다중 선형 회귀분석을 수행하고 통계적으로 유의한 여러 예측 변수를 포함하는 모델을 결정했다면, 자연스레 어떤 변수가 가장 중요한가 라는 질문이 떠오르게 됩니다.

이 질문은 생각보다 복잡합니다. 우선 '가장 중요하다'는 말의 정의는 대부분의 경우 해당 분야와 목표에 따라 달라집니다. 또한 표본 데이터를 수집하고 측정하는 방법도 각 변수의 명백한 중요도에 영향을 줄 수 있습니다.

이러한 문제에 유의하면서 이 질문에 대한 답을 찾아보도록 하겠습니다. 먼저 좀 의외겠지만 중요도에 관한 답변을 제공하지 않는 통계를 살펴본 다음 회귀 모형에서 가장 중요한 변수를 파악하기 위한 통계적 방법과 통계적이지 않은 방법을 모두 알아보겠습니다.

변수의 중요도를 파악하기 위해 일반 회귀 계수를 비교하지 마세요

일반 회귀 계수는 각 예측 변수와 반응의 관계를 나타냅니다. 계수 값은 예측 변수가 1단위 증가할 때 반응의 평균이 얼마나 변화하는지 나타냅니다. 결과적으로 계수가 큰 변수가 보다 큰 응답 변화를 나타내므로 더 중요하다고 생각하기 쉽습니다.

하지만 여러 변수 유형에 따라 단위가 다르므로 직접적인 비교는 불가능합니다. 예를 들어, 온도, 무게 또는 화학성분 농도에서 1단위 변화의 의미는 매우 다릅니다.

게다가 이 문제는 각 측정 유형 내에서도 여러 단위가 존재하므로 더욱 복잡해집니다. 예를 들어, 무게는 그램(g)과 킬로그램(kg) 단위로 측정할 수 있습니다. 같은 데이터 세트에 대해 어떤 모형에서는 그램을 사용하고 다른 모형에서는 킬로그램을 사용하여 모형을 적합화하면 모형 의 적합화가 그대로임에도 불구하고 무게 계수가 최대 1,000배까지 변화할 수 있습니다. 즉, 변수의 중요도는 동일한데 계수 값이 크게 변화하게 됩니다.

요점: 계수가 크다고 해서 예측 변수가 더 중요한 것은 아닙니다.

변수의 중요도를 파악하기 위해 P값을 비교하지 마세요

이제 계수 값이 변수의 중요도를 나타내는 것이 아니라는 것을 알았습니다. 하지만 변수의 p값은 어떨까요? 어쨌든 우리는 모형에 변수를 포함할지 여부를 정하기 위해 낮은 p값을 살펴보곤 하니까요.

p값 계산은 다양한 속성을 포함하지만, 중요도 측정은 여기에 포함되지 않습니다. 매우 낮은 p값은 아주 정확도가 높은 추정치와 큰 표본 크기 등 중요도 외 다른 속성을 나타낼 수 있습니다.

현실에서는 사소한 효과도 p값이 매우 낮을 수 있습니다. 즉, 통계적으로 유의미한 결과가 현실에서는 실제로 유의미하지 않을 수 있습니다.

요점: 낮은 p값이 반드시 실제로 중요한 예측 변수를 나타내는 것은 아닙니다.

변수의 중요도를 파악하기 위해 몇 가지 통계를 비교하세요

우선 변수의 중요도를 평가할 수 없는 몇 가지 분명한 통계는 제외했습니다. 다행히 회귀 모형에서 가장 중요한 예측 변수를 파악하는 데 도움이 되는 통계가 몇 가지 있습니다. 이러한 통계는 각각 '가장 중요한'을 정의하는 방식이 조금씩 다르기 때문에 그 결과가 서로 일치하지 않을 수 있습니다.

표준화된 회귀 계수

위에서(일반 회귀 계수는 척도가 서로 달라서 직접 비교가 불가능하다고 언급했습니다만) 같은 척도를 기반으로 하도록 회귀 계수를 표준화하면 비교가 가능합니다.

표준화된 계수를 도출하려면 모든 계량형 예측 변수의 값을 표준화하세요. Minitab에서는 기본 회귀분석 대화 상자에서 '코드화' 버튼만 클릭하면 간편하게 표준화할 수 있습니다. '계량형 예측 변수 표준화'에서 '평균값을 뺀 후 표준 편차로 나누기'를 선택하세요.

표준화된 예측 변수를 사용하여 회귀 모델을 적합한 다음 코드화된 계수(표준화된 계수)를 확인하세요. 이 코드화를 통해 서로 다른 예측 변수의 척도를 동일한 척도에 놓고 해당 계수를 직접 비교할 수 있습니다. 표준화된 계수는 예측 변수의 표준편차가 1만큼 변화하는 경우 반응의 평균이 얼마나 변화하는지 나타냅니다.

요점: 표준화된 계수의 절대값이 가장 큰 예측 변수를 찾으세요.

변수가 모형에 마지막으로 추가되었을 때 R-제곱의 변화

Minitab 보조도구의 다중 회귀에는 깔끔한 분석이 포함되어 있습니다. 이 분석은 다른 모든 변수가 이미 포함된 모형에 각 변수를 추가하는 경우 R-제곱이 얼마나 증가하는지 계산합니다.

R-제곱 분석의 변화는 각 변수를 모형에 마지막으로 입력된 변수로 취급하므로, 변화는 모형의 다른 변수들은 설명할 수 없지만 특정 변수는 설명할 수 있는 분산의 비율을 나타냅니다. 즉, R-제곱 변화는 각 변수가 모형의 다른 변수를 넘어 설명하는 고유한 분산의 양을 나타냅니다.

요점: R-제곱의 가장 큰 증가와 관련된 예측 변수를 찾으세요.

회귀 모형에서 가장 중요한 변수를 파악하기 위해 통계를 사용한 예

아래 결과는 예측 변수가 3개인 회귀 모형을 나타냅니다. 텍스트 결과는 Minitab의 일반 회귀 분석에 의해 도출되었습니다. 이 예에서는 코드화된 계수로 분류된 표준화된 계수를 확인할 수 있도록 코드화 대화상자를 사용하여 연속 예측 변수를 표준화했습니다. 이 분석은 Minitab의 **통계분석 > 회귀 분석 > 회귀 분석 > 적합 회귀 모형 메뉴에 있습니다.**

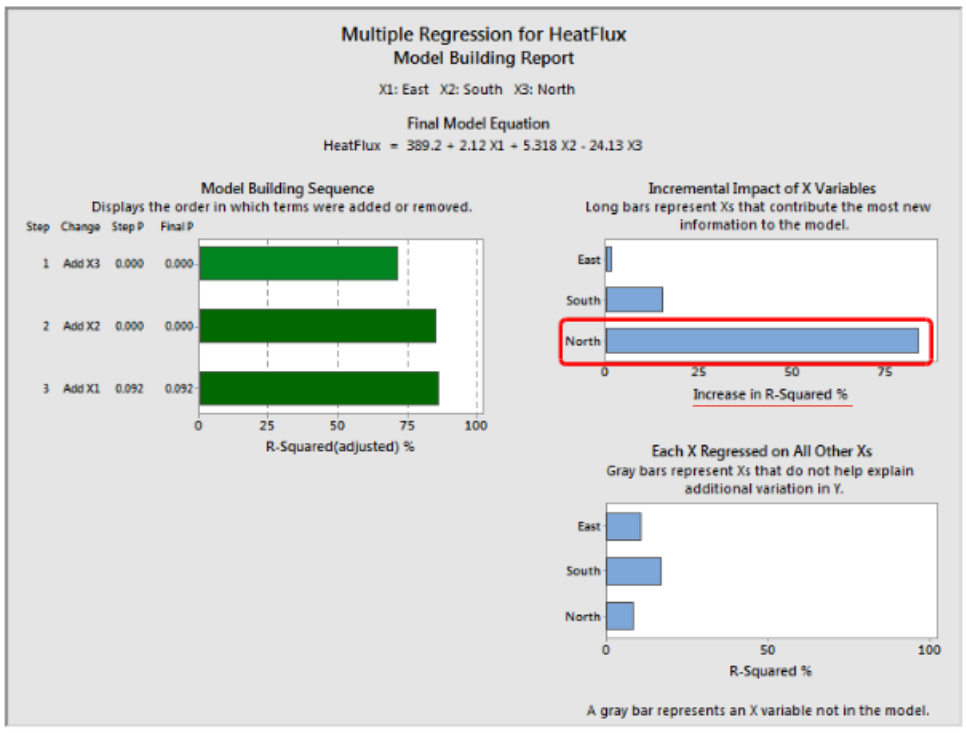
그래프가 포함된 보고서는 보조 메뉴의 다중 회귀분석에서 생성했습니다. 이 분석은 Minitab의 **보조 도구 > 회귀 분석 > 다중 회귀 분석 메뉴에 있습니다.**

```
Coded Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  249.64     1.60   156.36    0.000
East        3.01     1.72     1.75    0.092  1.12
South       9.85     1.78     5.52    0.000  1.21
North     -21.92     1.70   -12.92    0.000  1.09

Regression Equation in Uncoded Units

HeatFlux = 389.2 + 2.12 East + 5.318 South - 24.13 North
```



표준화된 계수를 보면, North의 표준화 계수 절대값이 가장 크며, South와 East가 차례로 그 뒤를 잇는 것을 확인할 수 있습니다. Incremental Impact 그래프는 North가 설명하는 고유 분산을 가장 많이 설명하고, South와 East가 차례로 그 뒤를 잇는다는 사실을 나타냅니다. 이 예에서는 두 통계량 모두 North가 회귀 모형에서 가장 중요한 변수임을 나타냅니다.

통계를 사용하여 중요한 변수를 파악할 때 주의사항

통계 측정을 통해 여러 예측 변수의 상대적 중요성을 확인할 수 있습니다. 그러나 이러한 측정을 통해 변수가 실질적으로도 중요한지 파악할 수는 없습니다. 실질적인 중요도를 확인하려면 해당 분야에 대한 지식을 활용해야 합니다.

표본 수집 및 측정 방법으로 인해 표본에 포함된 변수의 명백한 중요도가 모집단에서의 실제 중요도에 비해 편향될 수 있습니다.

관찰에서 무작위로 표본을 추출하면 표본의 예측 변수 값의 변동성이 모집단의 변동성을 반영할 가능성이 높습니다. 이 경우 표준화된 계수와 R-제곱 값의 변화는 모집단 값을 반영할 가능성이 높습니다.

그러나 표본으로 제한된 범위의 예측 변수를 선택하면 두 통계가 해당 예측 변수의 중요도를 과소평가하는 경향이 나타납니다. 반대로 예측 변수의 표본 변동성이 모집단의 변동성보다 크면 두 통계가 해당 예측 변수의 중요도를 과대평가하는 경향이 나타납니다.

또한 예측 변수 측정의 정확도와 정밀도는 명백한 중요도에 영향을 줄 수 있으므로 반드시 고려해야 합니다. 예를 들어 측정의 품질이 낮으면 변수가 실제보다 예측 가능성이 낮아 보일 수 있습니다.

반응 평균을 변경하는 것이 목표라면 예측 변수와 반응 간 단순한 상관관계가 아닌 인과관계가 존재한다고 확신해야 합니다. 상관관계는 있지만 인과관계가 없는 경우, 통계적 중요도 측정과는 관계없이 예측 변수 값을 의도적으로 변경하더라도 반드시 원하는 반응 변화가 발생하는 것은 아닙니다.

인과관계가 성립하는지 확인하려면 보통 관찰 연구보다는 설계된 실험을 수행해야 합니다.

중요한 변수를 파악하기 위한 비통계적 고려 사항

대부분의 경우 '가장 중요한'의 정의는 목표와 해당 분야에 따라 다릅니다. 통계학을 사용해 회귀 모형에서 가장 중요한 변수를 파악할 수는 있으나, 통계 분석의 모든 측면에 해당 분야에 대한 전문 지식을 활용하는 것이 중요합니다. 실제로 회귀 모형에서 가장 중요한 변수를 파악하는 데에는 현실적인 문제가 영향을 미치기 마련입니다.

예를 들어 여러분의 목표가 예측 변수값을 변경하여 반응을 바꾸는 것이라면, 변경하기 가장 적합한 변수를 파악하는 데 여러분의 전문 지식을 활용해야 합니다. 변경하기가 비교적 어렵거나 많은 비용이 소요되는 변수가 존재할 수 있으며, 일부 변수는 변경이 불가능할 수도 있습니다. 어떤 경우에는 변수 하나를 크게 변경하는 것이 다른 변수의 작은 변화보다 보다 더 실용적일 수도 있습니다.

'가장 중요한'것은 주관적이고 상황에 따라 변화할 수 있는 특성입니다. 통계학을 활용하여 회귀 모형에서 가장 중요한 변수가 될 수 있는 변수를 파악할 수는 있으나, 이 경우에도 해당 분야에 대한 전문 지식을 함께 활용해야 한다는 사실을 잊지 마세요.