**SAS 데이터 분석 입문 8장**

**2019020650 김형욱**

**\* 8장 예제문제**

**<예 8-1>**

**data** sasadv.adsales;

input company adver sales @@;

cards;

01 11 23 02 19 32 03 23 36 04 26 46 05 56 93

06 62 99 07 29 49 08 30 50 09 38 65 10 39 70

11 46 71 12 49 89

;

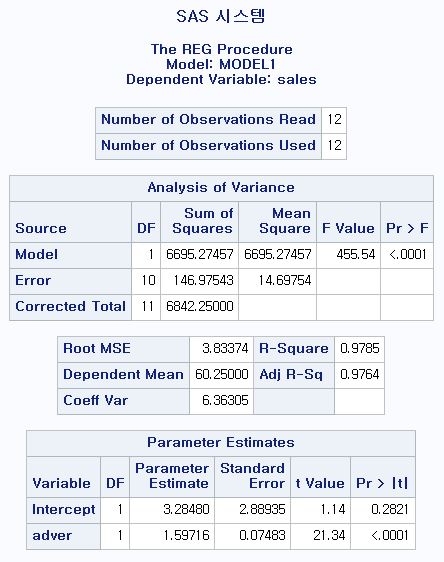
**run**;

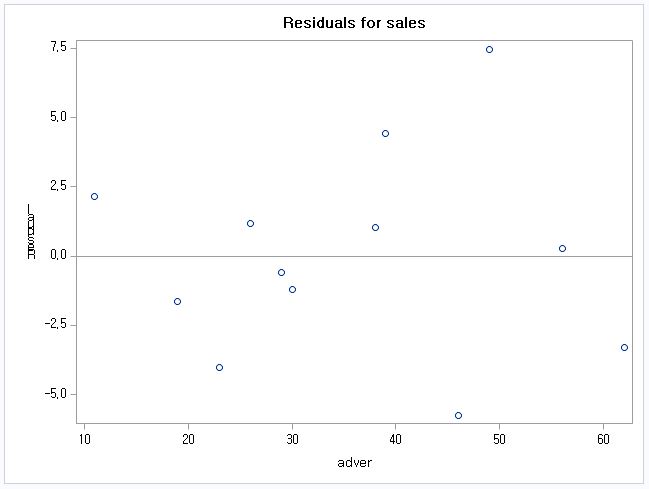
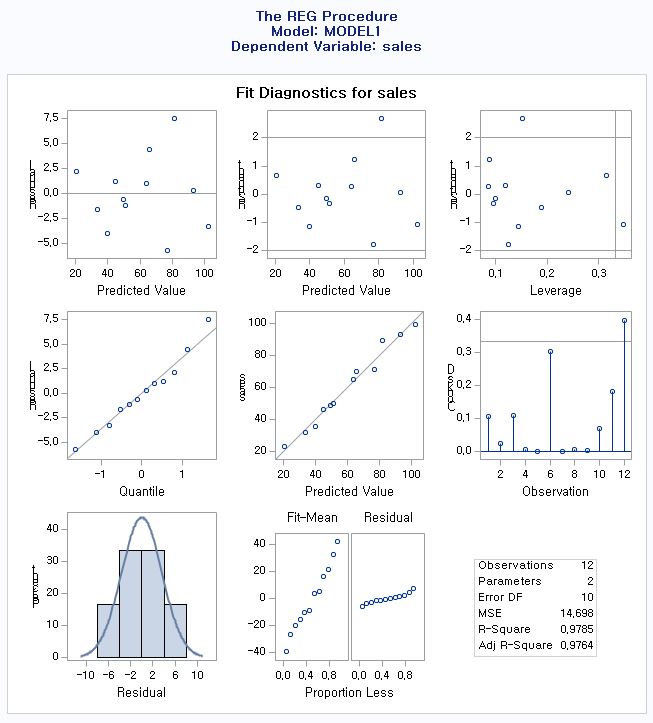
**proc** **reg** data=sasadv.adsales;

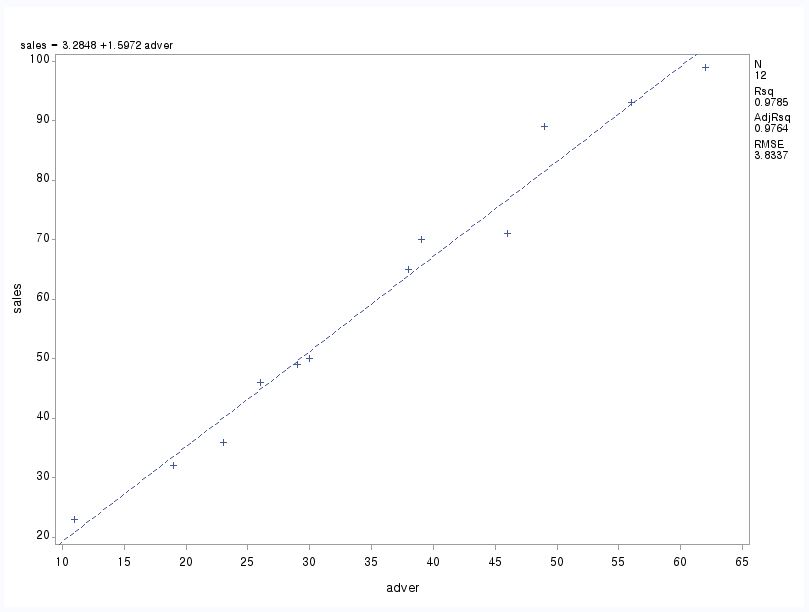
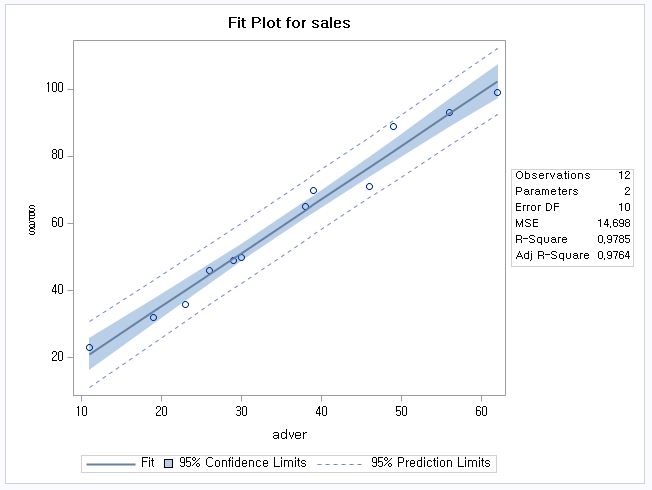
model sales=adver;

plot sales\*adver;

**run**;







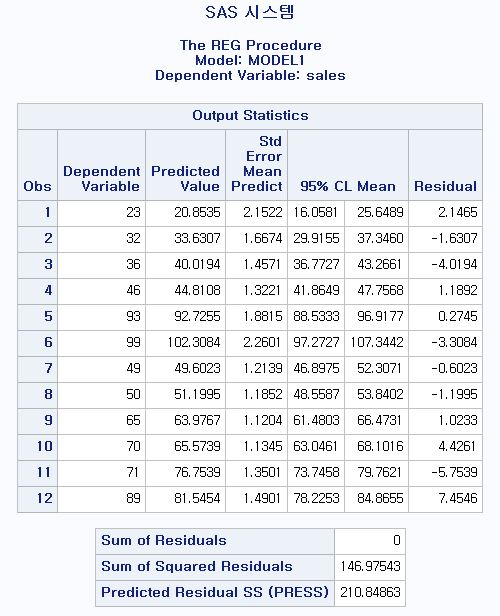
해석 : 선형회귀분석 결과, 추정된 회귀식은 DRW0000274823e1이다. 즉, ‘광고비(x)가 1단위 증가하면 매출액(y)이 1.597단위 증가한다’는 것을 의미한다. 또한 기울기 Beta에 대한 t-값은 21.34이고 유의확률이 0.0001이므로 유의수준 5% 하에서 귀무가설(Beta=0)을 기각한다. 따라서 ‘광고비(x)가 매출액(y)에 유의한 영향을 준다’고 할 수 있다.

**<예 8-2>**

**proc** **reg** data=sasadv.adsales;

model sales=adver / p clm;

**run**;



해석 : 위 결과는 <예 8-1>에서 주어진 ‘매출액’ 데이터에서 독립변수 adver의 각 값에 대응하는 종속변수 sales의 예측값 및 신뢰구간을 구한 결과이다. 출력 결과에서 ‘Predicted Value’는 판매액의 예측값을 나타내고, ‘95% CL Mean’은 판매액의 평균값(기대값)에 대한 95% 신뢰구간을 나타낸다.

**<예 8-3>**

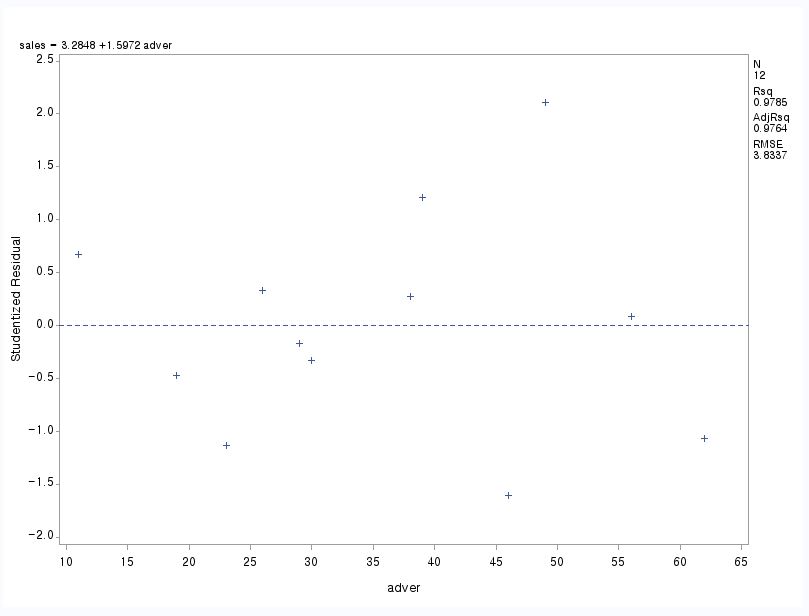
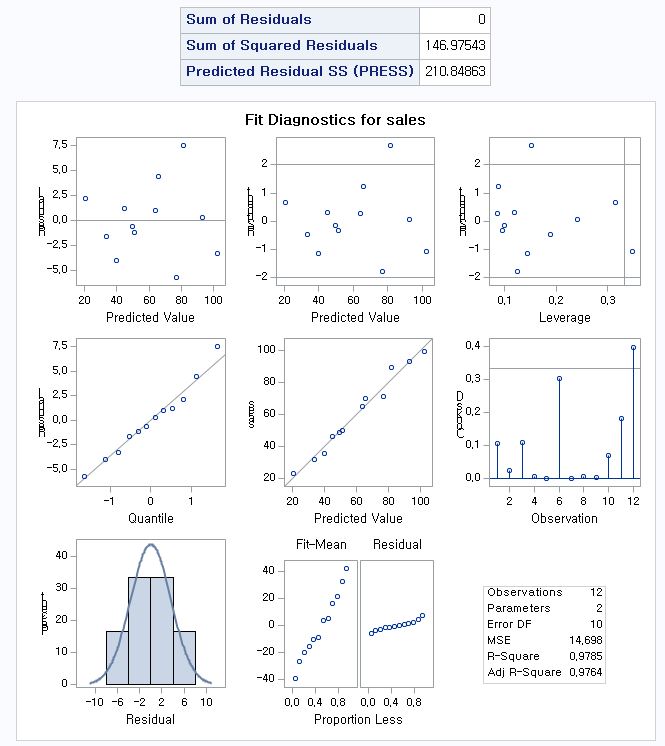
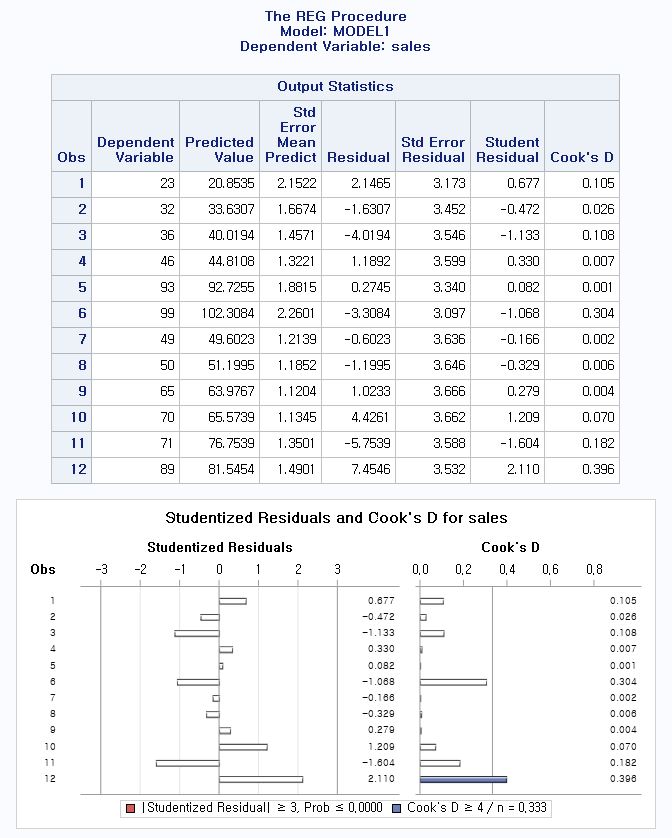
**proc** **reg** data=sasadv.adsales graphics;

model sales=adver / r;

output out=regout student=std\_r;

plot student.\*adver;

**run**;



해석 : <예 8-1>에서 주어진 데이터로 표준화잔차와 잔차도표를 출력해보았다. 출력 결과에서 ‘Student Residual’은 표준화잔차를 나타내는 것으로, 이 결과와 ‘표준화잔차 도표를 살펴보면 절대값이 큰(2 또는 3을 벗어나는) 표준화잔차가 눈에 띄지 않으며 특별한 패턴을 보이고 있지 않다는 것을 알 수 있다.

**<예 8-4>**

**proc** **univariate** data=regout;

histogram std\_r / normal(mu=est sigma=est)

cfill=grey

midpoint=-**3** to **3** by **1**;

inset normal(cvm cvmpval ad adpval);

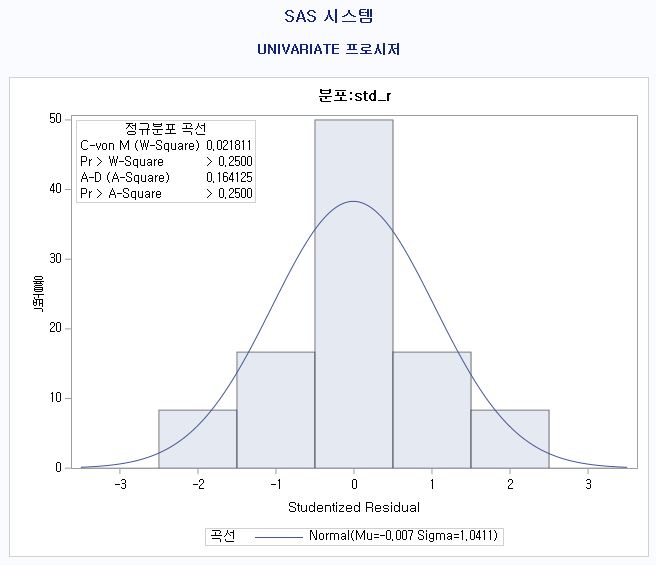
**run**;

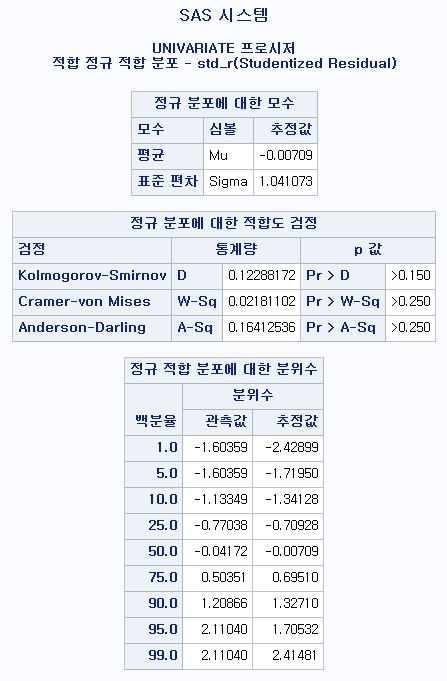
**proc** **univariate** data=regout;

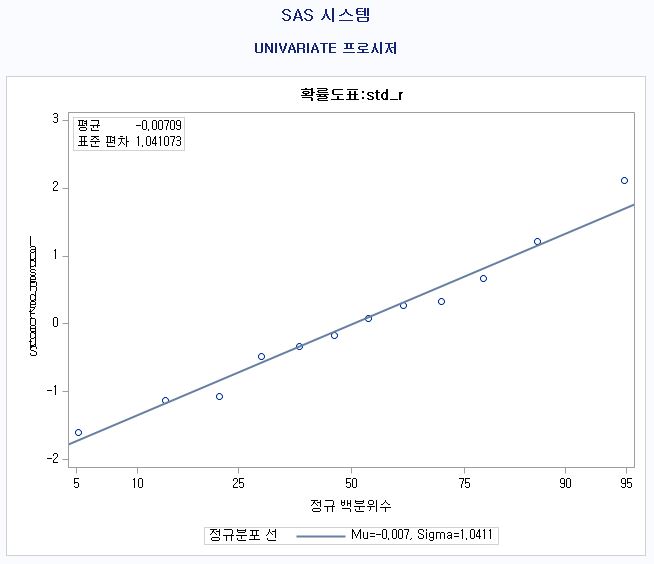
probplot std\_r / normal(mu=est sigma=est);

inset mean std;

**run**;







해석 : 잔차의 정규성을 검토하기 위해, 표준화잔차의 히스토그램과 표준화잔차의 정규확률도표를 그려보았다. 먼저 히스토그램을 살펴보면, 표준화잔차들이 ‘0’을 중심으로 좌우대칭적인 모습을 보이고 있고, 정규분포와 아주 흡사한 형태임을 알 수 있다. 또한 정규확률도표에서도 표준화잔차들이 대체적으로 일직선을 이루고 있어서 정규성 가정에 큰 문제가 없음을 알 수 있다. 그리고 Shapiro-Wilk test나 Kolmogorov-Smirnov test를 이용해서도 잔차의 정규성을 검토할 수 있다. 이 예제에서 사용한 Kolmogorov-Smirnov test 결과를 살펴보면, 유의확률이 0.150보다 크므로 유의수준 5% 하에서 귀무가설(잔차가 정규성을 따른다)을 기각하지 못한다. 따라서 잔차가 정규성을 따른다고 할 수 있다.

**<예 8-5>**

**proc** **reg** data=sasadv.adsales;

model sales=adver / dw;

**run**;



해석 : 잔차의 독립성 검토를 위해 더빈-왓슨 검정을 실시하였다. 그 결과, 더빈-왓슨 통계량 값이 2.470로 비교적 2에 가까워서 잔차의 독립성 가정에 큰 문제가 없음을 알 수 있다.

**<예 8-6>**

**data** sasadv.satisfac;

infile 'F:\SAS 독학\예제로 배우는 SAS 데이터 분석 입문 데이터\satisfaction.txt';

input id **3.** x1 **1.** x2 **1.** x3 **1.** x4 **1.** y **1.**;

label x1 = '디자인'

x2 = '편리성'

x3 = '성능'

x4 = '견고성'

y = '구입의향';

**run**;

**proc** **corr** data=sasadv.satisfac;

var y x1-x4;

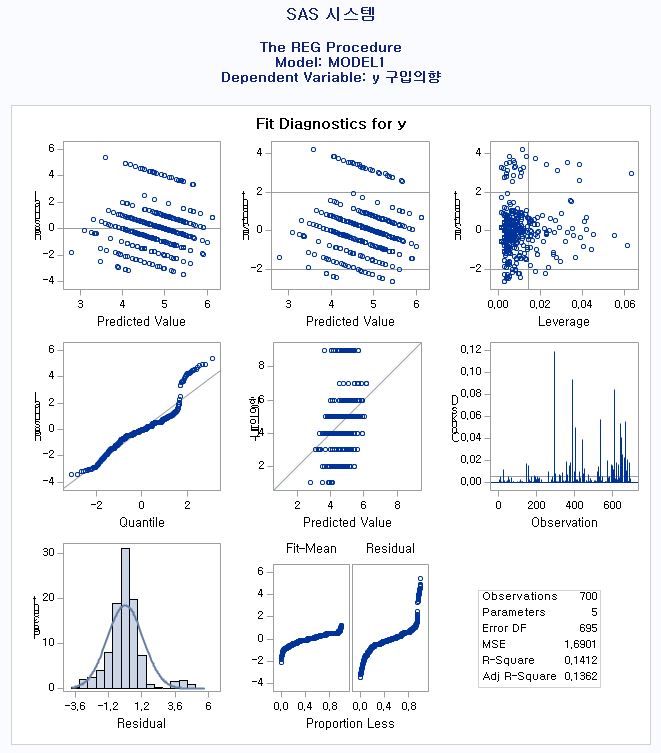
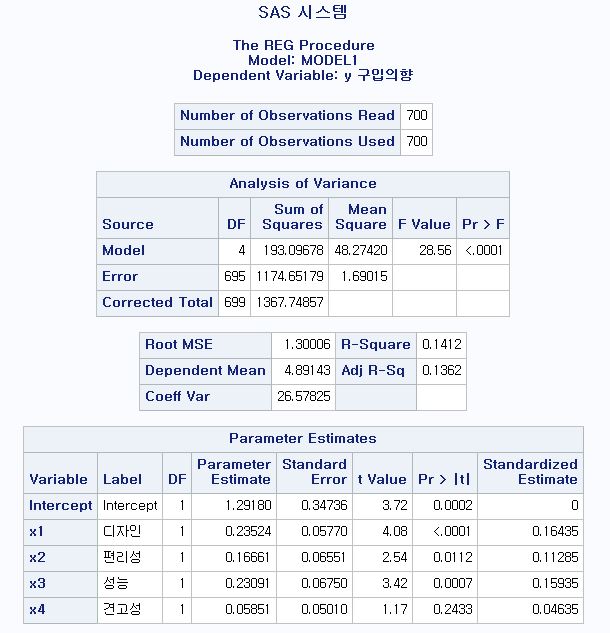
**run**;

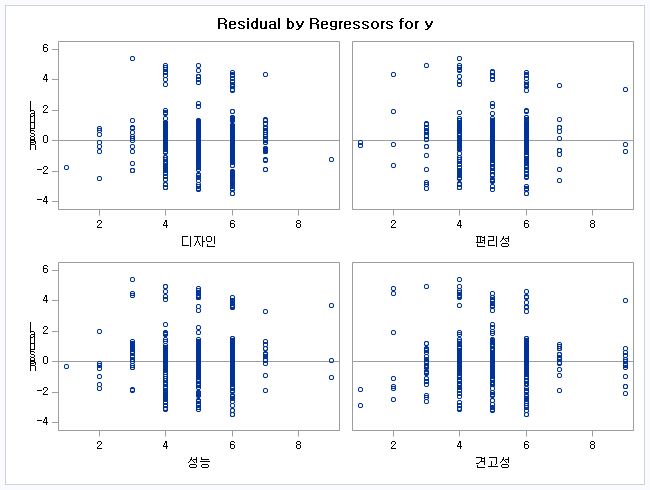
**proc** **reg** data=sasadv.satisfac;

model y=x1-x4 / stb;

**run**;







해석 : 변수 y(구입 의향)을 종속변수로 하고 x1(디자인 만족도), x2(사용 편리성 만족도), x3(성능 만족도), x4(고장 및 견고성 만족도)를 독립변수로 하여 다중 회귀분석을 한 결과이다. 추정된 회귀식의 설명력이 어느 정도 되는지를 보여주는 결정계수 값이 0.141이므로, 4개의 독립변수로 종속변수의 변이 중 약 14% 정도를 설명할 수 있다는 것을 알 수 있다. 또한 분산분석표를 보면, 유의확률이 0.0001보다 작으므로 유의수준 5% 하에서 귀무가설(Beta1=Beta2=Beta3=Beta4=0)을 기각하여 ‘4개의 독립변수 중 적어도 하나 이상의 변수는 종속변수를 설명하는데 유의하게 기여한다’고 할 수 있다. 그리고 각 회귀계수에 대한 t-통계량 및 p-value를 살펴보면, 변수 x4(고장 및 견고성 만족도)에 대한 회귀계수 Beta4만 유의수준 5% 하에서 귀무가설(Beta4=0)을 기각하지 못한다. 따라서 변수 x4(고장 및 견고성 만족도)를 제외한 나머지 3개의 독립변수들은 종속변수를 설명하는데 유의하게 기여한다고 할 수 있다. 분명 상관분석 결과에서는 독립변수 x4(고장 및 견고성 만족도)와 종속변수 y(구입 의향)는 상관계수가 0.207이고 유의확률이 0.0001로 유의하다는 것을 볼 수 있다. 그러나 변수 x4(고장 및 견고성 만족도)와 다른 독립변수들 간에 밀접한 연관성이 존재하여, 변수 x4(고장 및 견고성 만족도)가 종속변수 y(구입 의향)를 설명하는데 유의하게 기여하지 않는다고 결과가 나온 것이다. 이러한 연관성을 고려할 때 다른 독립변수들의 값이 모두 같도록 통제된다면 변수 x4(고장 및 견고성 만족도)는 종속변수 y(구입 의향)에 유의한 영향을 주지 않는다고 해석할 수 있다.

**<예 8-7>**

**data** sasadv.fitness;

input oxygen age weight runtime rstpulse runpulse maxpulse @@;

cards;

44.609 44 89.47 11.37 62 178 182 45.313 40 75.07 10.07 62 185 185

54.297 44 85.84 8.65 45 156 168 59.571 42 68.15 8.17 40 166 172

49.874 38 89.02 9.22 55 178 180 44.811 47 77.45 11.63 58 176 176

45.681 40 75.98 11.95 70 176 180 49.091 43 81.19 10.85 64 162 170

39.442 44 81.42 13.08 63 174 176 60.055 38 81.87 8.63 48 170 186

50.541 44 73.03 10.13 45 168 168 37.388 45 87.66 14.03 56 186 192

44.754 45 66.45 11.12 51 176 176 47.273 47 79.15 10.60 47 162 164

51.855 54 83.12 10.33 50 166 170 49.156 49 81.42 8.95 44 180 185

40.836 51 69.63 10.95 57 168 172 46.672 51 77.91 10.00 48 162 168

46.774 48 91.63 10.25 48 162 164 50.388 49 73.37 10.08 67 168 168

39.407 57 73.37 12.63 58 174 176 46.080 54 79.38 11.17 62 156 165

45.441 52 76.32 9.63 48 164 166 54.625 50 70.87 8.92 48 146 155

45.118 51 67.25 11.08 48 172 172 39.203 54 91.63 12.88 44 168 172

45.790 51 73.71 10.47 59 186 188 50.545 57 59.08 9.93 49 148 155

48.673 49 76.32 9.40 56 186 188 47.920 48 61.24 11.50 52 170 176

47.467 52 82.78 10.50 53 170 172

;

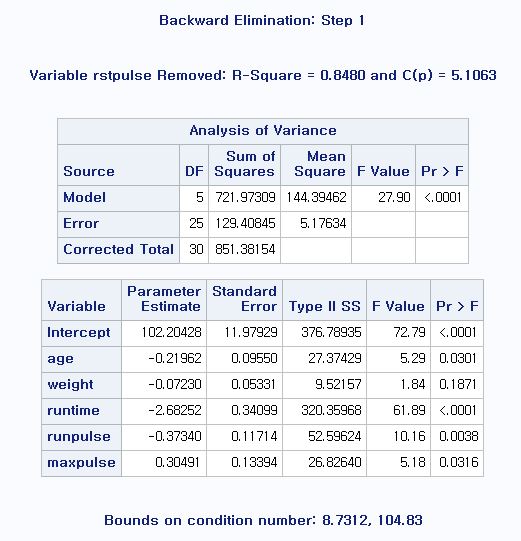
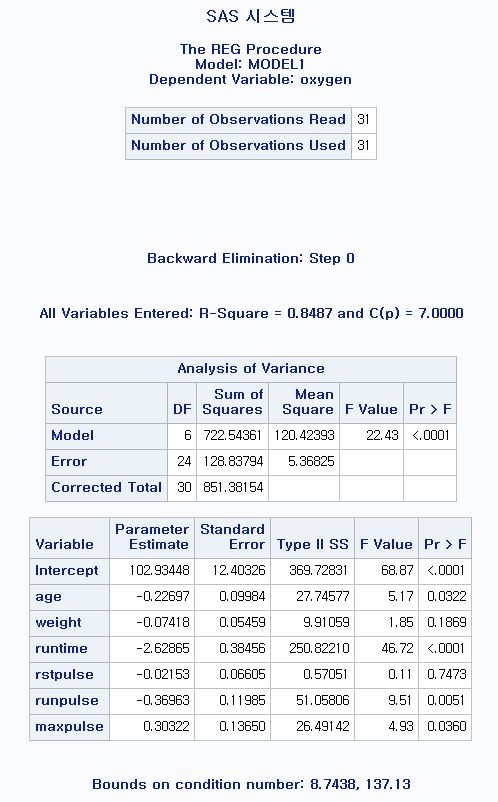
**run**;

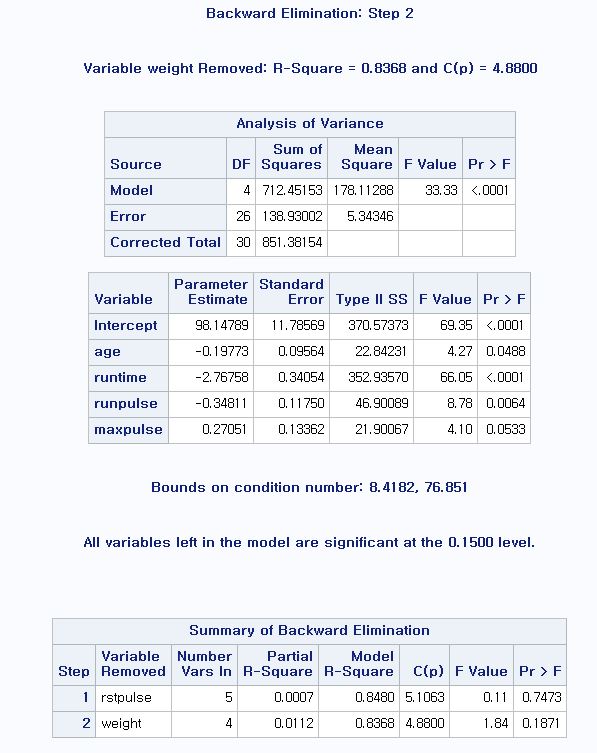
**proc** **reg** data=sasadv.fitness;

model oxygen=age weight runtime rstpulse runpulse maxpulse

/ selection=backward slstay=**0.15**;

**run**;





해석 : 위 예제는 에어로빅 적합성을 알아보기 위해 31명으로부터 oxygen(산소섭취율), age(나이), weight(체중), runtime(1.5마일을 주행하는데 소요된 시간), rstpulse(휴식중의 맥박수), runpulse(주행중의 맥박수), maxpulse(주행중의 최대 맥박수)를 측정하여 얻은 데이터를 사용하였다. 종속변수를 oxygen로 하고 나머지 6개의 변수들을 독립변수로 두었는데, 분석에 필요한 변수들만 선택하기 위해 ‘후진제거법’을 사용하였다. 그 결과, 변수 rstpulse(첫 번째 단계에서 제거)와 변수 weight(두 번째 단계에서 제거)가 제거된 모형이 선택되었다.

**<예 8-8>**

**data** sasadv.multico;

input y x1 x2 x3 x4 @@;

cards;

4.00 4.00 4.00 4.00 4.00 4.00 4.00 4.00 3.00 3.50

4.00 4.00 3.00 4.50 4.00 4.00 4.00 4.00 3.00 3.50

5.00 5.00 5.00 4.00 4.50 4.00 4.00 5.00 3.00 3.50

4.00 4.00 4.00 3.00 3.50 5.00 5.00 4.00 4.00 4.50

4.00 4.00 4.00 3.50 4.00 4.00 4.00 5.00 3.00 3.50

3.00 4.00 4.00 3.00 3.50 4.00 3.00 3.00 3.00 3.00

4.00 4.00 4.00 4.00 4.00 4.00 4.00 4.00 3.00 3.50

4.00 4.00 4.00 3.00 3.50 4.00 4.00 3.00 2.50 3.00

4.00 4.00 3.00 3.00 3.50 4.00 4.00 5.00 5.00 4.50

5.00 4.00 4.00 5.00 4.50 5.00 5.00 4.00 4.00 4.50

;

**run**;

**proc** **reg** data=sasadv.multico;

model y=x1-x4 / vif collin;

**run**;



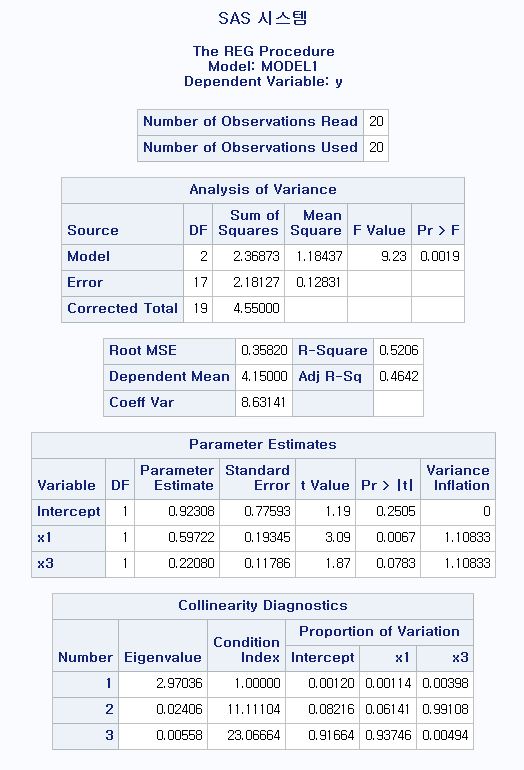
해석 : 위 예제는 다중공선성이 존재하는 데이터를 다중회귀분석한 결과이다. 먼저 분산분석표를 살펴보면, F-통계량에 대한 유의확률이 0.015로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 적어도 하나 이상의 독립변수가 종속변수 y를 설명하는데 기여하고 있다고 할 수 있다. 그러나 각 독립변수들의 t-통계량에 대한 유의확률을 살펴보면, 유의확률이 모두 0.05보다 크므로 어느 변수에 대해서도 통계적으로 유의하지 않다. 이는 다중공선성이 심각하여 회귀계수 추정치의 분산이 매우 커져서 나타나는 일반적인 현상이다. 정말로 다중공선성이 발생했는지에 대한 여부를 확인하기 위해 분산확대인자(VIF)와 상태지수를 살펴보았다. 그 결과, 변수 x3과 x4에 대한 분산확대인자(VIF)의 값이 10보다 크고 변수 x3과 x4에 대한 상태지수 또한 매우 큰 값이 나왔다. 따라서 다중공선성이 존재한다고 할 수 있는데, 가장 큰 상태지수 값인 120.382에 대응되는 분산비율을 살펴보면 변수 x1, x3, x4에 대한 분산비율들이 매우 크므로 이 세 변수들이 밀접한 선형관계를 가지고 있다는 것을 알 수 있다.

**<예 8-9>**

**proc** **reg** data=sasadv.multico;

model y=x1 x3 / vif collin;

**run**;



해석 : 위 예제는 <예 8-8>의 데이터에서 변수 x1과 x3만을 사용하여 다중공선성 분석을 수행한 결과이다. 결과를 살펴보면, 결정계수는 0.5206으로 약간 작아졌으나 수정결정계수는 0.4642로 오히려 좋아졌다. 또한 두 변수에 대한 유의확률이 각각 0.0067, 0.0783이므로 유의수준 10% 하에서 유의하다는 것을 알 수 있다. 그리고 분산확대인자(VIF) 값과 상태지수 또한 작아진 것을 확인할 수 있다.

**<예 8-10>**

**data** sasadv.dummy;

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 28 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** sasadv.dummy1;

set sasadv.dummy;

d1=**0**; d2=**0**;

if region=**1** then d1=**1**;

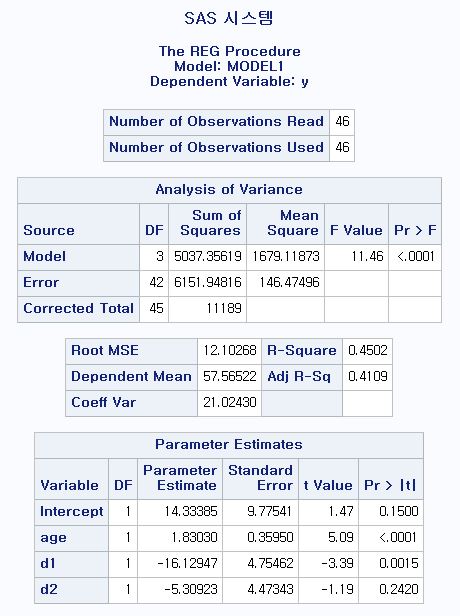
if region=**2** then d2=**1**;

**run**;

**proc** **reg** data=sasadv.dummy1;

model y=age d1 d2;

**run**;



해석 : 위 예제는 지역구분을 나타내는 질적변수 region을 처리하여 가변수 d1과 d2를 생성시킨 가변수 데이터에 대한 분석 결과이다. 변수 age(연령)에 대한 계수 추정치가 1.830이므로 연령이 증가할수록 선호도가 증가한다고 할 수 있으며, 변수 d2에 대한 t-검정의 유의확률이 0.2420이므로 연령이 동일한 경우 ‘중소도시’와 ‘읍면’간의 선호도 차이는 유의하지 않다고 할 수 있다.

**<예 8-11>**

**data** sasadv.mobile;

input year t yt @@;

yts=log(yt);

cards;

1984 1 27 1985 2 47 1986 3 71

1987 4 103 1988 5 204 1989 6 397

1990 7 800 1991 8 1662 1992 9 2719

1993 10 4718 1994 11 9600 1995 12 16410

1996 13 28900 1997 14 45700

;

**run**;

/\* 데이터의 산점도 확인 \*/

**proc** **gplot** data=sasadv.mobile;

plot yt\*t;

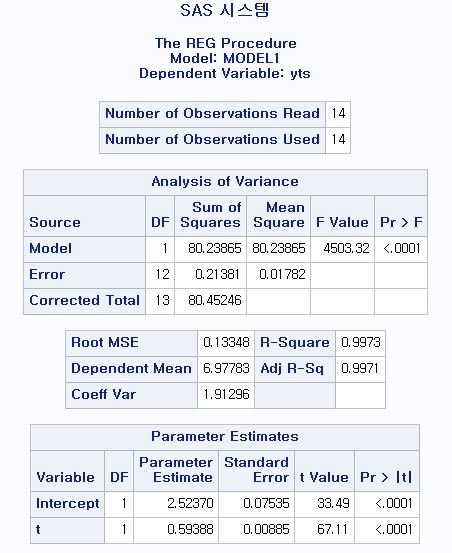
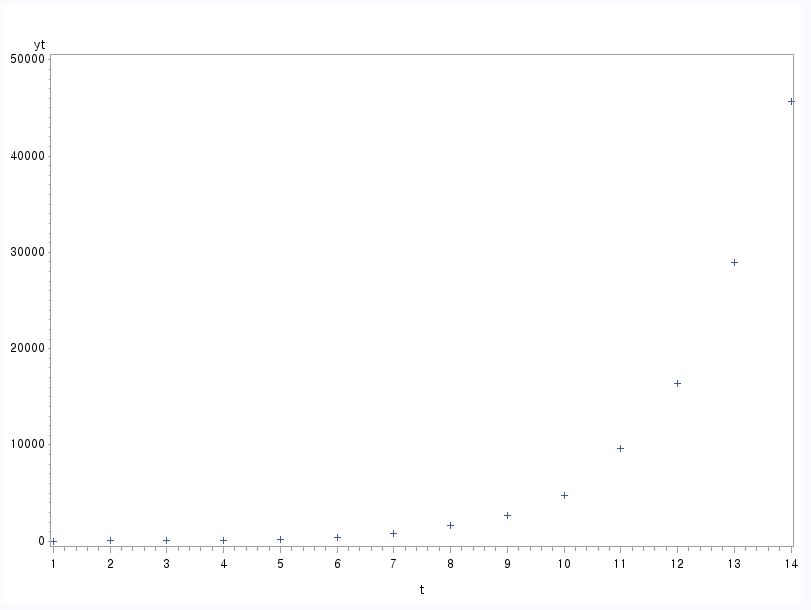
**run**;

/\* 변수변환에 의한 곡선추정 \*/

**proc** **reg** data=sasadv.mobile;

model yts=t;

**run**;



해석 : 주어진 데이터의 산점도를 그려보면, 위와 같이 비선형적인 관계임을 확인할 수 있다. 그러므로 변수변환(여기서는 로그변환)을 통해 곡선을 추정해보았다. 그 결과, 추정된 지수모형은 다음과 같다.

DRW0000274823f5

또한 추정된 지수모형에 대한 결정계수는 0.9973으로 설명력이 매우 높다는 것을 알 수 있다.

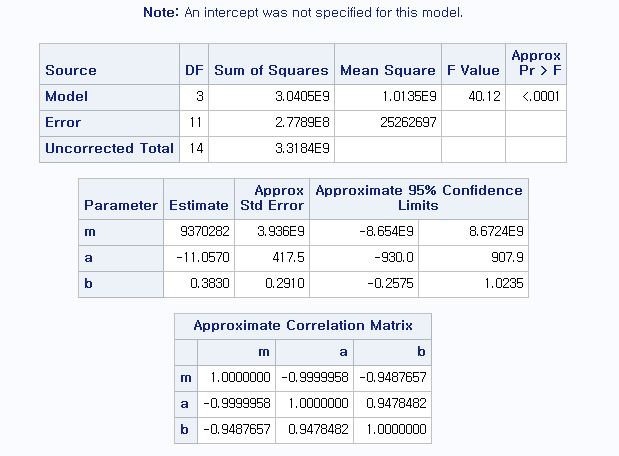
**<예 8-12>**

**proc** **nlin** data=sasadv.mobile;

parms m=**100000** a=-**5** b=**0.5**;

model yt=(m\*exp(a+b\*t)) / (**1**+exp(a+b\*t));

**run**;



해석 : 위 예제는 <예 8-11>의 이동통신 데이터에 대해 비선형 회귀분석을 수행한 결과이다. 여기서는 로지스틱모형을 사용하여 분석했고, 각 모수들에 대한 초기값 또한 지정해주었다. 그 결과 추정된 로지스틱모형 추정식은 다음과 같다.

DRW0000274823ed

이와 같이 비선형 회귀분석을 하는 경우에는 각 모수에 대해 적절한 시작값(초기값)을 지정하는 것이 특히 중요하다.

**\* 8장 연습문제**

**<연습문제 8-1>**

**data** sasadv.ex8\_1;

input x y @@;

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** **corr** data=sasadv.ex8\_1;

var y x;

**run**;

**proc** **reg** data=sasadv.ex8\_1;

model y=x / r dw;

output out=regout8\_1 student=std\_r8\_1;

plot y\*x;

plot student.\*x;

**run**;

/\* 표준화잔차에 대한 히스토그램 및 정규확률도표 \*/

**proc** **univariate** data=regout8\_1;

histogram std\_r8\_1 / normal(mu=est sigma=est)

cfill=grey

midpoint=-**3** to **3** by **1**;

inset normal(cvm cvmpval ad adpval);

**run**;

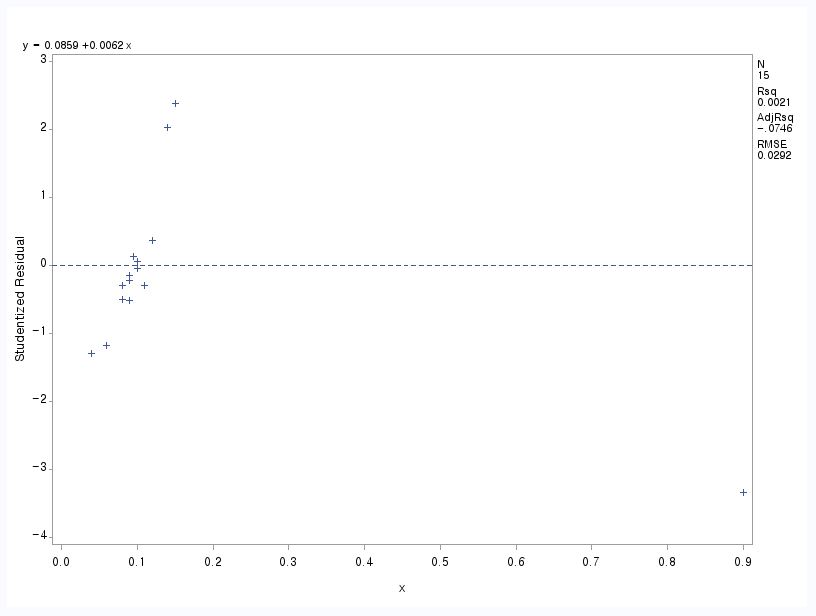
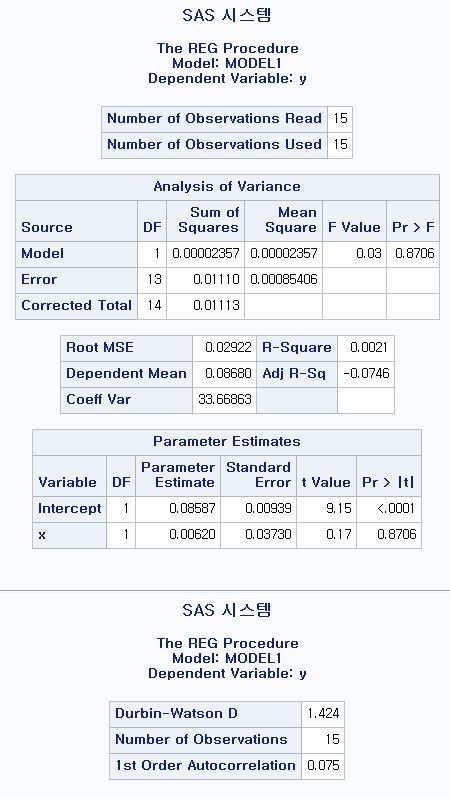
**proc** **univariate** data=regout8\_1;

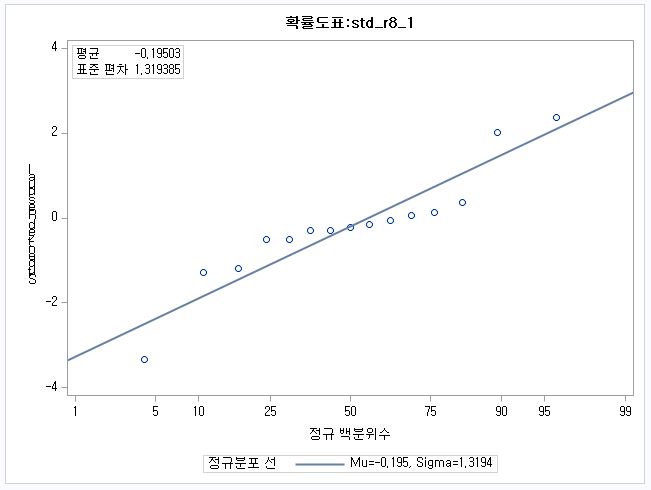
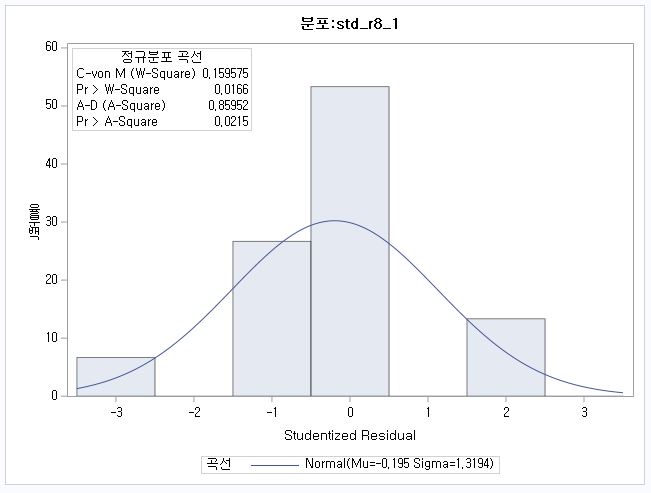
probplot std\_r8\_1 / normal(mu=est sigma=est);

inset mean std;

**run**;







해석 : 위 문제는 15명의 음주 운전자에 대해 혈액채취(y)와 디지털 측정기(x)에 의해 각각 알코올 농도를 측정한 데이터로 회귀분석 및 잔차분석을 수행한 결과이다. 먼저 두 변수 x와 y간의 피어슨 상관계수를 보면 0.04603으로 매우 낮기 때문에 두 변수는 매우 약간 선형적 관계를 보인다고 추론할 수 있다. 그리고 추정된 회귀직선식은 다음과 같다.

DRW000035c444af

다음으로 잔차의 독립성 검토를 위해 더빈-왓슨 검정을 해 본 결과, 더빈-왓슨 통계량 값이 1.424로 비교적 2와 가까우므로 독립성 가정을 만족한다고 볼 수 있다. 두 번째로 잔차의 등분산성 검토를 위해 잔차도표를 그려보니, 표준화잔차가 ‘0’을 기준으로 랜덤하게 퍼져있지 않으므로 등분산성 가정을 만족하지 않는다고 볼 수 있다. 마지막으로 잔차의 정규성 검토를 위해 표준화잔차의 히스토그램과 정규확률도표를 그려보았다. 분포의 모양이 정규분포를 따르지 않고 정규확률도표에서 표준화잔차가 일직선 상에 위치하지 않는 것으로 보아, 정규성 가정을 만족하지 않는다고 볼 수 있다.

**<연습문제 8-3>**

**data** sasadv.ex8\_3;

input x1 x2 x3 x4 y;

cards;

21 1 71.0 12.7 170

22 6 56.5 8.0 120

24 5 56.0 4.3 125

24 1 61.0 4.3 148

25 1 65.0 20.7 140

27 19 62.0 5.7 106

28 5 53.0 8.0 120

28 25 53.0 0.0 108

31 6 65.0 10.0 124

32 13 57.0 6.0 134

33 13 66.5 8.3 116

33 10 59.1 10.3 114

34 15 64.0 7.0 130

35 18 69.5 7.0 118

35 2 64.0 6.7 138

36 12 56.5 11.7 134

36 15 57.0 6.0 120

37 16 55.0 7.0 120

37 17 57.0 11.7 114

38 10 58.0 13.0 124

38 18 59.5 7.7 114

38 11 61.0 4.0 136

38 11 57.0 3.0 126

39 21 57.5 5.0 124

39 24 74.0 15.7 128

39 14 72.0 13.3 134

41 25 62.5 8.0 112

41 32 68.0 11.3 128

41 5 63.4 13.7 134

42 12 68.0 10.7 128

43 25 69.0 6.0 140

43 26 73.0 5.7 138

43 10 64.0 7.0 118

44 19 65.0 7.7 110

44 18 71.0 4.3 142

45 10 60.2 3.3 134

47 1 55.0 4.0 116

50 43 70.0 11.7 132

54 40 87.0 11.3 152

;

**run**;

/\* 다중공선성 진단 \*/

**proc** **corr** data=sasadv.ex8\_3;

var y x1 x2 x3 x4;

**run**;

**proc** **reg** data=sasadv.ex8\_3;

model y=x1 x2 x3 x4 / r dw vif collin;

**run**;

/\* 변수선택 - 단계적 선택법(x2, x3 선택됨) \*/

**proc** **reg** data=sasadv.ex8\_3;

model y=x1 x2 x3 x4 / selection=stepwise;

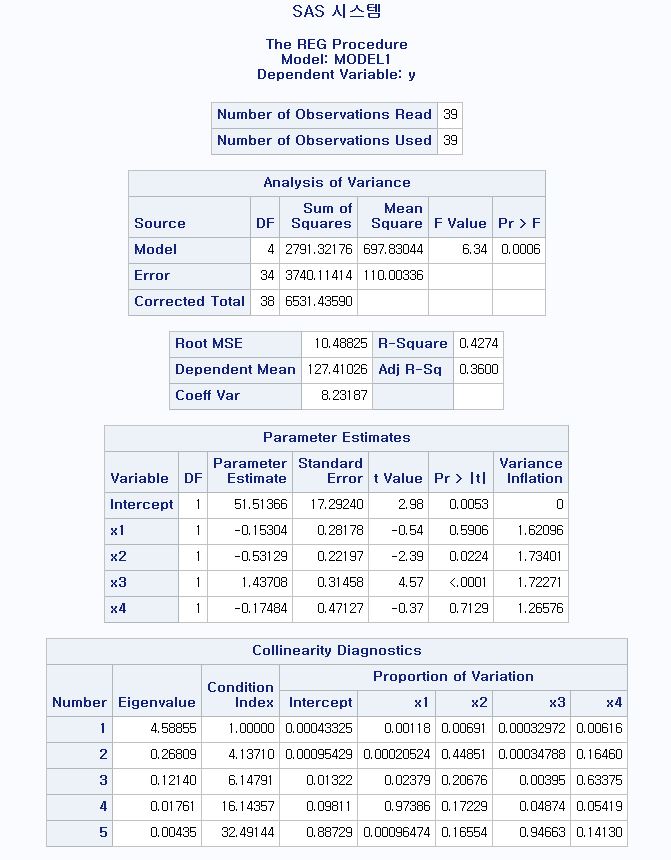
**run**;

**proc** **reg** data=sasadv.ex8\_3;

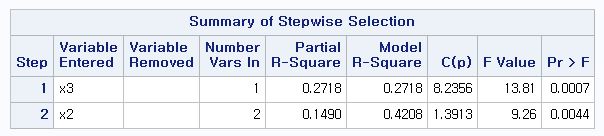
model y=x2 x3 / r dw;

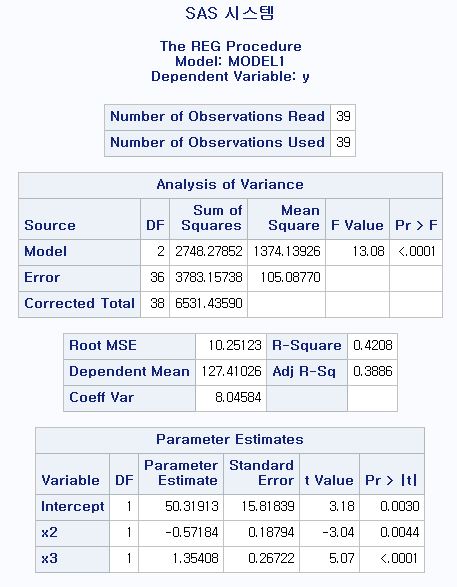
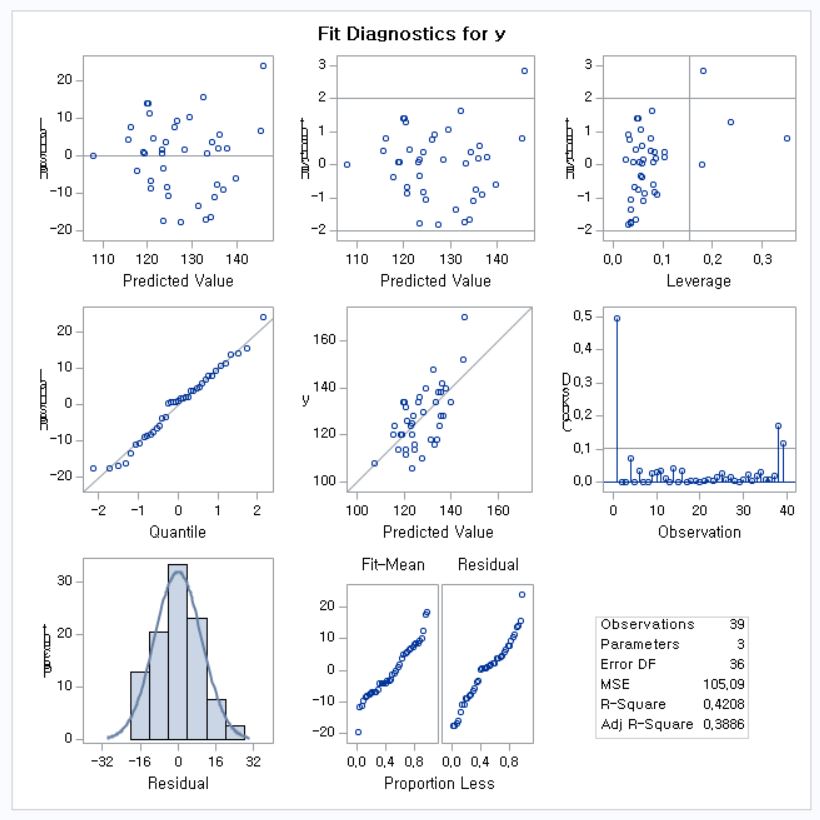
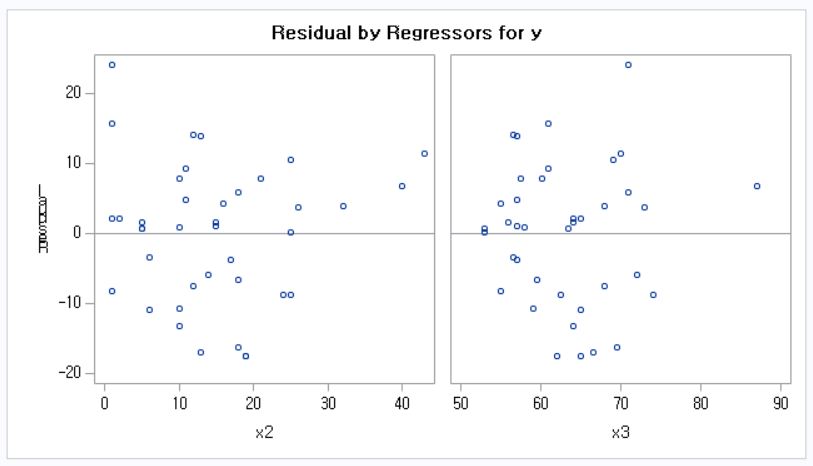
**run**;

**1. 변수선택을 하지 않고, 모든 변수들을 다 포함한 모형**



**2. 변수선택법(Stepwise)으로 선택된 변수 x2와 x3만을 포함한 모형**



해석 : 위 문제는 최고혈압(y)과 신체적 특성들인 나이(x1), 이주후 경과기간(x2), 몸무게(x3), 복부 피부두께(x4)에 관한 데이터로 다중 회귀분석을 수행한 결과이다. 먼저 변수선택법을 사용하지 않고 모든 변수들을 다 포함한 모형에 대한 결과를 보면, 독립변수들 간의 상관계수도 그리 높지않고 분산확대인자(VIF)와 상태지수 또한 높지 않아서 다중공선성 문제는 없어보인다. 그런데 독립변수 x1과 x4의 t-통계량에 대한 유의확률이 유의하지 않은 것으로 보아 변수선택법을 사용하여 분석에 사용할 변수들을 선택해주는 과정이 필요할 것으로 보인다. 그래서 변수선택법을 통해 변수 선택을 해보니, 실제로 모형에 포함된 변수들이 변수 x2와 x3인 것을 확인할 수 있다. 두 변수 x2와 x3의 t-통계량에 대한 유의확률이 모두 유의하고, 결정계수 값이 좀 줄어들긴 했지만 수정결정계수 값이 더 높아졌다는 것도 확인할 수 있다. 그리고 잔차의 정규성(P-P plot에 일직선으로 분포), 등분산성(잔차도표가 랜덤하게 퍼져있음), 독립성(더빗-왓슨 통계량이 2에 가까움)에 대한 가정이 모두 만족한다. 따라서 변수선택법을 통해 선택된 변수들만 포함한 모형(변수 x2와 x3를 포함한 모형)이 더 설명력이 높고 좋은 모형이라고 할 수 있다.

**<연습문제 8-5>**

**data** sasadv.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**;

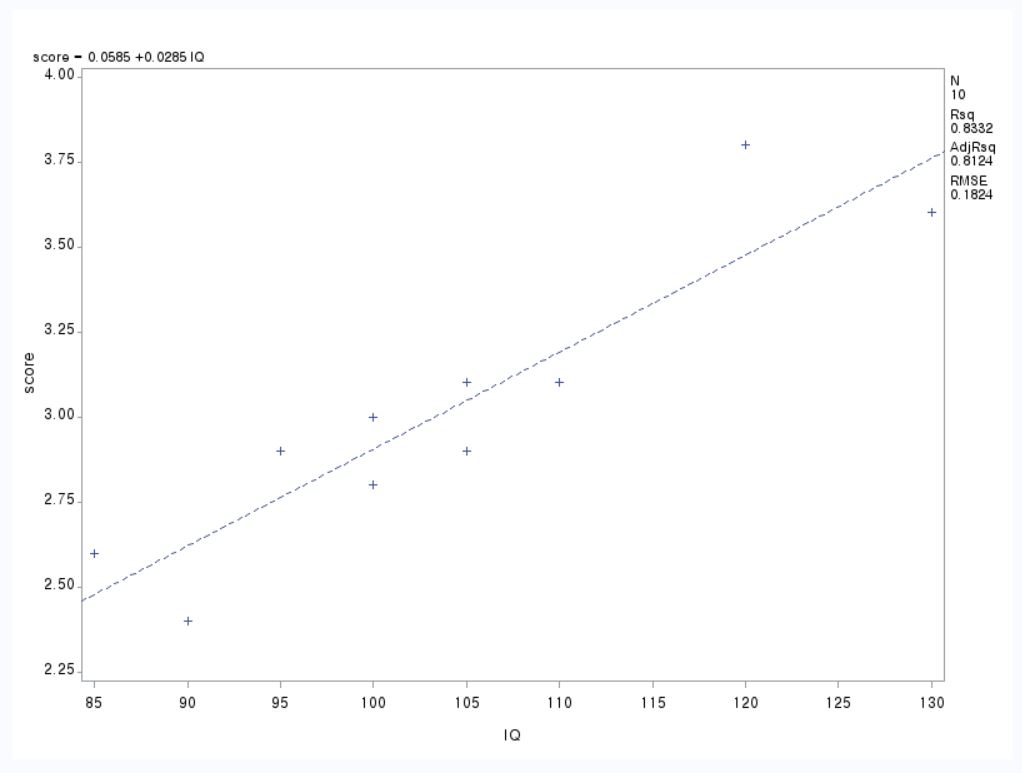
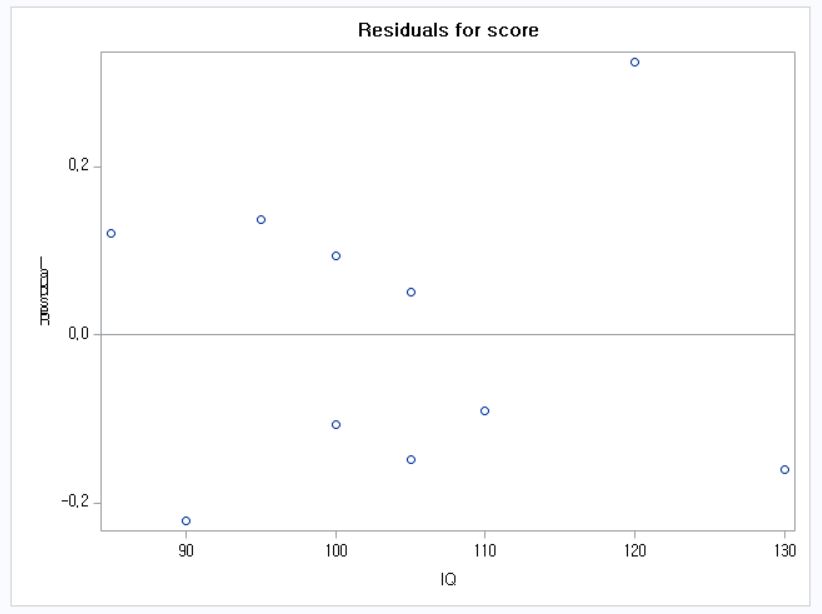
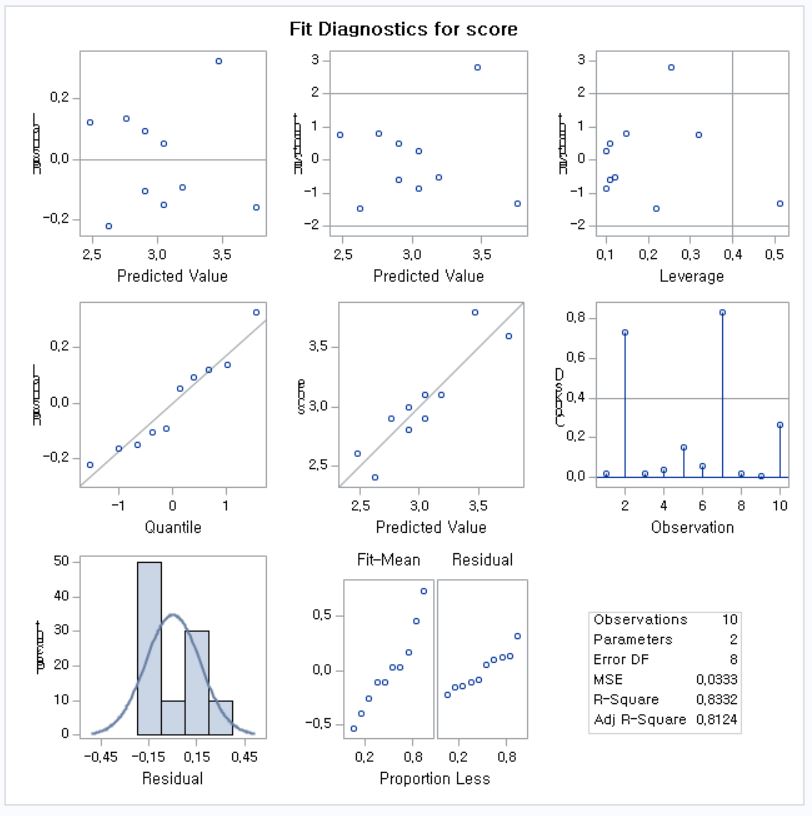
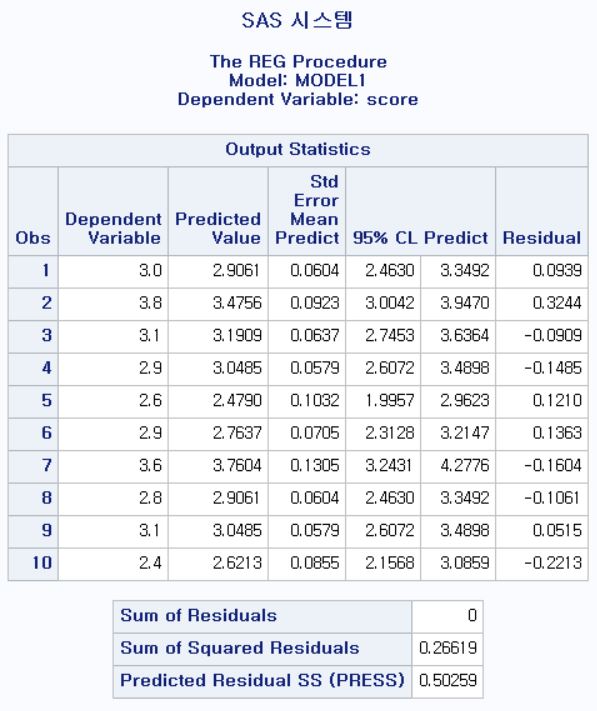
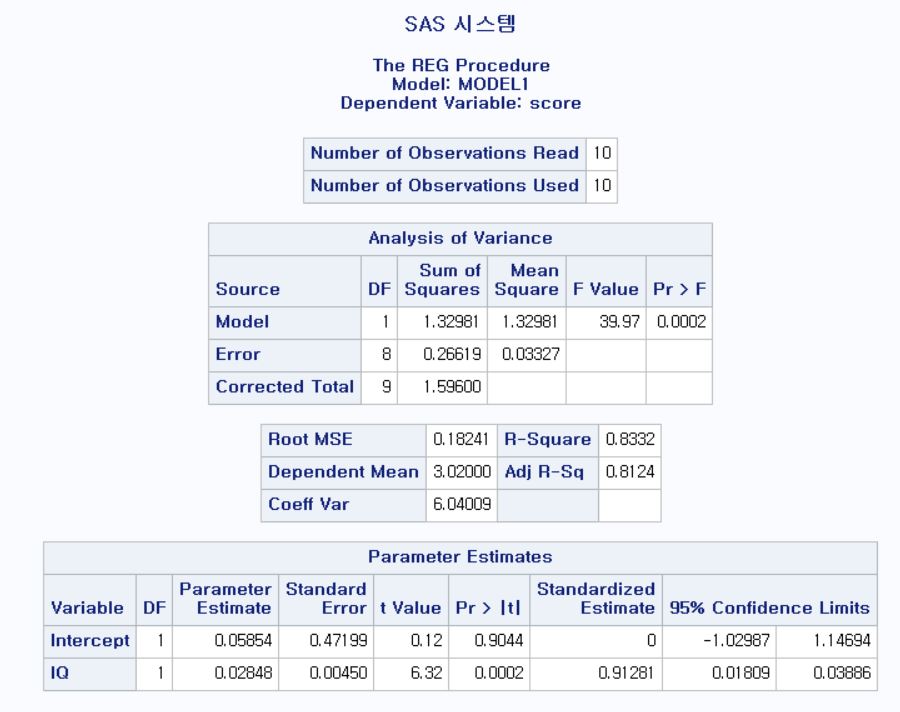
/\* (가), (나), (다), (라), (마) \*/

**proc** **reg** data=sasadv.ex8\_5;

model score=IQ / p cli clb stb;

plot score\*IQ;

**run**;



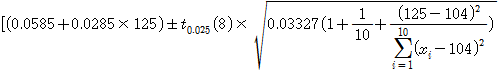
(가) 주어진 데이터를 이용하여 추정된 회귀직선식은 다음과 같다.

Score=0.0585+0.0285\*IQ

(나) 회귀직선의 기울기에 대한 95% 신뢰구간은 다음과 같다.

(0.01809, 0.03886)

(다) IQ가 125인 학생의 평균평점의 95% 예측구간은 다음과 같다.



(라) 단순 선형회귀분석의 경우, 결정계수의 값과 상관계수의 값이 동일하다. 따라서 상관계수의 값은 0.9128이다. 즉, IQ와 1학년 평균평점(score)사이에는 큰 양의 상관이 존재한다는 것을 의미한다.

(마) 결정계수의 값은 0.8332이다. 즉, 추정된 회귀직선식의 설명력이 약 83%라는 것을 의미한다.