R을 이용한 기초 통계 분석: 상관분석/회귀분석

Jinseog Kim
Dongguk University
jinseog.kim@gmail.com

2017-04-12

Agenda

- **상관분석 (corelation analysis)**
- **회귀분석(regression analysis)**
  - 선형회귀분석 (linear regression analysis)
  - 로지스틱 회귀분석 (logistic regression analysis)

상관분석 (corelation analysis)

1. 공분산 : 두 연속형 변수간의 산포의 정도를 측정

$$Cov(X, Y) = E\Big[(X - E[X])(Y - E[Y])\Big] = E[XY] - E[X]E[Y]$$

2. 상관계수 (correlation coefficient) : 연속형 변수간의 선형적 관계의 정도
   - population

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

   - estimate

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

   - property & interpretation
     - $-1 \leq \rho \leq 1$
     - 절대값이 1에 가까울수록 강한 직선관계
     - 부호에 따라 양(음)의 상관관계

상관분석 (conti)

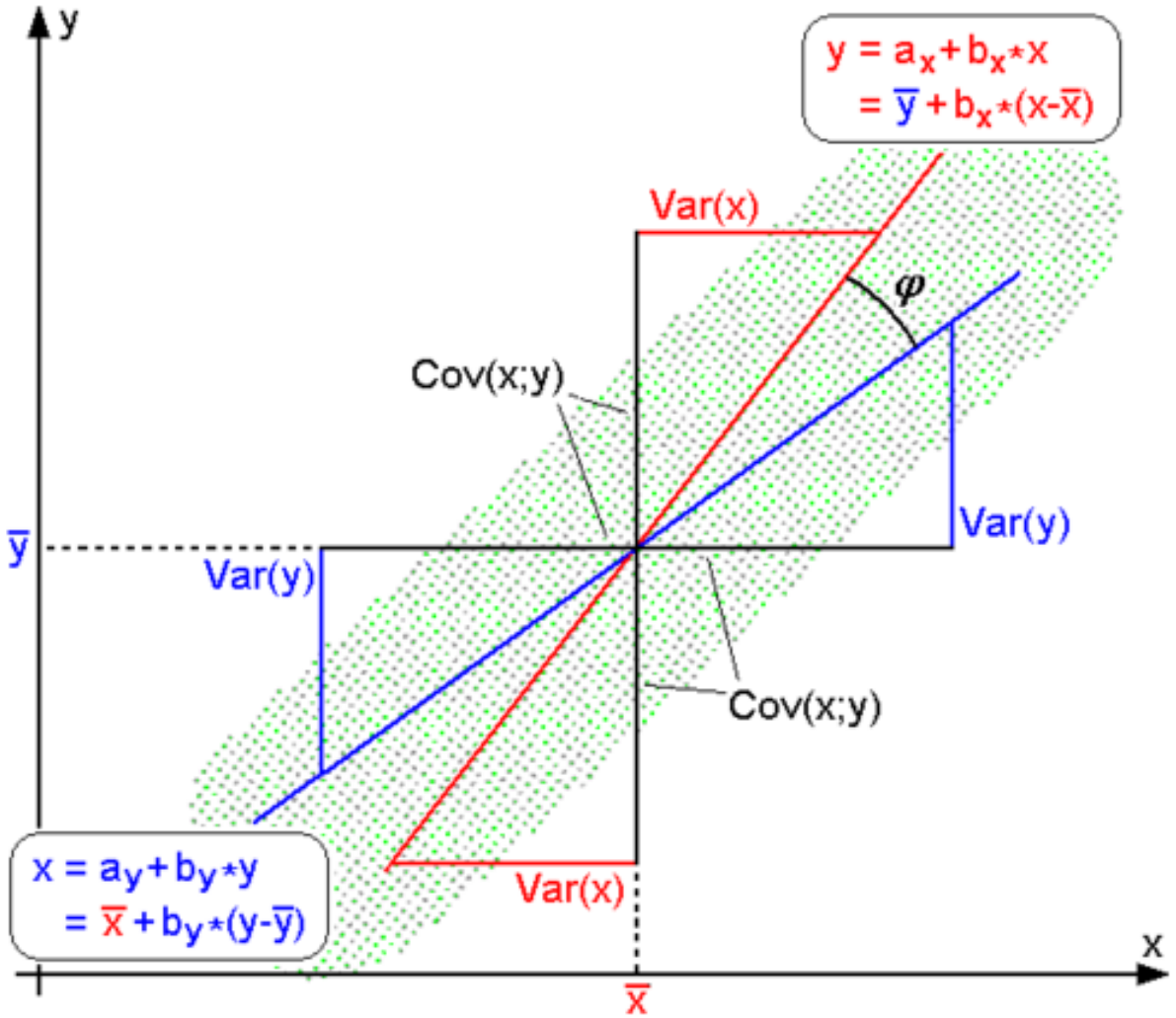


Figure: 기하학적 의미

상관분석 (conti)

- Test for correlation: $H_0 : \rho = 0$
- Test statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2) \text{ under } H_0 \text{ is true}$$

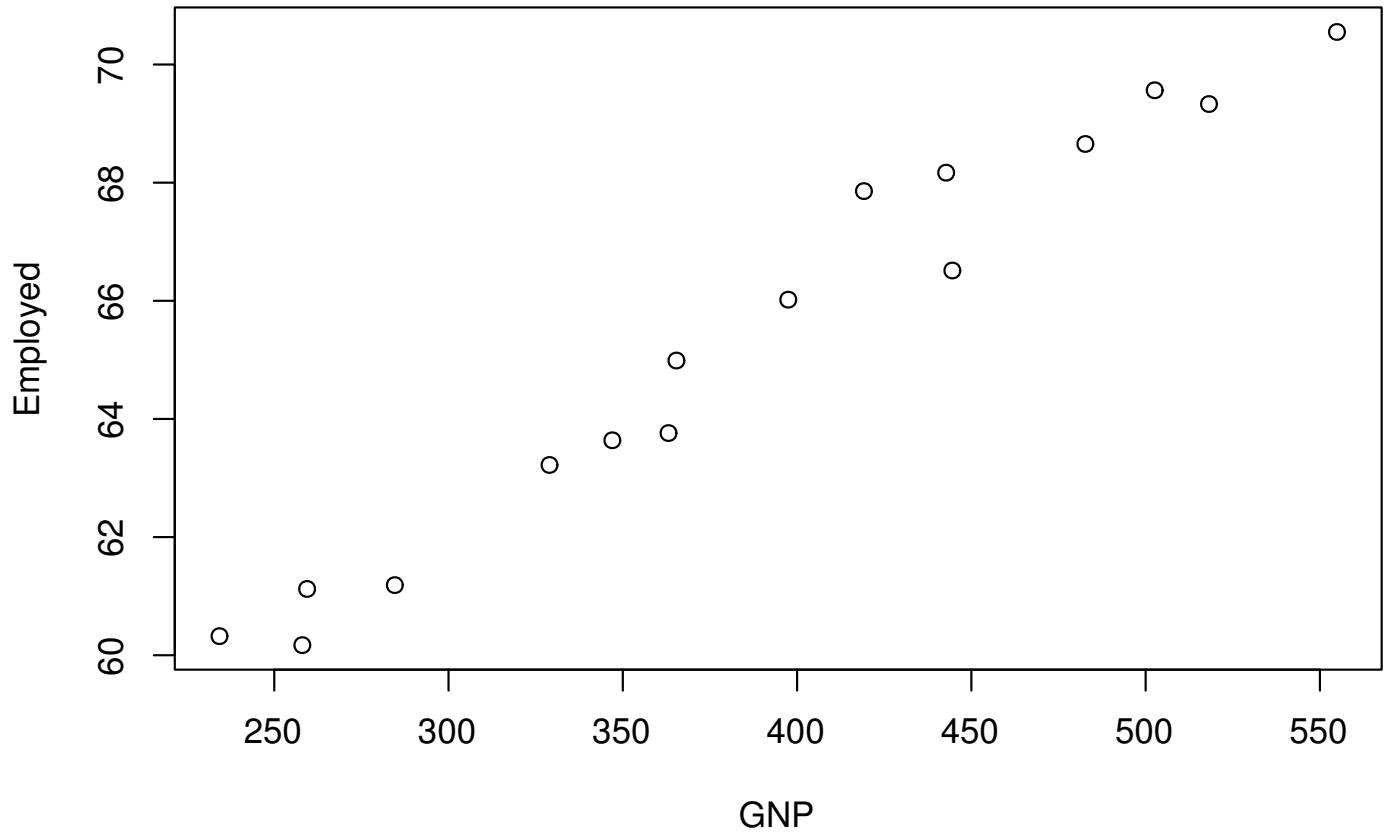
예제: 상관분석

- longley data : 7가지 경제지표 (1947~1962)
- 연도별 GNP와 취업률의 관계

```
x <- longley[, c("GNP", "Employed")]
x
```

```
##          GNP Employed
## 1947 234.289   60.323
## 1948 259.426   61.122
## 1949 258.054   60.171
## 1950 284.599   61.187
## 1951 328.975   63.221
## 1952 346.999   63.639
## 1953 365.385   64.989
## 1954 363.112   63.761
## 1955 397.469   66.019
## 1956 419.180   67.857
## 1957 442.769   68.169
## 1958 444.546   66.513
## 1959 482.704   68.655
## 1960 502.601   69.564
## 1961 518.173   69.331
```

예제: 상관분석 (conti)



예제: 상관분석 (conti)

- covariance matrix

```
cov(x)
```

```
##                GNP  Employed
## GNP       9879.3537 343.33021
## Employed   343.3302  12.33392
```

- correlation and test

```
cor.test(x[,1], x[,2])
```

```
##
##  Pearson's product-moment correlation
##
## data:  x[, 1] and x[, 2]
## t = 20.374, df = 14, p-value = 8.364e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9519914 0.9944238
## sample estimates:
##       cor
```

선형회귀분석 (linear regression analysis)

- 확률변수간의 함수관계를 추정하는 통계분석 방법

1. 설명변수 (독립변수):
2. 반응변수 (종속변수):
3. 회귀계수

회귀모형

- $x_1, \ldots, x_p$: 설명변수, y: 반응변수

$$y = f(x_1, \ldots, x_p) + \epsilon = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon,$$

  where $\epsilon \sim N(0, \sigma^2)$.
- Let $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ and $x = (x_1, \ldots, x_p)^T$

$$y = f(x) = \beta_0 + \beta^T x.$$

통계적 추론

- 회귀계수(β_j)의 추정: 최소제곱법
  - 아래의 식을 최소화

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}))^2$$

  - 결과 : $\hat{\beta} = (X^T X)^{-1} X^T y$
- 회귀계수에 대한 t 검정: ($H_0 : \beta_j = 0$)

$$t = \frac{\hat{\beta}_j}{s.e(\beta_j)} \sim t(n-p)$$

통계적 추론

- 적합도 평가(goodness of fit):
  - 결정계수(Coefficient of determination)

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

  - 수정결정계수

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

  - F 검정

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

Variable selection (변수선택법)

- 다중회귀모형에서는 (의미가 있던 또는 없던지) 설명변수가 많이 포함될 수록 R^2 이 커짐 (과적합)
- 변수선택법 : 유의한 설명 변수를 찾는 방법
  - All possible subset regression
  - 전진선택법 (Forward selection)
  - 후진소거법 (Backward elimination)
  - 단계적선택법 (Stepwise selection)
- 변수선택의 판정기준 (Selection Criterion)
  - F-test
  - Akaike Information Ceriterion (AIC)
  - Bayesian Information Ceriterion (BIC)

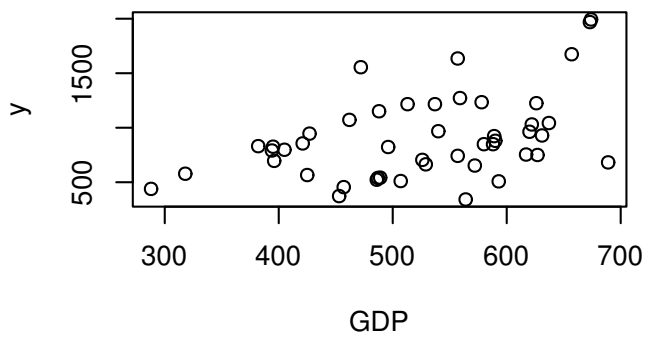
예제: 회귀분석 (UScrime)

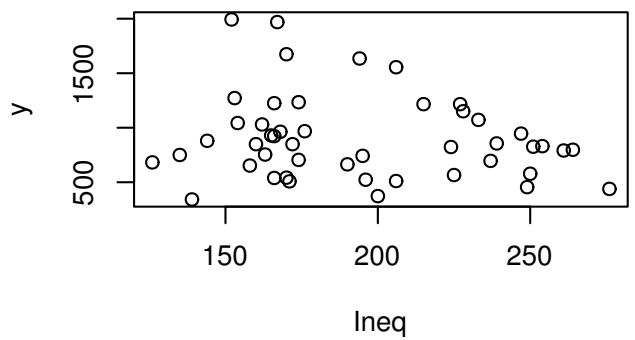
- objective: 처벌정책이 범죄율에 미치는 영향 연구
- data: 1960년 미국 47개주의 데이터
- 변수소개: 설명변수(15), 반응변수(y)

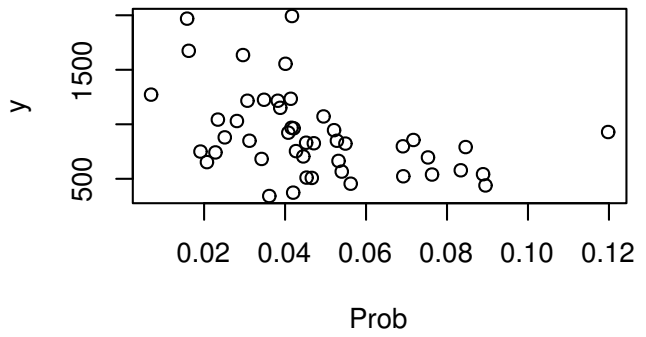
변수명	변수설명
M	percentage of males aged 14-24
So	indicator variable for a southern state
Ed	mean years of schooling
Po1	police expenditure in 1960
Po2	police expenditure in 1959
LF	labour force participation rate
M.F	number of males per 1000 females
Pop	state population
NW	number of nonwhites per 1000 people
U1	unemployment rate of urban males 14-24
U2	unemployment rate of urban males 35-39
GDP	gross domestic product per head
Ineq	income inequality
Prob	probability of imprisonment
Time	average time served in state prisons
y	rate of crimes in a particular category per head of population

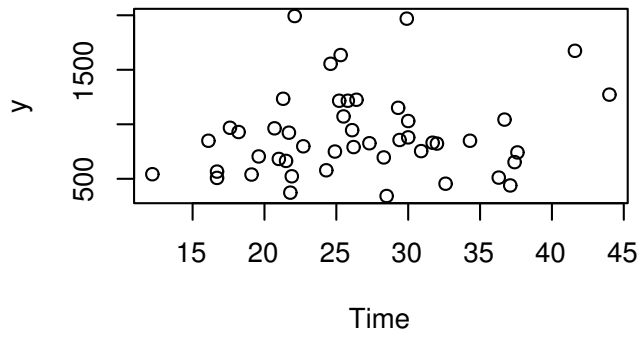
예제(UScrime): scatter plot

```
library(MASS)
par(mfrow=c(2,2))
for(i in 12:15)
plot(UScrime[, c(i,16)])
```









예제(UScrime)

- 상관계수

```
x <- UScrime[,12:16]
cor(x)
```

```
##                 GDP        Ineq        Prob          Time
## GDP    1.0000000000 -0.8839973 -0.5553347  0.0006485587
## Ineq  -0.8839972758  1.0000000  0.4653219  0.1018228182
## Prob  -0.5553347075  0.4653219  1.0000000 -0.4362462614
## Time   0.0006485587  0.1018228 -0.4362463  1.0000000000
## y      0.4413199490 -0.1790237 -0.4274222  0.1498660617
##                  y
## GDP    0.4413199
## Ineq  -0.1790237
## Prob  -0.4274222
## Time   0.1498661
## y      1.0000000
```

예제(UScrime)

- AVOVA table

```
m2 <- lm(y~GDP+Ineq+Prob+Time, data=x)
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## GDP        1 1340152 1340152 14.5517 0.0004409 ***
## Ineq       1 1403081 1403081 15.2350 0.0003377 ***
## Prob       1  248260  248260  2.6957 0.1080897
## Time       1   21396   21396  0.2323 0.6323064
## Residuals 42 3868038   92096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

예제(UScrime)

- 회귀계수의 추정 및 검정 / 모형의 적합도

```
summary(m2)
```

```
##
## Call:
## lm(formula = y ~ GDP + Ineq + Prob + Time, data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -587.30 -175.27  -11.71  116.08  757.78
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3008.562   1028.933  -2.924 0.005549 **
## GDP             4.590      1.058   4.338 8.84e-05 ***
## Ineq            9.362      2.463   3.801 0.000459 ***
## Prob        -4596.040   2783.448  -1.651 0.106156
## Time           -3.661      7.595  -0.482 0.632306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

단계적선택법 (AIC)

```
s <- step(m2, trace=0)
coef(summary(s))
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -3087.688952 1006.644979 -3.067307 3.728496e-03
## GDP             4.587014    1.048568  4.374553 7.611052e-05
## Ineq            9.104082    2.382697  3.820915 4.230197e-04
## Prob        -3893.644780 2350.238295 -1.656702 1.048598e-01
```

Model comparisons

- initial model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3008.562230	1028.933034	-2.9239631	0.0055488
GDP	4.589500	1.058067	4.3376256	0.0000884
Ineq	9.361878	2.463027	3.8009650	0.0004594
Prob	-4596.040204	2783.448432	-1.6512036	0.1061559
Time	-3.660756	7.594888	-0.4820026	0.6323064

- final model (stepwise selection)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3087.688952	1006.644979	-3.067307	0.0037285
GDP	4.587014	1.048568	4.374553	0.0000761
Ineq	9.104082	2.382697	3.820915	0.0004230
Prob	-3893.644780	2350.238295	-1.656702	0.1048598

Model comparisons (conti.)

Measures	initial	final
Multiple R-squared	0.4379	0.4348
Adjusted R-squared	0.3843	0.3953
F-statistic	8.179	11.02
DF	(4, 42)	(3, 43)
p-value	5.69e-05	1.7e-05

범주형 변수가 포함된 경우의 회귀분석

- 가변수(dummy variable)의 이용
- 범주의 수가 K개인 범주형 변수: $(K-1)$개의 가변수(dummy variable) $z_1, \ldots, z_{K-1}$로 코딩
  - $K = 3$인 경우

범주	z_1	z_2
1	1	0
2	0	1
3	0	0

- 가변수를 이용한 선형모형

$$y = \beta_0 + \sum_{k=1}^{K-1} \beta_k z_k + \epsilon$$

Advanced Regression methods

- penalized regression :
  - lasso, glmnet, elastic net, . . . : 변수선택과 계수추정을 동시에 하는 방법들
- non-linear or non-parametric models
  - random forest, svm, deep learning (multi-layer neural network)

Logistic regression (로지스틱회귀)

- Regression model : 반응변수가 연속형인 경우

$$E(y|x_1, x_2, \ldots, x_p) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

- Logistic regression model : 반응변수가 이진형인 경우 ($y \in \{0, 1\}$)

$$\log\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

$$P(Y=1|x) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}$$

로지스틱 회귀에서 오즈비 (odds ratio)

	smoking(x=1)	non-smoking(x=0)
lung cancer(y=1)	a	c
non (y=0)	b	d

- Risk Ratio(RR): 흡연의 폐암 발생률을 비흡연의 폐암 발생률로 나눈 값

$$RR = \frac{P(Y=1|x=1)}{P(Y=1|x=0)} = \frac{a/(a+b)}{c/(c+d)}$$

- 오즈:
  - 흡연(비흡연)의 폐암 확률을 흡연(비흡연)의 폐암이 발생하지 않을 확률로 나눈 값
  - 특정 리스크에 노출될 경우의 위험도로 해석

$$odds(x=1) = \frac{P(Y=1|x=1)}{P(Y=0|x=1)} = \frac{a/(a+b)}{b/(a+b)} = \frac{a}{b}$$

로지스틱 회귀에서 오즈비 (odds ratio)

- 오즈비
  - 설명변수 $x = 1$에서의 오즈와 $x = 0$에서의 오즈의 비
  - x가 한 단위 증가할 때 $y = 1$일 위험과 $y = 0$일 위험의 비의 증가율
  - 특정 리스크에 노출될 경우, 그렇지 않은 경우에 대한 상대적 위험도

$$\frac{P(Y = 1|x = 1)/P(Y = 0|x = 1)}{P(Y = 1|x = 0)/P(Y = 0|x = 0)} = \exp(\beta_1).$$

- RR vs OR
  - $P(Y = 1|x = 0) \approx 0$
  - 즉, 리스크에 노출되지 않을 경우 질병에 걸릴 확률)이 아주 작으면 (희귀성의 가정)
  - $RR \approx OR$

로지스틱 회귀에서 오즈비 (odds ratio)

- 로그 오즈비 : 오즈비에 로그를 취한 값으로 회귀계수와 일치
- 예: x는 흡연 유무이고 y는 폐질환 여부 (1, 0)
  - $\hat{\beta} = 3.72 \rightarrow$ odds ratio $= \exp(3.72) = 42$
  - 흡연자의 폐질환에 대한 위험이 비흡연자의 위험에 비해 **42**배 증가하는 것으로 해석

예제: 전립선암 (Prostate Cancer) 양성 여부

- 림프절이 전립선암에 대해 양성인지 여부
- 53명의 환자자료
- 변수 설명

변수명	변수 설명
aged	환자의 연령
stage	질병 단계: 질병이 얼마나 진행되어 있는지 나타내는 측도
grade	종양의 등급: 진행의 정도
xray	X-선 결과
acid	혈청인산염(serum acid phosphatase) 특정한 부위에 종양이 전이되었을 때 상승되는 혈청의 인산염값
r	전립선암 양성(1) 여부

예제: 전립선암 (Prostate Cancer) 양성 여부

```
library(boot)
x<-nodal[,-1]
head(x, 20)
```

```
##    r aged stage grade xray acid
## 1  1    0     1     1    1    1
## 2  1    0     1     1    1    1
## 3  1    0     1     1    1    1
## 4  1    0     1     1    1    1
## 5  1    0     1     1    1    1
## 6  0    0     1     1    1    1
## 7  1    0     0     0    0    1
## 8  0    0     0     0    0    1
## 9  0    0     0     0    0    1
## 10 0    0     0     0    0    1
## 11 0    0     0     0    0    1
## 12 0    0     0     0    0    1
## 13 0    1     1     1    0    0
## 14 0    1     1     1    0    0
## 15 0    1     1     1    0    0
## 16 0    1     1     1    0    0
## 17 1    1     1     0    0    1
```

예제: 전립선암 (Prostate Cancer) 양성 여부

- initial model

```
gfit = glm(r~., data=x, family="binomial")
coef(summary(gfit))
```

```
##                Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -3.0793806  0.9867696 -3.1206684 0.001804411
## aged        -0.2917427  0.7540054 -0.3869239 0.698812567
## stage        1.3729295  0.7838488  1.7515235 0.079855775
## grade        0.8719723  0.8155785  1.0691457 0.285004012
## xray         1.8008141  0.8104165  2.2220847 0.026277583
## acid         1.6839295  0.7914741  2.1275863 0.033371400
```

예제: 전립선암 (Prostate Cancer) 양성 여부

- stepwise selection

```
m <- step(gfit, trace =0)
coef(summary(m))
```

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -3.051787  0.8420409 -3.624273 0.0002897749
## stage        1.645346  0.7296744  2.254905 0.0241392977
## xray         1.911626  0.7771336  2.459842 0.0138998172
## acid         1.637778  0.7539433  2.172283 0.0298343277
```

예제: 전립선암 (Prostate Cancer) 양성 여부

- 유의한 변수들(유의확률: $p < 0.05$)은 전립선암에 영향을 줌
  - 질병의 단계(`stage`)가 심화
  - X-선 결과(`xray`)가 좋지 않을수록
  - 혈청인산염 값(`acid`)이 높을수록
- `stage`(질병의 단계)의 오즈비
  - $\exp(1.645346) = 5.18280$
  - 질병의 진행단계가 심화된 그룹은 그렇지 않은 그룹에 비해 전립선암에 노출 위험이 약 5.2배 정도 높음