1. 단순 회귀분석의 실시

(1) "wage.RData"를 이용하여 다음의 변수를 종속변수와 독립변수로 하는 회귀분석을 실시하고 해당 명령문 script를 별도의 box로 처리하여 보고할 것(3점)

- 종속변수: 임금(wage)
- 독립변수: 연령(age)

```
##1-1 "wage.RData"를 이용하여 다음의 변수를 종속변수와 독립변수로 하는 회귀분석을 실시

model1<- lm(wage~age, data=data)

summary(model1)
```

(2) 단순회귀분석의 결과를 이용하여 아래의 질문에 대한 답을 제시하시오.(10점/각 5점)

가. 모형의 설명력을 나타내는 통계량의 값을 제시하고 그 결과를 해석하시오. 특히, 해당 통계량이 의미하는 것(어떤 모형과 비교한 결과인지를 반드시 제시할 것)

Summary를 통해, 회귀모델에 대한 계수를 산출해낼 수 있다.

```
summary(model1) #F-statistic = 32.05
```

```
> summary(model1) #F-statistic = 32.05

Call:
lm(formula = wage ~ age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-16.907  -6.015  -1.677   3.469  64.348

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.29416    1.85574   3.931 9.70e-05 ***
age          0.35036    0.06189   5.661 2.58e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.618 on 486 degrees of freedom
Multiple R-squared:  0.06186,   Adjusted R-squared:  0.05993
F-statistic: 32.05 on 1 and 486 DF,  p-value: 2.578e-08
```

모형에 대한 설명력을 나타내는 통계량은 F-statistic 결과이다. F-statistic의 경우, f 검정 통계량이라고 불리는데, t-test과 더불어 가설 검정을 위해 사용된다. F 검정량의 경우, 분산비 test이라고도 불리며, 귀무가설의 경우 축소모형이 적절하다, 즉 설명력이 같다. 대립가설의 경우, 완전모형이 적절하다, 즉 설명력이 같지 않다는 것이다. Model1의 경우, f 검정량의 수치가 32.050이다. 유의수준이 0.05(가장 흔하게 사용됨)인 경우, model 1의 p-value가 2.578e-08로 0.05인 유의수준보다 낮기 때문에 귀무가설을 기각한다. 결국, model1은 설명력이 같지 않다, 즉 완전모형이 적절하다는 결론이 도출된다. 결국, wage의 경우, age에 따른 설명력 차이가 있다는 것을 알 수 있다.

모형에 대한 설명력인 F-statistic의 경우, 상수모형과 비교한 것이다. 둘의 비교를 위해, model0를 설정하여, 상수모형을 설정하였다. 그 후, model0과 model1의 anova 검정을 실시하였다.

```
##comparing it with model0
model0<- lm(wage~1,data=data)
anova(model1,model0)  #you can see that F-statistic of model1 & anova's F is same
```

아노바 검정을 실시한 결과값으로, 두 모형(model0, model1) 간의 설명력을 비교할 수 있다.

```
> anova(model1,model0)  #you can see that F-statistic of model1 & anova's F is same
Analysis of Variance Table

Model 1: wage ~ age
Model 2: wage ~ 1
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    486 44959
2    487 47924 -1   -2964.6 32.046 2.578e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

아노바 검정의 f 값으로 32. 046이 도출되었고, 위에 기재된 model 1의 f 검정량의 수치와 동일한 것을 볼 수 있다. (model1의 f 값: 32.05)

나. 임금에 대한 연령이 어떤 효과성이 있는지를 설명할 것. 설명시 반드시 그 이유를 제시할 것

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.29416    1.85574   3.931 9.70e-05 ***
age          0.35036    0.06189   5.661 2.58e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.618 on 486 degrees of freedom
Multiple R-squared:  0.06186,   Adjusted R-squared:  0.05993
F-statistic: 32.05 on 1 and 486 DF,  p-value: 2.578e-08
```

t-검정량의 경우, 모수에 대한 가설 검정으로, 귀무가설이 $H_0 : \beta_1 = \beta_1^0$(두 모집단은 평균 간의 차이가 없다.)이고, 대립가설은, $H_1 : \beta_1 \neq \beta_1^0$(두 모집단은 평균 간의 차이가 있다.)이다. T 검정량의 경우에도, p-value로서 귀무가설과 대립가설을 채택할 수 있다.

summary로 model1의 계수를 다시 확인해보면, age이 한 살 증가할 때, 임금은 0.35036 증가할 것으로 기대된다는 것을 알 수 있다. (Age과 wage의 관계를 살펴보면, wage=7.29+0.35*age로 도출되는 것을 알 수 있다.) 이 경우, t 값의 p-value 값이 유의수준(0.05)보다 작아, 귀무가설을 기각하고 대립가설을 채택한다. 즉, age가 증가할 때, wage에 대한 차이가 있다고 할 수 있고, age는 유의한 변수로 정의된다. 결국 model1의 경우, age는 유의미한 변수로서 wage를 설명할 수 있다.

2. 다중 회귀분석의 실시

(1) "wage.RData"를 이용하여 다음의 변수를 종속변수와 독립변수로 하는 회귀분석을 실시하고 해당 명령문 script를 별도의 box로 처리하여 보고할 것(3점)

- 종속변수: 임금(wage)
- 독립변수: 연령(age), 경력(tenure)

```
##2-1 "wage.RData"를 이용하여 다음의 변수를 종속변수와 독립변수로 하는 회귀분석을 실시
model2<- lm(wage~age+tenure, data=data)
summary(model2) ##F-statistic=60.67
```

(2) 임금에 대한 연령과 경력이 어떤 효과성이 있는지를 설명할 것. 설명시 반드시 그 이유를 제시할 것(10점)

```
> summary(model2) ##F-statistic=60.67

Call:
lm(formula = wage ~ age + tenure, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-22.978  -5.088  -1.693   3.459  60.912

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.28568    1.79986   6.826 2.62e-11 ***
age          0.07489    0.06464   1.159    0.247
tenure       1.10348    0.12052   9.156  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.89 on 485 degrees of freedom
Multiple R-squared:  0.2001,    Adjusted R-squared:  0.1968
F-statistic: 60.67 on 2 and 485 DF,  p-value: < 2.2e-16
```

Model2의 summary를 살펴보면, age의 경우, age가 한 살 증가할 때(tenure가 고정되어 있다고 가정하고), wage가 0.074 증가한다. Age의 t 값의 p value는 0.24로 유의수준인(0.05) 보다 크기에, 귀무가설이 채택된다. 즉, model2의 경우, age에 따른 wage의 차이가 없음을 알 수 있다. Tenure의 경우, tenure가 한 단위 증가할 때, (age가 고정되어 있음을 가정) wage가 1.10 증가하는 것을 알 수 있다. 이 경우의 pvalue는 2e-16보다 작은 값으로 대립가설을 채택한다. 즉, tenure에 따른 wage 차이가 있음을 알 수 있고, tenure은 유의미한 변수임을 알 수 있다.

3. 단순회귀분석과 다중회귀분석 간 모형의 비교

(1) 단순회귀분석과 다중회귀분석 결과를 각기 다른 객체로 저장해서 두 모형이 임금에 대해 가지고 있는 설명력(분산)에 대한 검증을 실시하시오.(4점)

```
anova(model1,model2)
```

(2) 검증 결과를 토대로 단순회귀분석과 다중회귀분석 중 어떤 모형이 더 유의미한지 설명하시오. 설명 시 그 근거를 제시하시오. 다만, 분산분석 결과만을 해석하지 말고 독립변수들의 회귀계수의 유의도를 종합적으로 판단하여 결론을 도출할 것(10점)

```
> anova(model1,model2)
Analysis of Variance Table

Model 1: wage ~ age
Model 2: wage ~ age + tenure
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    486 44959
2    485 38333  1    6626.1 83.834 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova 검정의 결과를 통해, model2(다중회귀분석)와 model1(단순회귀분석)은 p-value값이 2.2e-16으로 유의수준(0.05)보다 낮기 때문에 대립가설을 채택하여, 두 모형의 설명력에 차이가 있음을 가시적으로 확인할 수 있다. 그러므로, tenure의 변수가 존재하는 것이 유의미함을 알 수 있다. 앞선 모델에 대한 f 값만을 비교를 한다면, model1은 f값이 32.05, model2은 f값이 60.67로 model2가 단순히 F값이 높기 때문에 model2가 더 유의미해 보일 수 있다. 하지만, 통상적으로 f 값은 독립변수의 수에 영향을 받는다. 이 때문에, 독립변수의 개수가 많으면 많을수록, f 값은 자연스레 높아진다. 따라서, 독립변수가 더 많은 model2가 f 값이 높을 수밖에 없다. 이런 이유로, 단순한 f 값 비교만으로는 유의미한 모형을 선정할 수 없고 anova검정을 실시하는 것이다.

Model1의 경우, 단순회귀분석으로 age는 유의미한 변수로 도출되었고, model2의 경우, 다중회귀분석으로 tenure만 유의미한 변수로 도출되었다. 다중회귀의 경우인 model2가 더 유의미한 모형임을 알 수 있다.

(+)유의미한 모형을 선정하는 데에는 기준이 필요하다. 한 기준으로는, adjusted R square값을 이용하는 것이 있다. adjusted R square의 값을 비교한다면, model1의 경우, 0.05, model2의 경우 0.19로 model2가 더 유의미함을 알 수 있다. 또다른 기준으로는 AIC를 비교하는 것이다. AIC 값이 적은 모형, 즉 적은 변수를 가지고 적절한 적합성을 보이는 모형이 선호되는 것이다. Model2가 AIC값이 낮음을 알 수 있다.

```
> AIC(model1,model2)
       df      AIC
model1  3 3598.206
model2  4 3522.399
```