초급회귀분석(단순/다중)

# 단순 선형 회귀분석

## 1. 회귀분석의 종류

|  |  |  |
| --- | --- | --- |
| 1차 구분요인 | 2차 구분요인 | 구분 |
| 독립변수의 수 | 1개 | 단순 회귀분석 |
|  | 2개 이상 | 다중 회귀분석 |
| 독립변수의 척도 | 명목척도, 서열척도 | 더미변수 회귀분석, 로지스틱 회귀분석 |
|  | 등간척도, 비율척도 | 일반 회귀분석 |
| 독립변수와 종속변수 관계 | 선형 | 선형 회귀분석 |
|  | 비선형 | 비선형 회귀분석 |

## 2. 산점도 확인

선형(linear)이란 그래프가 직선으로 나타난다는 뜻이다. SPSS Statistics에서는 점들의 분포를 보고 ‘선형이다’ 혹은 ‘비선형이다’ 라고 판단한다.

## 3. 회귀분석에서의 독립변수와 종속변수

회귀분석을 실시하기 위해서는 우선 변수를 확인해야 한다. 단순 회귀분석에서는 독립변수가 1개이며, 독립변수가 종속변수에 영향을 미친다. 이를 ‘광고비, 매출액’ 의 2개의 변수를 시준으로 살펴보자. 광고비가 매출액에 미치는 영향을 확인하기 위해 독립변수로는 ‘광고비’를 , 종속변수로는 ‘매출액’을 사용해야 한다.

## 4. 회귀식에서의 비표준화 계수(B)와 표준화계수(ß)

회귀식을 토출하려면 [계수]표를 이용해야 한다. [계수] 표에는 비표준화 계수와 표준화 계수가 표시되고 있는데, 회귀식에서는 측정 데이터와 관련이 있는 비표준화 계수를 사용한다. (독립변수가 1개인 단순회귀 모형의 경우, 광고비의 영향만 받기 때문에 측정 데이터 단위를 고려할 필요가 없기 때문에, 비표준화 계수를 사용한다.)

표준화 계수를 사용할 경우를 살펴보자. cm나 m 단위로 측정한 ‘키’라는 변수와 kg이나 g으로 측정한 ‘몸무게’라는 2개의 독립변수가 있다고 가정할 때, 두 변수의 단위가 다르면 크기 또한 다르기 때문에 각각의 수치가 똑같이 2.0으로 나왔다 하더라도 두변수는 같다고 할 수 없다. 이처럼 단위가 다른 변수가 두 개 이상일 경우 각 단위를 통일 시킨 계수로써 표준화 계수를 사용한다. 즉 회귀식을 만들 때는 비표준화 계수가 사용되지만, 독립변수 간의 계수를 비교하여 확인할 때문 표준화된 계수를 사용하여 각 변수들의 영향력을 비교, 판단해야 한다.

# 다중 회귀 분석 1 : 변수 계산을 활용한 분석

## 1. 공차와 VIF

‘공선성 통계량’은 ‘공차와 VIF’ 값으로 다중공선성을 확인할 수 있도록 해주는 결과표이다. 공차는 공차한계(tolerance)를 의미하고, VIF는 분산팽창계수(variance inflation factor)를 의미한다. VIF는 공차와 역수관계이다. VIF는 1~∞(1에서 무한대)의 값을 가지는데, 10 미만이면 다중공선성의 문제가 없다고 판단한다. 여기에서 다중공선성은 연구모형에 포함된 변수들 간의 상관관계를 의미하는데, 상관계수와의 차이는 변수들 간의 다중적인 상관관계를 의미한다. 즉 어느 한 변수의 고유분산에 대해 총분산과의 표준화된 비율을 말한다. 기준이 되는 10을 넘는다는 것은 한 변수가 다른 변수에 의해 중첩되어 나타난다는 의미이다.

# 다중회귀 분석 2 : 요인분석을 통한 분석

## 1. 변수 투입 방법

회귀분석 실행을 위해서는 변수 투입 방법을 선택해야 하는데, 회귀식에서 변수가 유의적인 기여를 하는지의 여부에 따라 선택하는 변수 투입 방법이 달라진다.

1. **입력** : 회귀분석을 실시할 때 독립변수를 모두 한 번에 투입하여 회귀모형을 추정하는 방법이다. 강제로 모든 변수를 투입하게 되므로 유의미한 변수와 유의하지 않는 변수에 대한 모든 정보가 산출된다.
2. **단계선택** : 회귀분석에 투입되는 독립변수들 가운데 설명력이 가장 높은 변수들로 회귀모델을 구성하는 방법이다. 첫 번째 단계에서는 종속변수 간 상관관계가 가장 높은 변수를 투입하고, 두 번째 단계에서는 종속변수간 편 상관관계가 있는 변수들을 투입한다. 그런 다음 각 단계별로 설명력이 높은 변수에 대한 유의성 검증을 시시하며, 유의하지 않는 변수는 제거한다.
3. **제거** : 회귀분석을 실시하면서 연구자가 선택한 변수들이 강제로 제거되어 분석된다.
4. **후진** : 회귀분석에서 모든 독립변수를 포함하여 통계적 기준에 따라 중요도가 가장 낮은 변수부터 하나씩 제거되면서 분석이 진행되는 방법이다. 더 이상 제거할 필요가 없을 때(통계적 기준치) 중단하며, 제거 후 남아 있는 변수들을 중 변수로 채택하여 분석한다.
5. **전진** : 회귀분석을 실시할 때 모든 독립변수를 포함하여 통계적 기준에 따라 중요도가 가장 높은 변수부터 하나씩 추가해 나가는 방법이다. 더 이상 중요한 변수가 없을 때(통계적 기준치) 중단된다. 후진은 ‘낮은 중요도를 제거’, 전진은 ‘높은 중요도를 추가’라고 기억해두자.

## 2. 방법을 ‘후진’으로 했는데도 ‘입력’으로 표시되는 이유

변수의 투입 방법을 선택할 때 ‘후진’을 선택했음에도 결과표에는 ‘입력’으로 되어 있다. 이때 혹시 뭔가 잘못된 것이 아닌가 하는 의심이 들 수도 있다. ‘후진’은 처음에는 모든 변수를 다 투입하는 ‘입력’의 방법을 자동으로 선택하여 분석하다가 회귀식에 유의하지 않은 변수가 있으면 하나씩 삭제하여 분석하는 방법이므로, 유의하지 않은 변수가 없는 경우에는 ‘후진’의 방법이 나타나지 않는다. 만약 ‘입력/제거된 변수’ 표에서 ‘FAC\_외관’이 유의하지 않는 변수라 가정하면 [입력/제거된 변수]표가 나타난다.

즉 첫 번째 단계(‘FAC\_외관’이 유의하지 않는 변수라고 가정할 때)에서는 모든 변수를 투입하는 ‘입력’의 방법으로 첫 번째 모형을 만들고, 유의하지 않는 변수인 ‘FAC\_외관’ 변수를 제외하고 새로운 모형인 두 번째 모형을 만들어서 다시 분석을 실시한다. 또한 실시하는 회귀분석은 이러한 1의 모형과 2의 모형에 대해 모두 분석을 실시하며, 둘 사이의 차이를 비교할 수 있도록 해준다. 이러한 이유로 인하여 전체 변수를 모두 투입하는 ‘입력’의 방법보다는 유의방법은 변경할 수 있다. 어떠한 방법을 선택할지는 연구자가 가장 잘 알 것이다.

## 3. [모형 요약] 표의 항목 의미

[모형 요약] 표에는 R과 R2 및 수정된 R2에 대해 분석된 값이 표시된다. 모두 R로 표기되어 혼동하기 쉬운데, 각각의 의미는 다음과 같다.

* R : 독립변수와 종속변수 간의 상관관계를 나타낸다.
* R2 : 회귀선의 설명력을 나타낸다. 즉 만족감을 나타내는 회귀선의 설명력을 의미한다.
* 수정된 R2 : 독립변수의 수와 표본의 크기를 고려하여 조정한 R2 값이다.

## 4. 지수형태의 계수 표시 방법

위 회귀식을 보면 상수항이 ‘-1.087E-07’ 으로 표기되어 있는데, 여기에 쓰인 ‘E’는 지수(Exponential)를 의미한다. 즉 ‘-1.087x10-7과 같은 표현으로, ‘-0.0000001087’로 이해하면 된다. 또한 ß0는 절편에 해당하므로 상수항이므로 유의하지 않은 값이라도 회귀식에 포함시켜야 한다. 때문에 유의확률은 p=1.000이지만 회귀식에 포함시킨다.

## 5. 고유값, 상태지수, 분산비율

고유값(eigen-value)은 독립변수들의 곱셈값에 대한 행렬을 요인분석하여 구하며, 상태지수는 ‘가장 큰 고유값’을 그에 해당하는 차원의 고유값으로 나눈 값의 제곱근이다. 상태지수값이 15보다 크면 다중공선성에 문제가 있다고 판단할 수 있다. 다중공선성에 문제가 생기면 분산비율을 확인해야 한다. 특히 하나의 고유값에 대한 두개 이상의 독립변수가 높게 설명된다면 이들 간에 다중공선성의 문제가 있다고 판단할 수 있다. (개념상으로 보면 어렵겠으나 상태지수를 기준으로 공선성에 대한 문제가 없다면 크게 신경쓰지 않아도 된다.)