분석프로그래밍 I.12
고급 통계 분석

국민대학교 경영학부 빅데이터경영통계학 전공

2018. 6. 04(월)

1 일반 선형 모형General Linear Model

- 분산분석ANOVA: ANalysis Of VAriance와 다중회귀Multiple Regression을 모두 포함하는 개념

```r
data(tips, package='reshape2')
## ANOVA
tipAnova <- aov(tip ~ day - 1, tips)
tipIntercept <- aov(tip ~ day, tips)

# 계수 비교
coef(tipAnova)
coef(tipIntercept)

# 계수 검정
library(lmtest) # coeftest
coeftest(tipAnova)
coeftest(tipIntercept)

## 다중회귀
tipLm <- lm(tip ~ day - 1, tips)
tipLmIntercept <- lm(tip ~ day, tips)

## 코딩 방법 바꾸기: 코딩 방법에 따라 계수의 의미가 달라진다.
tips2 <- tips
```

```r
levels(tips2$day)
tips2$day <- relevel(tips2$day, ref="Sun"); # new reference,
    Sunday
coef(lm(tip ~ day, tips2))
coef(lm(tip ~ day, tips2, contrasts = list(day = contr.sum)))
coef(lm(tip ~ day, tips2, contrasts = list(day = MASS::contr.sdif
    )))
```

Listing 1: `aov`와 `lm`의 분석 결과 비교

2 다중 선형 회귀Multiple Linear Regression

- 여러 개의 예측 변수predictors로 하나의 결과 변수outcome variable을 예측하고자 한다.
- 결과 변수 y 의 조건부 평균은 계수 $\beta_0, \beta_1, \cdots, \beta_k$ 의 선형 결합linear combination으로 나타낼 수 있다.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$
$$\mathbb{E}[y_i | x_{i1}, x_{i2}, \cdots, x_{ik}] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

  이를 행렬로 나타내면 $Y = X\beta + \epsilon$ 이 되고, $\hat{\beta} = (X^T X)^{-1} X^T Y$ 를 구할 수 있다.

- 기본적인 가정 : LINE(Linear, Independent, Normal, Equal variance)
- `lm` 함수를 사용한다. `lm(`*formula*`, data)`에서 *formula*를 통해 모형을 정한다. 이때 *formula*에 쓰이는 `+`, `-`, `*`, `/`는 산술적인 가감승제가 아니라 Wilkinson-Rogers 표기법을 따른다.

```r
dat = data.frame(y= , x1= , x2= , x3= )
lm(y ~ x1 + x2, data=dat)
lm(y ~ x1 + x2 + 1, data=dat)
lm(y ~ x1 + x2 - 1, data=dat)
lm(y ~ I(x1 + x2), data=dat)
lm(exp(y) ~ x1 + x2, data=dat)
lm(y ~ exp(x1) + x2, data=dat)
lm(y ~ x1 * x2, data=dat)  # x1 + x2 + x1:x2
```

```r
lm(y ~ (x1 + x2 + x3)^2 - x1:x2, data=dat)  # x1 + x2 + x3 + x1:x3 + x1:x2
lm(y ~ x1 / x2, data=dat)  # x1 + x2:x2, x1 + x2 %in% x1
fit <- lm(y ~ ., data=dat)
update(fit, formula = . ~ . - x1)
```

Listing 2: 함수 `lm`의 *formula* 예

- 교재의 housing 자료에 대한 분석의 예

```r
# 교재 p.356
housing <- read.table('https://www.jaredlander.com/data/housing.csv',
                      sep = ",",
                      header = TRUE,
                      stringsAsFactors = FALSE)
head(housing)
names(housing) <-c("Neighborhodd", "Class", "Units", "YearBuilt",
                   "SqFt", "Income", "IncomePerSqFt", "Expense",
                   "ExpensePerSqFt", "NetIncome", "value",
                   "ValuePerSqFt", "Boro")

#p364
house1 <- lm(ValuePerSqFt ~ Units + SqFt + Boro, data=housing)
fHouse1 <- formula(ValuePerSqFt ~ Units + SqFt + Boro)
summary(house1)
house1$coefficients
coef(house1)
coefficients(house1)

library(coefplot) #install.packages('coefplot')
coefplot(house1)

# coding
housing2 <- housing # reference: Bronx
levels(housing2$Boro)
```

```
housing2$Boro <- relevel(housing2$Boro, ref="Manhattan"); # new reference, Manhattan
coef(lm(fHouse1, housing2))
coef(lm(fHouse1, housing2, contrasts = list(Boro = contr.sum)))
coef(lm(fHouse1, housing2, contrasts = list(Boro = MASS::contr.sdif)))

# coefficients
coef(house1)
confint(house1) # confidence intervals of the parameter estimates
lmtest::coefci(house1)
# compare the speed of confint and coefci for glm.
lmtest::coeftest(house1)
vcov(house1) # estimated variance-covaraince matrix for parameter estimates

# fitted values, new predicitions
fitted(house1) # same as predict(house1, housing) or predict(house1, house1$data)
# fitted values
predict(house1, data.frame(Units=45, SqFt = 554000, Boro = "Manhattan")) # prediction for new data

# residuals
residuals(house1)
resid(house1)
car::residualPlots(house1) # car(Companion to Applied Regression)

# fit
deviance(house1) # residual sum of squares(RSS) for linear models
logLik(house1)
AIC(house1)
BIC(house1)

# variance inflation factors
car::vif(house1)
```

```

# diagnostics
plot(house1, which=1:6)
car::qqPlot(house1)
car::inverseResponsePlot(house1)

# outliers
car::outlierTest(house1)
car::influenceIndexPlot(house1)
car::influenceIndexPlot(house1, id = list(n=5)) # 5 most influential obs.

# New models
house2 <- lm(ValuePerSqFt ~ Units * SqFt + Boro, data=housing) # Nesting Model
house3 <- lm(ValuePerSqFt ~ Units : SqFt + Boro, data=housing) # Non-Nestting Model

# model comparison
anova(house1, house2)
lmtest::waldtest(house1, house2)
lmtest::encomptest(house1, house3)

multiplot(house1, house2)
car::compareCoefs(house1, house2, se=F)
car::compareCoefs(house1, house2)

AIC(house1, house2, house3)
BIC(house1, house2, house3)

# Stepwise variable selection p.415
nullModel <- lm(ValuePerSqFt ~ 1, data = housing)
fullModel <- lm(ValuePerSqFt ~ Units + SqFt * Boro + Boro * Class, data =housing)

houseStep <- step(nullModel, scope = list(lower = nullModel,
```

```
    upper = fullModel), direction = 'both')
```

Listing 3: 교재의 housing 자료에 대한 분석 예

3 일반화 선형 모형 Generalized Linear Model

- 결과 변수가 $-\infty \sim \infty$ 의 연속값을 갖지 않는 경우 (예. 개체수, 사망/생존 등) 에 사용할 수 있다.
- 결과 변수의 기댓값은 계수의 선형 결합에 선형/비선형 함수를 적용한 결과로 결정된다.[1]

  $$\mathbb{E}[Y|X] = g^{-1}(X\beta)$$

  로 표현할 수도 있다.
- 교재의 예

```
# 20. Generalized linear model. p376
acs <- read.table('http://jaredlander.com/data/acs_ny.csv', sep='
    ,', header=T, stringsAsFactors=F)
head(acs)
acs$Income <- with(acs, FamilyIncome >= 150000)
library(ggplot2)
library(useful)
ggplot(acs, aes(x=FamilyIncome)) +
  geom_density(fill = 'grey', color='grey') +
  geom_vline(xintercept=150000) +
  scale_x_continuous(labels=multiple.dollar, limits=c(0,1000000))

income1 <- glm(Income ~ HouseCosts + NumWorkers + OwnRent +
    NumBedrooms + FamilyType, data = acs, family=binomial(link='
    logit'))
summary(income1)

invlogit <- function(x) {
  1/(1+exp(-x))
```

[1] $g\left[\mathbb{E}[Y|X]\right] = X\beta$

```r
}

invlogit(income1$coefficients)

# p380
ggplot(acs, aes(x=NumChildren)) + geom_histogram(binwidth=1)
children1 <- glm(NumChildren ~ FamilyIncome + FamilyType +
    OwnRent, data = acs, family = poisson(link="log"))
summary(children1)
coefplot(children1)

z <- (acs$NumChildren - children1$fitted.values) / sqrt(children1
    $fitted.values)
sum(z^2)/children1$df.residual
pchisq(sum(z^2), children1$df.residual)

children2 <- glm(NumChildren ~ FamilyIncome + FamilyType +
    OwnRent, data = acs, family = quasipoisson(link='log'))
summary(children2)
coefplot::multiplot(children1, children2)

deviance(children1) #
logLik(children1)
```

Listing 4: 교재의 housing 자료에 대한 분석 예

4 조절 효과Moderation Effect

- 조절 효과의 예: 남녀에 따른 키와 체중의 관계

```r
gender <- rep(c("M","F"), each=100)
height <- c(rnorm(100, 175, 13), rnorm(100, 160, 10))
weight <- rep(NA, 200)
weight[1:100] <- 0.4*height[1:100] + 8 + rnorm(100)*10
weight[101:200] <- 0.25*height[101:200] + 18 + rnorm(100)*10

plot(weight ~ height, col=factor(gender))
```

```r

fit1 <- lm(weight ~ height)
fit2 <- lm(weight ~ gender)
fit3 <- lm(weight ~ height : gender)
fit4 <- lm(weight ~ height * gender)

car::compareCoefs(fit1, fit2, fit3, fit4)
# fit1
plot(weight ~ height, col=factor(gender))
abline(fit1, col="black")
# fit2
plot(weight ~ height, col=factor(gender))
abline(h = coef(fit2)["(Intercept)"])
abline(h = sum(coef(fit2)), col="red")
# fit3
plot(weight ~ height, col=factor(gender))
abline(a = coef(fit3)["(Intercept)"], b= coef(fit3)["height:genderF"])
abline(a = coef(fit3)["(Intercept)"], b= coef(fit3)["height:genderM"], col='red')
# fit4
plot(weight ~ height, col=factor(gender))
abline(a = coef(fit4)["(Intercept)"], b= coef(fit4)["height"])
abline(a = sum(coef(fit4)[c("(Intercept)", "genderM")]),
       b = sum(coef(fit4)[c("height", "height:genderM")]), col='red')

coplot(weight ~ height | gender)
```

Listing 5: 조절 효과의 예

5 선형 혼합 효과 모형 Linear Mixed Effect Model

- 집단별로 다른 선형 회귀선을 고려하여 분석하는 방법이다.
- 집단별로 다른 선형 계수가 특정한 분포(예. 정규분포)를 띈다고 가정하고 분석한다.

6 요인 분석Factor Analysis과 문항 반응 이론Item Response Theory

- 여러 변수에 나타나는 상관관계를 설명할 수 있는 잠재 변수를 가정한다.
- 변수가 연속이면 요인 분석, 변수가 이항 또는 다항이라면 문항 반응 이론으로 분석할 수 있다.

```
library(mirt)
library(psych)
data(Science)

# EFA
fit0 <- fa(Science)
print(fit0)
summary(fit0)

# IRT(Graded Response Model)
fit1 <- mirt(Science, 1)
summary(fit1)
fscores(fit1)
personfit(fit1)
itemfit(fit1)
itemplot(fit1, 4)
testinfo(fit1, Theta=0)

##
library(lavaan)
data(HolzingerSwineford1939)
dat <- HolzingerSwineford1939[,paste0("x",1:9)]

# determining the number of factors
fa.parallel(dat, fm='ml')
fit0 <- fa(dat, nfactors=3)
summary(fit0)
diagram(fit0)
```

Listing 6: 요인 분석과 문항 반응 분석의 예

7 경로 분석 Path Analysis

- 선형 모형에서 계수는 모형(설명 변수의 종류)에 따라 달라진다.
- 예측 모형은 인과 관계를 나타내지 않는다.
- 경로 분석은 여러 변수 사이의 인과 관계를 분석하기 위해 사용된다.
- 예. Simpson's paradox, Berkinson's paradox

8 구조 방정식 모형 Structural Equation Model

- 요인 분석 모형과 경로 분석 모형을 합쳐 놓은 것이다.
- 여러 변수에 나타나는 상관관계를 설명할 수 있는 잠재 변수를 가정한다.
- 변수가 연속이면 요인 분석, 변수가 이항 또는 다항이라면 문항 반응 이론으로 분석할 수 있다.

```
library(lavaan)
model <- '
# measurement model
  ind60 =~ x1 + x2 + x3
  dem60 =~ y1 + y2 + y3 + y4
  dem65 =~ y5 + y6 + y7 + y8
# regression model
  dem60 ~ ind60
  dem65 ~ ind60 + dem60
# residual correlations
  y1 ~~ y5
  y2 ~~ y4 + y6
  y3 ~~ y7
  y4 ~~ y8
  y6 ~~ y8'
fit <- sem(model, data=PoliticalDemocracy)
summary(fit, standardized=T)
semPlot::semPaths(fit)
```

Listing 7: SEM 분석의 예