# 로지스틱 회귀분석 보고서

**202255651 채수원**

## 1. 분석 목적

본 분석의 목적은 1974년도에 제시된 "여성은 가정을 보살피고 국가를 운영하는 일은 남자에게 맡겨두어야 한다."라는 주장에 대한 찬성 비율을 예측하는 것입니다. 주어진 데이터는 각 응답자의 **성별**과 **교육 기간**을 바탕으로 찬성 비율(ratio)을 설명하려고 합니다. 이를 위해 로지스틱 회귀분석을 사용하여 성별과 교육 기간이 찬성 비율에 미치는 영향을 분석했습니다.

## 2. 분석 방법

### 2.1. 데이터 설명

주어진 데이터는 41명의 응답자로부터 수집된 자료입니다

**education**: 응답자의 교육 기간 (연 단위)

**sex**: 응답자의 성별 (여성: Female, 남성: Male)

**agree**: 찬성한 응답자의 수

**disagree**: 반대한 응답자의 수

**ratio**: 찬성 비율 (찬성 인원 / 전체 응답자 수)

```
data_wrole = sm.datasets.get_rdataset("womensrole", package="HSAUR")
df_wrole = data_wrole.data
df_wrole["ratio"] = df_wrole.agree / (df_wrole.agree + df_wrole.disagree)
df_wrole.tail()
```

595ms 2024.12.02 04:21:17에 실행되었습니다

5 행 5 열 x 5 행

| | education | sex | agree | disagree | ratio |
|---|---|---|---|---|---|
| 37 | 16 | Female | 13 | 115 | 0.101562 |
| 38 | 17 | Female | 3 | 28 | 0.096774 |
| 39 | 18 | Female | 0 | 21 | 0.000000 |
| 40 | 19 | Female | 1 | 2 | 0.333333 |
| 41 | 20 | Female | 2 | 4 | 0.333333 |

### 2.2. 로지스틱 회귀분석

로지스틱 회귀분석은 찬성/반대 와 같은 이진 분류에 적합합니다. 이번 분석에서는 찬성 비율(ratio)을 **이진 종속 변수**로 모델링하여, 성별(sex)과 교육 기간(education)이 찬성 비율에 미치는 영향을 분석했습니다. **찬성 비율**을 종속 변수로, **성별**과 **교육 기간**을 독립 변수로 설정하여 회귀 모델을 구축하였습니다.

이번 분석의 로지스틱 회귀모델은 다음과 같이 표현됩니다.

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot \text{sex}+\beta_2 \cdot \text{education})}}$$

독립 변수들(sex, education)이 찬성확률에 영향을 끼칠 조건부확률을 나타냅니다.

### 2.3. 모델 적합도

- **회귀 분석 방법**: 최대우도법(Maximum Likelihood Estimation, MLE)

  MLE는 주어진 데이터에서 모델의 파라미터 값을 추정하기 위해 사용되는 방법입니다.

$$LL(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

  여기서 p = $\frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{education})}}$

- **모델의 설명력** - Pseudo $R^2$ 활용.

  이진 분류에서는 $R^2$를 활용할 수 없기 때문에 Pseudo $R^2$를 활용합니다. 일반적으로 값이 클수록 모델이 데이터를 잘 설명한다고 판단합니다. 로지스틱 회귀 모델에서 일반적으로 0.2~0.4 사이면 좋은 모델로 간주됩니다.

  다른 평가 지표와 함께 보조적으로, 모델 비교 및 설명력 평가에 사용됩니다.

- **모델 유의성 검정** - LLR p-value 활용.

  LLR p-value는 우도 비율 검정의 결과로써, 기본 모델(절편만 포함된 모델)과 비교하여, 독립 변수들이 모델에 추가될 때 유의미한 개선을 가져오는지 평가하는 데 사용됩니다.

# 3. 분석 결과

## 3.1. 첫 번째 로지스틱 회귀분석 결과

첫 번째 분석에서는 성별과 교육 기간을 모두 독립 변수로 포함한 모델을 사용했습니다.

```
model_wrole = sm.Logit.from_formula("ratio ~ education + sex", df_wrole)
result_wrole = model_wrole.fit()
print(result_wrole.summary())
```

15ms 2024.12.02 04:21:17에 실행되었습니다

```
Optimization terminated successfully.
         Current function value: 0.448292
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                  ratio   No. Observations:                   41
Model:                          Logit   Df Residuals:                       38
Method:                           MLE   Df Model:                            2
Date:                Mon, 02 Dec 2024   Pseudo R-squ.:                  0.3435
Time:                        04:21:17   Log-Likelihood:                -18.380
converged:                       True   LL-Null:                       -27.997
Covariance Type:            nonrobust   LLR p-value:                 6.660e-05
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.0442      0.889      2.299      0.022       0.302       3.787
sex[T.Male]   -0.1968      0.736     -0.267      0.789      -1.640       1.247
education     -0.2127      0.071     -2.987      0.003      -0.352      -0.073
==============================================================================
```

| 변수 | 회귀계수(coef) | 표준오차(std err) | z-value | p-value | 신뢰구간(95%) |
|---|---|---|---|---|---|
| Intercept | 2.0442 | 0.889 | 2.299 | 0.022 | [0.302, 3.787] |
| sex[T.Male] | -0.1968 | 0.736 | -0.267 | 0.789 | [-1.640, 1.247] |
| education | -0.2127 | 0.071 | -2.987 | 0.003 | [-0.352, -0.073] |

- LLR p-value - 6.660e-05로 매우 작아, 모델이 통계적으로 유의미하다는 것을 나타냅니다.
- Pseudo $R^2$ - 0.3435로써 로지스틱 회귀모델의 적정 범위인 0.2~0.4 사이에 있으므로 적절한 설명력을 가지고 있다고 평가됩니다.
- 성별 (sex[T.Male]) – p-value가 0.789로 0.5보다 크며, 신뢰구간에 0이 포함되므로 성별은 찬성 비율에 미치는 영향이 통계적으로 유의미하지 않다고 볼 수 있습니다. 따라서 성별 변수를 모델에서 제외하였습니다.
- 교육 기간 (education) – p-value가 0.003으로써 0.5보다 작고, 신뢰구간에 0이 포함되지 않으며, 신뢰구간이 음수 구간에만 존재하기 때문에 매우 강력한 음의 통계적 상관성을 나타내고 있습니다.

## 3.2. 두 번째 로지스틱 회귀분석 결과

1차 회귀분석 결과에 따라 독립변수에서 성별을 제외하고 두 번째 로지스틱 회귀분석을 진행하였습니다.

```
model_wrole2 = sm.Logit.from_formula("ratio ~ education", df_wrole)
result_wrole2 = model_wrole2.fit()
print(result_wrole2.summary())
```

16ms 2024.12.02 04:21:17에 실행되었습니다

```
Optimization terminated successfully.
         Current function value: 0.449186
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                  ratio   No. Observations:                   41
Model:                          Logit   Df Residuals:                       39
Method:                           MLE   Df Model:                            1
Date:                Mon, 02 Dec 2024   Pseudo R-squ.:                  0.3422
Time:                        04:21:17   Log-Likelihood:                -18.417
converged:                       True   LL-Null:                       -27.997
Covariance Type:            nonrobust   LLR p-value:                 1.202e-05
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.9345      0.781      2.478      0.013       0.405       3.464
education     -0.2117      0.071     -2.983      0.003      -0.351      -0.073
==============================================================================
```

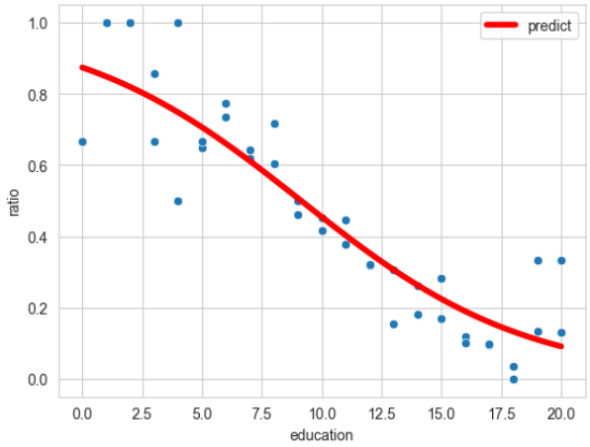
| 변수 | 회귀계수(coef) | 표준오차(std err) | z-value | p-value | 신뢰구간(95%) |
|---|---|---|---|---|---|
| Intercept | 1.9345 | 0.781 | 2.478 | 0.013 | [0.405, 3.464] |
| education | -0.2117 | 0.071 | -2.983 | 0.003 | [-0.351, -0.073] |

- LLR p-value - 1.202e-05로 매우 작아, 모델이 통계적으로 유의미하다는 것을 나타냅니다.
- Pseudo $R^2$ - 0.3422로써 로지스틱 회귀모델의 적정 범위인 0.2~0.4 사이에 있으므로 적절한 설명력을 가지고 있다고 평가됩니다.
- 교육 기간 (education) – p-value와 신뢰구간은 1차 모델과 비교하여 거의 변동이 없었습니다. 또한, 회귀계수 -0.2127은 교육 기간이 1년 증가할 때 찬성 비율의 로그 오즈가 -0.2127만큼 감소한다는 것을 의미합니다. 즉, 교육 기간이 길어질수록 찬성 비율이 낮아지는 경향이 나타납니다.

$$p = \frac{1}{1+e^{-1.9345}} \approx \frac{1}{1+0.145} \approx 0.875$$

- 절편 (Intercept) - 교육 기간이 0일 때 찬성 비율의 로그 오즈가 1.9345임을 나타냅니다. 이는 교육 기간이 0일 때 87.5%의 확률로 찬성할 수 있음을 알 수 있습니다.

## 4. 결론




<- 실제 데이터 산포도 + 완성된 모델 예측 그래프

본 로지스틱 회귀분석 모델은 교육 기간과 성별이 1974년도에 제시된 주장에 대한 찬성 비율에 미치는 영향을 분석했습니다.

- **성별의 영향**: 첫 번째 모델에서 성별 변수의 p-value는 0.789로, 이는 통계적으로 유의미하지 않다는 것을 의미합니다. 신뢰구간 또한 0을 포함하고 있어, 성별은 모델에서 제외되었습니다.
- **교육 기간의 영향**: 교육 기간은 두 모델 모두에서 유의미한 부정적인 상관관계를 보였습니다. 회귀계수는 -0.2127로 교육 기간이 1년 증가할 때 찬성 비율의 로그 오즈가 0.2127만큼 감소한다는 것을 나타냅니다. 이는 교육기간이 1년 증가할 때 찬성 비율이 약 14.7% 감소하는 것을 의미합니다.
- **모델 설명력 및 유의미성**: Pseudo $R^2$ 값은 0.3422로, 로지스틱 회귀모델의 적정 범위인 0.2~0.4 사이에 해당하므로 모델이 적절한 설명력을 가지고 있다고 평가할 수 있습니다. LLR p-value값은 0.05 이하이며, 매우 작다는 점에서 두 모델 모두에서 유의미하다고 평가할 수 있습니다.

결론적으로 교육 기간과 1974년도의 주장의 찬성 비율은 강력한 음의 통계적 상관관계를 나타냈으며, 성별은 유의미한 통계적 상관관계가 없었다는 점을 알 수 있었습니다.

## 6. 참고 문헌 및 데이터셋


- 데이터 셋 출처 - HSAUR 패키지 (womensrole dataset)
- 데이터 사이언스 스쿨, 로지스틱 회귀분석. (https://datascienceschool.net/intro.html)