.001395	32.809	<2e-16 ***
facebook	0.188530	0.008611	21.893	<2e-16 ***
newspaper	-0.001037	0.005871	-0.177	0.86

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956
F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16

```
Call:
```

```
lm(formula = sales ~ youtube + facebook, data = marketing)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-10.5572	-1.0502	0.2906	1.4049	3.3994

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.50532	0.35339	9.919	<2e-16 ***
youtube	0.04575	0.00139	32.909	<2e-16 ***
facebook	0.18799	0.00804	23.382	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.018 on 197 degrees of freedom
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8962
F-statistic: 859.6 on 2 and 197 DF, p-value: < 2.2e-16

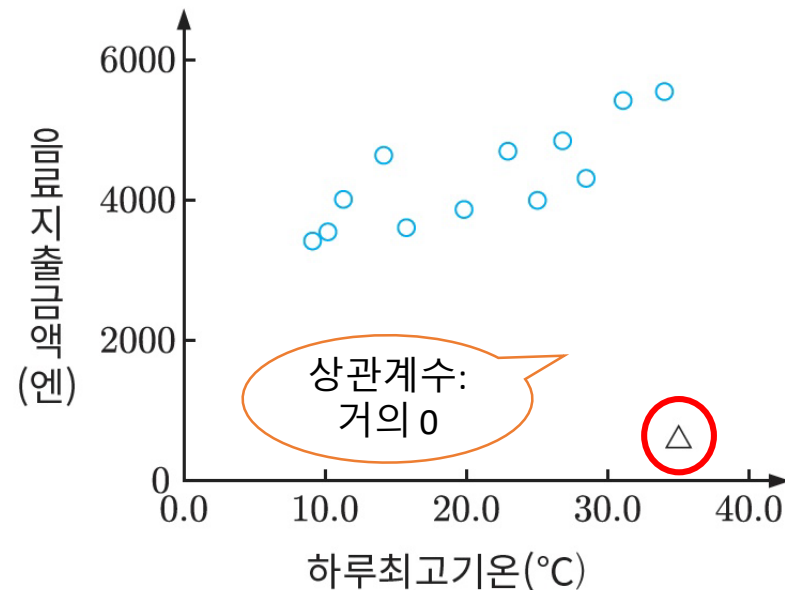
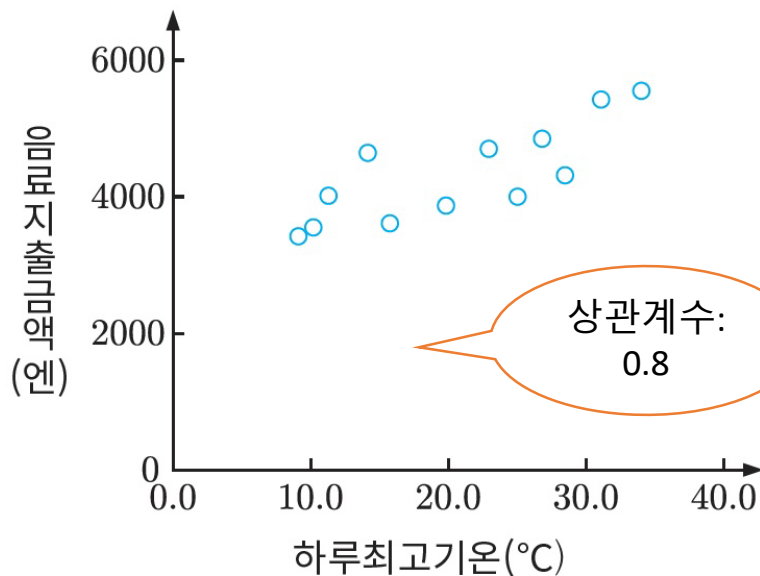
02-5 이상값의 영향

• 이상값 (outlier)

- 상관계수와 회귀분석의 결과에 커다란 영향을 줌
- 분석 대상에서 제외하는 경우가 많음
- 정말로 제외시켜도 좋은지 충분히 생각해 볼 필요가 있음

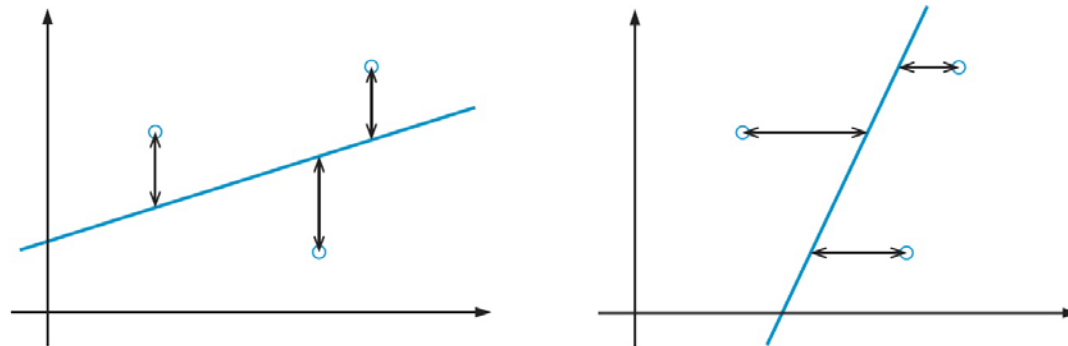
예: 대규모 지진 분석

- 이상값의 제외 여부는 데이터의 특성과 분석 목적에 따라 결정되어야 한다



02-6 역 회귀 분석

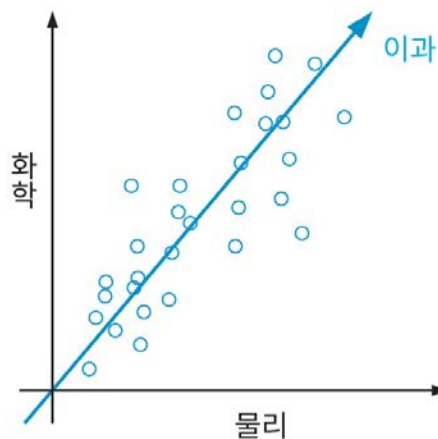
- 평균수명의 예에서 설명변수와 목적변수
 - 흡연율: 독립(설명)변수
 - 평균수명: 종속(반응)변수
 - 어느쪽을 설명변수로 하는가에 의해 회귀분석의 결과가 달라짐
- 최소제곱법
 - 설명변수가 (가로축) → 세로방향의 거리에 대한 제곱의 전체합을 최소화
 - 설명변수가 (세로축) → 가로 방향의 거리에 대한 제곱의 전체합을 최소화
- 분석의 목적에 맞추어 설명/반응 변수를 선택할 필요가 있다.



〈그림 3.4〉 역 회귀 분석

02-7 주성분 분석에 의한 설명변수의 합성

- 독립(설명)변수의 개수가 지나치게 많으면 변수들 사이의 관계가 복잡해진다.
- 주성분 분석(PCA: Principal Component Analysis)
 - 유사한 변수들을 모아 새로운 변수를 만들
 - 특징량: 기계학습에서 주성분 분석 및 기타 방법들을 사용하여 만든 데이터의 특징을 나타내는 유용한 변수
 - 빅데이터 분석에서 변수가 너무 많은 경우, 주성분분석을 통하여 변수의 수를 줄이는 (차원을 낮추는) 작업을 수행함.



〈그림 3.5〉 주성분 분석

03 로지스틱 회귀분석

- 로지스틱 회귀분석: 분석 대상들이 두 집단 혹은 그 이상의 집단으로 나누어진 경우에 개별 관측치들이 어느 집단에 분류될 수 있는가를 분석하고 예측하는 모델
- 선형회귀 분석과 로지스틱 회귀분석의 비교

	일반선형회귀분석	로지스틱 회귀분석
종속변수	연속형 변수	이산형 변수
모델 탐색 방법	최소제곱법	최대우도법(maximum likelihood method) 가중최소제곱법
모델 검정	F-test, t-test	χ^2 test

- 로지스틱 회귀분석 과정
 - 각 집합에 속하는 확률의 추정치를 예측. 이진 분류의 경우에는 집단 1에 속하는 확률 $P(Y=1)$ 의 추정치를 얻는다.
 - 확률값 \rightarrow 분류 기준값 (cut-off) 적용 \rightarrow 특정 집단으로 분류
예) $P(Y=1) \geq 0.5 \rightarrow$ 집단 1로 분류
 $P(Y=1) < 0.5 \rightarrow$ 집단 0으로 분류