Data Mining with R

3. 회귀 분석 (Regression)

Contents

- 회귀 분석의 개념
- Linear Regression
- 기타 Regression

회귀 분석 (regression)

- 회귀 분석
  - 독립변인 (input)이 종속변인 (target)에 영향을 미치는지 알아보고자 할 때 실시하는 분석방법
  - 연속형 자료에 따른 연속형 자료의 영향력을 검증할 때 사용

영향을 주는 변수 input	영향을 받는 변수 target	통계분석 방법
범주형 자료	범주형 자료	카이제곱 검정
	연속형 자료	T검정 분산분석
연속형 자료	연속형 자료	회귀분석 구조방정식
	범주형 자료	로지스틱회귀분석

회귀 분석 예시

- 데이터
  - 연속형 변수 : 커피 맛, 가게 인테리어, 직원 친절도 (7점 척도)
  - 목표 : 고객 만족도에 영향을 미치는 변수 파악

  - 독립변수 (input) : 연속형 자료 – 커피의 맛, 가게 인테리어, 직원 친절도
  - 종속변수 (target) : 연속형 자료 – 만족도

회귀 분석의 종료

- 단순 회귀 분석
  - 영향을 주는 변수가 1개
- 다중 회귀 분석
  - 영향을 주는 변수가 2개 이상
- 분석 도구에서는 큰 차이 없음

회귀 분석의 결과 분석

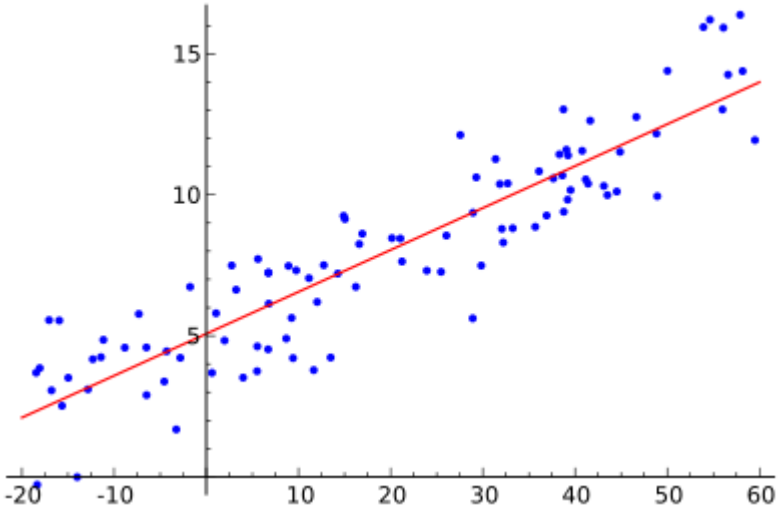
- R제곱 (R-Squared)
  - 식의 설명력
  - 독립 변수가 종속 변수를 얼마나 설명하느냐를 판단
- F-Statistics
  - 모형 적합도
  - p-Value가 0.05보다 작으면 이 모형이 적합함

Contents

- 회귀 분석의 개념
- Linear Regression
- 기타 Regression

Linear Regression

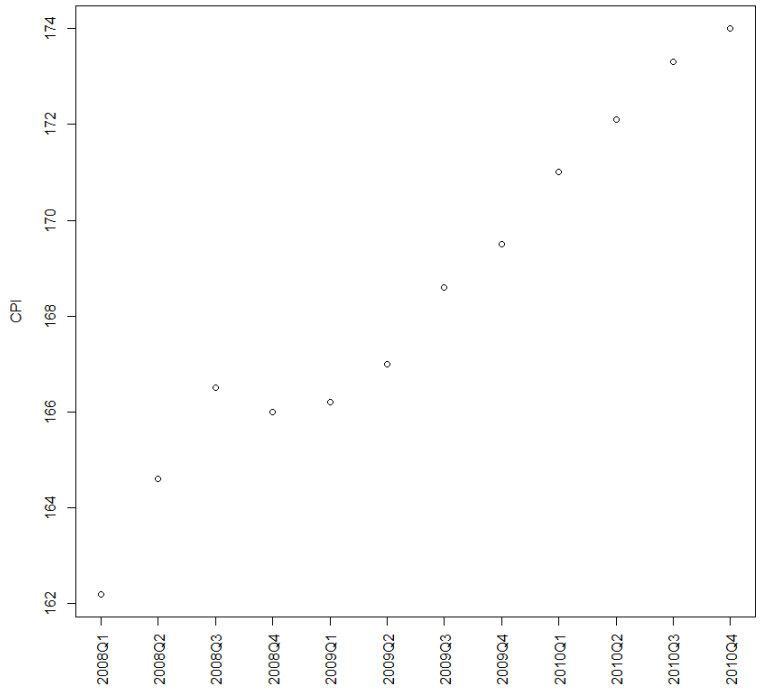
- Linear Regression은 다음과 같은 선형 방정식을 이용하여 목표 값을 예측하는 것이다.
- $y = c_0 + c_1 x_1 + c_2 x_2 + \cdots + c_k x_k$
- Target(y)와 Input($x_1, x_2, \ldots, x_k$)을 이용하여 $c_0, c_1, c_2, \ldots, c_k$을 찾아내는 것이 목표



The CPI Data

- Australian CPI (Consumer Price Index) data

```
> year <- rep(2008:2010, each = 4)
> quarter <- rep(1:4, 3)
> cpi <- c(162.2,164.6,166.5,166,166.2,167,168.6,169.5,171,172.1,173.3,174)
> plot(cpi, xaxt = "n", ylab = "CPI", xlab = "")
> axis(1, labels = paste(year, quarter, sep = "Q"), at = 1:12, las = 3)
```



Train by Linear Regression

```
> cor(year, cpi)
[1] 0.9096315548
> cor(quarter, cpi)
[1] 0.3738027922
> fit <- lm(cpi ~ year + quarter)
> fit

Call:
lm(formula = cpi ~ year + quarter)

Coefficients:
  (Intercept)          year       quarter
-7644.487500      3.887500      1.166667
```

$$cpi = -7644.487500 + 3.887500 * year + 1.166667 * quarter$$

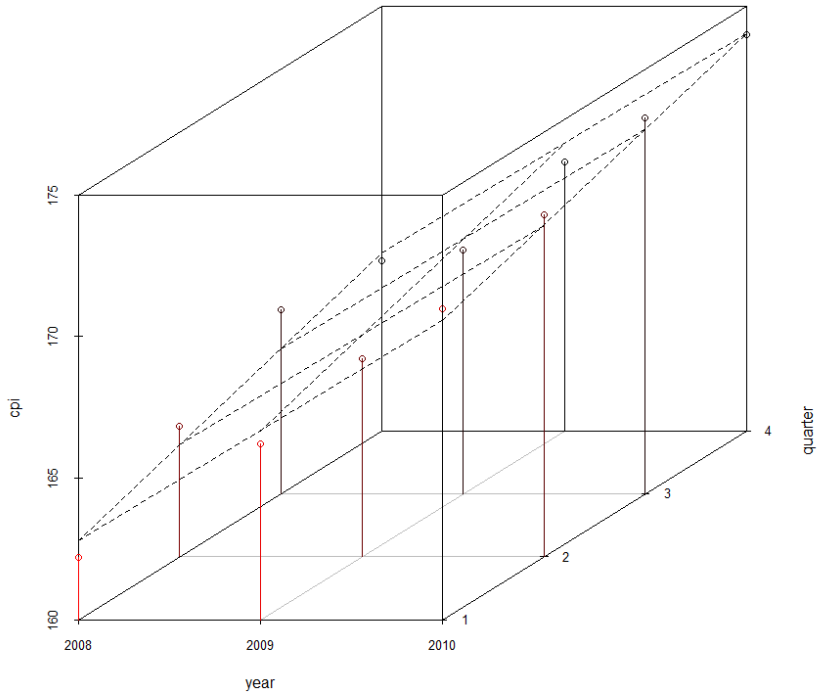
```
> cpi2011 <- fit$coefficients[[1]]+fit$coefficients[[2]]*2011 + fit$coefficients[[3]]*(1:4)
> cpi2011
[1] 174.4416667 175.6083333 176.7750000 177.9416667
```

Train by Linear Regression

```
> attributes(fit)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values"
 [6] "assign"        "qr"            "df.residual"   "xlevels"       "call"
[11] "terms"         "model"

$class
[1] "lm"

> fit$coefficients
    (Intercept)            year         quarter
-7644.487500000     3.887500000     1.166666667
> residuals(fit)
             1              2              3              4              5
-0.57916666667  0.65416666667  1.38750000000 -0.27916666667 -0.46666666667
             6              7              8              9             10
-0.83333333333 -0.40000000000 -0.66666666667  0.44583333333  0.37916666667
            11             12
 0.41250000000 -0.05416666667
>
> summary(fit)

Call:
lm(formula = cpi ~ year + quarter)

Residuals:
       Min         1Q     Median         3Q        Max
-0.8333333 -0.4947917 -0.1666667  0.4208333  1.3875000

Coefficients:
                Estimate    Std. Error   t value      Pr(>|t|)
(Intercept) -7644.4875000   518.6542802 -14.73908 0.00000013137 ***
year            3.8875000     0.2581653  15.05818 0.00000010908 ***
quarter         1.1666667     0.1885373   6.18799    0.00016117 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7302016 on 9 degrees of freedom
Multiple R-squared:  0.9671581, Adjusted R-squared:  0.9598599
F-statistic: 132.5201 on 2 and 9 DF,  p-value: 0.0000002108277
```

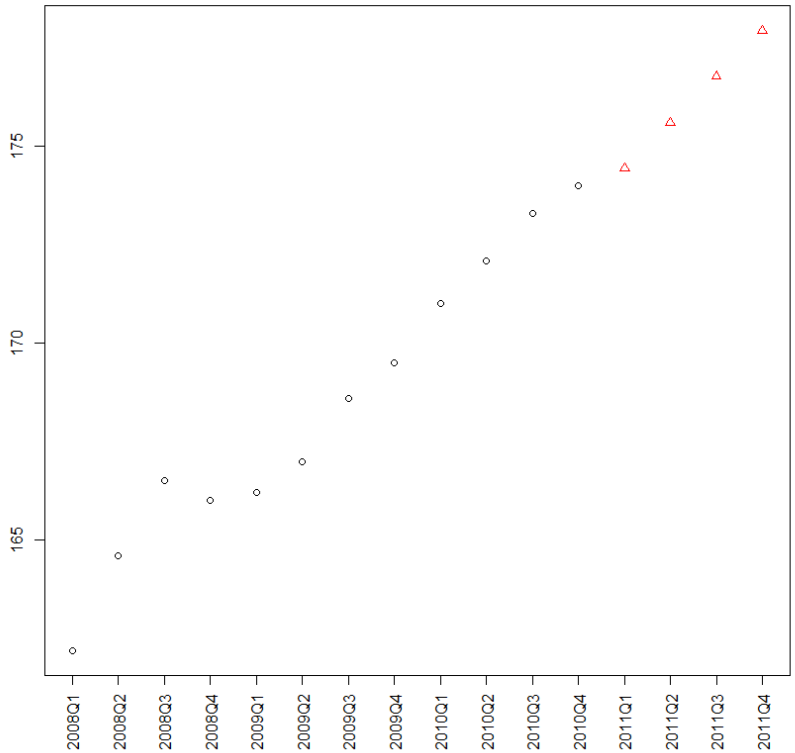
3D Plot of the Fitted Mode

```
> library(scatterplot3d)
> s3d <- scatterplot3d(year, quarter, cpi, cpi, highlight.3d = T, type = 'h', lab = c(2,3))
Warning message:
In scatterplot3d(year, quarter, cpi, cpi, highlight.3d = T, type = "h",  :
  color is ignored when highlight.3d = TRUE
> s3d$plane3d(fit)
```



Prediction of CPIs in 2011

```
> data2011 <- data.frame(year = 2011, quarter = 1:4)
> cpi2011 <- predict(fit, newdata = data2011)
> style <- c(rep(1,12), rep(2,4))
> plot(c(cpi, cpi2011), xaxt = 'n', ylab = 'CPI', xlab = '', pch = style, col = style)
> axis(1, at = 1:16, las = 3, labels = c(paste(year, quarter, sep='Q'), '2011Q1', '2011Q2',
 '2011Q3','2011Q4'))
```



Contents

- 회귀 분석의 개념
- Linear Regression
- 기타 Regression

Generalized Linear Model (GLM)

- GLM
  - Linear Regression: 종속변수(target)의 정규 분포와 분산의 동등성 가정
  - GLM: 자료의 독립성 가정
  - 광범위한 비정규분포 자료의 사용 허용
  - Input과 target의 관계가 선형 및 비선형인 경우에도 연결 함수 (link function)을 이용하여 모형의 선형성 충족

Build a Generalized Linear Model

```
> data('bodyfat', package='TH.data')
> myFormula <- DEXfat ~ age + waistcirc + hipcirc + elbowbreadth + kneebreadth
> bodyfat.glm <- glm(myFormula, family = gaussian('log'), data = bodyfat)
> summary(bodyfat.glm)

Call:
glm(formula = myFormula, family = gaussian("log"), data = bodyfat)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-11.5688297  -3.0064808   0.1265767   2.8310259  10.0966155

Coefficients:
              Estimate  Std. Error t value    Pr(>|t|)
(Intercept)  0.734292854 0.308948998 2.37674   0.0204204 *
age          0.002129249 0.001445584 1.47293   0.1455952
waistcirc    0.010488836 0.002478841 4.23135 0.000074391 ***
hipcirc      0.009702132 0.003231061 3.00277   0.0037944 **
elbowbreadth 0.002354959 0.045685757 0.05155   0.9590478
kneebreadth  0.063188120 0.028193167 2.24126   0.0284310 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 20.31432539)

    Null deviance: 8535.9838  on 70  degrees of freedom
Residual deviance: 1320.4319  on 65  degrees of freedom
AIC: 423.0247

Number of Fisher Scoring iterations: 5
```

Prediction with GLM

```
> pred <- predict(bodyfat.glm, type = 'response')
> plot(bodyfat$DEXfat, pred, xlab = 'Observed', ylab = 'Prediction')
> abline(a = 0, b = 1, col = 'red', lwd = 2)
```