Assignment 2

Jiseong Yang

2018년 12월 26일

1. 단순회귀분석

(1) 임금(wage)를 종속변수, 연령(age)를 독립변수로 하는 단순 회귀 분석을 실시하고 해당 명령문 script를 별도의 box로 처리하여 보고

```
# Create a model
mod_sim = lm(wage ~ age, data = data)
```

(2) 단순회귀분석의 결과를 이용하여 다음 질문에 대한 답 제시

- 단순회귀분석 결과

```
summary(mod_sim)
```

```
##
## 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
```

가. 모형의 설명력을 나타내는 통계량의 값을 제시하고 그 결과 해석

- t값과 p-value
  - 독립변수 "age"의 검정통계량 t값은 5.661로서 순수한 오차에 대한 해당 독립변수의 영향력의 비율을 나타낸다.
  - t값의 절대값보다 더 큰 값이 t분포에서 관측될 확률이 바로 p-value로서 그 값은 2.58e-08(0.0000000258)이다.
  - t값이 임계값보다 커 기각역에 위치하고, p-value가 유의 수준보다 작기 때문에 이는 해당 독립변수의 영향력이 단순한 우연의 결과라고 보기는 힘들다는 것을 의미한다.
  - 따라서 독립변수의 영향력이 없다는 영가설을 기각하고 독립변수가 유의하다는 대립가설을 채택한다.
- 결정계수와 수정된 결정계수
  - R^2의 값은 0.06186으로서 총 제곱합 중에서 이 회귀 모델에 의해서 설명되는 제곱합의 비율을 의미하며, 총 변동량 중에서 이 모델이 설명하는 부분은 약 6%정도 된다는 의미이다.
  - 수정된 R^2는 독립변수가 증가하면 같이 높아지는 R^2의 특성으로 인하여, 그것을 보정해주기 위해 자유도로 나눠준 것이다.
- F값과 p-value
  - F값은 32.05이고 그 p-value는 2.578e-08(0.00000002578)이기 때문에 기각역에서 검정통계량이 관측되었음을 알 수 있다. 이는 이 모델에 비하여 상수 밖에 없는 모델이 더 유의하다는 영가설을 기각할 수 있게 한다. 따라서 **상수 모델**보다는 이 모델이 더 유의하다는 대립가설을 채택한다.

나. 임금에 대한 연령이 어떤 효과성이 있는지 설명

- 회귀식
  - 임금(wage) = 7.29416 + 0.35036 X 연령(age)
- 해석
  - 회귀식의 계수는 독립변수의 영향력(가중치)를 의미한다.
  - 독립변수(연령)이 1단위 증가할 때마다 종속변수(임금)은 0.35036만큼 증가한다.

2. 다중회귀분석

(1) 임금(wage)를 종속변수, 연령(age)과 경력(tenure)을 독립변수로 하는 다중 회귀 분석을 실시하고 해당 명령문 script를 별도의 box로 처리하여 보고

```
# Create a model
mod_mul = lm(wage ~ age + tenure, data = data)
```

- 다중회귀분석 결과

```
summary(mod_mul)
```

```
##
## 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
```

(2) 임금에 대한 연령과 경력이 어떤 효과성이 있는지 설명

- 회귀식
  - 임금(wage) = 12.28568 + 0.07489 X 연령(age) + 1.10348 X 경력(tenure)
- 가설검정
  - 독립변수1(연령)의 t값은 1.159이고, p-value는 0.247로서 유의하지 않다.
  - 독립변수2(경력)의 t값은 9.156이고, p-value는 2e-16로서 유의하다.
- 해석
  - 회귀식의 계수는 독립변수의 영향력(가중치)를 의미한다.
  - 독립변수1(연령)은 종속변수(임금)에 유의미한 영향을 주지 않는다.
  - 독립변수2(경력)이 1단위 증가할 때마다 종속변수(임금)은 1.10348만큼 증가한다.

3. 단순회귀분석과 다중회귀분석 간 모형의 비교

(1) 단순회귀분석과 다중회귀분석 결과를 각기 다른 객체로 저장해서 두 모형이 임금에 대해 가지고 있는 설명력(분산)에 대한 검증 실시

```
# ANOVA
anova(mod_sim, mod_mul)
```

```
## 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
```

(2) 검증결과를 토대로 단순회귀분석과 다중회귀분석 중 어떤 모형이 더 유의미한지 설명

- 분산분석 결과 해석
  - 가설
    - 영가설: 1개의 독립변수(age)만을 취하는 모델이 2개의 독립변수(age, tenure)를 취하는 모델보다 유의미하다. (축소모형이 적절하다)
    - 대립가설: 1개의 독립변수(age)만을 취하는 모델이 2개의 독립변수(age, tenure)를 취하는 모델보다 유의미하지 않다. (완전모형이 적절하다)
  - 해석
    - 영가설의 검정통계량인 F값이 83.834로서 기각역 내에서 관측되고 p-value가 2.2e-16으로 유의수준보다 작아 통계적으로 유의하다.
    - 따라서 영가설을 기각하고 대립가설을 채택한다. (완전모형이 더 유의하다)
- 모델 비교

	독립변수 개수	유의미한 독립변수	t값	t값 유의확률	결정계수	수정된 결정계수	F값	F값 유의확률	MSE
단순회귀모델	1	연령(age)	5.661	2.58e-08	0.0619	0.0599	32.05	2.578e-08	93.8544
다중회귀모델	2	경력(tenure)	2.793	0.00542	0.2033	0.1983	41.16	2.2e-16	80.316

  - 단순회귀모델와 다중회귀모델은 독립변수의 개수가 각각 1개와 2개로 차이가 있으나, t검정 결과 유의미한 독립변수는 두 모델 모두 한 개이다.
  - 유의 수준 0.001에서 단순회귀모델의 유의미한 독립변수(연령)은 유의하나, 다중회귀모델의 유의미한 독립변수(경력)은 그렇지 않다. 그러나 유의수준 0.01에서는 모두 유의하다.
  - F값과 그 유의확률을 보았을 때 두 모델 모두 독립변수의 영향력이 0이라고 가정한 상수모델보다는 유의미한 것으로 확인된다.
  - 결정계수와 수정된 결정계수를 보았을 때 다중회귀모델은 단순회귀모델보다 3배 이상의 높은 설명력을 가지고 있다.
  - 8:2의 비율로 훈련데이터와 검정데이터를 분할한 뒤 테스트를 하여 계산된 제곱합 평균(Mean Squred Error, MSE)은 대체로 다중회귀모델이 단순회귀모델보다 낮다.
- 결론
  - 다중회귀모델이 단순회귀모델보다 독립변수가 1개가 더 많지만 모든 변수들이 유의미한 것은 아니기 때문에 모델을 선택하기에 앞서 다른 측면도 고려해야한다.
  - 단순회귀모델이 더 적은 변수를 가지고 원본 데이터를 설명하기 때문에 더 효율적이라고 볼 수도 있을 것이다.
  - 그러나 분산분석 결과와 결정계수 및 제곱합 평균(MSE) 등을 종합적으로 고려하였을 때, 다중회귀모델이 더욱 유의하고 원본 데이터를 더 잘 설명한다고 할 수 있다.