**회귀분석**

1. 회귀분석

변수 간의 관계를 모델링하고 예측하는 통계적 기법 중 하나로 주로 연속형 종속 변수와 1개 이상의 범주형/연속형 독립 변수 간의 관계를 파악하고 설명하는 데 사용

1) 회귀분석 유형

|  |  |  |  |
| --- | --- | --- | --- |
| 유형 | 종속변수 | 독립 변수 | 종속 변수와 독립 변수 간 관계의 선형성 |
| 단순 선형 회귀  (simple linear regression) | 1개 | 1개 | 선형 |
| 다중 선형 회귀  (multiple linear regression) | 1개 | 2개 이상 | 선형 |
| 일반 선형 모형  (general linear model) | 2개 이상 | 1개 이상 | 선형 |
| 비선형 회귀  (non-linear regression) | 1개 | 1개 이상 | 비선형 |

2. 단순 선형 회귀(simple linear regression)

- : Independent and Identically Distributed random variable

- 회귀계수(regression coefficient)

: 기울기(slope) , : 절편(intercept)

- : 번째 관측치 오차

**\* 오차항의 4가지 가정**

① 선형성(linearity) : 종속 변수 와 독립변수 는 선형 관계에 있음

② 독립성(independence) : 각각의 오차항은 서로 독립

③ 등분산성(constant variance) : 오차항의 분산은 모두 같음

④ 정규성(normality) : 오차항은 정규분포를 따름

1) 회귀계수 추정(OLS)

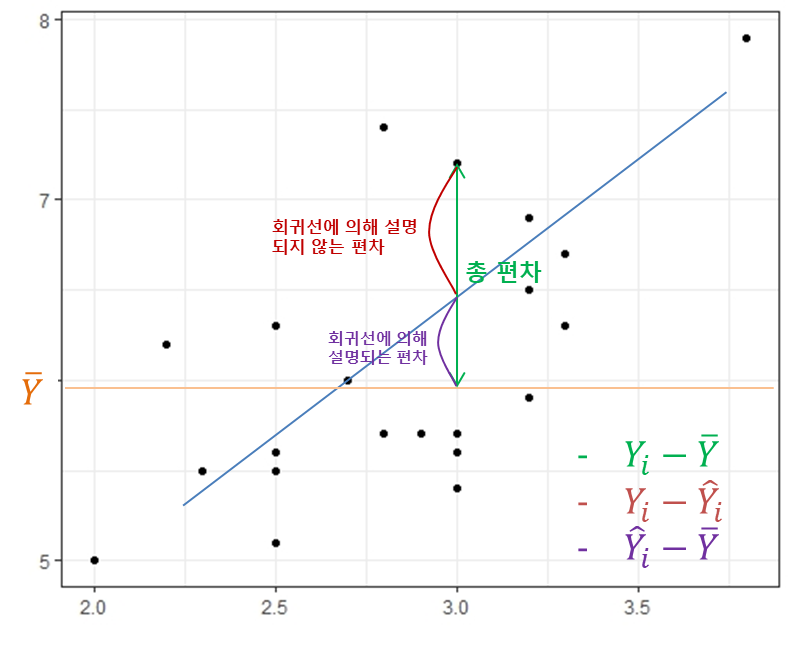
- 최소제곱법(least square method)

관측된 데이터와 모델의 예측값 간의 오차의 제곱을 최소화하여 최적의 회귀계수를 찾는 방법

오차의 제곱합 식인 을 각 회귀계수에 대해 편미분하여 그 값이 0이 되는 과 의 값을 도출하면 아래와 같음

추정된 회귀계수를 통해 회귀식을 추정하면 아래와 같음

**\* 잔차(residual)**

실제 오차를 알 수 없기 때문에 잔차를 이용해 모형을 평가하고 가정을 검토함

TSS = SSR = SSE =

TSS(total sum of squares) = SSR(regression sum of squares) + SSE(residual sum of squares)

[표] 유의성 검정을 위한 분산 분석 표

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 요인(Factor) | 자유도  (Degree of Freedom) | 제곱합  (Sum of Square) | 평균제곱  (Mean Square) | F-value |
| 회귀(regression) | 1 | SSR |  |  |
| 잔차(Residual) | n-2 | SSE |  |  |
| 총(Total) | n-1 | SST |  |  |

검정통계량 : 기각역 :

2) 유의성 검정

[1] 모형 유의성 검정

회귀식에서 기울기()가 0이면 모형이 유의하지 않음. 이를 확인하기 위해 가설검정을 하면 귀무가설과 채택가설이 아래와 같음

vs

통계패키지 결과 Analysis of Variance table을 얻을 수 있음

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Analysis of Variance | | | | | |
| Source | DF | Sum of Squares | Mean Square | F value | Pr>F |
| Model | 1 | SSR | MSR | F | p-value |
| Error | n-2 | SSE | MSE |  |  |
| Corrected Total | n-1 | SST |  |  |  |

p-value < 유의수준()이면 귀무가설 기각 ⇒ 모형이 통계적으로 유의함

[2] 회귀계수 유의성 검정

각 변수의 중요도 및 유의성 파악 가능하며 이를 확인하기 위해 가설검정을 하면 귀무가설과 채택가설이 아래와 같음

vs

검정통계량 : (단, ) 기각역 :

통계패키지 결과 Parameter Estimates table을 얻을 수 있음

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameter Estimates | | | | | |
| Variable | DF | Parameter Estimate | Standard Error | t-value | Pr>|t| |
| Intercept | 1 |  | S.E | T | p-value |
| 변수명 | 1 |  | S.E | T | p-value |

변수의 p-value < 유의수준()이면 귀무가설 기각 ⇒ 회귀계수가 통계적으로 유의함

3) 회귀모형 평가

[1] 평균제곱근 오차(Root Mean Square Error)

: 회귀선을 기준으로 실제값이 평균적으로 얼마나 벗어나 있는지 나타내는 값으로 작을수록 좋은 모형임

**\* RMSE는 값의 단위에 영향을 받으므로 절대적인 측도는 아님**

[2] 결정계수(coefficient of determination)

: 총 편차 중에 회귀선에 의해 설명되는 편차의 비율로 1에 가까울수록 모형의 설명력이 높음

**\* 이면 총 편차 중 80%가 모형에 의해 설명할 수 있음**

3. 다중 선형 회귀(multiple linear regression)

**\* 다중 선형 회귀분석 고려사항**

① 다중공선성(multicollinearity)

: 독립변수 간의 선형 종속이 심한 경우로 차원 축소나 Ridge regression 등을 통해 해결

- VIF가 10보다 크면 다중공선성 문제가 있는 것으로 판단

② 범주형 설명 변수(qualitative predictors)

: 범주형 설명 변수를 어떻게 다룰 것인가에 관한 문제

③ 변수 선택(variable selection)

: 많은 변수로 인한 모형 과적합(overfitting) 문제를 변수 제거를 통해 해결

1) 유의성 검정

[1] 모형 유의성 검정

회귀식에서 기울기가 0이면 모형이 유의하지 않음. 이를 확인하기 위해 가설검정을 하면 귀무가설과 채택가설이 아래와 같음

vs

[표] 유의성 검정을 위한 분산 분석 표

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 요인(Factor) | 자유도  (Degree of Freedom) | 제곱합  (Sum of Square) | 평균제곱  (Mean Square) | F-value |
| 회귀(regression) | p | SSR |  |  |
| 잔차(Residual) | n-p-1 | SSE |  |  |
| 총(Total) | n-1 | SST |  |  |

검정통계량 : 기각역 :

F-test에서 p-value < 유의수준()이면 귀무가설 기각 ⇒ 모형이 통계적으로 유의함

2) categorical 변수를 포함한 모형

Categorical 변수 내용 추가 필요

4. 회귀모형의 가정 진단

|  |  |  |
| --- | --- | --- |
| 구분 | 진단 방법 | 해결 방안 |
| 선형성(linearity) | 산점도, 상관계수 | 변수 변환, 비선형 모형 적합 |
| 독립성  (independence) | Durbin-Watson,  ACF(Auto Correlation Function),  잔차 그래프 | ARMA와 같은 시계열 모형 이용 |
| 정규성(normality) | 첨도와 왜도, Q-Q plot, 정규성 검정(Shapiro-Wilk or K-S test) | 변수 변환, 새로운 변수 투입, 모형 수정 |
| 등분산성  (constant variance) | 잔차 등분산 그래프, White test 등 | 변수 변환, 가중회귀분석 |

1) 선형성 진단

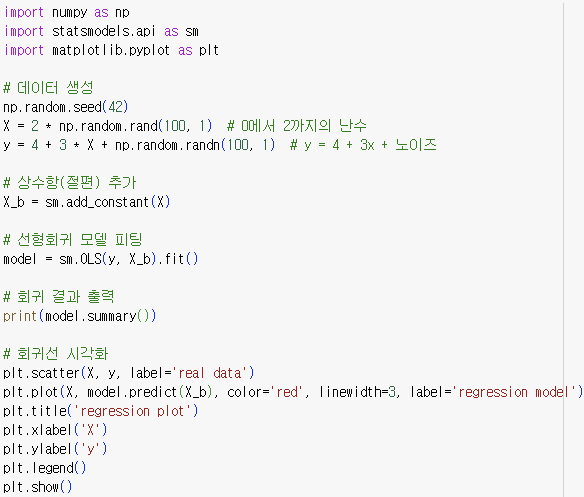
잔차와 변수들 간의 산점도를 그렸을 때 어떠한 패턴이 보이면 선형성 가정이 깨졌다고 진단할 수 있음

독립변수와 종속변수 간 산점도를 그려 선형 관계인지 진단할 수 있음

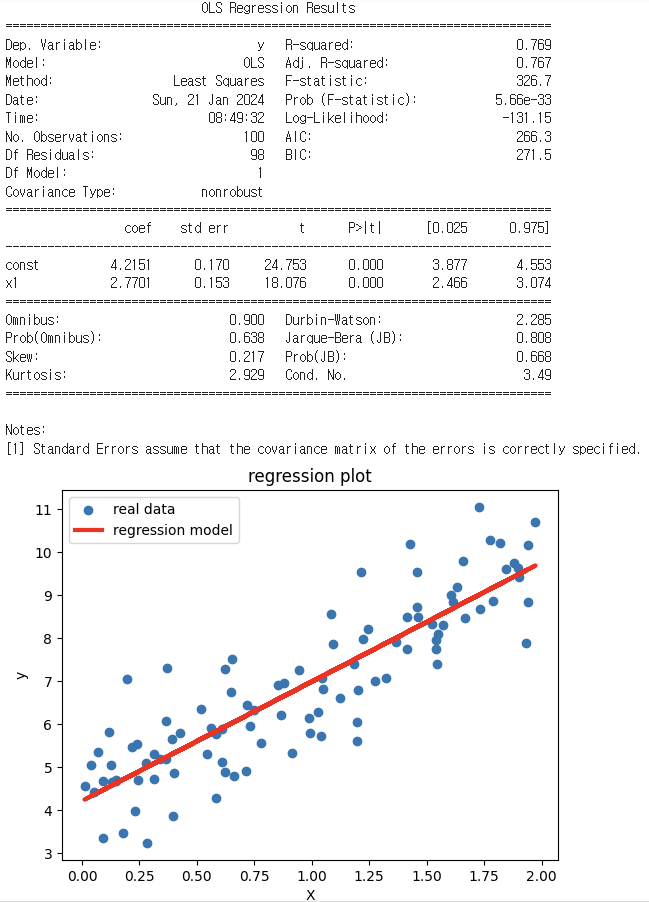
(상관계수를 이용하면 선형 관계의 정도 및 유의성은 파악 가능하지만 비선형 상관관계를 알 수 없음)

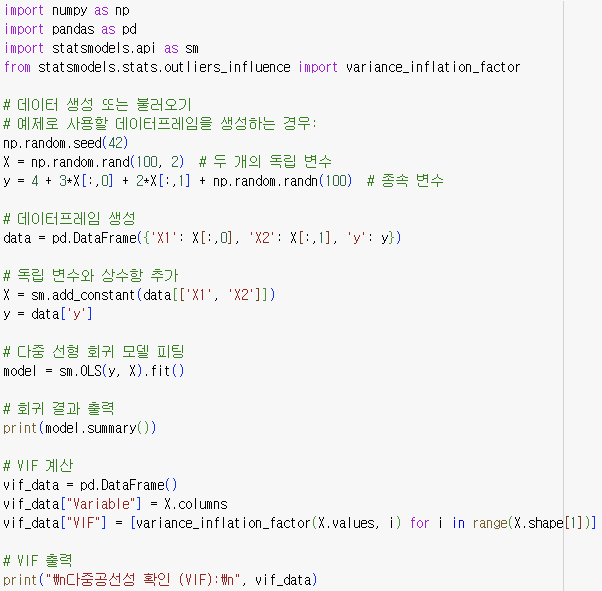
선형성 가정을 만족하지 않는 경우 변수를 제곱 형태 등으로 변환하여 선형성 유도 가능

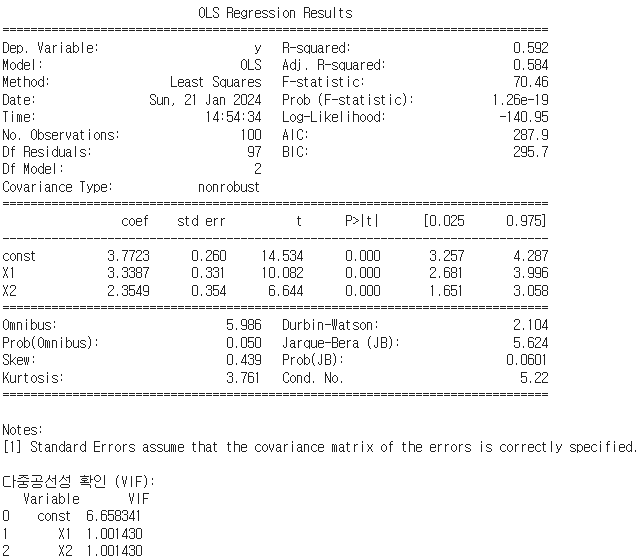
2) 독립성 진단

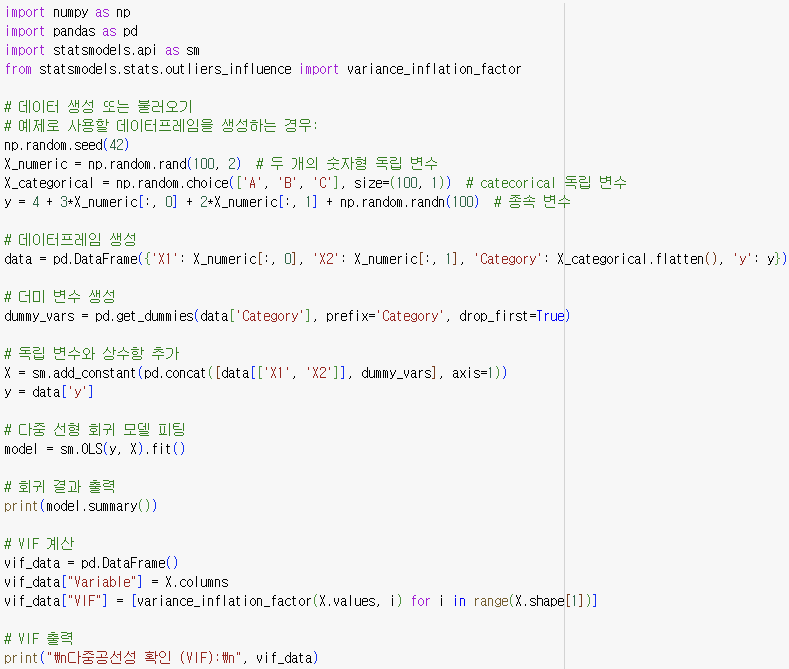
[단순 선형 회귀분석 실습]

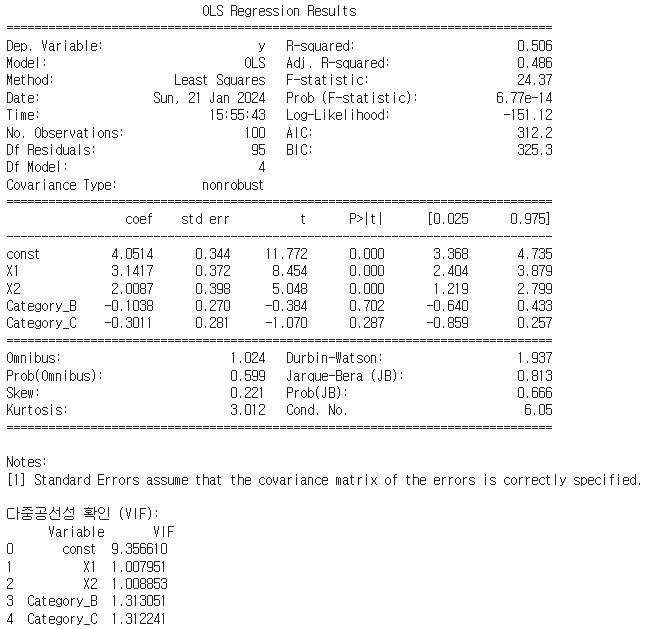
난수 발생시키는 코드보단 실제 데이터 활용하는 코드로 실습 코드 수정 필요

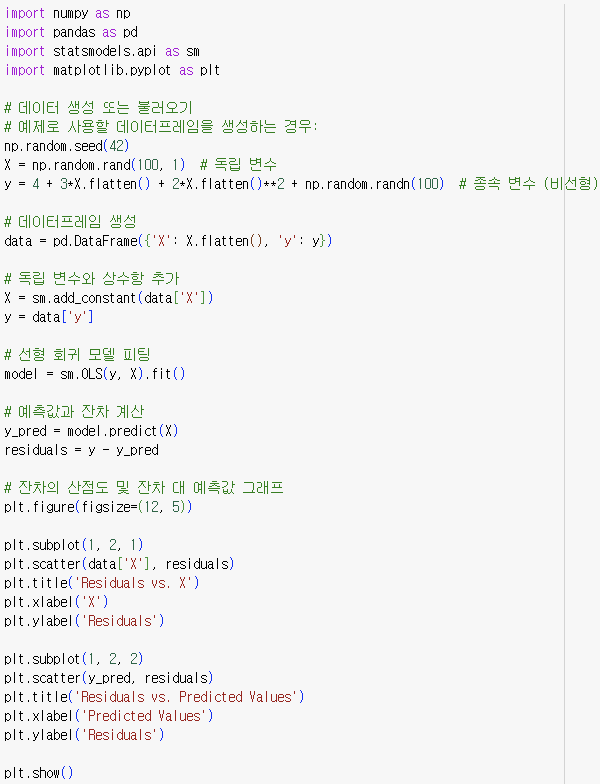


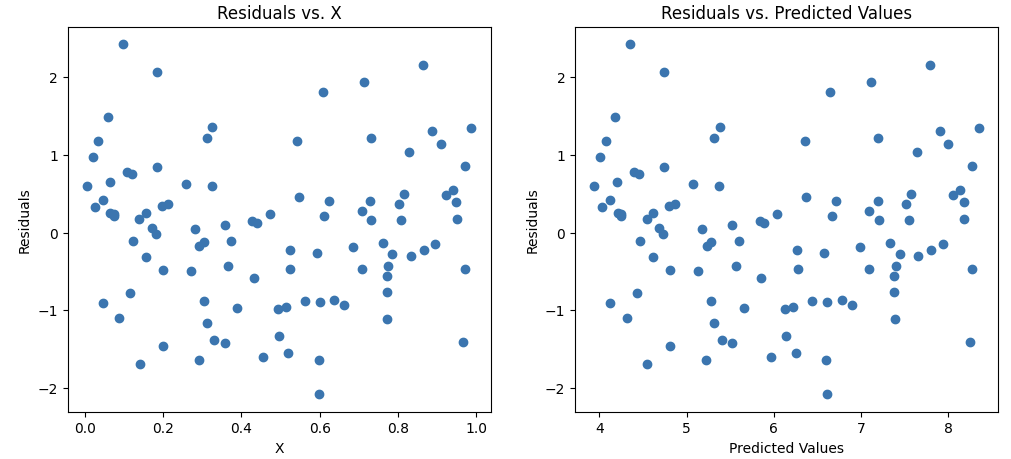
[다중 선형 회귀분석 실습]

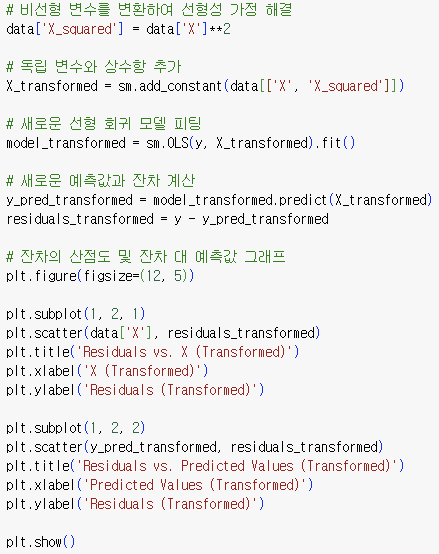


[categorical 변수가 포함된 다중 선형 회귀분석 실습]



[선형성 진단 실습]



[선형성 해결 실습]

