lab3

JiminLee_2229027

2023-10-02

Consider the dataset bears.txt.

This contains several variables measured on n=141 “bear capturing” occasions, with the following variables: ID: Identification number Age: Bear’s age, in months Month: Month when the measurement was made. Sex. 1 = male 2 = female Head.L: Length of the head, in inches Head.W: Width of the head, in inches Neck.G: Girth (distance around) the neck, in inches Length: Body length, in inches Chest.G: Girth (distance around) the chest, in inches Weight: Weight of the bear, in pounds Obs.No: Observation number for this bear. For example, the bear with ID=41 (Bertha) was measured on four occasions. The value of Obs.No goes from 1 to 4 for these observations Name: The names of the bears given to them by the researchers.

```
setwd('/Users/jimin/Desktop/데스크탑 - 이지민의 MacBook Air/ㅈㅣㅁㅣㄴ/ewha/202
3-2/Regression/week3/')
bears = read.table('bears.txt', header=TRUE)
head(bears, 10)
```

```
##     ID Age Month Sex Head.L Head.W Neck.G Length Chest.G Weight Obs.No
## 1  598  NA     4   1   13.5    7.0   24.5   62.0      41    248      1
## 2  578  NA     4   1   18.5    8.5   23.5   67.5      42    204      1
## 3   83 124     4   1   17.5    8.0   32.0   75.0      55    478      3
## 4  549  16     4   1   11.0    4.0   16.0   50.5      28     90      2
## 5  179 100     4   2   13.0    7.0   21.0   70.0      41    220      1
## 6  555  16     4   1   11.5    5.5   17.5   52.5      30    104      2
## 7  253  51     4   1   13.5    8.0   27.0   68.5      49    360      1
## 8   47  NA     4   1   15.5    7.0   29.3   76.0      53    416      1
## 9  592  NA     4   2   13.0    7.0   21.0   59.0      34    146      1
## 10 592  NA     4   2   15.5    6.0   20.5   60.0      35    152      2
##           Name
## 1       Albert
## 2         Bill
## 3      Charlie
## 4  Christopher
## 5       Fannie
## 6         Gary
## 7         John
## 8       Palmer
## 9      Vanessa
## 10     Vanessa
```

The observations are not independent, because the same bear may have been captured more than once (see variables “Name”, “ID” and “Obs.No”).

For each bear, select only the first observation, so that the new dataset will contain only independent observations. Why is that important for linear regression? How many bears do we have in the dataset?

linear regression 의 데이터는 서로 독립이란 전제에서 성립한다. 중복 관측 데이터가 있으면 독립조건이 성립되지 않는다. 따라서 첫 번째 관측치만을 반영함으로써 독립 조건을 만족한다.

```
bears_indep=bears[bears$Obs.No==1,]
print(nrow(bears_indep)) # 99 : 중복되지 않은 곰 마릿수

## [1] 99

head(bears_indep, 10)

##      ID Age Month Sex Head.L Head.W Neck.G Length Chest.G Weight Obs.No
Name
## 1  598  NA     4   1   13.5    7.0   24.5   62.0      41    248      1  Al
bert
## 2  578  NA     4   1   18.5    8.5   23.5   67.5      42    204      1
Bill
## 5  179 100     4   2   13.0    7.0   21.0   70.0      41    220      1  Fa
nnie
## 7  253  51     4   1   13.5    8.0   27.0   68.5      49    360      1
John
## 8   47  NA     4   1   15.5    7.0   29.3   76.0      53    416      1  Pa
lmer
## 9  592  NA     4   2   13.0    7.0   21.0   59.0      34    146      1 Van
essa
## 11 589  16     4   1   10.0    4.0   15.5   48.0      26     60      1   W
ille
## 12 590  16     4   1   10.0    5.0   15.0   41.0      26     64      1
XRay
## 13 596  NA     4   1   15.5    9.0   29.0   79.0      50    400      1
Zack
## 16 280  53     5   2   12.5    6.0   18.0   58.0      31    144      1   C
lara
```

Consider the variables y=“Weight”, x1=“Chest.G” and x2=“Head.W”. Fit two separate simple regression models for y=“Weight” on x1=“Chest.G”, and y=“Weight” on x2=“Head.W” (you can use the lm function or the equations).

```
y= bears_indep$Weight
x1 = bears_indep$Chest.G
x2 =  bears_indep$Head.W
```

```r
cat('x1 : Chest.G, y : Weight linear regression \n')

## x1 : Chest.G, y : Weight linear regression

cat('-----------------------------------------\n')

## -----------------------------------------

lm.chest = lm(y~x1)
summary(lm.chest)

##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -77.75 -18.32  -0.63  17.22  97.78
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -266.6770    13.2722  -20.09   <2e-16 ***
## x1            12.6462     0.3586   35.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.77 on 97 degrees of freedom
## Multiple R-squared:  0.9276, Adjusted R-squared:  0.9269
## F-statistic:  1244 on 1 and 97 DF,  p-value: < 2.2e-16

cat('x1 : Head.W, y : Weight linear regression \n')

## x1 : Head.W, y : Weight linear regression

cat('-----------------------------------------\n')

## -----------------------------------------

lm.head = lm(y~x2)
summary(lm.head)

##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -186.60  -40.84  -11.71   26.70  223.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -201.697     34.906  -5.778 9.13e-08 ***
## x2            61.482      5.372  11.446  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.61 on 97 degrees of freedom
## Multiple R-squared:  0.5746, Adjusted R-squared:  0.5702
## F-statistic:   131 on 1 and 97 DF,  p-value: < 2.2e-16
```

Do the estimated regression slopes suggest positive or negative relationships? Is there a meaningful interpretation for the regression intercepts?

Both regression line have positive relationships. (12.64, 61.482) regression intercepts : represents the mean value of the response variable when all of the predictor variables in the model are equal to zero. (출처 : https://www.statology.org/intercept-in-regression/) Both regression line are more than -200 when the predictor value are nearly 0.

Using the equation, estimate the variance sigma^2 of the error term for the two models (you can check the result with the summary function, but you need to compute it using the equation).

```
# regression line
chest.weight.line = function(x){
  est.y = -266.677 + 12.6462*x
  return (est.y)
}

head.weight.line = function(x){
  est.y = -201.697 + 61.482*x
  return (est.y)
}
# estimated y
est.weight.c = apply(bears_indep[c('Chest.G')], 2, chest.weight.line)
est.weight.h = apply(bears_indep[c('Head.W')], 2, head.weight.line)

est.sigmaSquare = function(est.y, y, n, p=2){
  sigmaSquare = sum((y - est.y)^2)/(n-p)
  return (sigmaSquare)
}

cat('x = chest, y = weight\n sigma^2 : ')

## x = chest, y = weight
##  sigma^2 :

est.sigmaSquare(est.y = est.weight.c,
                y = bears_indep[c('Weight')],
                n = length(bears_indep),
                p = 2)

## [1] 9182.238
```

```
cat('x = head, y = weight\n sigma^2 : ')

## x = head, y = weight
##  sigma^2 :

est.sigmaSquare(est.y = est.weight.h,
                y = bears_indep[c('Weight')],
                n = length(bears_indep),
                p = 2)

## [1] 53992.09
```

Using the equation, compute the coefficient of determination R2 for both regressions (you can check the result with the summary function, but you need to compute it using the equation). What is their interpretation?

```
cal.R2 = function(y, est.y){
  R2 = 1 - sum((y - est.y)^2) / sum((y - mean(y))^2)
  return (R2)
}

R2.chest = cal.R2(y = y, est.y = est.weight.c)
R2.head = cal.R2(y = y, est.y = est.weight.h)

cat('R2.chest : ', R2.chest,'\n')

## R2.chest :  0.9276481

cat('R2.head : ',R2.head)

## R2.head :  0.5745668
```

chest-weight, head-weight 는 모두 설명 변수와 반응 변수가 선형관계가 있다. 그 중 1 에 더욱 근접한 chest-weight 에 더욱 확실한 선형 관계가 드러난다.

Between x1="Chest.G" and x2="Head.W", which appears to be the best predictor for y="Weight"? (Address this comparing the coefficients of determination R2 of the two regressions).

R^2 의 값이 더욱 1 에 가까운 chest 가 head 보다 더 좋은 설명 변수다.

Fit a multiple linear regression model with predictors x1="Chest.G" and x2="Head.W".

Using the equation, estimate the variance sigma2 of the error term for the new model (you can check the result with the summary function, but you need to compute it using the equation).

```
X = cbind(x1, x2)
summary(lm(y~X))

##
## Call:
## lm(formula = y ~ X)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.365 -17.478   2.572  18.953 100.887
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -277.6130    14.6721  -18.92   <2e-16 ***
## Xx1           11.9565     0.5429   22.02   <2e-16 ***
## Xx2            5.6343     3.3536    1.68   0.0962 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.48 on 96 degrees of freedom
## Multiple R-squared:  0.9297, Adjusted R-squared:  0.9283
## F-statistic: 634.9 on 2 and 96 DF,  p-value: < 2.2e-16
```

```r
X = cbind(1, x1, x2)

est.y = X%*%solve(t(X)%*%X)%*%t(X)%*%y # estimated y 를 계산

est.sigmaSquare(est.y = est.y,
                y = bears_indep[c('Weight')],
                n = length(bears_indep),
                p = 3)
```

```
## [1] 9911.064
```

Using the equation, compute the coefficient of determination R2 for the new regression (you can check the result with the summary function, but you need to compute it using the equation). What is its interpretation?

```r
R2.multi = cal.R2(y = y, est.y = est.y)
R2.multi
```

```
## [1] 0.9297148
```

R^2 의 값이 0.9297148 으로, 거의 1 에 수렴하는 모습을 보인다. 이는 설명 변수 head, chest 와 반응 변수 weight 사이에 선형 관계가 있음을 의미한다.

Do you think this model is better that the one with only x1? Why?

좋다. 더 많은 설명변수를 갖고 있음에도 선형적 관계가 있음을 나타낼 수 있기 때문이다.