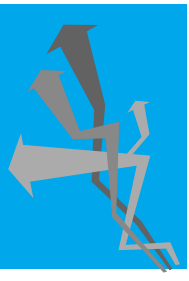
| 제14장 |

회귀분석



제13장에서 두 변수 사이의 연관성을 상관분석을 통해서 생각해 보았다. −1과 1 사이의 값을 갖는 상관계수라는 측도를 이용하여 변수들의 연관성의 정도를 측정하였다. 이때 상관관계가 밀접하다고 해서 두 변수 간에 인과관계(causation)가 있다고는 할 수 없다. 제13장에서의 자료분석 결과 '신문을 보는 시간과 TV를 시청하는 시간 사이에는 부적인 상관관계가 존재해서 신문을 보는 시간이 많을수록 TV를 시청하는 시간이 적어진다. 또는 TV를 시청하는 시간이 많을수록 신문을 보는 시간이 적어진다'라고 하였다. 그러나 '신문을 보는 시간이 5.5시간일 때 TV를 시청하는 시간은 몇 시간일까?'라는 물음에 대한 답은 상관분석을 통해서는 알 수가 없다. 이때 '신문을 보는 시간'은 'TV를 시청하는 시간'에 영향을 미치는 원인변수에 해당하고 'TV를 시청하는 시간'은 '신문을 보는 시간'에 따라 값이 달라지는 결과변수라고 할 수 있다.

이제 변수 간의 인과관계를 알아보기 위한 분석방법의 하나인 회귀분석을 살펴보도록 하자. 회귀분석은 1개 또는 그 이상의 독립변수와 종속변수 사이의 관계를 수학적인 함수식을 이용하여 규명하고자 하는 분석법으로, 독립변수의 변화에 따른 종속변수의 변화를 예측하는 데 사용된다. 예를 들면, 사람들의 키(독립변수)와 몸무게(종속변수)의 관계를 하나의 함수식으로 표현하고 이 함수식을 이용하여 특정 키에 대한 몸무게를 예측하고자 할 때 회귀분석을 적용할 수 있다.

1. 회귀분석의 개념

회귀분석(regression analysis)은 변수들 사이의 인과관계를 규명하고자 하는 분석방법이기 때문에 변수의 역할설정이 중요하다. 앞에서 '신문을 보는 시간이 TV를 시청하는 시간에 어떤 영향을 미치는가?'라는 물음에서 '신문을 보는 시간'은 원인변수에 해당하고 'TV를 시청하는 시간'은 결과변수였다. 이를 뒤바꾸어 'TV를 시청하는 시간'을 원인변수로 하고 '신문을 보는 시간'을 결과변수로 한다면 원래 의도했던 연구목적에 위배되는 것이다.

회귀분석에서 다른 변수에 영향을 주는 원인에 해당하는 변수를 **독립변수**(independent variable) 또는 **설명변수**(explanatory variable)라고 하며, 영향을 받는 결과에 해당하는 변수를 **종속변수**(dependent variable) 또는 **반응변수**(response variable)라고 한다. 물론 이러한 독립변수와 종속변수의 개념은 앞에서 살펴보았던 t-검증이나 분산분석 등에서의 독립변수, 종속변수의 개념과 동일하다.

회귀분석에서는 일반적으로 종속변수는 하나이고 이에 영향을 미치는 독립변수는 여러 개인 분석을 많이 하게 된다. 독립변수와 종속변수가 각각 하나일 때의 분석을 **단순회귀분석**(simple regression analysis)이라고 하고 종속변수는 1개이면서 독립변수가 2개 이상일 때의 분석을 **중회귀분석**(multiple regression analysis, 다중회귀분석, 중다회귀분석)이라고 한다.

회귀분석은 독립변수와 종속변수 사이의 구체적인 함수식을 찾아내고, 독립변수로부터 종속변수를 예측하는 데 그 목적이 있다. 이때 함수식은 제일 단순한 1차 직선식이 될 수도 있고 2차식, 3차식, 로그식, 지수식 등 다양한 형태의 곡선식이 될 수도 있다.

먼저 독립변수가 1개이고, 독립변수와 종속변수의 관계가 1차 직선인 경우의 **단순선형회귀분석**(simple linear regression analysis)을 통해 일반적인 회귀분석의 개념과 절차를 알아보도록 한다.

(1) 자료의 구조 및 산점도

회귀분석에서의 변수들은 기본적으로 모든 변수가 양적 자료이어야 한다. 질적 자료를 회귀분석에 포함시키고자 할 때에는 **가변수**(dummy variable)를 도입하여 분석할 수 있다.

회귀분석을 실시하기에 앞서 상관분석에서와 마찬가지로 변수 간에 산점도를 그려서

변수 간의 관계를 대략적으로 파악해야 한다. 즉, 산점도에서 변수 간의 관계가 직선관계라고 판단되면 선형회귀분석을 할 수 있지만 곡선관계로 판단되면 선형회귀분석을 해서는 안 될 것이다.

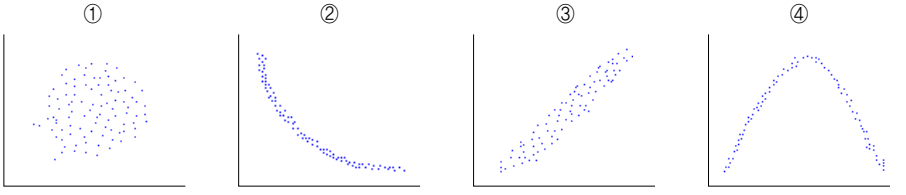


그림 14.1 여러 형태의 산점도

〈그림 14.1〉의 ①은 두 변수 간에 예측되는 관계가 없으므로 회귀분석을 할 필요가 없음을 나타내며, ②의 산점도로부터는 로그관계 또는 역수의 관계를 고려해 볼 수 있다. ③의 산점도는 두 변수의 관계가 직선(선형)관계임을 시사하며 ④의 산점도는 2차 곡선관계임을 나타낸다.

다음 〈그림 14.2〉는 중·고등학교에서 교장의 인간지향적 분위기의 정도와 교사들의 직무성과와의 산점도이다. 인간지향적 분위기의 정도가 높을수록 교사들의 직무성과도 높아진다고 볼 수 있다. 이 자료의 상관계수는 0.330이다.

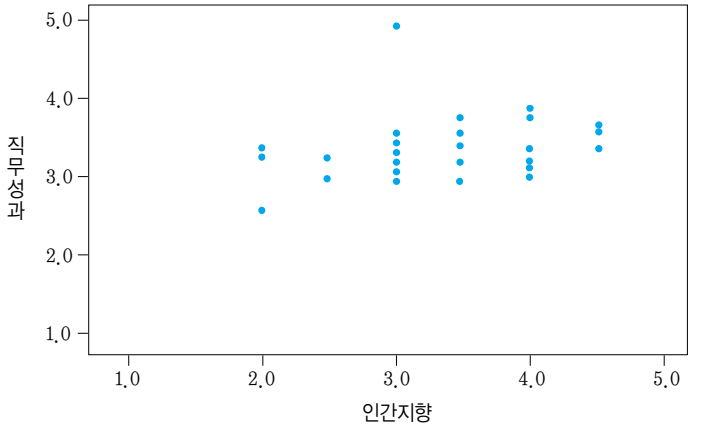


그림 14.2 인간지향성의 정도와 직무성과의 산점도

〈그림 14.2〉로부터 인간지향적 분위기의 정도와 교사들의 직무성과와의 사이에는 직선(선형)의 관계가 있을 것으로 짐작이 되므로 인간지향적 분위기의 정도를 독립변수로, 교사들의 직무성과를 종속변수로 하여 단순선형회귀분석을 해 보도록 하자. 즉 〈그림 14.2〉에 두 변수 간의 관계를 설명할 수 있는 직선을 아래 〈그림 14.3〉처럼 그리고 그 직선식을 구하고자 한다.

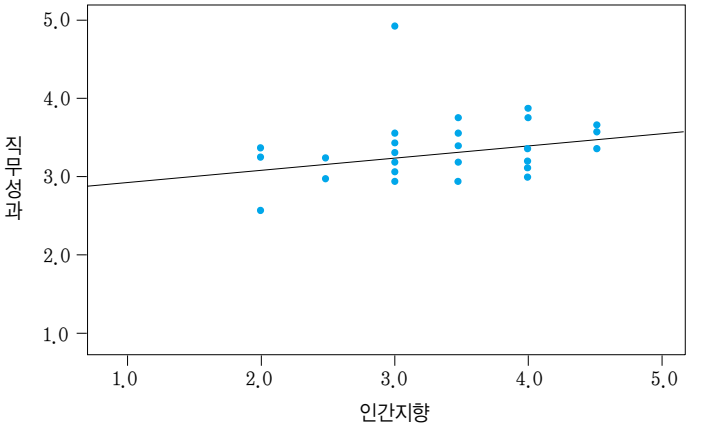


그림 14.3 산점도와 회귀식

(2) 단순선형회귀모형

독립변수 X와 종속변수 Y의 관계가 1차 직선이라고 예상되는 경우에 설정하게 되는 단**순선형회귀모형**(simple linear regression model)은 다음과 같다.

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad \epsilon_i \sim iid\, N(0, \sigma^2)$$
$$i = 1,\ 2,\ \cdots,\ n$$

여기에서 $x_1,\ x_2,\ \cdots,\ x_n$는 독립변수 X의 관측값이고, $y_1,\ y_2,\ \cdots,\ y_n$는 종속변수 Y의 관측값이며 ϵ_i는 오차항이다. α, β는 **모회귀계수**라고 하며 회귀분석을 통해 추정될 값이다.

이 모형의 의미는 다음과 같다.

① $y_i = \alpha + \beta x_i + \epsilon_i$: 두 변수 간의 관계는 선형이다.

② $\epsilon_i \sim iid\, N(0, \sigma^2)$: 오차항의 분포는 평균이 0이고 분산이 σ^2 인 정규분포이다. 이때 분산은 σ_i^2가 아니고 σ^2으로 모든 오차에 대해 동일하다. iid는 independent & identical distribution의 약자로 오차들이 서로 독립이고 동일한 분포를 갖는다는 뜻이다.

위의 ①과 ②로부터 단순선형회귀모형에 내포된 가정은 **선형성**, **정규성**, **등분산성**, **독립성**이라고 한다. 위의 단순선형회귀모형을 그림으로 표시하면 다음 〈그림 14.4〉와 같다. 즉, 관측값 y_i는 주어진 x_i에서 관측될 수 있는 (정규분포에 따르는) 여러 값 중 하나가 표본으로 추출된 것으로 본다.

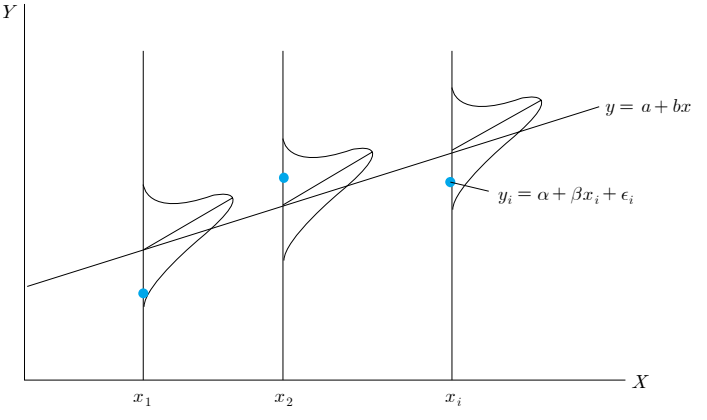


그림 14.4 단순선형회귀모형

(3) 회귀식의 추정

회귀분석의 1차적인 목적은 표본으로부터 모회귀계수 α, β를 추정하여 추정된 회귀식을 만드는 것이다.

$$\underbrace{y_i = \alpha + \beta x_i + \epsilon_i}_{\text{회귀모형}} \quad \rightarrow \quad \underbrace{\hat{y_i} = a + b x_i}_{\text{추정된(적합된) 회귀식}}$$

추정된 회귀식에서 a, b는 모회귀계수 α, β의 추정값으로 **표본회귀계수**라고 한다.

α, β를 표본으로부터 추정하는 방법을 생각해 보자.

추정된 회귀식의 의미는 〈그림 14.3〉에서 보듯이 두 변수 간의 관계를 원래 산점도의 모든 점 대신 직선식으로 대체하여 설명하겠다는 뜻이다. 따라서 원래의 관측값(y_i)과 직선상의 값($\hat{y_i}$)과는 차이가 있을 수밖에 없는데 이 차이는 되도록이면 작아야 한다. 다음과 같은 통계량을 생각해 보자.

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

S는 오차제곱합으로 이 S를 최소화하는 α, β를 추정하는 방법을 **최소제곱법**(least squares method)이라 하며 회귀계수의 추정법으로 가장 많이 이용되고 있다. 오차제곱합 S를 최소화하는 구체적인 방법은 생략하고 그 결과만 기술하기로 한다.

$$b \equiv \hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$
$$a \equiv \hat{\alpha} = \bar{y} - b\,\bar{x}$$

최소제곱법에 의해 추정된 추정량 a, b를 **최소제곱추정량**(least squares estimator, LSE)이라고 한다. 추정된 회귀식은 다음과 같이 표현되며 두 변수의 평균 $(\bar{x}, \bar{y})$점을 지나는 것을 알 수 있다.

$$\hat{y} = a + bx = \bar{y} + b(x - \bar{x})$$

(4) 회귀계수의 의미

회귀계수 a, b는 추정된 회귀식에서의 절편과 기울기이다. 절편은 독립변수의 값이 0일 때의 종속변수의 값으로 독립변수의 값이 0이면 종속변수도 반드시 0이어야 하는 자료분

석에서는 모형설정 시에 절편이 존재하지 않는 '원점을 지나는 회귀모형'을 고려할 수도 있다. 예를 들면, 키와 몸무게의 관계에서 키가 0이면 몸무게도 0인 경우이다. 그러나 소득과 지출의 관계에서는 소득이 없어도 지출이 있을 수 있으므로 절편이 포함되어야 한다. 실제적인 자료분석에서는 특별한 언급이 없으면 항상 절편을 포함한다. 즉, 키와 몸무게의 관계에서도 절편을 포함하는 것이 전체적인 회귀식 추정에 더 적절하다. 다음 〈그림 14.5〉를 보자. 실선은 절편을 포함한 회귀선이고 점선은 절편이 없는 회귀선으로 원점을 지나야 한다는 조건 때문에 관측값들로부터 멀리 떨어진 회귀선이 추정되었다.

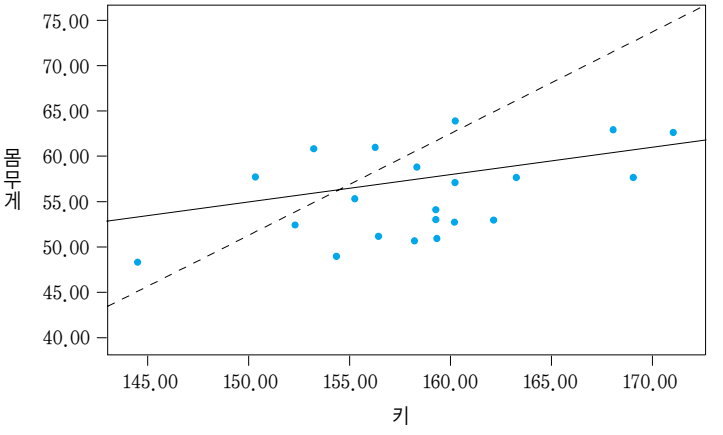


그림 14.5 키와 몸무게의 회귀선

기울기는 독립변수의 값이 1단위 증가할 때 늘어나는 종속변수의 값을 의미한다. b가 0에 가깝다면 독립변수가 아무리 값이 변해도 종속변수에 거의 영향을 미치지 못하고 회귀분석은 의미가 없을 것이다. 따라서 직접적으로 독립변수와 종속변수의 관계를 표현하는 계수인 b가 a보다 상대적으로 중요하다.

(5) 회귀식의 정도

앞에서 산점도를 그려 보고 회귀모형을 설정한 후 이에 따라 회귀식을 추정하였다. 이제 추정된 회귀식이 정말 원래의 관측값들을 잘 대표하는지 점검해 볼 필요가 있다. 모형을 잘못 설정했을 수도 있고 회귀분석의 기본 가정이 충족되지 않았을 수도 있다. '추정된 회귀식이 원래의 관측값들을 어느 만큼 대표하는지'를 **회귀식의 정도**(precision)라고 한

다. 추정된 회귀식의 정도를 측정하는 측도로 **결정계수**(coefficient of determination)가 가장 많이 이용된다. 결정계수의 개념과 구하는 방법을 알아보자.

관측값 y_i의 편차 $y_i - \bar{y}$를 다음과 같이 분해해 보자. 즉 독립변수 X의 관측값 x_i와는 전혀 관계없는 양$(y_i - \bar{y})$을 x_i와 관련된 양$(\hat{y}_i)$을 포함한 식으로 분해한 것이다.

$$\underbrace{y_i - \bar{y}}_{①} = \underbrace{(y_i - \hat{y}_i)}_{②} + \underbrace{(\hat{y}_i - \bar{y})}_{③}$$

① 편차
② 회귀식에 의해 설명되지 않는 편차
③ 회귀식에 의해 설명되는 편차

②의 $y_i - \hat{y}_i$를 **잔차**(residual)라고 하며 잔차가 작고 ③이 크면 회귀식의 정도가 좋다고 할 수 있다. 다음 〈그림 14.6〉은 이를 그림으로 표현한 것이다.

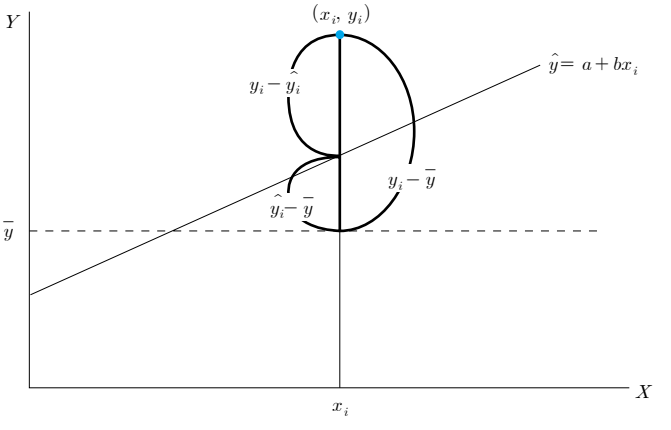


그림 14.6 편차의 분해

위의 식을 모든 편차들에 대해 제곱한 후 합하면 다음과 같다.

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$$SST = SSE + SSR$$

(총제곱합) (잔차제곱합) (회귀제곱합)

SSE가 작을수록, 또 SSR이 클수록 회귀식의 정도가 좋다. 즉, 회귀모형이 자료들을 잘 설명해 주고 있다. 따라서 SSE, $\frac{SSR}{SSE}$ 등을 정도의 측도로 사용할 수 있다. 총제곱합 SST 중에서 회귀제곱합 SSR이 차지하는 비율을 결정계수라고 하며, 다음과 같이 표현된다.

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

결정계수의 범위는 0과 1 사이이며 단순선형회귀의 경우 결정계수는 상관계수의 제곱과 같다. 결정계수의 값이 1에 가까울수록 추정된 회귀식 주위에 자료가 밀집되어 있으므로 자료를 잘 대표하고 있다고 할 수 있으며 독립변수가 종속변수를 잘 설명한다고 할 수 있다. 결정계수는 총제곱합 중 추정된 회귀식에 의해 설명되는 제곱합이므로 결정계수를 **회귀식의 기여율**이라고도 부른다. 그러나 추정된 회귀식이 자료를 충분히 잘 설명하는지를 정확히 파악하기 위해서는 결정계수뿐만 아니라 산점도와 잔차의 검토가 종합적으로 이루어져야 한다.

독립변수의 수가 증가하면 결정계수는 항상 증가한다. 따라서 결정계수의 기준으로 보면 독립변수의 수는 무조건 많으면 좋다는 결론이 나오는데 독립변수의 수가 많으면 여러 가지 문제가 발생한다. 따라서 독립변수의 수에 따라 무조건 증가하지 않는 수정결정계수를 정의하여 회귀식의 정도를 측정하는 측도로 사용하기도 한다.

(6) 잔차의 검토

잔차 $e_i = y_i - \hat{y}_i$ 는 위에서 살펴본 것처럼 회귀식의 정도를 알아볼 수 있는 기본개념이기도 하며 회귀모형에서의 가정들(선형성, 독립성, 정규성, 등분산성)에 대한 정보를 제공해주는 통계량이다. 회귀분석을 실시한 후에 잔차의 산점도를 그려 봄으로써 이러한 가

정을 확인해 볼 수 있다. 다음 〈그림 14.7〉은 잔차들을 독립변수 또는 추정된 종속변수 등에 대하여 그렸을 때 나올 수 있는 산점도이다.

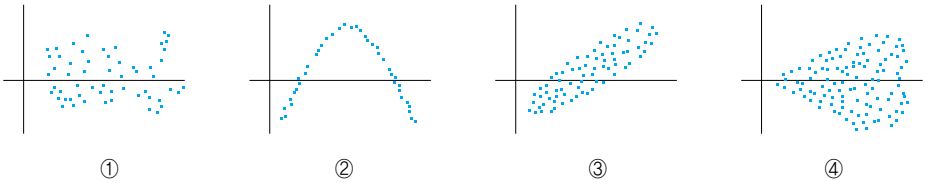


그림 14.7 잔차의 산점도

①은 잔차들이 랜덤하게 분포되어 있으므로 가정들이 만족된다고 볼 수 있고 ②는 선형회귀모형보다는 2차곡선회귀모형이 더 타당함을 나타낸다. 회귀분석을 하기 전에 독립변수와 종속변수의 산점도를 그려보았다면 알 수 있었던 결과이다. ③은 어떤 선형항이 추가되는 것이 좋음을 뜻하고 ④는 독립변수의 값이 커짐에 따라 잔차의 분산도 커지므로 등분산성이 만족되지 않음을 뜻한다. 소득이 적은 가구들은 지출액의 편차가 적은데 소득이 많은 가구 중에는 지출은 적고 저축을 많이 하는 가구가 있는 반면, 모두 다 지출해 버리는 가구도 있어 편차가 커질 수 있다. 이 경우에는 가중회귀분석을 실시하는 것이 바람직하다.

(7) 회귀진단

회귀진단(regression diagnostics)은 넓은 의미로는 추정된 회귀식의 전반적인 검토를 의미하며 위에서 살펴본 대로 추정된 회귀식의 정도, 잔차분석을 통한 회귀모형에 대한 가정의 검토 등을 포함한다. 그러나 좁은 의미로서의 회귀진단은 '영향력 있는 관측값의 탐색'을 의미하며 주로 회귀식의 추정에 큰 영향을 미치는 극단값을 찾아내는 것을 의미한다.

(8) 상관계수와 회귀계수의 관계

표본상관계수 r과 추정된 회귀계수 b와의 관계를 알아보자. 추정된 회귀계수 b를 r의 함수로 표현해 보면 다음과 같다. 이 식에서 s_x는 변수 X의 표준편차이며 s_y는 변수 Y의 표준편차이다.

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=r\frac{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}=r\frac{s_y}{s_x}$$

상관계수 r과 회귀계수 b의 부호가 같다. $b>0$이면 양의 상관관계, $b<0$이면 음의 상관관계, $b=0$이면 상관관계가 없음을 알 수 있다.

상관계수 r을 알고 있을 경우, 추정된 회귀식은 다음과 같이 표현된다.

$$\hat{y}-\bar{y}=r\frac{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}(x-\bar{x})=r\frac{s_y}{s_x}(x-\bar{x})$$

(9) 표준화된 회귀계수

독립변수가 여러 개인 중회귀분석에서 종속변수에 미치는 독립변수의 상대적인 영향력의 크기를 어떻게 알아볼 수 있는 있는지 생각해 보자. 회귀계수가 크면 그 변수의 영향력이 큰 것일까? 추정된 회귀식에서 회귀계수는 자료의 단위와 밀접한 관계가 있다. 같은 자료라 해도 독립변수의 단위가 달라지면(키를 cm나 inch로 측정) 회귀계수는 달라진다. 측정단위가 크다면(키를 cm가 아닌 m로 측정) 회귀계수의 값은 상대적으로 아주 작아지게 된다. 그러므로 회귀계수의 크기로 독립변수들의 상대적인 영향력의 크기를 말할 수는 없다.

표준화된 회귀계수는 원자료를 표준화(각각 평균을 빼고 표준편차로 나누는)한 후 회귀분석을 했을 때 나오는 회귀계수이다. 표준화된 회귀계수는 단위에 무관하기 때문에 독립변수들의 상대적인 영향력을 비교할 수 있으며 표준화된 회귀계수가 큰 독립변수의 종속변수에 대한 영향력이 크다고 말할 수 있다.

추정된 회귀식을 구성할 때에는 (비표준화된)회귀계수를 사용하고 독립변수들 사이의