【특강 2】 SPSS를 활용한 가설검정의 실례

1. 교차분석(Cross Analysis)

① 두 범주형 변수간의 연관성을 검정하는 겨우 카이제곱 검정통계량을 이용한다.
② 카이제곱 검정은 교차표(분할표)로 주어지기 때문에 교차분석이라고도 한다.

(1) 카이제곱 독립성 검정

두 범주형 변수 사이의 연관성을 검정하는 것	
가설	H_0 : A와 B는 서로 독립이다 (A와 B는 서로 연관성이 없다) H_1 : A와 B는 서로 독립이 아니다 (A와 B는 서로 연관성이 있다)
p값-검정	유의수준 α > 유의확률 p값 : H_0 기각 유의수준 α < 유의확률 p값 : H_0 채택

〖예1〗 성별과 영화장르 선호도에 대한 400명의 자료가 아래와 같이 교차표의 형태로 구성되어 있다. 성별과 영화장르 선호도에 연관성이 있는지 유의수준 0.05로 검정하여라.

성별	영화장르				합계
	액션	코믹	멜로	공포	
남자	95	40	80	25	240
여자	65	50	40	5	160
합계	160	90	120	30	400

H_0 : 성별과 영화장르 선호도는 서로 독립이다.
H_1 : 성별과 영화장르 선호도는 서로 독립이 아니다.

[결과] Pearson 카이제곱 검정 결과

	값	자유도	점근 유의확률 (양측검정)
Pearson 카이제곱	18.128[a]	3	.000
우도비	18.811	3	.000
선형 대 선형결합	5.510	1	.019
유효 케이스 수	400		

a. 0셀(.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다.
최소 기대빈도는 12.00입니다.

위의 검정 결과 Pearson 카이제곱 검정통계량 값이 18.128이고 유의확률 p값이 0.000으로 유의수준 0.05보다 작으므로 귀무가설을 기각한다. 즉, 유의수준 0.05에서 성별과 영화장르 선호도는 연관성이 있다고 할 수 있다.

〚예2〛 회사직원 300명, 한 달 동안 업무와 관련하여 만난 인원수를 조사한 자료로부터 직위와 부서에 대한 이차원 교차표를 작성하였다. 두 변수 간의 독립성 여부에 대한 검정을 유의수준 0.05에서 실시하시오.

[결과]

직위*부서 교차표

			부서		전체
			영업부	관리부	
직위	사원	빈도	130	77	207
		직위 중 %	62.8%	37.2%	100.0%
	팀장	빈도	41	24	65
		직위 중 %	63.1%	36.9%	100.0%
	차장	빈도	15	13	28
		직위 중 %	53.6%	46.4%	100.0%
전체		빈도	186	114	300
		직위 중 %	62.0%	38.0%	100.0%

Pearson 카이제곱 검정 결과

	값	자유도	점근 유의확률 (양측검정)
Pearson 카이제곱	.933[a]	2	.627
우도비	.914	2	.633
선형 대 선형결합	.533	1	.465
유효 케이스 수	300		

a. 0셀(.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다.
최소 기대빈도는 10.64입니다.

교차분석 결과의 카이제곱 검정표를 보면 Pearson 카이제곱 검정통계량 값이 0.933이고 유의확률 p값이 0.627로 유의수준 0.05보다 크기 때문에 귀무가설을 채택한다. 즉, 유의수준 0.05 하에서 직위와 부서는 서로 독립이라고 할 수 있다.

(2) 카이제곱 동일성(동질성) 검정

<table>
<tr><td colspan="2">하나의 특성에 대하여 몇 개의 범주로 분류된 자료가 주어졌을 때 여러 모집단들이 주어진 특성에 대하여 서로 동일한 분포를 하는지 검정하는 것</td></tr>
<tr><td>가설</td><td>H_0 : 각 집단이 변수 B의 범주에 대해 동일한 비율을 가진다.
$(p_{11}, p_{12}, \cdots, p_{1c})=(p_{21}, p_{22}, \cdots, p_{2c})=...=(p_{r1}, p_{r2}, \cdots, p_{rc})$
H_1 : 각 집단이 변수 B의 범주에 대해 동일한 비율을 갖지 않는다.</td></tr>
</table>

① 카이제곱 동일성 검정 절차는 가설을 제외하고 카이제곱 독립성 검정 절차와 동일하다.
② 단, 가설 설정과 자료를 모으는 단계 및 결과를 해석하는 부분에 차이가 있다. 실제 분석에서 카이제곱 독립성 검정과 카이제곱 동일성 검정을 분류하는 것은 쉽지 않으며 자주 혼용해서 사용한다.
→ 카이제곱 동일성 검정 “차이가 있는지를 검정하여라”
→ 카이제곱 독립성 검정 “연관성, 관련성, 독립성이 있는지를 검정하여라”

[예] 다음 자료는 연령대별 TV 프로그램 선호도를 조사한 교차표이다. 연령대별 TV 프로그램 선호도에 차이가 있는지 유의수준 0.05에서 검정하여라.

연령대	TV 프로그램			합계
	A	B	C	
20대	120	30	50	200
30대	10	75	15	100
40대	10	30	60	100
합계	140	135	125	400

H_0 : 연령대에 따라 TV 프로그램 선호도에 차이가 없다.
H_1 : 연령대에 따라 TV 프로그램 선호도에 차이가 있다.

[결과] Pearson 카이제곱 검정 결과

	값	자유도	점근 유의확률 (양측검정)
Pearson 카이제곱	180.495[a]	4	.000
우도비	177.248	4	.000
선형 대 선형결합	74.149	1	.000
유효 케이스 수	400		

a. 0셀(.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다.
최소 기대빈도는 31.25입니다.

Pearson 카이제곱 검정통계량 값이 180.495이고 유의확률 p값이 0.000으로 유의수준 0.05보다 작으므로 귀무가설을 기각한다. 즉, 유의수준 0.05하에서 연령대에 따라 TV 프로그램 선호도에 차이가 있다고 할 수 있다.

(3) 카이제곱 적합성 검정

범주에 속한 관찰도수가 범주에 속한 확률과 적합한지 검정하는 것 (i번째 범주에 속할 확률 p_i가 미리 주어진 확률 π_i와 같은지 검정)

범주	1	2	...	k	합계
관찰도수	O_1	O_2	...	O_k	n
범주에 속할 확률	p_1	p_2	...	p_k	1

가설	귀무가설(H_0) : $p_1 = \pi_1, \ldots, p_k = \pi_k$ 대립가설(H_1) : 귀무가설이 아니다.

[예] 어떤 종류의 완두콩을 색깔과 모양에 따라 노랗고 둥근형, 노랗고 뾰족한 형, 초록색에 둥근형, 초록색에 뾰족한 형의 네 가지로 구분할 때 각 종류에 속할 비율이 9 : 3 : 3 : 1이 된다. 완두콩 100개를 조사, 각 형태에 속하는 관측수가 54, 20, 16, 10으로 나타났다고 하면 이 결과는 이 법칙에 따른다고 할 수 있는지를 유의수준 0.05에서 검정하여라.

범주	1	2	3	4	합계
관찰도수	54	20	16	10	100
범주에 속할 확률	9/16	3/16	3/16	1/16	1

H_0 : $p_1 = 9/16,\ p_2 = 3/16,\ p_3 = 3/16,\ p_4 = 1/16$
H_1 : 귀무가설이 아니다.

[결과] 카이제곱 적합성 검정 결과

	관측도수
카이제곱	2.827[a]
자유도	3
근사 유의확률	.419

a. 0셀(.0%)은(는) 5보다 작은 기대빈도를 가집니다. 최소 기대빈도는 6.3입니다.

카이제곱 검정 통계량 값이 2.827이고 유의확률 p값이 0.419로 유의수준 0.05보다 크므로 귀무가설을 채택한다. 즉, 유의수준 0.05에서 완두콩은 멘델의 법칙 ($p_1 = 9/16,\ p_2 = 3/16,\ p_3 = 3/16,\ p_4 = 1/16$)에 따라 생산된다고 할 수 있다.

2. t검정

① 두 집단간의 모평균에 차이가 있는지를 검정하고자 할 때 T검정을 사용한다.
 ㉠ 독립표본 t검정 : 두 집단이 서로 독립
 ㉡ 대응표본 t검정 : 두 집단이 서로 종속(짝 비교)

(1) 독립표본 t검정

두 집단이 각각 $N(\mu_1, \sigma_1^{\ 2})$과 $N(\mu_2, \sigma_2^{\ 2})$인 정규분포를 따르고 서로 독립이라는 가정 하에 두 집단 간 모평균에 차이가 있는지를 검정하는 것	
가설	H_0 : $\mu_1 = \mu_2$ H_1 : $\mu_1 \neq \mu_2$

【참고】 독립표본 t검정은 등분산 검정결과 두 집단간 모분산이 같은지 다른지에 따라 t검정통계량이 다르게 적용된다.

〚예1〛 어느 회사직원 300명에 대해 부서에 따라 근무년 수를 조사하였다. 부서(1=영업부, 2=관리부)에 따라 근무년 수에 차이가 있는지 유의수준 0.05에서 검정하여라.
H_0 : 부서에 따라 근무년 수에 차이가 없다.
H_1 : 부서에 따라 근무년 수에 차이가 있다.

[결과] 독립표본 검정

		Levene의 등분산 검정		평균의 동일성에 대한 t검정		
		F	유의확률	t	자유도	유의확률 (양쪽)
근무 년수	등분산이 가정됨	.441	.507	-.621	298	.535
	등분산이 가정되지 않음			-.615	231.795	.539

㉠ Levene의 등분산 검정 ; 「검정통계량 F값=0.441」, 「유의확률 p값= 0.507」, 유의수준 0.05보다 크기 때문에 귀무가설을 채택
㉡ '등분산이 가정됨' 의 t검정 이용 ; 「검정통계량 t값=-0.621」, 「유의확률=0.535」, 유의수준 0.05보다 크므로 귀무가설을 채택. 즉, 유의수준 0.05에서 부서에 따라 근무년 수에 차이가 없다고 할 수 있다.

〚예2〛 어느 회사직원 300명에 대해 한 달 동안 업무와 관련하여 만난 인원수를 조사하였다. 직위범주(1=일반직, 2=관리직)에 따라 업무상 만나는 사람의 수에 차이가 있는지를 유의수준 0.01에서 검정하여라.

H_0 : 직위범주에 따라 업무상 만나는 사람의 수에 차이가 없다.

H_1 : 직위범주에 따라 업무상 만나는 사람의 수에 차이가 있다.

[결과]

독립표본 검정

		Levene의 등분산 검정		평균의 동일성에 대한 t검정		
		F	유의확률	t	자유도	유의확률(양쪽)
만난수	등분산이 가정됨	274.806	.000	-18.493	298	.000
	등분산이 가정되지 않음			-12.664	94.255	.000

㉠ Levene 등분산검정 ;「검정통계량 F값=274.806」, 「유의확률 p값=0.000」, 유의수준 0.01보다 작기 때문에 귀무가설을 기각.

㉡ 등분산이 가정되지 않음의 t검정 결과를 이용 : 「검정통계량 t값=-12.664」 , 「유의확률=0.000」로 유의수준 0.01보다 작으므로 귀무가설을 기각. 즉, 유의수준 0.01에서 직위범주에 따라 업무상 만나는 사람의 수에는 차이가 있다고 할 수 있다.

(2) 대응표본(paired sample) t검정

- 두 집단이 서로 독립이라는 가정을 필요로 하지 않으며 서로 짝을 이룬 자료일 때, 두 집단간 모평균에 차이가 있는지를 검정하는 것 - 각 쌍의 차이를 $D_i = X_{1i} - X_{2i}$ $(i = 1,\ 2,\ \dots,\ n)$이라 하고 $D_i \sim N(\mu_D,\ {\sigma_D}^2)$을 따른다고 가정	
가설	H_0 : 두 집단간 모평균에 차이가 없다 $(\mu_1 - \mu_2 = \mu_D = 0)$ H_1 : 두 집단간 모평균에 차이가 있다 $(\mu_1 - \mu_2 = \mu_D \neq 0)$

[예] 어느 고등학교 학생 400명, 토론식 수업을 하기 전의 국어점수와 토론식 수업을 한 후의 국어점수의 평균에 차이가 있는지 여부를 유의수준 0.05에서 검정하여라.

H_0 : 토론식 수업 전과 후의 국어점수의 평균에 차이가 없다.
H_1 : 토론식 수업 전과 후의 국어점수의 평균에 차이가 있다.

[결과]

대응표본 검정

		t	자유도	유의확률(양쪽)
대응	토론후국어-토론전국어	4.896	396	.000

㉠ 대응표본 t검정 결과 ; 「검정통계량 t값=4.896」, 「유의확률 p값=0.000」으로 유의수준 0.05보다 작으므로 귀무가설을 기각. 즉 유의수준 0.05에서 토론식 수업 전과 후의 국어점수의 평균에 차이가 있다고 할 수 있다.

㉡ 단, '토론전국어-토론후국어'를 분석하면 「검정통계량 t값=-4.896」이 된다. 대응표본 t검정은 서로 짝을 이룬 자료의 두 집단간 모평균의 차이가 있는지 없는지를 양측검정하기 때문에 어떤 검정통계량을 사용해도 결과 분석은 동일하다.

【참고】 독립표본 T검정과 대응표본 T검정의 비교

독립표본 T검정	대응표본 T검정
조사 대상 개체가 다름	조사 대상 개체가 같음
두 표본의 숫자가 다를 수 있음	반드시 짝이 있음
다른 집단을 비교하는 경우	전후 개념이 있는 경우가 많음
서로 독립이라는 가정이 필요	서로 독립이라는 가정이 필요치 않음
- 도시 지역과 시골지역의 평균 가족 수에 차이가 있는지 비교 - 흑인과 백인간의 지능 지수에 차이가 있는지 비교 - 대졸사원의 성별에 따라 월별 초임에 차이가 있는지 비교	- 동일한 운전자에게 기존 휘발유와 새로 개발한 휘발유의 평균 주행거리에 차이가 있는지 비교 - 10명의 학생들에게 새로운 교수법을 실시하여 이전 성적과 새로운 교육법을 실시한 이후의 성적이 같은지 비교 - 오른발에는 새로 산 운동화, 왼발에는 기존 운동화를 신고 걸은 후 운동화의 마모도가 같은지 비교

3. 분산분석

① 집단의 수가 2개인 경우 모평균에 차이가 있는지를 검정하기 위해 t검정을 사용한다. 집단의 수가 3개 이상인 경우의 모평균을 비교하고자 할 때는 분산분석을 사용한다.
② 집단을 나타내는 변수인 요인의 수가 1개인 경우 일원배치 분산분석이라 한다.
③ 변수의 요인의 수가 2개인 경우 이원배치 분산분석, 요인의 수가 3개 이상인 경우 다원배치 분산분석이라고 한다.

(1) 일원배치 분산분석

요인의 수가 1개이고 하위에 수준(level)이 k인 경우의 일원배치 분산분석표

요인	제곱합	df	평균제곱	검정통계량 F
집단간	SS_B	$k-1$	$MS_B=\dfrac{SS_B}{k-1}$	$F=\dfrac{MS_B}{MS_W}$
집단내	SS_W	$n-k$	$MS_W=\dfrac{SS_W}{n-k}$	
합계	SS_T	$n-1$		

가설	귀무가설(H_0) : $\mu_1=\mu_2=\cdots=\mu_k$ 대립가설(H_1) : 모든 μ_i가 같은 것은 아니다.

〖예〗 어느 고등학교를 졸업하는 1(문과계열) 학과와 2, 3(이과계열) 학과의 졸업생 300명의 성적을 조사하였다. 학과(1=문과, 2=이과, 3=이과)에 따라 수능점수에 차이가 있는지를 유의수준 0.05에서 검정하여라(단, 분산의 동일성 검정도 실시하여라).

H_0: 학과에 따라 수능점수에 차이가 없다($\mu_1=\mu_2=\mu_3$).

H_1: 학과에 따라 수능점수에 차이가 있다(모든 μ_i가 같은 것은 아니다)

[결과] 분산의 동질성에 대한 검정

Levene 통계량	df_1	df_2	유의확률
2.739	2	297	.066

일원배치 분산분석

요인	제곱합 (SS)	df	평균제곱	검정통계량 F	$p-$값
집단간	11416.312	2	5708.156	8.060	.000
집단내	210337.994	297	708.209		
합계	221754.306	299			

「검정통계량 F값=8.060」, 「유의확률 p값=0.000」, 유의수준 0.05보다 작으므로 귀무가설을 기각한다. 즉, 유의수준 0.05에서 학과에 따라 수능점수에 차이가 있다고 할 수 있다.

(2) 사후비교(다중비교)

① 분산분석 결과 귀무가설이 기각 된 경우, 적어도 한 쌍의 모평균이 같지 않다고 판단되는 경우에 어느 집단의 모평균 사이에 차이가 있는지를 분석한다.
② 다중비교 방법 ; 모든 대비를 이용한 Scheffé 검정법, 최소유의차를 이용하여 분석하는 LSD 검정법, Duncan 다중범위 검정법, Tukey 검정법, Dunnet 검정법 등
③ 다중비교 방법의 결과가 모두 동일하게 나오는 것은 아니므로 어느 방법이 좋다고 판단할 수는 없으며 이들 각 방법의 결과를 종합하여 결론을 내리는 것이 바람직하다.

〚예〛 유의수준 0.05에서 학과(1=문과, 2=이과, 3=이과)에 따라 수능점수에 차이가 있다는 것을 알았다. 어떤 그룹에 차이가 있는지를 유의수준 0.05에서 사후검정(Scheffé, LSD, Tukey, Duncan)을 실시하시오.

[결과] Tukey, Scheffé, LSD 방법에 의한 다중비교 결과
위의 다중비교 결과 Tukey, Scheffé, LSD 방법 모두 유의수준 0.05에서 집단 1과 집단 2 및 집단 1과 집단 3이 수능점수의 모평균에 유의한 차이가 있는 것으로 나타났다.

일원배치 분산분석

	학과	N	유의수준=.05에 대한 부집단 1	유의수준=.05에 대한 부집단 2
Tukey HSDa	1	100	327.796	
	2	100		340.518
	3	100		341.218
	유의확률		1.000	.981
Duncana	1	100	327.796	
	2	100		340.518
	3	100		341.218
	유의확률		1.000	.853
Scheffea	1	100	327.796	
	2	100		340.518
	3	100		341.218
	유의확률		1.000	.983

⇒ 동일 집단군 결과 집단 2와 집단 3을 동일 집단군으로 분류

4. 상관분석 (피어슨 상관계수)

<table>
<tr><td colspan="2">

· 피어슨 상관계수를 흔히 상관계수라 함
· 상관계수 r이란 공분산이 두 확률변수가 취하는 값의 단위에 의존하기 때문에 이러한 단위에 대한 의존도를 없애주기 위하여 공분산을 두 확률변수의 표준편차의 곱으로 나눈 값
· 확률분포가 정규분포를 따른다고 가정
· 표본으로부터 구한 피어슨 상관계수 r을 이용하여 모상관계수 ρ를 검정
</td></tr>
<tr><td>가설</td><td>H_0: 두 변수 간에 선형 연관성은 없다($\rho=0$).
H_1: 두 변수 간에 선형 연관성이 있다($\rho\neq 0$).</td></tr>
</table>

[예] 어느 고등학교 400명의 수능과 졸업평점과의 상관계수를 구하고 유의수준 0.05에서 검정하여라.

H_0 : 수능과 졸업평점 간에 선형 연관성이 없다($\rho=0$).
H_1 : 수능과 졸업평점 간에 선형 연관성이 있다($\rho\neq 0$).

[결과]

상관계수

		졸업평점	수능
졸업평점	Pearson 상관계수	1	.227**
	유의확률(양쪽)		.000
	N	300	300
수능	Pearson 상관계수	.227**	1
	유의확률(양쪽)	.000	
	N	300	300

**. 상수관계는 0.01수준(양쪽)에서 유의합니다.

수능과 졸업평점 간에 「상관계수=0.227」「유의확률 p값=0.000」유의수준 0.05보다 작으므로 귀무가설을 기각한다. 즉, 유의수준 0.05 하에서 수능과 졸업평점 간에 선형 연관성이 있다고 할 수 있다.

5. 회귀분석

① 회귀분석이란 독립변수(설명변수)와 종속변수(반응변수) 간에 존재하는 연관성을 분석하기 위하여 관측된 자료에서 이들 간의 함수적 관계식을 통계적으로 추정하는 방법이다.
② 단순회귀분석 : 독립변수가 1개일 경우
다중회귀분석 : 독립변수가 2개 이상일 경우
③ 선형관계의 여부에 따라 선형회귀분석과 비선형회귀분석으로 구분한다.
④ 잔차(=오차=관측값-예측값; $e_i = y_i - \hat{y_i}$)가 작으면 작을수록 좋은 회귀선이다.

(1) 단순회귀모형의 유의성 검정

· 단순회귀모형의 분산분석표

요인	제곱합	df	평균제곱	검정통계량 F	$p-$값
회귀	SSR	1	$MSR=\dfrac{SSR}{1}$	$F=\dfrac{MSR}{MSE}$	
잔차	SSE	$n-2$	$MSE=\dfrac{SSE}{n-2}$		
합계	SST	$n-1$			

SSR(sum of squares due to Regression) 회귀제곱합
SSE(sum of squares due to Residual Errors) 잔차제곱합
SST(Total sum of squares) 총제곱합

가설	H_0 : 회귀모형은 유의하지 않다($\rho=0$). H_1 : 회귀모형은 유의하다($\rho\neq 0$).

(2) 단순회귀계수의 유의성 검정

· t검정통계량의 제곱이 F검정통계량이 되므로 단순회귀계수의 유의성 검정과 단순회귀모형의 유의성 검정은 동일함. 즉, 단순회귀에서는 귀무가설을 검정하기 위해 t검정통계량과 F검정통계량 중 어느 것을 사용해도 무방

가설	H_0 : 회귀계수 ρ는 유의하지 않다($\rho=0$). H_1 : 회귀계수 ρ는 유의하다($\rho\neq 0$).

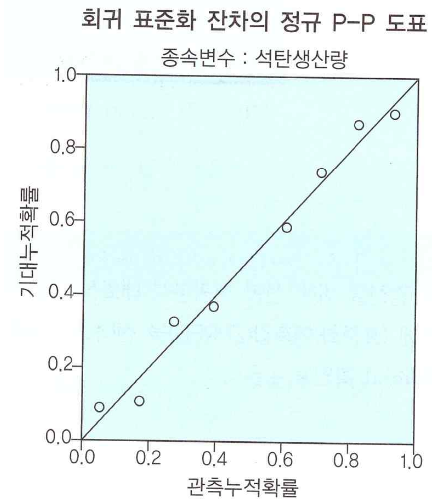
〖예〗 다음은 2003년부터 2011년까지의 우리나라 석탄 생산량(단위: 만톤)을 조사한 자료이다. 변환된 연도를 이용하여 회귀모형의 적합성 및 유의성 검정과 회귀계수의 유의성 검정을 유의수준 0.05에서 분석하여라. 그리고 회귀식을 구하여라.

우리나라 연도별 석탄 생산량 자료

연도	2003	2004	2005	2006	2007	2008	2009	2010	2011
변환된 연구(x)	0	1	2	3	4	5	6	7	8
석탄 생산량(y)	330	320	283	280	289	277	252	208	208

[결과]

㉠ 회귀모형의 적합성



모형 요약[b]

모형	R	R제곱	수정된 R제곱
1	.945[a]	.892	.877

a. 예측값 : (상수), 변형된 연도
b. 종속변수 : 석탄생산량

: 결정계수 값 R^2이 0.892로 총변동 중에서 회귀선에 의해 설명되는 비율이 89.2%이다.

㉡ 회귀모형의 유의성 검정

분산분석[b]

모형	제곱합	df	평균제곱	F	유의확률
회귀	13172.017	1	13172.017	58.104	.000[b]
잔차	1586.872	7	226.696		
합계	14758.889	8			

a. 예측값 : 석탄생산량
b. 종속변수 : (상수), 변형된 연도

: 분산분석표 결과 「검정통계량 F값=58.104」, 「유의확률 p값=0.000」으로 유의수준 0.05보다 작으므로 귀무가설을 기각한다. 즉, 유의수준 0.05에서 회귀모형은 유의하다고 할 수 있다.

㉢ 회귀계수의 유의성 검정을 유의수준 0.05에서 분석

계수[a]

모형	비표준화 계수		표준화 계수	t	유의확률
	B	표준오차	베타		
1 (상수)	331.156	9.254		35.784	.000
변형된연도	-14.817	1.944	-.945	-7.623	.000

a. 종속변수 : 석탄생산량

: 회귀계수 b의 「검정통계량 t값=-7.623」, 「유의확률 p값=0.000」으로 유의수준 0.05보다 작으므로 귀무가설을 기각. 유의수준 50.05에서 회귀계수 b는 유의하다고 할 수 있다.

④ 추정된 회귀식은 $\hat{y}=331.156-14.817x$. 연도가 1년 증가할 때 석탄 생산량은 14.817만 톤 감소함