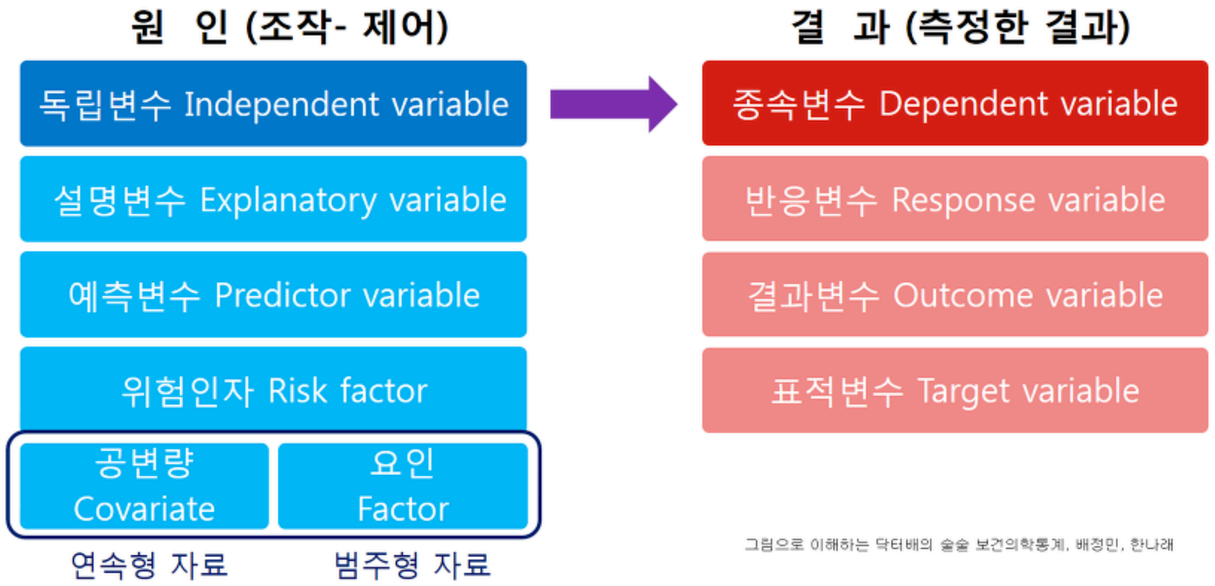
회귀분석 기초 개념

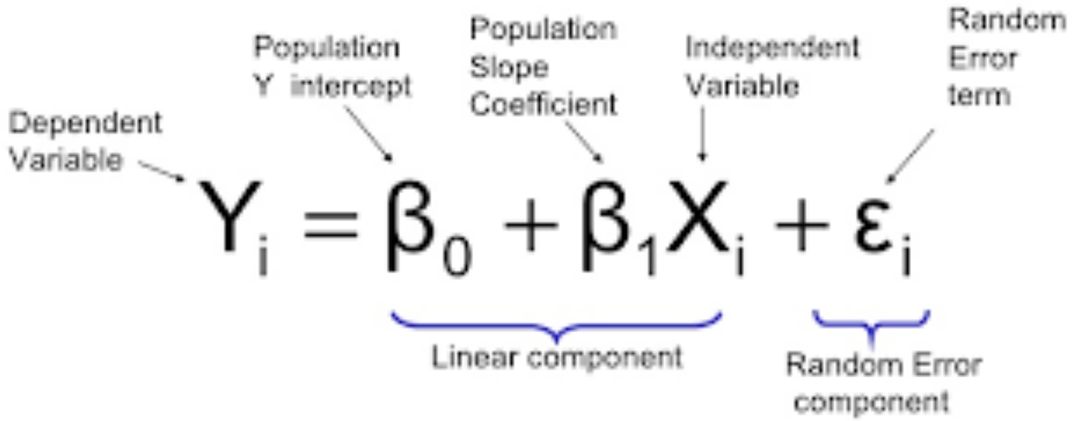
선형회귀

데이터 분석 모델링반 (ML1)

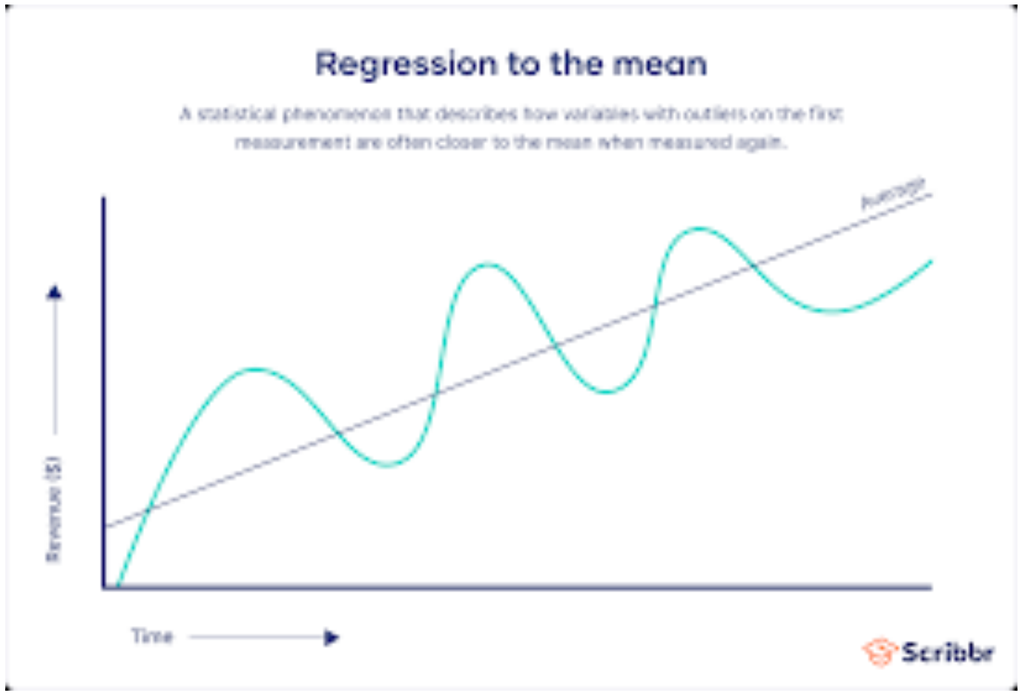
변수와의 관계 및 함수

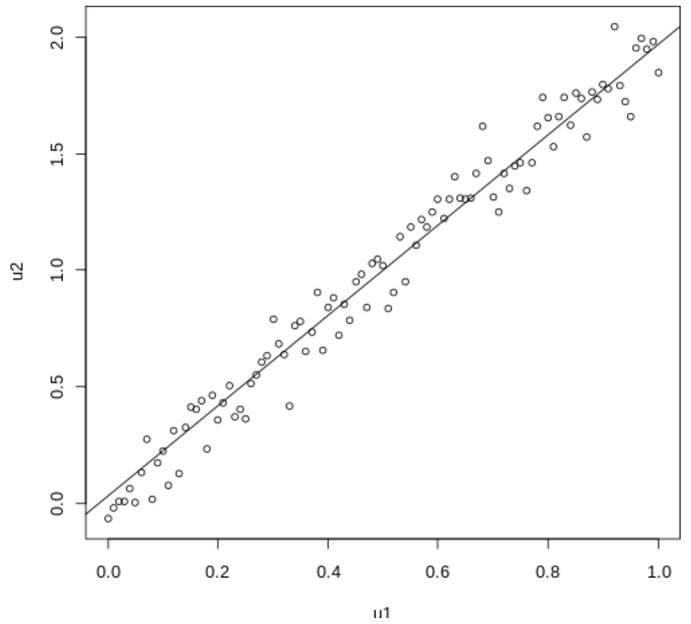


그림으로 이해하는 닥터배의 술술 보건의학통계, 배정민, 한나래



변수간의 선형관계 (linear relation)





산점도 (Scatter plot)

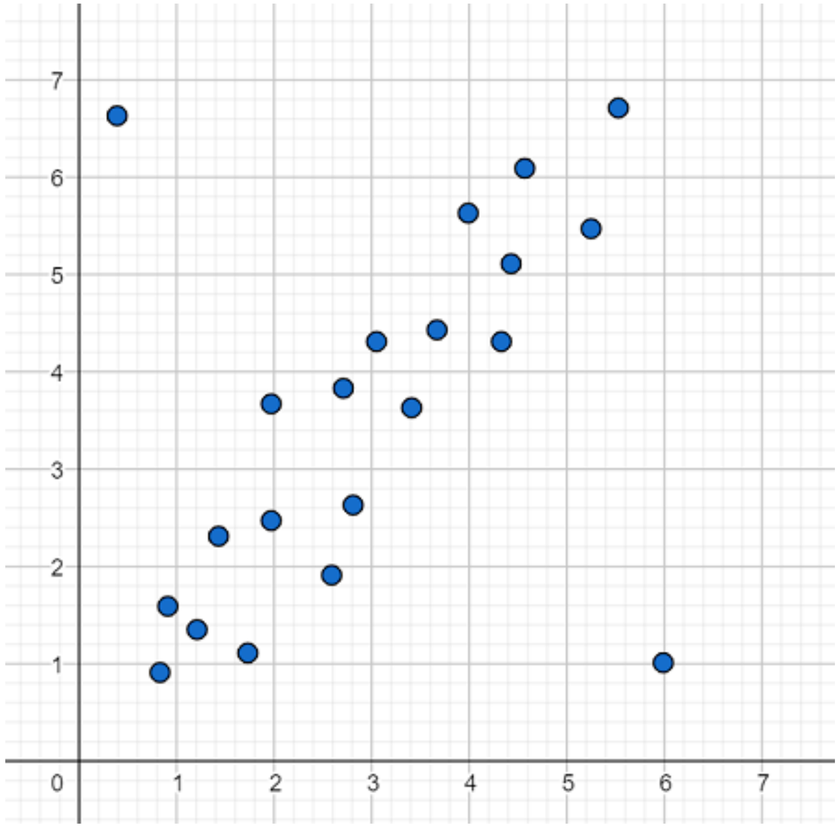
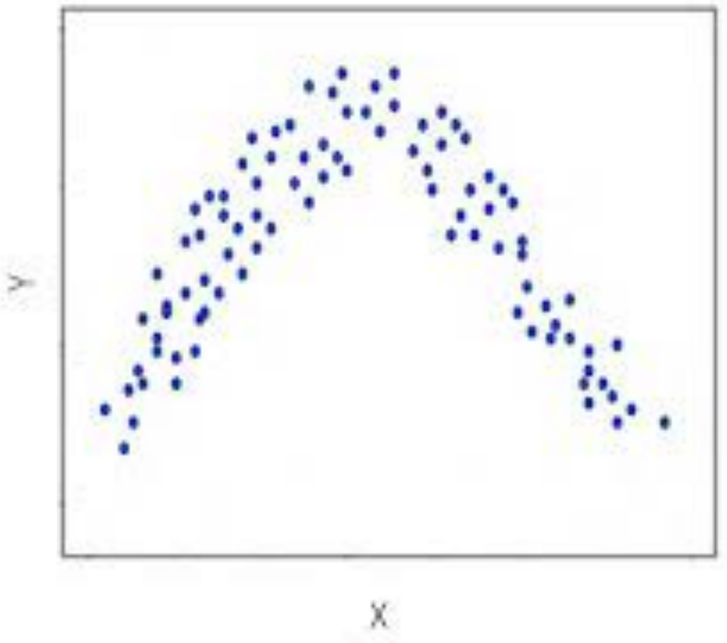


그림5 비선형적 관계 그래프



선형회귀 기본 가정(basic assumptions)

선형성 (Linearity)

설명변수 X와 반응변수 Y사이에 존재하는 관련성은, 주어진 X=x의 값에서 Y의 기대값을 $E(Y|X = x) = Bo + B1x$ 해당 선형식 회귀 계수에 대한 선형식

둘의 선형관계가 지속적이어야 하며, 진단 방법은 산점도, 잔차 예측값을 통해 확인

독립성 (Independence) - 오차항은 서로 독립적이어야 한다. 다른 값에 영향을 주지 않음

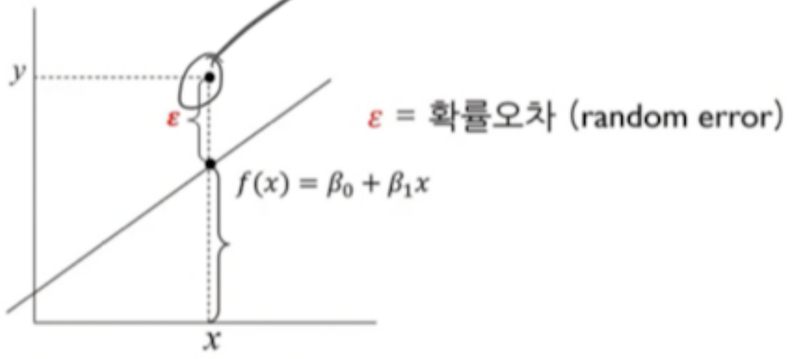
-시간의 순서에 따라 Durbin-Watson 통계량 계산하여 오차항 독립성 검사, 0에 가까우면 양의 자기상관, 4에 가까우면 음의 자기상관, 일반적 1.5~2.5 사이 독립

등분산성 (Homoscedasticity) - 모든 독립변수의 값에 대한 오차의 분산이 일정, 독립변수 크기에 상관없이 일관된 변동
- 잔차 대 적합값 플롯 사용하여 등분산성 평가, 플롯이 패턴을 보이거나 퍼짐 변화가 있으면 이 가정 위반

정규성 (Normality) -오차항은 정규 분포를 따라야 한다. 작은 표본에서 중요함 (특히), 큰 표본은 중심극한정리 완화

- Q-Q플롯, 정규성 검정 사용하여 오차항의 분포가 정규 분포 따르는지 확인 (Kolmogorov-Smirnov, Shapiro-Wilk테스트 등)

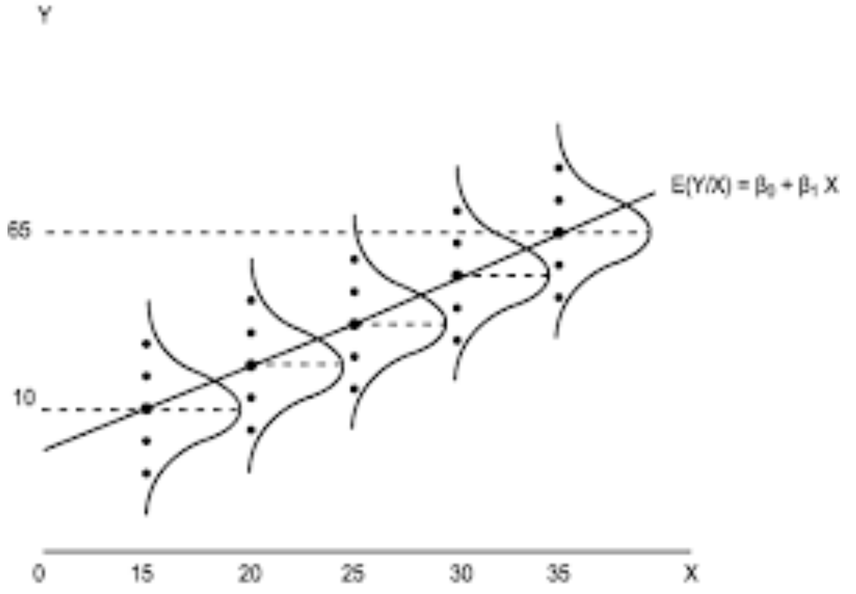
$y = \beta_0 + \beta_1 x + \varepsilon$



ε = 확률오차 (random error)

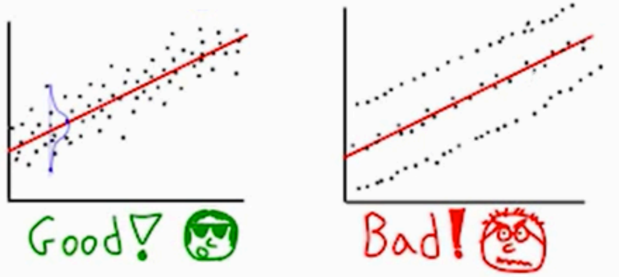
확률오차 가정 : ε_i~정규분포 $E(\varepsilon_i) = 0 \quad V(\varepsilon_i) = \sigma^2$ for all *i*.

$\varepsilon_i \sim N(0, \sigma^2),\ i = 1, 2, \ldots, n$

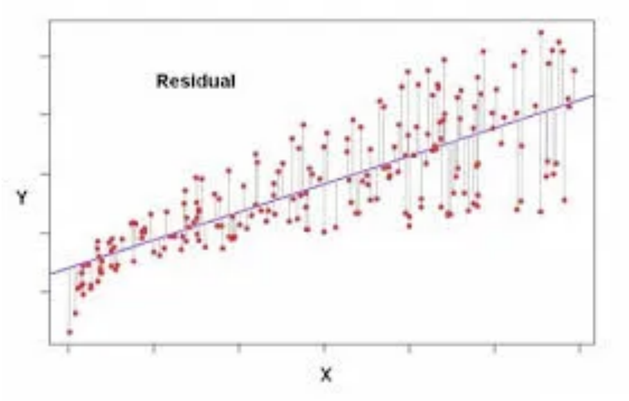


"잔차의 정규성"

오차항은 정규분포를 따르는가?
오차항에 대한 정규성 가정이 무너지면 t-검정, F-검정을 할 수가 없게 된다.



최소제곱법 (Method of least squares)



$$\sum_{i=1}^{n} residual^2$$

$$\sum_{i=1}^{n} (y_i - f(x_i, \beta))^2$$

$$f(x_i, \beta) = ax_i + b$$

$$\frac{\partial}{\partial a} \sum_{i=1}^{n} (y_i - ax_i + b)^2 = 0$$

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - ax_i + b)^2 = 0$$

$$\sum_{i=1}^{n} (y_i - (ax_i + b))^2 = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} ({y_i}^2 - 2y_i(ax_i + b) + (ax_i + b)^2) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} ({y_i}^2 - 2ax_iy_i - 2by_i + a^2{x_i}^2 + 2ax_i + b^2) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} [(a^2{x_i}^2 + 2ax_i - 2abx_iy_i) + ({y_i}^2 - 2by_i + b^2)]$$

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}[(a^2{x_i}^2+2abx_i-2ax_iy_i)+({y_i}^2-2by_i+b^2)]=0$$

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}[(b^2+2abx_i-2by_i)+(a^2x_i^2+{y_i}^2-2ax_iy_i)]=0$$

$$\vdots$$

$$\sum_{i=1}^{n}(2a{x_i}^2+2bx_i-2x_iy_i)=0$$

$$\sum_{i=1}^{n}(2b-2ax_i-2y_i)=0$$

$$\vdots$$

$$2(\sum_{i=1}^{n}a{x_i}^2+\sum_{i=1}^{n}bx_i-\sum_{i=1}^{n}x_iy_i)=0$$

$$2(\sum_{i=1}^{n}b-\sum_{i=1}^{n}ax_i-\sum_{i=1}^{n}y_i)=0$$

$$\vdots$$

$$\sum_{i=1}^{n}a{x_i}^2+\sum_{i=1}^{n}bx_i=\sum_{i=1}^{n}x_iy_i$$

$$\sum_{i=1}^{n}b-\sum_{i=1}^{n}ax_i=\sum_{i=1}^{n}y_i$$

$$\begin{bmatrix}\sum_{i=1}^{n}x_i^n & \sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n}1 & \sum_{i=1}^{n}x_i\end{bmatrix}\begin{bmatrix}a\\b\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{n}x_iy_i\\ \sum_{i=1}^{n}y_i\end{bmatrix}$$

$$\vdots$$

$$\begin{bmatrix}a\\b\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{n}x_i^n & \sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n}1 & \sum_{i=1}^{n}x_i\end{bmatrix}^{-1}\begin{bmatrix}\sum_{i=1}^{n}x_iy_i\\ \sum_{i=1}^{n}y_i\end{bmatrix}$$

정규 방정식

모든 j에 대해 $\frac{\partial J}{\partial \beta_j}=0$을 설정하면, 다음과 같은 연립 방정식을 얻을 수 있습니다:

$$X^T(Y-X\beta)=0$$

여기서 X는 설계 행렬(design matrix)이며, 각 열은 독립 변수를 나타냅니다. 이 방정식을 β에 대해 풀면:

$$\beta=(X^TX)^{-1}X^TY$$

이 식은 정규 방정식(Normal Equation)으로 알려져 있으며, 선형 회귀 모델의 회귀 계수를 찾는 데 사용됩니다.

결정계수 (Coefficient of determinant)

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

계산하기 전에 SST, SSE, SSR에 대해서 먼저 알 필요가 있습니다.

1 - SST(Total Sum of Squares)

$$SST = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$$

SST는 관측값에서 관측값의 평균(혹은 추정치의 평균)을 뺀 결과의 총합입니다.

2 - SSE(Explained Sum of Squares)

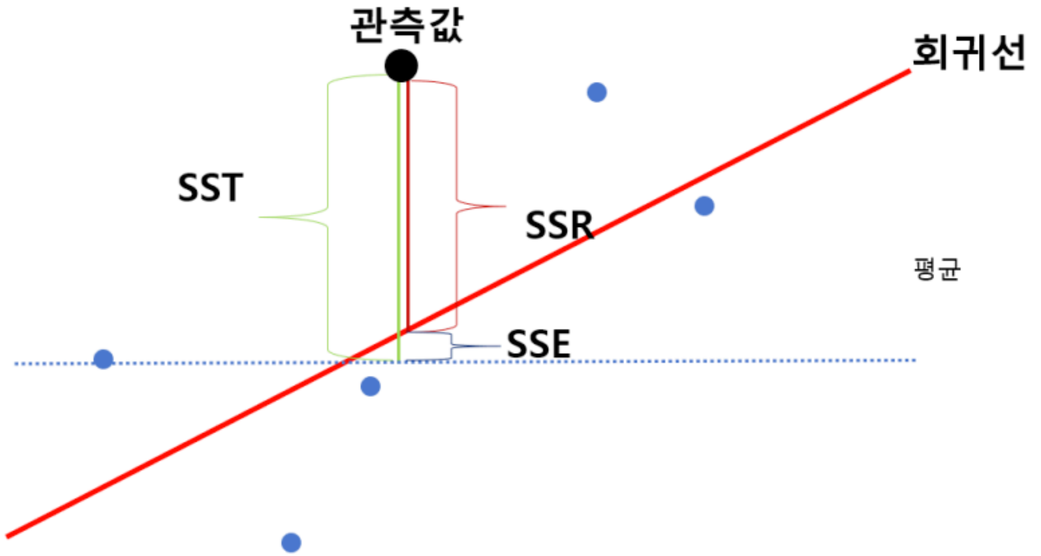
$$SSE = \sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2$$

SSE는 추정값에서 관측값의 평균(혹은 추정치의 평균)을 뺀 결과의 총합입니다.

3 - SSR(Residual Sum of Squares)

$$SSR = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

SSR은 관측값에서 추정값을 뺀 값. 즉 잔차(Residual)의 총합입니다.



다중공선성 Multicollinearity

독립 변수들 간에 높은 상관관계가 있을 때 발생하는 현상 회귀 모델의 정확성과 신뢰성을 저하시킬 수 있는 중요한 문제

다중공선성의 문제점

1. **계수 추정의 불안정성**: 다중공선성이 있는 경우, 작은 데이터의 변화에도 회귀 계수(β)가 크게 변동할 수 있습니다. 이로 인해 모델의 예측력이 불안정해지고, 데이터에 과적합할 위험이 커집니다.
2. **계수 해석의 어려움**: 변수들 간의 높은 상관관계로 인해, 각 독립 변수의 영향력을 분리해 해석하기 어렵습니다. 즉, 어떤 변수가 종속 변수에 영향을 미치는지 명확하게 구분이 어려워집니다.
3. **통계적 유의성 감소**: 다중공선성이 있는 변수들은 통계적으로 유의미하지 않게 나타날 수 있습니다. 이는 실제로 중요한 변수임에도 불구하고, 모델에서 제외될 위험이 있습니다.

다중공선성의 진단 방법

다음과 같은 방법으로 다중공선성을 진단할 수 있습니다:

- **상관 행렬(Correlation Matrix)**: 독립 변수들 간의 상관계수를 계산하여, 높은 상관관계를 보이는 변수 쌍을 확인합니다.
- **분산 팽창 요인(Variance Inflation Factor, VIF)**: VIF 값이 5 이상이면, 해당 변수는 다른 독립 변수들과 높은 상관관계를 가지고 있음을 의미합니다. 일부 기준은 VIF가 10 이상일 때 심각한 다중공선성이 있다고 봅니다.

다중공선성의 해결 방법

다중공선성 문제를 해결하기 위한 몇 가지 방법은 다음과 같습니다:

1. **변수 제거**: 높은 다중공선성을 보이는 변수 중 하나를 모델에서 제거합니다.
2. **주성분 분석(PCA)**: 주성분 분석을 통해 변수들을 변환하고, 변환된 변수들로 회귀 모델을 구축할 수 있습니다. 이 방법은 원본 변수들의 주요 정보를 요약하여 사용함으로써 다중공선성 문제를 줄입니다.
3. **릿지 또는 라쏘 회귀**: 이들 회귀 방법은 정규화를 통해 계수의 크기를 축소시킵니다. 이는 과적합을 방지하고, 다중공선성의 영향을 줄이는 효과가 있습니다.

모델의 정확성과 신뢰성 저하?

특이 행렬(Singular Matrix)	높은 선형 상관관계 -> XTX (공분산행렬) 선형 종속적인 관계 즉, 행렬의 Rank 감소하여 Full Rank가 되지 못하여 역행렬을 구할 수 없는 특이 행렬
Poorly Conditioned Matrix	변수 사이의 미묘한 변화가 계수 추정에 큰 변화를 일으키게 하여 계수 추정이 불안정 회귀 모델 예측 정확도도 심각하게 저하
통계적 유의성 왜곡	개별 회귀 계수의 표준 오차증가 -> 계수의 통계적 유의성 평가하는 데 사용되는 t- 통계량 신뢰도 저하,통계적으로 유의하지 않은 것으로 잘못 판단