



# SAS,SPSS를 활용한 회귀모형 적합

정보통계학과 김성수교수

## ✓ 학습목차

1

SAS 사용법

2

SAS를 이용한 회귀모형 적합

3

SPSS 사용법

4

SPSS를 이용한 회귀모형 적합

# **1** SAS 사용법

---

# SAS 소개

- ✓ 미국 North Carolina 주립대학에서 출발하여 종합정보처리시스템으로 발전.  
세계에서 가장 널리 쓰이는 통계패키지중의 하나임.
  - 다양한 형태의 데이터 처리가 용이
  - Excel 등 다양한 데이터베이스 데이터 처리가 용이
  - 대량 자료의 처리가 용이
  - 공인된 거의 모든 통계분석을 포함
- 
- ✓ SAS 소프트웨어 무료제공 : SAS University Edition 제공  
([http://www.sas.com/ko\\_kr/software/university-edition.html](http://www.sas.com/ko_kr/software/university-edition.html))

# SAS 활용 예-1

〈성적 데이터〉

이름	번호	성별	과	중간시험	기말시험
민영	1	여자	통계	50	50
민지	2	여자	컴퓨터	50	50
콩쥐	6	여자	컴퓨터	20	40
팔쥐	3	여자	통계	20	45
홍부	5	남자	컴퓨터	25	25
놀부	4	남자	통계	29	28



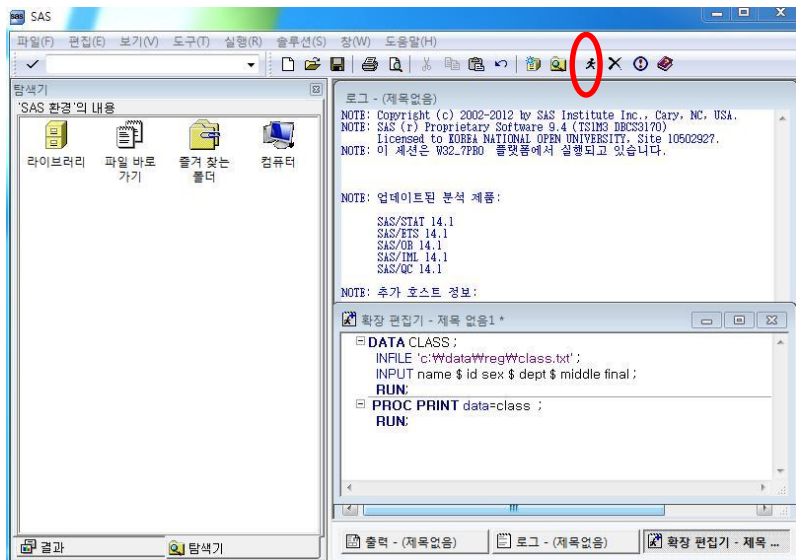
class - 메모장

파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)
Minyoung	1	F	Stat	50 50
Minji	2	F	Computer	50 50
Kongjui	6	F	Computer	20 40
Patjui	3	F	Stat	20 45
Heungboo	5	M	Computer	25 25
Nolboo	4	M	Stat	29 28

c:\wdata\wreg\class.txt

- ✓ 이 자료를 읽어 데이터 출력하기
- ✓ 중간시험과 기말시험 및 두 점수 합의 평균, 표준편차 구하기
- ✓ 학과 별로 구분하여 평균, 표준편차 구하기

# 데이터 출력하기



✓ SAS 화면

✓ 실행결과

```
DATA CLASS ;  
  INFILE 'c:\wdata\wreg\class.txt' ;  
  INPUT name $ id sex $ dept $ middle final ;  
RUN;  
PROC PRINT data=class ;  
RUN;
```

✓ 프로그램



OBS	name	id	sex	dept	middle	final
1	Minyoung	1	F	Stat	50	50
2	Minji	2	F	Computer	50	50
3	Kongjui	6	F	Computer	20	40
4	Patjui	3	F	Stat	20	45
5	Heungboo	5	M	Computer	25	25
6	Nolboo	4	M	Stat	29	28

# 중간, 기말 및 두 합의 평균, 표준편차 구하기

```
DATA CLASS ;  
  INFILE 'c:\data\reg\class.txt' ;  
  INPUT name $ id sex $ dept $ middle final ;  
RUN;  
PROC PRINT data=class ;  
RUN;  
DATA NCLASS ;  
  SET CLASS ;  
  total = middle + final ;  
RUN;  
PROC MEANS data=nclass ;  
  var middle final total ;  
RUN;
```

✓ 실행결과

SAS

파일(F) 편집(E) 보기(V) 이동(G) 도구(T) 솔루션(S) 창(W) 도움말(H)

결과

결과

Print: SAS 시스템

Means: SAS 시스템

Results Viewer - SAS Output

1	Minyoung	1	F	Stat	50	50
2	Minji	2	F	Computer	50	50
3	Kongjui	6	F	Computer	20	40
4	Patjui	3	F	Stat	20	45
5	Heungboo	5	M	Computer	25	25
6	Nolboo	4	M	Stat	29	28

SAS 시스템

MEANS 프로시저

변수	N	평균	표준편차	최솟값	최댓값
middle	6	32.3333333	14.0949163	20.0000000	50.0000000
final	6	39.6666667	10.8934231	25.0000000	50.0000000
total	6	72.0000000	22.2261108	50.0000000	100.0000000

결과

탐색기

출력 - (재확인)...

로그 - (재확인)...

확장 편집기 - ...

Results View...

# 학과별로 구분하여 평균, 표준편차 구하기(출력 자리수 조정)

```
PROC MEANS fw=6 data=nclass ;  
    class dept ;  
    var middle final total ;  
RUN;  
PROC SORT out=sclass ;  
    BY dept;  
RUN;  
PROC MEANS fw=6 data=sclass;  
    by dept ;  
    var middle final total ;  
RUN;
```

SAS 시스템

MEANS 프로시저

dept	관측값 수	변수	N	평균	표준편차	최솟값	최댓값
Computer	3	middle	3	31.67	16.07	20.00	50.00
		final	3	38.33	12.58	25.00	50.00
		total	3	70.00	26.46	50.00	100.0
Stat	3	middle	3	33.00	15.39	20.00	50.00
		final	3	41.00	11.53	28.00	50.00
		total	3	74.00	22.87	57.00	100.0

✓ class dept 실행결과

SAS 시스템

MEANS 프로시저

dept=Computer

변수	N	평균	표준편차	최솟값	최댓값
middle	3	31.67	16.07	20.00	50.00
final	3	38.33	12.58	25.00	50.00
total	3	70.00	26.46	50.00	100.0

dept=Stat

변수	N	평균	표준편차	최솟값	최댓값
middle	3	33.00	15.39	20.00	50.00
final	3	41.00	11.53	28.00	50.00
total	3	74.00	22.87	57.00	100.0

✓ by dept 실행결과



## SAS 활용 예-2

tscore - 메모장

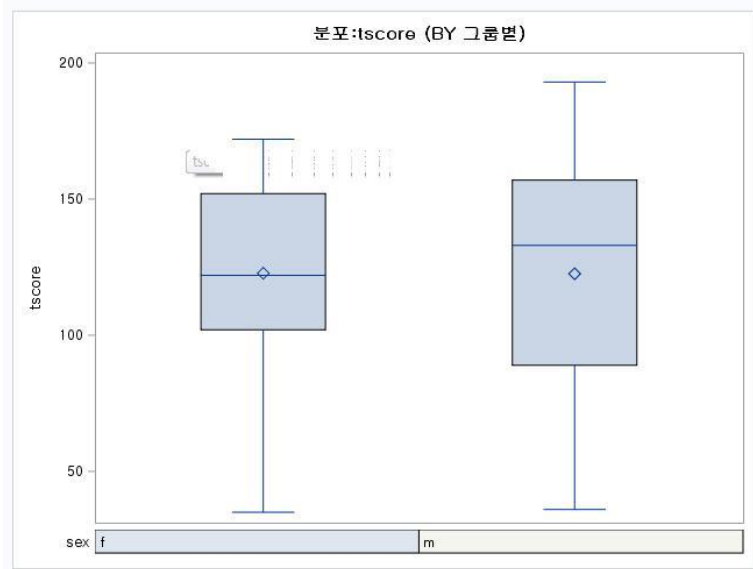
파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)							
13001	f	60	80	13003	m	90	72	13013	m	76	85
13020	f	74	89	13041	f	50	53	13065	m	86	64
70007	m	82	83	70011	m	100	93	70013	m	82	91
70028	m	77	86	70031	m	100	89	70032	m	80	87
70001	f	59	63	70003	f	33	33	70004	m	90	65
70008	m	83	52	70016	m	64	70	70019	m	72	64
70020	m	46	20	70021	m	90	87	70022	m	62	51
70025	m	80	86	70026	m	87	70	70028	m	75	89
70031	f	77	80	70032	m	32	38	70045	f	100	65
70048	m	85	31	70049	f	84	88	70052	f	59	52
70056	m	88	54	70039	m	22	63	70040	f	84	68
70001	m	65	19	70002	m	45	21	70003	f	64	84
70004	f	47	26	70005	f	67	87	70006	f	52	45
70007	m	66	40	70008	m	52	75	70009	f	50	40
70010	m	29	7	70011	m	41	48	70012	m	75	47
70013	m	66	13	70014	m	62	80	70017	m	70	58
70018	f	71	47	70019	f	50	48	70021	m	67	36

```
DATA class2 ;  
  INFILE 'c:\wdata\reg\wtscore.txt' ;  
  INPUT id sex $ midterm final @@;  
RUN;  
DATA tclass2 ;  
  SET class2 ;  
  tscore = midterm + final ;  
RUN;
```

- ✓ 남녀별로 midterm + final 합이 히스토그램,  
상자그림, 줄기-잎 그림 그리기

# 히스토그램 줄기-잎 그림 상자그림 그리기 (그룹기준 sex)

```
DATA class2 ;  
  INFILE 'c:\data\wreg\tscore.txt' ;  
  INPUT id sex $ midterm final @@;  
RUN;  
DATA tclass2 ;  
  SET class2 ;  
  tscore = midterm + final ;  
RUN;  
PROC SORT data=tclass2 OUT=tsclass2;  
  by sex;  
RUN;  
PROC UNIVARIATE data=tsclass2 PLOT ;  
  var tscore ;  
  by sex;  
RUN;
```



✓ 결과 (일부)

# SAS 활용 예-3 : 범주형변수의 라벨

```
data test;
  input region party @@;
datalines;
  1 2 1 3 1 4 2 2 2 3 2 4 3 1 3 2 3 3 1 3 1 2 2
  3 2 3 2 4 3 4
run;
proc print data=test(obs=3);
run;
proc format;
  value regname
    1="seoul"
    2="busan"
    3="others" ;
  value partyname
    1 = "A정당"
    2 = "B정당"
    3 = "C정당"
    4 = "D 정당" ;
run;
```

```
proc freq data=test;
  tables region party;
  tables region*party / CHISQ ;
  format region regname. ;
  format party partyname. ;
run;
```

# SAS 활용 예-3 : 범주형변수의 라벨 (출력결과)

SAS 시스템

FREQ 프로시저

region	빈도	백분율	누적 빈도	누적 백분율
seoul	5	33.33	5	33.33
busan	6	40.00	11	73.33
others	4	26.67	15	100.00

party	빈도	백분율	누적 빈도	누적 백분율
A정당	1	6.67	1	6.67
B정당	4	26.67	5	33.33
C정당	6	40.00	11	73.33
D 정당	4	26.67	15	100.00

빈도  
백분율  
행 백분율  
칼럼 백분율

테이블: region \* party

region	party				
	A정당	B정당	C정당	D 정당	합계
seoul	0	2	2	1	5
	0.00	13.33	13.33	6.67	33.33
	0.00	40.00	40.00	20.00	
	0.00	50.00	33.33	25.00	
busan	0	1	3	2	6
	0.00	6.67	20.00	13.33	40.00
	0.00	16.67	50.00	33.33	
	0.00	25.00	50.00	50.00	
others	1	1	1	1	4
	6.67	6.67	6.67	6.67	26.67
	25.00	25.00	25.00	25.00	
	100.00	25.00	16.67	25.00	
합계	1	4	6	4	15
	6.67	26.67	40.00	26.67	100.00

region \* party 테이블에 대한 통계량

통계량	자유도	값	Prob
카이제곱	6	3.8750	0.6936
우도비 카이제곱	6	3.7833	0.7060
Mantel-Haenszel 카이제곱	1	0.1716	0.6787
파이 계수		0.5083	
우발성 계수		0.4531	
크래머의 V		0.3594	
WARNING: 100%개의 셀이 5보다 적은 기대빈도를 가지고 있습니다. 카이제곱 검정은 올바르지 않을 수 있습니다.			

표본 크기 = 15

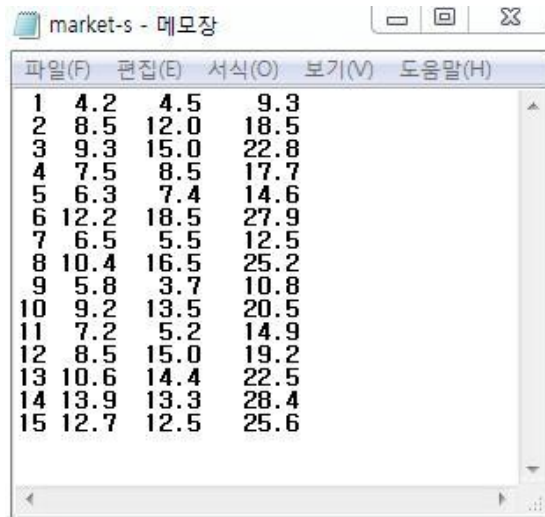
## 2 SAS 회귀모형

---

# 자료

## <표본상점의 총판매액 자료>

상점번호	광고료 (단위:100만원)	상점의 크기 (단위:10 )	총판매액 (단위:1000만원)
1	4.2	4.5	9.3
2	8.5	12.0	18.5
3	9.3	15.0	22.8
4	7.5	8.5	17.7
5	6.3	7.4	14.6
6	12.2	18.5	27.9
7	6.5	5.5	12.5
8	10.4	16.5	25.2
9	5.8	3.7	10.8
10	9.2	13.5	20.5
11	7.2	5.2	14.9
12	8.5	15.0	19.2
13	10.6	14.4	22.5
14	13.9	13.3	28.4
15	12.7	12.5	25.6



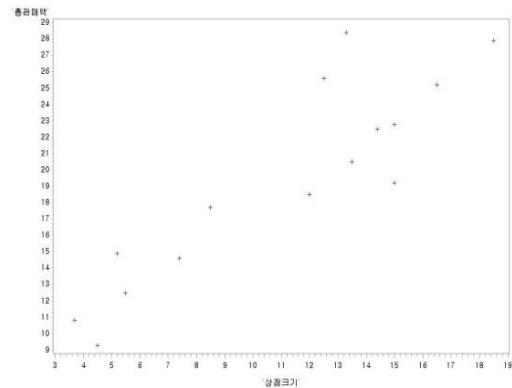
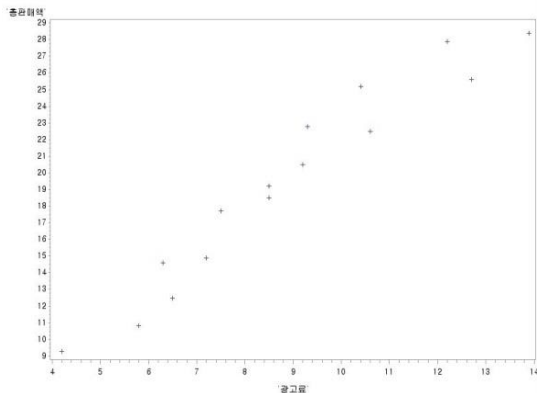
market-s - 데모장

파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)
1	4.2	4.5	9.3	
2	8.5	12.0	18.5	
3	9.3	15.0	22.8	
4	7.5	8.5	17.7	
5	6.3	7.4	14.6	
6	12.2	18.5	27.9	
7	6.5	5.5	12.5	
8	10.4	16.5	25.2	
9	5.8	3.7	10.8	
10	9.2	13.5	20.5	
11	7.2	5.2	14.9	
12	8.5	15.0	19.2	
13	10.6	14.4	22.5	
14	13.9	13.3	28.4	
15	12.7	12.5	25.6	

c:\data\reg\market-s.txt

# 1) 산점도 그리기

```
DATA market2 ;  
  INFILE "c:\data\reg\market-s.txt" ;  
  INPUT ID X1 X2 Y ;  
  LABEL  X1='광고료'  
         X2='상점크기'  
         Y='총판매액' ;  
  
RUN;  
PROC GPLOT ;  
  PLOT Y*X1 ;  
  PLOT Y*X2 ;  
RUN;
```



## 2) 중회귀모형 적합하기

```
PROC REG;
```

```
MODEL Y=X1 X2/I STB;
```

```
OUTPUT OUT=RES P=PRED R=RESID;
```

```
RUN;
```

```
PROC PRINT DATA=RES;
```

```
RUN;
```



# 중회귀모형 출력결과 : 회귀계수, 분산분석표

## SAS 시스템

The REG Procedure  
Model: MODEL1  
Dependent Variable: Y '총판매액'

Number of Observations Read	15
Number of Observations Used	15

### X'X Inverse, Parameter Estimates, and SSE

Variable	Label	Intercept	X1	X2	Y
Intercept	Intercept	0.8248328342	-0.098104723	0.0100049224	0.8504050674
X1	'광고료'	-0.098104723	0.0252062844	-0.011334282	1.5581063354
X2	'상점크기'	0.0100049224	-0.011334282	0.0081880293	0.4273559072
Y	'총판매액'	0.8504050674	1.5581063354	0.4273559072	10.418474877

### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	507.87753	253.93876	292.49	<.0001
Error	12	10.41847	0.86821		
Corrected Total	14	518.29600			

Root MSE	0.93178	R-Square	0.9799
Dependent Mean	19.36000	Adj R-Sq	0.9765
Coeff Var	4.81289		

### Parameter Estimates

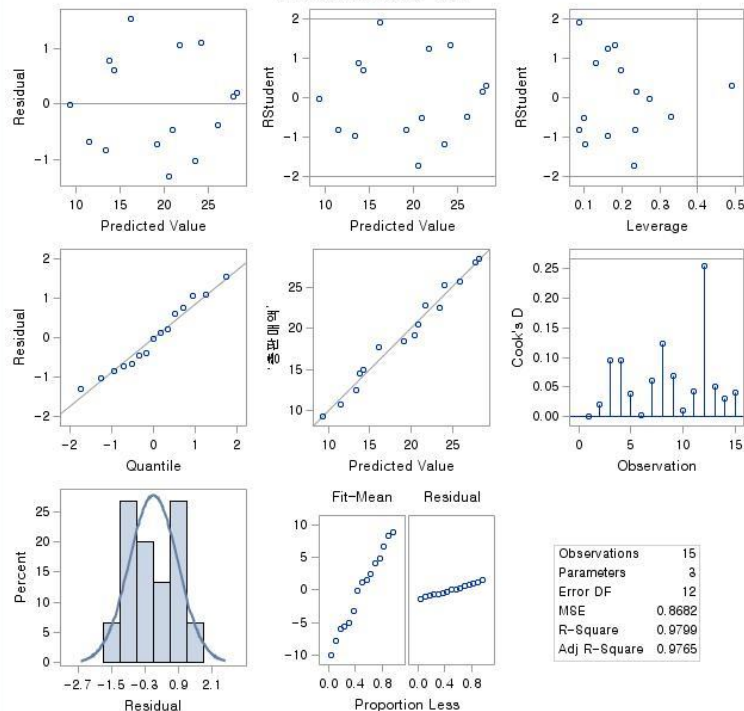
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Standardized Estimate
Intercept	Intercept	1	0.85041	0.84624	1.00	0.3348	0
X1	'광고료'	1	1.55811	0.14793	10.53	<.0001	0.70156
X2	'상점크기'	1	0.42736	0.08431	5.07	0.0003	0.33761

# 중회귀모형 출력결과 : 모형진단 그림

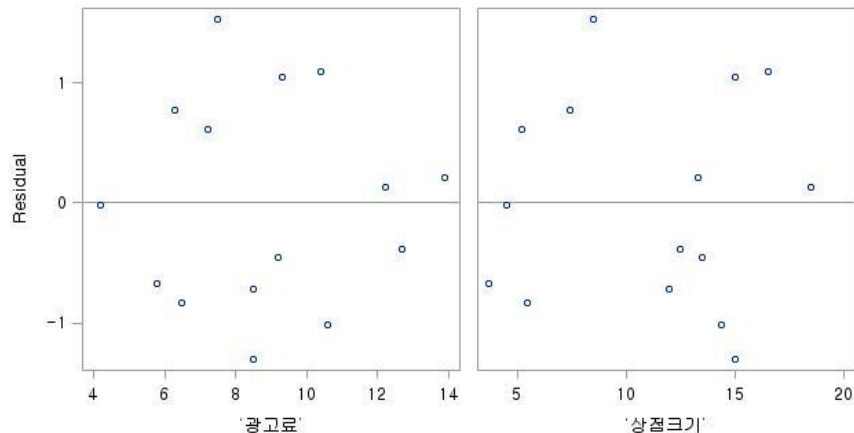
SAS 시스템

The REG Procedure  
Model: MODEL1  
Dependent Variable: Y '총판매액'

Fit Diagnostics for Y



Residual by Regressors for Y



### 3) 잔차 출력

```
PROC REG;  
  MODEL Y=X1 X2/I STB;  
  OUTPUT OUT=RES P=PRED R=RESID;  
RUN;  
  
PROC PRINT DATA=RES;  
RUN;
```

SAS 시스템

OBS	ID	X1	X2	Y	PRED	RESID
1	1	4.2	4.5	9.3	9.3176	-0.01755
2	2	8.5	12.0	18.5	19.2226	-0.72258
3	3	9.3	15.0	22.8	21.7511	1.04887
4	4	7.5	8.5	17.7	16.1687	1.53127
5	5	6.3	7.4	14.6	13.8289	0.77109
6	6	12.2	18.5	27.9	27.7654	0.13461
7	7	6.5	5.5	12.5	13.3286	-0.82855
8	8	10.4	16.5	25.2	24.1061	1.09392
9	9	5.8	3.7	10.8	11.4686	-0.66864
10	10	9.2	13.5	20.5	20.9543	-0.45429
11	11	7.2	5.2	14.9	14.2910	0.60898
12	12	8.5	15.0	19.2	20.5046	-1.30465
13	13	10.6	14.4	22.5	23.5203	-1.02026
14	14	13.9	13.3	28.4	28.1919	0.20808
15	15	12.7	12.5	25.6	25.9803	-0.38030

## 4) 잔차의 합

```
PROC REG;  
  MODEL Y=X1 X2/I STB;  
  OUTPUT OUT=RES P=PRED R=RESID;  
RUN;  
  
PROC PRINT DATA=RES;  
RUN;  
  
DATA RESPRED;  
  SET RES;  
  RP = PRED*RESID;  
  X1R = X1*RESID;  
  X2R = X2*RESID;  
RUN;  
  
PROC MEANS DATA=RESPRED;  
  VAR RP X1R X2R;  
RUN;
```

### SAS 시스템

#### MEANS 프로시저

변수	N	평균	표준편차	최솟값	최댓값
RP	15	6,229091E-14	16,7334503	-26,7513377	26,3700442
X1R	15	3,410605E-14	7,4299836	-11,0895040	11,4845415
X2R	15	1,900702E-14	10,5569403	-19,5697129	18,0496235

### **3** SPSS 사용법

---

# SPSS 소개

✓ 미국 시카고(Chicago) 대학의 National Opinion Research Center 에서 설문조사, 시장조사 등의 표본조사 데이터 분석을 중심으로 SPSS가 개발됨. 2009년 IBM이 인수하여 IBM SPSS 로 명명됨.

- 엑셀과 같은 스프레드 시트를 취하고 있어 자료의 처리가 용이
- GUI 방식으로 되어 있어 데이터분석, 처리가 용이
- 일반사용자들도 쉽게 사용
- 편리한 그래프 기능

✓ Real Easy Real Stat

# SPSS 활용 예

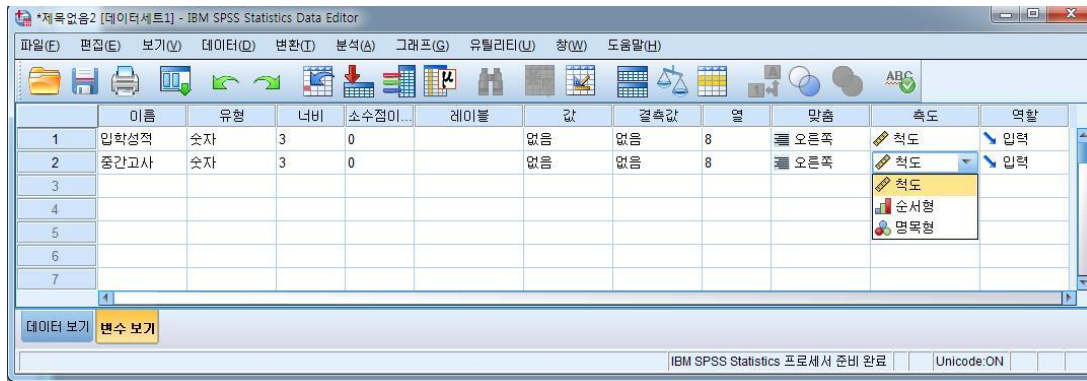
대학교 1학년생 20명을 임의 추출하여 이들의 대학 입학성적과 1학년 1학기 중간고사 성적과의 관계를 분석하고자 한다. 자료는 다음과 같다.

입학성적	170 147 166 125 182 133 146 125 130 179 174 128 152 157 174 185 171 102 150 192
중간고사	698 518 725 485 745 538 485 625 471 798 645 578 625 558 698 745 611 458 538 778

- ✓ 두 변수의 평균, 표준편차 등 기술통계량 구하기
- ✓ 두 변수의 산점도 그리기
- ✓ 두 변수의 상관계수 구하기

# 자료 입력 및 변수정의

✓ 하단의 **변수보기**를 누른 후, 이름, 측도 등을 정의



✓ **데이터보기**를 누른 후, 자료 입력

IBM SPSS Statistics Data Editor - Data View

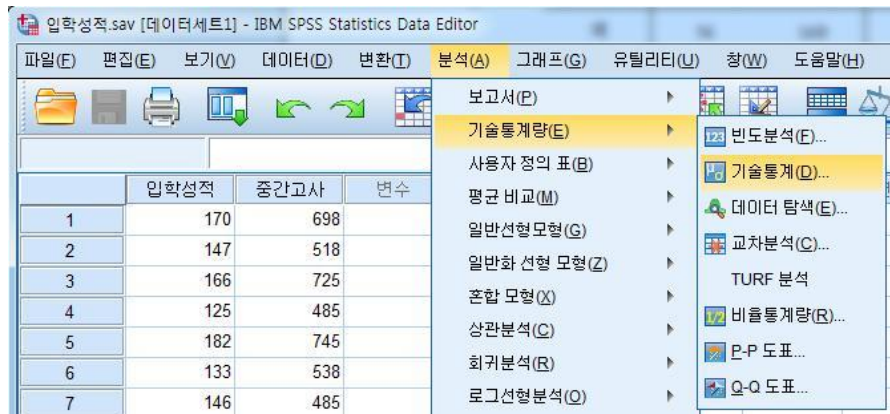
	입학성적	중간고사	변수	변수	변수
1	170	698			
2	147	518			
3	166	725			
4	125	485			
5	182	745			
6	133	538			
7	146	485			
8	125	625			
9	130	471			
10	179	798			
11	174	645			
12	128	578			
13	152	625			
14	157	558			
15	174	698			
16	185	745			
17	171	611			
18	102	458			
19	150	538			
20	192	778			

데이터 보기 변수 보기



# 기술통계량 구하기

## ✓ 분석 - 기술통계량 - 기술통계 선택



기술통계

## ✓ 변수를 선택한 후, 확인



## ✓ 출력 결과

[데이터세트1] C:\data\statpackage\입학성적.sav

기술통계량

	N	최소값	최대값	평균	표준편차
입학성적	20	102	192	154.40	24.624
중간고사	20	458	798	616.10	109.038
유효 N(목록별)	20				

# 산점도 그리기

✓ 그래프-레거시 대화상자-산점도/점도표 선택

IBM SPSS Statistics Data Editor

파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) **그래프(G)** 유틸리티(U) 창(W) 도움말(H)

20. 입학성적 192

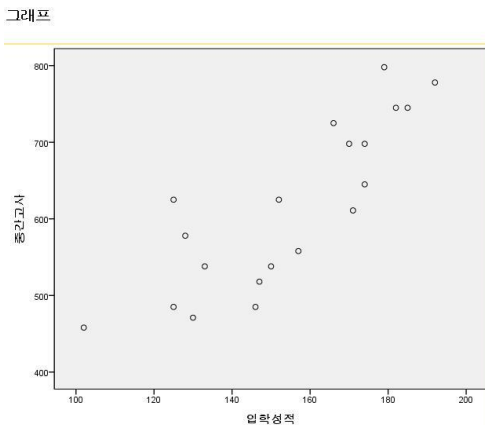
	입학성적	중간고사	변수	변수
1	170	698		
2	147	518		
3	166	725		
4	125	485		
5	182	745		
6	133	538		
7	146	485		
8	125	625		
9	130	471		
10	179	798		
11	174	645		
12	128	578		
13	152	625		
14	157	558		
15	174	698		

도표 작성기(C)...  
 그래프보드 양식 선택기(S)...  
 Weibull 도표...  
 부집단 비교  
 회귀 변수 도표  
**레거시 대화 상자(L)**  
 막대도표(B)...  
 3차원 막대도표...  
 선도표(L)...  
 영역도표(A)...  
 원도표(E)...  
 상한-하한 도표(H)...  
 상자도표(Q)...  
 오차막대도표(O)...  
 인구 피라미드(P)...  
**산점도/점도표(S)...**  
 히스토그램(H)...

산점도/점도표

단순 산점도 행렬 산점도 단순 점도표  
 겹쳐그리기 산점도 3차원 산점도

정의 취소 도움말



단순 산점도

Y축: 중간고사  
 X축: 입학성적

표식 기준(S):  
 케이스 레이블 기준(C):

패널 기준  
 행(R):  
 변수 중첩(변 없음)(N)  
 열(C):  
 변수 중첩(변 없음)(E)

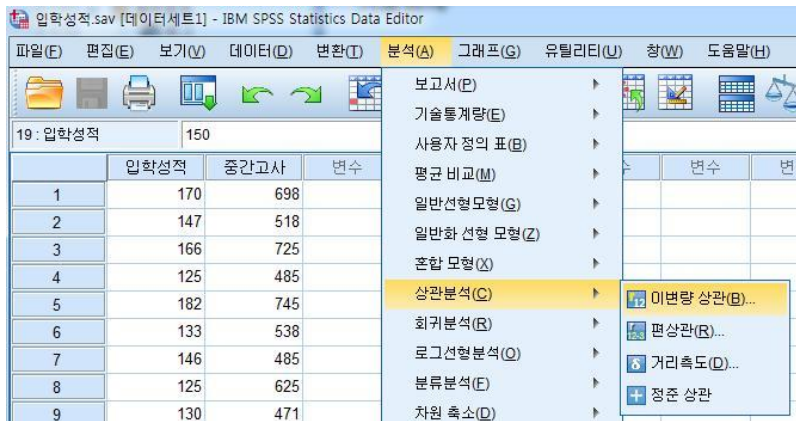
양식  
☒ 도표 양식으로 적용할 파일(F):  
 파일(F)...

확인 붙여넣기(P) 재설정(R) 취소 도움말

✓ 출력 결과

# 상관계수 구하기

## ✓ 분석 - 상관분석 - 이변량 상관 선택



상관관계

## ✓ 출력 결과

상관관계

	입학성적	중간고사
입학성적	1	.839**
Pearson 상관		
유의확률 (양측)		.000
N	20	20
중간고사	.839**	1
Pearson 상관		
유의확률 (양측)	.000	
N	20	20

\*\* 상관계수가 0.01 수준에서 유의합니다(양측).

## ✓ 변수를 선택한 후, 확인



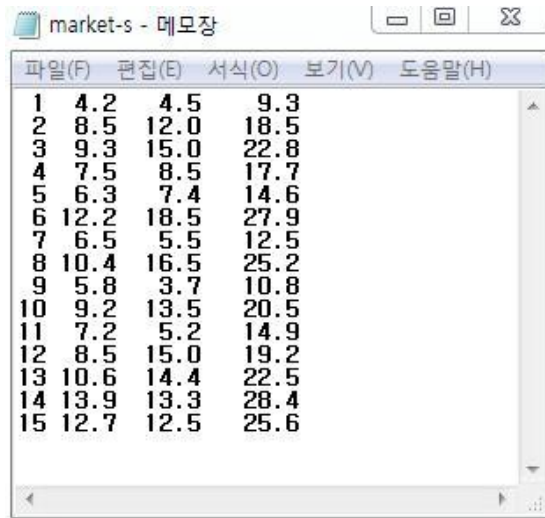
## **4** SPSS 회귀모형

---

# 자료

## <표본상점의 총판매액 자료>

상점번호	광고료 (단위:100만원)	상점의 크기 (단위:10 )	총판매액 (단위:1000만원)
1	4.2	4.5	9.3
2	8.5	12.0	18.5
3	9.3	15.0	22.8
4	7.5	8.5	17.7
5	6.3	7.4	14.6
6	12.2	18.5	27.9
7	6.5	5.5	12.5
8	10.4	16.5	25.2
9	5.8	3.7	10.8
10	9.2	13.5	20.5
11	7.2	5.2	14.9
12	8.5	15.0	19.2
13	10.6	14.4	22.5
14	13.9	13.3	28.4
15	12.7	12.5	25.6



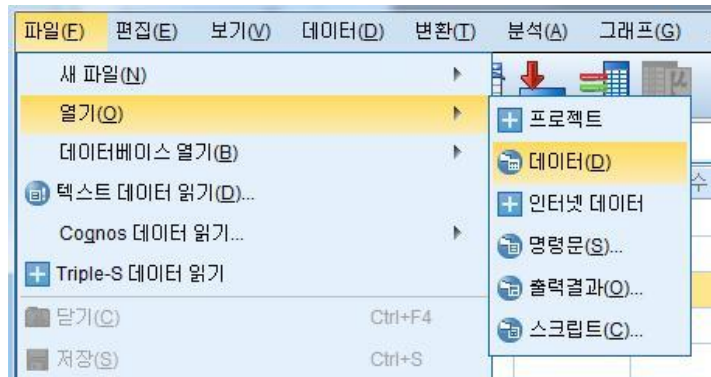
market-s - 데모장

파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)
1	4.2	4.5	9.3	
2	8.5	12.0	18.5	
3	9.3	15.0	22.8	
4	7.5	8.5	17.7	
5	6.3	7.4	14.6	
6	12.2	18.5	27.9	
7	6.5	5.5	12.5	
8	10.4	16.5	25.2	
9	5.8	3.7	10.8	
10	9.2	13.5	20.5	
11	7.2	5.2	14.9	
12	8.5	15.0	19.2	
13	10.6	14.4	22.5	
14	13.9	13.3	28.4	
15	12.7	12.5	25.6	

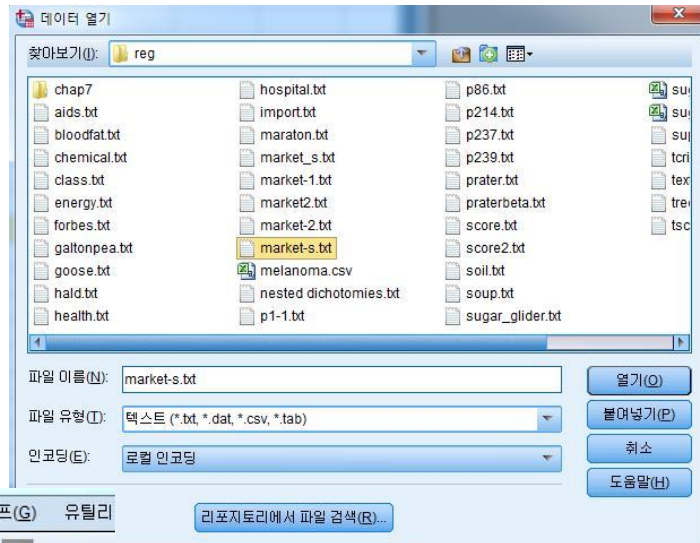
c:\data\reg\market-s.txt

# 1) 자료 불러오기

## ✓ 파일 - 열기 - 데이터 선택



## ✓ 데이터 열기



## ✓ 하단의 변수보기를 눌러 변수이름 등을 정의

	ID	광고료	상점크기	총판매액	변수
1	1	4.2	4.5	9.3	
2	2	8.5	12.0	18.5	
3	3	9.3	15.0	22.8	
4	4	7.5	8.5	17.7	
5	5	6.3	7.4	14.6	

## 2) 중회귀모형 적합하기

✓ 분석 - 회귀분석 - 선형 선택



✓ 종속변수, 독립변수를 선택한 후, 확인



# 중회귀모형 출력결과 : 회귀계수, 분산분석표

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준 오차
1	.990 <sup>a</sup>	.980	.977	.9318

a. 예측자: (상수), 상점크기, 광고료

ANOVA<sup>a</sup>

모형		제곱합	자유도	평균제곱	F	유의확률
1	회귀	507.878	2	253.939	292.487	.000 <sup>b</sup>
	잔차	10.418	12	.868		
	전체	518.296	14			

a. 종속변수: 총판매액

b. 예측자: (상수), 상점크기, 광고료

계수<sup>a</sup>

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	.850	.846		1.005	.335
	광고료	1.558	.148	.702	10.532	.000
	상점크기	.427	.084	.338	5.069	.000

a. 종속변수: 총판매액





다음시간 안내

## 7강. 변수선택