회귀분석 정의와 사례

https://m.blog.naver.com/PostView.nhn?blogId=starb0ard&logNo=80139391237&proxyReferer=https%3A%2F%2Fwww.google.com%2F

정의

회귀분석(regression analysis)이란 둘 또는 그 이상의 변수들간의 관계를 파악함으로써 어떤 특정한 변수(종속변수)의 값을 다른 한 개 또는 그 이상의 변수(독립변수)들로부터 설명하고 예측하는 통계적 기법이다.

회귀분석의 **어원**은 다음과 같다.1

회귀(Regress)의 원래 의미는 옛날 상태로 돌아가는 것을 의미한다. 영국의 유전학자 프란시스 갈톤(Francis Galton)은 부모의 키와 아이들의 키사이의 연관관계를 연구하면서 부모와 자녀의 키사이에는 선형적인 관계가 있고 키가 커지거나 작아지는 것보다는 전체 키 평균으로 돌아가려는(**회귀하려는**) 경향이 있다는 가설을 세웠으며 이를 분석하는 방법을 "회귀분석"이라고 하였다.

다른 자료를 통해 설명을 보충해 보았다.

회귀분석이란 상관관계의 연관성(association)과 인과모형의 인과성(causation)을 종합한 개념으로 정리할 수 있다. 또한 회귀분석은 계량적 종속변수와 하나 혹은 그 이상의 독립변수들간의 관련성을 분석하는 데 있어 매우 강력한 분석력을 갖고 있으며, 또한 적응성이 뛰어난 특성을 가지고 있다. 회귀분석의 일반적인 형태는 1차 방정식의 함수관계로 나타난다.2

회귀분석이 활용되는 사례

- 어떤 연관성을 가지고 있는 종속변수의 유의적인 변동이나 분산을 설명하기 위하여, 종속변수와 관계 있는 독립변수들 중 각각의 독립변수가 어느 정도의 설명력을 가지고 있는가를 결정할 때

- 강한 관련성을 갖고 있는 독립변수가 그와 관련된 종속변수를 어느 정도 설명하고 있는지를 결정하려 할 때

- 독립변수와 종속변수들 간의 수학적 방정식에 의해 그 관련성의 구조(structure)나 형태(form)를 파악하고자 할 때

- 종속변수의 미래 가치를 예측하고자 할 때

- 어떤 특별한 변수나 혹은 변수들의 집합(set)에 대한 기여도를 평가하는데 있어, 다른 독립변수를 통제하려고 할 때

설명

좀더 쉬운 사례를 들어 생각해 보자.

최고기온과 최근 빙과류 판매 사이에 다음과 같은 상관관계가 있다고 가정하면,

25도: 260만개

26도: 270만개

27도: 280만개

28도: 290만개

29도: 300만개

그렇다면 기온이 32도일때 판매량은 얼마라고 예상할 수 있을까? 그것을 알아내는 것이 바로 회귀분석의 목적이다.

위의 경우라면 330만개라고 쉽게 추측할 수 있다. 위의 데이터에서 **"판매량 = (최고기온 X 10만) + 10만"**이라는 식을 쉽게 도출해 낼 수 있기 때문이다. (물론 실제 세계에서 이렇게 극단적으로 선형적인 상황은 존재하지 않는다)

기업에서는 회귀분석을 통해 무엇을 하고자 할까? 몇가지 사례를 들면 다음과 같은 것이 있을 수 있다.

1) 매출액에 영향을 미치는 변수는 무엇인지

2) 이들 변수들이 어느 정도로 영향을 미치고 있는지를 알고 싶어하며

3) 나아가 미래 매출액을 예측할 수 있기를 바란다.

기업의 회귀분석 활용 사례를 추가한다.3

생존의 기로에 선 백화점 업계의 노력

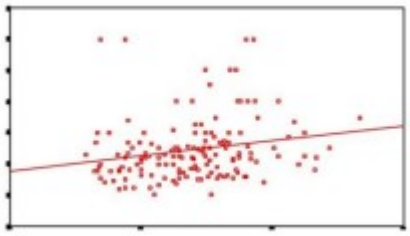
국내에서 최고의 품격을 자랑하는 A백화점에서는 여러가지 브랜드의 대형 할인매장의 등장으로 인하여 경영상 새로운 전환점을 맞이하게 되었다. 물론 백화점과 할인매장의 특성상 고객계층이 다르기 때문에 크게 관계가 없다는 자세를 견지하고는 있지만, 내심 가격할인에 따른 소비자들의 반응을 예의 주시하지 않을 수 없었다. A백화점에서는 마케팅 전략기획 회의를 통해 기존에 자사의 유통체인을 이용하는 고객들을 세분화하고, 그들이 자사 백화점에 충성도를 가지고 있는 것에 어떤 특성이 있는가를 파악하여 이를 보다 강화하는 전략을 수립하기로 결정하였다.

자사 백화점 고객의 충성도 모델이라는 명칭으로 개발된 모형은 자사의 고객들이 자사 백화점을 찾는 이유를 8가지 범주의 차원으로 구성하여 설명하고 있다. 종속변수는 자사 백화점에 대한 충성도의 정도를 7점 척도로 평가하였다. 독립변수는 1) 판매되는 제품이나 상품의 품질, 2) 진열된 상품의 다양성과 구색, 3) 교환 및 환불 시스템, 4) 매장 내에서의 인적인 서비스 품질(응대성과 친절성, 적극성 등), 5) 가격특성, 6) 쇼핑객을 위한 제반 편의시설 구성과 위치, 7) 쇼핑객의 동선을 고려한 점포의 배열, 8) 신용카드와 영수증 발급시스템 등으로 구성하고 역시 7점 척도를 사용하여 측정하였다. 그 결과 8개의 선택범주 중에서 매장 내에서의 인적인 서비스 품질이 백화점의 충성도를 높이는 데 결정적인 기여를 하고 있는 것으로 분석되었다. 모든 변수들의 회귀 계수는 모두 정(+)의 값을 취하고 있었으며, 개별적인 차원의 세부평가요소들 역시 자사 백화점을 선택하는 데 있어 통계적으로 유의한 영향을 미치는 것으로 나타나고 있었다. 이 모델은 자사 백화점을 이용하는 고객들의 선호도와 충성도를 예측하는 데 매우 적합한 모델로 판며오디었으며, 그 후 A백화점은 이러한 평가모델을 활용하여 지속적인 고객 선호도와 충성도를 평가하여 개선하고 있으며, 국내에서 최고수준의 격조 잇고 품위 있는 그리고 가장 믿을 만한 백화점으로 성장하였다.

회귀분석은 **단순회귀분석**(simple regression analysis)과 **다중회귀분석**(multiple regression analysis)으로 나눈다.
단순회귀분석은 한 개의 독립변수를 이용하여 종속변수를 설명, 예측하는 것으로 회귀분석의 가장 단순한 형태이다.

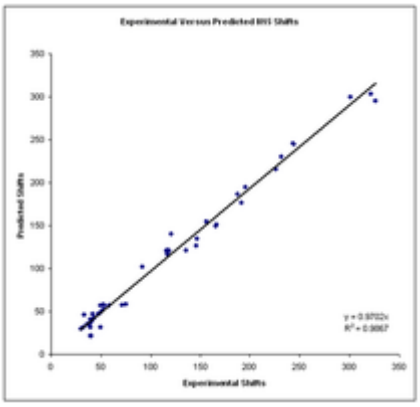
중회귀분석은 여러개의 독립변수와 종속변수 사이의 관계를 설명, 예측하고자 할 때 사용할 수 있는 분석 방법이다.

다음 그림은 단순회귀분석의 사례들이다.



위와 같이 무질서해 보이는 분포에서도 회귀분석이 가능하다.

붉은 직선이 위의 분포에서 회귀분석을 통해 얻은 결과이다.



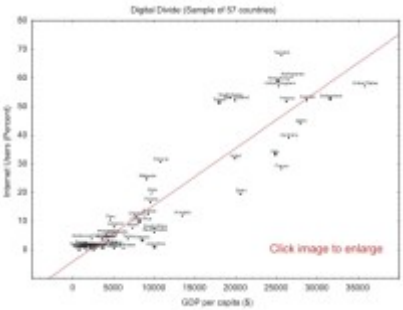
Y = 0.9702X 라는 직선식을 도출하였다.

R제곱은 **결정계수**(coefficient of determination)라고 부른다.

결정계수는 도출한 직선이 실제 데이터 분포를 얼마나 적합하게 반영하고 있는가를 말해준다.

R제곱이 0이면 실제 변화량을 직선이 전혀 반영하고 있지 못한것이고,

R제곱이 1 이면 위의 날씨와 빙과류 판매의 경우처럼 실제 변화량을 직선이 정확히 반영하고 있는 경우이다.



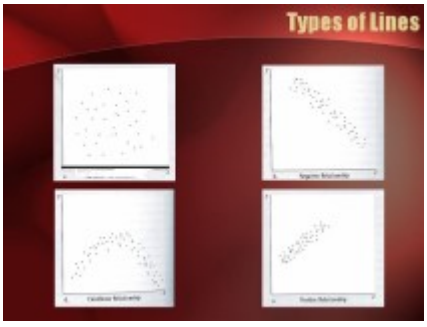
국가들의 1인당 GDP와 인터넷사용자 비율 간의 관계 분포와 회귀분석 결과

양의 상관관계가 있음을 알 수 있다.

회귀분석한 직선이 분포를 꽤 잘 설명하고 있다.

구글 검색을 통해 파워포인트 파일을 찾아 보았다. 그 중 기본이 되는 개념만 정리해 보았다. 파일 다운로드 받으려면 다음 링크를 클릭하면 된다.

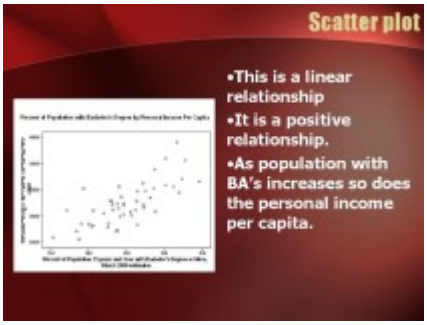
http://web4.uwindsor.ca/users/l/lmiljan/Scopeand.nsf/9d019077a3c4f6768525698a00593654/e639e0cdf0d162c985256bb2004c8fde/$FILE/Regression%20Analysis.ppt



상관관계에는 위와 같은 형태들이 있다.

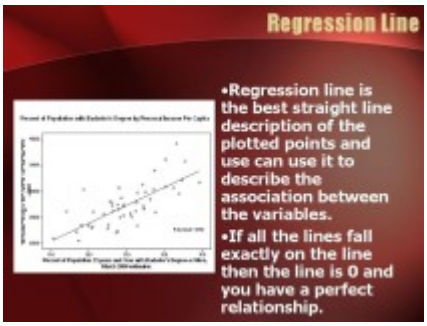
좌측상단은 서로 상관관계가 없는 것이고, 우측상단은 음의 상관관계,

좌측하단은 곡선적인 상관관계 우측하단은 양의 상관관계를 보인다.



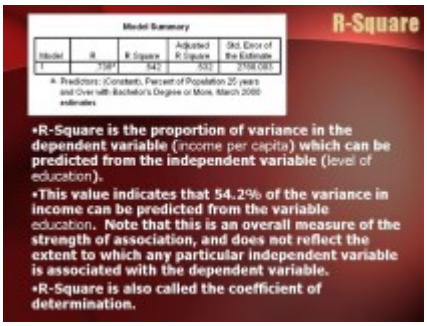
대졸 이상의 학력자 수와 1인당 소득의 관계를 나타낸 것이다.

아직 직선을 긋지 않았지만 한 눈에 선형적인 상관관계가 있음을 짐작할 수 있다.



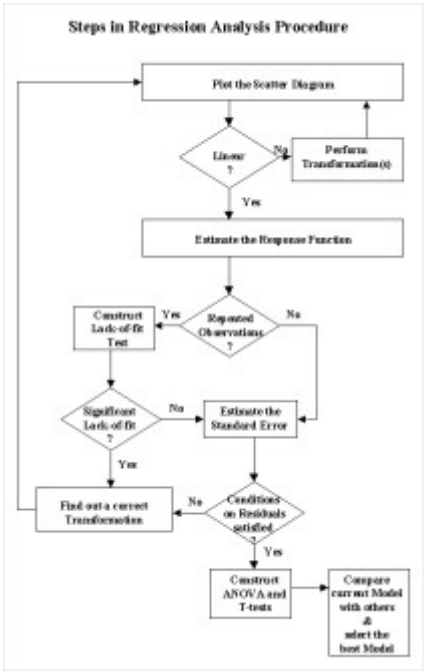
선을 그었다. 그 선을 regression line이라고 한다.

둘 사이의 상관관계를 가장 잘 설명할 수 있는 선이다.



R제곱 즉 결정계수(coefficient of determination)에 대한 설명이다.

다음은 회귀분석의 절차를 나타낸 순서도이다.



회귀분석의 **단계**

모든 위의 자료는 마케터배씨의 사이트에서 2010.05에 가져왔음을 밝힙니다.

http://blog.naver.com/sako71/130087110672