Lab4

As you did in Lab3, consider only the first observation for each bear (bears_indep=bears[bears$Obs.No==1,]).

```
bears_df = read.table("/Users/jimin/Desktop/데스크탑 - 이지민의 MacBook Air/ᄌ
ᅵᄆ ᅵᆫ/ewha/2023-2/Regression/week3/bears.txt",header = TRUE)
head(bears_df, 5)

##    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
##          Name
## 1      Albert
## 2        Bill
## 3     Charlie
## 4 Christopher
## 5      Fannie

bears_indep=bears_df[bears_df$Obs.No==1,] # 중복되는 관측값 삭제하기
```

Consider the simple linear regression model with response y="Weight" and predictor x1="Neck.G" (you can fit the model using the lm function or the equations).

```
y = bears_indep$Weight
x1 = bears_indep$Neck.G
model.1 = lm(y~x1)
summary(model.1)

##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.011 -19.446  -3.831  15.644 168.594
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -244.7885    15.4956   -15.8   <2e-16 ***
## x1            20.5900     0.7148    28.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 37.01 on 97 degrees of freedom
## Multiple R-squared:  0.8953, Adjusted R-squared:  0.8942
## F-statistic: 829.7 on 1 and 97 DF,  p-value: < 2.2e-16
```

Provide the estimated regression slope. Does it suggest a positive or negative relationship?

```
coef = model.1$coefficients

cat("y = ",coef[1] ,"+", coef[2],"x + e")

## y =  -244.7885 + 20.58996 x + e

# It has positive relationship  because slope is over 0.
```

• Using the equation, estimate the variance of the errors (i.e. compute sigma2_hat) and the standard error of the estimated slope (i.e. compute se(beta1_hat)). You can check the result with the summary function, but you need to compute it using the equation.

```
RSS.1 = sum(residuals(model.1)^2)
df = length(residuals(model.1)) - 2 # there are 2 parameters
sigma2_hat = RSS.1 / df
cat('the variance of the error : ',sigma2_hat,'\n')

## the variance of the error :  1369.55

X = matrix(x1)
X = cbind(1, X)

se.beta1.hat = sqrt(sigma2_hat*solve(t(X)%*%X)[2,2])
cat('the standard error of the estimated slope : ', se.beta1.hat)

## the standard error of the estimated slope :  0.7148337
```

• Perform a test for the hypotheses: H0: beta1 = 0 vs H1: beta1 != 0. Provide: o Distribution of the test statistic under H0 (type and parameters of the distribution) 자유도가 n-p 인 t 분포를 따른다.

o Value of the test statistics, computed using the equation (you can check the result with the summary function, but you need to compute it using the equation)

```
beta1_hat = coef[2]
test_statistics = beta1_hat / se.beta1.hat
cat("test_statistics : ", test_statistics)

## test_statistics :  28.80384
```

o P-value, computed using the equation (you can check the result with the summary function, but you need to compute it using the equation)

```
pval.1 = 2*pt(abs(test_statistics), df, lower.tail = FALSE)
round(pval.1, 4) # pval of test_statistics
```

```
## x1
##  0
```

o Interpretation of the result. pval 의 크기가 0 이기 때문에 어떤 유의수준에서라도 H0 는 기각된다.

Fit a multiple linear regression model with response y="Weight" and predictors x1=" Neck.G" and x2="Head.W" (you can fit the model using the lm function or the equations).

```
x2 = bears_indep$Head.W
model.2 = lm(y~x1+x2)
summary(model.2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -96.572 -19.677  -4.368  16.749 169.096
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -242.713     17.562 -13.821   <2e-16 ***
## x1            20.840      1.215  17.159   <2e-16 ***
## x2            -1.157      4.527  -0.255    0.799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.19 on 96 degrees of freedom
## Multiple R-squared:  0.8954, Adjusted R-squared:  0.8932
## F-statistic: 410.9 on 2 and 96 DF,  p-value: < 2.2e-16
```

• Using the equation, compute TSS, RSS and SS_reg.

```
TSS = sum((y - mean(y))^2)
RSS = sum(residuals(model.2)^2)
SS_reg = TSS - RSS
cat('TSS : ', TSS,'\nRSS : ',RSS, '\nSS_reg : ',SS_reg)
```

```
## TSS :  1269109
## RSS :  132756.1
## SS_reg :  1136353
```

• Perform a test on all the predictors, i.e. test the hypotheses H0: beta1 = beta2 = 0 vs H1: at least one != 0. Provide: o Distribution of the test statistic under H0 (type and parameters of the distribution) df1 = p-1, df2 = n-p 인 F 분포를 따른다.

o Value of the test statistics, computed using the equation (you can check the result with the summary function, but you need to compute it using the equation)

```
p = 3 # b0,b1,b2
n = length(y) # 99
t_stat = ((SS_reg) / (p-1)) / (RSS / (n-p))
cat("t_stat : ",t_stat)

## t_stat :  410.8658
```

o P-value, computed using the equation (you can check the result with the summary function, but you need to compute it using the equation)

```
pval.2 = pf(t_stat,p-1, n-p, lower.tail = FALSE )
round(pval.2, 5)

## [1] 0
```

o Interpretation of the result. p val 이 0 이기 때문에 모든 유의수준에서 H0 이 기각된다.

Consider a multiple linear regression model with response y="Weight" and predictors x1="Head.L", x2="Head.W", x3="Neck.G", x4="Length" and x5="Chest.G".

```
model.3 = lm(data = bears_indep, Weight~Head.L+Head.W+Neck.G+Length+Chest.G)
summary.2 = summary(model.3)
summary.2

##
## Call:
## lm(formula = Weight ~ Head.L + Head.W + Neck.G + Length + Chest.G,
##     data = bears_indep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.457 -17.969  -2.059  14.432  99.239
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -258.3771    20.8837 -12.372  < 2e-16 ***
## Head.L        -7.5230     3.3596  -2.239   0.0275 *
## Head.W         0.3087     3.3965   0.091   0.9278
## Neck.G         8.5812     1.7639   4.865 4.65e-06 ***
## Length         1.3305     0.7425   1.792   0.0764 .
## Chest.G        7.8844     1.0190   7.738 1.19e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.24 on 93 degrees of freedom
## Multiple R-squared:  0.9456, Adjusted R-squared:  0.9427
## F-statistic: 323.5 on 5 and 93 DF,  p-value: < 2.2e-16
```

After fitting the model, perform a t test for each single beta_i, and a global F test for all predictors to answer to the following questions (you can use the lm function and the summary function). o Is the model significant?

```
beta.i = coefficients(model.3)
print(beta.i)

##  (Intercept)       Head.L       Head.W       Neck.G       Length      Ches
t.G
## -258.3771145   -7.5230466    0.3087305    8.5811503    1.3305341    7.8844
494

cat("\nfstatistic : ", summary.2$fstatistic[1])

##
## fstatistic :  323.4895

cat("\np value : ", 0)

##
## p value :  0
```

모델의 유의수준이 0 임으로 모든 ALPHA 값에서 귀무가설이 기각된다. 따라서 이 모델은 적어도 하나의 설명변수가 반응변수와 선형적 관계가 있다고 해석할 수 있으므로 모델은 significant 하다.

o Which of the predictors are NOT significant in the model at alpha=0.05, if you consider them individually? Head.W, Length 는 significant 하지 못한 설명 변수다. pvalue 가 alpha 값을 넘어 각 설명변수의 파라미터 베타가 0 이라는 가설을 기각하지 못하기 때문이다.

Consider the subset of predictors that are not significant at alpha=0.05. Perform a test to see if they are significant or not, when you consider them simultaneously; i.e. test the hypotheses H0: betaq = ... = betap-1 = 0 vs H1: at least one != 0. Provide: o Distribution of the test statistic under H0 (type and parameters of the distribution) df1 = p-q, df2 = n - p 인 F 분포를 따른다.

o RSS of the full model and RSS of the reduced model, computed using the equation (you can check the result with the anova function, but you need to compute it using the equation)

```
lm_full = lm(data = bears_indep, Weight~Head.L+Head.W+Neck.G+Length+Chest.G)
lm_sub = lm(data = bears_indep, Weight~Head.L+Neck.G+Chest.G)

cal_RSS = function(residual){
  RSS = sum(residual^2)
  return (RSS)
}
```

```
full_RSS = cal_RSS(lm_full$residuals) # 69003.64
sub_RSS = cal_RSS(lm_sub$residuals) # 71386.92
full_RSS

## [1] 69003.64

sub_RSS

## [1] 71386.92
```

o Value of the test statistics, computed using the equation (you can check the result with the anova function, but you need to compute it using the equation)

```
cal_F = function(full_RSS, sub_RSS, df1, df2){
    F.stat = ((sub_RSS - full_RSS) / df1) / (full_RSS / df2)
    return (F.stat)
}
n = length(y)
p = 6
q = 4

F_stat = cal_F(full_RSS, sub_RSS, df1 = p-q, df2 = n - p)
F_stat

## [1] 1.606035
```

o P-value, computed using the equation (you can check the result with the anova function, but you need to compute it using the equation)

```
pval.3 = pf(F_stat, p-q, n-p, lower.tail = FALSE)
pval.3 # pvalue

## [1] 0.2061971
```

o Interpretation of the result. p val 이 0.2 이므로 일반적으로 사용하는 제 1 종 요류의 확률인 1%, 5%, 10%보다 크다. 이는 귀무가설이 기각되지 못한다는 것을 의미한다. 따라서 Head.W, Length 는 significant 하지 못한 설명 변수라고 해석할 수 있다.