회귀분석(Regression Analysis)

http://contents.kocw.or.kr/document/LN09_1.pdf
http://contents.kocw.or.kr/document/LN10_1.pdf

회귀분석: 특정 변수(독립변수)가 다른 변수(종속변수)에 어떠한 영향을 미치는 가를 분석하는 방법
인과관계가 있는지를 분석하는 방법
Where
인과관계: 변수 A가 변수 B의 값을 변하게 하는 원인이 되는 관계. 이때 변수 A를 독립변수, 변수 B를 종속변수로 지칭

한 변수의 값을 가지고 다른 변수의 값을 예측해 주는 분석방법

상관관계 분석 vs. 회귀분석 차이점

1) 상관관계 분석: 변수 간의 관련성 분석
2) 회귀분석: 변수 간의 인과관계 분석

회귀분석의 특징

1) 가장 강력하고 사용범위가 넓은 분석 방법
2) 독립변수가 종속변수에 영향을 미치는 변수를 규명하고, 이들 변수에 의해서 회귀방정식(Y=a+bX where a: 상수, b: 회귀계수, X: 독립변수, Y: 종속변수)을 도출하여 회귀선을 추정
3) 회귀계수는 단위시간에 따라 변하는 양(기울기)이며, 회귀선을 추정함에 있어 최소자승법을 이용
4) 독립변수와 종속변수가 모두 등간척도 또는 비율척도로 구성되어 있어야 한다.

더 알아보기 (최소자승법)
잔차(오차): 관측치와 예측치의 차
최소자승법: 잔차들의 제곱의 합이 최소가 되도록 정하는 방법

2.1 회귀방정식의 이해

회귀선(Regression Line):

1) 한 변수의 증감이 다른 변수의 단위증가에 대해 어느 정도인가를 나타내는 선
2) 두 집단의 분포에서 잔차(각 값들의 편차)들의 제곱의 합을 최소화시키는(최소자승법) 회귀방정식에 의해 만들어진다.
3) 두 변수 간의 예측 관계에 있어서 한 변수에 의해서 예측되는 다른 변수의 예측치들이 그 변수의 평균치로 회귀하는 경향이 있다고 하여 갈튼(Galton)에 의해서 명명되었다.

2.2 단순 회귀분석

독립변수와 종속변수가 각각 한 개 일 때 독립변수가 종속변수에 미치는 인과관계를 분석하고자 할 때 사용.
회귀분석 수행 시 기본 가정이 충족되어야 사용 가능

회귀분석의 기본 가정:
선형성: 독립변수와 종속변수가 선형적. 회귀선 확인
잔차 정규성: 잔차(종속변수의 관측값과 회귀모델의 예측값 간의 차이)의 정규성. 정규성 검정
잔차 독립성: 잔차들은 서로 독립적. 더빈-왓슨 값 확인
잔차 등분산성: 잔차들의 분산이 일정. 표준잔차와 표준예측치 도표
다중 공선성: 독립변수 간의 강한 상관관계로 인한 문제 발생 여부. 분산팽창요인(VIF) 확인

회귀분석 절차:
1단계: 회귀분석의 기본 가정 충족 여부 확인
2단계: 분산분석의 F값으로 회귀모형의 유의성 여부 판단
3단계: 독립변수와 종속변수 간의 상관관계와 회귀모형의 설명력 확인
4단계: 가설의 채택 여부 결정
5단계: 회귀방정식을 적용하여 회귀식을 수립하고 결과 해석

귀무가설: 제품적절성이 제품만족도에 영향을 미친하고 볼 수 없다.
대립가설: 제품적절성이 제품만족도에 영향을 미친다고 볼 수 있다.

실습 (단순 선형 회귀분석 수행)

1단계: 데이터 가져오기

```
product <- read.csv("product.csv", header = TRUE)
str(product)
```

2단계: 독립변수와 종속변수 생성

```
y = product$제품_만족도
x = product$제품_적절성
df <- data.frame(x, y)
```

단순 선형회귀 분석은 lm()함수 이용
형식: lm(formula = Y ~ X, data)
Where
X: 독립변수
Y: 종속변수
data: 데이터프레임

3단계: 단순 선형회귀 모델 생성

```
result.lm <- lm(formula = y ~ x, data = df)
```

4단계: 회귀분석의 절편과 기울기

```
result.lm
```

5단계: 모델의 적합값과 잔차 보기

```
names(result.lm)
```

5-1단계: 적합값 보기

```
fitted.values(result.lm)[1:2]
```

fitted.values: 모델이 예측한 적합값

5-2단계: 관측값 보기

```
head(df, 1)
```

5-3단계: 회귀방정식을 적용하여 모델의 적합값 계산

```
Y = 0.7789 + 0.7393 * 4
Y
```

5-4단계: 잔차(오차) 계산

```
3 - 3.735963
```

5-5단계: 모델의 잔차 보기

```
residuals(result.lm)[1:2]
```

residuals: 모델의 잔차

5-6단계: 모델의 잔차와 회귀방정식에 의한 적합값으로부터 관측값 계산

```
-0.7359630 + 3.735963
```

실습 (선형 회귀분석 모델 시각화)

1단계: xy 산점도

```
plot(formula = y ~ x, data = product)
```

2단계: 선형 회귀모델 생성

```
result.lm <- lm(formula = y ~ x, data = product)
```

3단계: 회귀선

```
abline(result.lm, col = "red")
```

더 알아보기(회귀선(regression line)과 회귀(regression)
회귀: 생물학적 연구에서 부모의 특이한 형질이 자식에게서는 약해지고 "평균으로 돌아가려는 경향" 때문에 "회귀"라는 용어를 붙였다고 한다.

실습 (선형 회귀분석 결과보기)

summary(result.lm)

결정계수: 독립변수에 의해서 종속변수가 얼마만큼 설명되었는가를 나타내는 회귀모형의 설명력 1에 가까울수록 설명변수(독립변수)가 설명을 잘한다 라고 판단

수정결정계수(Adjusted R-sqaured): 오차를 감안하여 조정된 R값. 실제 분석에서는 이 값을 이용
F-statistic: F-검정통계량으로 회귀모형의 적합성(회귀선이 모형에 적합)

[표 15.3] 제품적절성에 따른 제품만족도 영향 분석
독립변수, 종속변수, 검통통계량, 유의확률, 분석통계량, 회귀식

2.3 다중 회귀분석

여러 개의 독립변수가 동시에 한 개의 종속변수에 미치는 영향을 분석할 때 이용하는 분석 방법

공차한계(Tolerance): 한 독립변수가 다른 독립변수들에 의해서 설명되지 않은 부분
분산팽창요인(Variance Inflation Factor, VIF): 공차한계의 역수

더 알아보기 (다중 공선성(Multicollinearity)문제)

다중 공선성: 한 독립변수의 값이 증가할 때 다른 독립변수의 값이 증가하거나 감소하는 현상
독립변수들이 강한 상관관계를 보이는 경우는 회귀분석의 결과를 신뢰하기 어렵다. 상관관계가 높은 독립 변수 중 하나 혹은 일부를 제거하거나 변수를 변형시켜서 해결
독립변수 제거를 통한 정보손실 vs. 다중 공선성 문제 해결 → 판단 필요

귀무가설: 적절성과 친밀도는 제품의 만족도에 영향을 미친다고 볼 수 없다.
대립가설: 적절성과 친밀도는 제품의 만족도에 영향을 미친다고 볼 수 있다.

실습 (다중 회귀분석)

1단계: 변수 모델링

```
y = product$제품_만족도
x1 = product$제품_친밀도
x2 = product$제품_적절성
df <- data.frame(x1, x2, y)
```

2단계: 다중 회귀분석

```
result.lm <- lm(formula = y ~x1 + x2, data = df)
result.lm
```

실습 (다중 공선성(Multicollinearity)문제 확인)

1단계: 패키지 설치

```
install.packages("car")
```

library(car)

car 패키지

2단계: 분산팽창요인(VIF)

vif(result.lm)

vif()함수
https://www.rdocumentation.org/packages/car/versions/3.0-10/topics/vif

분산팽창요인(VIF)값이 10이상인 경우 다중 공선성 문제를 의심
10이 절대값은 아님.

실습 (다중 회귀분석 결과보기)

summary(result.lm)

결과 제시 방법
가설, 분석결과, 가설검정, 회귀모형 결정계수, 수정결정계수, 회귀모형의 적합성, 독립변수 설명

[표 15.4] 제품친밀도와 제품적절성이 제품만족도에 미치는 영향분석
종속변수, 독립변수, 표준오차, 베타, 검정통계량, 유의확률, 분산팽창요인, 분석통계량, 회귀식

2.4 다중 공선성 문제 해결과 모델 성능평가

다중 공선성 문제 해결 방법

실습순서:
다중 공선성 문제 해결 → 회귀모델 생성 → 예측치 생성 → 모델 성능평가

(1) 다중 공선성 문제 해결

다중 공선성 문제: 독립변수 간의 강한 상관관계로 인하여 회귀분석의 결과를 신뢰할 수 없는 현상
강한 상관관계를 갖는 독립변수를 제거하여 해결

실습 (다중 공선성 문제 확인)

1단계: 패키지 설치 및 데이터 로딩

```
install.packages("car")
library(car)

data(iris)
```

car 패키지 설치

2단계: iris 데이터 셋으로 다중 회귀분석

```
model <- lm(formula = Sepal.Length ~ Sepal.Width +
                    Petal.Length + Petal.Width, data = iris)
vif(model)
sqrt(vif(model)) > 2
```

3단계: iris 변수 간의 상관계수 구하기

```
cor(iris[ , -5])
```

상관계수로 변수간의 강한 상관관계 구분

(2) 회귀모델 생성

동일한 데이터 셋을 7:3 비율로 학습데이터와 검정데이터로 표본 추출한 후 학습데이터를 이용하여 회귀모델을 생성

실습 (데이터 셋 생성과 회귀모델 생성)

1단계: 학습데이터와 검정데이터 표본 추출

```
x <-sample(1:nrow(iris), 0.7 * nrow(iris))
train <- iris[x, ]
test <- iris[-x, ]
```

sample()함수 이용하여 70% 데이터 추출하여 학습데이터, 나머지 데이터는 검정데이터로 설정

다중 공선성 문제가 발생하는 Petal.Width변수를 제거한 후 학습데이터를 이용하여 회귀모델 생성

2단계: 변수 제거 및 다중 회귀분석

```
model <- lm(formula = Sepal.Length ~ Sepal.Width + Petal.Length, data = train)
model
summary(model)
```

(3) 회귀방정식 도출

절편, 기울기, 독립변수(x)의 관측치를 이용하여 회귀방정식을 도출

실습 (회귀방정식 도출)

1단계: 회귀방정식을 위한 절편과 기울기 보기

```
model
```

2단계: 회귀방정식 도출

```
head(train, 1)
```

*sample data에 따라서 회귀방정식은 상이할 수 있음.

```
# 다중 회귀방정식 적용(예시)
Y = 2.3826 +   0.5684 * 2.9 + 0.4576 * 4.6
Y
6.6 - Y
```

(4) 예측치 생성

검정데이터를 이용하여 회귀모델의 예측치를 생성.
학습데이터에 의해 생성된 회귀모델을 검정데이터에 적용하여 모델의 예측치를 생성

predict()함수
형식: predict(model, data)
where
model: 회귀모델(회귀분석 결과가 저장된 객체)
data: 독립변수(x)가 존재하는 검정데이터 셋

https://www.rdocumentation.org/packages/car/versions/3.0-10/topics/Predict
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/predict

실습 (검정데이터의 독립변수를 이용한 예측치 생성)

```
pred <- predict(model, test)
pred
```

*데이터에 따라 예측치는 상이할 수 있음.

(5) 회귀모델 평가

모델평가는 일반적으로 상관계수를 이용
모델의 예측치(pred)와 검정데이터의 종속변수(y)를 이용하여 상관계수(r)를 구하여 모델의 분류정확도를 평가한다.
상관관계가 높다면 분류정확도가 높다고 볼 수 있음.

실습 (상관계수를 이용한 회귀모델 평가)

```
cor(pred, test$Sepal.Length)
```

2.5 기본 가정 충족으로 회귀분석 수행

회귀분석은 선형성, 다중 공선성, 잔차의 정규성 등 몇가지 기본 가정이 충족되어야 수행할 수 있는 모수 검정방법

회귀분석의 기본 가정을 충족하는지 확인

실습 회귀분석의 기본 가정 충족으로 회귀분석 수행

1단계: 회귀모델 생성

1-1단계: 변수 모델링

```
formula = Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width
```

1-2단계: 회귀모델 생성

```
model <- lm(formula = formula, data = iris)
model
```

2단계: 잔차(오차)분석

2-1단계: 독립성 검정 – 더빈 왓슨 값으로 확인

```
install.packages('lmtest')
library(lmtest)
dwtest(model)
```

lmtest패키지
더빈 왓슨 값의 p-value가 0.05이상(DW값 1-3범위)이면 잔차에 유의미한 자기 상관이 없다고 볼 수 있다. 즉 독립성이 있다고 볼 수 있다.

2-2단계: 등분산성 검정 – 잔차와 적합값의 분포

```
plot(model, which = 1)
```

잔차(residual) 0을 기준으로 적합값(fitted values)의 분포가 좌우 균등하면 잔차들은 등분산성과 차이가 없다고 볼 수 있다.

2-3단계: 잔차의 정규성 검정

```
attributes(model)
res <- residuals(model)
shapiro.test(res)
par(mfrow = c(1, 2))
hist(res, freq = F)
qqnorm(res)
```

Shapiro.test()함수 이용하여 정규성 검정
hist()함수로 히스토그램과 qqnorm()함수를 통해 Normal Q-Q plot으로 정규성 확인 가능

3단계: 다중 공선성 검사

```
library(car)
sqrt(vif(model)) > 2
```

Petal.Length와 Petal.Width변수 간에는 다중 공선성 문제가 의심스러워 두 변수 중 하나 제거하여 회귀모델 다시 생성

4단계: 회귀모델 생성과 평가

```
formula = Sepal.Length ~ Sepal.Width + Petal.Length
model <- lm(formula = formula, data = iris)
summary(model)
```

Petal.Width제거 후 회귀모델 생성
모델이 유의하고 모델의 설명력(Adjusted R square)가 높음

ch15 연습문제 1-2번