# 단순 회귀 분석

## 개요

보조 도구의 단순 회귀 분석 절차는 최소 제곱 추정 방법을 사용하여 연속형 예측 변수(X)가 하나이고 연속형 반응 변수(Y)가 하나인 선형 및 2차 모형을 적합합니다. 사용자는 모형 유형을 선택하거나 보조 도구에서 가장 적합한 모형을 선택하도록 할 수 있습니다. 이 백서에서는 보조 도구에서 회귀 모형을 선택하기 위해 사용하는 기준에 대해 설명합니다.

또한 Minitab에서는 유효한 회귀 모형을 얻는 데 중요한 여러 요인을 조사합니다. 먼저, 표본은 검정력이 충분하고 X와 Y의 관계의 강도 추정치에 대한 정밀도가 충분할만큼 커야 합니다. 그런 다음에는 분석 결과에 영향을 미칠 수 있는 비정상적인 데이터를 식별하는 것이 중요합니다. 또한 오차항이 정규 분포를 따른다는 가정도 고려하고 비정규성이 전체 모형 및 계수의 가설 검정에 미치는 영향을 평가합니다. 마지막으로, 모형이 유용한지 확인하려면 선택된 모형 유형이 X와 Y의 관계를 정확히 반영해야 합니다.

보조 도구에서는 이러한 요인에 따라 데이터에 대해 다음과 같은 검사를 자동으로 수행하고 검사 결과를 보고서 카드에 표시합니다.

- 데이터 양
- 비정상적인 데이터
- 정규성
- 모형 적합

이 문서에서는 이러한 요인이 실제 회귀 분석과 어떤 관계가 있는지 조사하고 보조 도구에서 해당 요인을 확인하기 위한 가이드라인을 정한 방법에 대해 설명합니다.

# 회귀 분석 방법

## 모형 선택

보조 도구의 회귀 분석은 연속형 예측 변수가 하나이고 연속형 반응 변수가 하나인 모형을 적합하며 두 가지 유형의 모형을 적합할 수 있습니다.

- 선형: $F(x) = \beta_0 + \beta_1 X$
- 2차: $F(x) = \beta_0 + \beta_1 X + \beta_2 X^2$

사용자는 분석을 수행하기 전에 모형을 선택하거나 보조 도구에서 모형을 선택하도록 할 수 있습니다. 여러 가지 방법을 사용하여 어느 모형이 데이터에 가장 적절한지 확인할 수 있습니다. 모형이 유용한지 확인하려면 선택된 모형의 유형이 X와 Y의 관계를 정확히 반영해야 합니다.

### 목적

Minitab에서는 모형 선택에 사용할 수 있는 여러 방법을 조사하여 보조 도구에 사용할 방법을 결정하고자 했습니다.

### 방법

Minitab에서는 모형 선택에 일반적으로 사용되는 세 가지 방법(Neter et al., 1996)을 조사했습니다. 첫 번째 방법은 유의한 최고차항을 식별하는 것입니다. 두 번째 방법은 $R^2_{\text{수정}}$ 값이 가장 높은 모형을 선택하는 것입니다. 세 번째 방법은 전체 F-검정이 유의한 모형을 선택하는 것입니다. 자세한 내용은 부록 A를 참조하십시오.

보조 도구에서 사용할 방법을 결정하기 위해 Minitab에서는 방법을 조사하고 계산을 서로 비교했습니다. 또한 품질 분석 전문가의 의견도 수집했습니다.

### 결과

조사 결과를 바탕으로, Minitab에서는 모형에서 최고차항의 통계적 유의성을 바탕으로 모형을 선택하는 방법을 사용하기로 결정했습니다. 보조 도구에서는 먼저 2차 모형을 조사하고 모형의 제곱 항($\beta_2$)이 통계적으로 유의한지 여부를 검정합니다. 해당 항이 유의하지 않은 경우에는 모형에서 2차 항을 빼고 선형 항($\beta_1$)을 검정합니다. 이 방법을 통해 선택된 모형은 모형 선택 보고서에 표시됩니다. 또한 보조 도구에서 선택한 것과 다른 모형을 선택한 경우에는 모형 선택 보고서와 보고서 카드에 이를 표시합니다.

Minitab에서는 부분적으로 유의하지 않은 모형을 제외한 보다 단순한 모형을 일반적으로 선호한다는 품질 전문가의 의견 때문에 이 방법을 선택했습니다. 또한 방법을 비교한 결과, 모형 내 최고차항의 통계적 유의성을 사용하는 것이 가장 높은 $R^2_{\text{수정}}$ 값을 바탕으로 모형을 선택하는 것보다 더 유용합니다. 자세한 내용은 부록 A를 참조하십시오.

최고차항의 통계적 유의성을 사용하여 모형을 선택하지만 $R^2_{\text{수정}}$ 값 및 모형에 대한 전체 F-검정도 모형 선택 보고서에 표시합니다. 보고서에 표시된 상태를 보려면 아래 모형 적합 데이터 검사 항목을 참조하십시오.

# 데이터 검사

## 데이터 양

검정력은 귀무 가설이 거짓일 때 귀무 가설을 기각할 확률과 관련이 있습니다. 회귀 분석의 경우 귀무 가설은 X와 Y 간에 관계가 없다는 것입니다. 데이터 집합이 너무 작은 경우 검정의 검정력이 실제로 존재하는 X와 Y 간의 관계를 탐지하기에 적절하지 않을 수 있습니다. 따라서 데이터 집합은 실제적으로 중요한 관계를 높은 확률로 탐지하기에 충분히 커야 합니다.

### 목적

Minitab에서는 데이터의 양이 X와 Y 간의 관계에 대한 전체 F-검정의 검정력 및 X와 Y 간의 관계의 강도 추정치, $R^2_{수정}$의 정밀도에 어떤 영향을 미치는지 확인하고자 했습니다. 이 정보는 데이터에서 관측된 관계의 강도가 실제 관계의 강도에 대한 믿을 만한 지표라고 확신할 수 있을 만큼 데이터 집합이 충분히 큰지 여부를 확인하는 데 중요합니다. $R^2_{수정}$에 대한 자세한 내용은 부록 A를 참조하십시오.

### 방법

전체 F-검정의 검정력을 조사하기 위해 Minitab에서는 $R^2_{수정}$ 값의 범위 및 표본 크기에 대해 검정력 계산을 수행했습니다. $R^2_{수정}$의 정밀도를 조사하기 위해 수정된 $R^2(\rho^2_{수정})$의 여러 모수 및 여러 표본 크기에 대해 $R^2_{수정}$의 분포를 시뮬레이트했습니다. 또한 $R^2_{수정}$ 이 $\rho^2_{수정}$에 가까우려면 표본이 얼마나 커야 하는지 확인하기 위해 $R^2_{수정}$ 값의 변동성을 조사했습니다. 계산 및 시뮬레이션에 대한 자세한 내용은 부록 B를 참조하십시오.

### 결과

Minitab에서는 적당히 큰 표본의 경우 X와 Y 간의 관계가 실제 중요할 정도로 충분히 강력하지 않더라도 회귀 분석의 검정력이 X와 Y 간의 관계를 탐지하기에 양호하다는 것을 알았습니다. 보다 구체적으로 다음과 같은 사항을 확인했습니다.

- 표본 크기가 15이고 X와 Y 간에 강력한 관계($\rho^2_{수정}$ = 0.65)가 있는 경우 통계적으로 유의한 선형 관계를 찾을 확률은 0.9969입니다. 따라서 검정에서 15개 이상의 데이터 점을 사용하여 통계적으로 유의한 관계를 찾지 못하는 경우 실제 관계가 매우 강력하지 않을 가능성이 있습니다($\rho^2_{수정}$ 값 < 0.65).
- 표본 크기가 40이고 X와 Y 간에 적당히 약한 관계($\rho^2_{수정}$= 0.25)가 있는 경우 통계적으로 유의한 선형 관계를 찾을 확률은 0.9398입니다. 따라서 X와 Y 간의 관계가 적당히 약한 경우에도 F-검정에서 40개의 데이터 점을 사용하여 X와 Y 간의 관계를 찾을 가능성이 높습니다.

회귀 분석에서는 X와 Y 간의 관계를 상당히 쉽게 탐지할 수 있습니다. 따라서 통계적으로 유의한 관계를 찾은 경우 $R^2_{수정}$을 사용하여 관계의 강도를 평가해야 합니다. Minitab에서는 표본 크기가 충분히 크지 않은 경우 $R^2_{수정}$를 매우 신뢰할 수 없으며 표본에 따라 크게 달라질

수 있다는 것을 알았습니다. 그러나 표본 크기가 40 이상인 경우에는 $R^2_{수정}$ 값이 매우 안정적이며 신뢰할 수 있습니다. 표본 크기가 40이면 실제 값 및 모형 유형(선형 또는 2차)에 관계 없이 $R^2_{수정}$의 관측치가 $\rho^2_{수정}$의 0.20 내에 들어간다는 것을 90% 신뢰할 수 있습니다. 시뮬레이션 결과에 대한 자세한 내용은 부록 B를 참조하십시오.

이러한 결과를 바탕으로 데이터 양을 확인하는 경우 보조 도구의 보고서 카드에는 다음과 같은 정보가 표시됩니다.

| 상태 | 조건 |
| --- | --- |
|  | **표본 크기 < 40**<br>표본 크기 가 관계의 강도에 대해 아주 정확한 추정치를 제공하기에 충분히 크지 않습니다. R-제곱 및 R-제곱(수정) 등 관계의 강도 측정값은 크게 다를 수 있습니다. 더 정확한 추정치를 얻으려면 더 큰 표본(일반적으로 40 이상)을 사용해야 합니다.<br>**표본 크기≥ 40**<br>표본이 관계의 강도에 대해 정확한 추정치를 얻기에 충분히 큽니다. |

# 비정상적인 데이터

보조 도구 회귀 분석 절차에서는 비정상적인 데이터를 표준화 잔차 또는 레버리지 값이 큰 관측치로 정의합니다. 이 방법은 일반적으로 회귀 분석에서 비정상적인 데이터를 식별하기 위해 사용됩니다(Neter et al., 1996). 비정상적인 데이터가 결과에 중대한 영향을 미칠 수 있기 때문에 유효한 분석을 위해 데이터를 수정해야 할 수도 있습니다. 그러나 공정의 본래 변동에 따라 비정상적인 데이터가 발생할 수도 있습니다. 따라서 이러한 데이터 점을 처리하는 방법을 정하려면 비정상적인 동작의 원인을 식별하는 것이 중요합니다.

## 목적

Minitab에서는 데이터 점이 비정상적이라는 신호를 보내기 위해 표준화 잔차 및 레버리지 값이 얼마나 커야 하는지 확인하고자 했습니다.

## 방법

Minitab에서는 Minitab의 일반적인 회귀 분석 절차에 따라 비정상적인 관측치를 식별하기 위한 기준을 개발했습니다(통계분석 > 회귀 분석 > 회귀 분석).

## 결과

### 표준화 잔차

표준화 잔차는 잔차 $e_i$를 해당 표준 편차의 추정치로 나눈 값과 같습니다. 일반적으로, 관측치는 표준화 잔차의 절대값이 2보다 큰 경우 비정상적인 것으로 간주됩니다. 그러나 이 지침은 약간 보수적입니다. 모든 관측치의 약 5%가 우연히 이 기준을 충족할 것으로 예상됩니다(오차가 정규 분포를 따르는 경우). 따라서 관측치가 실제로 비정상적인지 확인하기 위해서는 비정상적인 반응치의 원인을 조사하는 것이 중요합니다.

### 레버리지 값

레버리지 값은 관측치의 X 값에만 관련이 있으며 Y 값에는 종속되지 않습니다. 관측치는 레버리지 값이 모형 계수의 수(p)를 관측치 수(n)로 나눈 값의 3배보다 크면 비정상적인 것으로 간주됩니다. 또한 일부 교과서에서는 $\frac{2 \times p}{n}$를 사용하지만 이 값이 일반적으로 기준 값으로 사용됩니다(Neter et al., 1996).

데이터에 높은 레버리지 점이 포함되어 있는 경우 이 점이 데이터를 적합하기 위해 선택된 모형의 유형에 불필요한 영향을 미치는지 여부를 확인하십시오. 예를 들어, 극단적인 X 값 하나 때문에 선형 모형 대신 2차 모형을 선택하게 될 수 있습니다. 2차 모형의 관측된 곡면성이 공정에 대한 여러분의 지식과 비교하여 일관성이 있는지 여부를 확인해야 합니다. 일관성이 없는 경우 더 간단한 모형을 데이터에 적합하거나 추가 데이터를 수집하여 공정을 더 철저히 조사하십시오.

비정상적인 데이터를 확인하는 경우 보조 도구의 보고서 카드에는 다음과 같은 상태가 표시됩니다.

| 상태 | 조건 |
|---|---|
|  | 비정상적인 데이터 점이 없습니다. 비정상적인 데이터 점은 결과에 심각한 영향을 미칠 수 있습니다. |
|  | 하나 이상의 큰 표준화 잔차 또는 하나 이상의 높은 레버리지 값이 있습니다.<br>점 위로 마우스를 움직이거나 Minitab의 브러시 기능을 사용하여 워크시트 행을 식별할 수 있습니다. 비정상적인 데이터는 결과에 심각한 영향을 미칠 수 있기 때문에 비정상적인 특성에 대한 원인을 식별해 보십시오. 데이터 입력이나 측정 오차가 있으면 수정하십시오. 특수 원인과 연관된 데이터를 제거하고 분석을 다시 실행하십시오. |

# 정규성

회귀 분석에서 일반적인 가정은 랜덤 오차($\varepsilon$)가 정규 분포를 따른다는 것입니다. 정규성 가정은 계수의 추정치($\beta$)에 대한 가설 검정을 수행하는 경우 중요합니다. 랜덤 오차가 정규 분포를 따르지 않는 경우에도 표본이 충분히 크면 검정 결과를 일반적으로 신뢰할 수 있습니다.

## 목적

Minitab에서는 정규 분포를 바탕으로 신뢰할 수 있는 결과를 제공하기 위해 필요한 표본 크기를 확인하고자 했습니다. 실제 검정 결과가 검정의 목표 유의 수준(알파 또는 제1종 오류율)과 얼마나 가깝게 일치했는지, 즉 여러 비정규 분포에 대해 검정이 예상된 것보다 더 자주 또는 덜 자주 귀무 가설을 잘못 기각했는지 확인하고자 했습니다.

## 방법

제1종 오류율을 추정하기 위해 정규 분포에서 크게 벗어난 치우친 분포, 두꺼운 꼬리를 갖는 분포 및 가는 꼬리를 갖는 분포를 사용하여 여러 시뮬레이션을 수행했습니다. 15의 표본 크기를 사용하여 선형 및 2차 모형에 대한 시뮬레이션을 수행했습니다. 전체 F-검정과 모형의 최고차 항에 대한 검정을 모두 조사했습니다.

각 조건에 대해 10,000번의 검정을 수행했습니다. Minitab에서는 각 검정에 대해 귀무 가설이 참이 되도록 랜덤 데이터를 생성했습니다. 그런 다음 0.05의 목표 유의 수준을 사용하여 검정을 수행했습니다. 10,000번 중에서 검정이 귀무 가설을 실제로 기각한 횟수를 집계하고 이 비율을 목표 유의 수준과 비교했습니다. 검정이 제대로 수행되는 경우 제1종 오류율은 목표 유의 수준에 매우 가깝습니다. 시뮬레이션에 대한 자세한 내용은 부록 C를 참조하십시오.

### 결과

전체 F-검정 및 모형 내 최고차항의 검정 모두에 대해 통계적으로 유의한 결과를 찾을 확률은 어떤 하나의 비정규분포에서 크게 다르지 않습니다. 제1종 오류율은 모두 0.038과 0.0529 사이로, 목표 유의 수준 0.05에 매우 가깝습니다.

표본 크기가 비교적 작아도 검정이 제대로 수행되기 때문에 보조 도구에서는 데이터의 정규성을 검정하지 않습니다. 대신, 보조 도구에서는 표본 크기를 확인하고 표본이 15보다 작은 경우 알려줍니다. 보조 도구의 보고서 카드에는 회귀 분석에 대해 다음과 같은 상태가 표시됩니다.

| 상태 | 조건 |
|---|---|
|  | 표본 크기가 15 이상이므로 정규성은 문제되지 않습니다. |
|  | 표본 크기가 15개 미만이므로 정규성이 문제가 될 수 있습니다. p-값을 해석할 때는 주의해야 합니다. 표본이 작으면 p-값의 정확성이 비정규 잔차에 민감합니다. |

## 모형 적합

회귀 분석을 수행하기 전에 선형 또는 2차 모형을 선택하거나 보조 도구에서 모형을 선택하도록 할 수 있습니다. 여러 가지 방법을 사용하여 적절한 모형을 선택할 수 있습니다.

### 목적

Minitab에서는 모형 유형을 선택하는 데 사용된 여러 방법을 조사하여 보조 도구에서 사용할 방법을 결정하고자 했습니다.

### 방법

Minitab에서는 모형 선택에 일반적으로 사용되는 세 가지 방법(Neter et al., 1996)을 조사했습니다. 첫 번째 방법은 유의한 최고차항을 식별하는 것입니다. 두 번째 방법은 $R^2_{\text{수정}}$ 값이 가장 높은 모형을 선택하는 것입니다. 세 번째 방법은 전체 F-검정이 유의한 모형을 선택하는 것입니다. 자세한 내용은 부록 A를 참조하십시오.

Minitab에서는 보조 도구에서 사용할 방법을 결정하기 위해 방법 및 계산을 서로 비교했습니다. 또한 품질 분석 전문가의 의견도 수집했습니다.

### 결과

Minitab에서는 모형에서 최고차항의 통계적 유의성을 바탕으로 모형을 선택하는 방법을 사용하기로 결정했습니다. 보조 도구에서는 먼저 2차 모형을 조사하고 모형의 제곱 항($\beta_3$)이

통계적으로 유의한지 여부를 검정합니다. 항이 유의하지 않은 경우에는 선형 모형의 선형 항($\beta_1$)을 검정합니다. 이 방법을 통해 선택된 모형은 모형 선택 보고서에 표시됩니다. 또한 보조 도구에서 선택한 것과 다른 모형을 선택한 경우에는 모형 선택 보고서와 보고서 카드에 이를 표시합니다. 자세한 내용은 위의 회귀 분석 방법 항목을 참조하십시오.

이런 결과를 토대로 보조 도구의 보고서 카드에는 다음과 같은 상태가 표시됩니다.

| 상태 | 조건 |
| --- | --- |
|  | **사용자의 모형이 보조 도구의 최적합 모형과 일치하는 경우**<br>목표의 관점에서 데이터와 적합한 모형을 평가해야 합니다. 적합선 그림에서 다음 사항을 확인하십시오.<br>• 표본이 X 값의 범위를 적절하게 포함합니다.<br>• 모형이 데이터의 곡면성을 적절하게 적합합니다(과적합 방지).<br>• 선이 특수 관심 영역에 잘 적합됩니다.<br><br>**사용자의 모형이 보조 도구의 최적합 모형과 일치하지 않는 경우**<br>모형 선택 보고서에 더 나은 선택이 될 수 있는 대체 모형이 표시됩니다. |

# 참고 문헌

Neter, J., Kutner, M.H., Nachtsheim, C.J., & Wasserman, W. (1996). *Applied linear statistical models*. Chicago: Irwin.

# 부록 A: 모형 선택

예측 변수 X와 반응 변수 Y의 관계를 보여주는 회귀 모형 형식은 다음과 같습니다.

$Y = f(X) + \varepsilon$

여기서 함수 f(X)는 주어진 X에 대한 Y의 기대값(평균)을 나타냅니다.

보조 도구에서는 두 가지 형식의 f(X)를 선택할 수 있습니다.

| 모형 유형 | f(X) |
|---|---|
| 선형 | $\beta_0 + \beta_1 X$ |
| 2차 | $\beta_0 + \beta_1 X + \beta_2 X^2$ |

계수 $\beta$의 값은 알려져 있지 않으며 데이터로부터 추정해야 합니다. 추정 방법은 최소 제곱으로 표본 내 잔차 제곱합을 최소화합니다.

$$\min \sum_{i=1}^{n} \left(Y_i - \hat{f}(X_i)\right)^2.$$

잔차는 관측된 반응 $Y_i$과 추정된 계수를 바탕으로 한 적합치 $\hat{f}(X_i)$ 간의 차이입니다. 이 제곱합의 최소값이 주어진 모형에 대한 SSE(오차 제곱합)입니다.

보조 도구에서 모형 유형을 선택하는 데 사용되는 방법을 결정하기 위해 세 가지 옵션을 평가했습니다.

- 모형 내 최고차항의 유의성
- 모형의 전체 F-검정
- 수정된 $R^2$ 값($R^2_{\text{수정}}$)

## 모형 내 최고차항의 유의성

이 방법에서, 보조 도구는 2차 모형으로 시작합니다. 보조 도구에서는 2차 모형 내 제곱항에 대한 가설을 검정합니다.

$H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

이 귀무 가설이 기각되는 경우 보조 도구에서는 제곱 항 계수가 0이 아니라는 결론을 내리고 2차 모형을 선택합니다. 그렇지 않은 경우 보조 도구에서는 선형 모형에 대한 가설을 검정합니다.

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

# 전체 F-검정

이 방법은 전체 모형(선형 또는 2차)의 검정입니다. 선택한 회귀 함수 f(X)의 형식에 대해 다음 사항을 검정합니다.

$H_0$: $f(X)$ 가 일정함

$H_1$: $f(X)$ 가 일정하지 않음

# 수정된 $R^2$

수정된 $R^2$($R^2_{수정}$)는 모형에서 반응의 변동성 중 X로 인한 부분을 측정합니다. X와 Y 간의 관측된 관계의 강도를 측정하는 데는 두 가지 일반적인 방법이 있습니다.

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

및

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

여기서

$$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

SSTO는 총 제곱합으로, 전체 평균 $\bar{Y}$에 대한 반응의 변동을 측정하며, SSE는 회귀 함수 f(X)에 대한 변동을 측정합니다. $R^2_{수정}$에서의 수정은 전체 모형 내 계수의 수(p)에 대한 것으로, $\varepsilon$의 분산을 추정할 수 있도록 n - p 자유도가 남게 됩니다. $R^2$은 모형에 계수가 추가되는 경우 감소하지 않지만, $R^2_{수정}$는 수정으로 인해 계수를 추가해도 모형이 개선되지 않는 경우 감소할 수 있습니다. 따라서 모형에 항을 추가해도 반응의 분산이 더 많이 설명되지 않는 경우 $R^2_{수정}$는 감소하며, 추가 항이 유용하지 않다는 것을 나타냅니다. 그러므로 수정된 측정값을 사용하여 선형 및 2차 모형을 비교해야 합니다.

# 모형 선택 방법 간의 관계

Minitab에서는 세 가지 모형 선택 방법 간의 관계, 계산 방식 및 서로에게 미치는 영향을 조사하고자 했습니다.

먼저, 전체 F-검정 및 $R^2_{수정}$이 계산되는 방식 간의 관계를 살펴보았습니다. 전체 모형의 검정에 대한 F-통계량은 $R^2_{수정}$ 계산에도 사용되는 SSE 및 SSTO의 항으로 표현할 수 있습니다.

$$F = \frac{(SSTO - SSE)/(p-1)}{SSE/(n-p)}$$

$$= 1 + \left(\frac{n-1}{p-1}\right)\frac{R^2_{수정}}{1 - R^2_{수정}}.$$

위의 공식을 보면 F-통계량이 $R^2_{수정}$의 증가 함수임을 알 수 있습니다. 따라서 검정은 $R^2_{수정}$이 검정의 유의 수준($\alpha$)에 의해 지정된 특정 값을 초과한 경우에만 $H_0$을 기각합니다. 이를 보여주기 위해 Minitab에서는 표 1에 표시된 여러 표본 크기에 대해 $\alpha = 0.05$에서 2차

모형의 통계적 유의성을 얻는 데 필요한 최소 $R^2_{수정}$을 계산했습니다. 예를 들어, n = 15인 경우 전체 F-검정이 통계적으로 유의하려면 모형의 $R^2_{수정}$ 값이 0.291877 이상이어야 합니다.

표 1 $\alpha = 0.05$에서 다양한 표본 크기의 2차 모형에 대한 전체 F 검정이 유의해지는 최소 $R^2_{수정}$

| 표본 크기 | 최소 $R^2_{수정}$ |
|---|---|
| 4 | 0.992500 |
| 5 | 0.900000 |
| 6 | 0.773799 |
| 7 | 0.664590 |
| 8 | 0.577608 |
| 9 | 0.508796 |
| 10 | 0.453712 |
| 11 | 0.408911 |
| 12 | 0.371895 |
| 13 | 0.340864 |
| 14 | 0.314512 |
| 15 | 0.291877 |
| 16 | 0.272238 |
| 17 | 0.255044 |
| 18 | 0.239872 |
| 19 | 0.226387 |
| 20 | 0.214326 |
| 21 | 0.203476 |
| 22 | 0.193666 |
| 23 | 0.184752 |
| 24 | 0.176619 |
| 25 | 0.169168 |

| 표본 크기 | 최소 $R^2_{\text{수정}}$ |
|---|---|
| 26 | 0.162318 |
| 27 | 0.155999 |
| 28 | 0.150152 |
| 29 | 0.144726 |
| 30 | 0.139677 |
| 31 | 0.134967 |
| 32 | 0.130564 |
| 33 | 0.126439 |
| 34 | 0.122565 |
| 35 | 0.118922 |
| 36 | 0.115488 |
| 37 | 0.112246 |
| 38 | 0.109182 |
| 39 | 0.106280 |
| 40 | 0.103528 |
| 41 | 0.100914 |
| 42 | 0.098429 |
| 43 | 0.096064 |
| 44 | 0.093809 |
| 45 | 0.091658 |
| 46 | 0.089603 |
| 47 | 0.087637 |
| 48 | 0.085757 |
| 49 | 0.083955 |
| 50 | 0.082227 |

다음으로, Minitab에서는 모형 내 최고차항의 가설 검정 및 $R^2_{수정}$ 간의 관계를 조사했습니다. 2차 모형의 제곱 항과 같은 최고차항에 대한 검정은 제곱합 또는 전체 모형(예: 2차)의 $R^2_{수정}$ 및 축소 모형(예: 선형)의 $R^2_{수정}$의 항들로 표현할 수 있습니다.

$$F = \frac{SSE(\text{축소}) - SSE(\text{완전})}{SSE(\text{완전})/(n-p)}$$

$$= 1 + \frac{(n-p+1)\left(R^2_{수정}(\text{완전}) - R^2_{수정}(\text{축소})\right)}{1 - R^2_{수정}(\text{완전})}.$$

공식을 보면 $R^2_{수정}(\text{축소})$의 고정된 값에 대해 F-통계량이 $R^2_{수정}(\text{완전})$.의 증가 함수임을 알 수 있습니다. 또한 검정 통계량이 두 $R^2_{수정}$ 값의 차이에 따라 어떻게 달라지는지 알 수 있습니다. 특히 통계적으로 유의한, 충분히 큰 F-값을 얻으려면 전체 모형에 대한 값이 축소 모형에 대한 값보다 커야 합니다. 따라서 최고차항의 유의성을 사용하여 가장 좋은 모형을 선택하는 방법이 가장 높은 $R^2_{수정}$이 있는 모형을 선택하는 방법보다 더 엄격합니다. 또한 최고차항 방법은 대부분의 사용자들이 더 간단한 모형을 선호하는 것과 일치합니다. 따라서 Minitab에서는 최고차항의 통계적 유의성을 사용하여 보조 도구에서 모형을 선택하기로 결정했습니다.

일부 사용자는 데이터를 가장 잘 적합하는 모형, 즉 $R^2_{수정}$가 가장 높은 모형을 선택하는 경향이 있습니다. 보조 도구의 모형 선택 보고서 및 보고서 카드에 이 값이 표시됩니다.

# 부록 B: 데이터 양

이 항목에서는 관측치의 수 n이 전체 모형 검정의 검정력 및 모형 강도의 추정치 $R^2_{수정}$의 정밀도에 미치는 영향에 대해 설명합니다.

관계의 강도를 정량화하기 위해 표본 통계량 $R^2_{수정}$에 대응되는 모수로 새로운 양 $\rho^2_{수정}$을 소개합니다. 다음 사항을 기억하십시오.

$$R^2_{수정} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

따라서 다음 사항을 정의합니다.

$$\rho^2_{수정} = 1 - \frac{E(SSE|X)/(n-p)}{E(SSTO|X)/(n-1)}$$

연산자 E(·|X)는 주어진 X 값에 대한 기대값 또는 랜덤 변수의 평균을 나타냅니다. 올바른 모형이 독립적으로 동일하게 분포된 ε이 있는 $Y = f(X) + \varepsilon$라고 가정하면 다음과 같은 결과를 얻게 됩니다.

$$\frac{E(SSE|X)}{n-p} = \sigma^2 = Var(\varepsilon)$$

$$\frac{E(SSTO|X)}{n-1} = \sum_{i=1}^{n} \frac{(f(X_i) - \bar{f})^2}{(n-1) + \sigma^2} + \sigma^2 \sum_{i=1}^{n} \frac{(f(X_i) - \bar{f})^2}{(n-1) + \sigma^2}$$

여기서$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$.

그러므로,

$$\rho^2_{수정} = \frac{\sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/(n-1)}{\sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/(n-1) + \sigma^2}$$

## 전체 모형 유의성

전체 모형의 통계적 유의성을 검정하는 경우, Minitab에서는 랜덤 오차 ε이 독립적이며 정규 분포를 따른다고 가정합니다. 따라서 Y의 평균이 일정하다 $(f(X) = \beta_0)$라는 귀무 가설 하에서 F-검정 통계량은 $F(p-1, n-p)$ 분포를 따릅니다. 대립 가설 하에서는 F-통계량이 비중심 모수를 가진 비중심 $F(p-1, n-p, \theta)$ 분포를 따릅니다.

$$\theta = \sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/\sigma^2$$

$$= \frac{(n-1)\rho^2_{수정}}{1 - \rho^2_{수정}}$$

$H_0$을 기각하는 확률은 비중심 모수와 함께 증가하며, 비중심 모수는 n 및 $\rho^2_{수정}$ 둘 다에 따라 증가합니다.

위의 공식을 사용하여 Minitab에서는 선형 및 2차 모형에 대해 n = 15일 때 $\rho^2_{수정}$ 값의 범위에 대한 전체 F-검정의 검정력을 계산했습니다. 결과는 표 2를 참조하십시오.

표 2 n=15일 때 여러 $\rho^2_{수정}$ 값을 사용한 선형 및 2차 모형의 검정력

| $\rho^2_{수정}$ | Θ | F의 검정력<br>선형 | F의 검정력<br>2차 |
|---|---|---|---|
| 0.05 | 0.737 | 0.12523 | 0.09615 |
| 0.10 | 1.556 | 0.21175 | 0.15239 |
| 0.15 | 2.471 | 0.30766 | 0.21896 |
| 0.20 | 3.50 | 0.41024 | 0.29560 |
| 0.25 | 4.667 | 0.51590 | 0.38139 |
| 0.30 | 6.00 | 0.62033 | 0.47448 |
| 0.35 | 7.538 | 0.71868 | 0.57196 |
| 0.40 | 9.333 | 0.80606 | 0.66973 |
| 0.45 | 11.455 | 0.87819 | 0.76259 |
| 0.50 | 14.00 | 0.93237 | 0.84476 |
| 0.55 | 17.111 | 0.96823 | 0.91084 |
| 0.60 | 21.00 | 0.98820 | 0.95737 |
| 0.65 | 26.00 | 0.99688 | 0.98443 |
| 0.70 | 32.667 | 0.99951 | 0.99625 |
| 0.75 | 42.00 | 0.99997 | 0.99954 |
| 0.80 | 56.00 | 1.00000 | 0.99998 |
| 0.85 | 79.333 | 1.00000 | 1.00000 |
| 0.90 | 126.000 | 1.00000 | 1.00000 |
| 0.95 | 266.00 | 1.00000 | 1.00000 |

전체적으로, Minitab에서는 X와 Y 간의 관계가 강력하고 표본 크기가 15 이상일 때 검정의 검정력이 높다는 것을 알았습니다. 예를 들어, 표 2를 보면 $\rho^2_{수정}$ = 0.65인 경우, $\alpha$ = 0.05에서 통계적으로 유의한 선형 관계를 찾을 확률이 0.99688임을 알 수 있습니다. F-검정을 사용하여 이러한 강력한 관계를 탐지하지 못하는 경우는 표본의 0.5% 미만입니다. 2차 모형의 경우에도 F-검정을 사용하여 관계를 탐지하지 못하는 경우는 표본의 2% 미만입니다. 따라서 검정에서 15개 이상의 관측치를 사용하여 통계적으로 유의한 관계를

찾지 못하면 실제 관계가 존재할 경우 $\rho^2_{수정}$ 값이 0.65보다 작을 가능성이 높습니다. $\rho^2_{수정}$가 0.65보다 크지 않아도 실제적으로 중요할 수 있습니다.

Minitab에서는 또한 표본 크기가 더 클 때(n=40) 전체 F-검정의 검정력을 조사하고자 했습니다. 표본 크기 n = 40이 $R^2_{수정}$의 정밀도에 대한 중요한 임계값이라는 것을 확인했으며(아래 관계의 강도 참조) 표본 크기에 대해 검정력 값을 평가하고자 했습니다. 선형 및 2차 모형에 대해 n = 40일 때 $\rho^2_{수정}$ 값의 범위에 대한 전체 F-검정의 검정력을 계산했습니다. 결과는 표 3을 참조하십시오.

표 3 n = 40일 때 여러 $\rho^2_{수정}$ 값을 사용한 선형 및 2차 모형의 검정력

| $\rho^2_{수정}$ | θ | F의 검정력<br>선형 | F의 검정력<br>2차 |
|---|---|---|---|
| 0.05 | 2.0526 | 0.28698 | 0.21541 |
| 0.10 | 4.3333 | 0.52752 | 0.41502 |
| 0.15 | 6.8824 | 0.72464 | 0.60957 |
| 0.20 | 9.7500 | 0.86053 | 0.76981 |
| 0.25 | 13.0000 | 0.93980 | 0.88237 |
| 0.30 | 16.7143 | 0.97846 | 0.94925 |
| 0.35 | 21.0000 | 0.99386 | 0.98217 |
| 0.40 | 26.0000 | 0.99868 | 0.99515 |
| 0.45 | 31.9091 | 0.99980 | 0.99905 |
| 0.50 | 39.0000 | 0.99998 | 0.99988 |
| 0.55 | 47.6667 | 1.00000 | 0.99999 |
| 0.60 | 58.5000 | 1.00000 | 1.00000 |
| 0.65 | 72.4286 | 1.00000 | 1.00000 |

Minitab에서는 X와 Y 간의 관계가 상당히 약한 경우에도 검정의 검정력이 높다는 것을 알았습니다. 예를 들어, 표 3을 보면 $\rho^2_{수정}$ = 0.25인 경우에도 $\alpha = 0.05$에서 통계적으로 유의한 선형 관계를 찾을 확률이 0.93980임을 알 수 있습니다. 관측치가 40개인 경우 관계가 상당히 약하더라도 F-검정에서 X와 Y 간의 관계를 탐지하지 못할 가능성이 없습니다.

## 관계의 강도

데이터에 통계적으로 유의한 관계가 있어도 X와 Y 간에 반드시 강력한 관계가 있는 것은 아닙니다. 따라서 많은 사용자들이 관계가 실제로 얼마나 강력한지 알려주는 $R^2_{수정}$과 같은

지표를 찾습니다. $R^2_{수정}$을 $\rho^2_{수정}$의 추정치로 고려하는 경우, 추정치가 참 $\rho^2_{수정}$ 값에 상당히 가깝다고 확신하고자 합니다.

$R^2_{수정}$ 및 $\rho^2_{수정}$ 간의 관계를 보여주기 위해 Minitab에서는 여러 $\rho^2_{수정}$ 값에 대한 $R^2_{수정}$의 분포를 시뮬레이트하여 $R^2_{수정}$가 여러 n 값에 대해 얼마나 가변적인지 확인했습니다. 아래 그림 1-4의 그래프는 시뮬레이트된 10,000개의 $R^2_{수정}$ 값에 대한 히스토그램을 보여줍니다. 각 히스토그램 쌍에서 $\rho^2_{수정}$의 값이 같으므로 표본 크기 15와 표본 크기 40인 경우에 대한 $R^2_{수정}$의 변동성을 비교할 수 있습니다. Minitab에서는 0.0, 0.30, 0.60 및 0.90의 $\rho^2_{수정}$ 값을 검정했습니다. 모든 시뮬레이션은 선형 모형으로 수행되었습니다.

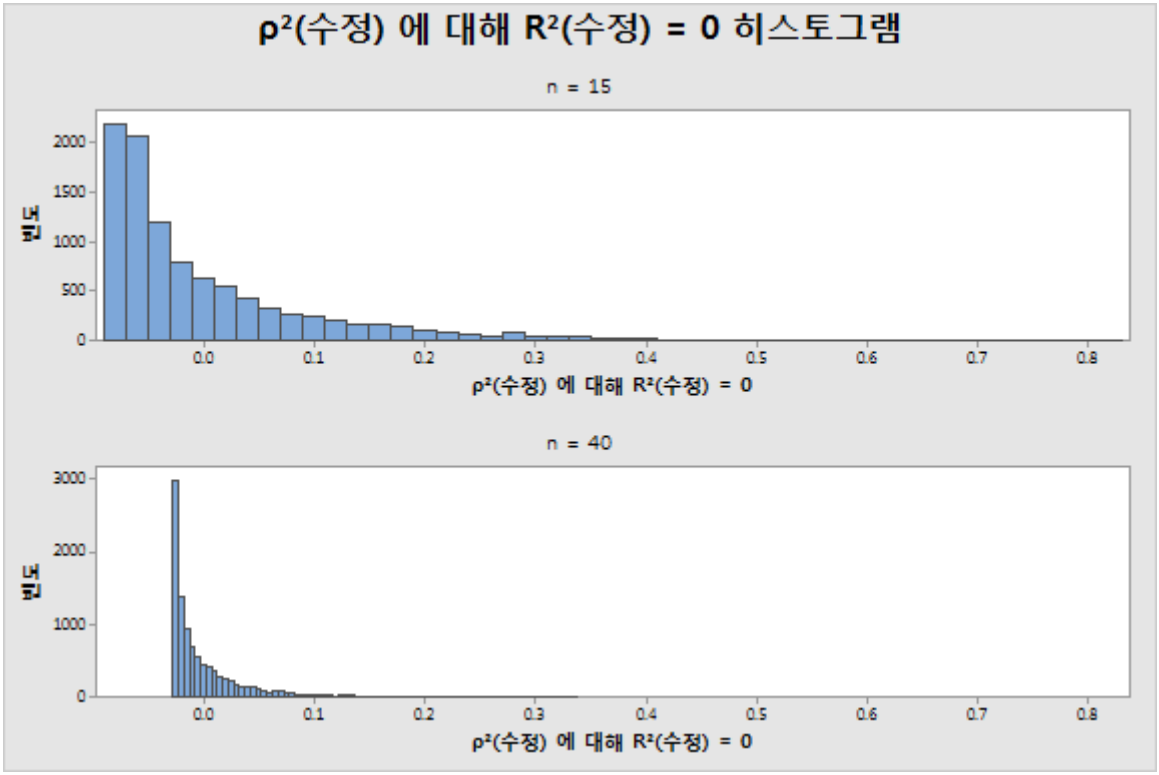


그림 1 n=15 및 n=40에서 $\rho^2_{수정}$ = 0.0에 대해 시뮬레이트된 $R^2_{수정}$ 값

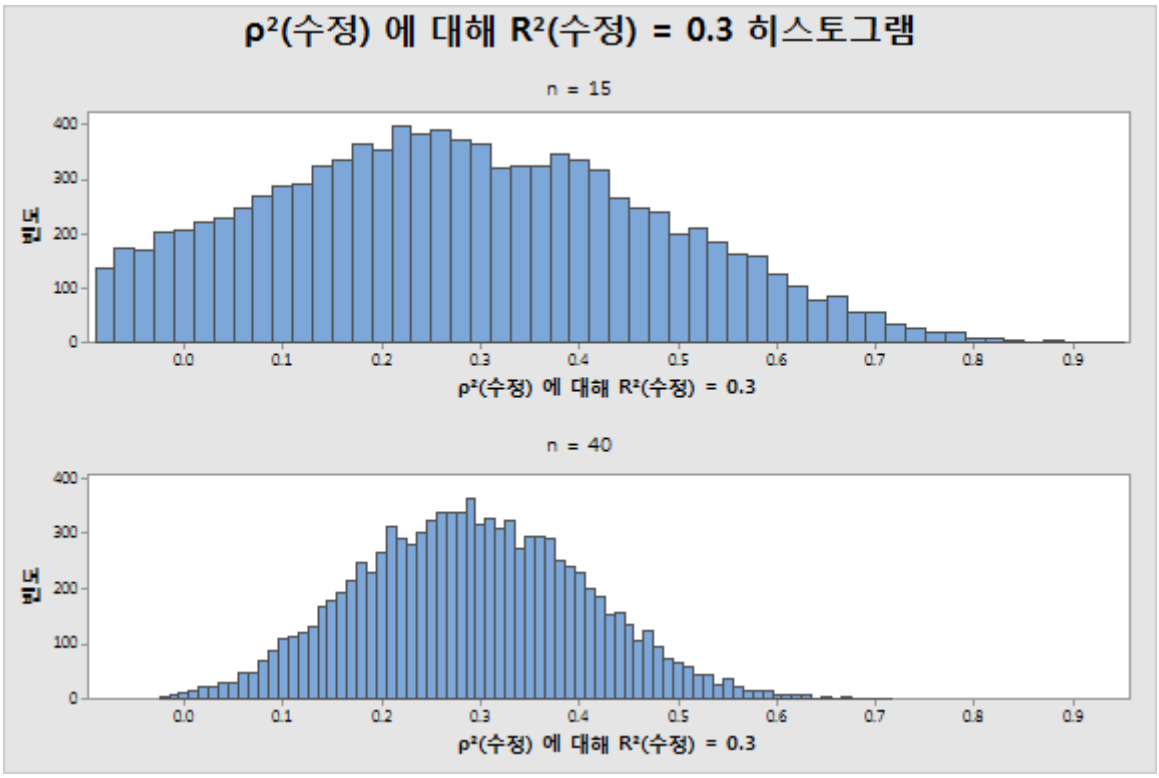


그림 2 n=15 및 n=40에서 $\rho^2_{수정}$ = 0.30에 대해 시뮬레이트된 $R^2_{수정}$ 값

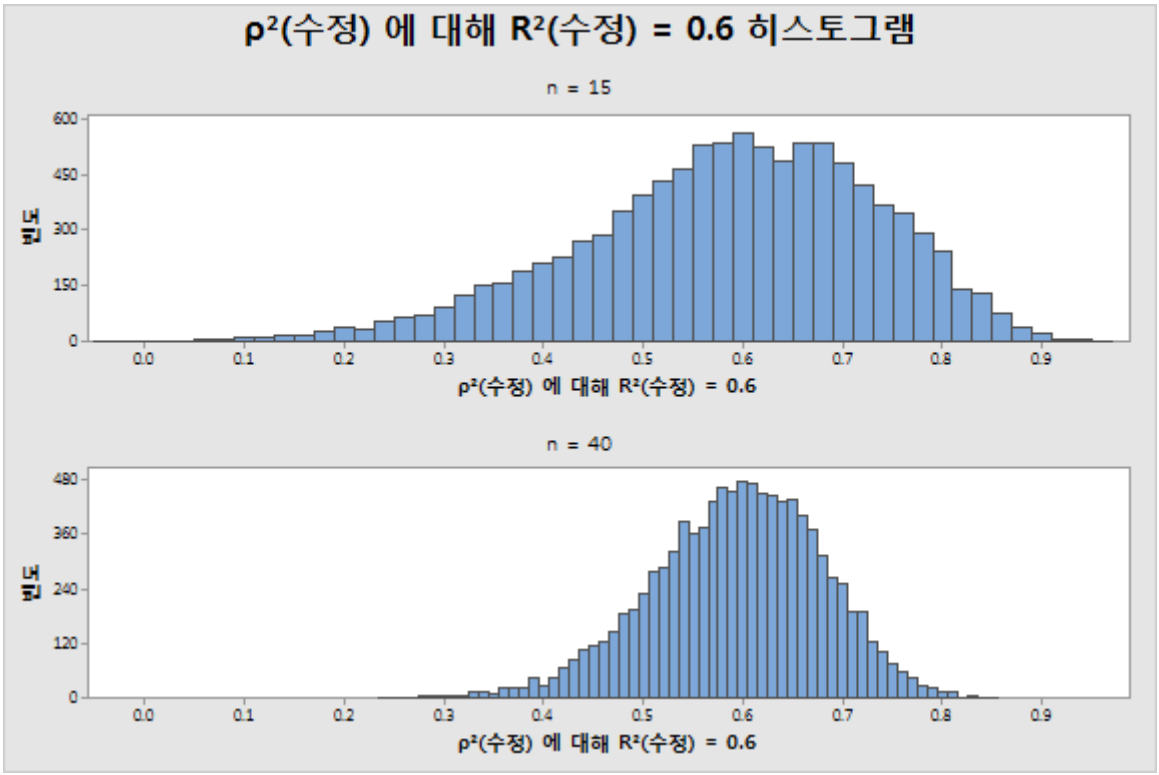


그림 3 n=15 및 n=40에서 $\rho^2_{수정}$ = 0.60에 대해 시뮬레이트된 $R^2_{수정}$ 값

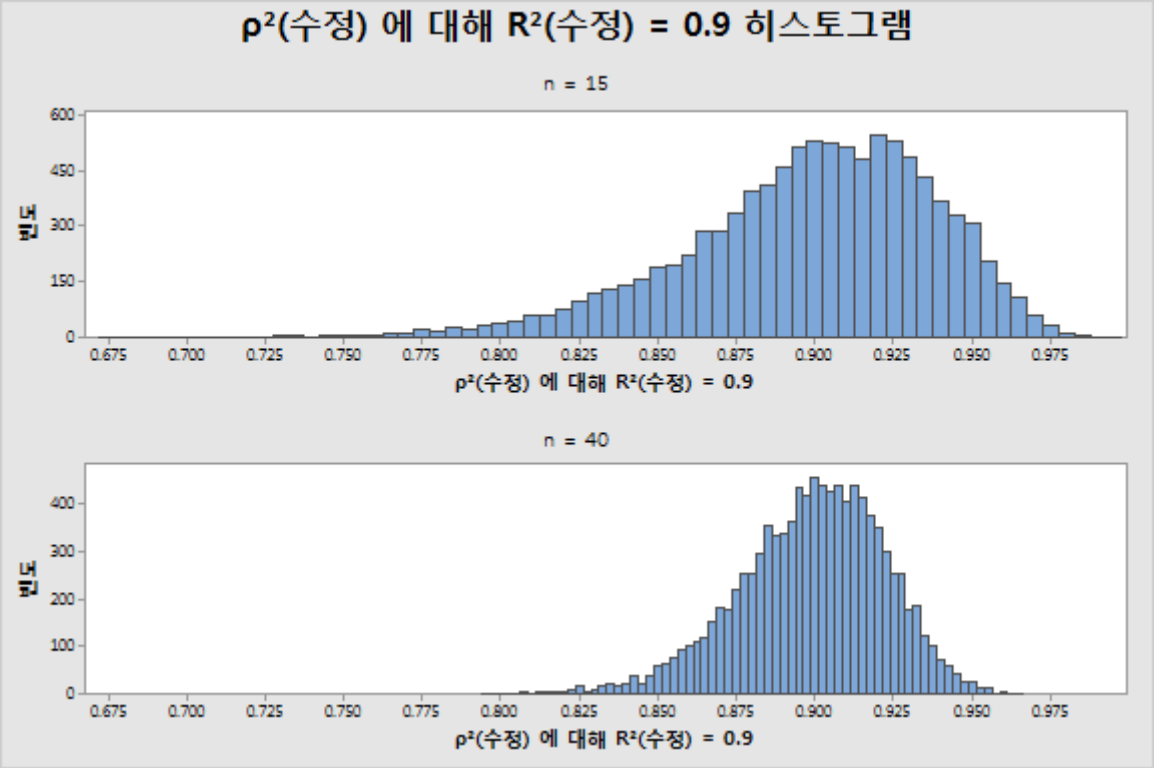


그림 4 n=15 및 n=40에서 $\rho^2_{수정}$ = 0.90에 대해 시뮬레이트된 $R^2_{수정}$ 값

전체적으로, 시뮬레이션 결과에 따르면 관계의 실제 강도($\rho^2_{수정}$)와 데이터에서 관측된 관계($R^2_{수정}$) 간에 상당한 차이가 있을 수 있습니다. 표본 크기를 15에서 40으로 증가시키면 차이의 가능한 범위가 크게 감소합니다. Minitab에서는 0.20보다 큰 절대 차이 $|R^2_{수정} - \rho^2_{수정}|$가 10% 이하의 확률로 발생하는 최소값을 식별하여 40개의 관측치가 적절한 임계값이라는 것을 확인했습니다. 이러한 사실은 어느 모형을 고려하는 경우에도 $\rho^2_{수정}$의 참 값과 관계가 없습니다. 선형 모형에서 가장 좋지 않은 경우는 $\rho^2_{수정}$ = 0.31으로, n = 36이 필요했습니다. 2차 모형에서 가장 어려운 경우는 $\rho^2_{수정}$ = 0.30으로, n = 38이 필요했습니다. 관측치가 40개인 경우, 값 및 모형 유형(선형 또는 2차)에 관계 없이 $R^2_{수정}$의 관측치가 $\rho^2_{수정}$의 0.20 내에 들어간다는 것을 90% 신뢰할 수 있습니다.

# 부록 C: 정규성

보조 도구에서 사용되는 회귀 모형 형식은 모두 다음과 같습니다.

$$Y = f(X) + \varepsilon$$

랜덤 항 $\varepsilon$에 대한 일반적인 가정은 항들이 독립적이며 동일하게 분포된 평균 0 및 일반 분산 $\sigma^2$을 갖는 정규 랜덤 변수라는 것입니다. $\beta$ 모수의 최소 제곱 추정치는 $\varepsilon$이 정규 분포를 따른다는 가정을 무시하는 경우에도 여전히 최적의 선형 불편 추정치입니다. 정규성 가정은 f(X)에 대한 가설 검정에서처럼 이러한 추정치에 확률을 추가하고자 하는 경우에만 중요하게 됩니다.

Minitab에서는 정규성 가정을 바탕으로 회귀 분석의 결과를 신뢰할 수 있는 n의 크기를 확인하고자 했습니다. 시뮬레이션을 수행하여 다양한 비정규 오차 분포 하에 가설 검정의 제1종 오류율을 조사했습니다.

표 4에는 10,000개 시뮬레이션 중에서 선형 및 2차 모형의 다양한 ε의 분포에 대해 $\alpha = 0.05$에서 전체 F-검정이 유의한 비율이 나와 있습니다. 이러한 시뮬레이션에서는 X와 Y 간에 관계가 없다는 귀무 가설이 참이었습니다. X 값은 전체 구간에 고르게 배치되어 있었습니다. 모든 검정에 대해 n=15의 표본 크기를 사용했습니다.

표 4 비정규 분포의 경우 n=15인 선형 및 2차 모형에 대한 전체 F-검정의 제1종 오류율

| 분포 | 선형적으로 유의함 | 2차적으로 유의함 |
|---|---|---|
| 정규 | 0.04770 | 0.05060 |
| t(3) | 0.04670 | 0.05150 |
| t(5) | 0.04980 | 0.04540 |
| Laplace | 0.04800 | 0.04720 |
| 균등 | 0.05140 | 0.04450 |
| Beta(3, 3) | 0.05100 | 0.05090 |
| 지수 | 0.04380 | 0.04880 |
| Chi(3) | 0.04860 | 0.05210 |
| Chi(5) | 0.04900 | 0.05260 |
| Chi(10) | 0.04970 | 0.05000 |
| Beta(8, 1) | 0.04780 | 0.04710 |

다음으로, 최적의 모형을 선택하는 데 사용된 최고차항의 검정을 조사했습니다. 각 시뮬레이션에 대해 제곱 항이 유의한지 여부를 고려했습니다. 제곱 항이 유의하지 않은 경우, 선형 항이 유의한지 여부를 고려했습니다. 이러한 시뮬레이션에서 귀무 가설이 참이고 목표 $\alpha = 0.05$ 및 n=15이었습니다.

표 5 비정규 분포의 경우 n=15인 선형 또는 2차 모형에 대한 최고차항의 검정의 제1종 오류율

| 분포 | 제곱 | 선형 |
|---|---|---|
| 정규 | 0.05050 | 0.04630 |
| t(3) | 0.05120 | 0.04300 |
| t(5) | 0.04710 | 0.04820 |
| Laplace | 0.04770 | 0.04660 |
| 균등 | 0.04670 | 0.04900 |
| Beta(3, 3) | 0.05000 | 0.04860 |
| 지수 | 0.04600 | 0.03800 |
| Chi(3) | 0.05110 | 0.04290 |
| Chi(5) | 0.05290 | 0.04490 |
| Chi(10) | 0.04970 | 0.04610 |
| Beta(8, 1) | 0.04770 | 0.04380 |

시뮬레이션 결과를 보면 전체 F-검정 및 모형 내 최고차항의 검정 모두에 대해 통계적으로 유의한 결과를 찾을 확률은 어떤 하나의 오차 분포에서 크게 다르지 않습니다. 제1종 오류율은 모두 0.038과 0.0529 사이입니다.