<8-1>

```
DATA MYLIB.EX8_1;
INPUT x y @@;
LABEL x='디지털 측정기' y='혈액채취';
CARDS;
0.150 0.154 0.090 0.082 0.110 0.078
0.100 0.085 0.090 0.072 0.120 0.097
0.900 0.079 0.090 0.080 0.100 0.088
0.140 0.144 0.095 0.090 0.060 0.053
0.080 0.078 0.040 0.050 0.080 0.072
;
RUN;

PROC REG DATA=MYLIB.EX8_1;
        MODEL y = x;
        PLOT y*x;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y 혈액채취

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00002357	0.00002357	0.03	0.8706
Error	13	0.01110	0.00085406		
Corrected Total	14	0.01113			

Root MSE	0.02922	R-Square	0.0021
Dependent Mean	0.08680	Adj R-Sq	-0.0746
Coeff Var	33.66863		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.08587	0.00939	9.15	<.0001
x	디지털 측정기	1	0.00620	0.03730	0.17	0.8706

<8-3>

```
DATA MYLIB.EX8_3;
INPUT x1 x2 x3 x4 y @@;
LABEL x1='나이' x2='경과기간' x3='몸무게' x4='복부 피부두께' y='최고혈압';
CARDS;
21 1 71.0 12.7 170 38 18 59.5 7.7 114
22 6 56.5 8.0 120 38 11 61.0 4.0 136
24 5 56.0 4.3 125 38 11 57.0 3.0 126
24 1 61.0 4.3 148 39 21 57.5 5.0 124
25 1 65.0 20.7 140 39 24 74.0 15.7 128
27 19 62.0 5.7 106 39 14 72.0 13.3 134
28 5 53.0 8.0 120 41 25 62.5 8.0 112
28 25 53.0 0.0 108 41 32 68.0 11.3 128
31 6 65.0 10.0 124 41 5 63.4 13.7 134
32 13 57.0 6.0 134 42 12 68.0 10.7 128
33 13 66.5 8.3 116 43 25 69.0 6.0 140
33 10 59.1 10.3 114 43 26 73.0 5.7 138
34 15 64.0 7.0 130 43 10 64.0 7.0 118
35 18 69.5 7.0 118 44 19 65.0 7.7 110
35 2 64.0 6.7 138 44 18 71.0 4.3 142
36 12 56.5 11.7 134 45 10 60.2 3.3 134
36 15 57.0 6.0 120 47 1 55.0 4.0 116
37 16 55.0 7.0 120 50 43 70.0 11.7 132
37 17 57.0 11.7 114 54 40 87.0 11.3 152
38 10 58.0 13.0 124
;
RUN;

PROC REG DATA=MYLIB.EX8_3;
        MODEL y = x1 x2 x3 x4;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y 최고혈압

Number of Observations Read	39
Number of Observations Used	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2791.32176	697.83044	6.34	0.0006
Error	34	3740.11414	110.00336		
Corrected Total	38	6531.43590			

Root MSE	10.48825	R-Square	0.4274
Dependent Mean	127.41026	Adj R-Sq	0.3600
Coeff Var	8.23187		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	51.51366	17.29240	2.98	0.0053
x1	나이	1	-0.15304	0.28178	-0.54	0.5906
x2	경과기간	1	-0.53129	0.22197	-2.39	0.0224
x3	몸무게	1	1.43708	0.31458	4.57	<.0001
x4	복부 피부두께	1	-0.17484	0.47127	-0.37	0.7129

PROC REG 결과 X1과 x4의 p-value가 유의수준보다 크므로, 변수선택을 실시한다.

```
PROC REG DATA=MYLIB.EX8_3;
        MODEL y = x1 x2 x3 x4 / STB
selection=backward;
RUN;
```

Backward Elimination: Step 1

Variable x4 Removed: R-Square = 0.4250 and C(p) = 3.1376

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2776.18060	925.39353	8.62	0.0002
Error	35	3755.25529	107.29301		
Corrected Total	38	6531.43590			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	52.82117	16.71960	1070.86703	9.98	0.0033
x1	-0.14097	0.27643	27.90208	0.26	0.6133
x2	-0.51868	0.21663	615.07568	5.73	0.0221
x3	1.38360	0.27615	2693.40484	25.10	<.0001

Bounds on condition number: 1.6934, 13.961

Backward Elimination: Step 2

Variable x1 Removed: R-Square = 0.4208 and C(p) = 1.3913

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2748.27852	1374.13926	13.08	<.0001
Error	36	3783.15738	105.08770		
Corrected Total	38	6531.43590			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	50.31913	15.81839	1063.39240	10.12	0.0030
x2	-0.57184	0.18794	972.89889	9.26	0.0044
x3	1.35408	0.26722	2698.29454	25.68	<.0001

Bounds on condition number: 1.3013, 5.205

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

Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4	복부 피부두께	3	0.0023	0.4250	3.1376	0.14	0.7129
2	x1	나이	2	0.0043	0.4208	1.3913	0.26	0.6133

The REG Procedure
Model: MODEL1
Dependent Variable: y 최고혈압

Number of Observations Read	39
Number of Observations Used	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2748.27852	1374.13926	13.08	<.0001
Error	36	3783.15738	105.08770		
Corrected Total	38	6531.43590			

Root MSE	10.25123	R-Square	0.4208
Dependent Mean	127.41026	Adj R-Sq	0.3886
Coeff Var	8.04584		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	50.31913	15.81839	3.18	0.0030	0
x2	경과기간	1	-0.57184	0.18794	-3.04	0.0044	-0.44026
x3	몸무게	1	1.35408	0.26722	5.07	<.0001	0.73320

<8-5>

```
DATA MYLIB.EX8_5;
INPUT IQ score @@;
CARDS;
100 3.0 120 3.8 110 3.1 105 2.9 85 2.6
95 2.9 130 3.6 100 2.8 105 3.1 90 2.4
;
RUN;

PROC REG DATA=MYLIB.EX8_5;
MODEL score=IQ;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.32981	1.32981	39.97	0.0002
Error	8	0.26619	0.03327		
Corrected Total	9	1.59600			

Root MSE	0.18241	R-Square	0.8332
Dependent Mean	3.02000	Adj R-Sq	0.8124
Coeff Var	6.04009		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.05854	0.47199	0.12	0.9044
IQ	1	0.02848	0.00450	6.32	0.0002

(가) 분산분석표에서 p-value가 0.0002 이므로 IQ는 유의미한 설명변수가 된다. 따라서 회귀직선은 $\hat{y} = 0.058454 + 0.02848x$ 이다.

(나) β 에 대한 $100(1-\alpha)\%$ 신뢰구간은 $(\hat{\beta} \pm t_{\frac{\alpha}{2}}(8)\,\hat{\sigma}/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2})$ 이다. $t_{0.025}(8) = 2.306$, $\hat{\sigma} = 0.18241$, $\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{1559}$ 이므로, 신뢰구간은 (0.01783, 0.03913) 이다.

(다) IQ가 125인 학생의 평균의 95% 신뢰구간은 (3.12391, 4.11317) 이다.

(라) 상관계수는 0.9128 이다.

(마) R^2는 0.8332 이므로 EHRFLQ변수로 쫑쑦변수를 약 83% 정도 설명할 수 있다.

<8-7>

```
DATA MYLIB.EX8_7;
INPUT id ver math gpa @@;
CARDS;
1 623 509 2.6 11 490 701 1.2
2 593 611 2.8 12 537 681 2.1
3 584 738 3.0 13 558 602 2.3
4 669 701 2.9 14 578 665 3.0
5 578 635 2.9 15 646 573 2.0
6 520 583 2.8 16 557 674 3.2
7 578 614 3.0 17 597 602 2.4
8 695 634 3.3 18 669 653 2.0
9 613 693 2.3 19 519 529 3.0
10 726 800 3.9 20 653 668 2.8
;
RUN;

PROC REG DATA=MYLIB.EX8_7;
MODEL gpa=ver math/stb;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: gpa

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1,11077	0,55538	1,75	0,2043
Error	17	5,40673	0,31804		
Corrected Total	19	6,51750			

Root MSE	0,56395	R-Square	0,1704
Dependent Mean	2,67500	Adj R-Sq	0,0728
Coeff Var	21,08236		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	0,09493	1,52257	0,06	0,9510	0
ver	1	0,00357	0,00216	1,65	0,1166	0,38193
math	1	0,00068221	0,00195	0,35	0,7309	0,08076

분산분석표에서 p-value가 0.2043으로 유의수준 보다 크므로, 어학능력점수와 수리능력점수는 평균학점의 좋은 설명변수가 되지 못한다. 이들 간에는 유의한 선형관계가 없다고 할 수 있다.

<8-8>

```
DATA MYLIB.EX8_8;
INPUT year GDP @@;
CARDS;
1 309.9 2 323.7 3 324.1 4 355.3 5 383.4 6 395.1
7 412.8 8 407.0 9 438.0 10 446.1 11 452.5 12
447.3
13 475.9 14 487.7 15 497.2 16 529.8 17 551.0 18
581.1
19 617.8 20 658.1 21 675.2 22 706.6 23 725.6 24
722.5
25 745.4 26 790.7
;
RUN;

PROC REG DATA=MYLIB.EX8_8;
MODEL GDP=year/R;
PLOT student.*year;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: GDP

Number of Observations Read	26
Number of Observations Used	26

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	515122	515122	741,75	<,0001
Error	24	16667	694,47213		
Corrected Total	25	531789			

Root MSE	26,35284	R-Square	0,9687
Dependent Mean	517,68462	Adj R-Sq	0,9674
Coeff Var	5,09052		

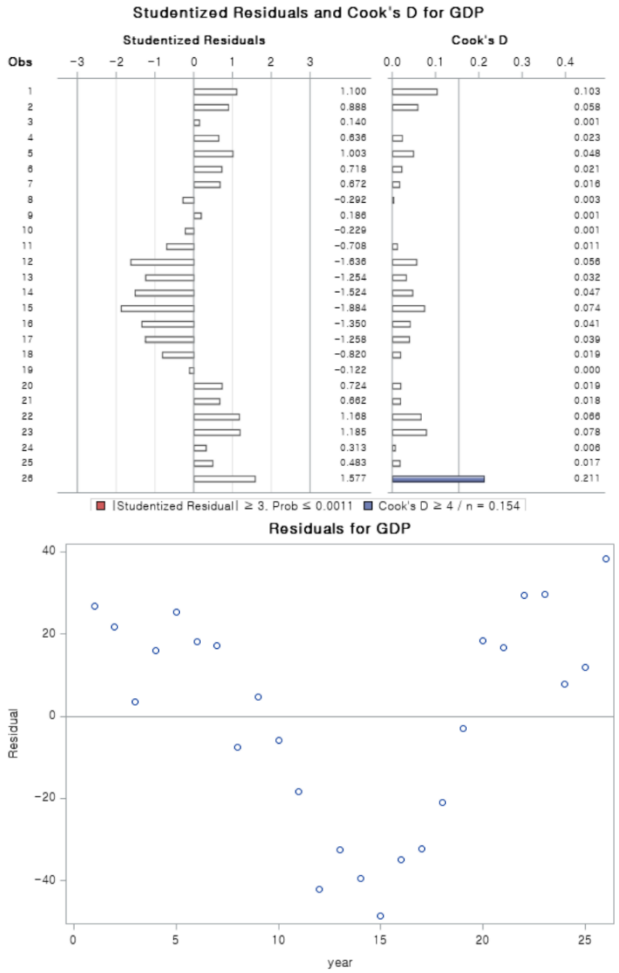
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	264,32308	10,64201	24,84	<,0001
year	1	18,76752	0,68910	27,24	<,0001

The REG Procedure
Model: MODEL1
Dependent Variable: GDP

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D
1	310	283,0906	10,0452	26,8094	24,363	1,100	0,103
2	324	301,8581	9,4610	21,8419	24,596	0,888	0,058
3	324	320,6256	8,8917	3,4744	24,807	0,140	0,001
4	355	339,3932	8,3406	15,9068	24,998	0,636	0,023
5	383	358,1607	7,8114	25,2393	25,169	1,003	0,048
6	395	376,9282	7,3090	18,1718	25,319	0,718	0,021
7	413	395,6957	6,8391	17,1043	25,450	0,672	0,016
8	407	414,4632	6,4090	-7,4632	25,562	-0,292	0,003
9	438	433,2308	6,0271	4,7692	25,654	0,186	0,001
10	446	451,9983	5,7033	-5,8983	25,728	-0,229	0,001
11	453	470,7658	5,4478	-18,2658	25,784	-0,708	0,011
12	447	489,5333	5,2706	-42,2333	25,820	-1,636	0,056
13	476	508,3009	5,1797	-32,4009	25,839	-1,254	0,032
14	488	527,0684	5,1797	-39,3684	25,839	-1,524	0,047
15	497	545,8359	5,2706	-48,6359	25,820	-1,884	0,074
16	530	564,6034	5,4478	-34,8034	25,784	-1,350	0,041
17	551	583,3709	5,7033	-32,3709	25,728	-1,258	0,039
18	581	602,1385	6,0271	-21,0385	25,654	-0,820	0,019
19	618	620,9060	6,4090	-3,1060	25,562	-0,122	0,000
20	658	639,6735	6,8391	18,4265	25,450	0,724	0,019
21	675	658,4410	7,3090	16,7590	25,319	0,662	0,018
22	707	677,2085	7,8114	29,3915	25,169	1,168	0,066
23	726	695,9761	8,3406	29,6239	24,998	1,185	0,078
24	723	714,7436	8,8917	7,7564	24,807	0,313	0,006
25	745	733,5111	9,4610	11,8889	24,596	0,483	0,017
26	791	752,2786	10,0452	38,4214	24,363	1,577	0,211



잔차의 절댓값이 모두 2 이하이지만 표준화잔차의 도표를 보면 일정한 모양의 구조를 띄는 것을 알 수 있다. 따라서 잔차의 등분산성이 위배된다. 그러므로 오차항의 등분산성 또한 위배된다. 그러므로 위의 결과에 의해서 구해진 선형 회귀모형은 유의미한 의미를 가지지 못한다.

<8-9>

```
DATA MYLIB.EX8_9;
INPUT id y age region @@;
CARDS;
1 46 21 1 2 39 21 3 3 62 21 3 4 38 21 2
5 39 21 3 6 70 22 2 7 39 22 2 8 35 22 1
9 41 22 3 10 41 23 2 11 50 23 1 12 71 23 2
13 66 23 3 14 38 24 1 15 68 24 3 16 44 24 3
17 43 24 2 18 44 25 2 19 46 25 3 20 53 25 1
21 41 26 1 22 71 26 3 23 46 26 2 24 76 26 2
25 57 27 1 26 49 28 2 27 58 25 1 28 74 28 3
29 45 28 1 30 48 30 1 31 53 30 2 32 77 30 3
33 79 30 2 34 85 31 2 35 50 31 1 36 56 32 2
37 81 32 3 38 53 33 1 39 88 33 2 40 60 34 2
41 86 35 3 42 93 36 2 43 63 36 2 44 58 36 1
45 64 37 2 46 64 40 1
;
RUN;

DATA MYLIB.EX8_9_1;
        SET MYLIB.EX8_9;
        d1=0; d2=0;
        IF region=1 THEN d1=1;
        IF region=2 THEN d2=1;
        z1=d1*age; z2=d2*age;
RUN;
option ls=80;
PROC REG DATA=MYLIB.EX8_9_1;
MODEL y = age d1 d2 z1 z2;
RUN;

PROC GLM DATA=MYLIB.EX8_9_1;
CLASS region;
MODEL y = age region age*region / SOLUTION;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	46
Number of Observations Used	46

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5584.95405	1116.99081	7.97	<.0001
Error	40	5604.35029	140.10876		
Corrected Total	45	11189			

Root MSE	11.83675	R-Square	0.4991
Dependent Mean	57.56522	Adj R-Sq	0.4365
Coeff Var	20.56234		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-17.71917	19.65892	-0.90	0.3728
age	1	3.08539	0.75897	4.07	0.0002
d1	1	40.08578	25.94616	1.54	0.1302
d2	1	27.35533	24.84715	1.10	0.2775
z1	1	-2.10619	0.96478	-2.18	0.0350
z2	1	-1.27665	0.92401	-1.38	0.1748

The GLM Procedure

Dependent Variable: y

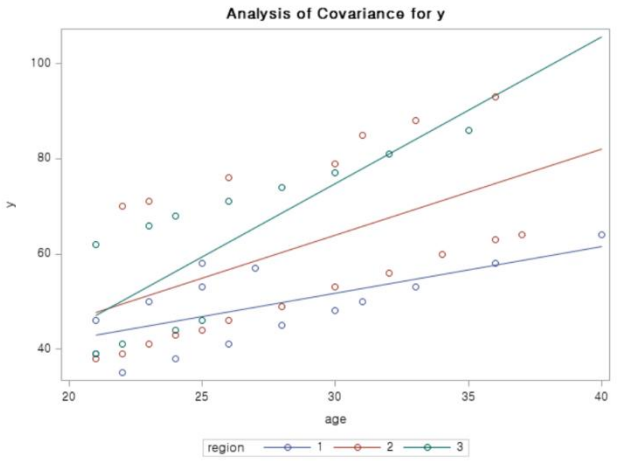
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5584.95405	1116.99081	7.97	<.0001
Error	40	5604.35029	140.10876		
Corrected Total	45	11189.30435			

R-Square	Coeff Var	Root MSE	y Mean
0.499133	20.56234	11.83675	57.56522

Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	3204.952151	3204.952151	22.87	<.0001
region	2	1712.067542	856.033771	6.11	0.0048
age*region	2	667.934360	333.967180	2.38	0.1052

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	3999.125824	3999.125824	28.54	<.0001
region	2	341.106843	170.553421	1.22	0.3068
age*region	2	667.934360	333.967180	2.38	0.1052

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-17.71916509	B	19.65892424	-0.90	0.3728
age	3.08538899	B	0.75896786	4.07	0.0002
region 1	40.08577749	B	25.94616307	1.54	0.1302
region 2	27.35532953	B	24.84714949	1.10	0.2775
region 3	0.00000000	B	.	.	.
age*region 1	-2.10618842	B	0.96478084	-2.18	0.0350
age*region 2	-1.27664525	B	0.92400959	-1.38	0.1748
age*region 3	0.00000000	B	.	.	.



(가) z1과 z2는 region에 따라 같은 집단에 속하는 age값들에 대해서만 회귀분석을 따로 고려해주기 위한 가변수이다. 따라서 z1과 z2를 추가하여 다중회귀분석을 실시하면, region별로 기울기가 같다는 가정 없이 각각의 집단에 대한 기울기를 값을 얻을 수 있다.

(나) 우선, GLM 프로시저의 출력결과에서 분산분석표(Type 3 SS) 중 상호작용 age*region에 대한 F-검정의 결과의 유의확률이 0.1052이므로 상호작용의 효과가 유의하지 않다. 따라서 귀무가설을 기각할 수 있으므로, 기울기의 동일성을 가정할 수 있다. 하지만 REG 프로시저에서는 z1과 z2의 구분에 따른 region 집단에 대한 기울기의 유의성에 대한 검정만을 알 수 있다. 그러므로 region에 따른 기울기의 동일성의 가정은 검정할 수 없다. 따라서 GLM 프로시저의 출력결과에 따라 기울기의 동일성을 가정 할 수 있다.

(다) 〈예 8.8〉 에서는 다중공정성을 검정하고 있으나, 변수 y를 구분하는 집단의 변수가 없으므로 기울기의 동일성에 대한 가정을 검정할 새로운 변수를 만들지는 않아도 된다. 따라서 가장 단순한 선형모델을 선택한다.

```
PROC GLM DATA=MYLIB.EX8_9_1;
CLASS region;
MODEL y = age region age*region / SOLUTION;
RUN;
```