회귀분석을 이용한 사회과학자료의 분석

서울대학교 심리학과, 인지과학협동과정 교수
조사연구 편집위원장
조사연구학회 이사
서울대 사회과학대학 교무부학장 역임

주요 연구 :
Self-efficacy in information security : Its influence on end users' information security practice behavior
When fit indices and residuals are incompatible.
LSA모형에서 다의어 의미의 표상

회귀분석을 이용한 사회과학자료의 분석

초판1쇄 발행 | 2011년 ♣월 ♣일

지은이 김청택

편집 · 인쇄 민속원
주소 서울 마포구 대흥동 337-25
전화 02) 804-3320, 805-3320, 806-3320(代) 팩스 02) 802-3346
이메일 minsok1@chollian.net 홈페이지 www.minsokwon.com

회귀분석을 이용한 사회과학자료의 분석

김청택

민 속 원

머리말

회귀분석은 독립변수와 종속변수 사이의 관계를 선형식(일차함수)을 사용하여 기술하고 예측하는 방법이다. 회귀분석을 통하여 두 변수 사이의 관계를 설명하는 함수를 추정할 수 있으며, 추정의 정밀도를 계산할 수 있다. 많은 사회과학연구에서 특정한 결과를 예측할 수 있는 변수를 찾는 것에 목적을 두고 있으므로 회귀분석은 사회과학연구에서 중요할 수 밖에 없다. 어떤 사회를 권위주의적으로 만드는 요인들이 무엇인지, 어떤 학생들이 다른 학생들보다 공부를 잘 하는지, 즉 개인의 어떤 특성들이 학생들의 성적을 예측할 수 있는지 등을 연구하기 위해서 회귀분석을 사용해야 한다. 회귀분석은 그 자체로도 많이 사용되는 통계적 기법 중에 하나지만, 또한 선형모형의 기본이 된다는 점에서 중요하다. 분산분석, 로지스틱 회귀분석, 위계적 선형모형, 시계열분석 등과 같은 통계기법을 이해하기 위해서는 회귀분석에 대한 이해가 필수적이다.

이 책은 짧은 시간 안에 회귀분석에 대한 기본 개념과 적용 방법을 이해하게 하는 목적으로 저술되었지만 회귀분석을 단순히 기계적으로 적용시키는 방법을 나열하고 있지는 않다. 소위 요리책과 같은 방식으로 통계적 분석법을 기술하면, 통계 방법에 대한 이해없이도 결과를 얻을 수 있지만, 대부분의 경우에는 잘못된 해석에 도달하게 된다. 여기에서는 회귀분석을 이해하고 결과를 정확하게 해석하기 위하여 필수적인 사실들과 논리들을 설명하고 있다. 이러한 점에서 이 책을 처음부터 끝까지 이해하는 것이 중요하다. 새로운 기계를 구입하면, 그 제품의 설명서가 나온다. 그 설명서에서 제품을 사용하기 전에 반드시 설명서를 읽어야 된다는 경고문을 자주 발견할 수 있다. 이 책도 회귀분석을 사용하기 전에 읽어야 되는 설명서 쯤으로 이해하면 좋을 것 같다. 다만, 읽지 않고 사용하였을 때의 치룰 수 있는 위험은 다른 제품보다도 훨씬 크다고 하겠다.

목차

일러두기

이 책은 통계학의 기본개념을 이해하고 있다는 가정 하에서 저술되었다. 즉 확률, 점추정, 구간추정, 가설검증의 개념을 이해하고 적용할 수 있어야지 이 책에 대한 이해가 가능하다. 그렇지 않은 독자들은 기초 통계학 책의 앞 부분을 참조하기 바란다. 회귀분석에 대한 이해를 돕기 위하여 몇 가지 자료가 사용되었지만, 이는 모두 실제 자료는 아니다. 또한 여기에서 사용된 SPSS 통계패키지는 예시를 하기 위하여 사용된 것이다, SAS, R, Stata 등의 통계패키지에서도 SPSS에서 분석할 수 있었던 내용을 분석할 수 있으며, 어떤 경우에는 더 많은 정보를 제공하기도 한다.

회귀분석

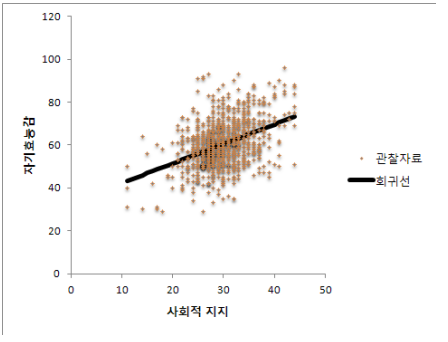
1. 변수들 사이의 관련성에 대한 연구

많은 사회과학연구들은 변수들 간의 관계에 대하여 관심을 가지고 있다. 예컨대, 학력이 높아지면 수입이 높아지는지, 수입이 높아지면 더 보수적이 되는지 등에 대하여 조사하고자 한다. 이러한 변수들 간의 관계를 탐구할 때, 특정한 변수를 설명하거나 예언하려고 하는 경우가 대부분이다. 한 연구자가 개인의 수입에 대하여 관심이 있다면, 수입을 예측하기 위하여 여러 가지 변수들을 찾을 수 있다. 다른 연구자가 사람들의 보수성향–진보성향에 대하여 관심이 있다면, 진보–보수를 예측할 수 있는 여러 가지 변수를 찾을 수 있고 그 중 하나가 수입이 될 수도 있다.

연구자가 설명하고자 하는 변수를 결과변수(outcome variable)라 하고 결과변수를 설명 혹은 예측하기 위하여 사용하는 변수를 예측변수(predictor)라 한다. 보다 엄격히 통제된 실험 연구에서는 전자를 종속변수(dependent variable), 후자를 독립변수(independent variable)라 부른다. 독립변수는 실험자의 설계에 의하여 변화하는 변수이고 종속변수는 독립변수의 변화에 종속되어 변하는 변수이다.

회귀분석은 독립변수(예측변수)를 이용하여 종속변수(결과변수)를 예언하는 분석 방법이다. 아래의 그래프를 잠시 살펴보기로 하자. 아래의 자료는 인터넷의 중독 행동에 대하여 연구하기 위하여 1,000명의 표본에서 수집된 자료이다. 연구를 위해 여러 가지 변수들이 측정되었지만, 여기에서는 사회적 지지(social support)의 정도가 인터넷에 대한 자기 효능감(effectiveness)에 영향을 미치는지를 연구하고자 한다. 사

회적 지지란 가족, 친지, 동료 등 주변의 사람들로부터 신체적 정서적 위안을 받는 것을 의미하고, 자기효능감이란 주어진 상황에서 원하는 결과를 얻기 위해 적절히 행동할 수 있다는 기대와 신념이다. 아래의 그래프에 1,000명의 자료가 제시되어 있다. 한 표본은 하나의 점으로 표시되어 있다. 그래프를 훑어보면, 대체적으로 사회적 지지가 증가함에 따라 자기 효능감이 증가하고 있는 것을 알 수 있다. 그리고 이 관계를 가장 잘 대표할 수 있는 직선을 하나 그었다. 이 직선을 회귀선이라 한다. 회귀분석은 이 회귀선을 찾아내고, 이 회귀선이 실제 자료를 얼마나 잘 나타내고 있는지에 대한 정보를 제공하기 위한 통계적 모형이다.



〈그림 1〉 자료의 표현과 회귀선

2. 상관계수

1) 공분산

두 변수의 관련성을 계산하기 위해서는 한 사람에게서 두 개의 자료를 구해야 된다. 예컨대 키와 몸무게, 지능지수와 성적 등의 두 변수를 한 사람에게서 관찰해야 한다. i 번째 사람의 두 자료를 X_i와 Y_i라 하면, 이 두 변수의 관련성을 나타내는 통계량(statistic)으로는 공분산(covariance)이 있다. 공분산의 공식은 다음과 같다.

$$Cov(X, Y) = \frac{\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})}{N}$$

이 공식의 분자를 살펴보면, X와 Y가 각각 X의 평균과 Y의 평균보다 크면 양수이고, X와 Y가 각각 X의 평균과 Y의 평균보다 작아도 양수가 된다. 반면, X가 X의 평균보다 크고 Y는 Y의 평균보다 작으면 음수이다. 이 공식의 특성상, X가 커질 때 Y도 커지고, X가 작아질 때 Y도 작아지면 양수가 되고, 한 변수가 커질 때 다른 변수가 작아지면 음수가 되며, 특정한 패턴이 없으면 0에 가까운 수치가 된다. 공분산은 두 변수의 관계를 기술하는 데 적합한 통계치이다.

공분산의 한 가지 문제점은 척도 의존적(scale-dependent)이라는 것이다. 척도가 바뀌면 공분산의 크기도 달라진다는 의미이다. 위의 공식을 이용하여 키와 몸무게의 공분산을 구하는 경우를 생각하여 보자. 키를 cm로, 몸무게를 g으로 계산한 경우와 키를 m로 몸무게를 kg으로 계산한 경우에 공분산의 크기는 달라진다. 전자의 공분산이 후자의 공분산보다 100,000배가 크게 된다. 따라서 공분산의 수치만으로 두 변수간의 관계를 추론하기는 불가능하다.

2) 상관계수(Correlation)

공분산이 척도의존적이라는 단점을 보완하기 위하여 사용되는 것이 상관계수이다. 상관계수는 공분산을 각 변수의 표준편차로 나눈 값이다.

$$r = \frac{Cor(X, Y)}{SD(X)SD(Y)}$$

여기에서 SD는 표준편차를 나타낸다. 위의 예로 돌아가서 cm를 m로 바꾸면 분자의 크기가 100분의 1로 줄어들지만 분모도 역시 100분의 1로 줄어들기 때문에 척도가 변경되어도 상관계수의 값은 변화하지 않는다.

상관계수의 크기는, 두 변수의 모든 관측치가 모두 XY 그래프에서 하나의 직선상에 놓여 있을 때 +1이나 −1이 된다(단 수평선의 경우는 제외). 이 때 한 변수의 값

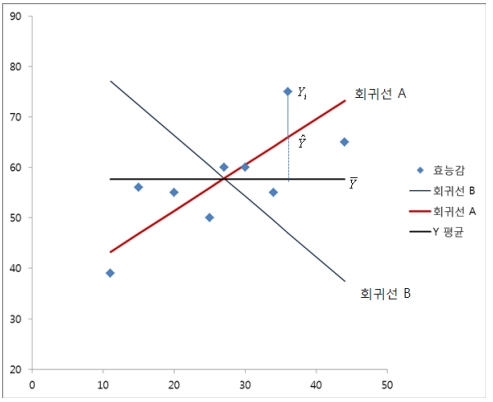
을 알면 다른 변수의 값을 정확하게 알 수 있다. 즉, 직선이 양의 기울기면 +1, 음의 기울기면 −1이다. 상관계수가 0의 경우는 X와 Y에 특정한 직선을 그릴 수 없다.[1] 0과 1 사이의 값은 회귀선이 양의 기울기를 가지며, 자료가 직선에 가깝게 형성되어 있으면 1에 가까운 상관계수를 가지고 그렇지 않은 경우에는 0에 가까운 상관계수를 가진다.

3. 단순 회귀분석

1) 회귀선

회귀분석은 〈그림 1〉에서 X와 Y의 관계를 가장 잘 기술할 수 있는 직선을 구하는 분석방법이다. 아래의 그래프를 다시 한 번 살펴보자. 자료의 설명을 위하여 회귀선 A와 회귀선 B를 그어 보았다. 어느 것이 좋은 것으로 판단되는가? 당연히 회귀선 A이다. 좋은 회귀선을 찾아내기 위해서는 "어떤 직선이 관계를 잘 기술한다"라는 말을 수리적으로 기술하여야 한다. 여러 가지 수리적 기술방법이 있겠지만 회귀분석에서는 X에 해당하는 실제 Y의 값과 X가 주어졌을 때 직선(일차함수) 상에 있는 Y값의 차이의 제곱합을 최소화하는 직선을 가장 관계를 잘 설명하는 직선으로 채택한다.

1 수평선을 그릴 수 있는 경우(Y=c, c는 상수)도 상관계수는 영이다.



〈그림 2〉 자료와 회귀선

〈그림 2〉에서 한 자료 y_i에서, y_i와 $\hat{y_i}$의 차이는 회귀선 A를 선택하였을 때, 회귀선으로 설명하지 못하는 부분이다. 만약에 모든 자료가 회귀선상에 있다면 이 값은 모두 0이 될 것이다. 이 차이를 최소화하는 직선을 구하면 된다.

먼저 회귀선을 $\hat{y_i} = a + bx_i$라 하면 회귀선에 의하여 설명되지 못하는 부분을 e라 하면 e는 다음과 같다.

$$e_i = (y_i - \hat{y_i}) = (y_i - a - bx_i)$$

e들의 제곱합을 최소화시키는 직선을 찾아서 회귀선으로 택하는 방식을 최소자승법(Ordinary Least Square; OLS)이라 한다. 대부분의 회귀분석에서는 이 방법으로 회귀식을 구하고 있다.

$$L = \sum_{i=1}^{N} e^2 = \sum_{i=1}^{N} (y_i - \hat{y_i})^2 = \sum_{i=1}^{N} (y_i - a - bx_i)^2$$

〈수식 1〉

위의 식에서 x_i와 y_i는 주어진 자료로, 알려져 있는 수치이다. OLS에서는 어떤 a와 b가 수식 1을 가장 최소로 만드는지를 수학적으로 찾는다. L을 a와 b에 대하여 미분하여 0이 되는 지점을 찾으면 된다.

여기서 이 책에서 사용할 표기방법에 대하여 간단히 약속을 하기로 하자

SD_x : x의 표준 편차, 따라서 SD_x^2은 x의 분산이 된다.

SD_{xy} : x와 y의 공분산이다. 이 공분산을 각각의 표준편차인 SD_x, SD_y로 나누면 상관계수가 된다.

r_{xy} : x와 y의 상관계수이다. 맥락에 따라 혼돈이 없는 경우에는 첨자를 생략할 것이다.

L을 a에 대하여 미분하여 0이 되는 지점을 찾으면 다음과 같다.

$$\frac{dL}{da} = -2\sum(y_i - a - bx_i) = 0$$

$$\sum y_i - an - b\sum x_i = 0$$

$$\bar{y} - a - b\bar{x} = 0 \quad (n\text{으로 나누기} : \frac{\sum y_i}{n} = \bar{y})$$

$$a = \bar{y} - b\bar{x}$$

〈수식 2〉

이 결과에서 알 수 있는 것은 회귀선은 $\bar{x}$와 $\bar{y}$를 지난다는 것이다. 즉 회귀선은 X의 평균과 Y의 평균을 항상 지나게 되어 있다.

L을 b에 대하여 미분하여 0이 되는 지점을 찾으면 다음과 같다.

$$\frac{dL}{db} = -2\sum(y_i - a - bx_i)x_i = 0$$

$$\sum x_i y_i - a\sum x_i - b\sum x_i^2 = 0$$

$$\sum x_i y_i - (\bar{y} - b\bar{x})\sum x_i - b\sum x_i^2 = 0$$

$$b(\sum x_i^2 - \frac{(\sum x_i)^2}{n}) = \sum x_i y_i - \bar{y}\sum x_i$$

$$bSD_x^2 = SD_{xy}$$

$$b = \frac{SD_{xy}}{SD_x^2} \quad \left(r = \frac{SD_{xy}}{SD_x SD_y}\right)$$

$$b = \frac{r SD_x SD_y}{SD_x^2} = r\frac{SD_y}{SD_x}$$

〈수식 3〉

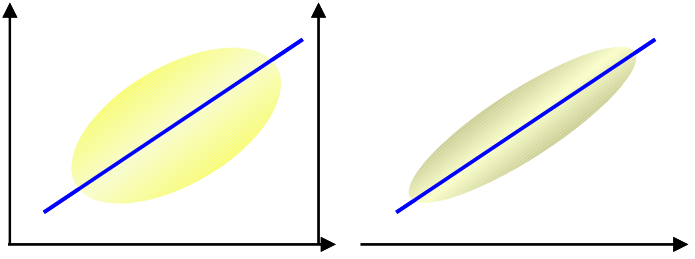
이 식에서 회귀식의 기울기가 구해진다. 기울기는 크게 세 가지 요소에 의하여 결정됨을 알 수 있다. 첫째는 두 변수간의 상관계수이다. 상관이 높을수록 기울기의 절대값은 커진다. 그 다음으로 x와 y의 표준편차의 비이다. 두 표준편차가 동일할 때는 기울기가 상관계수와 같고 y의 표준편차가 x의 표준편차보다 클수록 기울기는 증가하게 된다.

요약하여 정리하면 다음과 같다. 회귀식은 $\hat{y_i} = a + bx_i$인데 이때 기울기는 $b = r\frac{SD_y}{SD_x}$이고 절편은 $a = \bar{Y} - b\bar{X}$이다.

2) 결정계수

회귀식이 정해지면, 회귀식이 얼마나 자료를 잘 설명할 수 있는지를 나타내는 통계치를 계산하는 것이 필요하다. 이를 결정계수(coefficient of determination)라 한다.

아래의 그림에서 보면 동일한 직선을 가지는 회귀식이라도 서로 다른 특성을 가진 자료에서 나왔다는 것을 알 수 있다. 오른쪽의 회귀식이 왼쪽의 회귀식보다 자료를 더 잘 설명하고 있다. 따라서 오른쪽 회귀식을 사용할 때가 왼쪽의 회귀식을 사용할 때보다 훨씬 정교한 예측을 할 수 있다.



결정계수는 y의 전체 분산 중에서 회귀식에 의하여 설명되는 부분이 얼마인지를 계산하여 정해진다. 〈그림 2〉를 다시 돌아가면 y_i의 편차는 다음과 같이 표현될 수 있다.

$$y_i - \bar{y} = (y_i - \hat{y_i}) + (\hat{y_i} - \bar{y})$$

이 편차의 제곱합을 n으로 나누면 분산이 된다. 먼저 제곱합을 계산하면 다음과 같다.

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y_i})^2 + \sum(\hat{y_i} - \bar{y})^2 + 2\sum(y_i - \hat{y_i})(\hat{y_i} - \bar{y})$$

$\sum(y_i - \hat{y_i})(\hat{y_i} - \bar{y}) = 0$((식2)로부터 0이 된다.)

양변을 n으로 나누면,

$$\frac{\sum(y_i - \bar{y})^2}{n} = \frac{\sum(y_i - \hat{y_i})^2}{n} + \frac{\sum(\hat{y_i} - \bar{y})^2}{n}$$

〈수식 4〉

위의 식에서 $\frac{\sum(\hat{y}_i-\bar{y})^2}{n}$은 회귀식에 의하여 설명될 수 있는 분산이고, $\frac{\sum(y_i-\hat{y}_i)^2}{n}$은 회귀식에 의하여 설명되지 않는 분산이다. 따라서 y의 분산은 회귀식에 의하여 설명되는 분산과 회귀식에 의하여 설명되지 않는 분산으로 분할될 수 있다. 그러면 회귀식에 의하여 설명되는 분산을 계산하여 보자.

$$\begin{aligned}
\frac{\sum(\hat{y}-\bar{y})^2}{n} &= \frac{\sum(a+bx_i-\bar{y})^2}{n} \\
&= \frac{\sum(\bar{y}-b\bar{x}+bx-\bar{y})^2}{n} \\
&= \frac{b^2\sum(x-\bar{x})^2}{n} \\
&= r^2\cdot\frac{SD_y^2}{SD_x^2}\cdot SD_x^2 \qquad \leftarrow b^2 = r^2\cdot\frac{SD_y^2}{SD_x^2} \\
&= r^2\cdot SD_y^2
\end{aligned}$$

〈수식 5〉

회귀선에 의하여 설명되는 분산, $r^2SD_y^2$는 y의 분산에 상관계수를 제곱하여 곱한 값이다. 결정계수는 y의 분산 중에서 회귀선에 의하여 설명되는 부분의 비율이기 때문에 $r^2SD_y^2$를 y의 분산으로 나누면 된다. 따라서 결정계수는 r^2이 된다. 독립변수가 하나인 단순회귀에서는 결정계수가 상관계수의 제곱으로 계산될 수 있다. 위의 예에서 자기효능감과 사회적지지의 상관계수가 .8이라면 자기 효능감의 분산 중 64%가 사회적지지에 의해 설명될 수 있다고 해석할 수 있다.[2]

〈수식 4〉로 돌아가면 다음과 같이 되어 회귀선에 의하여 설명되지 않는 y의 분산 비율은 $1-r^2$가 된다.

2 엄격한 의미에서 해석하면, 자기 효능감 분산의 64%가 사회적 지지와 자기효능감의 선형적인 관계에 의하여 설명될 수 있다는 것이다.

$$\frac{\sum(y_i-\bar{y})^2}{n}=\frac{\sum(y_i-\hat{y_i})^2}{n}+\frac{\sum(\hat{y_i}-\bar{y})^2}{n}$$

$$SD_y^2=(1-r^2)SD_y^2+r^2\cdot SD_y^2$$

〈수식 6〉

3) 표준회귀계수와 회귀현상

(1) 표준회귀계수의 계산

회귀분석에서 가장 관심이 있는 것은 회귀선의 기울기와 결정계수일 것이다. 이 두 통계치가 두 변수의 관계를 설명하고 있는 것이다. 그런데 회귀선의 기울기는 x와 y의 표준편차에 의하여 변한다. 예를 들면, 키로 몸무게를 예측하는 모형에서 몸무게를 g으로 나타낼 때와 kg으로 나타낼 때 각각 기울기가 달라지기도 한다. 사회과학의 많은 연구에서 x와 y 척도의 절대적인 크기는 의미가 없는 경우가 많다. 위의 예에서 자기 효능감 점수가 25점이라는 것이 얼마나 높은 점수인지에 대하여 추론할 수 없다. 이때에는 평균과 표준편차가 주어지거나 표준점수의 형식으로 제시되어야 크기를 추론할 수 있다. 마찬가지로 기울기가 2 혹은 100이라는 것을 해석하기 위해서는 자기 효능감의 표준편차, 사회적지지의 표준편차의 정보가 제공되어야 한다. 이러한 불편을 덜어주기 위하여 사회과학에서는 표준회귀계수를 많이 사용한다.

표준회귀계수는 X와 Y를 평균이 0이고 표준편차가 1인 표준점수로 변환한 다음 회귀분석을 하여 구한 기울기이다. 만약 여기에서 기울기가 0.8이라면, X가 1 표준편차만큼 증가하면, Y는 0.8 표준편차만큼 증가한다고 해석할 수 있다. 표준회귀계수를 구하는 방법은 간단하다. 먼저 기울기를 구하면 다음과 같다. 기울기는 $b=r\frac{SD_y}{SD_x}$인데 x와 y의 표준편차가 1이므로, 기울기는 상관계수와 동일하게 된다. 또한 회귀식은 $(\bar{X},\bar{Y})$를 지나게 되어 있으므로 (0, 0)을 지나야 되고 절편은 0이 된다. 따라서 표준회귀식을 나타내면 다음과 같다.

$$\hat{Z}_{y_i}=rZ_{x_i}$$

(2) 회귀현상

회귀식의 기울기를 보면 중요한 특징이 있다. X와 Y의 표준편차가 동일할 때 기울기는 상관계수와 동일하게 되므로, 기울기는 절대로 1보다 클 수 없다는 것이다. 예컨대 아버지의 키로 아들의 키를 예측하면 기울기는 1보다 클 수 없다는 것이다.[3] 즉 아버지의 키가 평균보다 10cm미터 크면, 아들의 키는 평균보다 10점만큼은 크지 않다는 것이다. 회귀선은 평균으로 회귀하는 현상이 있다.

이는 회귀분석기법을 발견한 Galton이 제시한 유전적인 설명으로는 타당하지만, 통계적인 모형의 관점에서 보면 바람직하지 못한 특성이다. 아들의 키로 아버지의 키를 예측하든지, 아버지의 키로 아들의 키를 예측하든지 상관없이 기울기는 항상 1보다 작다. 예컨대 아들의 키가 아버지의 키를 예측하는 기울기가 0.9라면 다음과 같은 관계가 성립한다.

$$\text{아들} = a + 0.9 * \text{아버지}$$

그렇다면 다음이 성립하여 아들이 아버지를 예측할 때는 기울기가 1보다 커야 한다.

$$\text{아버지} = \frac{a}{0.9} + \frac{1}{0.9} * \text{아들}$$

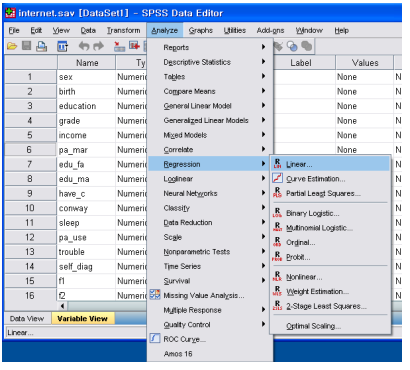
그런데, 아들로 아버지를 회귀식으로 예측하면 여전히 1보다 작은 기울기가 나온다. 회귀분석에서 독립변수와 종속변수를 어떻게 설정하느냐에 따라 결과가 다르게 나올 수 있음을 보여주는 현상이다. 이러한 점을 항상 염두에 두고 자료를 해석해야 한다.

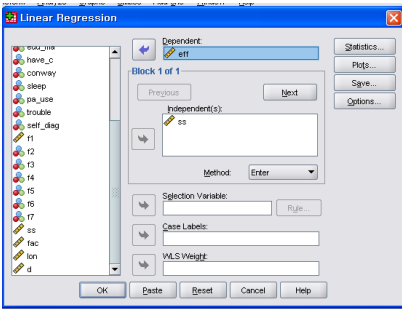
3 여기에서 아들과 아버지의 키 분포에서 표준편차는 동일하다고 가정한다.

4) SPSS를 통한 분석

SPSS, SAS, Stata, R 등의 소프트웨어에서 회귀분석 프로그램은 모두 포함되어 있고 이를 적용시키는 방법은 매우 단순하다. 여기에서는 SPSS를 통하여 회귀분석을 하는 방법을 살펴보기로 하겠다.

SPSS 자료 파일에 효능감이 eff로 사회적지지가 ss로 코딩되어 있다고 하자. 먼저 Analyze → Regression → Linear Menu로 들어가면 다음과 같이 메뉴가 나타난다.





왼쪽 창에 나타난 변수들 중에서 독립변수와 종속변수를 정한 다음 OK 단추

를 누르면 다음과 같은 회귀분석 결과가 나타난다.

Model Summary

Model	R	R Square	Adjusted R Square	Std.Error of the Estimate
1	.413[a]	.171	.170	9.428

a .Predictors:(Constant) ss

위의 결과표는 결정계수(보통 R^2으로 표기)를 제시하고 있다. 이 결과는 사회적 지지가 효능성 분산의 17.1%를 설명하고 있다는 것을 나타낸다. R은 결정계수의 제곱근이며, 이 경우에는 두 변수의 상관계수와 동일하다. Adjusted R square는 뒷부분에서 설명하기로 하겠다.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18254.154	1	18254.154	205.346	.000[a]
	Residual	88716.802	998	88.895		
	Total	106970.956	999			

a. Predictors:(Constant), ss
b. Dependent Variable:eff

두 번째 결과표는 회귀식에 대한 분산분석(ANOVA)의 결과를 보여주고 있다. 이는 "회귀식이 자료를 전혀 설명하지 못하다"는 영가설에 대한 가설검증이다. 이는 R^2의 모수치인 P^2이 0이라는 영가설에 대한 가설 검증 혹은 회귀식의 기울기가 0이라는 영가설에 대한 가설검증이다. 이 두 영가설은 회귀분석에서는 동일하다. p-값이 .05보다 작아서 영가설을 기각하여야, 회귀식이 X와 Y사이의 관계를 설명할 수 있는 부분이 있다는 것이다. 위의 결과에서는 p-값(Sig.)이 .05보다 작으므로, 회귀식이 종속변수의 분산을 일부 설명할 수 있다고 해석할 수 있다.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std.Error	Beta		
1	(Constant)	33.267	1.899		17.520	.000
	ss	.909	.063	.413	14.330	.000

a. Dependent Variable:eff

마지막으로 회귀계수가 제시된다. B열에 절편과 기울기를 나타낸다. 여기에서 회귀식은 $\hat{y_i} = 33.267 + .909x_i$이다. 그다음 열은 회귀계수들에 대한 표준오차(standard error)가, 그다음 열은 표준화된 회귀계수가 제시되어 있다. 이 값은 X와 Y의 상관계수와 동일하며, R과도 동일함을 알 수 있다. 이는 단순회귀식에서만 나타나는 현상이다. 다음과 t값과 t값에 대한 p-값(Sig.)이 제시되어 있다. 이 검증은 각 회귀계수가 0이라는 영가설을 검증하는 것이다. 따라서 공식은 $t = \frac{B-0}{SE_B}$이다.

$$t = \frac{B-0}{SE_B} = \frac{33.267-0}{1.899} = 17.520 \text{ (절편)}$$

$$t = \frac{B-0}{SE_B} = \frac{.909-0}{.063} = 14.330 \text{ (기울기)}$$

특히 기울기가 0과 같지 않다는 가설검증에 관심을 가지게 되는데, 이 기울기에 대한 가설검증과 회귀식에 대한 가설검증은 단순회귀에서는 동일하다. t-값을 제곱하면 위에서 ANOVA의 F값과 동일하게 된다. 위의 결과에서 표준회귀계수를 해석하면, 사회적지지가 1표준편차만큼 증가하면, 자기 효능감은 .413 표준편차만큼 증가한다. 그리고 이 기울기는 통계적으로 0과 다르다.

4. 다중 회귀분석

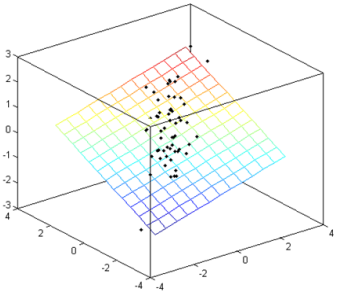
1) 회귀선

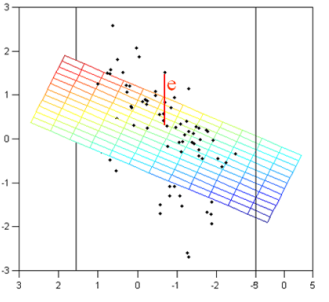
단순회귀분석에서는 종속변수를 예측하기 위하여 하나의 독립변수를 사용하였는데 반해 다중회귀분석에서는 두 개 이상의 독립변수를 사용한다. 회귀식은 다음과 같다.

$$\hat{y_i} = b_0 + b_1 x_{1i} + b_{2i} x_{2i} + \ldots + b_k x_{ki}$$

〈수식 7〉

여기에서 b_0는 절편, b_1에서 b_k는 k개의 독립변수들의 기울기에 해당된다. 즉 다시 말하면 b_2는 두 번째 독립변수인 x_2가 1점 증가할 때 종속변수가 b_2만큼 증가한다는 것을 의미한다. 기하학적으로 보면, 각 개인은 독립변수 k개와 종속 변수 1개로 이루어진 (k+1)차원 상에서 한 점으로 표현된다. 회귀분석은 이 자료를 k차원상의 공간으로 투사시키는 것이다. 예컨대 독립변수가 두 개인고 종속변수가 하나인 자료는 아래의 그림에서 보는 바와 같이 삼차원상에서 표시된다. 회귀분석에서는 이 자료들을 2차원인 평면상에서 자료를 투사시키게 된다. 삼차원 상의 자료를 2차원 상에서 표현하면 설명하지 못하는 부분이 생기기 마련이다. 이를 아래의 그래프에서 e로 표시되어 있다.





단순회귀분석과 동일하게 e의 제곱합을 최소화시키는 방식으로 축소된 공간을 찾아내게 된다. 이 경우에는 평면을 찾아내게 되는 것이다. 즉 다음의 L를 최소화 시키게 된다.

$$L=\sum_{i=1}^{N} e^2 = \sum_{i=1}^{N} (y_i - \hat{y_i})^2 = \sum_{i=1}^{N} (y_i - b_0 - b_1 x_{1i} - \ldots - b_k x_{ki})^2$$

〈수식 8〉

이에 대한 해를 구하는 방식도 기본적으로 단순회귀분석에서 구하는 방식과 동일하게 OLS를 사용한다. 행렬에 대한 미분이 필요하기 때문에 여기에서는 생략하기로 한다. 각 기울기는 컴퓨터 프로그램을 통해서 쉽게 계산될 수 있다.

다중회귀분석에서는 기울기를 해석할 때는 다소 주의를 요한다. 예컨대 b_2의 기울기는 다른 독립변수가 모두 고정되어 있을 때 x_2가 1점 증가할 대 종속변수가 얼마만큼 증가하는지를 나타내는 것이다. 그런데 x_2가 다른 독립변수와 전혀 상관이 없는 경우에는 다른 독립변수와 무관하게 x_2가 종속변수에 어떻게 영향을 미치는 지로 해석할 수 있지만, x_2가 다른 독립변수와 상관이 있는 경우에는 해석에 유의하여야 한다. 좀 더 자세한 이해를 위하여 독립변수가 두 개인 회귀분석에서 기울기를 구하는 공식을 살펴보자.

$$b_1^* = \frac{r_{YX_1} - r_{X_1X_2} r_{YX_2}}{1 - r_{X_1X_2}^2}$$

$$b_2^* = \frac{r_{YX_2} - r_{X_1X_2} r_{YX_1}}{1 - r_{X_1X_2}^2}$$

$$b_j = b_j^* \left(\frac{SD_y}{SD_{X_j}} \right)$$

〈수식 9〉

여기에서 b_j는 j번째 독립변수의 기울기를 나타내며, b_j^*는 표준화된 기울기를 나타낸다. r은 상관계수, SD는 표준편차이다.

먼저 b_1^*을 살펴보자. 이 기울기를 결정하는 데 x_1과 Y의 상관계수가 영향을 미친다. 이는 단순회귀분석의 기울기와 같다. 그런데 그 밖에도 x_1과 x_2의 상관계수, x_2와 Y의 상관계수도 X_1의 기울기에 영향을 미치고 있다. 즉 x_1과 Y의 관계를 나타내는 기울기에 다른 변수도 관여하고 있다.

두 독립변수의 상관계수가 0일 때만 b_1^*이 X_1과 Y에 의 상관계수에 의해서만 결정된다. 또한 x_1과 x_2의 상관계수가 1일 때는 두 기울기가 계산될 수 없다는 것도 기억해야 한다. 뒤에 공선성(collinearity)의 주제에 대하여 다룰 때 다시 언급될 것이다.

2) 결정계수

다중회귀분석에서도 종속변수의 전체 분산 중에서 회귀식에 의하여 설명될 수 있는 분산의 비율인 결정계수가 단순회귀와 동일한 방식으로 계산될 수 있다. 단순회귀에서는 X와 Y의 상관계수의 제곱이 결정계수였는데, 다중회귀분석에서는 y와 $\hat{y}$의 상관계수의 제곱이 결정계수가 된다.

결정계수가 통계적으로 유의미한지는 다시 말하면 Y를 회귀식에 의하여 설명할 수 있는 부분이 있는지를 검증하는 ANOVA 검증으로 할 수 있다. 이때 F-값은 다음과 같이 되고, 자유도는 (k, n-k-1)이다(여기서 k는 독립변수의 수이다). 이 공식은 물론 단순회귀분석에도 적용될 수 있다.

$$F = \frac{R^2}{SD_R^2} = \frac{R^2 / k}{(1-R^2)/(n-(k+1))}$$

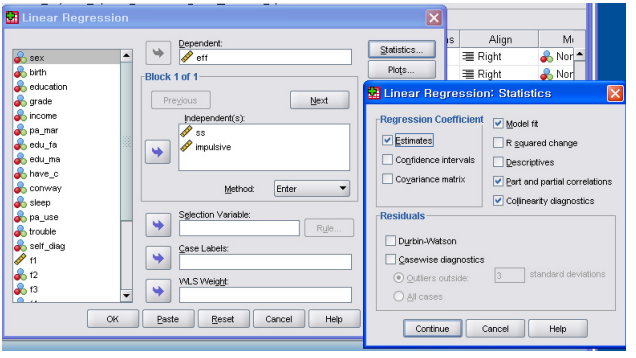
다중회귀분석의 R^2는 모집단의 파라미터인 P^2에 대한 편중 추정치로, 과대추정하게 된다. 또한 독립변수가 많아질수록 과대추정의 정도는 심해진다. 불편추정치인, $\widetilde{R^2}$는 다음과 같이 계산될 수 있다.

$$\widetilde{R}^2 = 1-(1-R^2)\frac{n-1}{n-k-1}$$

$\widetilde{R}^2$는 독립변수가 많아질수록 R에 비하여 축소되는 정도가 심해지기 때문에 Shrinkage R이라고도 한다.

3) SPSS를 통한 분석

SPSS를 통한 회귀분석을 하는 방법은 단순회귀와 동일한데 다만 독립변수의 수가 하나 이상 정의하는 것이 다르다. 아래의 그림에서 단순회귀와 마찬가지로 종속변수를 자기효능감(eff)로 정의하고 독립변수로 사회적지지(ss)와 함께 충동성(impulsive)을 추가로 정의하였다. 이렇게 정의하여 OK단추를 누르면 회귀분석이 계산될 것이다. 여기에서는 이에 더해서 Statistics메뉴로 들어가서 Part and partial correlation과 Collinearity diagnostics를 체크하였다. 이는 뒤에 설명할 유용한 정보를 제공할 것이다.



SPSS의 결과표에서 가장 먼저 등장하는 것이 결정계수에 대한 부분이다. 아래의 표에서 보는 바와 같이 R^2는 .422이고 이에 대한 불편추정치는 Adjusted R

square로 표기되어 있고 .420이다. 불편추정치는 원래의 R^2보다 항상 작으며, 독립변수의 수가 많아질수록 작아지는 정도가 더 커진다는 점은 이미 설명한 바 있다.

Model Summary

Model	R	R Square	Adjusted RSquare	Std.Error of the Estimate
1	.649[a]	.422	.420	7.878

a. Predictors : (Constant), i mpulsive, ss

결과표의 두 번째 부분은 R^2의 모수치에 해당하는 P^2이 0인지에 대한 가설검정이다. 이 영가설을 기각하면, 회귀분석이 Y 자료를 설명하는 부분이 있다고 해석할 수 있다. 아래의 표에서 보는 바와 같이 p-값(.000)이 .05보다 작으므로 영가설을 기각할 수 있다.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45092.338	2	22546.169	363.268	.000[a]
	Residual	61878.618	997	62.065		
	Total	106970.956	999			

a. Predictors : (Constant), impulsive, ss
b. Dependent Variable : eff

그 다음으로 회귀식에 대한 정보가 제시되어 있다. 이 회귀식을 수식으로 나타내면 다음과 같다.

$$\hat{y} = 84.448 + .417x_1 - .692x_2$$

여기에서 x_1, x_2는 각각 사회적 지지와 충동성 점수를 나타낸다. 이 두 독립변수들의 기울기인 .417과 -.692는 통계적으로 유의미한 결과이고 ($p<.05$)이다. 이

결과를 해석하면, 사회적지지가 증가할수록, 충동성이 감소할수록 자기효능감이 증가한다.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std.Error	Beta		
1	(Constant)	84.448	2.928		28.839	.000
	ss	.417	.058	.189	7.180	.000
	impulsive	-.693	.033	-.549	-20.795	.000

a. Dependent Variable : eff

다중회귀분석에서 기울기를 해석할 때, 연구자들이 알고자 하는 정보 중의 하나가 여러 가지 독립변수 중에서 어느 변수가 종속변수에 상대적으로 영향을 많이 주는지 이다. 당장 생각해 볼 수 있는 것이 기울기이다. 기울기가 커지면 Y에 영향을 많이 주는 것이다. 그러나 단순한 기울기(회귀계수)를 비교할 수는 없다. 왜냐하면 독립변수마다 회귀계수의 척도가 모두 다르게 때문이다. 예컨대 SS는 1부터 10까지 변하는 변수이고 impulsive는 1부터 1000까지 변하는 변수라면, 독립변수 1점에 종속변수가 몇 점 증가하는지를 나타내는 두 기울기가 동일한 잣대를 가지고 있지 않다. 두 독립변수가 동일한 잣대(척도)를 가지게 만드는 방법이 두 점수를 표준 점수로 바꾸는 것이다. 사실 회귀분석 프로그램은 표준회귀계수에 대한 기울기도 제공하고 있다. 위의 결과에서 Beta라 표기된 열이다. 결과에 따르면, ss의 기울기는 .189이고 impulsive의 기울기는 -.549이다. 즉 ss가 1 표준편차만큼 증가하면 종속변수 eff는 .189 표준편차만큼 증가하고, impulsive가 1표준편차만큼 증가하면, eff는 .549 표준편차만큼 감소한다. 따라서 충동성이 자기효능감이 사회적지지보다 자기 효능감에 더 큰 영향을 미친다고 해석할 수 있다.

사회과학에서 사용하는 대부분의 척도들은 절대적인 크기가 존재하지 않는다. 애국심, 사회성, 보수성, 진보성 등과 같은 변수는 상대적인 크기를 다루는 것이다. 따라서 표준화되지 않는 회귀계수보다 표준화된 회귀계수가 실질적으로 더 많

이 사용된다. 상관계수와 공선성에 대한 결과도 같이 제시되나 다음 부분에서 다루기로 하겠다.

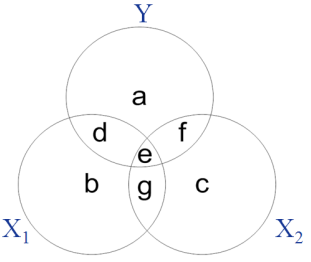
4) 편상관계수와 부분상관계수

다중회귀분석에서는 독립변수가 다수이기 때문에 각 독립변수가 독자적으로 종속변수의 분산을 얼마나 설명할 수 있는지에 대한 관심을 가진다. 이때 사용하는 것이 부분상관계수(semi-partial correlation)와 편상관계수(partial correlation)이다. 이들 상관계수는 아래와 같이 SPSS에서 Patial과 Part(부분상관계수)로 표기되어 제시된다.

Model		Correlations		
		Zero-order	Partial	Part
1	(Constant)			
	ss	.413	.222	.173
	impulsive	-.626	-.550	-.501

a. Dependent Variable : eff

이들 상관계수에 대한 이해를 돕기 위해 벤다이어그램을 이용하기로 하겠다. 각 원은 변수들의 분산을 나타내고, 두 원이 겹치는 부분은 두 변수들 간의 공분산을 나타낸다. 아래의 그림에서 회귀분석에서 사용되는 분산은 다음과 같이 표현될 수 있다.



- Y의 분산 : a+d+e+f
- x_1 에 의해서 설명되는 Y의 분산 : d+e
- x_2 에 의해서 설명되는 Y의 분산 : e+f
- x_1 과 x_2 에 의해서 설명되는 Y의 분산 : d+e+f

x_1과 x_2으로 Y를 설명하는 다중회귀분석에서 R^2는 다음과 같이 표현될 수 있다.

$$R^2 = \frac{d+e+f}{a+d+e+f}$$

다음으로 하나의 독립변수에 의하여 설명되는 부분을 구하여 보자. x_1에 의해서만 설명되는 부분은 d이고 x_2에 의해서만 설명되는 부분은 f이다. 그런데 e는 두 변수에 의해 모두 설명되는 부분이다.

x_1과 Y, x_2와 Y의 부분상관계수의 제곱은 각각 다음과 같다.

$$sr_1^2 = \frac{d}{a+d+e+f} \qquad sr_2^2 = \frac{f}{a+d+e+f}$$

x_1과 Y, x_2와 Y의 편상관계수의 제곱 각각 다음과 같다.

$$pr_1^2 = \frac{d}{a+d} \qquad pr_2^2 = \frac{f}{a+f}$$

공식에서 보는 바와 같이 부분상관계수와 편상관계수에서는 한 변수의 효과를 제거한 다음 다른 변수의 효과를 보고 있다. 예컨대, 사회적 지지와 자기효능감의 부분상관 혹은 편상관계수는 충동성의 효과를 제거하였을 때, 사회적 지지와 자기효능감과의 상관계수를 구하는 것이다. 여기서 유념할 점은 위의 벤다이어그램에서 e는 사실 사회적 지지에 의하여 설명될 수 있는 부분이 되기도 하지만 부분/편상관계수의 계산에서는 제외한다는 것이다.

부분상관계수와 편상관계수가 다른 점은 X_1을 통제한 다음 X_2와 Y의 상관계수를 구할 때, 부분상관계수는 x_1이 Y에 미치는 효과만을 제거하는데, 편상관계수에서는 x_1이 Y에 미치는 효과뿐만 아니라 x_1이 x_2에 미치는 효과까지도 제거한다는 것이다. 이러한 특성 때문에 편상관계수가 부분상관계수 보다 크게 된다.

5) 공선성의 문제

다수의 독립변수를 사용할 때 고려해야 하는 또 하나의 문제는 공선성(collinearity)이다. 이는 독립변수들 사이에 높은 상관이 관찰되는 경우를 말하는 것이다. 〈수식 9〉에서 기울기를 구하는 식에서 두 독립변수의 상관계수가 1이면 기울기가 구해지지 않는다. 이 경우가 완전한 공선성이 발생하는 경우이다. 비록 기울기가 1이 아니라 하더라도 1에 가까운 숫자이면 분모가 아주 작은 숫자가 되어 분자에서 조금만 변하더라도 기울기는 큰 값으로 변하게 된다. 이 경우 기울기의 추정치가 불안정하게 된다.

공선성은 다양한 형태로 나타난다. 한 변수와 다른 변수의 상관이 높은 경우뿐만 아니라, 다수의 독립변수들이 한 독립변수들을 예측할 수 있으면 이것도 공선성에 해당되게 된다(이 경우를 다중 공선성이라 이름 붙인다). 공선성은 독립변수의 상관계수를 관찰하는 것만으로 발견되지 않는 경우가 있다. 대부분의 통계 패키지에는 공선성에 대한 통계치를 제공하고 있다. 가장 대표적인 것이 VIF(Variance Inflation Factor)이다.

$$VIF(b_j) = \frac{1}{1 - R_j^2}$$

여기서 R_j^2는 j번째 독립변수를 j번째 독립변수를 제외한 다른 독립변수들을 이용하여 예측하는 회귀에서의 결정계수이다. 대체적으로 이 값이 5보다 크면 공선성을 의심해야 되고 10보다 크면 공선성이 있다고 판단할 수 있다. 그러나 5와 10은 절대적인 기준이라기 보다는 경험적인 기준이므로 연구자들이 상황에 따라서 판단해야 한다.

아래에 SPSS의 결과가 제시되어 있다. Tolerance는 VIF의 역수이다. 이 결과에 따르면 두 변수 모두 공선성을 걱정할 필요는 없는 것으로 보인다.

Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	ss	,834	1,199
	impulsive	,834	1,199

a. Dependent Variable: eff♣표에 a의 위치가 표시되어 있지 않습니다. 확인해 주세요.

6) 단계적 회귀분석과 위계적 회귀분석

다중회귀분석에서는 여러 개의 독립변수를 사용하게 됨에 따라 단순회귀분석에서 할 수 없었던 새로운 정보를 계산할 수 있었다. 또한 새로운 문제들과 이슈들이 등장하였다. 공선성에서는 상관이 높은 독립변수를 사용할 때 회귀분석의 결과를 불안정하게 만들 수 있다는 문제점이 나타났으며, 부분상관계수와 편상관계수에서는 여러 독립변수들 중에서 한 변수가 독자적으로 종속변수를 얼마나 설명할 수 있는지를 구하는 방법들을 다루었다.

다중회귀분석에서 여러 개의 독립변수를 사용할 때 또 하나 해결해야 하는 것은 적절한 독립 변수를 선택하는 문제이다. 최종적인 모형에서 종속변수를 잘 설명할 수 있는 독립변수는 남겨두고 종속변수를 설명하는데 도움이 되지 않는 변수를 제거하면, 단순하면서도 자료를 잘 설명할 수 있는 모형이 된다. 변수를 선택하는 문제는 연구방법의 문제, 회귀분석의 문제 등과 함께 매우 복잡한 이슈이다. 여기에서는 단계적 회귀분석(stepwise regression)과 위계적 회귀분석(hierarchical regression)에 대하여 살펴보기로 하겠다.

(1) 단계적 회귀분석

단계적 회귀분석은 엄격한 사회과학에서는 사용할 수 없는 방법이다. 이 방법은 먼저 가용한 모든 변수를 독립변수로 선택한 다음, 이 중에서 종속변수를 설명하는데 도움이 되는 변수를 통계적으로 찾아내는 것이다. 단계적 회귀분석에는 전방 단계적 분석과 후방 단계적 분석 그리고 최적 집합 단계적 분석의 세 가지 방

법이 있다.

다음과 같은 상황을 가정하자. 한 연구자가 Y를 예측하는데 어떤 변수가 영향을 주는지에 대하여 가설을 가지고 있지 않으며 잘 알지 못하는 상태이다. 현재 자신이 가지고 있는 자료에서는 Y와 함께 k개의 변수가 있다. 이를 $X_1 \cdots, X_k$라 하자. Y를 잘 설명할 수 있는 X가 어떤 X인지를 찾으려고 한다.

① 전방 단계적 분석(forward stepwise regression)

이 방법은 단계별로 진행된다. 첫 번째 단계에서는 Y에 대하여 하나의 독립변수를 가지는 모든 회귀분석을 실시한다. 즉 k번의 회귀분석을 한다. 그런 다음 가장 R^2이 가장 큰 회귀분석을 첫 번째 회귀분석결과로 삼는다. 예컨대 독립변수 X_3를 채택하는 모형이 가장 높은 R^2을 가진다면, 그 모형은 다음과 같다.

$$\hat{Y} = b_0 + b_1 X_3$$

두 번째 단계에서는 첫 번째 단계 채택한 독립변수를 모형에 그대로 유지하게 한 채 다른 독립변수를 하나씩 추가적으로 회귀모형에 투입한다. 즉 (X_3, X_1), (X_3, X_2), $\cdots$, (X_3, X_k)의 (k-1)개의 회귀분석을 한 다음 R^2을 구한다. 이 중 R^2가 가장 큰 모형을 두 번째 단계의 모형으로 선택한다.

그 다음 단계들에서도 동일한 방식으로 독립변수의 선택이 진행된다. 예컨대 (X_3, X_7)을 독립변수로 가지는 모형이 가장 R^2가 높다고 가정하면 그 다음 단계에서는 (X_3, X_7)를 포함하는 세 개의 독립변수를 가지는 모형을 적용시켜 R^2를 계산한다.

이런 방식으로 진행하면 k단계까지 갈 수 있다. 그러나 이 단계에 가기 전에 이 절차를 중단해야만 종속변수를 설명하는 독립변수 몇 개가 선택되는 것이다. 그렇지 않으면 모든 독립변수를 선택하는 단계까지 가게 된다. 중단하는 원칙은 기본적으로 새로운 변수가 투입되더라도 종속변수를 설명하는 양이 증가하지 않으면 그 변수의 투입을 중단한다는 것이다. 예컨대 7개의 독립변수를 사용한 R^2가 .40이었는데, 8개의 독립변수를 사용한 R^2도 .40이라면 8번째에 투입한 변수

는 R^2를 전혀 증가시키지 못하므로 모형에 포함시킬 이유가 없다. 각 단계에서 R^2의 변화량이 0인지 아닌지를 가설검증하여 영가설을 기각하지 못할 때(즉 증가량이 0과 다르지 않을 때) 단계적 회귀분석을 멈추게 멈추고 이전 통계모형을 최종 회귀모형으로 선택한다.

② 후방 단계적 분석(backward stepwise regression)

이 방법은 전방 단계적 분석과 반대 방향으로 진행된다. 즉 일단 k개의 모든 변수를 회귀식에 투입한다. 그 다음 단계에서 하나씩 변수를 제거하여 (k-1)개의 독립변수를 가진 k개의 회귀분석을 하여 R^2을 구한다. 이 중에서 R^2가 가장 높은 것을 선택한다. 바꾸어 말하면, 독립변수 하나를 제거하였을 때 원래의 R^2보다 변화가 가장 적은 모형을 선택한다. 논리는 변수를 제거하더라도 R^2에 영향을 미치지 않는 변수를 제거한다는 것이다. 이러한 방식으로 변수를 하나씩 제거하게 된다.

여기에서도 중단법칙이 필요하다. 위와 같이 한 단계의 모형에서 다른 단계의 모형으로 진행될 때 R^2의 변화량을 구한다. 이 변화량이 0이라면, 변수를 제거하더라도 설명량이 줄지 않으므로 변수를 제거한다. 그러나 변화량이 0보다 크면, 한 변수를 제거하면 종속변수를 설명하는 양이 적어지게 된다. 이 경우에는 변수를 제거하는 것은 좋지 않다. 정리하여 말하면, 각 단계에서 R^2변화량이 0이라는 영가설을 기각하면 변수를 제거할 수 없고, 기각하지 않으면 변수를 제거할 수 있다. 따라서 기각되는 시점에서 단계는 멈추게 된다.

③ 최적 집합 단계적 회귀분석(best set stepwise regression)

전방 단계적 회귀분석에서는 독립변수가 일단 포함되게 되면 다음 단계들에서 제거되지 않는다. 또한 후방 단계적 회귀분석에서는 독립변수가 일단 제거되면 다음 단계에서 다시 포함될 수 없다. 이러한 방식 때문에 전방 방법과 후방 방법이 동일한 결과를 산출하지 않는다. 또한 이들 방법들에서 정해진 최종 모형이 독립변수가 다섯 개인 모형이라고 가정하면 이 다섯 개의 독립변수는 이들로 구성된 모든 회귀모형 중에서 종속변수를 가장 잘 설명하는 독립변수 집합이 아닐 수 있

다. 다음의 예를 들어 보자.

위에서 제시된 전방 단계적 회귀분석에서 처음에 X_3을 선택하고 두 번째 X_8을 선택하였다고 가정하자. 즉 두 번째 단계에서 (X_3, X_8)의 독립변수를 가가지는 회귀식을 선택하게 된다. 그러나 (X_4, X_8)이 (X_3, X_8)보다 더 높은 설명량 R^2을 가질 수 있다. 모든 독립변수들의 상관이 0이 경우에는 이러한 일이 발생하지 않으나 독립변수들간의 상관이 높을 때는 충분히 발생할 수 있다. 전방과 후방 단계적 회귀분석방법인 경우에는 이 (X_4, X_8)를 찾아내지 못한다. 한번 투입되거나 제거된 변수는 그 이후의 모형에서 계속하여 투입되거나 제거되기 때문이다.

최적 집합 단계적 회귀분석에서는 j개의 독립변수를 사용하는 경우에는 모든 가능한 조합들에 대한 회귀분석을 시행하고 그 중 R^2이 가장 높은 독립변수의 집합을 선택한다. 이 경우에 ${}_kC_j$개만큼의 회귀분석을 하여야 되기 때문에 전방이나 후방 단계적 회귀분석보다는 계산량이 많아진다. 사실 전방과 후방 단계적 회귀분석은 계산 속도가 낮은 컴퓨터를 사용하던 시절에 개발된 것이다. 그러나 계산속도가 빨라지고 계산에 대한 비용이 매우 낮은 현대에서는 계산량은 크게 문제가 되지 않는다.

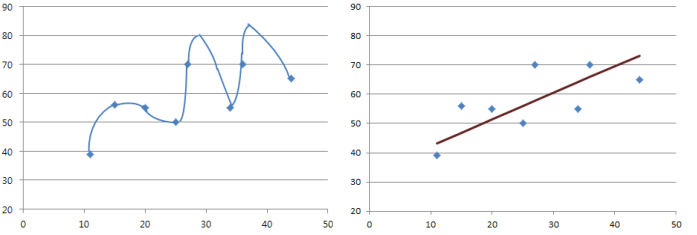
④ 단계적 회귀분석의 문제점

단계적 회귀분석은 한마디로 말하면 사용해서는 안 되는 방법이다. 피치 못하게 이 방법을 사용하는 경우에는 결과 해석에 매우 주의해야 하며, 추가적인 분석으로 이 결과를 보충하기 전까지는 그것을 신뢰할 수 없다. 그 이유는 다음과 같다.

단계적 회귀분석과 같은 분석 방법을 자료주도적 분석(data-driven analysis)라 한다. 즉 자료를 사용하여 통계적 모형을 구하거나 수정해 나가는 방법이다. 이러한 방법은 분석의 결과를 일반화할 수 없게 한다. 일반화란 한 표본을 사용하여 얻어낸 결론은 다른 표본에도 적용시킬 수 있는 것을 의미한다. 자료주도적 분석은 현재 표본 자료에 있는 패턴을 잘 추려낼 수 있는 방식으로 통계적 모형을 만들어낸다. 이렇게 하면 소위 과잉 적합(over fitting)현상이 발생하게 된다. 표본집단은 모집단에서 존재하는 진眞패턴보다 더 복잡한 모형을 형성하기 쉽다. 즉, 모집단에 존재하는 패턴 뿐 만 아니라 현재 다루는 표본에만 존재하는 독특한 패턴까지도 잡아내

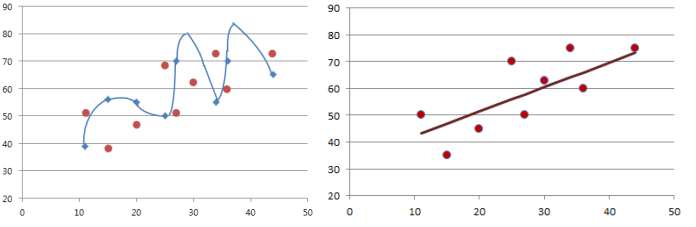
게 된다. 만약 이러한 일이 일어난다면 새로운 표본에 이 통계적 모형을 적용시키면 예측력이 나빠지게 된다.

아래에는 8개의 자료를 설명하기 위한 두 가지 통계적 모형들이 제시되어 있다. 왼쪽의 모형은 8개의 자료를 완벽하게 설명할 수 있는 모형인 반면에, 오른쪽 모형은 자료를 완벽하게 설명할 수는 없지만 상당히 많이 설명할 수 있는 모형이다.



종속변수를 많이 설명하는 모형을 선택한다면, 둘 중에 왼쪽 모형을 선택할 것이다. 자료주도적 분석을 하면 왼쪽과 같은 회귀식을 찾아내게 된다. 그러나 왼쪽 모형은 원래 자료를 생성하는 진모형이기 보다는 주어진 자료에만 있는 독특한 특성을 반영하는 모형일 수도 있다.

새로운 자료를 수집하여 위의 두 모형을 적용시키면 다음 그래프들과 같은 결과가 나타난다. 오른쪽의 선형모형은 새로운 자료를 수집하더라도 여전히 자료를 잘 설명하고 있다. 그러나 왼쪽의 모형은 새로운 자료를 전혀 설명하지 못하는 모형이 된다.



사회과학에서 추구하는 것은 표본에 독특한 자료의 특성을 기술하는 것이 아니라 연구에서 관찰되지 않았던 자료까지 그것을 확장시켜서 해석하는 것이다. 따라서 자료주도적 분석은 현재의 자료를 설명하는 데는 도움을 줄 수 있지만, 사회과학적 자료를 축척하는 데는 도움을 주지 못한다.

다시 한 번 강조하지만, 단계적 회귀분석과 같은 자료주도적 분석은 가능한 한 피해야 된다. 그러나 연구 초기에 구체화된 연구가설이 없고 아이디어가 부족하여 여러 가지 가능성을 탐색해야 하는 경우도 있다. 이때에는 단계적 회귀분석을 하는 것을 피할 수 없을 것이다. 일단 탐색적인 단계적 회귀분석을 수행하고 난 다음, 반드시 새로운 자료를 수집하여 단계적 회귀분석에서 결정한 모형이 새로운 자료에도 적용되는지를 확인해야 한다. 이를 교차타당화(cross-validation)이라 한다.

(2) 위계적 회귀분석

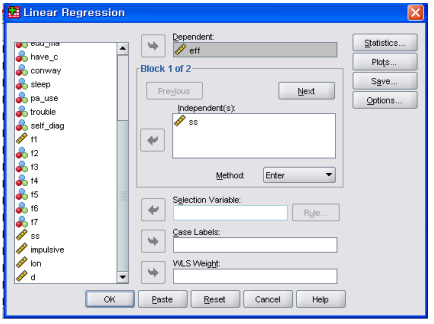
위에서 단계적 회귀분석은 가능한 한 피해야 하는 분석방법이라 강조하였다. 여기에서는 대안으로 위계적 회귀분석방법을 추천한다. 이 방법은 단계적 회귀분석과 유사하지만, 변수를 선택하는 주체가 연구자인 점에서 다르다.

예컨대 진보 성향이 영향을 미치는 요인들에 대하여 연구한다고 하자. 종속변수는 진보성향 점수가 될 것이고 독립변수로 여러 가지를 생각해 볼 수 있다. 먼저 개인의 인구 통계학적 변수들이 진보 성향에 영향을 미칠 수 있을 것이다. 연령, 성별, 출신지역을 먼저 독립변수로 고려할 수 있다. 그 다음으로 수입을 독립변수로 포함시키고, 마지막으로 학력을 독립변수로 포함시키고자 한다.

제 1 단계 : (종속변수) 진보성향점수
(독립변수) 연령, 성별, 출신지역
제 2 단계 : (종속변수) 진보성향점수
(독립변수) 연령, 성별, 출신지역, 수입
제 3 단계 : (종속변수) 진보성향점수
(독립변수) 연령, 성별, 출신지역, 수입, 학력

각 단계별로 회귀분석을 실시하고 R^2를 계산한다. 단계에서 단계로 진행될 때 R^2의 변화량이 통계적으로 의미가 있는지를 검증한다. 예컨대, 1단계에서 2단계 분석으로 넘어가면 R^2는 증가하게 되어 있다. 이 증가량이 0과 다른지에 대하여 가설검증을 한다. 만약 R^2의 변화가 통계적으로 유의미하면, 수입은 연령, 성별, 출신지역이 설명하지 못하는 부분을 설명할 수 있다는 의미이다. 2단계와 3 단계의 R^2 차이가 통계적으로 유의하면, 연령, 성별, 출신지역, 수입이 설명하지 못하는 부분을 학력이 설명할 수 있다는 것이다.

SPSS에서 이를 간단히 처리하는 방법은 SPSS 회귀분석 창에서 Dependent와 Independent를 입력하면 두 번째 열의 두 번째 행의 창에 Next라는 단추가 있다. Next라는 단추를 누리면 새로운 독립변수를 추가할 수 있다. 또한 Statistics의 단추에서 R-square Change 단추를 누르면 첫 번째 모형과 두 번째 모형의 R^2변화량이 통계적으로 유의미한지를 검증하여 준다.



위계적 회귀분석은 또한 혼입변수를 통계적으로 통제할 때도 많이 사용된다. 예컨대 남녀 간이 임금의 차이가 있는지를 회귀분석을 통하여 검증하려고 한다. 독립변수를 성별, 종속변수를 임금으로 하여 회귀분석을 하게 된다. 이때 한 가지 고려되어야 할 점은 남녀 간에 차이 있는 여러 가지 변수들이 존재할 것이고 이 변수들에 의하여 임금의 차이가 나는데 남녀 간의 임금차이로 보일 수 있는 가능성이 있다는 것이다. 예컨대 남자들이 여자들보다 근무연한이 높고 초과근무시간

이 많다면, 임금차는 성별에 따른 차이가 아니고 근무연한과 초과근무시간의 차이 때문일 수 있다. 이때에는 근무연한과 초과근무시간을 통제하였을 때도 남녀 간의 임금 차이가 있는지를 검증해야 한다.

위계적 회귀분석에서 다음과 같이 두 변수를 통제할 수 있다. 1단계에서 근무연한과 초과근무시간을 독립변수로 넣고, 2단계에서 근무연한과 초과근무시간 그리고 성별을 독립변수로 넣어서 회귀분석을 한 다음, 1단계와 2단계의 R^2차이를 구한다. R^2차이는 근무연한이나 초과근무시간으로는 설명할 수 없고 오직 성별의 차이로만 설명할 수 있는 분산이 되고 이 분산이 통계적으로 유의하면, 근무연한과 초과근무시간을 통제하였을 때도 성별의 차이가 난다고 해석할 수 있다.

7) 범주변수의 처리

지금까지의 회귀분석모형에서는 종속변수와 독립변수가 모두 연속변수인 경우만을 고려하였다. 종속변수는 반드시 연속변수여야 하지만 독립변수는 범주변수여도 회귀분석이 가능하다. 범주변수를 더미 코딩dummy-coding이나 효과 코딩(effect coding)으로 변경하여서 처리하면 된다. 예컨대 독립변수가 임상집단이고 우울성집단, 편집성집단, 충동성집단, 그리고 정상집단의 네 개의 수준을 가지고 있다고 하자. 집단 수보다 하나 작은 세 개의 더미 변수 D1, D2, D3를 만든다. 우울성집단은 D1만 1이고 나머지는 0으로, 편집성집단은 D2만 1이고 나머지는 0으로, 충동성집단은 D3만 1이고 나머지는 0으로, 정상집단은 세 변수 모두 0으로 입력하다. 그런 다음 D1, D2, D3를 독립변수로 하여 회귀분석을 한다.

	D1	D2	D3
우울성집단	1	0	0
편집성집단	0	1	0
충동성집단	0	0	1
정상집단	0	0	0

즉 회귀모형은 다음과 같이 된다.

$$\hat{Y} = b_0 + b_1 D_1 + b_2 D_2 + b_3 D_3$$

이렇게 하여 회귀분석을 하면, 하나의 절편과 세 개의 기울기가 구해질 것이다. 각 집단에 해당하는 $\hat{Y}$는 다음과 같다.

$$\begin{aligned} \text{우울성 집단} &: \hat{Y} = b_0 + b_1 \\ \text{편집성 집단} &: \hat{Y} = b_0 + b_2 \\ \text{충동성 집단} &: \hat{Y} = b_0 + b_3 \\ \text{정상 집단} &: \hat{Y} = b_0 \end{aligned}$$

따라서 절편, b_0는 정상집단의 평균을 나타내며, b_1은 정상집단과 우울성 집단의 평균차, b_2은 정상집단과 편집성 집단의 평균차, b_3은 정상집단과 충동성 집단의 평균차를 나타낸다. 각 기울기가 통계적으로 의미가 있으면, 각 기울기가 의미하는 바대로 해석하면 된다. 또한 세 개의 더미 변수에 의하여 설명되는 분산(R^2)은 집단 간 차이에 의해 설명되는 분산으로 해석하면 된다.

효과코딩은 위에서 더미코딩과 동일한데 마지막 집단을 0으로 할당하는 것이 아니라 -1로 할당하는 것만 다르다. 그러면 각 집단에 해당하는 $\hat{Y}$는 다음과 같다.

$$\begin{aligned} \text{우울성 집단} &: \hat{Y} = b_0 + b_1 \\ \text{편집성 집단} &: \hat{Y} = b_0 + b_2 \\ \text{충동성 집단} &: \hat{Y} = b_0 + b_3 \\ \text{정상 집단} &: \hat{Y} = b_0 - b_1 - b_2 - b_3 \end{aligned}$$

위의 네 집단의 $\hat{Y}$을 다시 평균하면 b_0가 된다. 따라서 절편은 전체 평균이고 b_1, b_2, b_3는 각각 각 집단과 전체평균의 차이가 된다. R^2는 더미코딩을 한 경우와 동일하다.

8) 조절 효과(moderation effect)

조절 효과란 독립변수 A가 종속변수에 미치는 영향의 정도가 다른 독립변수 B의 크기에 따라 달라지는 것을 말한다. 예컨대 부모의 사회경제적 지위(SES)가 학생의 성적에 미치는 영향력을 분석할 때, 학생들의 IQ가 낮아지면 SES에 영향을 많이 받고 높아지면 SES에 영향을 적게 받을 수 있다. 즉 학생의 지능수준에 따라 달라지면 성적에 대한 SES의 기울기도 달라진다. 이때 SES가 성적에 미치는 효과를 지능이 조절한다고 이야기 한다.

조절효과를 회귀분석에 포함시키는 방법은 간단하다. 두 개의 독립변수 X_1와 X_2가 있으면, 이 둘을 곱한 값을 회귀식에 포함시키면 된다. 통계패키지를 이용할 때는 실제로 두 변수를 곱한 값을 새로운 변수에 할당시키면 된다. 이때 회귀식은 다음과 같이 된다.

$$\begin{aligned}\hat{Y} &= b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2 \\ &= b_0 + (b_1 + b_3X_2)X_1 + b_2X_2 \\ &= b_0 + b_1X_1 + (b_2 + b_3X_1)X_2\end{aligned}$$

식에서 보는 바와 같이 X_1의 기울기는 X_2의 크기에 따라서 변하고, X_2의 기울기는 X_1의 크기에 따라서 변한다. 다만 b_3가 0인 경우에는 이러한 조절 효과가 없어진다. 따라서 조절효과는 b_3가 0인지 아닌지를 가설검증하면 된다. b_3가 통계적으로 유의미하면 조절효과가 있는 것이고 그렇지 않으면 없는 것이다.

5. 회귀분석모형에 대한 정리

회귀분석모형은 다수의 독립변수를 이용하여 하나의 종속변수를 예언하는 통계적 기법이다. 이 분석을 통하여 종속변수를 예언하는 선형식인 회귀식을 구할

수 있으며, 회귀식이 종속변수을 얼마나 잘 설명하는지에 대한 통계치인 결정계수를 구할 수 있다. 다중 회귀분석의 경우에는 다수의 독립변수를 사용함으로써 얻는 이점과 문제점들이 있다. 공선성, 변수의 선택, 조절 효과 등이 이에 해당한다.

회귀분석은 모든 선형모형의 기본이 되는 모형이다. 이 모형을 기본으로 하여 위계적 선형모형, 일반화 선형모형, 로지스틱 모형 등이 발전된다. 회귀분석에서 사용된 개념들과 기법들이 이러한 모형에서도 그대로 적용될 수 있다. 회귀분석모형의 이해는 고급 통계로 가기 전에 반드시 거쳐야 하는 단계이다.

회귀분석의 가정들, 회귀분석에서의 여러 가지 가설 검증 등에 대하여서는 여기서 다루지 않았다. 보다 자세한 이해를 원하면 회귀분석 교과서를 참조할 수 있다.

찾아보기