CHAPTER 10

회귀 분석

10.1. 회귀 분석 개념

변수간의 관계 분석에는 교차분석, 분산분석, 회귀분석(상관분석 포함)이 있다. 어떤 분석을 사용할 것인가는 변수의 특성(분류형인가 측정형인가)에 의해 결정된다. 이를 표로 정리하면 다음과 같다.

종속변수 / 설명변수	분류형(범주형)	측정형(연속형)
분류형(범주형)	교차분석 Log-linear(설명변수 2 개 이상)	(다원) 분산분석
측정형(연속형)	Logistic regression	회귀분석, 상관분석

두 측정형 변수들의 관계를 살펴보는 방법은 상관 분석(correlation analysis)과 회귀 분석(regression analysis)으로 나뉘는데 상관 분석은 두 변수간의 선형(linear) 관계 정도를 알아보는 것이고 회귀분석은 두 변수간의 인과 관계(casual relationship)가 있는지(주로 선형 인과 관계) 알아보는 분석 방법이다.

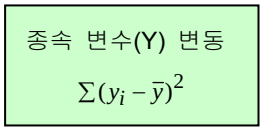
예를 들면, 키와 몸무게의 상관 관계, 부모와 자녀 IQ 간의 상관 관계를 보는 경우 상관 분석, “키가 큰 사람은 몸무게가 많이 나가는가?” “부모의 IQ 는 자식의 IQ 에 영향을 미치는가?” 등을 알아보려면 회귀분석을 시행해야 한다. 회귀분석에서 원인이 되는 변수를 독립변

수 혹은 설명 변수라 하고 (키, 부모 IQ) 결과로 나타나는 변수를 반응 변수 혹은 종속 변수라 한다. 회귀 모형(종속변수와 설명 변수간 선형 관계) 설정은 학문적 이론이나 자신의 연구 과제 혹은 경험에 근거하게 된다. 이 경우 그 모형 자체가 타당성을 가져야 한다. 예를 들면 맥주 소비량에 영향을 주는 요인(원인)을 찾아내려고 한다. 이에 맥주 소비량을 종속 변수로 기독교인 수를 설명 변수로 하여 회귀 분석을 실시하면 매우 유의 하다는 결론에 도달할 것이다. 그렇다고 기독교인들이 맥주를 많이 소비하여 이런 현상이 생겼다고 말할 수 있는가? 회귀 분석에서 인과 관계 설정은 이론이나 현실적인 타당성에 근거해야 한다.

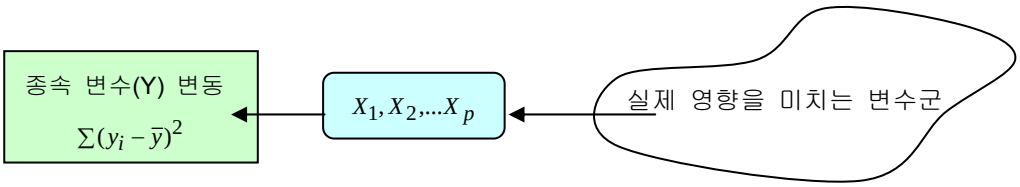
10.1.1. 개요

회귀 분석은 변수들간의 인과 관계를(종속 변수, 설명변수) 설정하여 ①변수들간의 함수 형태 ②영향을 미치는 변수 ③종속 변수에 대한 예측 값을 얻는 분석 방법이다. 이런 과정에서 다음과 같은 문제에 봉착하게 된다.

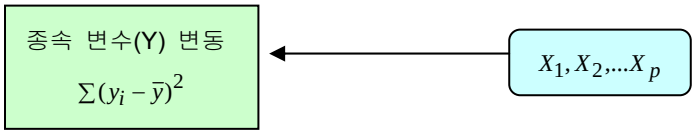
(1)종속 변수의 무엇을 설명할 것인가? ➔ 종속 변수가 가진 정보를 어떻게 표현 할 것인가? 이 경우 종속 변수가 가진 정보를 변수의 변동(분산)으로 생각했다. 종속 변수의 변동은 분산 분석(ANOVA)



(2)어떤 설명 변수(독립 변수)가 종속 변수에 영향을 미칠까? 변수의 선택은 이론이나 경험에 의해 분석자가 선택하게 된다.



(3)설명 변수와 종속 변수는 어떤 함수 관계를 갖는가? 실제 함수 f 의 형태는 알 수 없거나 이론 모형은 복잡하다(nonlinear: 비선형 회귀 모형) 그리하여 함수 관계를 단순화 하여 다루기 쉽고 해석이 용이한 선형 함수를 선택하게 된다.



$$Y = f(X_1, X_2, ... X_p)?$$

그리하여 다음의 선형 함수 관계를 설정하여 1)설명 변수는 종속 변수에 영향을 미치나? 2) 미친다면 어떻게 영향을 미치나? (회귀 계수 $\alpha, \beta_1, \beta_2, ... \beta_p$) 3)설명 변수 값에 따라 종속 변수 값을 예측한다.

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + e_i$$

통계적 모형은 과학적 진실이기 보다는 사실의 대표적 모형이다. 그러므로 설명 변수에 의해 설명되지 못하는 부분에 대해 오차항(e_i)으로 보고 $iid \sim N(0, \sigma^2)$ 을 가정한다.

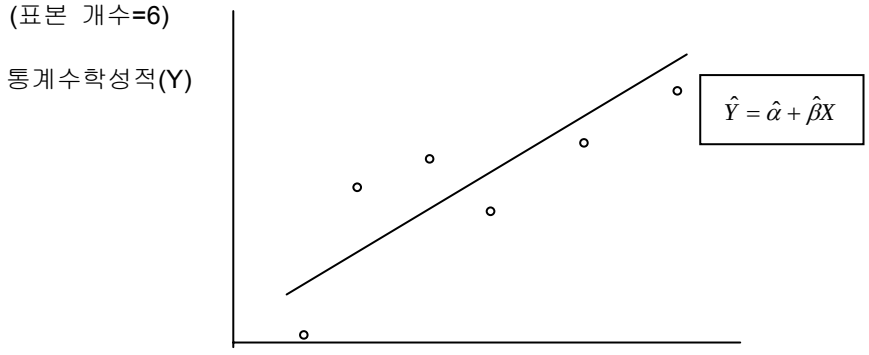
데이터 $(y_i, x_{1i}, x_{2i}, ..., x_{ni})$, $i = 1,2,..., p$ ➔ 모형 추정 $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_p X_{pi}$

종속 변수(Y) 변동	=	모형이 설명하는 변동	+	모형이 설명 못하는 변동
$\Sigma(y_i - \bar{y})^2$		$\Sigma(\hat{y}_i - \bar{y})^2$		$\Sigma(y_i - \hat{y}_i)^2$

분산 분석에서는 요인(인자 factor)이 설명하는 변동은 $\sum\sum(\bar{y}_i - \bar{y})^2$ 으로 계산되었다. 즉 $\hat{y}_i$ 대신 $\bar{y}_i$ 가 들어간 것 밖에는 달라진 점이 없다.

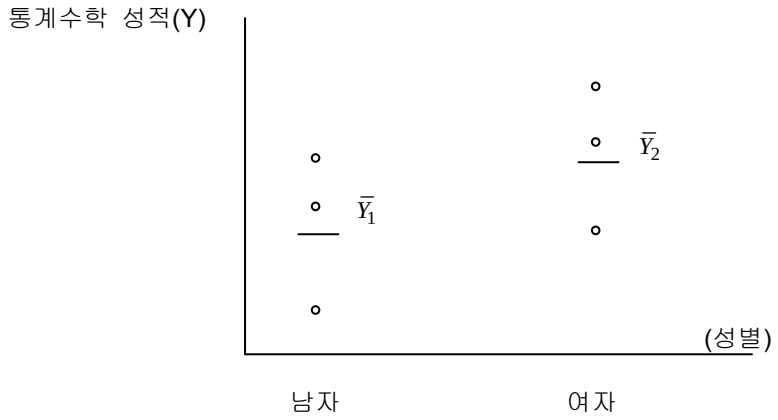
(예제)통계 수학 성적(Y)에 영향을 미치는 요인은? 수능에서 수학 성적(X)을 생각해 보자.

(표본 개수=6)



직선을 어떻게 그을 것인가? 어떤 직선이 데이터 값에 가장 근접한 직선일까? 통계학에서는 이에 대한 방법으로 OLS(Ordinary Least Square 최소 자승법) 방법을 발견하였다.

만약 통계 수학 성적에 영향을 미치는 요인으로 성별을 생각하였다고 하자. 각 집단(요인의 수준) 평균을 이용하여 성별이 통계 수학 성적에 미치는 영향에 대한 유의성을 검정하게 된다.



10.1.2. 기원

회귀(regress)의 사전적 의미는 “go back to an earlier and worse condition” (옛날 상태로 돌아감)을 의미하게 되는데 이런 용어를 사용하게 된 것은 영국의 유전학자 Francis Galton(1822-1911)의 연구에 기인한다. Galton 은 (처음에는 sweat pea) 부모의 키와 자녀의 키 사이 관계를 연구하면서 928 명의 성인 자녀 키(여자는 키에 1.08 배)와 부모 키(아버지 키와 어머니 키의 평균)를 조사하여 다음 표를 얻게 되었다.

성인 자녀	부모 키											
	<64.0	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	>73.0	Totals
>73.7	—	—	—	—	—	—	5	3	2	4	—	14
73.2	—	—	—	—	—	3	4	3	2	2	3	17
72.2	—	—	1	—	4	4	11	4	9	7	1	41
71.2	—	—	2	—	11	18	20	7	4	2	—	64
70.2	—	—	5	4	19	21	25	14	10	1	—	99
69.2	1	2	7	13	38	48	33	18	5	2	—	167
68.2	1	—	7	14	28	34	20	12	3	1	—	120
67.2	2	5	11	17	38	31	27	3	4	—	—	138
66.2	2	5	11	17	36	25	17	1	3	—	—	117
65.2	1	1	7	2	15	16	4	1	1	—	—	48
64.2	4	4	5	5	14	11	16	—	—	—	—	59
63.2	2	4	9	3	5	7	1	1	—	—	—	32
62.2	—	1	—	3	3	—	—	—	—	—	—	7
<61.7	1	1	1	—	—	1	—	1	—	—	—	5
Totals	14	23	66	78	211	219	183	68	43	19	4	928

부모 키와 자녀의 키 간에는 직선 관계가 있음을 발견하였고 또한 자녀의 키는 평균 키를 중심으로 회귀(무한정 커지거나 작아지지는 않는다.)하려는 경향이 있음을 언급하였다. Galton 은 경험적 연구를 통하여 회귀 분석 개념을 도출하였다면 Karl Pearson(1903) 1078 명의 부자 키를 조사하여 아버지 키와 아들 키 간에 다음 선형 함수 관계가 있음을 보였다.

$$Y(\text{자녀키}) = 33.73 + 0.516X(\text{부모키})$$

10.1.3. 모형 및 가정

(1)모형

$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i$ for $i = 1,2,...,n$ (관측치 수) $e_i \underset{\sim}{iid} N(0, \sigma^2)$

종속변수 관측치=(설명변수에 의해 설명되는 부분 $a + bx_i$)+설명되지 않는 부분(e_i)

(cf) iid: independently(독립이고) and identically distributed(동일한 분포)

▪종속변수와 설명변수의 관계는 직선이다.

(2)가정

오차항에 대한 3 가지 가정이 있다.

①오차항은 독립이다. ➔ 독립성 가정

▪시계열 자료(자료가 시간 순서를 갖는 경우)에서만 독립성 가정을 판단한다. 시계열 자료가 아닌 자료를 횡단 자료(cross-sectional data)라 한다.

▪오차항이 독립이 성립하지 않으면 설명되지 않는 부분에 일정한 패턴(관계)이 있다는 것이다. 오차항이 독립이 아닌 경우 가장 간편한 해결 방법은 1 차 차분(1st difference: $y_t - y_{t-1}$) 이용하는 것이다.

②정규분포를 따른다. ➔ 정규성 가정

▪오차항이 정규분포이면 y_i도 정규분포를 따른다. 회귀 변동(SSR), 오차 변동은 (SSE) y_i의 제곱 형태이므로 χ^2 분포를 따른다. $\frac{SSR/df_1}{SSE/df_2}$ 은 F-분포를 따른다. 그러므로 모형 검정을 위해 F-검정을 할 수 있다.

▪정규성이 성립하지 않으면 변수 변환(가장 자주 사용되는 것이 Log 변환)인데 정규성 문제는 추정과 검정에 큰 영향을 미치지는 않는다.

③분산이 같다. ➔ 등분산(σ^2) 가정

▪분산이 다르면 각 관측치(y_i)의 값들이 모형에서 벗어나는 것이 모형의 부적합성 뿐 아니라 다른 요인이 있는 것이므로 반드시 해결해야 한다.

▪해결 방법으로는 WLS(Weight Least Square) 방법이 있다.

오차항에 대한 3 가지 가정이 성립하는지에 대해 분석하는 것을 잔차 분석 (residual analysis)이라 한다. 설문 분석에서 회귀분석을 이용하는 경우 이런 가정을 검정하지는 않는다. 그리고 일반적으로 리커드 척도 문항은 정규 분포를 따르고 등분산 가정을 만족한다. 왜냐하면 사람들의 응답 분포는 3 점을 중심으로 좌우 대칭이고 각 응답 점수의 분산은 동일할 가능성이 높다. 그러나 이산형 자료인 경우에는 정규성 검정, 등분산성 검정이 제대로 되지 않을 가능성이 많다.

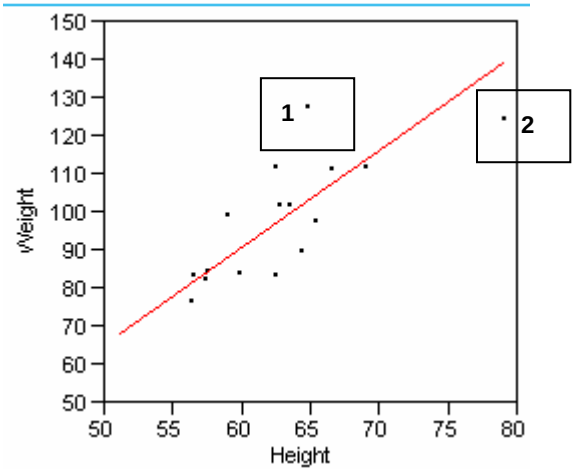
설문 조사에서는 회귀 분석은 리커드 척도 문항 간 관계가 존재하는가? 어떤 리커드 척도 문항의 영향이 가장 큰가? 를 보기만 하면 된다.

10.1.4. 산점도 (scatter plot)

산점도는 두 변수간의 (함수) 관계를 나타내는 이차원 그래프로 종속 변수는 Y 축, 설명 변수는 X 축으로 (인과 관계가 존재할 때) 하여 관측치 쌍을 그린다. 산점도는 두 변수 (X, Y) 간의 함수 관계를 쉽게 파악할 수 있으므로 일변량 분석의 시작이 Stem and Leaf plot 과 Box-whisker plot 이라면 다변량(이변량) 분석의 시작은 산점도이다.

(1)종속변수(y)와 설명변수(x1, x2, x3, …) 간의 선형 관계가 존재하는지 미리 알 수 있다.

(2)설명 변수 간 상관 관계가 높아 발생하는 다중 공선성 (multicollinearity) 문제를 미리 파악할 수 있다.



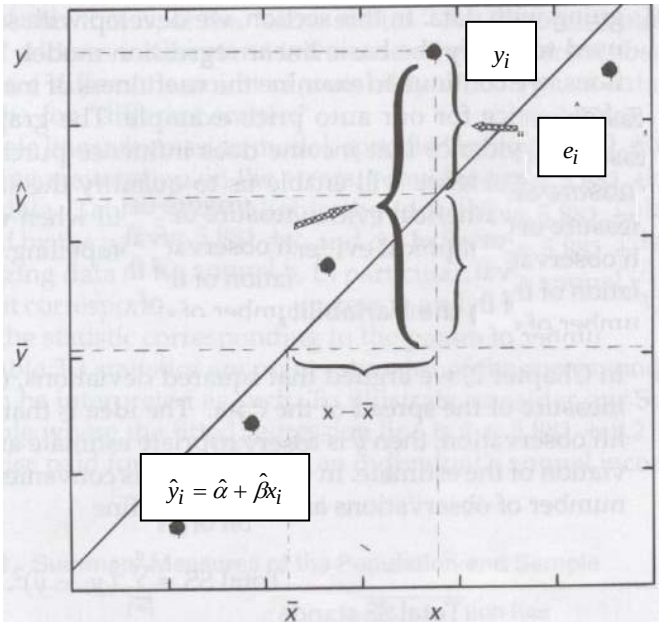
산점도를 살펴 되면(그림 2, 물론 붉은 선은 추정 회귀선) 두 가지 특이한 관측치가 발견된다. 1 관측치는 같은 키의 다른 사람에 비해 몸무게가 많이 나가는 것을 알 수 있다.[이상치가 아닐까] 2 관측치는 키와 몸무게의 관계가 선형이 아니라 이차식 관계가 성립하지 않을까 하는 의심을 갖게 한다. 이처럼 산점도는 (1)두 변수의 함수 관계의 형태 물론 (2)특이한 관측치가 존재하는지를 알 수 있으므로 회귀 분석의 시작이다. 설명 변수가 2 개 이상이 경우에는 scatter plot matrix 를 그리면 된다.

설문 분석에서 산점도는 그리지 않는다. 종속 변수와 설명 변수가 가질 수 잇는 값이 연속형이라기 보다는 이산형(1, 2, 3, 4, 5)이므로 산점도의 의미는 크게 축소된다. 그러므로 설문 분석에서 상관 계수나 회귀 분석이 이용되려면 리커드 척도 각 문항에 대한 분석보다는 요인 분석 결과 묶여진 그룹을 하나의 변수(강의실 만족도, 정보 시설 만족도)로 사용하여 회귀 분석을 실시하는 것이 바람직하다. 묶을 수 없다면 할 수 없지만...

10.1.5. 회귀 계수 추론

회귀 모형을 추정한다는 것은 수집된 데이터(산점도)에 가장 적절한 회귀 직선을 구하는 것이다. 방법으로는 **OLS(Ordinary Least Square: 최소 자승법)**과 **MLE(MLE: Maximum Likelihood Estimator: 최대 우도 추정법, 최우 추정법)** 방법이 있다.

각 관측치에 가장 적합한 회귀 직선은 회귀 직선과 관측치의 벗어난 정도(오차: e_i)가 가장 적은 직선일 것이다. 그런데 $\sum_{i=1}^{n} e_i = 0$ 이므로 $\sum_{i=1}^{n} e_i^2$ (절대값 대신 제곱하는 이유는 (1)다루기 쉽고 (2)멀리 떨어질수록 더 큰 페널티를 부여)을 최소화 하는 α, β 를 추정하는 방법을 최소 자승법(OLS)라 한다.



$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$ 을 최소화 하는 추정치 $\hat{\alpha}, \hat{\beta}$ 를 OLS 추정치라 한다. OLS 추정치를 구하려면 Q 를 α, β 에 대해 각각 편미분(partial derivative) 하게 된다. OLS 추정 방법에는 오차항에 대한 가정이 전혀 필요하지 않다.

단순 회귀 모형의 경우 $\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ 이 OLS(최소 자승법) 추정치이다. 적합 된 회귀 모형(데이터에 가장 적절한 선형 회귀식, 추정 회귀 모형)은 다음과 같다.

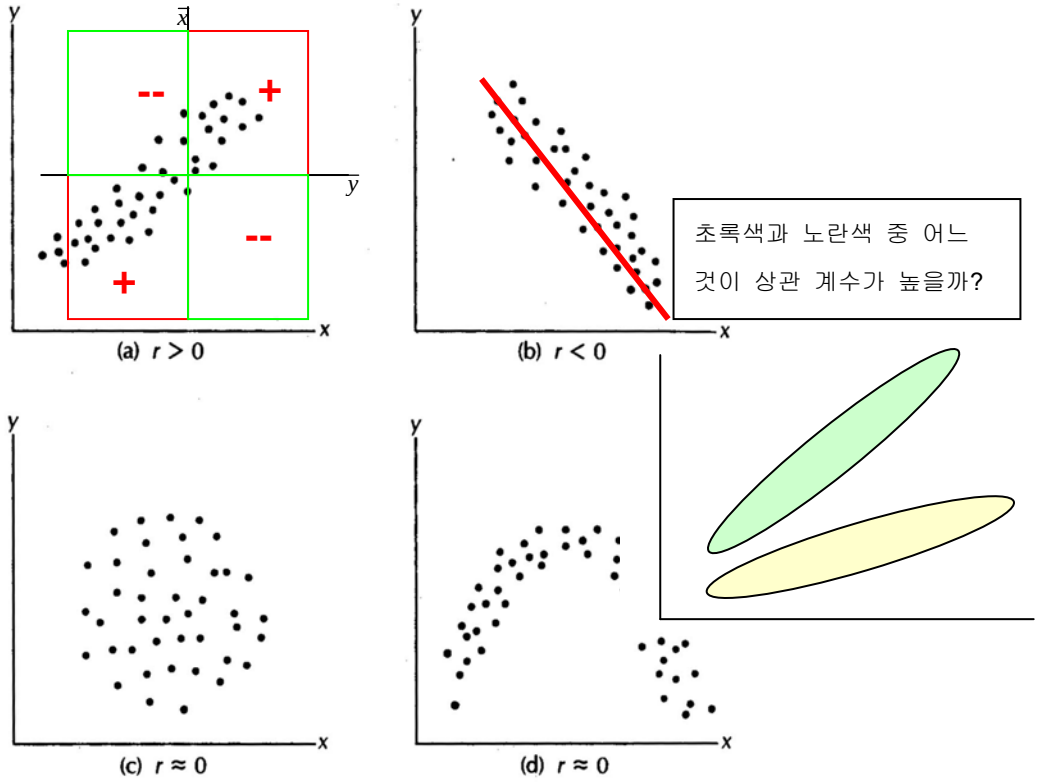
- $\hat{y} = \hat{\alpha} + \hat{\beta}x_i$ ---(적합 된 회귀식)
- 회귀 계수에 대한 추론:

$$\frac{\hat{\beta} - \beta}{s(\hat{\beta})} \sim t(n-2), \quad s^2(\hat{\beta}) = \frac{MSE}{\sum (X_i - \bar{X})^2}, \quad MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n - p - 1}$$

10.1.6. 상관 분석과의 관계

상관 분석은 두 변수 간의 선형(직선) 관계가 존재하는 알아보는 방법이다. 회귀 분석과 유사하지만 인과 관계에 대한 분석은 아니다. 상관 계수 구하는 식은 $r = \frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$ 이다.

그러므로 $Sxx = \sum(x_i - \bar{x})^2, Syy = \sum(y_i - \bar{y})^2$ 이라 하면 회귀 모형에서 기울기 회귀 계수 추정치와 상관 계수는 관계는 $\hat{\beta} = \sqrt{\frac{Syy}{Sxx}}r$ 이다. 이러므로 기울기의 부호와 상관 계수의 부호는 같다.



상관계수가 1 에 가까우면 한 변수가 증가(감소)하면 다른 변수 값도 증가(감소)하고 –1 에 가까우면 한 변수 값이 증가(감소)하면 다른 변수의 값이 감소(증가)한다는 것을 의미하며, 1(양의)과 –1(음의)에 가까울수록 상관관계는 높다고 한다. 관측치에 대한 타원의 폭이 좁을수록 상관 관계가 높고 회귀 모형의 직선 적합도는 높아진다.

다음은 상관 계수의 유의성 검정 방법을 정리한 것이다.

H_O: 모집단 상관계수=ρ_0 =0	H_O: 모집단 상관계수=$\rho_0 \neq 0$
$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} \sim T(n-2)$ 단순 회귀 모형에서 상관 계수 검정 결과나 회귀 계수 검정 결과는 일치 한다. (3 장 참고)	$z^* = 0.5 \ln \frac{1+r}{1-r} \sim N(0.5 \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3})$ $h1 = 0.5 \ln \frac{1+r}{1-r} - 1.96/\sqrt{n-3})$ $h2 = 0.5 \ln \frac{1+r}{1-r} + 1.96/\sqrt{n-3})$ 모집단의95%신뢰구간 $(\frac{e^{h1}-e^{-h1}}{e^{h1}+e^{-h1}}, \frac{e^{h2}-e^{-h2}}{e^{h2}+e^{-h2}})$
두 상관계수 차이 검정 (독립인 경우) $z(x) = 0.5 \ln \frac{1+r_x}{1-r_x}, z(y) = 0.5 \ln \frac{1+r_y}{1-r_y}$ $z = \frac{z(x)-z(y)}{\sqrt{1/(n_x-3)+1/(n_y-3)}} \sim N(0,1)$	

10.1.7. 분산 분석

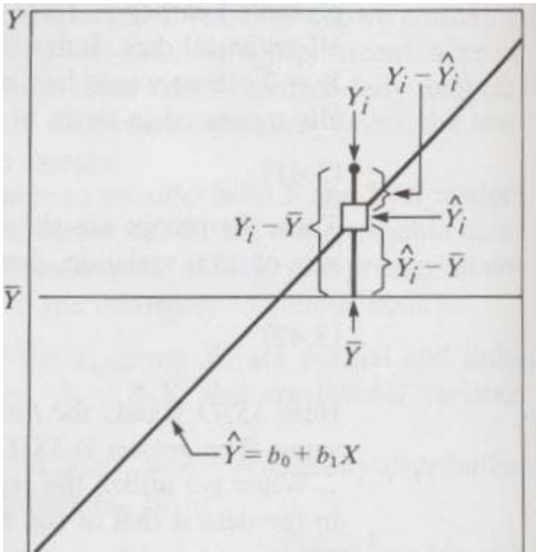
(1)분산분석 접근

회귀 분석을 분산 분석적 측면에서 다루는 것은 단순 회귀에서는 새로운 것이 없으나(단순 회귀 모형에서는 회귀 계수에 대한 t-검정은 분산 분석의 F-검정과 동일) 보다 복잡한 회귀 모형을 다루는데 도움을 얻을 것이다. 분산분석 접근은 종속 변수 Y 에 관련된 총변동과 자유도 분할에 근거한다. 총변동(SSTO, SST)은 종속 변수의 관측치와 평균의 편차(deviation) $(Y_i - \bar{Y})$ 제곱합을 의미하며 이는 종속 변수가 가진 정보이다.

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

회귀 모형에서 데이터에 포함된 불확실성 (uncertainty)은 적합 회귀선(fitted regression line, 추정 회귀식)으로부터 관측치가 얼마나 벗어나 있나를 의미하며 $(Y_i - \hat{Y}_i)$ 이것들의 제곱합을 오차 변동 (Error Sum of Squares, SSE) 혹은 오차 제곱합이라 하며 적합 회귀식에 의해 설명되지 않는 변동에 해당된다.

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$



두 변동의 차이를 회귀 변동(Regression Sum of Squares, SSR) 혹은 모형 변동(Model SS)이라 하며 적합 된 회귀식이 데이터의 관계를 얼마나 잘 설명하는지 나타낸다.

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 \quad \rightarrow \quad SST = SSE + SSR$$

▣자유도 분할

총 변동의 자유도(관측치 중 자유로운 개수, 관측치 하나 하나는 독립적이고 정보를 갖고 있다)는 (n-1)이고 SSE 의 자유도는 (n-2)이다. 왜냐하면 총변동의 경우는 $\bar{Y}$ 가 하나 추정되었고 SSE 의 경우에는 $\hat{\alpha}, \hat{\beta}$ 가 두 개 추정되었기 때문이다. SSR 의 자유도는 SST 자유도로부터 SSE 자유도를 뺀 값으로 1 이다.

▣평균

변동 합(제곱합)을 자유도로 나눈 값을 평균 변동이라 한다.

$$MSR = \frac{SSR}{1} \text{ (Mean Sum of squares of Regression 회귀 평균 변동)}$$

$$MSE = \frac{SSE}{(n-2)} \text{ (Mean Sum of squares of Error 오차 평균 변동)}$$

▣EMS (Expected Mean Squares 기대 평균 변동)

$$E(MSE) = \sigma^2, \quad E(MSR) = \sigma^2 + \beta^2 \sum (X_i - \overline{X})^2$$

이다.

그러므로 β가 0 이 아니면 회귀 모형은 유의해진다. 즉 단순 회귀에서 $H_0 : \beta = 0$을 검정하는 것과 아래 F-검정을 하는 것과는 동일하다.

▣F-검정 (F-test)

회귀 모형에서 설명 변수의 유의성 검정($H_0 : \beta = 0$)을 위하여 다음 검정 통계량을 생각해 보자.

$$F^* = \frac{MSR}{MSE}$$

$H_0 : \beta = 0$가 채택되면 기대 평균 변동에서 알 수 있듯이 $F^* = 1$ 이고 $H_0 : \beta = 0$가 기각되면 F^*는 1 보다 커지게 된다. 귀무가설 $H_0 : \beta = 0$ 하에서는 SSR/σ^2 과 SSE/σ^2 가 서로 독립이므로

$$F^* = \frac{MSR}{MSE} = \frac{\frac{SSR}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(n-2)} = \frac{\chi^2(1)/1}{\chi^2(n-2)/(n-2)} \sim F(1, n-2)$$

이다. 그러므로 $F^* \le F(1-\alpha; 1, n-2)$ 이면 귀무가설 $H_0 : \beta = 0$ (설명 변수는 종속 변수에 영향을 미치지 않는다) 채택하고 $F^* > F(1-\alpha; 1, n-2)$ 이면 귀무가설을 기각한다.

▣t-검정과의 관계

$SSR = \beta^2 \sum (X_i - \bar{X})^2$ 이고 $s^2(\hat{\beta}) = \frac{MSE}{\sum (X_i - \bar{X})^2}$ 이므로 다음이 성립하므로 분산 분석의 F-검정과 기울기 회귀 계수에 대한 t-검정은 동일하다.

$$F^*(1, n-2) = \frac{\beta^2 \sum (X_i - \bar{X})^2}{MSE} = \frac{\hat{\beta}^2}{s^2(\hat{\beta})} = (\frac{\hat{\beta}}{s(\hat{\beta})})^2 = t^2(n-2)$$

(예제)다음은 자동차 구매 가격 데이터의 회귀 분석 결과(Price=Income)이다.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1186892153	1186892153	196.26	<.0001
Error	60	362851718	6047529		
Corrected Total	61	1549743871			

$F = t^2 \Rightarrow (14.01)^2 = 196.26$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	5866.33390	749.81441	7.82	<.0001	4366.48178	7366.18603
INCOME	1	0.21132	0.01508	14.01	<.0001	0.18115	0.24150

▣분산 분석표 (Analysis of Variance Table): 귀무가설 $H_0 : \beta = 0$ 를 검정

변동 (source)	SS(자승합)	df(자유도)	MS(평균 자승합)	EMS(기대 평균 자승합)
Regression (모형, 회귀)	$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	1	$MSR = SSR/1$	$E(MSR) = \sigma^2 + \beta^2 \sum (X_i - \bar{X})^2$
Error (오차)	$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	n-2	$MSE = SSE/(n-2)$	$E(MSE) = \sigma^2$
Total (총변동)	$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	n-1		$F = \frac{MSR}{MSE} \sim F(1, n-1)$

▣결정 계수

회귀 계수 추정과 검정, 종속 변수 관측치에 대한 예측치($\hat{Y}_{new}$), 평균 예측치($\hat{E(Y_0)}$)에 대해 살펴 보았으나 두 변수 간의 선형 관계 정도를 나타낸 통계량은 없었다. 이에 다음과 같이 결정 계수를 (Coefficient of Determination) 정의한다. $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ 결정 계수는 두 변수 간의 선형 관계 정도가 높으면 (관측치들이 직선 가까이에 모여 있다는 것을 의미) 결정 계수는 1 에 가까워진다. 특히 단순 회귀 모형에서는 상관계수는 $r = \pm\sqrt{R^2}$ 이 성립한다. (참고: $\beta = \sqrt{\frac{S_{yy}}{S_{xx}}} r$) 결정 계수는 단순히 선형 관계 정도를 나타내는 수치일 뿐 검정할 수 있는 검정 통계량이 존재하지 않아 단지 지표로 사용될 뿐이다. 특히 설명 변수가 이산형(설문지 Likert 척도)인 경우 매우 낮아지는 경향이 있고 관측치가 많아지면 커지는 경향이 있어 선형 관계 정도를 나타내는 좋은 지표는 아니다.

10.2. 설문 분석에 회귀 분석 적용

설문 분석에서 상관 분석이나 회귀 분석이 가능한 문항은 리커드 척도 문항이다. 상관 분석은 두 리커드 척도 문항간 상관 관계 정도를 알아보는 것이고 회귀 분석은 종속 변수로 설정된 리커드 척도 문항에 설명 변수로 설정된 리커드 척도 문항이 영향을 미치는지 미친다면 어떤 문항이 가장 많이 영향을 미치는지 알아보는 것이다.

리커드 척도의 값은 1, 2, 3, 4, 5이므로(물론 요인 분석 결과 2개 이상의 리커드 척도 문항을 묶는다면 문제는 다소 해결되겠지만) 상관 계수 값은 일반적으로 0.5를 넘지 못하고 회귀 분석의 결정 계수 값은 0.3(당황하지 말자. 몰래 고치지도 말자. 당연한 것이다)을 제대로 넘지 못한다. 상관 분석의 경우 비모수 분석 방법이 있으니 회귀 분석은 계산이 복잡하고 이에 대한 통계 소프트웨어가 지원되지 않는다. 그러므로 마땅한 해결책이 없어 많은 사회 조사 연구에서는 이런 단점이 있음에도 불구하고 리커드 척도 문항과 관계 분석으로 활용되고 있다.

상관 분석은 리커드 척도 문항과 선형관계가 있는지에 관심이 있으므로 상관 계수와 유의확률만 계산하면 된다. 상관 관계가 유의하고 부호가 +이면 한 문항의 만족도가 높아지면 다른 문항의 만족도도 높아짐을 의미하며 –이면 다른 문항의 만족도는 낮아진다는 것을 의

미한다.

회귀 분석 절차는 산점도 그리기 ▶회귀 모형 유의성 검정(F-검정) ▶유의한 설명 변수 선택 ▶다중공선성과 영향치 진단 ▶잔차 분석 ▶최종 회귀 모형 제시 및 해석 순으로 진행되나 설문 조사 분석에서는 회귀 모형 유의성 검정(F-검정) ▶유의한 설명 변수 선택 ▶최종 회귀 모형 제시 및 해석만 실시하면 된다.

예제 설문에서 경상대 시설에 대한 만족도(Q4-Q12) 간 상관 관계가 존재하는가를 분석하고 경상대 시설 만족도(Q4-Q12) 중 경상대 시설 전체 만족도(Q13)에 영향을 미치는 시설은 무엇이며 어느 시설 만족도 영향이 가장 큰지 알아보고자 회귀 분석을 실시해 보자.

10.3. 통계 소프트웨어 사용

10.3.1. SAS

▣경상대 시설 만족도 간 상관 관계가 존재하는가?

NOSIMPLE 옵션은 각 변수의 기초 통계량(평균, 분산)을 출력하지 말라는 옵션이다.

```
PROC CORR DATA=SURVEY NOSIMPLE;
    VAR Q4-Q12;
RUN;
```

유의 확률이 0.05 보다 작으면 귀무가설(두 변수간 상관 관계는 없다. 모집단 상관 계수 $\rho = 0$)이 기각 되어 상관 관계가 존재한다고 한다. 만약 유의하고 부호가 양이면 양의 상관 관계, 유의하고 부호가 음이면 음의 상관 관계가 존재한다.

	Q4	Q5	Q6	Q7	Q8	Q9
Q4	1.00000 129	0.48442 <.0001 129	0.35678 <.0001 129	0.47492 <.0001 129	0.31608 0.0003 129	0.43002 <.0001 128
Q5	0.48442 <.0001 129	1.00000 130	0.33808 <.0001 130	0.46952 <.0001 130	0.34472 <.0001 130	0.46121 <.0001 129

상관 계수
유의확률
n: 자료 수

변수간 상관 관계가 모두 유의하고 양의 부호를 갖는다. Q4(건물 공간 만족도) 만족도가 높은 응답자는 Q5(휴식 공간 만족도) 만족도도 높게(낮게) 응답한다. 만약 음의 부호이면 OO 만족도가 높은 응답자는 △ 만족도는 낮게 응답한다. 설문 조사에서 상관 계수가 음

의 의미는 두 문항의 척도가 반대임을 의미한다. 부정적인 성향을 묻는 문항과 긍정적인 성향을 묻는 문항이 섞여 있는 경우 이런 경우가 발생한다. 부정적인 견해를 묻는 문항을 부정 문항, 부적 문항이라고 한다.

▣부정 문항이 들어 있는 경우의 예

①Q5. 당신은 외모에 만족합니까? (5 점 척도)

②Q6. 당신은 혼자 있는 것이 편하십니까? (5 점 척도)

③Q7. 당신은 능력이 있다고 생각합니까? (5 점 척도)

④Q8. 당신은 친구 관계가 원만하다고 생각합니까? (5 점 척도)

	Q5	Q6	Q7	Q8
Q5	1.00000	-0.33808 <.0001	0.46952 <.0001	0.34472 <.0001
Q6	-0.33808 <.0001	1.00000	-0.44755 <.0001	-0.25000 0.0041
Q7	0.46952 <.0001	-0.44755 <.0001	1.00000	0.30147 0.0005
Q8	0.34472 <.0001	-0.25000 0.0041	0.30147 0.0005	1.00000

유사 개념에 대한 만족도를 측정한 4 개 문항 Q5-Q8 을 묶을 수 있는지 알아 보려면 요인 분석(factor analysis)을 실시해 보자.

```
PROC FACTOR DATA=ZZZ ROTATE=VARIMAX;
    VAR Q5-Q8;
RUN;
```

Factor Pattern

	Factor1
Q5	0.75881
Q6	-0.70522
Q7	0.79112
Q8	0.62252

▶Q5, Q6, Q7를 하나의 문항으로 묶는 것이 적당하다.

Q6 번 문항은 다른 문항과 개념이지만 반대 점수로 측정되고 있다. 즉 Q6=1(매우 만족), ..., 5(매우 불만족)이다. 그러므로 이를 다른 문항과 합쳐(그룹화) 사용할 때는 다음과 같이 한다.

```
DATA ZZZ;
    set ZZZ;
    q6=6-q6;
    GROUP1=MEAN(Q5, Q6, Q7);
RUN;
```

Q5, Q6, Q7 문항의 내적 일치도(크론바흐 알파, 신뢰도 계수)를 구하면 다음과 같다.

```
PROC CORR DATA=ZZZ NOSIMPLE NOCORR ALPHA;
    VAR Q5 Q6 Q7;
RUN;
```

Cronbach의 α계수

변수	α계수
원데이터	0.670726
표준화	0.683350

만약 부정 문항이 있다면 신뢰도 계수가 –가 나오는 경우가 발생한다. 이래와 같이 나오면 Q6가 부정 문항임을 인지한다. 그런데 이는 요인 분석에서 밝혀질 것이다.

변수를 제외했을때의 Cronbach 계수

	데이터 변수		표준화된 변수	
삭제한 변수	합계와의 상관 계수	α계수	합계와의 상관	α계수
Q5	0.097730	-1.61297	0.125043	-1.62027
Q6	-.447373	0.621031	-.458268	0.639011
Q7	0.105136	-.985364	0.019090	-1.02153

▣Spearman rank correlation coefficient (순위 상관 계수)

변수들의 상관 관계를 정도를 자료의 순위 값에 의하여 계산하는 방법으로 비모수 상관계수의 한 방법이다. 자료의 수가 적거나 치우침이 큰 경우 사용한다.

①가정: 두 변수에 대한 표본 관측치는 (x_i, y_i) 이고 각 변수는 크기 순으로 정렬이 가능하다.

②가설: 귀무가설: 두 변수 (x, y) 는 서로 독립이다. 대립가설: 독립이 아니다.

④검정통계량: $r_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$ where $\sum d_i^2 = \sum_{i=1}^{n}[R(x_i) - R(y_i)]^2$

$R(x_i)$ 는 x 변수의 i 번째 관측치의 순위이다. $R(y_i)$ 는 y 변수의 i 번째 관측치의 순위이다.

Large Sample approximation: $z = r_S\sqrt{n-1} \sim Normal(0,1)$

▣Kendall's Tau(τ)

비모수 방법으로 순서형 변수들의 상관 관계를 정도를 계산한다. 설문 조사에서 리커드 척도 문항 각각의 상관 계수를 구할 때는 Kendall 상관 계수를 이용하는 것을 권한다. 단 요인 분석에 의해 리커드 척도 문항을 묶으면 Kendall 상관 계수를 사용할 필요는 없다.

①가정: 두 변수에 대한 표본 관측치는 (x_i, y_i) 이다.

②가설: 귀무가설: 두 변수 (x, y) 는 서로 독립이다. 대립가설: 독립이 아니다.

③검정통계량:

$$T = \frac{\sum_{i<j} \mathrm{sgn}(x_i - x_j)\,\mathrm{sgn}(y_i - y_j)}{\sqrt{(T_0 - T_1)(T_0 - T_2)}}, where \begin{bmatrix} T_0 = n(n-1)/2 \\ T_1 = \sum t_i(t_i - 1)/2 \\ T_2 = \sum u_i(u_i - 1)/2 \end{bmatrix}$$

n = 쌍의 관측치 수, t_i = 주어진 순위에서 동일한(tied) X 관측치 수

u_i = 주어진 순위에서 동일한(tied) Y 관측치 수 $\quad \mathrm{sgn}(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z = 0 \\ -1 & \text{if } z < 0 \end{cases}$

```
PROC CORR DATA=SURVEY NOSIMPLE PEARSON SPEARMAN KENDALL;
    VAR Q4-Q12;
RUN;
```

스피어만 상관 계수
H0: Rho=0 검정에 대한 Prob > |r|
관측치 개수

	Q4	Q5	Q6	Q7	Q8	Q9
Q4	1.00000 129	0.46184 <.0001 129	0.35826 <.0001 129	0.48885 <.0001 129	0.31649 0.0003 129	0.39608 <.0001 128
Q5	0.46184 <.0001 129	1.00000 130	0.30484 0.0004 130	0.46593 <.0001 130	0.36784 <.0001 130	0.47006 <.0001 129

켄달 타우 b 상관 계수
H0: Rho=0 검정에 대한 Prob > |r|
관측치 개수

	Q4	Q5	Q6	Q7	Q8	Q9
Q4	1.00000 129	0.38103 <.0001 129	0.29627 <.0001 129	0.40490 <.0001 129	0.26672 0.0002 129	0.32985 <.0001 128
Q5	0.38103 <.0001 129	1.00000 130	0.25370 0.0004 130	0.39769 <.0001 130	0.30708 <.0001 130	0.39844 <.0001 129

리커드 척도 문항과 같은 순서형 변수의 경우 상관 계수 값은 일반적으로 Kendal<Spearman<Pearson 순서이다. 그러나 값의 차이와는 달리 유의 확률은 거의 차이가 없다.

▣경상대 시설 만족도(Q4-Q12) 중 경상대 전체 만족도(Q13)에 영향을 미치는 시설은?

Q4-Q12 시설 만족도 중 요인 분석 결과 묶인 것은 묶인 그룹을 변수로 이용하여 회귀분석을 실시한다.

```
DATA SURVEY0;
    SET SURVEY;
        LECTURE=MEAN(OF Q5-Q8);
        INFORMATION=MEAN(Q10, Q11);
RUN;
```

```
PROC REG DATA=SURVEY0;
    MODEL Q13=Q4 LECTURE INFORMATION Q9 Q12/SELECTION=BACKWARD SLS=0.1;
RUN;
```

SELECTION 옵션은 유의한 설명 변수를 선택하는 방법에 관한 것이다. 방법에는 STEPWISE, FORWARD, BACKWARD 가 있다. SLS 는 Significant Level for Stay (잔류 유의수준)으로 0.1 로 해 주자. 의미는 유의 수준 0.1 에서 유의한 설명 변수만을 선택하기 위함이다.

▣변수 선택(Variable selection)

1)Backward

①고려된 설명변수를 모두 삽입한 후 설명변수 중 가장 유의하지 않은 설명변수를 제외한다. 가장 유의하지 않다는 것은 F 값이 가장 작은(유의 확률이 가장 큰) 설명변수를 의미한다. 물론 제외되는 설명 변수의 유의 확률은 설정된 유의 수준(일반적으로 0.05 그러나 0.1 도 적당하다)보다 커야 한다.

②모든 설명변수가 유의할 때까지 ① 과정을 반복한다.

2)Forward

①고려된 설명변수 중 설명력이 가장 높고(F-값이 가장 크거나 유의 확률이 가장 적은) 유의한 변수를 선택한다.

②이미 선택된 설명변수의 설명 부분을 제외하고 설명력이 가장 크고 유의한 경우 설명 변수를 선택한다.

③유의한 설명변수가 없을 때까지 [2]을 반복한다.

3)Stepwise

Forward 방법과 유사하지만 한 번 선택된 설명 변수에 대해서는 유의성 검정을 다시 실시한다는 점이 다르다.

①고려된 설명변수 중 설명력이 가장 높고(F-값이 가장 크거나 유의 확률이 가장 적은) 유의한 변수를 선택한다.

②이미 선택된 설명변수의 설명 부분을 제외하고 설명력이 가장 크고 유의한 설명 변수를 선택한다.

③새로 선택된 변수의 설명 부분을 제외한 부분에 대해 이미 존재한 설명 변수의 유의성을 검정하여 유의하지 않으면 제외한다.

④순서 ②, ③을 반복한다.

4)설문 조사에서 설명 변수 선택 방법은 Backward 방법을 사용하면 된다. 다음은 Backward 방법에 의한 출력 결과의 일부이다.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	120.29472	30.07368	53.45	<.0001
Error	123	69.20528	0.56264		
Corrected Total	127	189.50000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.65265	0.30072	2.65005	4.71	0.0319
Q4	0.33689	0.07021	12.95354	23.02	<.0001
LECTURE	0.45454	0.11018	9.57582	17.02	<.0001
INFORMATION	0.27840	0.06085	11.77679	20.93	<.0001
Q12	0.17060	0.04663	7.52937	13.38	0.0004

Bounds on condition number: 1.6162, 21.891

All variables left in the model are significant at the 0.1000 level.

유의한 설명변수만 남고 유의하지 않은 Q9 변수(문항)는 제외되었다. 아래 요약 결과처럼 Q9 의 유의 확률은 0.1039 로 유의수준 0.1 에서 유의하지 않았다.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Q9	4	0.0079	0.6348	6.6849	2.68	0.1039

Q4(건물 안 공간 만족도), 강의실 만족도(Lecture), 정보 시설 만족도(information), Q12(화장실 시설)는 전반적인 만족도에 영향을 미친다. (아래 수식) 회귀 계수가 모두 양이므로 각 시설 만족도가 높으면 시설 전반에 관한 만족도가 높아진다. 그럼 어떤 시설 만족도(설명 변수)가 가장 영향이 큰가? 이에 대한 대답을 하려면 표준화 회귀 계수를 이용해야 한다.

$$Y(overall) = -0.65 + 0.34 * Q4 + 0.45 * Lecture + 0.27 * Information + 0.17 * Q12$$

▣영향이 가장 큰 시설은 어디인가?

```
PROC REG DATA=SURVEYO;
    MODEL Q13=Q4 LECTURE INFORMATION Q12/STB;
RUN;
```
유의한 변수만 넣었다.

STB(Standardized Beta)는 표준화 회귀 계수를 구하라는 옵션이다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-0.67738	0.29982	-2.26	0.0256	0
Q4	1	0.34557	0.06971	4.96	<.0001	0.34123
LECTURE	1	0.45978	0.11009	4.18	<.0001	0.25804
INFORMATION	1	0.26827	0.06006	4.47	<.0001	0.27440
Q12	1	0.17686	0.04624	3.82	0.0002	0.22875

Q4 만족도의 영향이 가장 크고 Q12 의 영향이 가장 적으므로 전반적인 시설 만족도를 많이 높이려면 건물 안 공간 만족도 높일 수 있는 방안을 우선적으로 강구해야 한다.

(8)STB(Standardized Beta Coefficient 표준화 회귀 계수)

각 설명 변수의 종속 변수에 대한 설명력은 회귀 계수에 의해 해석된다. 그러나 설명 변수의 측정 단위가 다르다면 설명 변수 한 단위당 종속 변수의 증가량으로 해석되는 회귀 계수의 값으로는 설명력을 측정할 수 없다.

설명 변수의 측정 단위가 다른 경우 설명 변수의 설명력을 비교하고자 하면 다음 방법에 의해 각 변수를 표준화한 후 이에 대한 회귀 모형을 추정하여 얻는 추정 회귀 계수를 표준화 회귀 계수라 한다. 표준화 회귀 계수는 각 설명 변수의 설명력을 비교할 때 사용된다.

①변수 표준화: $Y^* = \frac{Y - \bar{Y}}{s_Y}$ (종속 변수), $X_k^* = \frac{X_k - \bar{X}_k}{s_{X_k}}$, $k = 1,2,\ldots, p$ (설명 변수)

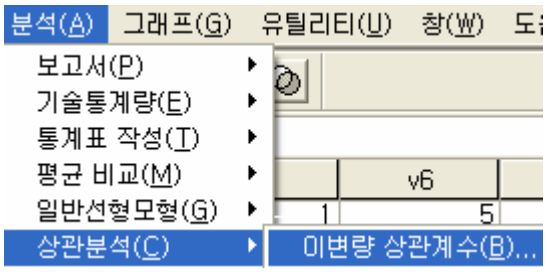
②표준화 회귀 모형: $Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_2^* + \cdots + \beta_p X_{pi}^* + e_i$ --- (*) (표준화 회귀 모형의 절편은 없다)

(*) 모형에 대한 OLS 추정치기 표준화 회귀 계수가 되고 이를 이용하여 측정 단위가 다른 설명 변수의 설명력을 비교하는데 사용된다.

10.3.2. SPSS 사용하기

(1)경상대 시설 만족도 간 상관 관계가 존재하는가?

상관 분석 절차는 다음과 같다.



상관계수

		V4	V5	V6	V7	V8
V4	Pearson 상관계수	1.000	.484**	.357**	.475**	.316*
	유의확률 (양쪽)	.	.000	.000	.000	.000
	N	129	129	129	129	129
V5	Pearson 상관계수	.484**	1.000	.338**	.470**	.345*
	유의확률 (양쪽)	.000	.	.000	.000	.000
	N	129	130	130	130	130

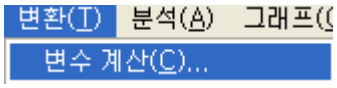
비모수 상관

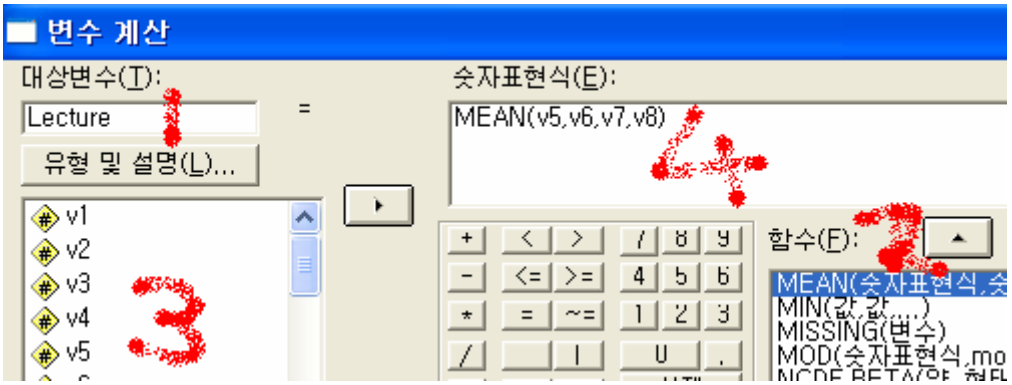
상관계수

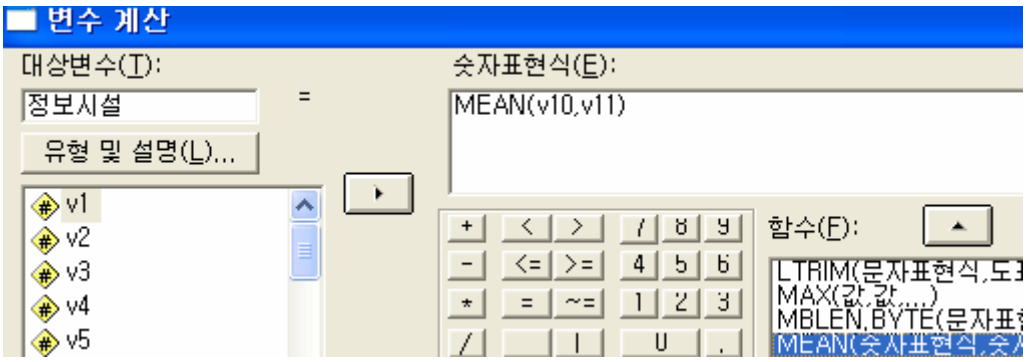
			V4	V5	V6
Kendall의 tau_b	V4	상관계수	1,000	,381**	,296**
		유의확률 (양쪽)	.	,000	,000
		N	129	129	129
	V5	상관계수	,381**	1,000	,254**
		유의확률 (양쪽)	,000	.	,000
		N	129	130	130

(2)경상대 시설 만족도(Q4-Q12) 중 경상대 시설(Q13) 전체 만족도에 영향을 미치는 시설은?

회귀 분석 절차는 다음과 같다. 우선 요인 분석에 의해 합칠 수 있는 문항을 합쳐 새로운 변수를 만든다.

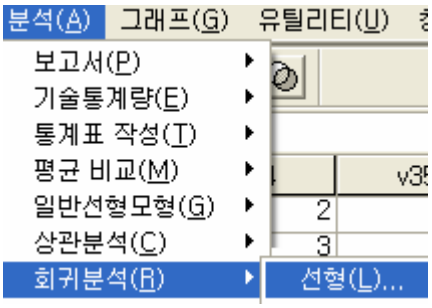


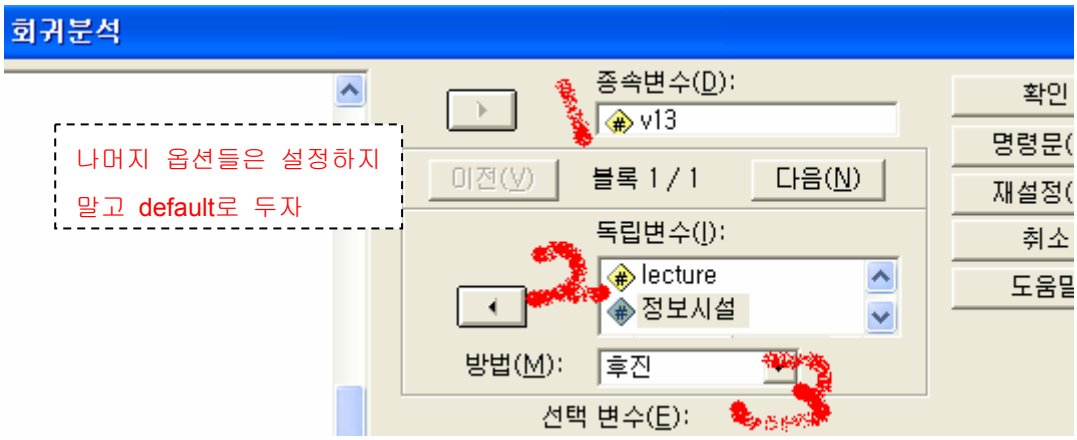




이렇게 하면 마지막 열에 변수 두 개 Lecture, 정보 시설 변수가 만들어진다.

v35	lecture	정보시설
5	2.75	1.50
.	1.00	1.00
.	2.00	1.00
6	2.50	4.00
5	4.25	4.00





변수 제거 결과로 SAS 결과와 동일하다.

진입/제거된 변수 [b]

모형	진입된 변수	제거된 변수	방법
1	정보시설, V12, LECTURE, V9, V4[a]	.	입력
2	.	V9	후진 (기준: 제거할 F의 확률 >= .100).

a. 요청된 모든 변수가 입력되었습니다.

b. 종속변수: V13

분산 분석 결과와 회귀 계수, 추정, 유의성 검정 결과(t-검정)도 동일하다.

분산분석 [c]

모형		제곱합	자유도	평균제곱	F	유의확률
1	선형회귀분석	118.739	5	23.748	40.944	.000[a]
	잔차	70.761	122	.580		
	합계	189.500	127			
2	선형회귀분석	118.024	4	29.506	50.776	.000[b]
	잔차	71.476	123	.581		
	합계	189.500	127			

a. 예측값: (상수), 정보시설, V12, LECTURE, V9, V4

b. 예측값: (상수), 정보시설, V12, LECTURE, V4

c. 종속변수: V13

계수[a]

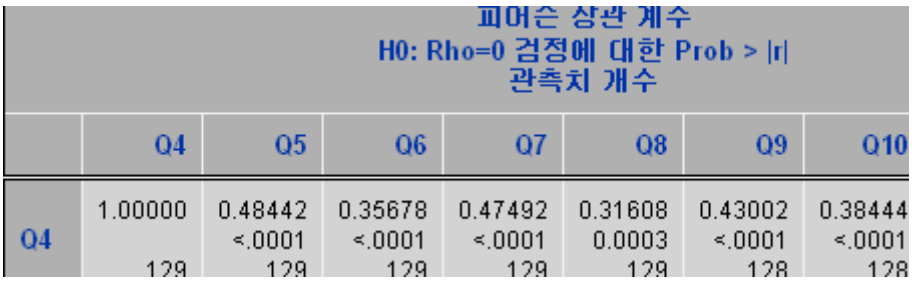
모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	-.137	.232		-.590	.556
	V4	.320	.074	.315	4.309	.000
	V9	7.550E-02	.068	.080	1.110	.269
	V12	.160	.047	.207	3.385	.001
	LECTURE	.277	.096	.217	2.890	.005
	정보시설	.240	.067	.248	3.611	.000
2	(상수)	-.109	.231		-.472	.638
	V4	.327	.074	.323	4.424	.000
	V12	.157	.047	.202	3.318	.001
	LECTURE	.317	.089	.248	3.546	.001
	정보시설	.264	.063	.273	4.186	.000

a. 종속변수: V13

10.4. 보고서 작성

10.4.1. 상관 계수

아래 웹 결과를 엑셀로 가져가 상관 계수와 유의 확률만 표로 적성하면 된다. (SPSS 는 결과 창에서 복사하여 적절히 표를 만들면 된다.



피어슨 상관 계수
H0: Rho=0 검정에 대한 Prob > |r|
관측치 개수

	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q4	1.00000 129	0.48442 <.0001 129	0.35678 <.0001 129	0.47492 <.0001 129	0.31608 0.0003 129	0.43002 <.0001 128	0.38444 <.0001 128

	건물안 공간	휴식공간	강의실 공간	강의실 시설	보조기자재
건물안 공간	1	0.48442 <.0001	0.35678 <.0001	0.47492 <.0001	0.31608 0.0003
휴식공간	0.48442 <.0001	1	0.33808 <.0001	0.46952 <.0001	0.34472 <.0001
강의실 공간	0.35678 <.0001	0.33808 <.0001	1	0.44755 <.0001	0.25 0.0041
강의실 시설	0.47492 <.0001	0.46952 <.0001	0.44755 <.0001	1	0.30147 0.0005
보조기자재	0.31608 0.0003	0.34472 <.0001	0.25 0.0041	0.30147 0.0005	1

문항간 상관 계수가 모두 양이므로 각 시설의 만족도가 높아지면 다른 시설에 대한 만족도도 높아짐을 알 수 있다. 이는 휴식 공간 만족도를 높일 수 있는 대책을 마련하여 시행하여 학생들의 휴식 공간 만족도가 높아진다면 다른 시설에 대한 만족도도 높아짐을 의미한다. 휴식 공간과 상관 관계가 가장 높은 건물 안 공간 만족도의 상승 정도가 가장 높을 것이다.

위의 경우 문항이 많아 유의 확률 값을 적어주면 표가 복잡해진다. 이런 경우 *(10% 유의 수준), **(5% 유의 수준), ***(1% 유의 수준, 매우 유의)을 이용하여 정리하면 된다.

	건물안 공간	휴식공간	강의실 공간	강의실 시설	보조기자재
건물안 공간		0.48442***	0.35678***	0.47492***	0.31608***
휴식공간			0.33808***	0.46952***	0.34472***
*: 10% 유의 수준, **: 5% 유의 수준, ***: 1% 유의 수준					

10.4.2. 회귀 분석

회귀 분석은 다음과 같이 작성하면 된다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-0.67738	0.29982	-2.26	0.0256	0
Q4	1	0.34557	0.06971	4.96	<.0001	0.34123

변수	추정치	추정 오차	T-값	유의확률	표준화 회귀계수
절편	-0.67738	0.29982	-2.26	0.0256	0
강의실 공간	0.34557	0.06971	4.96	<.0001	0.34123
강의실	0.45978	0.11009	4.18	<.0001	0.25804
정보시설	0.26827	0.06006	4.47	<.0001	0.2744
화장실	0.17686	0.04624	3.82	0.0002	0.22875

경상 대학 시설 만족도 영향을 미치는 것은 강의실 공간 만족도, 강의실 만족도, 정보 시설 만족도, 화장실 만족도이고 추정치가 +이므로 각 시설 만족도가 높을수록 시설 만족도는 높아짐을 알 수 있다. 강의실 공간 표준화 회귀계수가 0.34 로 가장 크므로 시설 만족도에 가장 큰 영향을 미친다. 그러므로 시설 만족도를 단기간에 높이려면 강의실 공간을 넓히는 대책을 강구해야 할 것이다.

[연습문제]

(1)예제 설문지에서 OO 대학교에 대한 전반적인 만족도(Q25)에 수준 만족도(Q22), 다니는 것에 대한 만족도(Q23), 입학 만족도(Q24) 가 영향을 미치는지 회귀 분석을 실시하시오.

(2)어느 만족도 문항이 전반적인 만족도에 가장 영향을 많이 미치는지 분석하시오.

(3)팀 프로젝트 설문지에서 회귀 분석이 가능하면 실시하시오.

(4)지금까지 실시한 팀 프로젝트 설문 조사 분석을 정리하여 보고서 형태로 만드시오. 보고서는 서론(조사 목적 및 조사 개요), 본론(분석 결과 정리), 결론으로 구성하시오.