수치 데이터 예측: 회귀 기법

이번 장에서 배울 내용

- 선형 회귀linear regression 기법의 통계 원리를 이용하여 데이터 간의 관계를 나타낸다.
- 회귀 분석, 선형 방정식, 회귀 모델 추정에 관한 R 사용법을 배운다.
- 회귀 트리와 모델 트리로 알려진 혼합hybred 모델을 어떻게 사용하고, 수치 예측에 관한 어떤 decision tree를 사용하는지 알아본다.

회귀란?

- 회귀(복귀, regression)은 평균으로 돌아간다라는 말에서 유래했다.
- 회귀는 하나의 종속 변수 y와 독립 변수 x(하나 이상)사이의 관계를 명시하는 것을 의미한다.
- 즉 $y = \alpha + \beta x$ (단순선형회귀)에서 가장 좋은 α 와 β 를 찾는 것이 학습(training)이다.

회귀분석 사용 예

- 경제, 사회학, 심리학, 물리학, 생태학 같은 분야에서 측정된 특성으로, 모집단과 개별이 어떻게 다른지 평가.
- 임상 약품 실험, 엔지니어링 안전 검사, 시장 연구 같은 사건과 결과 간의 인과 관계를 수량화.
- 보험 청구, 자연 재해 손해, 선거 결과, 범죄율 예측 처럼, 주어진 기준으로 미래의 행위를 예측하기 위해 사용됨.

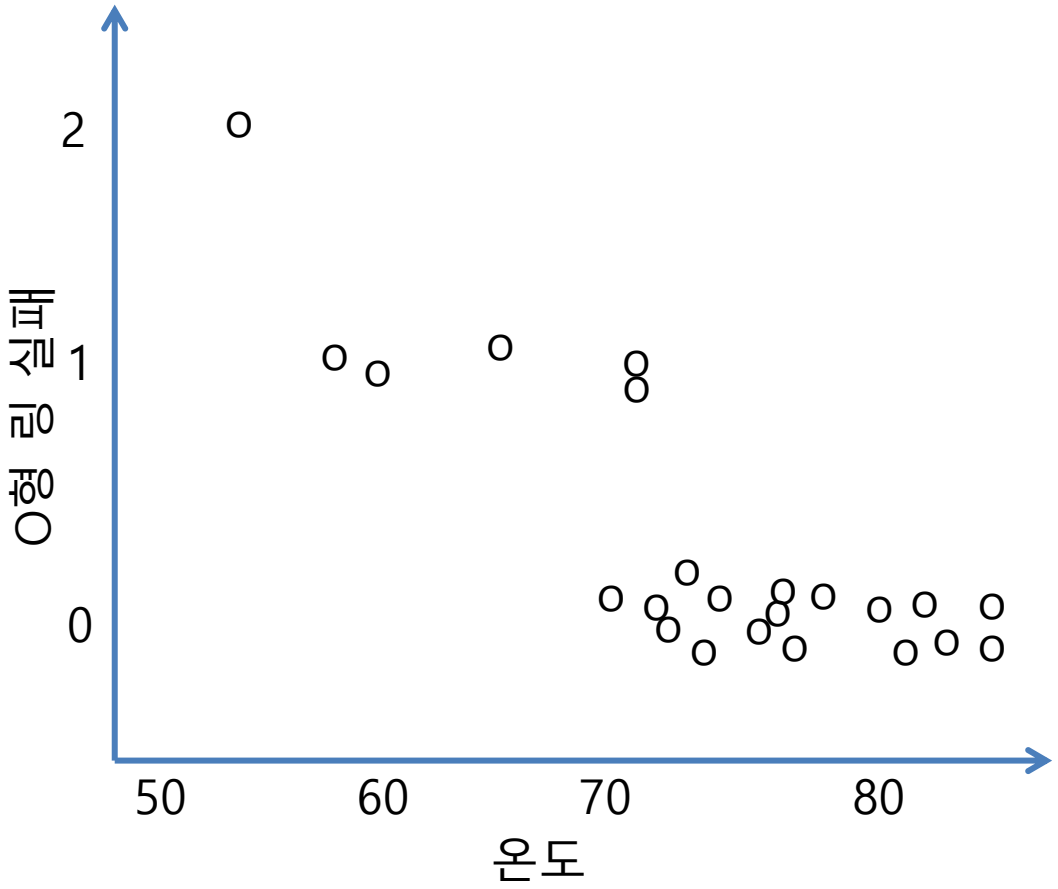
선형 회귀linear regression

- 단순 선형 회귀simple linear regression: 하나의 독립변수
- 다중 회귀multiple regression: 여러 개의 독립변수
- 이 두 모델의 독립변수는 연속적continuous 이다.
- 로지스틱 회귀: 이진 범주형 결과를 모델화하는데 사용됨
- 푸아송 회귀Poission regression: generalized linear model의 일종(log-linear model)

온도와 O형 링의 파손사이 관계

- 1986년 1월 28일, 추진 로켓과 연결하는 O형 링이 파손됐고 돌발 폭발이 원인이 돼 미국의 챌린저호가 폭발하였다.
- 그 원인의 주범으로 낮은 온도(그 당시 화씨 31도)를 꼽았다. 따라서 온도와 O형 링의 파손 여부를 밝히기 위하여 23개의 O형 링을 조사하였다.

산포도



$y = \alpha + \beta x$ 풀기

- 만약에

α=4.30

β=-0.057

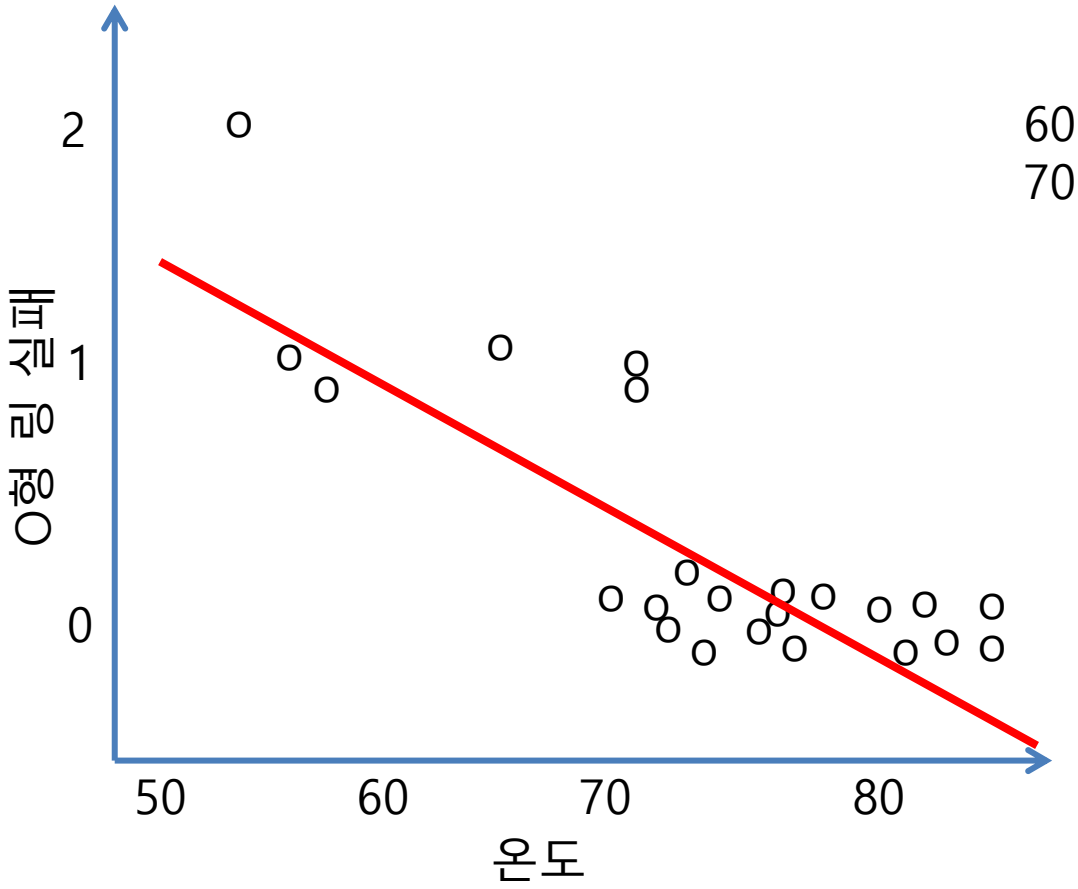
이라고 하자(왜 그런지는 좀 있다가 설명한다)

- 그러면 직선의 방정식은

$y = 4.3 - 0.057x$

이 되고 이를 산포도에 그려 본다.

산포도



60일 경우 실패 개수 0.8
70인 경우 실패 개수 0.3

얼마나 위험했는지 분석

- 날씨 값 $x = 31.5$를 대입하면

- $y = 4.3 - 0.057 * 31 = 2.50$

- 즉 2.5개의 O형 링이 파손됐을 것이다.
- 60도인 경우보다 2.5/0.8=3배 정도
- 70도인 경우보다 2.5/0.3=8배 정도
  의 위험도로 O형 링이 판손됨을 알 수 있다.

Ordinary least squares (정규 최소 제곱)

- 아마도 least squares 방법을 들었을 것이다.
- 즉 $y = \alpha + \beta x$라는 식으로 시작하자.
- 각 각의 x_i 값을 대입하면 추정치 $(y_i)'$ 값이 나오는데 이 값을 그래프에 그려진 실제 값 y_i 와의 차이를 제곱한 후 더하여 최소의 값이 나오도록 하는 α, β를 선택한도록 한다.

중요한 parameters α, β 구하기

- Exercise:

  Show that $\beta = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

  (단 $\bar{x}(\bar{y})$는 $x(y)$들의 평균)

Answer to the previous exercise

α, β에 관하여 각각 미분하고 0이라고 둔다

$$f = \Sigma u_i^2 = \Sigma(y_i - \alpha - \beta x_i)^2$$
$$\frac{\partial f}{\partial \alpha} = -2\Sigma(y_i - \alpha - \beta x_i) = 0$$
$$\frac{\partial f}{\partial \beta} = -2\Sigma(y_i - \alpha - \beta x_i)x_i = 0$$

Call the solutions to these equations $\hat{\alpha}$ and $\hat{\beta}$. Solving we get:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$
$$\hat{\beta} = \frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2}$$

Where $\bar{y} = \frac{\Sigma y_i}{n}$ and $\bar{x} = \frac{\Sigma x_i}{n}$. Computing these results can be left as an exercise.

summarize

- $\alpha = \bar{y} - \beta\bar{x}$
- $\beta = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$

challenger.csv

o_ring_ct	distress_ct	temperature	pressure	launch_id
6	0	66	50	1
6	1	70	50	2
6	0	69	50	3
6	0	68	50	4
6	0	67	50	5
6	0	72	50	6
6	0	73	100	7
6	0	70	100	8
6	1	57	200	9
6	1	63	200	10
6	1	70	200	11
6	0	78	200	12
6	0	67	200	13
6	2	53	200	14
6	0	67	200	15
6	0	75	200	16
6	0	70	200	17
6	0	81	200	18
6	0	76	200	19
6	0	79	200	20
6	0	75	200	21
6	0	76	200	22
6	1	58	200	23

Pearson's correlation coefficient (피어슨의 상관 관계)

- 피어슨의 상관 관계

$$\rho_{x,y} = Corr(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

- $\rho_{x,y}$은 -1에서 1의 값을 갖는다. -1에 가까우면 강한 역상관 관계 1에 가까우면 강한 상관 관계를 갖는다.

강한 역상관 관계

#Pearson's correlation coefficient 상관관계

```
r <- cov(launch$temperature, launch$distress_ct)/
       (sd(launch$temperature)* sd(launch$distress_ct))
r

cor(launch$temperature, launch$distress_ct)
```

Output:

```
>r
 [1] -0.725671
 > cor(launch$temperature, launch$distress_ct)
[1] -0.725671
```

다중 선형 회귀

- 실제 회귀 문제는 하나 이상의 독립 변수를 사용한다.
- 따라서 다중 선형 회귀를 사용한다.

다중 선형 회귀 장단점

장점	단점
• 수치 데이터를 모델화하기 위한 가장 일반적인 접근법	• 데이터에 대한 가장 강한 가정을 만듬
• 거의 모든 데이터를 모델화 함	• 사전에 모델의 형태가 사용자로부터 명시 되어야 함
• 속성과 결과 간 관계의 견고성과 크기를 추정할 수 있음	• 수치 속성만 작동하고 범주형 데이터는 부가 처리가 필요함
	• 모델을 이해하려면 통계적 일부 지식이 필요함

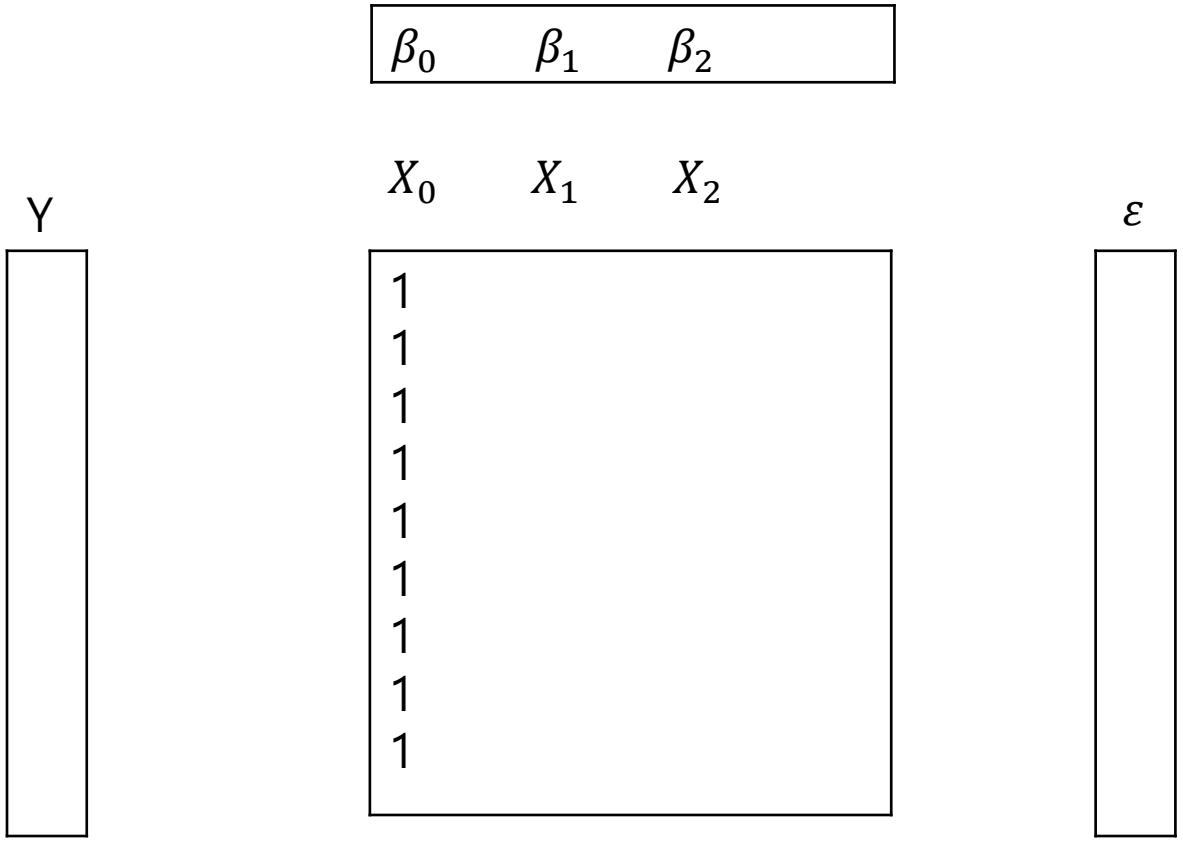
다중 선형 회귀 공식

- $y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$

- 간단히

$$Y = X\beta + \varepsilon$$

$$Y = X\beta + \varepsilon$$

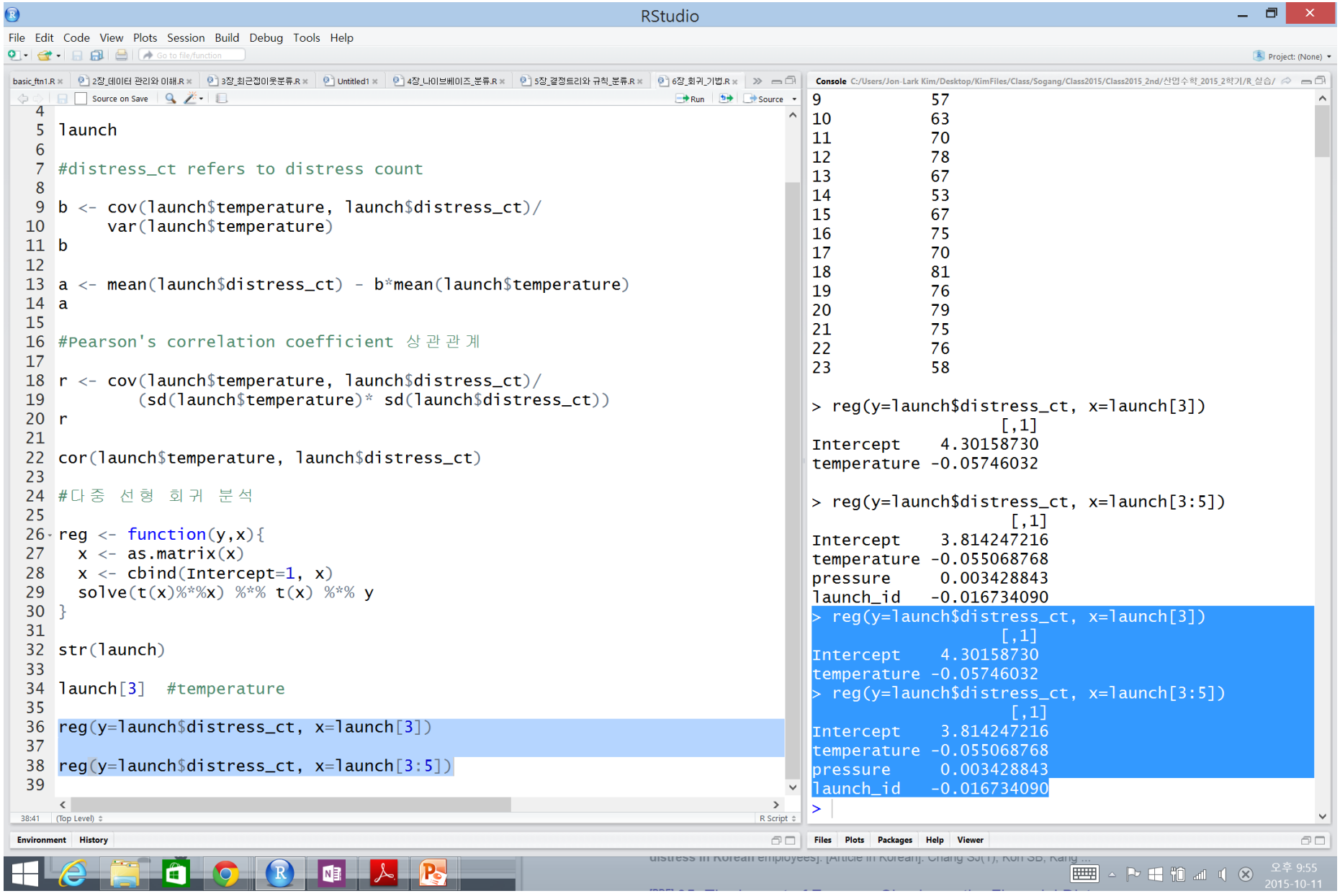


$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T Y}$$

#다중 선형 회귀 분석

```
reg <- function(y,x){
  x <- as.matrix(x)
  x <- cbind(Intercept=1, x)
  solve(t(x)%*%x) %*% t(x) %*% y
}

str(launch)

launch[3]   #temperature

reg(y=launch$distress_ct, x=launch[3])

reg(y=launch$distress_ct, x=launch[3:5])
```

다중 회귀 분석 결과(온도만 사용할 때, 온도, 압력, ID사용할 때)



```
> reg(y=launch$distress_ct, x=launch[3])
                          [,1]
Intercept          4.30158730
temperature    -0.05746032
```

즉 **y=4.3-0.057x**

```
>reg(y=launch$distress_ct, x=launch[3:5])
[,1]
Intercept          3.814247216
temperature    -0.055068768
pressure           0.003428843
launch_id         -0.016734090
```

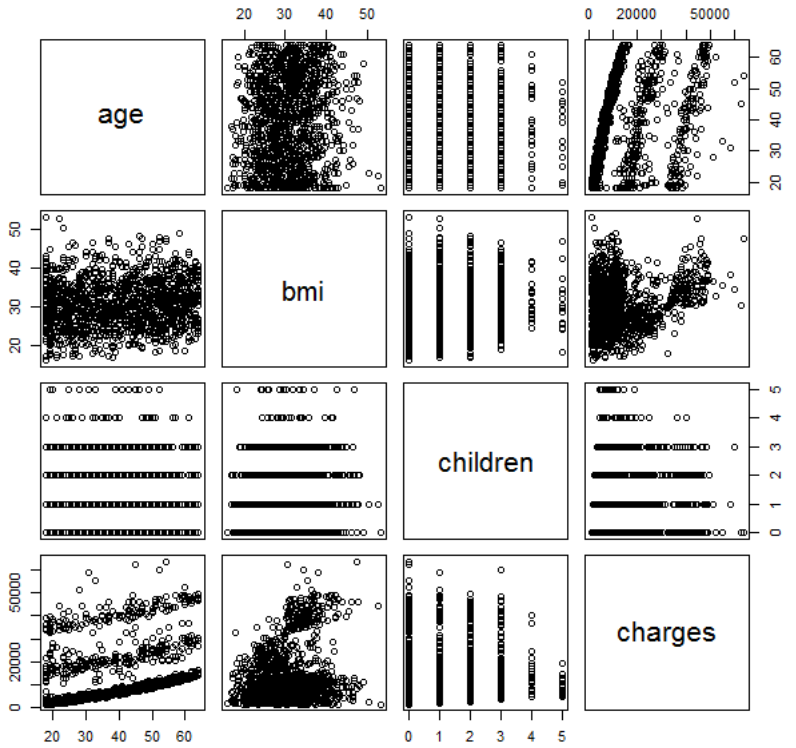
즉 **y=3.8 - 0.055 x_1 + 0.0034 x_2 – 0.0167 x_3**

질문: 왜 id가 증가하면 파손의 정도가 감소할까?

예제: 선형 회귀를 사용한 의료비 예측

```
insurance <- read.csv("insurance.csv", stringsAsFactors = F)
str(insurance)
 #bmi: Body mass index = weight in kg /(height in m)^2
 # insurance.csv은 사이버 캠퍼스에 있음

summary(insurance$charges)

hist(insurance$charges)

table(insurance$region) #northeast northwest southeast southwest

cor(insurance[c("age", "bmi", "children", "charges")])

pairs(insurance[c("age", "bmi", "children", "charges")])
```

pairs(insurance[c("age", "bmi", "children", "charges")]) 상관관계



```
install.packages("psych")
library(psych)
pairs.panels(insurance[c("age", "bmi", "children", "charges")])
```