데이터 통계 분석 기초 3

환경생태데이터사이언스 실습 November 26, 2019

오늘의 학습 목표

1. 다변수 회귀분석

다중 선형 회귀분석 (Multiple linear regression)

다중 선형 회귀분석

- 다중 회귀 분석[1]
  - 중선형 회귀, 또는 다변수 선형 회귀 등으로 불림.
  - 여러 독립변수가 사용된 형태의 선형 모형.
  - 다음과 같은 식으로 표현
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$

[1]출처: 서민구 (2014). R을 이용한 데이터 처리 & 분석 실무. 길벗

다중 선형 회귀 모형의 생성

- 선형 회귀 모형과 마찬가지로 lm( ) 함수를 이용하여 생성.
  - 'lm(y ~ x1 + x2, data = yourData)'
  - 붓꽃 자료를 이용한 다중 선형 회귀 분석
    - sepal: 꽃받침
    - petal: 꽃잎
    - Iris setosa : 부채붓꽃
    - Iris versicolor: 북방푸른꽃창포
    - Iris virginica :

```
data(iris)  # Load iris data
str(iris)   # Check structure of this data
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1
```

다중 선형 회귀 모형의 생성

```r
# Create multiple linear model with iris data
irisModel <- lm(Sepal.Length ~
                  Sepal.Width + Petal.Length + Petal.Width,
                  data = iris)
# Check the model
irisModel
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width,
##     data = iris)
##
## Coefficients:
##  (Intercept)   Sepal.Width  Petal.Length   Petal.Width
##       1.8560        0.6508        0.7091       -0.5565
```

다중 선형 회귀 모형 평가

```r
summary(irisModel) # Check detailed information of the model
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width,
##     data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82816 -0.21989  0.01875  0.19709  0.84570
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.85600    0.25078   7.401 9.85e-12 ***
## Sepal.Width   0.65084    0.06665   9.765  < 2e-16 ***
## Petal.Length  0.70913    0.05672  12.502  < 2e-16 ***
## Petal.Width  -0.55648    0.12755  -4.363 2.41e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3145 on 146 degrees of freedom
## Multiple R-squared:  0.8586, Adjusted R-squared:  0.8557
## F-statistic: 295.5 on 3 and 146 DF,  p-value: < 2.2e-16
```

다중 선형 회귀 모형 평가

- `summary`를 통해 나온 통계값
  - Pr(>|t|): t-test를 통해 얻은 검정 값.
    - 통계적으로 절편 혹은 변수의 계수 값이 0인지 아닌지 검정.
    - Pr(>|t|) < 0.05: 절편 혹은 계수가 통계적으로 유의하다.
  - Multiple R-squared: 결정계수
  - Adjusted R-squared: 수정 결정계
  - 다중 회귀 변수에서 F 통계량을 구하기 위한 귀무가설
    - H_0: 모든 계수가 0이다.
    - H_1: 하나 이상의 설명 변수의 계수가 0이 아니다.

범주형 변수를 포함한 다중 선형 회귀 분석

- 범주형 변수를 포함한 분석
  - 범주형 변수도 다른 변수와 마찬가지로 + 기호를 통해 추가.

```
irisModel2 <- lm(Sepal.Length ~
                   Sepal.Width + Petal.Length + Petal.Width + Species,
                   data = iris)
irisModel2
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width +
##     Species, data = iris)
##
## Coefficients:
##       (Intercept)        Sepal.Width       Petal.Length
##            2.1713             0.4959             0.8292
##       Petal.Width  Speciesversicolor   Speciesvirginica
##           -0.3152            -0.7236            -1.0235
```

범주형 변수를 포함한 다중 선형 회귀 분석

- 범주형 변수는 각 요인별로 가변수를 생성하여 분석한다.

Species	Speciesversicolor	Speciesvirginica
setosa	0	0
versicolor	1	0
virginica	0	1

- Iris versicolor의 꽃받침 길이
- $Sepal.Length(versicolor) = 2.171 + 0.496 \cdot Sepal.Width + 0.829 \cdot Petal.Length - 0.315 \cdot Petal.Width - 0.724$
- Iris virgicana의 꽃받침 길이
- $Sepal.Length(virgicana) = 2.171 + 0.496 \cdot Sepal.Width + 0.829 \cdot Petal.Length - 0.315 \cdot Petal.Width - 1.024$
- Iris setosa의 꽃받침 길이
- $Sepal.Length(setosa) = 2.171 + 0.496 \cdot Sepal.Width + 0.829 \cdot Petal.Length - 0.315 \cdot Petal.Width$

범주형 변수를 포함한 다중 선형 회귀 분석 모형의 평가

```r
anova(irisModel2)
```

```
## Analysis of Variance Table
##
## Response: Sepal.Length
##                Df Sum Sq Mean Sq  F value    Pr(>F)
## Sepal.Width     1  1.412   1.412  15.0011 0.0001625 ***
## Petal.Length    1 84.427  84.427 896.8059 < 2.2e-16 ***
## Petal.Width     1  1.883   1.883  20.0055 1.556e-05 ***
## Species         2  0.889   0.444   4.7212 0.0103288 *
## Residuals     144 13.556   0.094
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

다중 회귀 분석을 위한 변수 선택

- 다중 회귀 분석에 필요한 독립변수 선택
  1. 경험을 기반으로 종속변수와 관련이 있는 독립변수 선택.
  2. 변수간 관계를 정확히 모르는 상황에서 회귀분석을 위한 독립변수를 선택.
     - 모든 변수를 선택.
     - 공선성이 있는 변수들을 제거하고 분석 시행.
     - 통계적으로 필수적이면서 최소의 독립변수를 선택.
  - 공선성 확인: `vif` 함수 (`HH` package)
  - 통계적 방법: `step` 함수를 이용하여 통계적으로 변수 선택.

다중 회귀 분석을 위한 변수 선택

- 다중공선성
  - 다중 회귀 분석 $Y \leftarrow X_1, X_2, \cdots, X_p$ 일 때 i 번째 독립변수에 대한 다중 회귀 분석 $X_i \leftarrow X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_p$의 결정계수를 R_i^2라 하자.
  - 분산팽창인자 (Variance Inflation Factor)
    - $VIF_i = \frac{1}{1 - R_i^2}$
  - 특정 독립변수의 분산팽창인자가 크다는 것은 다른 독립변수에 의해 많은 부분이 설명되고 있다는 것을 의미 = 공선성이 존재한다.
  - 일반적으로 5 혹은 10보다 크면 다중공선성이 크다고 판단할 수 있다.

```
# install.packages("HH")
library(HH)
vif(irisModel)
```

```
##  Sepal.Width Petal.Length  Petal.Width
##     1.270815    15.097572    14.234335
```

다중 회귀 분석을 위한 변수 선택

- 변수를 더하거나 빼면서 통계적 수치 비교 (AIC: Akaike information criteria)
- 아카이케의 정리
    - 모형의 예측 정확도에 대한 편향되지 않은 추정값 = 모형이 자료를 얼마나 잘 설명하는지 (설명력) - 모형이 가지고 있는 가변변수의 수 (복잡성)
    - AIC 값은 Kullbak-Leibler 정보 측도를 기반으로 하며 값이 작을수로 더 나은 모형이라 할 수 있다.
- 일반적인 변수 선택법
    - 전진 선택 (forward selection)
    - 후진 선택 (backward selection)
    - 모두 선택 (Bidirectional elimination)

다중 회귀 분석을 위한 변수 선택

```r
nullModel <- lm(Sepal.Length ~ 1, data = iris)
fullModel <- lm(Sepal.Length ~ ., data = iris)
# forwardR <- step(nullModel,
#                  scope = list(lower=nullModel, upper=fullModel),
#                  direction="forward")
# backwardR <- step(fullModel,  direction="backward")
# BidirecR  <- step(nullModel, scope=list(upper=fullModel), direction
```

다중 회귀 분석 결과 시각화

- 공선성 제거 후 모형 (irisModel2)

```
vif(irisModel2)
```

```
##       Sepal.Width      Petal.Length       Petal.Width
##          2.227466         23.161648         21.021401
## Speciesversicolor  Speciesvirginica
##         20.423390         39.434378
```

```
irisModel3 <- lm(Sepal.Width ~ Sepal.Length + Petal.Width + Species, data=iris)
vif(irisModel3)
```

```
##      Sepal.Length       Petal.Width Speciesversicolor
##          3.024496         16.215838          7.652090
##  Speciesvirginica
##         18.490299
```

```
irisModel4 <- lm(Sepal.Width ~ Sepal.Length + Species, data=iris)
vif(irisModel4)
```

```
##      Sepal.Length Speciesversicolor  Speciesvirginica
##          2.622646          2.073395          3.474819
```

다중 회귀 분석 결과 시각화

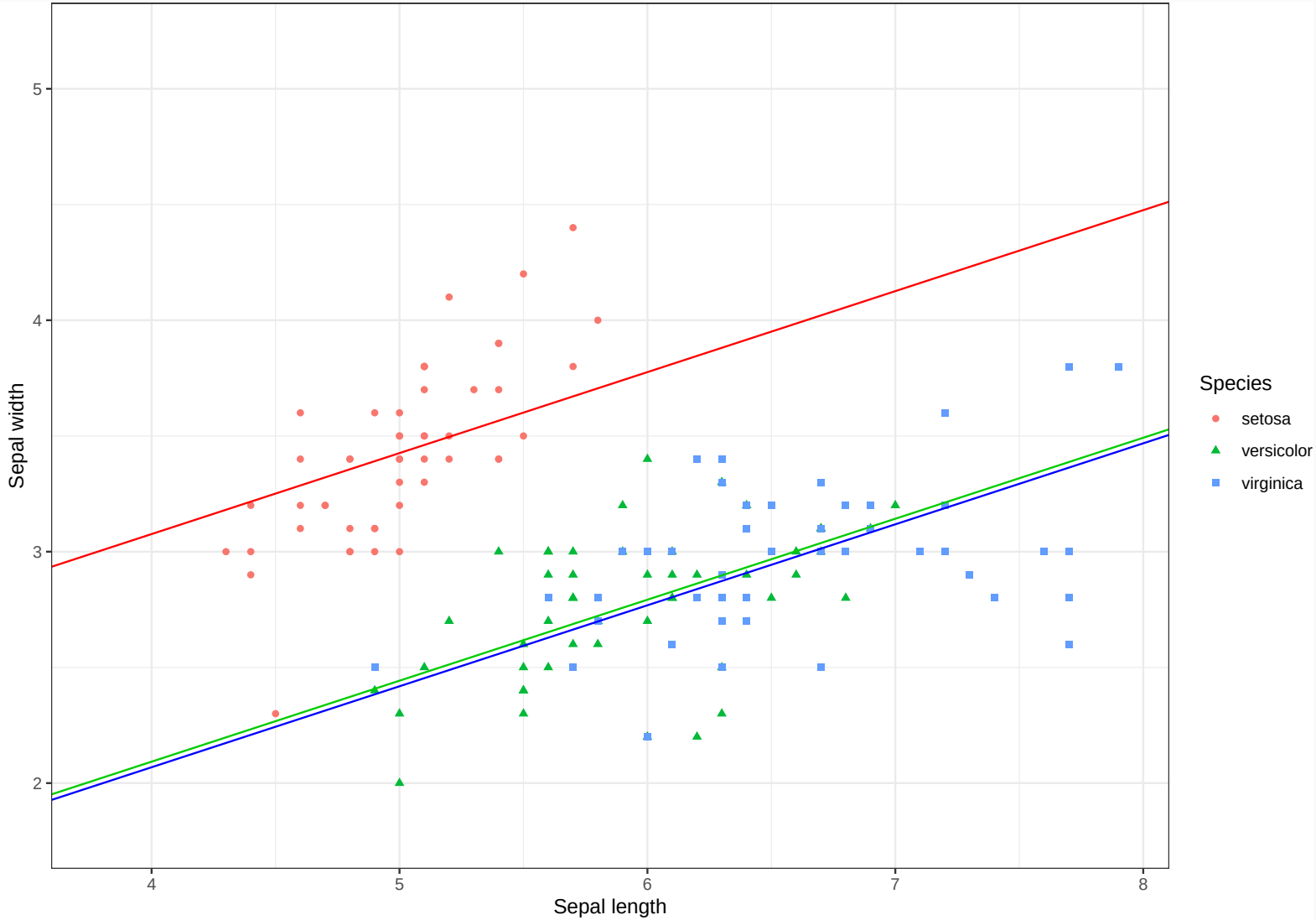
```
# equation
equation_setosa <- function(x){ coef(irisModel4)[1] +
                                coef(irisModel4)[2]*x}
equation_versicolor <- function(x){ coef(irisModel4)[1] +
                                    coef(irisModel4)[2]*x +
                                    coef(irisModel4)[3]}
equation_virgicana <- function(x){  coef(irisModel4)[1] +
                                    coef(irisModel4)[2]*x +
                                    coef(irisModel4)[4]}
# parameters
s_iris <- coef(irisModel4)[2]
i_setosa     <- coef(irisModel4)[1]
i_versicolor <- coef(irisModel4)[1] + coef(irisModel4)[3]
i_virgicana  <- coef(irisModel4)[1] + coef(irisModel4)[4]
```

다중 회귀 분석 결과 시각화

```
irisPlot <- ggplot(iris,
                   aes(y=Sepal.Width, x=Sepal.Length,
                       color=Species, shape=Species, fill=Species))
    geom_point() +
    geom_abline(slope = s_iris, intercept = i_setosa, color=2)+
    geom_abline(slope = s_iris, intercept = i_versicolor, color=3)+
    geom_abline(slope = s_iris, intercept = i_virgicana, color=4)+
    #geom_smooth(method="lm") +
    theme_bw() +
    xlab("Sepal length") + ylab("Sepal width") +
    expand_limits(y=c(1.8,5.2), x=3.8) +
    scale_y_continuous(breaks = 2:5) +
    scale_x_continuous(breaks = 4:8)
```

다중 회귀 분석 결과 시각화

```
irisPlot
```



```
irisPlot2 <- ggplot(iris,
                    aes(y=Sepal.Width, x=Sepal.Length,
                        color=Species, shape=Species, fill=Species))
    geom_point() +
    geom_abline(slope = s_iris, intercept = i_setosa, color=2)+
    geom_abline(slope = s_iris, intercept = i_versicolor, color=3)+
    geom_abline(slope = s_iris, intercept = i_virgicana, color=4)+
    geom_smooth(method="lm") +
    theme_bw() +
    xlab("Sepal length") + ylab("Sepal width") +
    expand_limits(y=c(1.8,5.2), x=3.8) +
    scale_y_continuous(breaks = 2:5) +
    scale_x_continuous(breaks = 4:8)
```

다중 회귀 분석 결과 시각화

```
irisPlot2
```