

Chapter 13 Correlation and Regression Analysis

담뱃세·소주세 인상, 판매율에 상관관계 있나?

[더팩트 | 신진환 기자] 폐암과 간암의 주범으로 알려진 담배와 소주. 최근 사람들은 건강에 대한 관심이 높아지면서 담배와 술의 가격 인상 여부가 사회적 화두로 떠올랐다.

최근 문형표 보건복지부 장관이 국민 건강증진을 위해 담뱃세와 주류세를 올리려는 의지를 드러냈다. 이처럼 정부는 '건강의 절대악'인 담배와 주류의 세금을 올려 강력한 금연·금주 정책으로 흡연율과 음주율을 낮추겠다는 방안이다. 세금 인상과 판매량의 상관관계가 있는 것일까.



9 일 업계와 통계청에 따르면 담배의 경우 담뱃세를 인상했을 때 흡연율은 떨어진 것으로 나타났다. 지난 2004 년 말 담뱃세를 500 원 올리자 판매량이 감소하면서 57.8%에 이르던 성인 남성의 흡연율(2004 년 9 월)이 44.1%(2006 년 12 월)까지 13%포인트 이상 떨어졌다.

그러나 지속적인 결과는 얻지 못했다. 2008 년 이후 흡연율 하락 추세가 정체 현상을 보이기 때문이다. 실제로 지난 2004 년 57.8%에 이르던 성인 남성의 흡연율이 2 년 후인 44.1%로 13%포인트 이상 낮아졌다.

담배 소비액도 떨어지는 추세를 보였다. 담뱃세 인상 등 요인으로 금연 분위기가 확산하면서 가구당 월평균 담배 소비액은 2006 년 2 만 2062 원부터 점차 떨어졌다. 담배 소비액이 늘었던 경우는 2010 년(0.7%)뿐이다.

세계은행 등의 연구결과에 따르면, 담배수요의 가격탄력성은 -0.3~-0.5 로 담뱃세를 10% 인상했을 때 소비량은 3~5% 감소한다. 즉, 담배가격이 흡연의 진입장벽 역할을 하고 잠재적 신규소비자의 진입을 억제하는 효과를 발생하는 것. 때문에 담뱃세를 올리면 담배 소비량은 감소하는 추이를 보인다.

술의 경우도 담배와 비슷한 현상을 보이고 있다. 소주 시장 점유율 50% 이상을 차지하고 있는 하이트진로의 '참이슬'은 지난 2012 년 12 월 말에 주정 가격 인상과 원부자재, 유가 상승 등으로 출고가(360 ml 병 기준)가 종전 888.9 원에서 961.7 원으로 8.19% 올랐다.

당시 가격 인상으로 2013 년 1 월 판매량이 급감했지만, 그해 1 분기 소주 판매량은 전년 같은 기간보다 10.6% 증가했다.

통계청의 '가계동향조사'를 보면, 지난해 가구당 월평균 주류 소비는 1 만 751 원으로 나타났다. 이는 통계 작성 시작 이래 가장 많은 수치이다. 2003 년 6359 원이었던 가구당 월평균 주류 소비는 2004 년 7000 원, 2009 년 8356 원, 2010 년 9021 원, 2011 년 9400 원, 2012 년 9779 원, 2013 년 1751 원으로 10 년간 매년 늘었다.

소주세 인상은 지난 2004 년 4 월(740 원→800 원), 2008 년 12 월(800 원→888.9 원)에도 있었다. 그럼에도 주류 소비액은 꾸준히 증가하고 있다. 즉, 소주세가 인상되더라도 판매율에 큰 영향을 끼치지 않는 것으로 보인다.

소주세가 오르더라도 담배보다 영향이 적은 이유는 상승 폭이 작기 때문이다. 더불어 소주세 인상으로 출고가가 오르더라도 소주를 취급하는 음식점에서 가격을 동결할 경우 소비자들은 이를 알기 어렵다. 소주세가 인상되더라도 소비자는 담뱃세보다 체감하는 물가 상승이 적기 때문에 판매율에 영향을 미치게 끼치는 것으로 풀이된다.

한 주류업계 관계자는 "주정 가격 인상 및 원부자재가격 등을 고려해 인상하더라도 실제 인상 금액은 50 원에서 100 원 사이"라며 "그래서 소주세가 인상돼도 소비자들은 큰 폭으로 올랐다고 느끼기 어려울 것"이라고 설명했다.

Biz focus 2014/09/09

지금까지 우리는 통계로 부터 가설로 설정된 내용이 맞는지를 확인하기 위한 여러가지 방법들을 알아 보았습니다. 통계의 또 다른 기능으로 예측이 있습니다. 이러한 내용은 담뱃세와 소주세 인상과 판매율에 상관 관계를 다룬 기사에서 볼 수 있습니다. 겉으로는 금연이나 음주량을 줄이기 위해 세금 인상을 말할 수 있지만, 과거 자료로 부터 상관관계를 조사해 보면 여러 가지 요인에 의해서 반드시 그렇지 않음을 알 수 있습니다. Marketing 에서는 상관 관계를 상쇄하기 위해서 Just-noticeable difference 을 이용하기도 합니다. 예를 들어 가격 인상을 눈치채지 못할 정도로 조금씩 올려서 판매율에 영향을 미치지 않게 하거나 반대로 내용은 같지만 제품에 부피가 커진것을 인지할 수 있을 정도로 증가하여 소비자에게 회사에 대한 좋은 이미지를 줄 수 있습니다.

상관 관계를 correlation 이라고 하고 측정된 값이 +1 에서 -1 사이에 값으로 형성됩니다. 가격 할인과 판매량에서 correlation 값이 0 이면 두 데이터는 관계가 없다는 의미이고, 양수이면 할인률의 증가는 판매량의 증가를 나타낼 경우, 음수이면 할인률 증가는 판매량 감소를 의미합니다. 음수일 경우에 일반적인 예는 운동량의 증가는 체중의 감소와 같은 경우에 해당 합니다.

이와 같은 관계를 알게 되면 할인률로 부터 판매량을 예측이 가능해집니다. 예를 들어 할인률을 10%증가하였더니 판매량이 20%증가하였고, 20%증가하였더니 40%증가하였다는 조사 결과가 나타나면, 할인률을 30%로 증가하였을 경우 판매량은 60%증가가 증가할 것이라고 예측을 할 수 있습니다. 이러한 관계는 수식으로 표현하면

$$y = 2x$$

와 같습니다. 여기서 변수 x 는 할인률이고 y 는 판매량 증가를입니다.

정리를 하면 이번 장에서는 이러한 correlation 과 예측을 위해 한개의 변수를 갖고만 하는 regression 을 학습하고 다음 장에 여러 변수로 예측을 하는 multiple regression model 을 학습하겠습니다. 이 곳에 다룰 내용들은 선형(linear)관계에서만 다룰 것이고 비 선형(non-linear) 관계는 다루지 않습니다.

Linear 라는 용어를 설명하면 다음과 같습니다.

$$\begin{aligned} y_1 &= f(ax_1) = af(x_1) \\ y_2 &= f(bx_2) = bf(x_2) \end{aligned}$$

Homogeneity 로 수식에서 보듯이 x 값의 증가는 y 에도 똑같이 증가를 발생합니다.

$$f(x_1 + x_2) = f(x_1) + f(x_2)$$

다음으로 Additivity 로 입력값의 합은 두 결과값의 합과 같아야 합니다. 이러한 조건이 성립이 되면 함수 f 는 linear 라고 말을 합니다.

Nonlinear 한 경우는 앞의 두 가지 조건을 성립하지 못하는 경우로 예를 들어 다음 수식을 보시면 이러한 관계가 성립이 되지 않습니다.

$$y = f(x) = x^2$$

x 의 c 배 증가는 y 에 c^2 배 증가를 만듭니다. 예를 들어 x 값이 2 이면 y 값은 $4a$ 이지만 이에 두배인 4 를 넣으면 $16a$ 로 4 배가 증가를 합니다. 이러한 경우를 nonlinear 라고 합니다.

1. Dependent and Independent variable

Independent variable 은 다른 변수에 영향을 받지 않는 변수이고 반대로 dependent variable 은 independent variable 에 영향을 받아 값이 변하는 변수입니다. 예를 들어 할인률이 판매량 증가에 영향을 미치는 경우 할인률은 independent variable 이고 판매량 증가율은 dependent variable 입니다.

그런데 independent variable 와 dependent variable 의 관계는 반대로 성립되지 않습니다. 예를 들어 판매량의 변화는 할인률 변화에 영향을 주지 않습니다. 기업들이 판매가 잘되는 제품에 일부러 이윤을 버리면서 판매량이 100%증가 했다고 할인률을 50%증가시킬 이유가 없기 때문입니다.

2. Correlation

Correlation 은 두 변수 사이에 선형 관계의 강도와 방향을 +1 ~ -1 사이 값으로 알려 줍니다. 예를 들어 정확하게 independent 변수의 2 배 증가가 dependent 변수에 2 배 증가를 하게 한다면 선형 관계가 성립이 되면서 증가 방향도 같기 때문에 correlation 값은 +1 이 됩니다. 하지만 할인률 증가에 따른 판매량 증가가 항상은 되지만 행사할 때 마다 증가율이 다르고 증가율에 일정하게 비례하여 증가가 이루어 지지 않은다면 선형 관계가 약해져 0 ~ 1 사이에 값이 됩니다.

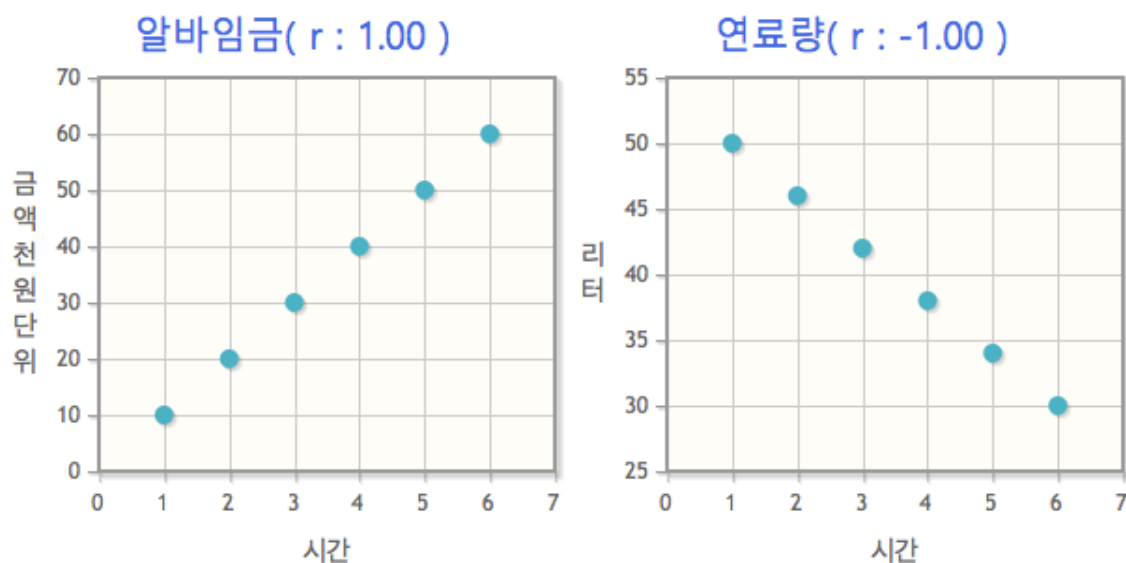
이와 반대로 운동량의 따른 체중의 변화가 반비례하지만 반비례값이 일정하지 않고 상황에 따른 다르다면 correlation 값은 -1 ~ 0 사이의 음수 값이 됩니다.

이러한 correlation 관계를 숫자로 표시한것을 correlation coefficient 라고 하고 수식은 다음과 같습니다.

$$r_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y} = \frac{E(xy) - \mu_x \mu_y}{\sqrt{E(x^2) - \mu_x^2} \sqrt{E(y^2) - \mu_y^2}} \quad (13.1)$$

그럼 경우에 따른 correlation coefficient 값의 차이를 그래프와 함께 알아 보도록 하겠습니다.

chapter13/corr1.html



첫번째 경우는 같은 종류의 알바 시간당 임금을 조사한 결과 모든 알바가 시간당 만원인 경우이고, 두번째 경우는 석유를 연료를 사용하는 난방장치에 시간당 남은 석유량으로 시간당 4 리터씩 소비하는 것으로 나타났습니다. 이러한 경우 시간과 임금그리고 시간과 석유 소비량의 correlation 은 완벽하게 강한 관계를 갖고 있어 절대값이 모두 1 이지만 관계의 방향은 반대 입니다.

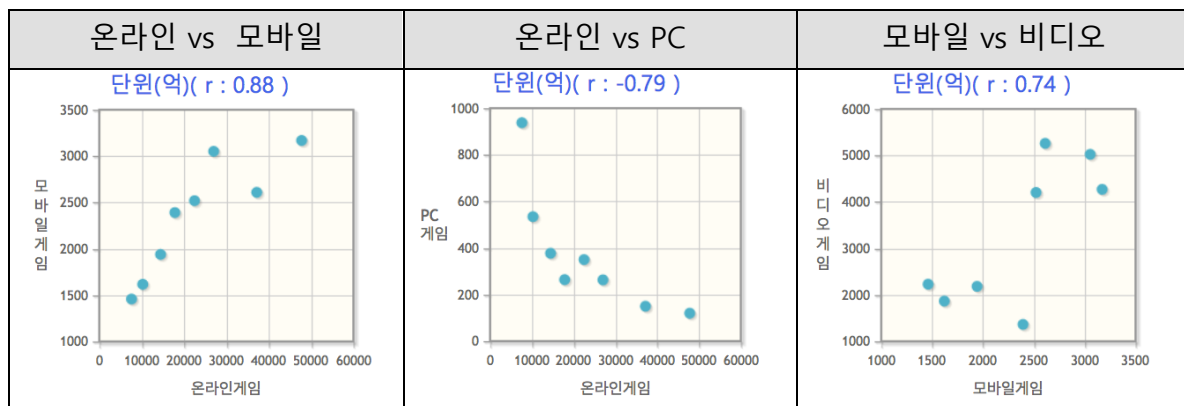
다음은 통계청에서 2003 년 부터 2010 년까지의 게임 산업의 매출액으로 게임 종류별 상관관계를 correlation coefficient 로 알아 보도록 하겠습니다.

	2010	2009	2008	2007	2006	2005	2004	2003
온라인게임	47,672	37,087	26,922	22,403	17,768	14,397	10,186	7,541
아케이드게임	715	618	628	352	7,009	9,655	2,247	3,118
비디오게임	4,268	5,257	5,021	4,201	1,365	2,183	1,866	2,229
모바일게임	3,167	2,608	3,050	2,518	2,390	1,939	1,617	1,458

pc 게임	120	150	263	350	264	377	534	937
-------	-----	-----	-----	-----	-----	-----	-----	-----

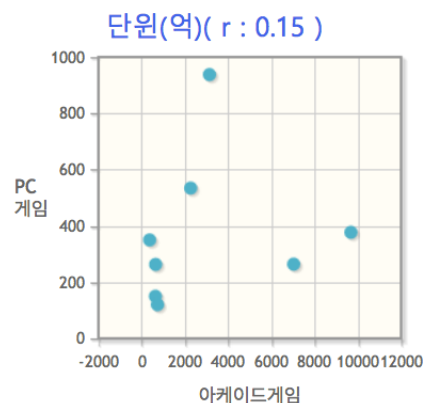
단위: 억원

chapter13/game.html



매출액 테이블에서 보는 것과 같이 온라인, 모바일, 비디오 게임은 매출이 계속 늘어나지만 PC 게임은 줄어 들고 있습니다. 그래서 온라인, 모바일, 비디오 게임의 매출액의 상관 관계는 양수로 나타나지만 이 3 가지 게임과 PC 게임의 상관관계는 반대로 움직이기 때문에 음수로 나타납니다.

다음 PC 게임과 아케이드게임의 매출액 관계를 보면 PC 게임의 매출액이 2006 년과 2008 년에 거의 비슷하지만 아케이드 게임은 매출액이 4 배가 증가하고 반대로 아케이드 게임의 매출액이 2003 년과 2005 년에 비슷하지만 PC 게임은 매출액이 3 배로 감소합니다. 즉 이 두 가지 게임의 매출액에는 상관관계를 찾기 어렵고 이것은 correlation coefficient 를 보면 알 수 있습니다.



수식 13.1 은 비록 계산을 할 때 사용된 표준편차값들이 population 의 표준편차를 계산하는 방식으로 sample 크기 n 으로 나누어 계산을 합니다.

이 값은 sample 로 부터 얻은 결과이기 때문에 값의 신뢰성을 알아보아야 합니다. 이를 위한 hypothesis 절차는 다음과 같습니다. 이를 위해서 온라인 게임과 PC 게임의 자료를 갖고 알아 보겠습니다.

1) Hypothesis 설정

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Population 의 correlation coefficient 가 0 인가 아닌가를 확인하여 sample 로 부터 얻은 correlation coefficient 를 통해서 두 변수간의 관계가 있는지 없지에 대한 확신을 얻는 것입니다.

2) Significance level: 0.05

3) Test statistic

$$t = \frac{r_{x,y}}{\sqrt{\frac{1 - r_{x,y}^2}{n - 2}}} \quad (13.2)$$

Student's t-distribution 을 이용한 합니다. 여기서 n 은 sample 의 크기 입니다. 게임 매출액의 예제에서 n 은 8, correlation coefficient 값은 -0.79 로 이를 적용하면

$$t = \frac{-0.79231}{\sqrt{\frac{1 - 0.79231^2}{8 - 2}}} = -3.18102$$

4) Critical value:

Two-tail 검사이므로 $t_{0.025} \sim t_{0.975}$ 사이에 값으로 degree of freedom 은 n-2 로 6 을 적용하여 Student's t-distribution 에 해당 값의 범위는 -2.4469 ~ 2.4469 입니다.

5) Confidence interval

이 계산을 위해서는 normal distribution 을 이용해서 값을 범위를 계산해야 합니다. 그런데 수식 13.2 는 $\rho = 0$ 일 경우에만 즉 null hypothesis 가 참일 때만 대칭적으로 분포되지만 그

외의 null hypothesis 인 경우는 분포는 skew 되어 normal distribution 에 형태를 갖추지 못합니다. 이를 보완한 것이 Fisher's Z transform 입니다.

$$\tanh(Q \pm z_{0.05}s)$$

여기서 Q 가 Fisher's Z transform 이고 이 값은 평균

$$Q = \tanh^{-1} r_{x,y} = \frac{1}{2} \ln \frac{1 + r_{x,y}}{1 - r_{x,y}} \quad (13.3)$$

중심으로 standard error 인 s 값

$$\frac{1}{\sqrt{n-3}}$$

으로 normal distribution 을 형성합니다.

이러한 과정으로 얻은 값을 r 로 변환하기 위해서 Fisher's Z transform 의 inverse 과정을 거치게 됩니다.

$$\tanh x = \frac{e^{2x} - 1}{e^{2x} + 1}$$

이렇게 수식을 적용하여 계산된 Confidence interval 은 -0.198 ~ -0.961 이 됩니다.

6) p-value

p-value 를 계산하기 위해서 다시 Student t-distribution 으로 부터 계산을 하여야합니다. 여기서 two tail 검사이기 때문에

$$2 \times \min \left(P(t < \bar{t}_{\alpha/2} | df), P(t > \bar{t}_{1-\alpha/2} | df) \right)$$

에 의해서 0.019 가 됩니다.

7) Null hypothesis reject 검사

Test statistics 는 critical value 보다 작고, CI 도 0 이 포함이 안되고, p-value 역시 0.05 보다 작기 때문에 null hypothesis 인 population 의 correlation coefficient 가 0 이라는 것을 주장하기에 충분한 근거를 찾지 못하게 되었습니다. 이를 해석하면 온라인 게임은 성장하고 PC 게임은 하향산업이 되고 있음을 알 수 있습니다.

만일 두개의 population coefficient 값이 같은가를 검사하는 방법은 Fisher's Z transform 으로 결정을 할 수 있습니다.

$$Q_1 - Q_2 = \frac{1}{2} \left(\ln \frac{1+r_1}{1-r_1} - \ln \frac{1+r_2}{1-r_2} \right)$$

이 값은 null hypothesis 인 평균 0 에 표준 편차

$$\sigma = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

으로 normal distribution 에 z score 를 이용하여 검사를 할 수 있습니다.

3. Simple Regression Analysis

Correlation 은 두 변수의 연관성 및 방향을 제시하지만 예측을 하는 경우에는 사용을 하지 못합니다. 이를 위해 이 장에서는 가장 간단한 예측 모델을 데이터로 부터 찾는 방법을 소개하도록 하겠습니다.

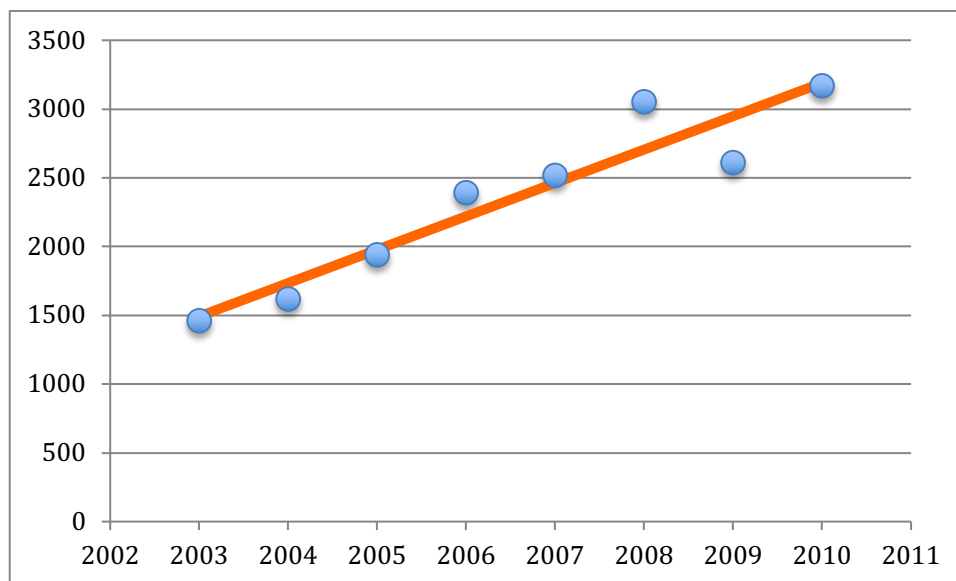
$$\hat{y} = a + bx \tag{13.4}$$

여기서 \hat{y} 는 independent variable 에 대한 예측되는 dependent variable 입니다.

이렇게 independent variable 이 한개인 경우를 simple regression 이라고 하고 여러개 있을 경우를 다음장에 배울 multiple regression 이라고 합니다.

Regression 은 사전적인 의미로 퇴행, 퇴보, 회귀라는 의미를 갖고 있습니다. 이럼 이 단어가 왜 사용하는가를 이해하면 regression analysis 를 이해하는데 도움이 될 것입니다. 진실은 하나이지만 말은 여러가지로 나타나고 어떤 말에서는 진실을 숨기려하고 어떤 말은 진실을 드러내려고 합니다. 예를 들어 사고가 나고 주변 사람들과 책임자들에게 사고 현장에서 발생한 일들을 물어 보게 됩니다. 이렇게 모아지 정보를 토대로 짜집기를 하게 되는데 점점

정보가 많아 질 수록 사고가 발생한 원인이 나타나게 되고 어떤 원인에 의해서 사고가 발생했음을 알게 됩니다. 다시 말해, 이렇게 진실의 정보를 담고 있는 관찰된 자료를 취합하여 조사를 하게되면 점점 사실에 가까워 집니다. 즉 정보가 사실로 되보, 회귀하게 됩니다. 이와 같이 regression analysis 는 관찰된 정보를 토대로 전체적인 데이터의 흐름 다시 말해 트렌드와 같은 정보를 얻게되어 결과를 토대로 앞으로 발생할 사건을 예측을 할 수 있습니다.



년도별 모바일 게임 매출액

이러한 Regression analysis 는 모바일 게임 매출액과 같은 트렌트를 조사하는데 중요한 역할을 합니다. 점들은 실제 매출액이고 이 점들 중간에 있는 선이 점들에 대한 관계 및 방향을 제시하면서 동시에 2011 년 이후 매출액을 예상할 수 있도록 합니다. 선에 대한 수식은 다음과 같습니다.

$$\hat{y} = 242.6071x - 484450$$

여기서 \hat{y} 는 예상 매출액이고 x 는 년도입니다. 예측값과 실제값의 차이를 residue 라고 합니다.

$$e_i = y_i - \hat{y}_i \quad (13.5)$$

이제 이러한 simple regression 을 수식을 얻는 방법을 학습하도록 하겠습니다.

기본 원리는 (x, y) 로 있는 값들에 가장 잘 맞는 즉 예측과 실제 y 값과의 차이들이 가장 작게 해주는 선을 찾는 방식으로 least square 방식을 이용합니다. 이 결과로 찾은 선을 regression line 이라고 합니다. 이를 통해 찾는 값은 수식 13.4 에 slope 값 a 와 y 와의 교차하는 b 값입니다.

그럼 least square 방식을 수식으로 표현하는 sum of square error 를 보시겠습니다.

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (13.6)$$

SSE 가 최소값으로 만드는 \hat{y} 를 만드는 a 와 b 값을 찾는 방법은 다음과 같습니다.

우선 SSE 를 최소로 만들기 위한 SSE 를 b 에 대해 미분한 값이 0 이 되는 지점을 찾습니다.

$$\begin{aligned} \left(\sum_{i=1}^n (y_i - bx_i - a)^2 \right)' &= -2 \sum_{i=1}^n (y_i - bx_i - a) = 0 \\ a &= \frac{1}{n} \left(\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i \right) = \bar{y} - b\bar{x} \end{aligned} \quad (13.7)$$

같은 방법으로 기울기 a 값을 계산을 하도록 하겠습니다.

$$\begin{aligned} \left(\sum_{i=1}^n (y_i - bx_i - a)^2 \right)' &= -2b \sum_{i=1}^n x_i (y_i - bx_i - a) = 0 \\ \sum_{i=1}^n x_i y_i - b \sum_{i=1}^n x_i^2 - a \sum_{i=1}^n x_i &= 0 \end{aligned}$$

여기에 수식 13.7 을 적용하면

$$n \sum_{i=1}^n x_i y_i - bn \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n y_i - b \left(\sum_{i=1}^n x_i \right)^2 = 0$$

$$b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{cov(x, y)}{\sigma_x^2} = r \frac{\sigma_y}{\sigma_x} \quad (13.8)$$

수식 13.7 과 13.8 을 이용해서 모바일게임 매출액 에대한 regression line 을 만들어 보면

$$\begin{aligned} \bar{x} &= 2006.5, \bar{y} = 2343.4 \\ \sigma_x &= 2.2913, \sigma_y = 586.8773 \\ r &= 0.9472 \end{aligned}$$

이 값을 통해 13.8 을 적용하여 기울기를 구하면

$$b = 0.9472 \frac{586.8773}{2.2913} = 242.6071$$

y 축과 만나는 지점 a 값은

$$a = 2343.4 - 242.6071 \times 2006.5 = -484450$$

그럼 이렇게 계산된 기울기와 y 축과 만나는 점에 대한 확신성에 대한 검사를 하는 방법들에 대해서 알아 보도록 하겠습니다.

3.1. Sum of Square 해석

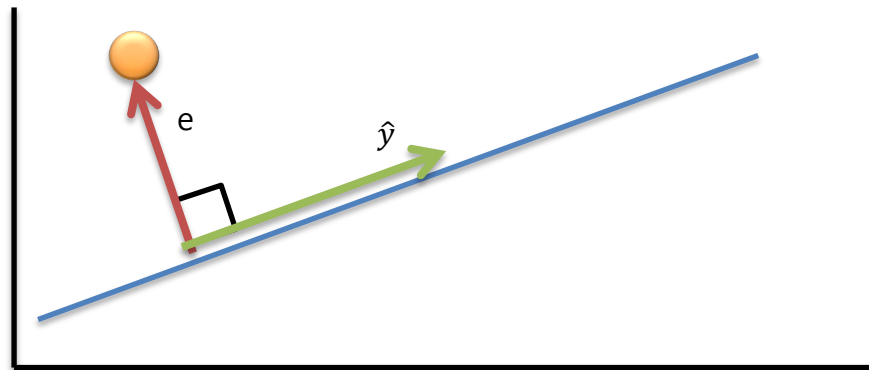
Total sum of square (SST)는 ANOVA 를 하면서 이미 학습한 내용으로 Regression 에서의 표현은 다음과 같습니다.

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (13.9)$$

여기서 y 값은 sample 로 부터 얻은 dependent variable 의 값이고 \bar{y} 은 이것에 평균입니다. 이 값은 표준 편차와 같이 값들이 응집되면 작아지고 퍼져 있으면 큰값을 갖게 됩니다. 이 값은 다음과 같이 분리가 됩니다.

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

여기서 중요한 내용은 residue 인 e_i 와 예측값 \hat{y}_i 은 수식 내부에서 vector 의 dot 연산의 결과로 나타나는데 이 값에 대한 vector 는 서로 수직 관계으로 있기 때문에 곱하게 되면 0 이 됩니다.

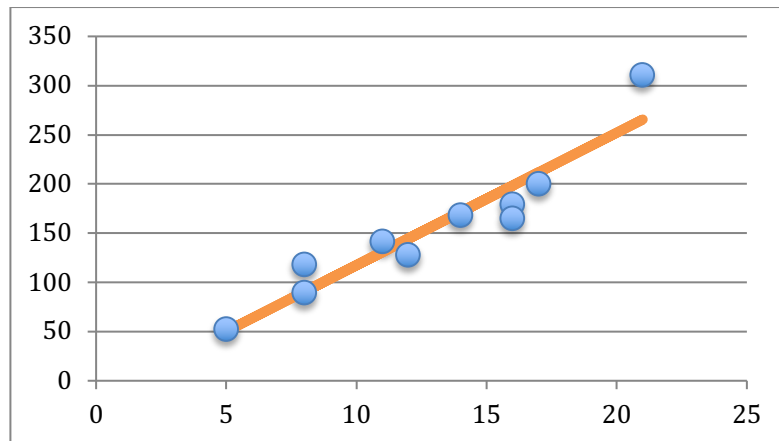


이러한 점을 이용하여 SSE 를 간략하게 바꿀 수 있습니다.

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n y_i(y_i - \hat{y}_i) - \sum_{i=1}^n \hat{y}_i(y_i - \hat{y}_i) = \sum_{i=1}^n y_i(y_i - \hat{y}_i) - \vec{\hat{y}} \cdot \vec{e} (= 0) = \sum_{i=1}^n y_i e_i = \vec{y} \cdot \vec{e}$$

설명을 위해 자동차가 소비한 연료에 이동거리를 이용한 설명으로 수식을 설명을 하도록 하겠습니다.

	연료소비량(리터)	이동 거리(km)	예측 거리(km)	평균연비(km/리터)
1	16	179.3353	198.2487	11.21
2	14	168.2739	171.3335	12.02
3	5	52.4114	50.2151	10.48
4	12	127.9184	144.4183	10.66
5	8	117.682	90.5879	14.71
6	21	310.394	265.5367	14.78
7	11	141.6365	130.9607	12.88
8	16	164.7824	198.2487	10.30
9	8	89.3912	90.5879	11.17
10	17	200.0185	211.7063	11.77
평균	12.8	155.1844	155.1844	11.9976
표준편차	4.6648	66.4615	62.7765	1.5557



소비된 연료량과 이동 거리에 대한 Correlation coefficient 값은 0.9446 으로 값들을 수식 13.8 에 적용하여 기울기를 구하고 13.7 로 y 축과 만나는 지점을 계산을 하면

$$\hat{y} = 13.4576x - 17.0729$$

$$SST(\approx 44171) = SSR(\approx 39409) + SSE(\approx 4762)$$

여기서 SSE 가 의미하는 것은 이동거리인 dependent variable 의 변화의 원인으로 independent variable 인 연료 소비량에 의한 것이 아닌 다른 요인에 의한 변화인데 이러한 결과는 평균연비가 다르게 나오는 결과를 만듭니다. 예를 들어 가다 서다를 반복하는 도심 주행을 많이 했다면 소비된 연료량 보다 실제 이동거리가 작게 나타날 것이고 고속도로와 같이 일정 속도로 계속 달리는 경우는 연비가 좋게 나타나게 됩니다. 즉 연료 소비량외에 다른 영향을 미치는 요인에 의한 이동거리의 측정값과 실제 값의 차이에 대한 정보가 SSE 입니다.

이와 반대로 SSR 은 연료 소비량에 의한 이동거리에 반영되는 정보에 대한 정보를 나타냅니다.

3.2. Coefficient of determination, R^2

Dependent variable 의 전체 변화의 합에 대한 percentage 를 측정한 값입니다.

$$R^2 = \frac{SSR}{SST} \quad (13.10)$$

R^2 값의 범위는 0 에서 1 까지의 값으로 1 이 가까울 수록 dependent variable 의 변화는 independent variable 의 변화에 의한 것을 의미하는 것이고 반대로 0 에 가까울 경우는 dependent

variable 의 변화는 independent variable 의 변화와 무관하게 되어 상관성이 없다는 것을 의미합니다.

자동차 연비의 예를 적용을 하게 되면

$$R^2 = \frac{39409}{44171} = 0.8922$$

이 값의 hypothesis test 절차는 다음과 같습니다.

1) Hypothesis test 설정

$$\begin{aligned} H_0: \rho^2 &\leq 0 \\ H_1: \rho^2 &> 0 \end{aligned}$$

여기서 ρ^2 값은 population 의 coefficient of determination 값으로 이 값이 0 에서 1 사이에 값이고 두 변수의 관련성을 측정하는 것이기 때문에 upper tail 검사를 실행 합니다.

2) Significance level: 0.05

3) Test statistic

$$F = \frac{SSR}{\left(\frac{SSE}{n-2}\right)} \quad (13.11)$$

여기서 n 값은 sample 크기로 수식 13.11 을 적용을 하게 되면 값은

$$F = \frac{39409}{\left(\frac{4762}{8}\right)} = 66.2058$$

4) Critical value

F-distribution 을 이용하기 때문에 두 개의 degree of freedom 값을 알아야 합니다. 첫번째 DF 값은 1 이고 두번째 DF 값은 n-2 로 여기서는 8-2=6 이 됩니다. 이를 적용하여 critical value 를 구하면 5.9873 이 됩니다.

5) Null hypothesis reject 검사

Critical value 가 test statistic 에 비해 매우 작기 때문에 population 의 coefficient of determination 값은 0 보다 크다는 근거가 나타나고 이러한 사실은 통계적 해석이 없이도 아는 내용입니다만 연비와 이동거리간에 연관성이 있음을 수치적으로 알려 줍니다.

4. \hat{y} 의 Confidence Interval

연료량에 따른 이동 거리의 예측값을 계산을 하였을 때 연료량을 알면 예상되는 이동 거리를 계산할 수 있습니다. 예를 들어 만약 10 리터의 양을 갖고 있다면 이동할 수 있는 거리는

$$\hat{y} = 13.4576 \times 10 - 17.0729 = 117.5031$$

이 값의 의미는 예측된 값이기 때문에 정확한 값은 되지 못합니다. 이를 보완하는 방법은 값 x 에 해당하는 평균 예측값의 confidence interval 을 구하여 실제 평균 값이 있을 영역을 찾는 것입니다. 이 계산을 위해서 필요한 것은 standard error 값입니다.

$$s_e = \sqrt{\frac{SSE}{n-2}} \quad (13.12)$$

이 수식의 의미는 SSE 즉 예측값과 실제값에 차이로 error 값이 클 수록 standard error 는 커지고 sample 의 개수가 적을 수록 또한 standard error 가 커지게 됩니다. 자동차 연비에 대한 계산에서 이 값을

$$s_e = \sqrt{\frac{4762}{8}} = 24.3988$$

그럼 standard error 값을 적용하여 예측값의 CI 계산하는 방법은 다음과 같습니다.

$$CI = \hat{y} \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = \hat{y} \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} \left(1 + \left(\frac{x - \bar{x}}{\sigma_x} \right)^2 \right)} \quad (13.13)$$

여기서 critical t-statistic 을 얻기 위한 degree of freedom 은 $n-2$ 가 됩니다.

연비량 10 리터를 넣었을 때 평균 이동거리에 대한 95% confidence interval 은 다음과 같습니다.

$$CI = 117.5031 \pm 2.306 \times 24.3988 \sqrt{\frac{1}{10} \left(1 + \left(\frac{10 - 12.8}{4.6648} \right)^2 \right)} = 117.5031 \pm 20.7513$$

$$UCL = 138.2544$$

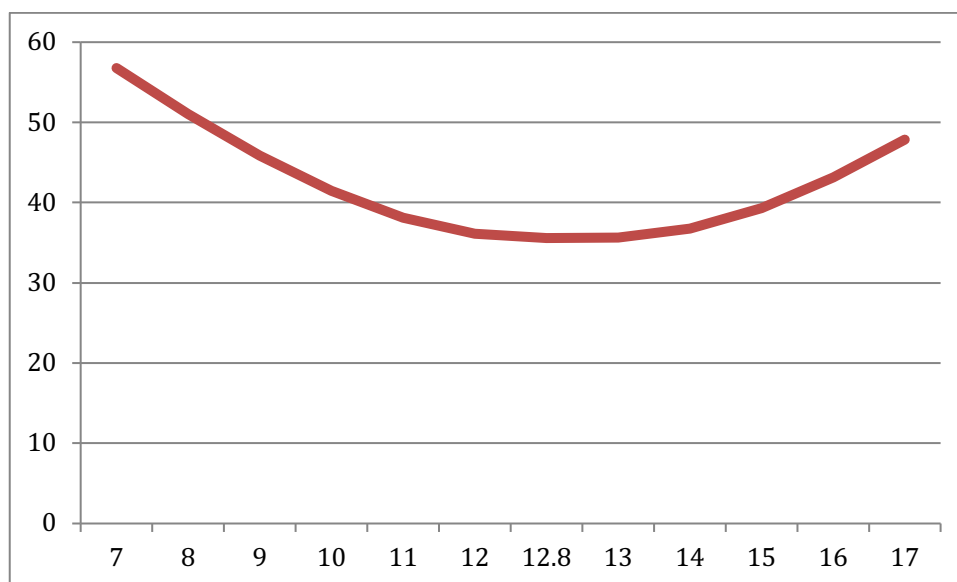
$$LCL = 96.7518$$

이 의미는 모든 동일 모델 차량에 10 리터를 주유하고 달렸을 때 나타나는 이동거리가 모두 117.5031 이 될 수 없지만, 이 값으로 10 리터를 주유한 차량에 평균 이동거리에 대한 95% CI 은

96.7518 에서 138.2544 km 가 된다는 것입니다. 이는 regression line 에 대한 confidence interval 입니다.

이러한 연료량에 대한 평균 이동거리에 CI 값에 폭은 예측하고자하는 연료량이 평균 연료량에 가까울 수록 작아 집니다.

연료량(리터)	예측값(km)	LCL	UCL	폭
7	77.1303	48.74251868	105.5180813	56.77556263
8	90.5879	65.05995188	116.1158481	51.05589625
9	104.0455	81.09819963	126.9928004	45.89460073
10	117.5031	96.75283484	138.2533652	41.50053032
11	130.9607	111.8908058	150.0305942	38.13978846
12	144.4183	126.3672424	162.4693576	36.10211512
12.8	155.18438	137.3930592	172.9757008	35.58264153
13	157.8759	140.0682347	175.6835653	35.61533061
14	171.3335	152.962935	189.704065	36.74113007
15	184.7911	165.120441	204.461759	39.34131808
16	198.2487	176.6736114	219.8237886	43.1501771
17	211.7063	187.7662391	235.6463609	47.88012182



수식 13.13 으로 부터 계산된 CI 는 연료량에 대한 평균 이동거리의 CI 이고, 연료량 예측되는 이동거리의 CI 의 계산은 수식 13.13 을 수정해야 합니다.

$$PI = \hat{y} \mp t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} \left(1 + \left(\frac{x - \bar{x}}{\sigma_x} \right)^2 \right)} \quad (13.14)$$

이 수식을 10 리터에 이동거리에 적용을 하게 되면

$$UPL = 177.469$$

$$LPL = 57.537$$

값의 폭이 CI 보다 더 넓어짐을 알 수 있습니다. 이러한 폭을 줄이는 방법은 sample 크기를 증가 시키는 것입니다.

5. Slope 와 Intercept 검사

Regression line 을 계산하기 위해서 slope 값을 얻었지만 이 값은 만일 x, y 가 관계가 없다면 coefficient 가 0 이기 때문에 slope 역시 0 이 됩니다. 그래서 slope 와 intercept 가 0 인지 아닌지 확인을 하고 이 값들의 CI 를 얻어는 방법을 알아 보도록 하겠습니다.

$$\hat{y} = \alpha x + \beta$$

여기서 α 와 β 는 population 으로 부터 얻은 값으로 이 값이 0 인가 아닌가를 갖고 검사를 하면 됩니다.

1) Hypothesis 설정

	Intercept	Slope
$H_0:$	$\alpha = 0$	$\beta = 0$
$H_1:$	$\alpha \neq 0$	$\beta \neq 0$

2) Significance level: 0.05

3) Test statistics

<div style="border-bottom: 1px solid black; margin-bottom: 5px;">Slope</div> $t_b = \frac{b - \beta}{s_b}$	<div style="border-bottom: 1px solid black; margin-bottom: 5px;">Intercept</div> $t_a = \frac{a - \alpha}{s_a}$
--	---

(13.14)

여기서 a, b 는 sample 로 부터 얻은 값들이고 각각의 standard error 값들은 다음과 같습니다.

<div style="border-bottom: 1px solid black; margin-bottom: 5px;">Slope</div> $s_b = \frac{s_e}{\sigma_x \sqrt{n}}$	<div style="border-bottom: 1px solid black; margin-bottom: 5px;">Intercept</div> $s_a = \frac{s_e}{\sqrt{n}} \sqrt{\frac{E(x^2)}{E(x^2) - \bar{x}^2}} = s_b \sqrt{E(x^2)}$
--	--

(13.15)

소비연료에 해당하는 이동 거리의 regression line 을 적용을 하면 standard error 는

<div style="border-bottom: 1px solid black; margin-bottom: 5px;">Slope</div> $s_b = \frac{24.3988}{4.6648\sqrt{10}} = 1.654$	<div style="border-bottom: 1px solid black; margin-bottom: 5px;">Intercept</div> $s_a = 1.654 \sqrt{\frac{1856}{10}} = 22.5335$
--	---

이 값들을 t statistic 에 적용을 하면

Slope	Intercept
$t_b = \frac{13.4576 - 0}{1.654} = 8.1363$	$t_a = \frac{-17.0729 - 0}{22.5335} = -0.7577$

4) Critical value

Degree of freedom 은 n-2 로 하면 됩니다. 이 예제에서 sample 의 크기가 10 이므로 8 로 하여 0.025~0.975 에 해당하는 t statistic 값은 -2.306 ~ 2.306 이 됩니다.

5) Null hypothesis reject 검사

결과를 보면 slope 는 null hypothesis 를 reject 할 수 있는 근거가 있지만 intercept 는 reject 을 하지 못하여 0 이 될 수 있습니다.

6) CI

Slope	Intercept
$CI_a = a \pm t_{\alpha/2} s_b$	$CI_b = b \pm t_{\alpha/2} s_a$

(13.16)

수식을 적용하게 되면 범위는

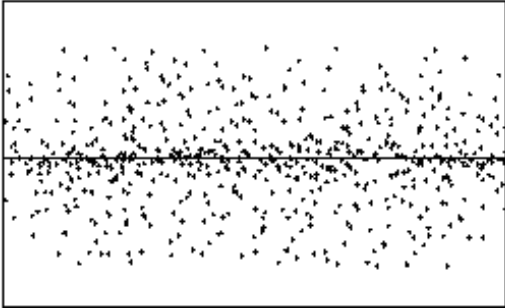
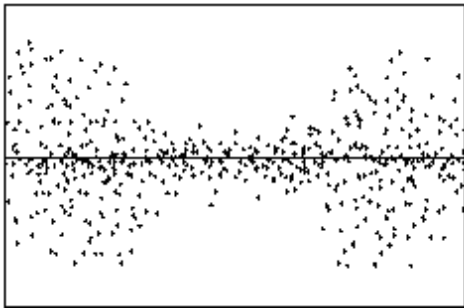
Slope	Intercept
$9.6434 \sim 17.2718$	$-69.0352 \sim 34.8894$

6. Regression 분석을 위한 가정과 주의 할점

가정들

- 1) dependent variable 과 independent variable 은 linear 관계에 있어야 합니다.
- 2) Residual 값들의 분포는 패턴이 없어야 합니다. 즉 residual 은 error 값들이기 때문에 이 값은 random 으로 분포가 되어 있어야 하지 어느 특정 형태를 갖추고 있으면 안됩니다.
- 3) Homoscedasticity. 각각의 independent variable 들에 해당하는 dependent variable 의 분포의 variation 은 같아야 합니다.

각 independent variable 에 대한 residue 값들의 분포가 다음과 같은 경우 (a)의 경우는 regression 분석을 위한 가정에 맞지만 (b)그렇지 못합니다.

 <p style="text-align: center;">(a)</p>	 <p style="text-align: center;">(b)</p>
<p>모든 independent variable 에 residue 의 random 으로 분포가 되어 있고, 이는 각 independent variable 에 해당되는 실제값 dependent variable 들의 분포가 같다는 것을 의미합니다.</p>	<p>Residue 값의 분포를 보면 양쪽으로는 넓게 퍼져있지만 가운데는 집중되는 형태를 갖추고 있고 이는 각 independent variable 에 해당하는 실제 값 dependent variable 의 분포가 다를 수 있습니다.</p>

주의사항

- 1) 예측 되는 값이 실제 범위를 넘을 수 있습니다. 예를 들어 성적과 같이 0 ~ 100 점 사이에 값이 있는데 예측되는 값이 100을 넘을 수 도 있습니다.
- 2) 두 변수의 관계가 correlation 으로 측정을 하여 통계적으로 관계를 발견했다 하지만 실제로 independent variable 이 dependent variable 을 변화 시킨다는 것을 증명하지 못합니다. 예를 들어 왼쪽, 오른쪽 발 크기를 갖고 한쪽을 independent variable 로 설정하여 보면 거의 1 에 가까운 r 값이 나타나지만 한쪽 발이 다른쪽 발 크기를 변화시키지는 못합니다.