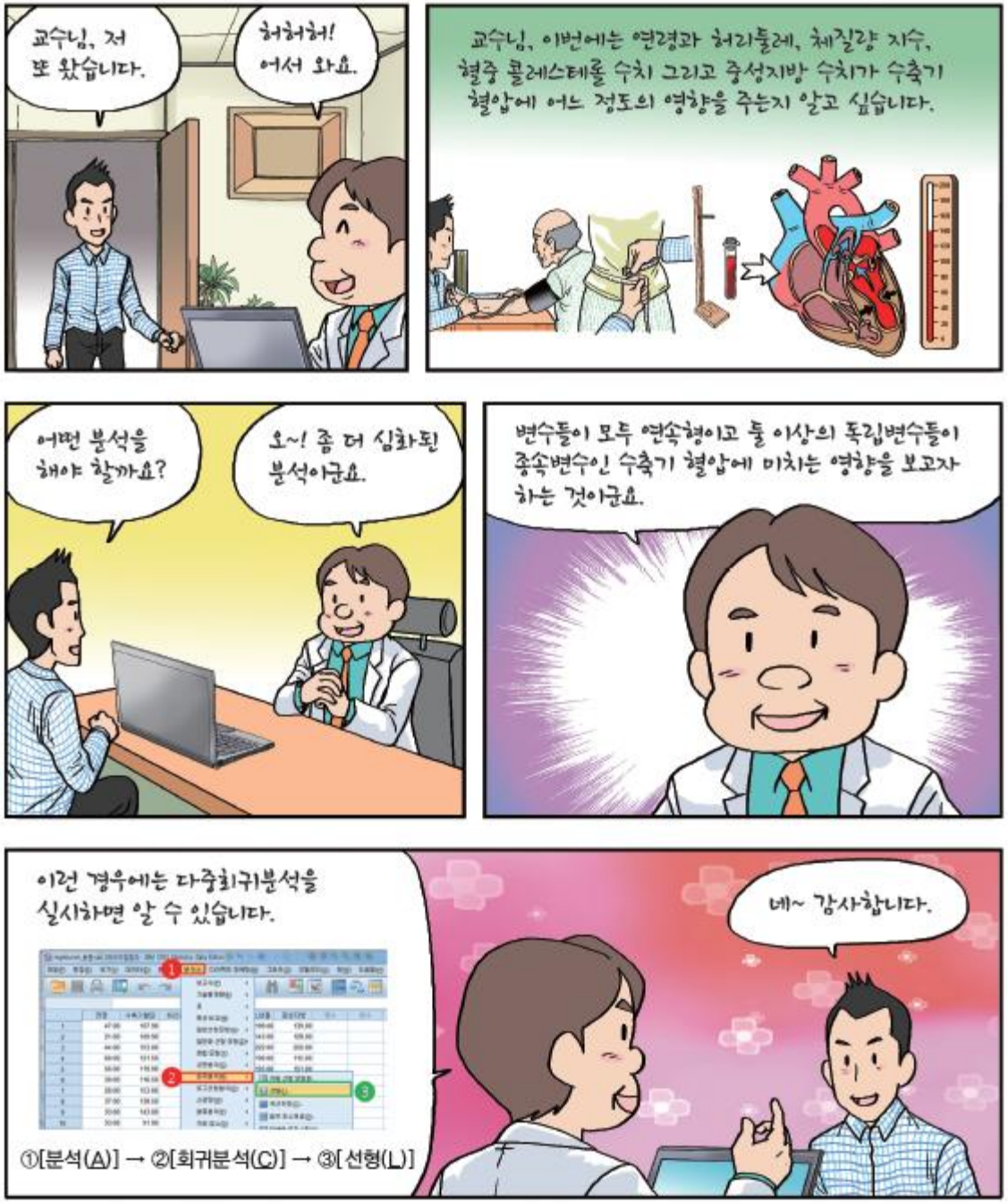
CHAPTER 알기 쉬운 **연구방법론**

14

회귀분석


교수님, 저 또 왔습니다.
허허허! 어서 와요.
교수님, 이번에는 연령과 허리둘레, 체질량 지수, 혈중 콜레스테롤 수치 그리고 중성지방 수치가 수축기 혈압에 어느 정도의 영향을 주는지 알고 싶습니다.
어떤 분석을 해야 할까요?
오~! 좀 더 심화된 분석이군요.
변수들이 모두 연속형이고 둘 이상의 독립변수들이 종속변수인 수축기 혈압에 미치는 영향을 보고자 하는 것이군요.
이런 경우에는 다중회귀분석을 실시하면 알 수 있습니다.
①[분석(A)] → ②[회귀분석(C)] → ③[선형(L)]
네~ 감사합니다.

- Y= f(x)의 함수관계에서 x는 독립변수이고 y는 종속변수라 할 때 x에 의해 y는 영향을 받는다.
- 회귀분석은 x와 y 의 관계를 찾아내는 분석방법으로 독립변인이 종속변인에 영향을 미치는지 알아보고자하는 방법
- 연속형 자료에 따른 연속형 자료의 영향력을 검증하고자 할 때

영향을 주는 변수	영향을 받는 변수	통계분석방법
범주형 자료	범주형 자료	카이제곱 검정
	연속형 자료	T검정 분산분석
연속형 자료	연속형 자료	회귀분석 구조방정식
	범주형 자료	로지스틱 회귀분석

- Y= f(x) 에서 x는 y에 영향을 미치고 이러한 관계 내에서 회귀분석은 이런 종속변수와 독립변수의 관계를 찾아내는 분석방법.
- X를 독립변수 = 설명변수
- Y는 종속변수 = 반응변수

- 예로 물품의 만족도는 가격, 포장, 맛 등이 영향을 준다. 이때 가격, 포장, 맛이 물품만족도에 어느 정도 영향을 주는가를 분석할 때 사용하고 이것을 식으로 나타낸 것을 단순회귀모형이라 한다.

만족도= 회귀계수 + (회귀계수 가격)+ 회귀계수 포장 + 회귀계수 맛 + 오차

이때 회귀계수는 독립변수 요소들이 종속변수인 만족도 어느 정도 영향을 주는가를 수치로 나타내는 것이라 할 수 있다.

회귀분석은 종속변수와 독립변수 둘 다 양적 변수가 되야 하고 만약 종속변수가 질적이고 독립변수가 양적 이면 **로지스트 회귀분석**을 사용해야 한다.

회귀분석은 어떤 연구문제에 적용할 수 있을까?

- 연구문제 예시

1. 부모의 수입이 성적에 미치는 영향
2. 키가 몸무게에 미치는 영향
3. 나이가 스마트폰 사용시간에 미치는 영향

01 회 귀 분 석

① 회귀분석은 변수 간에 서로 종속적인 관계에 있을 때 독립변수의 변화량에 따라 종속변수의 변화량이 어느 정도인지를 일차방정식의 함수관계로 파악할 수 있는 분석방법(선형 회귀식)

② 측정된 값들의 분포와 가장 근사한 직선방정식을 찾아낼 수 있는데 이를 직선회귀방정식이라 하며 이는 측정된 값들 사이로 임의로 선을 긋고 측정값들로 부터 임의로 그은 선까지 수직으로 내려 그은 거리의 제곱 R^2의 총 합이 최소가 되는 방정식을 의미하며 식으로 표현하면

$y = \alpha + \beta\chi + \varepsilon$ (종속변수에 영향을 주는 독립변수가 하나)

$y = \alpha + \beta_1\chi_1 + \beta_2\chi_2 + \beta_3\chi_3 + \cdots + \varepsilon$ (종속변수에 영향을 주는 독립변수가 둘 이상)

α 는 상수, β 는 회귀계수(기울기), ε 는 오차

참고)) 회귀분석과 상관관계

- **상관관계와 회귀분석의 차이점**

두 변수 X, Y에 대하여 X가 독립변수, Y가 종속변수라고 하면

상관관계 : X ↔ Y　　회귀분석 : X → Y

상관관계는 관련성(한 변수가 커질 때 다른 변수 값이 높거나 낮아지는 경향)에 목적을 둔 것이라면

회귀분석은 예측을 목적(두 변수간의 관계식, 혹은 여러 변수와의 관계식(회귀모형)을 구해야함)으로 한다.

예측 : 어떤 알고 있는 변수로부터 그에 대응하는 다른 변수의 값을 알아보는 것

③ 종속변수에 영향을 주는 독립변수가 하나- 단순회귀분석.

종속변수에 영향을 주는 독립변수가 둘 이상이면 다중회귀분석.

④ 독립변수의 변화에 따른 종속변수의 변화량을 가장 근사한 값으로 나타내는 방정식이기 때문에 기울기가 같더라도 측정된 값들의 분포는 달라 질 수 있다.

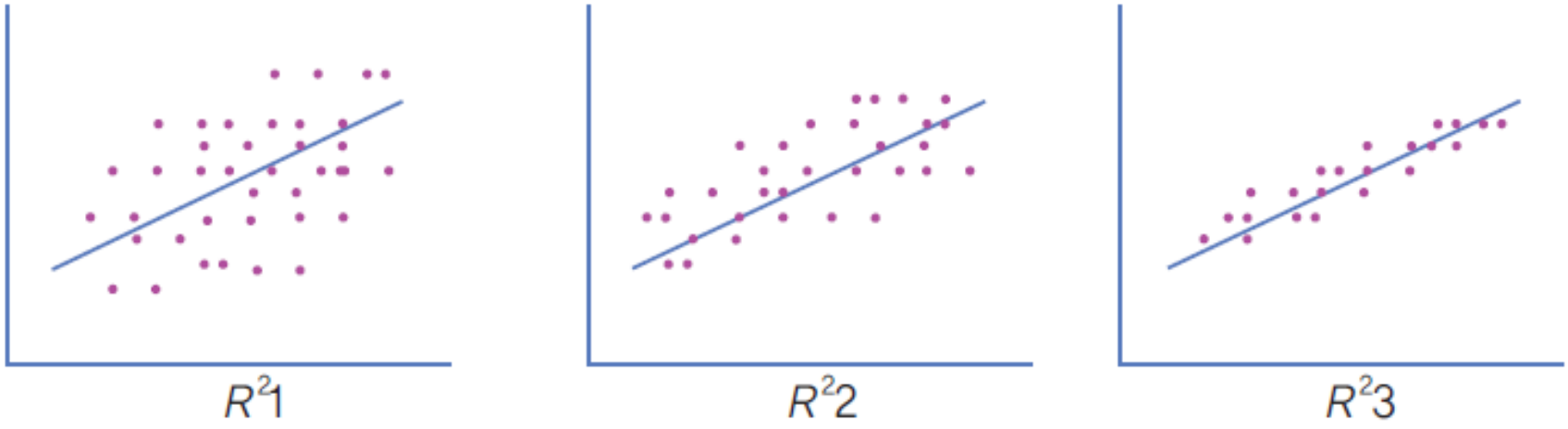
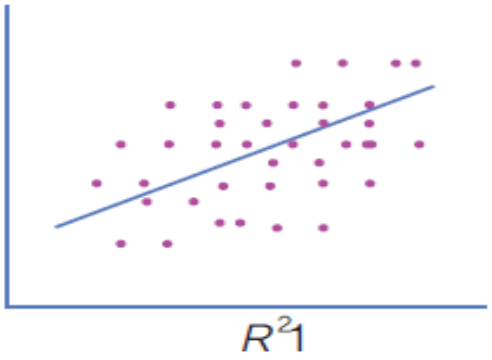


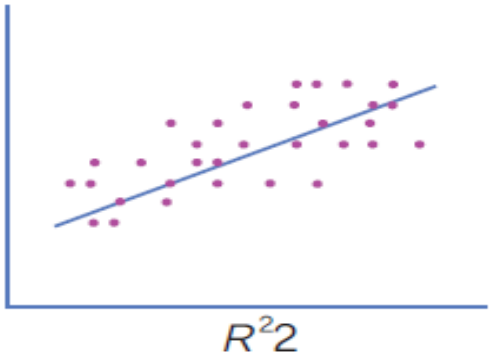
그림 14–1. 측정값들의 분포에 따른 변수의 설명력

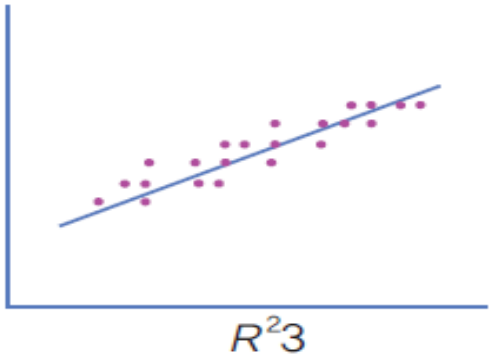
⑤ 종속변수에 대하여 독립변수들이 어느 정도 설명할 수 있는가를 파악해야 하는데 이를 변수의 설명력이라 하고 상관계수의 제곱으로 표현하며 이를 결정계수(R^2)

종속변수에 대한 독립변수들이 갖는 변수의 설명력의 크기는 $R^2 3 > R^2 2 > R^2 1$ 순이다.

단순회귀 분석일 때	다중회귀분석일 때
영가설(H_0): $\beta_1=0$	영가설(H_0): $\beta_1=\beta_2=\beta_3=\cdots\cdots=0$
대립가설(H_A): $\beta_1\neq0$,	대립가설(H_A): 모든 회귀계수가 0은 아니다.







회귀분석의 가설

H_0 : 독립변수 X가 Y에 영향을 주지 않는다.

H_1 : 독립변수 X가 Y에 영향을 준다.

회귀분석, 결과 분석하기

① R^2의 의미

R^2 를 적합도의 척도로 사용할 수 있다는 것은 알았지만 R^2 의 정확한 의미는 무엇일까? (R^2 는 추정된 직선이 자료를 얼마나 잘 설명할 수 있는지를 나타낸다.)

두 변수간에 회귀방정식을 구했다고 하자. R^2 =0라면 기울기는 0이다. R^2 가 1이라하면 오차가 0이다. 만약 0.52라면 52%를 설명한다고 할 수 있다.

② F값의 이해

- **F값은 모형 적합도를 나타냄**

P값이 0.05보다 작으면 이 모형이 적합하다고 할 수 있다.

P값이 0,05보다 크면 이 모형을 부적합하다고 할 수 있다.

R제곱과 F값

R제곱 : 독립변수가 종속변수의 몇 퍼센트인가를 설명하는 수치

F값 : 회귀식의 적합도 (p<0.05보다 작아야 회귀식이 유의미함)

③ B값과 β값

- **B는 적대적 영향력의 크기를 알려주며 표준화되지 않는 영향력을 판단할 때 사용하고,**

β는 이 영향력의 상대적인 차이를 비교할 때 사용합니다.

B와 β

B :비표준화 계수
절대적인 영향력의 크기

β :표준화 회귀계수
상대적인 영향력의 크기,
종속변수에 가장 큰 영향을 미치는 변수가 무엇인가를 판단할 때 활용

④ 공선성

- **공선성**은 한 독립변수의 값이 증가할 때 다른 독립변수의 값이 증가, 감소하는 현상을 말한다. 이 말은 독립변수 서로가 상관관계가 있어 독립적이지 않고 종속적이라는 말이기에 회귀분석을 하지 말아야 한다.
- 체크하기 위한 것이 분산팽창계수(VIF)와 공차한계(공차)이다.
- VIF의 수치 중 10 이상의 항목과 공차의 수치 중 0.1 이하의 항목은 버리거나 요인분석을 통해 항목을 합쳐 다시 회귀분석을 해야 한다. 즉 이 항목은 종속적이라는 말.

⑤잔차의 독립성 확인

- Durbin-Watson 통계량으로 확인 (0에 가까우면 음의 상관; 4에 가까우면 양의 상관- 결국 0과 4에 가깝지 않아야 한다)
- 2전 후값을 갖을 때

단순회귀분석의 과정

회귀방정식을 구하는 것이 끝이 아니라 검정이 필요하다.

((단순회귀직선의 적합도 검정))

가설의 설정 $Y_i = \beta_0 + \beta_1 X_i$

영가설 : $\beta_1 = 0$

대립가설 : $\beta_1 \neq 0$

H_0 : 독립변수 X가 Y에 영향을 주지 않는다.

H_1 : 독립변수 X가 Y에 영향을 준다.

검정통계량은 F값으로 본다.

회귀직선의 진단 시 고려해야 할 것

①독립변수들 간의 다중 공선성- 독립변수간에 상관이 높은 것을 피한다

②이상값에 대한 점검

Cook의 통계량의 값이 1.0이상이면 이상점으로 간주.

③잔차(residual) – 등분산성, 독립성, 정규성

잔차의 독립성의 경우, **Durbin-Watson 통계량으로 확인 (0에 가까우면 음의 상관; 4에 가까우면 양의 상관- 결국 0과 4에 가깝지 않아야 한다)**

잔차의 정규성의 경우 – 정규확률도표를 그려 확인.

(2) 잔차의 분석

- 잔차는 독립성, 정규성, 등분산성을 만족해야 하며, 그 검정을 위해 아래 그래프를 그릴 수 있다. 회귀모형을 이용한 추론이나 예측이 타당성을 얻으려면 회귀모형의 가정들을 만족해야 한다. Y_i 혹은 X_i에 대한 잔차 ei의 산점도를 이용하여 회귀모형의 가정에 대한 검토를 할 수 있다.

B병원의 중환자실 환자 20명 중에서 인공호흡을 하는 사람은 15명.
흡입된 산소의 분률과 ABGA로 측정된 동맥혈 산소 분압을 기록한 자료이다.

회귀분석의 가설

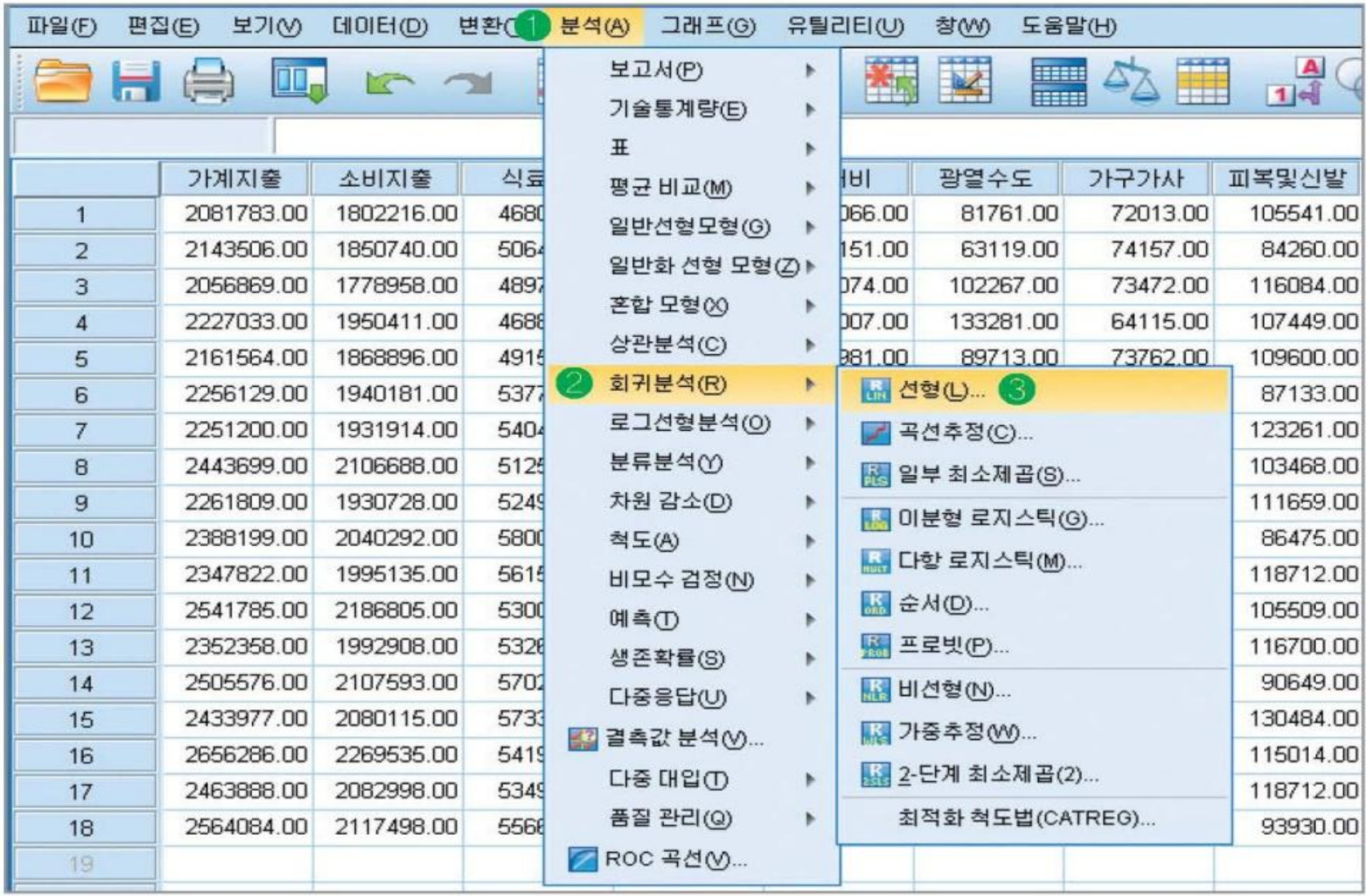
H_0 : 독립변수 X가 Y에 영향을 주지 않는다.

H_1 : 독립변수 X가 Y에 영향을 준다.

통계청 자료 분석을 통해 각 각 대입

가계지출 = f(소비지출 + 식료품 + 외식 + 주거비 + 가구가사 + 광열 수도 + 피복 및 신발)

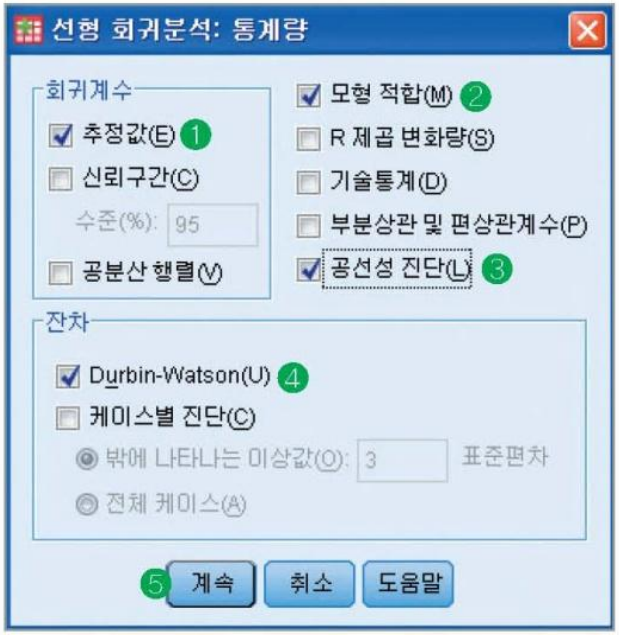
① 분석 → ② 회귀분석 → ③ 선형



① 가계지출은 종속변수

② 소비지출, 식료품, 외식, 주거비, 가구가사, 광열수도, 피복 및 신발은 독립변수에 삽입

③ 통계량 클릭



① 추정값 체크
② 모형 적합 체크
③ 공선성 진단 체크
④ Durbin-Watson 체크
⑤ 계속 클릭
⑥ 도표 클릭

① y축으로 이동
② x축으로 이동
③ 히스토그램 체크
④ 정규확률도표 체크
⑤ 계속 클릭

결과 보고서 분석

모형 요약[b]

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차	Durbin-Watson
1	.998[a]	.996	.993	14722.78596	2.230

a. 예측값: (상수), 피복 및 신발, 식료품, 광열수도, 주거비, 식료품외식, 가구가사, 소비지출
b. 종속변수: 가계지출

R 제곱은 결정계수(설명력)로 독립변수의 변화에 의해 설명될 수 있는 종속변수의 변화량을 나타낸다.

R 제곱은 회귀선의 기여율이다. 0에서 1 사이에 존재하며, 0에 가까우면 회귀분석의 유용 가치가 없어지며, 1에 가까울수록 회귀분석의 유용 가치가 높아진다.

Durbin-Watson 계수는 자동 상관에 대한 검정 통계량이다.

자동 상관은 관측 자료에 오류가 포함되는 것을 말한다.

Durbin-Watson 계수는 0에서 4 사이의 값이 나타나는데, 2에 가까우면 자동 상관을 무시해도 된다.

분산분석[b]

모형		제곱합	자유도	평균 제곱	F	유의확률
1	회귀 모형	5.048E11	7	7.211E10	332.660	.000[a]
	잔차	2.168E9	10	2.168E8		
	합계	5.069E11	17			

a. 예측값: (상수), 피복 및 신발, 식료품, 광열수도, 주거비, 식료품외식, 가구가사, 소비지출
b. 종속변수: 가계지출

- **분산분석은 회귀모형이 유의성을 검정하는 분석으로 가설은 다음과 같다.**

H_0 : 회귀모형이 필요하지 않다.

H_1 : 회귀모형이 필요하다.

유의 확률이 0.05보다 작아야 회귀모형이 이룬다.

계수[a]

모형	비표준화 계수		표준화 계수	t	유의확률	공선성 통계량	
	B	표준오차	베타			공차	VIF
1 (상수)	−284757.485	112678.805		−2.527	.030		
소비지출	1.212	.112	.936	10.777	.000	.057	17.635
식료품	−.203	.316	−.040	−.642	.535	.111	9.022
식료품외식	.267	.355	.034	.754	.468	.206	4.854
주거비	3.332	1.575	.097	2.115	.061	.203	4.936
광열수도	−.272	.492	−.043	−.552	.593	.070	14.265
가구가사	1.180	1.205	.057	.980	.350	.128	7.807
피복 및 신발	−.502	.605	−.040	−.830	.426	.188	5.309

a. 종속변수: 가계지출

회귀계수(비표준화 계수, B)는 독립변수와 종속변수와의 관계를 설명한다.

표준화계수(베타)는 독립변수가 종속변수에 미치는 상대적 영향력으로, 독립변수의 절대값이 가장 큰 변수가 가장 많은 영향을 미친다.

- **VIF의 수치 중 10 이상의 항목과 공차의 수치 중 0.1 이하의 항목은 버리거나 요인분석을 통해 항목을 합쳐 다시 회귀분석을 해야 한다. 즉 이 항목은 종속적이라는 말.**
- **이 연구에서 소비지출과 광열수도는 공차한계 0.1이므로 제외하고 다시 회귀분석을 해야 한다.**

- **방정식 구성**

Y(가계지출)= -284757.485(상수) + 1.212(소비치출) + 오차

- Y(가계지출)= -284757.485(상수) + 1.212(소비지출) + 오차

여기에서 통계학적인 의미에서 회귀계수가 양의 값인 소비지출이 1원 중가 할 수록 가계지출은 1.212원이 증가한다.

여기에서 유의확률 중 0.05이하는 소비지출뿐이다. 하지만 소비지출은 공차한계가 0.1이하이므로 제외하거나 요인분석을 통해 종속적인 요인을 합해 다시 회귀분석을 해야 한다.

책에서 예제

Tip

회귀분석에 사용되는 변수들 가운데 연속형이 아닌 명목형 즉 범주형 변수도 포함할 수 있다. 다만 자료의 값이 0과 1로 표현된 가변수로 치환된 경우에만 해당한다. 하지만 이 책에서는 가변수가 포함되지 않는 분석만을 예시로 보여준다.

예시

연령과 혈중 콜레스테롤, 허리둘레, 체질량지수, 중성지방이 수축기 혈압에 어떤 영향을 주는지 상관분석에서 활용 한 데이터를 이용하여 SPSS 프로그램을 이용하여 회귀분석을 실습해보도록 하자.

(1) 통계학적 가설

H_0: 연령과 허리둘레, 체질량 지수, 혈중 콜레스테롤 수치, 중성지방은 수축기 혈압에 영향을 주지 않는다.

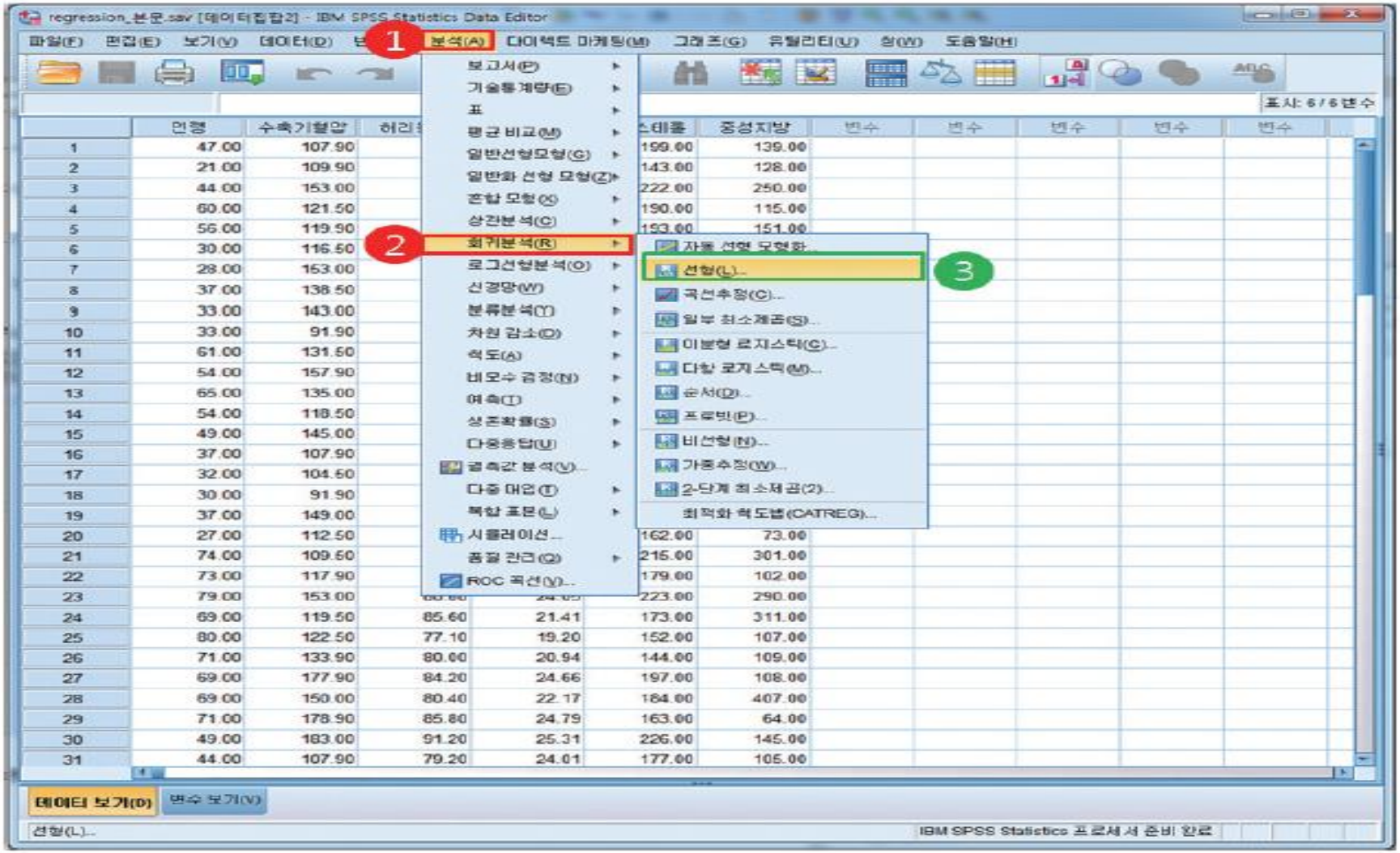
H_A: 연령과 허리둘레, 체질량 지수, 혈중 콜레스테롤 수치, 중성지방은 수축기 혈압에 영향을 줄 것이다.

(2) 자료 입력하기

자료의 입력은 모두 연속형 변수로 입력한다

(3) 회귀분석 대화상자 열기

메뉴에서 ①[분석(A)] →②[회귀분석(R)] →③[선형(L)]를 클릭한다.



(4) 변수 옮기기

대화상자에서 종속변수(D)에 수축기 혈압을 넣고, 독립변수(I)에 연령, 허리둘레, 체질 량 지수, 콜레스테롤, 중성지방을 입력한다.

(5) 분석 실행하기

방법(M)에서 유의성의 세기에 관계없이 모든 변수들이 종속변수에 미치는 영향을 한 번에 보고 싶다면 입력을, 유의성이 큰 변수들을 우선으로 하여 순차적인 모형을 만들고 싶다면 단계선택으로 바꿔주면 된다. [확인]을 클릭한다.

(5) 분석 실행하기


선형 회귀분석
연령
허리둘레
체질량지수
콜레스테롤
중성지방
종속변수(D):
수축기혈압
블록(B)1대상1
이전(V)
다음(N)
독립변수(I):
연령
허리둘레
체질량지수
방법(M):
입력
선택변수(C):
규칙(U)...
케이스 설명(C):
WLS 가중값(H):
통계량(S)...
도표(T)...
저장(S)...
옵션(O)...
붓스트랩(B)...
확인
붙여넣기(P)
재설정(R)
취소
도움말

(6) 결과 해석하기

진입/제거된 변수[a]

모형	진입된 변수	제거된 변수	방법
1	중성지방, 연령, 콜레스테롤, 체질량지수, 허리둘레[b]		입력

a. 종속변수: 수축기혈압

b. 요청된 모든 변수가 입력되었습니다.

모형요약

모형	R	R 제곱	수정된 R의 제곱	추정값의 표준오차
1	.615[a]	.378	.341	18.23995

a. 예측값: (상수), 중성지방, 연령, 콜레스테롤, 체질량지수, 허리둘레

(6) 결과 해석하기

수축기 혈압에 영향을 주는 변수들을 직선회귀 방정식으로 표현하는 방법

수축기 혈압 = 29.2171+(0.321 X 연령)-(0.003 X 허리둘레)+(1.637 X 체질량 지수)
+(0.180 X 콜레스테롤)+(0.072 X 중성지방)

(6) 결과 해석하기

분산분석[a]

모형		제곱합	자유도	평균 제곱	F	유의확률
1	회귀모형	17000.672	5	3400.134	10.220	.000[b]
	잔차	27946.449	84	332.696		
	합계	44947.121	89			

a. 종속변수: 수축기혈압
b. 예측값: (상수), 중성지방, 연령, 콜레스테롤, 체질량지수, 허리둘레

계수[a]

모형		비표준화 계수		표준화계수	t	유의확률
		B	표준오차	베타		
1	(상수)	29.217	23.793		1.228	.223
	연령	.321	.135	.238	2.371	.020
	허리둘레	−.003	.494	−.001	−.006	.995
	체질량지수	1.637	1.330	.196	1.231	.222
	콜레스테롤	.180	.070	.241	2.549	.013
	중성지방	.072	.031	.241	2.348	.021

a. 종속변수: 수축기혈압

논문에 실제로 적용하기

다른 변수를 통제한 상태에서 수축기 혈압은 연령이 한 살 증가할 때마다 0.321 mmHg 씩 증가하였고, 콜레스테롤 이 한 단위 증가할 때마다 0.180 mmHg씩 증가하였으며, 중성지방이 한 단위 증가할 때마다 0.072 mmHg씩 증가하였고 통계적으로 유의하였다(p,0.05). 하지만 체질량 지수와 허리둘레는 수축기 혈압에 대해 유의한 변수로 작용하지 못하였다. 이 회귀식에서 독립변수들이 수축기 혈압에 대하여 갖는 변수의 설명력은 34.1%였으며 분석 결과는 다음과 같다.

표 14-1. 수축기 혈압에 영향을 주는 요인에 대한 다중회귀 분석

변수	B	표준오차	베타	t	*p*값
(상수)	29.217	23.793		1.228	.223
연령	.321	.135	.238	2.371	.020
허리둘레	−.003	.494	−.001	−.006	.995
체질량지수	1.637	1.330	.196	1.231	.222
콜레스테롤	.180	.070	.241	2.549	.013
중성지방	.072	.031	.241	2.348	.021

종속변수: 수축기혈압

R^2: 0.341

다음의 자료는 식도암으로 사망한 사람들의 진단시 병기(Stage)와 생존개월수 (Month)를 나타낸 것이다. 다음 자료를 이용해 진단시 병기와 생존개월 수의 회귀방정식을 세워보자

Index	1	2	3	4	5
Stage	1	2	2	2	3
Month	50	48	37	35	43
Index	6	7	8	9	10
Stage	3	3	4	4	4
Month	22	19	15	7	5

모형 요약[b]

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차	통계량 변화량					Durbin-Watson
					R 제곱 변화량	F 변화량	df1	df2	유의확률 F 변화량	
1	.884[a]	.781	.754	8.258	.781	28.589	1	8	.001	2.554

a. 예측값: (상수), Stage

b. 종속변수: Month

1. 회귀분석에서 귀한 회귀선의 설명력 = 0.781로 상당히 높다

분산분석[a]

모형		제곱합	자유도	평균 제곱	F	유의확률
1	회귀 모형	1949.400	1	1949.400	28.589	.001[b]
	잔차	545.500	8	68.188		
	합계	2494.900	9			

a. 종속변수: Month

b. 예측값: (상수), Stage

2. 유의성 검정으로 0.05보다 작음으로 회귀직선은 유의하다.

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률	공선성 통계량	
		B	표준오차	베타			공차	VIF
1	(상수)	68.000	7.906		8.601	.000		
	Stage	-14.250	2.665	-.884	-5.347	.001	1.000	1.000

a. 종속변수: Month

3. 유의확률이 0.001임으로 stage는 month를 설명해주는 유의한 변수이고 회귀식은 비표준화 계수의 값을 읽어서 month = 68 - 14.25 X stage

표준화 계수는 모든 변수값을 평균 0, 표준편차를 1로 맞춘 값.
독립변수가 여러 개거나 변수간의 단위가 다를 때 변수의 영향력을 비교할 수 있다.
표준화계수의 값이 최대인 변수가 가장 영향력이 큰변수이다.