제7강: 연관성 분석 & 회귀 분석

금융 통계 및 시계열 분석

TRADE INFORMATIX

2014년 1월 28일

목차

통계적 연관성

- 통계적 연관성 (Statistical Association)
  - 두 개의 확률사건이 독립적이 아닐 때 통계적 연관성을 가진다.
  - 상관관계 (correlation) 는 통계적 연관성의 한 종류
- 확률사건의 독립 (event independence)
  - 두 개의 확률사건 A, B 가 동시에 일어날 확률이 각각의 확률 사건이 일어날 확률의 곱인 경우

$$P(A \cap B) = P(A) \cdot P(B) \tag{1}$$

- 확률변수의 독립 (random variable independence)
  - 두 개의 확률변수 X, Y 에 대해 모든 확률사건 $P\{X \leq a\}$, $P\{Y \leq b\}$ 가 독립이면 그 두 확률변수는 독립
  - 이를 확률분포로 나타내면 각각의 확률분포함수의 곱이 joint 확률분포와 같으면 두 확률변수는 독립

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y) \tag{2}$$

확률변수의 유형

1. 카테고리 값 (categirical value). Nominal Value
   - ❑ 이산적 (discrete) 인 값. 정수 (integer) 값으로 대표할 수는 있지만 크기의 비교가 불가능
   - ❑ 특수한 경우로 success/pass 의 두 가지 상태값만 가지는 경우 (binary)
   - ❑ 분할표 (table) 를 이용하면 특정한 값의 그룹에 속하는지 아닌지를 사용하여 count 정수로 나타낼 수 있음
2. 순서값. Ordinal Value
   - ❑ 이산적 (discrete) 인 경우도 있고 연속적 (continuous) 인 경우도 있음. 정수 (integer) 값 혹은 실수 (real value) 으로 대표할 수는 있으며 크기의 상대적인 비교가 가능
3. 실수값. Real Value
   - ❑ 임의의 연속적 (continuous) 인 값. 양수/음수 모두 가능한 경우와 duration 값과 같이 양수만 가능한 경우가 있음

확률변수의 유형에 따른 연관성 분석

1. 독립변수와 종속변수가 모두 카테고리값인 경우
   - ❑ 각각의 카테고리에 해당하는 자료의 수 (count) 를 분할표 (table) 로 분석
   - ❑ Pearson's Chi-squared test
   - ❑ proportion test
     - ► 확률적인 양 중 하나는 binary value이고 나머지가 카테고리값인 경우
2. 독립변수가 카테고리값이고 종속변수가 실수인 경우
   - ❑ ANOVA (Analysis of Variance)
     - ► One-way ANOVA : 두 값중 하나는 실수이고 하나는 카테고리값
     - ► Two-way ANOVA : 세 값중 하나는 실수이고 나머지 두 값은 카테고리값
3. 독립변수와 종속변수가 모두 실수인 경우
   - ❑ Pearson product-moment linear correlation coefficient. 일반적인 상관도 (correlation) 정의
   - ❑ Spearman's ρ
   - ❑ Kendall's τ
4. 종속변수가 카테고리값이고 독립변수가 실수인 경우
   - ❑ Classification, Clustering
     - ► 일반적인 패턴인식의 방법론

Binomial Test

문제 1 : 상승확률 비교

삼성전자의 주가가 전일 대비 상승할 확률이 p인 Bernoulli trial 일때 2011년에는 247일중 120일 상승하고 2012년에는 247일중 126일 상승하였다. 상승확률 p가 변화하였는가?

```
> library(rquantbook)
> df1 <- get_quantbook_data("krx_stock_daily_price", ticker="005930",
+           date_start="2011-01-01", date_end="2011-12-31")
> df2 <- get_quantbook_data("krx_stock_daily_price", ticker="005930",
+           date_start="2012-01-01", date_end="2012-12-31")
> p1 <- df1$close
> p2 <- df2$close
> d1 <- p1[-1] > p1[-length(p1)]
> d2 <- p2[-1] > p2[-length(p2)]


> c(length(d1[d1==TRUE]), length(d1))
[1] 120 247
> c(length(d2[d2==TRUE]), length(d2))
[1] 126 247
```

Binomial Test

- binomial 분포의 확률값에 대한 검정
- 성공확률이 p, 실패확률이 $q = 1 - p$인 경우, 전체 n개의 시도에서 K번의 성공이 나올 확률은

$$Z = \frac{K - np}{\sqrt{npq}} \tag{3}$$

- $n > 25$ 인 경우에 Z는 표준 정규 분포로 수렴

$$z \propto N(0, 1) \tag{4}$$

Binomial Test in R

binom.test

- ❑ `binom.test(x, n, p=0.5)`
  - ▶ x : 성공 카운트
  - ▶ n : 전체 카운트
  - ▶ p : 테스트 하려는 확률값

```
> binom.test(126, 247, 120/247)

Exact binomial test

data:  126 and 247
number of successes = 126, number of trials = 247,
p-value = 0.4459
alternative hypothesis: true probability of success is not equal to 0.48583
95 percent confidence interval:
 0.4459603 0.5740368
sample estimates:
probability of success
             0.5101215
```

Chi-Squared Test (Case 1)

문제 2 : 카테고리 비율 비교

삼성전자, 현대차, 포스코 세 종목 중 당일 가장 많이 상승한 종목을 우승 종목으로 하였을 때 각각 우승한 횟수는 2011년에 87, 94, 66번이다. 각 종목의 우승 확률은 같다고 할 수 있는가?

```
> library(rquantbook)
> api <- "krx_stock_daily_price"
> d11<-"2011-01-01";d12<-"2011-12-31";d21<-"2012-01-01";d22<-"2012-12-31";
> df11 <- get_quantbook_data(api, ticker="005930", date_start=d11, date_end=d12)
> df12 <- get_quantbook_data(api, ticker="005380", date_start=d11, date_end=d12)
> df13 <- get_quantbook_data(api, ticker="005490", date_start=d11, date_end=d12)
> best_count <- function(df1, df2, df3) {
+   p1 <- df1$close; p2 <- df2$close; p3 <- df3$close
+   d1 <- (p1[-1] - p1[-length(p1)])/p1[-length(p1)]
+   d2 <- (p2[-1] - p2[-length(p2)])/p2[-length(p2)]
+   d3 <- (p3[-1] - p3[-length(p3)])/p3[-length(p3)]
+   table(max.col(cbind(d1, d2, d3)))
+ }



> best_count(df11, df12, df13)

 1  2  3
87 94 66
```

Chi-Squared Test (Case 1)

- k 개의 카테고리 결과가 나올 수 있는 프로세스에 대해 각각의 카테고리 결과가 나올 확률이 $(p_1, p_2, \cdots, p_k)$ 인지 테스트
- n 번 시도 중 각각의 카테고리 결과가 나온 횟수 X_i 에 대해 기대값과의 오차의 제곱의 합은 자유도 $k-1$ 인 Chi-Squared 분포

$$\sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \propto \chi^2_{k-1} \tag{5}$$

Chi-Squared Test in R (Case 1)

chisq.test

- 복수개의 카테고리 자료의 확률분포에 대한 검정
- `chisq.test(x, p=rep(1/length(x), length(x)), correct=TRUE)`
  - `p` : 테스트 하려는 확률 비율
  - `correct` : cotinuity correction

```
> chisq.test(c(87, 94, 66))
Chi-squared test for given probabilities

data:  c(87, 94, 66)
X-squared = 5.1579, df = 2, p-value = 0.07585
```

Chi-Squared Test (Case 2)

문제 3 : 카테고리 비율 비교

삼성전자, 현대차, 포스코 세 종목 중 당일 가장 많이 상승한 종목을 우승 종목으로 하였을 때 각각 우승한 횟수는 2011년에 87, 94, 66번이고 2012년에 97, 78, 72 번이다. 2012년의 우승 비율은 2011년과 달라졌는가?

```
> library(rquantbook)
> api <- "krx_stock_daily_price"
> d11<-"2011-01-01";d12<-"2011-12-31";d21<-"2012-01-01";d22<-"2012-12-31";
> df11 <- get_quantbook_data(api, ticker="005930", date_start=d11, date_end=d12)
> df12 <- get_quantbook_data(api, ticker="005380", date_start=d11, date_end=d12)
> df13 <- get_quantbook_data(api, ticker="005490", date_start=d11, date_end=d12)
> df21 <- get_quantbook_data(api, ticker="005930", date_start=d21, date_end=d22)
> df22 <- get_quantbook_data(api, ticker="005380", date_start=d21, date_end=d22)
> df23 <- get_quantbook_data(api, ticker="005490", date_start=d21, date_end=d22)
> best_count <- function(df1, df2, df3) {
+   p1 <- df1$close; p2 <- df2$close; p3 <- df3$close
+   d1 <- (p1[-1] - p1[-length(p1)])/p1[-length(p1)]
+   d2 <- (p2[-1] - p2[-length(p2)])/p2[-length(p2)]
+   d3 <- (p3[-1] - p3[-length(p3)])/p3[-length(p3)]
+   table(max.col(cbind(d1, d2, d3)))
+ }



> best_count(df11, df12, df13); best_count(df21, df22, df23)

 1  2  3
87 94 66

 1  2  3
97 79 71
```

Chi-squared Test (Case 2)

- $r \times c$개의 contingency table에 대해 행과 열의 카테고리 분포가 독립적인지 테스트
- 즉, 결과값이 j번째 카테고리가 나올 확률이 행 i에 따라 달라지는지 테스트
- 실제 결과값과 독립적이라고 가정한 경우의 기대치의 오차의 제곱의 합은 Chi-Squared 분포로 수렴

$$H_0 : P(i, j) = P(i) \cdot P(j) \tag{6}$$

Chi-squared Test in R (Case 2)

chisq.test

- 복수개의 카테고리 자료의 확률분포에 대한 검정
- `chisq.test(x, correct=TRUE)`
  - `p` : 테스트 하려는 contingency table

```
> chisq.test(rbind(c(97,78,72),c(87,94,66)))

Pearson's Chi-squared test

data:  rbind(c(97, 78, 72), c(87, 94, 66))
X-squared = 2.2927, df = 2, p-value = 0.3178
```

Difference in Means between Groups

문제 4 : 카테고리별 평균 비교

고객중 성별에 따른 연령의 차이가 있는지 테스트

```
> df <- read.csv("client.csv",
+          fileEncoding="CP949", encoding="UTF-8",
+          stringsAsFactors=FALSE)
> data <- split(df, df[,2])
> df1 <- as.data.frame(t(sapply(data, function(d) {
+            c(d[[5]][1],d[[7]][1],d[[6]][1])})),
+          stringsAsFactors=FALSE)
> rownames(df1) <- NULL
> colnames(df1) <- c("city", "gender", "age")
> df2 <- df1[-as.numeric(rownames(df1[df1$city=="",])),]
> colnames(df2) <- c("city", "gender", "age")
> df2$city <- factor(df2$city)
> df2$gender <- factor(df2$gender)
> df2$age <- as.numeric(df2$age)
```

ANOVA (Analysis of Variance)

- 전체 n개의 샘플이 k개의 카테고리로 구분 가능할 때 샘플 그룹간에 평균의 차이가 존재하는지 테스트
- 각 샘플그룹은 분산의 크기가 같은 정규분포이어야 한다.
- 전체 샘플평균은 $\bar{x}$, 각 샘플그룹의 샘플갯수는 n_i, 샘플평균은 x_i, 샘플분산은 v_i 일 때
- 그룹내 분산 (within-group variance) V_W

$$V_W = \frac{1}{n-c} \sum_{i=1}^{c} (n_i - 1) v_i \tag{7}$$

- 그룹간 분산 (between-group variance) V_B

$$V_B = \frac{n}{c-1} \sum_{i=1}^{c} (x_i - \bar{x})^2 \tag{8}$$

- 이 때, 그룹내 분산과 그룹간 분산의 비율은 자유도 $(c-1, n-c)$인 F 분포를 이룬다.

$$\frac{V_B}{V_W} \sim F_{c-1,n-c} \tag{9}$$

ANOVA in R

aov

- `aov(formula, data)`
  - `formula` : 연관성 테스트를 위한 모델 포뮬라
  - `data` : dataframe

```
> result <- aov(age ~ gender, data=df2)
> result

Call:
   aov(formula = age ~ gender, data = df2)

Terms:
                  gender Residuals
Sum of Squares  1008.475  9454.169
Deg. of Freedom        1        57

Residual standard error: 12.87877
Estimated effects may be unbalanced

> summary(result)

            Df Sum Sq Mean Sq F value Pr(>F)
gender       1   1008  1008.5    6.08 0.0167 *
Residuals   57   9454   165.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA 결과 in R

model.tables

- `model.tables(result, type)`
  - `result` : ANOVA 결과
  - `type` : `effects`이면 ANOVA 계수에 대한 결과, `means`이면 그룹 평균에 대한 결과 표시

```
> model.tables(result, type="effects")
Tables of effects

 gender

     4.658 -3.67
rep 26.000 33.00
> model.tables(result, type="means")
Tables of means
Grand mean

45.45763

 gender

    50.12 41.79
rep 26.00 33.00
```

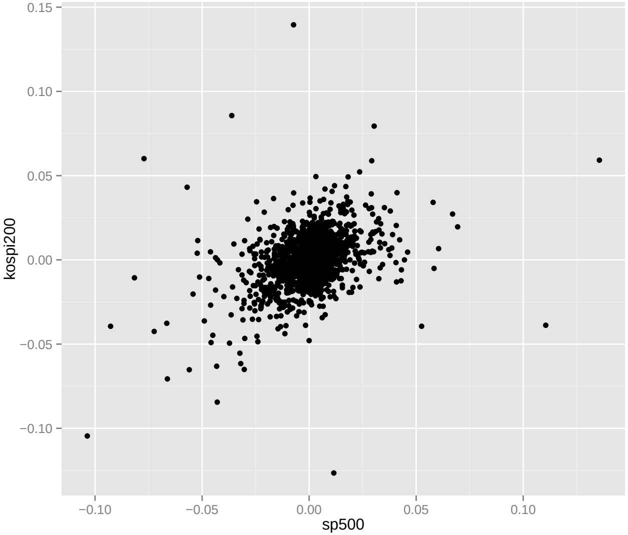
Simple Correlation

문제 5 : 연속변수의 상관관계

SP500(SPDR ETF) 과 KOSPI200(KODEX 200 ETF) 사이의 수익률의 상관관계는?

```
> library(quantmod)
> d1 <- getSymbols("NYSEARCA:SPY", src="google", auto.assign=FALSE)
> d2 <- getSymbols("KRX:069500", src="google", auto.assign=FALSE)
> d <- merge(lag(d1,1), d2)
> r <- ROC(d)
> x <- coredata(r[,4])
> y <- coredata(r[,9])
> xy <- data.frame(x,y)
> names(xy) <- c("sp500", "kospi200")
```

S&P 500 vs KOSPI 200



상관계수 (Correlation)

상관계수

- ❑ 두 확률변수의 선형 상관관계를 나타내는 척도 (Pearson Correlation)
  - ▶ $\rho = 1$: 완전 선형 상관 관계
  - ▶ $\rho = 0$: 무상관 (독립과는 다름)
  - ▶ $\rho = -1$: 완전 선형 반상관 관계
- ❑ 비선형 상관관계는 측정 불가능
- ❑ 실제 물리적인 상관관계가 없어도 spurious correlation 이 나타날 수 있음

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sqrt{E[(X-\bar{X})(Y-\bar{Y})]}}{\sqrt{E[(X-\bar{X})^2]}\sqrt{E[(X-\bar{Y})^2]}} \tag{10}$$

Spearman correlation & Kendall correlation

- Spearman's rank correlation coefficient ρ_s
  - 두 변수를 순위(rank)로 변환한 후에 순위에 대해 Pearson Correlation을 구함
  - 비선형함수라도 단조함수(monotonic function)이면 상관관계 계산 가능
- Kendall tau rank correlation coefficient τ
  - 두 변수를 순위(rank)로 변환한 후에 두 변수의 순위가 같은 concordant 짝의 수를 이용하여 계산

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$$

Correlation Test

Correlation Test

- 상관계수가 유의미한 값 즉 0이 아닌 값을 가지는지 검정

$$H_0 : \rho = 0 \tag{11}$$

- test statistics : Student-t 분포

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2} \tag{12}$$

Correlation in R

cor

- cor(x, y, method)
  - x, y : 두 확률변수의 샘플집합
  - method : correlation 정의. "pearson", "kendall", "spearman"
- cor.test(x, y, alternative, method)
  - alternative : two.sided", "less", "greater"

```
> cor(x,y)
                    KRX.069500.Close
NYSEARCA.SPY.Close                NA
> cor.test(x,y)
Pearson's product-moment correlation

data:  x and y
t = 16.5047, df = 1605, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3383172 0.4219541
sample estimates:
      cor
0.3809146
```

선형 회귀

- 반응변수 y의 기대값 μ를 설명변수 x의 선형 조합으로 설명하려는 시도

$$y \sim N(\mu, \sigma) = N(b_1 x + b_0, \sigma) \tag{13}$$

- 반응변수 y와 설명변수에 의한 예측값의 오차 e는 정규분포

$$y - (b_1 x + b_0) = e \sim N(0, \sigma) \tag{14}$$

Solution 1 : Generalized method of moments

- GMM 조건 : 추정 오차와 설명 변수는 무상관관계

$$\mathsf{E}\left[\, x_i e_i \,\right] = \mathsf{E}\left[\, x_i (y_i - x_i' \beta) \,\right] = 0. \tag{15}$$

$$\hat{b}_1 = \frac{\mathsf{Cov}[x, y]}{\mathsf{Var}[x]} \tag{16}$$

Solution 2 : OLS (Ordinary least squares)

❑ 실제 샘플의 값 $\{y_i\}$과 선형 회귀로 인한 예측치 $\{b_1 x_i + b_0\}$의 관계를 선형대수방정식으로 표시

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \tag{17}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \\ x_n & 1 \end{pmatrix}, \quad \beta = \begin{pmatrix} b_1 \\ b_0 \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}, \tag{18}$$

❑ $\{y_i\}$와 $\{b_1 x_i + b_0\}$사이의 오차 제곱의 합을 최소화

$$\hat{\beta} = \arg\min \sum_{i=1}^{N} (y_i - b_1 x_i - b_0)^2 = \arg\min (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \tag{19}$$

❑ 계수 추정치 $\hat{\beta}$

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \tag{20}$$

선형회귀계수 검정

- 선형회귀계수 b_1, b_0 는 Student-t 분포

$$t = \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} \sim t_{n-2} \tag{21}$$

- standard error $s_{\hat{\beta}}$ 는

$$s_{\hat{\beta}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{22}$$

결정계수 (coefficient of determination)

❑ 추정된 선형회귀모형이 실제 자료를 설명할 수 있는 능력의 척도

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} \tag{23}$$

$$\text{Unexplained Variation} = \sum (y_i - \hat{y}_i)^2 \tag{24}$$

$$\text{Total Variation} = \sum (y_i - \bar{y})^2 \tag{25}$$

❑ multiple R : $R = \sqrt{R^2}$

- ▶ 설명변수가 1개인 simple regression 에서는 correlation과 일치

Linear Regression in R

lm

- `lm(formula, data)`
  - `formula` : 모형 포뮬라
  - `data` : 모형에 사용된 자료가 dataframe인 경우

```
> m <- lm(kospi200 ~ sp500, data=xy)
> summary(m)

Call:
lm(formula = kospi200 ~ sp500, data = xy)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.131324 -0.007368  0.000141  0.007000  0.142199 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.0002237  0.0003590   0.623    0.533    
sp500       0.3980029  0.0241146  16.505   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01439 on 1605 degrees of freedom
  (230 observations deleted due to missingness)
Multiple R-squared: 0.1451, Adjusted R-squared: 0.1446
F-statistic: 272.4 on 1 and 1605 DF,  p-value: < 2.2e-16
```

Linear Model Object

- `coef` : 회귀모형 계수
- `confint` : 회귀모형 계수의 신뢰구간
- `fitted` : 회귀모형 fitting 결과
- `predict` : 회귀모형 예측 결과
- `residuals` : 회귀모형 오차

```
> coef(m)
(Intercept)       sp500
 0.00022365  0.39800288
> confint(m)
                    2.5 %       97.5 %
(Intercept) -0.0004804584 0.0009277584
sp500        0.3507034890 0.4453022630
```

Linear Model Plot

```
> plot(x, y,
+      xlab="S&P500",
+      ylab="KOSPI200",
+      main="Returns")
> abline(m, lwd=3)
> pr <- predict(m,
+         interval="confidence",
+         level=0.999)
> c1 <- cbind(m$model$sp500, pr[,2])
> c2 <- cbind(m$model$sp500, pr[,3])
> lines(c1[order(c1[,1]),], lty=2)
> lines(c2[order(c2[,1]),], lty=2)
```