Chapter 3 잔차분석

이론이나 경험에 의해 변수 간의 회귀모형을 설정하고 $y_i = \alpha + \beta x_i$ (선형: linearity), 관측치가 (x_i, y_i), $i = 1,2,...,n$ 얻어지면 이를 이용하여 회귀분석을 실시한다. 설정된 회귀모형에는 오차항에 대한 3가지 가정 $e_i \sim iidNormal(0, \sigma^2)$ 을 한다.

(정규성 normality, 등분산성 homoscadicity, 독립성, independence)

①관측치를 이용하여 OLS 추정치 $\hat{\alpha}, \hat{\beta}$ 을 구하고, 예측치(predicted) $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, 잔차(residual) $r_i = \hat{e}_i = y_i - \hat{y}_i$, 그리고 $\hat{\sigma}^2 = MSE$ 을 구한다.

이때까지는 오차에 대한 가정이 필요없다.

②분산분석 접근을 이용하여 $H_0 : \beta = 0$ 검정할 때는 $F^* = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE} \sim F(1, n-2)$ 이용한다. 단순회귀(설명변수가 하나)분석에서는 ②와 ③이 동일하다.

③ $\hat{\beta}$ 의 분포를 이용하여 $H_0 : \beta = 0$ (설명 변수가 종속 변수에 영향을 미치지 않는다. 선형관계가 존재하지 않는다) 가설 검정을 $T^* = \frac{\hat{\beta} - \beta_0 (= 0)}{s(\hat{\beta})} \sim t(n-2)$ 이용한다.

2장에서 살펴본 것은 $H_0 : \beta = 0$ 의 유의성을 검정하여 설명변수가 종속변수의 변동을 설명하는 정도가 "유의하다"는 가설을 검정하였다. 이런 가설 검정은 오차의 3가지 가정과 선형성 하에서 이루어졌다. 그러므로 이런 가정이 성립해야 회귀분석 결과가 타당한 것이다. 이에 대한 분석을 잔차(residual)분석이라 한다.

회귀 모형에서 종속변수는 오차항의 가정을 그대로 따르므로 종속변수에 대한 일변량 분석(stem-leaf plot, box-whisker plot, Shapiro-Wilk W-통계량)을 다루는 책도 있으나 잔차분석만으로 충분하므로 본 강의에서는 제외한다.

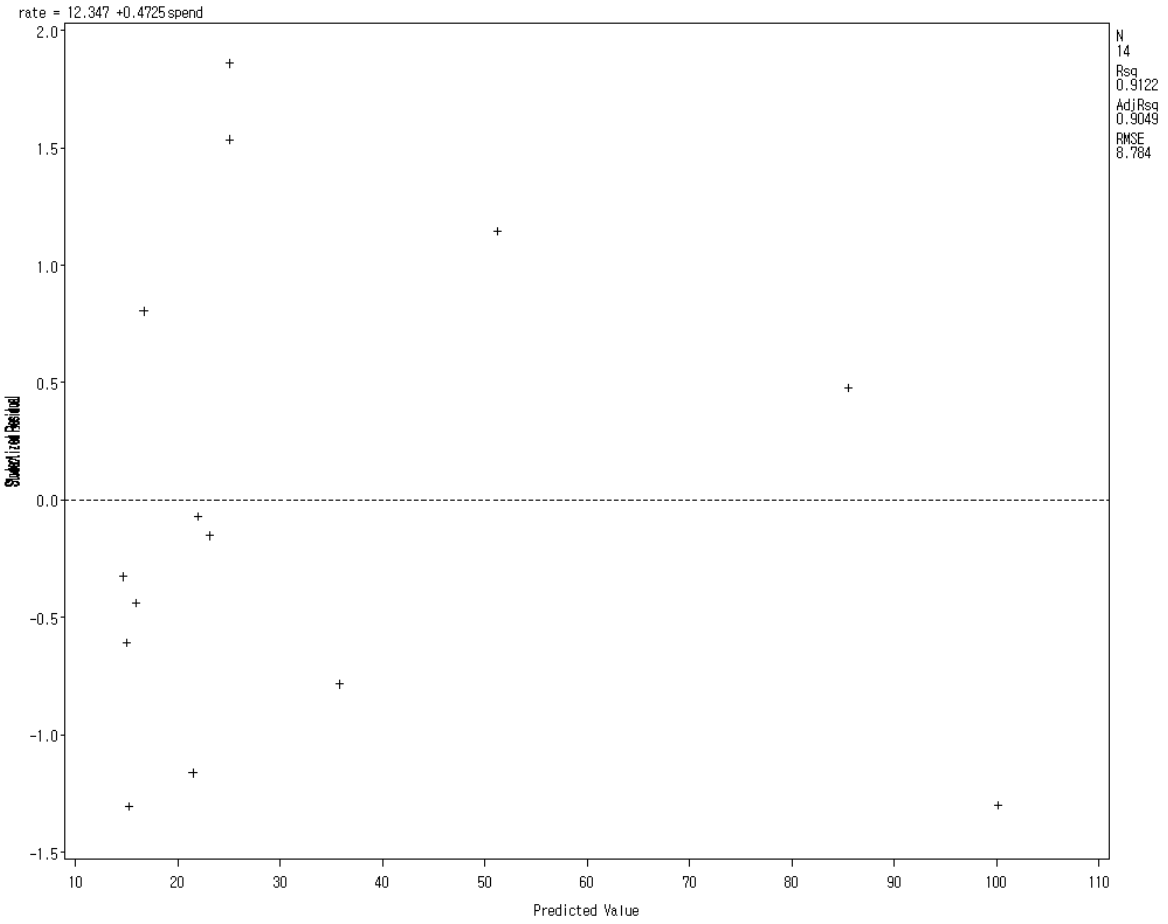
3.1 잔차

회귀 모형에서 오차항은 측정할 수 없으므로 오차항($e_i = Y_i - E(Y_i)$)에 대한 추정치가 필요한데 이를 잔차라 한다. 오차의 가정은 잔차에 의해 성립 여부가 판단된다. 잔차는 다음과 같이 정의된다. $r_i = \hat{e}_i = y_i - \hat{y}_i$

참고

SAS에서 스튜던트 잔차(Studentized residual)와 예측치($\hat{Y}_i$)의 산점도를 PROC REG에서 그릴 수 있다.

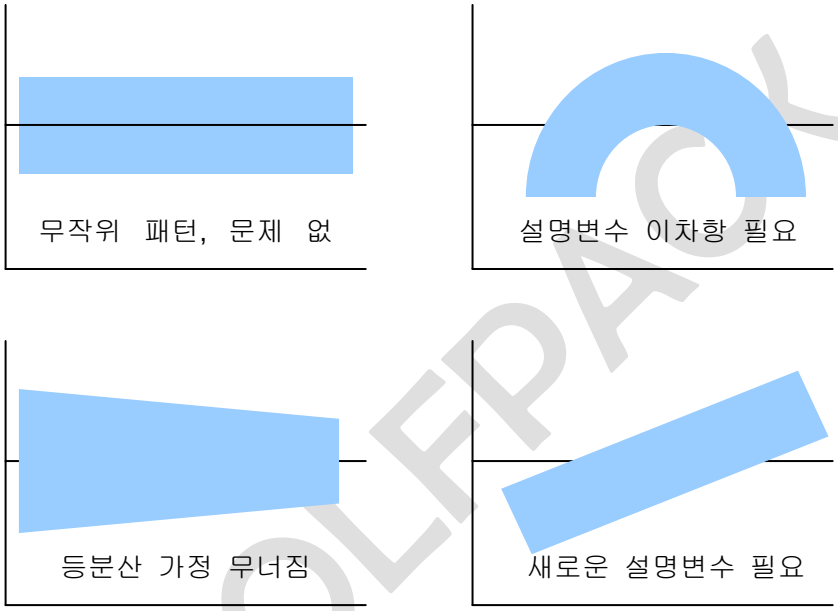
```
proc reg data=ad;
    model rate=spend;
    reweight obs.=1; reweight obs.=9;
    reweight obs.=5; reweight obs.=19;
    reweight obs.=7;
    reweight obs.=11;
    plot student.*predicted.;
    output out=out1 predicted=yhat residual=res student=sres;
run;
```



3.4 가정 파괴와 해결책

선형회귀 분석이란 **선형모형**을 설정하고 수집된 데이터를 이용하여 **회귀계수를 추정**하고(OLS 방법) t-검정이나 분산 분석에 의해 **설명변수의 유의성**(단순회귀모형에서는 회귀계수의 유의성과 동일)을 검정한다. 그리고 얻어진 적합(fitted) 회귀모형에 의해 주어진 설명변수의 값에 대한 종속변수의 예측치를 얻는다.

이런 과정에서 회귀모형은 선형이고 오차는 독립성(시계열 자료만), 등분산성, 정규성을 가정한다. 이제 이런 가정을 진단하는 방법과 파괴되었을 때 해결책을 살펴보기로 하자. 다음은 잔차와 예측치의 산점도를 그린 것이다.



3.3.1 선형성(linearity)

진단방법 ①(설명변수와 종속변수) 산점도 ➔ 이차 함수 형태

②잔차와 예측치 산점도 ➔ 이차 함수 형태

해결방법 ①설명 변수의 이차항이나 다차항을 삽입한다.

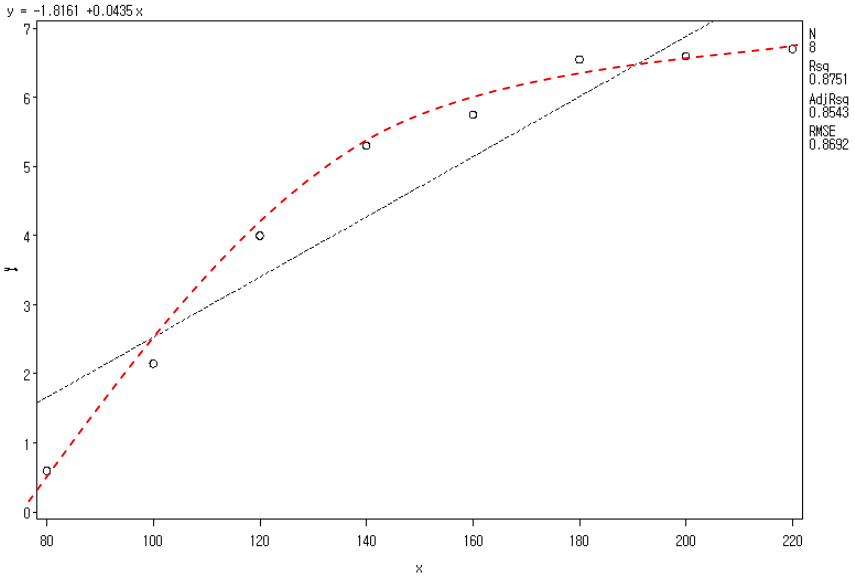
산점도를 보면 종속변수와 설명변수의 직선(산형) 관계를 진단할 수 있다. 잔차와 예측치의 산점도가 일정한 함수 형태를 가지면(일반적으로 이차 함수) 선형성이 무너지게 되는데 이를 해결하려면 설명변수의 이차항을 설명변수로 추가한다. 이차항을 추가할 때는 설명변수를 표준화 한 후 넣으면 다중공선성 문제가 완화된다. (다음 페이지 참고)

EXAMPLE 3-2 선형성 파괴: 이차 관계

```
data quard;
    input y x @@;
    cards;
0.6 80 6.7 220 5.3 140 4 120
6.55 180 2.15 100 6.6 200 5.75 160
run;

proc reg data=quard;
    model y=x;
    plot y*x;
    plot student.*predicted.;
run;
```

종속변수와 설명변수의 산점도를 보면 직선 관계라고 보기 어렵다. 이차 함수 관계에 가깝다.



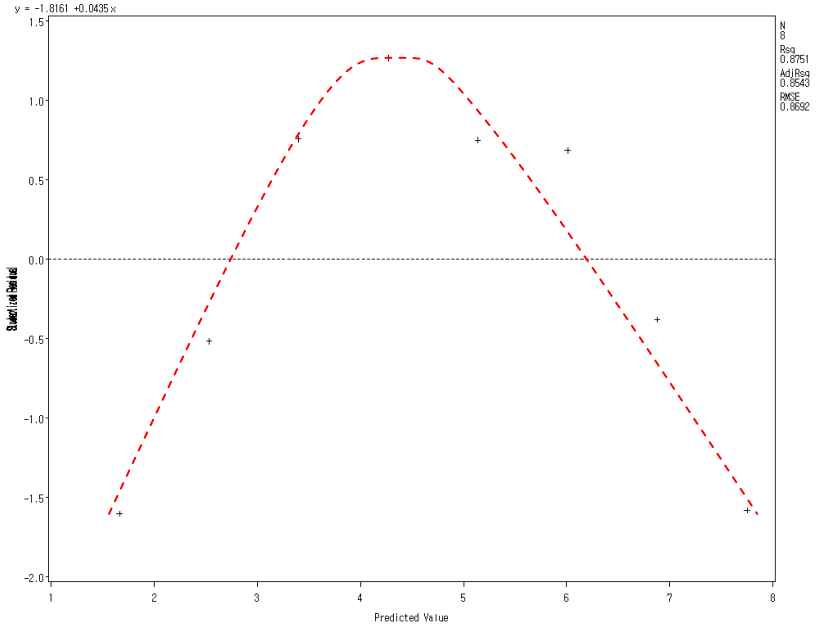
그러나 설명변수 X 의 회귀계수 유의성 검정 결과는 매우 유의하므로 잔차 분석을 하지 않는다면 설명변수와 종속변수 간에는 직선관계가 성립한다고 결론 내리게 된다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.81607	1.05185	-1.73	0.1350
X	1	0.04348	0.00671	6.48	0.0006

잔차와 예측치 산점도를 살펴보자. 무작위 패턴이 아니라 이차 함수 형태를 가지므로 설명변수의 제곱 항이 필요하다. 이는 앞의 산점도에서도 예상되었던 일이다. 이처럼 종속변수와 설명변수의 산점도는 회귀분석 결과를 미리 예상할 수 있게 하는 주요 도구이다.

Plot of Stud.Res*Yhat



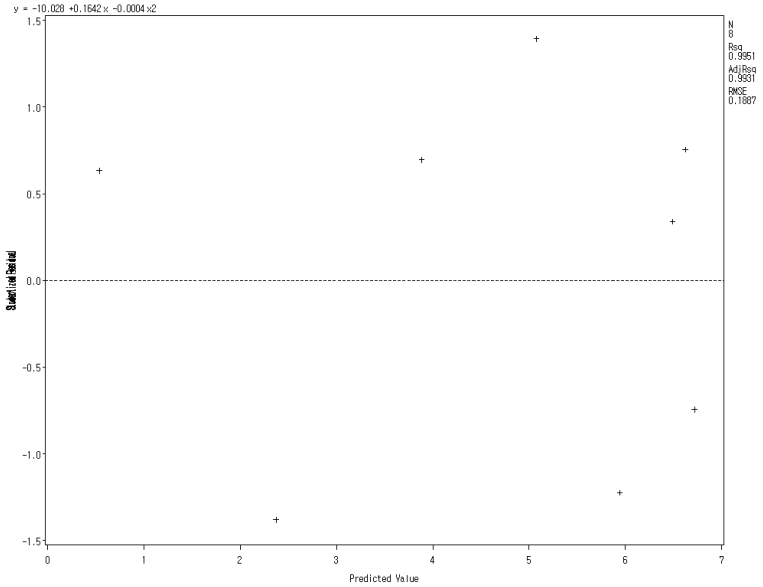

```
data quard1;
    set quard;
    x2=x**2;
run;
```

```
proc reg data=quard1;
    model y=x x2;
    plot student.*predicted.;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-10.02768	0.77696	-12.91	<.0001
X	1	0.16424	0.01102	14.91	<.0001
X2	1	-0.00040253	0.00003640	-11.06	0.0001

Plot of Stud.Res*Yhat



회귀 계수에 대한 t-검정 결과 설명 변수 X, X^2 모두 유의하고 잔차 분석 결과 잔차가 패턴을 갖지 않으므로 최종 회귀 모형은 $\hat{Y} = -10.03 + 0.16X - 0.0004X^2$ 이다. 회귀계수도 유의하고 잔차에도 아무 문제가 없어 보인다. 그러나...

설명변수의 일차항과 이차항을 회귀모형에 동시에 넣으면 다중공선성 문제가 발생한다는 것이다. 다중공선성이란 설명변수들 간의 높은 상관 관계로 인하여 회귀계수 추정치의 분산이 커져 추정치의 부호까지 바뀌는 심각한 문제를 의미한다. 설명변수를 표준화한 후 넣으면 다중공선성 문제가 다소 해결할 수 있다.

정말 다중공선성 문제가 발생하는지 알아보자. 다중공선성에 대한 자세한 다음에 다루기로 하고 여기서 간단히 언급하겠다. VIF, Condition Index를 출력하기 위하여 VIF, COLLIN 옵션을 사용하였다.

```
proc reg data=quard1;
     model y=x x2/vif collin;
run;
```

VIF, condition Index 가 10(일반적으로)이상이므로 X, X^2 은 다중공선성 문제를 일으킨다.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-10.02768	0.77696	-12.91	<.0001	0
x	1	0.16424	0.01102	14.91	<.0001	57.25000
x2	1	-0.00040253	0.00003640	-11.06	0.0001	57.25000

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: Intercept	x	x2
1	2.86923	1.00000	0.00083195	0.00018091	0.00048203
2	0.12980	4.70160	0.03124	0.00005057	0.01447
3	0.00096570	54.50822	0.96793	0.99977	0.98504

이제 설명변수 X 을 표준화 한 후 설명변수로 사용해 보자. STANDARD procedure는 평균(M=0)과 표준편차(STD=1)에 의해 표준화하는 문장이다.

```
proc standard data=quard m=0 std=1 out=quard2;
     var x;
run;

data quard3;
     set quard2;
     x2=x*x;
run;

proc reg data=quard3;
     model y=x x2/vif collin;
run;
```

다중공선성 문제는 해결되고 회귀계수도 유의하였다. 여기에는 보여주지 않았지만 잔차와 예측치의 산점도에도 아무 문제가 없었다. 최종 회귀모형은 다음과 같다.

$\hat{Y}_i = 5.55 + 2.13X_i^* - 0.97X_i^{*2}$, where $X^* = \frac{X - \bar{X}}{S_X}$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	5.55156	0.10147	54.71	<.0001	0
x	1	2.13018	0.07134	29.86	<.0001	1.00000
x2	1	-0.96607	0.08737	-11.06	0.0001	1.00000

HOMEWORK #4-2 DUE 3월 30일(수)

다음 자료에서 설명변수 X가 종속변수 Y에 선형적인 영향을 미치는지 분석하시오. 잔차분석결과 문제가 있으면 해결하고 최종 회귀 모형을 제시하시오. SPSS 사용

```
data hw5;
    input y x @@;
    cards;
2.5 1 2.6 1.6 2.7 2.5 5 3 5.3 4
9.1 4.6 14.8 5 17.5 5.7 23 6 28 7
run;
```

3.3.2 등분산성(homoscedasticity)

진단방법 ①잔차와 예측치 산점도, 나팔 모양

해결방법 ①가중최소자승법, WLS(Weighted Least Square) 사용한다.

②종속변수변환. 일반적으로 LOG 변환을 하는 것이 일반적이다.

잔차와 예측치 산점도에서 나팔 모양이면 오차의 분산이 예측치가 커짐에 따라 커지거나 작아지고 있음을 의미하므로 등분산 가정이 무너지게 된다. 이런 경우 가중최소자승 추정치를 이용하거나 종속변수변환을 실시한다. 등분산의 경우 일반적으로 오차의 분산은 $V(e_i) = \sigma_i^2 = \sigma^2 / w_i$ 으로 가정되고 가중최소자승가중치로 $w_i = 1/y_i^2$, 혹은 $w_i = 1/x_i^2$ 을 주로 사용한다.

WLS(Weighted Least Square)

$\min_{\alpha,\beta} \sum w_i (y_i - \alpha - \beta x_i)^2$ 인 $\hat{\alpha}, \hat{\beta}$ 을 WLS 추정치라 한다. 일반적으로 가중치 w_i 는 $1/\sigma_i{}^2$ ($\sigma_i{}^2$

을 알고 있을 때, 그러나 실제 알지 못한다) 혹은 $1/x_i^2$, $1/\hat{y}_i^2$ 등을 사용한다. 단순회귀의 잔차분석은 잔차와 예측치 산점도에 주로 의존하므로 $1/\hat{y}_i^2$을 주로 사용한다. 다중회귀에서는 문제가 되는 설명변수를 이용한 가중치$1/x_i^2$을 사용하기도 하지만 판단이 쉽지 않아 다중회귀모형에서도 $1/\hat{y}_i^2$을 사용한다.

가중회귀 추정치를 구하는 문제는 다음과 같이 생각할 수 있다. 종속변수가 y_i^*, 설명변수가 $1/x_i$ 인 회귀모형의 OLS 구하는 문제와 동일하다.

$$\min_{\alpha,\beta} \sum \frac{1}{x_i^2}(y_i - \alpha - \beta x_i)^2 = \min_{\alpha,\beta} \sum (\frac{y_i}{x_i} - \frac{\alpha}{x_i} - \beta)^2 = \min_{\alpha,\beta} \sum (y_i^* - \frac{1}{x_i}\alpha - \beta)^2$$

가중치를 $1/\hat{y}_i^2$ 사용했을 때는 다음 정규방정식에 의해 추정치를 구할 수 있다. 이를 가중회귀추정치이다.

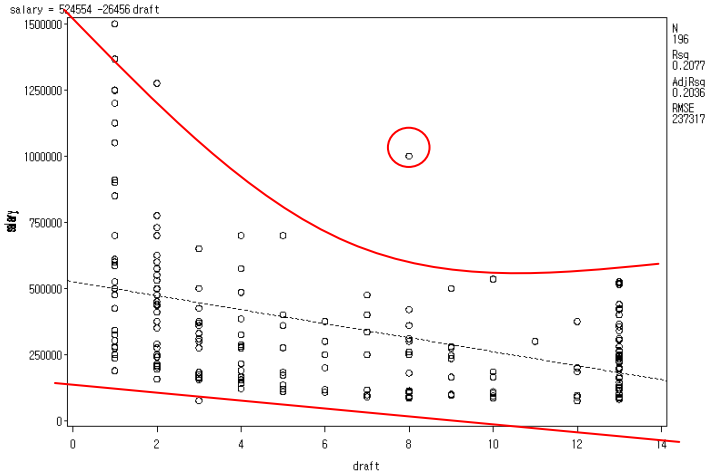
$$\alpha \sum w_i + \beta \sum w_i x_i = \sum w_i y_i$$
$$\alpha \sum w_i x_i + \beta \sum w_i x_i^2 = \sum w_i x_i y_i$$

EXAMPLE 3-3 이분산성 문제

NFL 선수 연봉 관련 자료이다.[💾**NFL.xls**] 연봉(salary), 포지션(position: 1=Offensive Back, 2=Defensive Back, 3=Lineman, 4=kicker/punter), Draft 순위(draft), 경력(yrs_exp), 출장 회수(played), 선발 출장 회수(started), 지역 인구(city_pop)를 조사하였다. Draft 순위가 연봉에 미치는 영향을 보기 위하여 선형 회귀 분석을 실시하자.

```
proc reg data=nfl;
    model salary=draft;
    plot salary*draft;
    plot student.*predicted.;
run;
```

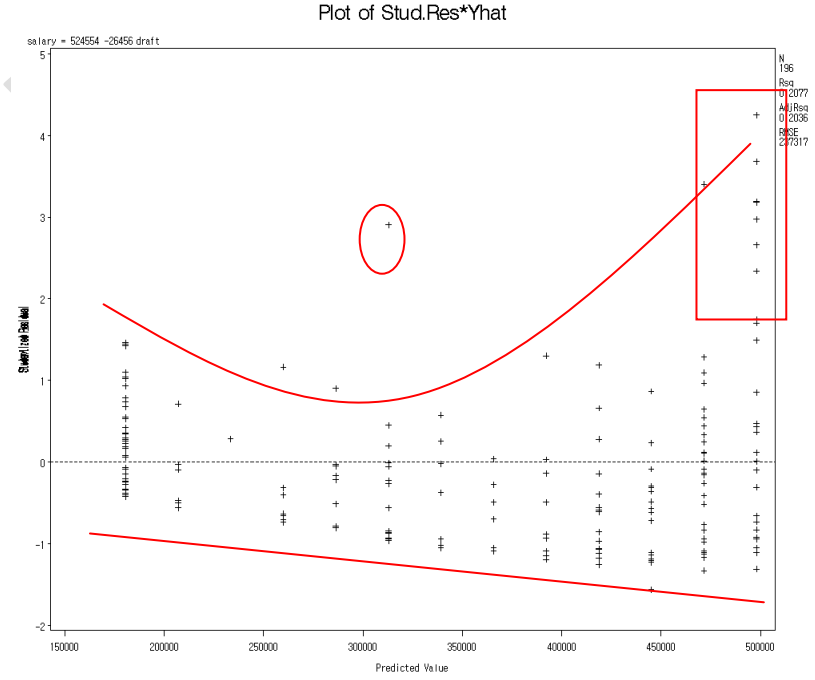
다음은 종속변수(salary)와 설명변수(draft)의 산점도이다. 여기서도 알 수 있듯이 설명변수의 각 값에서 보면 종속변수의 변동이 다름을 알 수 있다. 산점도를 통하여서도 종속변수의 변동이 다름을 알 수 있다. 아마 이분산(heteroscedasticity) 문제가 발생할 것이라는 것을 예상할 수 있다. 그런데 문제가 심각해 보인다. 이분산 문제가 부채꼴의 형태가 아니라는 것이다. 왜냐하면 가중자승추정은 부채꼴의 이분산만 해결할 수 있기 때문이다. 산점도에 의하면 이상치도 존재하는 것 같다.



회귀 계수 검정에 의하면 추정된 회귀 모형은 적합하다.(유의확률<0.001) 문제가 없어 보인다. 잔차분석 전까지는...

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	524554	29150	17.99	<.0001
draft	1	-26456	3709.54993	-7.13	<.0001

불행히도 잔차와 예측치의 산점도는 나팔 모양을 갖는다. 이는 이분산 문제가 발생했음을 말해준다. 이분산 문제로 인하여 네모 상자 부분의 관측치들이 이상치로 간주되고 있다. 산점도를 보면 종속변수의 값이 커짐에 오차의 분산이 커지므로 이분산 문제를 해결하기 위해서는 가중치로 $1/\hat{y}_i^2$을 사용하면 된다.



다음 프로그램은 가중치를 $1/\hat{y}_i^2$ 하여 WLS 추정치를 구한 프로그램이다.

```
proc reg data=nfl;
    model salary=draft;
    output out=out1 p=yhat;
run;
```

```
data out2;
    set out1;
    w=1/yhat**2;
run;
```

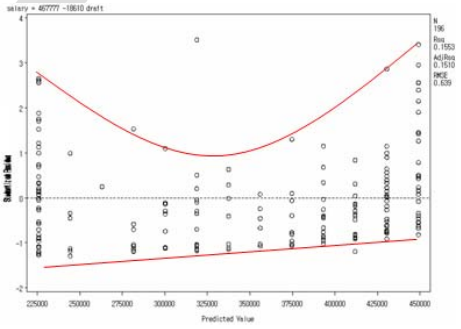
($\hat{y}$ 값 이용하여 가중치 계산)

```
proc reg data=out2;
    weight w;
    model salary=draft;
    plot rstudent.*predicted.;
run;
```

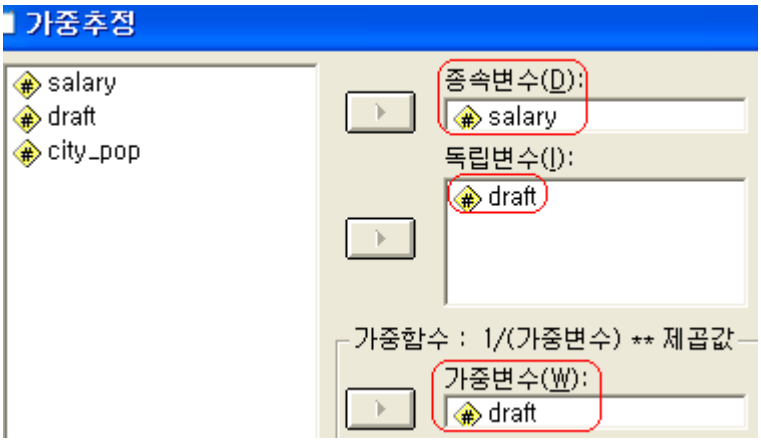
회귀계수 추정치가 OLS 추정치와 다르다. 회귀계수는 매우 유의하다. 그러나 잔차와 예측치 산점도를 보면 여전히 이분산 문제가 있는 것으로 나타났다.(아래 산점도) 이는 종속변수(연봉)와 설명변수(DRAFT)의 산점도에서 살펴 본 것 같이 나팔 모양이 아니라 양쪽이 넓어짐을 알 수 있다.

이로 인하여 $1/\hat{y}_i$을 가중치로 이용한 WLS 추정방법은 문제 해결을 하지 못했다.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	467777	33703	13.88	<.0001
draft	1	-18610	3115.75806	-5.97	<.0001



SPSS 가중회귀분석 절차는 분석(A) 회귀분석(R) 가중추정(W)...이다. 가중치를 종속변수의 예측치로 사용하려면 우선 일반 회귀분석을 시행하여 종속변수의 예측치를 변수로 저장하여 구한 후 가중함수에 종속 변수 예측치(변수명: PRE_1)를 지정하면 된다.



HOMEWORK #4-3 DUE 3월 30일(수)

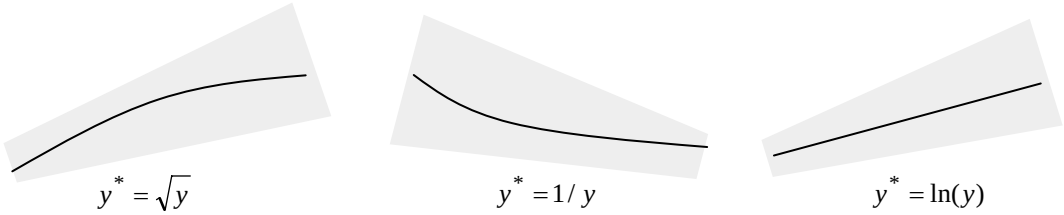
다음 자료는 나이가 혈압에 영향을 미치는지 알아보고자 조사한 자료이다. 잔차 분석 결과 문제가 있으면 해결하고 최종 회귀 모형을 제시하시오. SAS 이용하기

Age X_i	Pressure Y_i	Age X_i	Pressure Y_i	Age X_i	Pressure Y_i
27	73	37	78	42	85
21	66	38	87	44	71
22	63	33	76	46	80
26	79	35	79	47	96
25	68	30	73	45	92
28	67	37	68	55	76
24	75	31	80	54	71
25	71	39	75	57	99
23	70	46	89	52	86
20	65	49	101	53	79
29	79	40	70	56	92
24	72	42	72	52	85
20	70	43	80	57	109
38	91	46	83	50	71
32	76	43	75	59	90
33	69	49	80	50	91
31	66	40	90	52	100
34	73	48	70	58	80

3.3.3 변수변환

회귀분석에서 변수변환(variable transformation)은 오차의 비정규성 문제, 종속 변수와 설명 변수간의 비선형 함수 관계 해결에 이용된다. 변수변환은 종속변수나 설명변수 모두 가능하나 일반적으로 종속변수에 하는 것이 적절하다. 왜냐하면 회귀계수의 의미는 설명변수 한 단위의 증가에 따른 종속변수 변화 량이므로 해석의 편리성 때문이다.

변수변환 방법 설정은 원 변수의 산점도나 잔차와 예측치의 산점도에 의해 결정한다. 다음은 **잔차와 예측치 산점도**의 형태에 따른 적절한 종속 변수변환 방법을 보여준다.



$y^* = \sqrt{y}$ $y^* = 1/y$ $y^* = \ln(y)$

EXAMPLE 3-4 변수 변환

NFL 선수 연봉 자료에서 Draft 순위가 연봉에 미치는 영향을 보기 위하여 선형 회귀 분석을 실시하였더니 이분산 문제가 발생하였다. 그러나 WLS 추정 방법에 의해 이분산 문제를 해결하지 못하였다. 변수변환에 의해 이분산 문제를 해결해 보자.

종속변수를 로그 변환한 후 회귀분석을 실시해 보자.

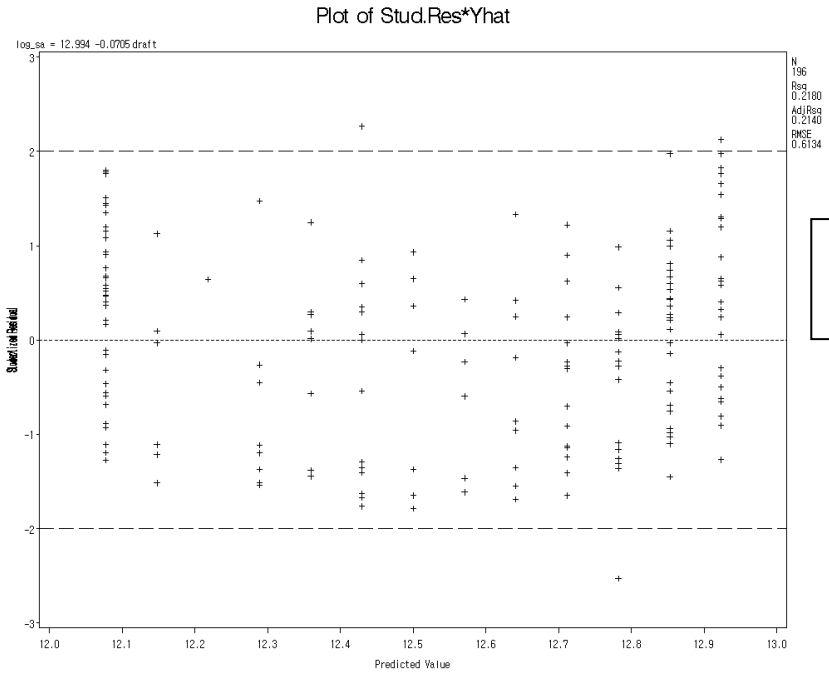
```
data nfl1;
    set nfl;
    log_sa=log(salary);
run;

proc reg data=nfl1;
    model log_sa=draft/r;
    plot student.*predicted./vref=2 vref=-2;
run;
```

회귀계수도 유의하고 잔차와 예측치 산점도에도 이상치만 몇 개 나올 뿐 이분산 문제는 해결되었다. 이처럼 로그 변환은 많은 문제의 해결책으로 빈번히 등장한다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.99407	0.07534	172.46	설명변수 유의
draft	1	-0.07051	0.00959	-7.35	



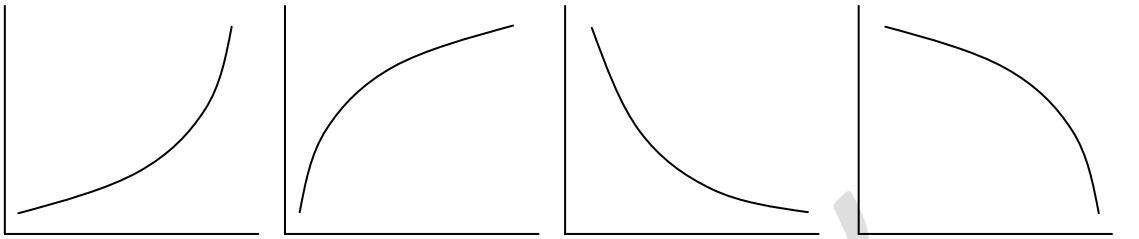
무작위로 보이므로 이분산 문제 해결

회귀계수 $H_0 : \beta = 0$ 가설 검정 결과 귀무가설이 기각되고, 잔차 분석 결과 문제(Random pattern)가 없으므로 최종 회귀모형을 얻을 수 있다. 아직 이상치는 제외하지 않았지만... 설명변수 Draft 순위는 선수 연봉에 영향을 미치되 음의 영향(추정 회귀 계수 $\hat{\beta}$ 부호 -)을 미친다. 즉, draft 순위가 늦을수록 연봉은 높아진다. 추정된 최종 회귀 모형은 다음과 같다.

$$\ln(\hat{salary}) = 12.99 - 0.0705 * draft$$

Draft 순위가 3위인 경우 연봉은 얼마인가? 354,512.9(= $e^{12.7785}$)($)이다.

다중회귀모형에서는 종속 변수와 설명 변수의 산점도를 보고 선형 관계가 아닌 설명변수가 존재하면 설명변수에 대한 변수변환을 실시한다. 일반적으로 변수변환은 종속변수와 설명변수 어느 것이나 가능하나 다중 회귀의 경우에는 설명변수를 단순회귀에서는 종속변수를 변환하는 것이 일반적이다.



$y^* = \log y, y^* = 1-/y$ $x^* = x^2,\ x^* = x^3$	$y^* = y^2, y^* = y^3$ $x^* = \log x, x^* = 1-/x$	$y^* = \log y, y^* = 1/y$ $x^* = \log x, x^* = 1-/x$	$y^* = y^2, y^* = y^3$ $x^* = x^2,\ x^* = x^3$

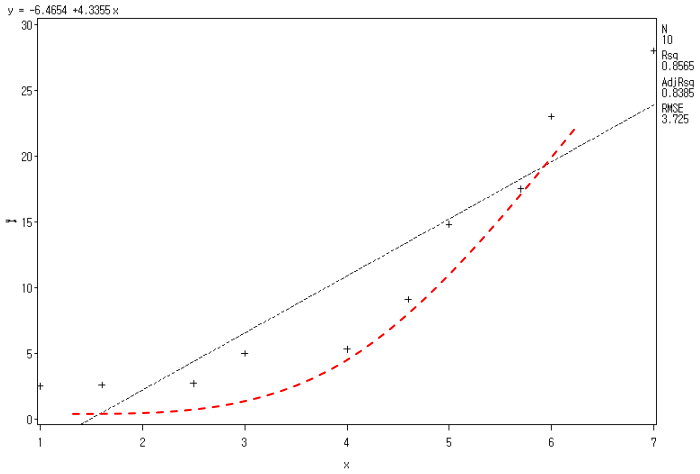
EXAMPLE 3-5 변수 변환(2)

다음 자료에 대한 회귀분석을 실시해 보자.

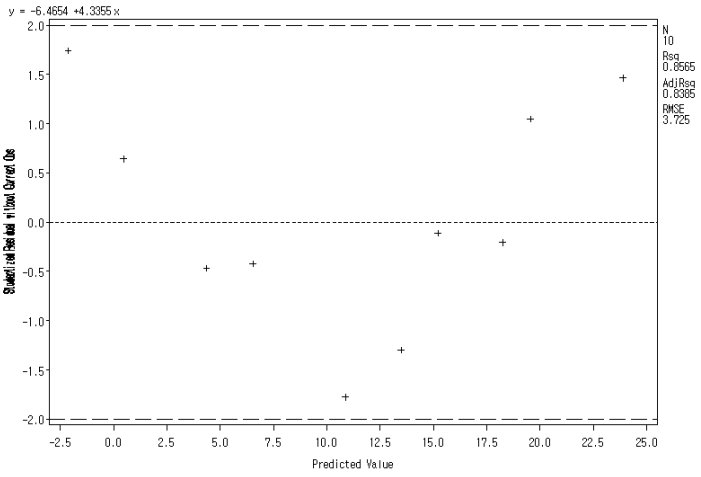
```
data one;
    input y x;
    cards;
    2.5 1
    2.6 1.6
    2.7 2.5
    5 3
    5.3 4
    9.1 4.6
    14.8 5
    17.5 5.7
    23 6
    28 7
run;
```

```
proc reg data=one;
    model y=x;
    plot y*x;
    plot  student.*predicted./vref=-2 vref=2;
run;
```

종속변수와 설명변수의 산점도를 보면 종속변수와 설명변수의 관계는 직선으로 생각하기에는 어렵다. 이 산점도는 페이지 70의 변수변환 보기 그림의 첫 번째와 비슷하므로 종속변수 $y^* = \log y$ 이나 설명변수의 $x^* = x^2$ 변환이 적절하다.



잔차와 예측치의 산점도는 마치 설명변수 이차항을 삽입해야 하는 형태이다. 페이지 70을 보라. 위 형태의 산점도를 보이면 이 문제를 해결하기 위하여 설명변수를 변환하는 경우 제곱 변환을 해야 한다고 설명하였다. 앞에서 언급하였듯이 종속변수 변환이 해석 용이, 일반적인 방법이므로 종속변수 변환을 사용하여 문제를 해결하였다.



직선 관계가 적절하지 않아도 선형회귀모형을 추정하면 유의하다. (기울기 회귀계수의 유의확률=0.0001) 그러므로 F-검정이나 t-검정에만 의존하여 모형의 유의성(회귀계수의 유의성, 설명변수의 유의성)을 검정하면 문제가 발생한다.

Parameter Estimates

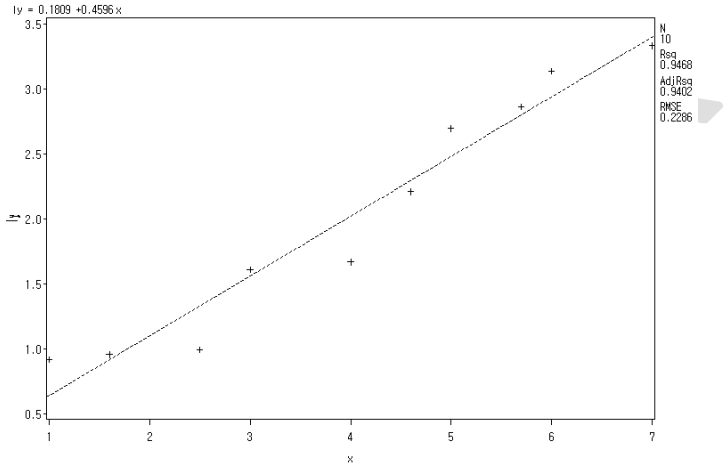
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.46538	2.79521	-2.31	0.0495
x	1	4.33549	0.62745	6.91	0.0001

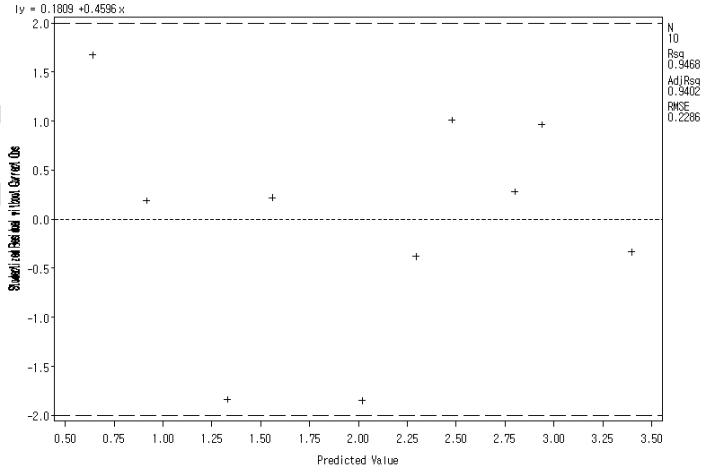
잔차와 예측치 산점도의 경우를 보면 설명변수의 이차항이 빠진 것으로 판단된다. 그러므로 설명변수의 이차항을 넣는 것도 고려할 수 있다. 그러나 이차항을 넣으면 설명변수 일차항과 다중공선성 문제가 발생하므로 사용하지 않는 것이 좋다. (3.3.1절 끝부분) 종속

변수변환 후 $y^* = \log y$ 회귀분석을 실시하자.

```
data two;
     set one;
     ly=log(y);
run;

proc reg data=two;
     model ly=x;
     plot ly*x;
     plot  student.*predicted./vref=-2 vref=2;
run;
```





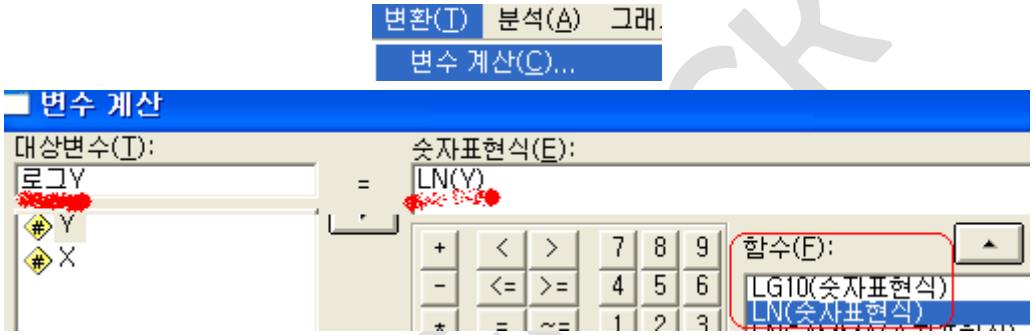
Log(y)와 X의 산점도는 직선 형태를 보이고, 잔차와 예측치의 산점도 문제가 없어 보이므로 추정 회귀 모형은 적합하다. 설명변수 X 가 한 단위 증가하면 종속변수 $\ln(Y)$ 는 0.46 증가한다. 그러므로 Y 는 $e^{0.46} = 1.584$ 증가한다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.18090	0.17154	1.05	0.3224
x	1	0.45955	0.03851	11.93	<.0001

최종회귀모형: $Ln(y) = 0.18 + 0.4596 * X$ $(p < 0.0001)$

SPSS 우선 변수변환을 실시한다.



데이터에 "로그Y" 변수가 계산된다. 이제 종속변수를 "로그Y", 설명변수를 C로 하여 회귀분석을 실시하면 된다.

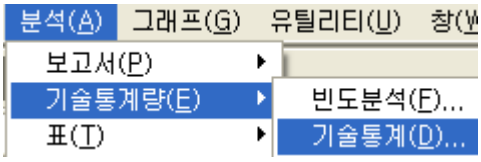
Y	X	로그Y
2.5	1.0	.92
2.6	1.6	.96
2.7	2.5	.99
5.0	3.0	1.61
5.3	4.0	1.67

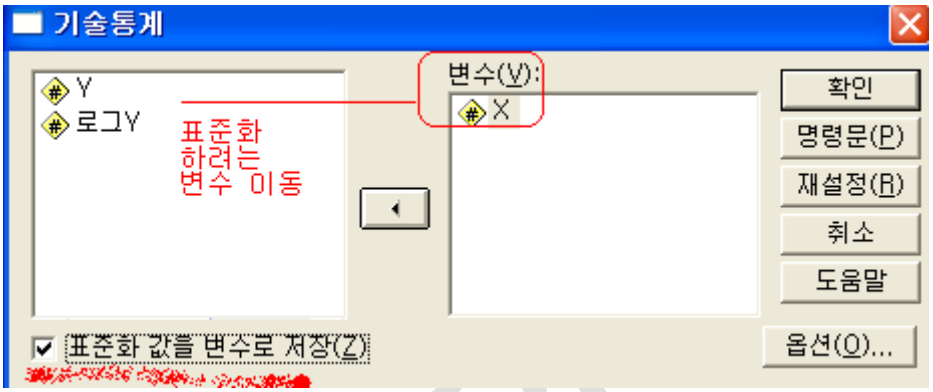
HOMEWORK #5-1 DUE 4월 6일(수)

HOMEWORK#4-2 문제를 변수변환 방법으로 해결하고 최종회귀모형을 구하고 해석하시오. SPSS 사용

참고

(1) SPSS에서 변수 표준화 하기.



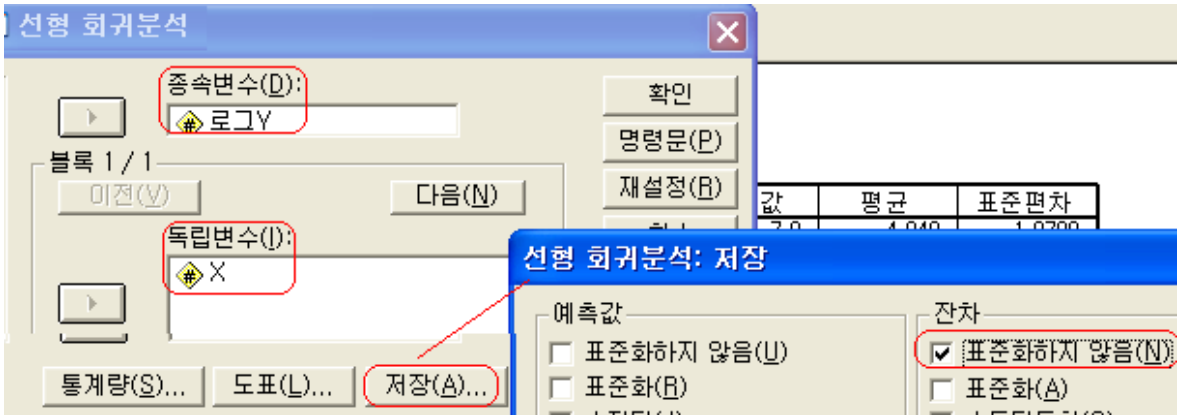


확인 을 누르면 데이터 마지막 열에 표준화 변수가 만들어진다.

Y	X	로그Y	ZX
2.5	1.0	.92	-1.53622
2.6	1.6	.96	-1.23302
2.7	2.5	.99	-.77821
5.0	3.0	1.61	-.52555
5.3	4.0	1.67	-.02021

(2) SPSS에서 잔차의 정규성 검정하기.

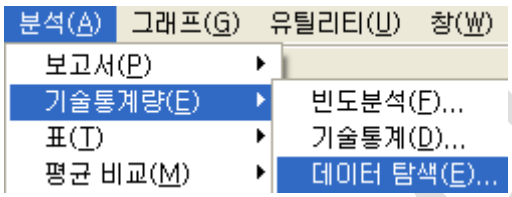
회귀분석을 실시하고 잔차를 변수로 저장한다.



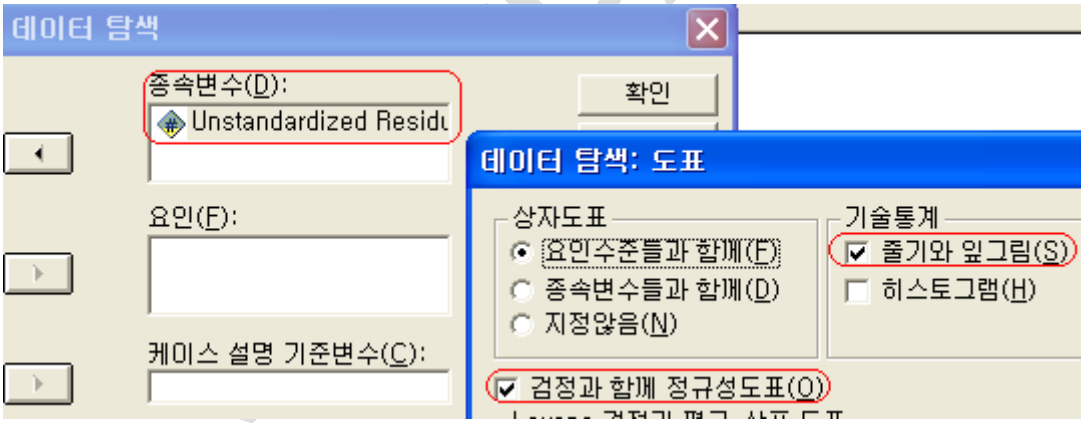
데이터 마지막 열에 잔차가 만들어진다.

Y	X	로그Y	ZX	RES_1
2.5	1.0	.92	-1.53622	.27583
2.6	1.6	.96	-1.23302	.03932
2.7	2.5	.99	-.77821	-.33654
5.0	3.0	1.61	-.52555	.04987
5.3	4.0	1.67	-.02021	-.35141
9.1	4.6	2.21	.28299	-.08658

다음 방법을 이용하여 잔차의 정규성을 검정하면 된다.



종속변수에는 잔차를 설정하고 "도표" 옵션을 아래와 같이 설정한다.



K-S나 S-W 검정통계량 모두 유의확률이 0.05보다 크므로 귀무가설(정규분포를 따른다)은 채택되어 정규성을 만족한다고 할 수 있다.

정규성 검정

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	통계량	자유도	유의확률	통계량	자유도	유의확률
Unstandardized Residual	.172	10	.200(*)	.915	10	.318

3.1.1 잔차의 성질

잔차의 평균은 0이고 분산은 MSE ($V(r_i)\frac{\sum(r_i-\bar{r})^2}{n-2}=\frac{\sum(r_i)^2}{n-2}=\frac{\sum(y_i-\hat{y}_i)^2}{n-2}=MSE$)이다. 실제 잔차들은 서로 독립은 아니다. 왜냐하면 잔차를 구하기 위해서는 $\hat{\alpha},\hat{\beta}$ 가 추정되는데 이 추정치에는 각 관측치 (x_i,y_i) 의 정보가 모두 있기 때문이다. 그러나 우리가 앞에서 증명하였듯이 $\sum x_i r_i=0$, $\sum \hat{y}_i r_i=0$ 이 성립하고 모수(α,β)의 개수에 비해 관측치의 개수(n)가 상대적으로 크면 잔차의 비독립성 효과는 줄어든다. 실제 횡단면 자료(시계열 자료가 아님)에 대한 회귀 분석에서는 오차(잔차)의 독립성 검정을 실시하지 않는다. 시계열 자료에서 오차의 독립성은 Durbin-Watson (DW) 통계량을 이용한다.

3.1.2 잔차분석 정의

잔차분석이란 오차의 추정치인 잔차를 이용하여 다음 정보를 얻어내는 과정을 의미한다.

(1)설명변수와 종속변수의 함수 관계는 선형인가?

(2)오차의 분산은 설명 변수의 값에 따른 변화는 없는가? (등분산성)

(3)오차항은 서로 독립인가? (독립성)

(4)이상치나 영향치가 존재하는가?

(5)오차항은 정규분포를 따르는가? (정규성)

(6)고려된 설명 변수 이외 다른 주요한 설명 변수가 존재하지는 않는가?

3.2 잔차의 종류

3.2.1 표준화 잔차

잔차의 표준화 값을 표준화 잔차(standardized residual)이라 하고 다음과 같이 정의한다.

$$z=\frac{r_i-\bar{r}}{\sqrt{MSE}}$$

위 식에서 알 수 있듯이 표준화 잔차는 추정 회귀식으로부터 관측치가 얼마나 떨어져 있나를 나타내는 것으로 ± 2 보다 크면 이상치(혹은 영향치)일 가능성이 높다. 잔차를 이용한 통계량은 표준화 잔차 이외에도 많이 존재하는데 이는 나중에 다루기로 한다.

3.2.2 스튜던트 잔차 (Studentized Residual)

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{MSE/1-h_{ii}}}, \quad h_{ii} = \underline{x}_i'(X'X)^{-1}\underline{x}_i$$

잔차를 t-분포를 따르는 통계량으로 만든 것으로 ±2 이면 이상치(혹은 영향치)로 판단하게 된다.

3.2.3 스튜던트 제외 잔차

자신의 관측치를 제외하고 회귀모형을 추정한 후 얻어진 잔차로 다음과 같이 정의한다.

$$r_{(i)} = \frac{y_i - \hat{y}_{(i)}}{\sqrt{MSE_{(i)}/1-h_{ii}}}, \quad h_{ii} = \underline{x}_i'(X'X)^{-1}\underline{x}_i$$

$\hat{y}_{(i)}$ 는 i-번째 관측치를 제외하고 얻은 추정 회귀모형으로부터 구한 예측치, $MSE_{(i)}$ 는 평균오차변동이다. 스튜던트화 제외 잔차가 ±2는 이 관측치는 이상치(혹은 영향치)로 판단하게 된다.

그럼 스튜던트 잔차와 스튜던트 제외 잔차 중 어는 것을 이용하여 이상치나 영향치를 판단하는가? 스튜던트 제외 잔차를 이용하는 것이 이상치를 더욱 많이 발견하게 되므로 일반적으로 스튜던트화 잔차를 이용한다.

3.3 잔차 진단

3.3.1 그래프

잔차분석의 6가지 이탈성에 대한 진단을 위하여 다음 그래프를 그릴 수 있다. 각각이 대응되는 것은 아님을 주의하기 바란다.

①잔차(Y-축)와 설명변수 산점도 Scatter plot of residual against independent variable: 설명변수와 잔차의 산점도는 함수 형태를 가져서는 안된다. 왜냐하면 오차와 설명변수가 종속변수를 설명하지 못하는 부분에 해당되기 때문이다. 단순회귀에서는 설명변수가 하나이므로 설명변수와 종속변수가 동일하므로(동일한 형태) 단순회귀에서는 이 산점도를 사용하지 않는다.

다중회귀에서는 오차의 등분산성 진단을 위하여 가끔 사용하기도 하지만 유효성이 의심되어 이 산점도는 다중회귀 잔차분석에서도 거의 사용하지 않는다.

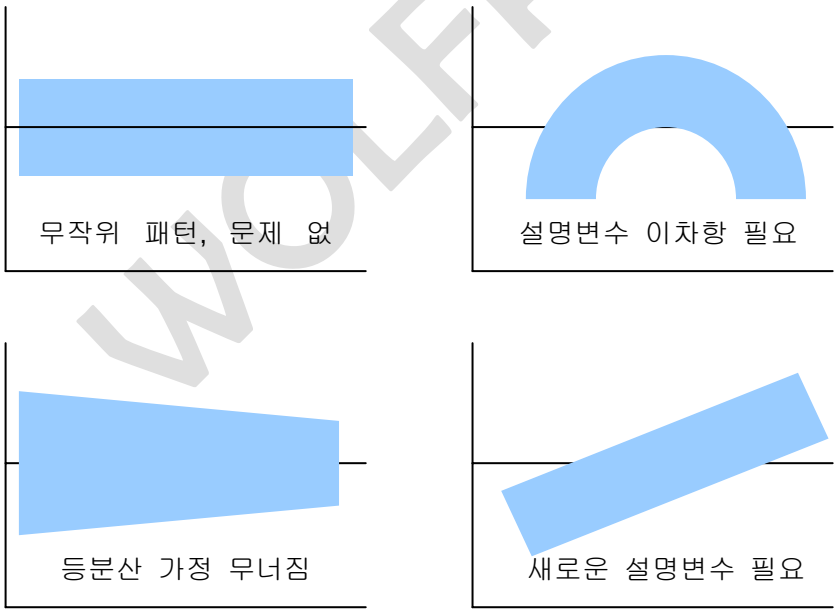
②**잔차와 종속변수 추정치 산점도**: 잔차를 Y축, 종속변수의 예측치를 X-축으로 하여 산점도를 그린다. 잔차는 추정된 회귀 모형이 종속 변수의 변동을 설명하지 못하는 부분에 해당하므로 산점도에 일정한 패턴이 있으면 안되고 평균 0을 중심으로 무작위(random) 하게 흩어져 있어야 한다. 그리고 잔차가 크다는 것은 그 관측치가 이상치 가능성 있다. 또한 이 산점도에 의해 등분산성, 선형성도 진단한다.

③잔차와 시간(time)의 시간도표(time plot): 시계열 데이터에만 국한된다.

④변수(관측치를 나누는 분류 변수) 수준별 잔차 그래프: 관측치를 분류할만한 변수가 있을 때에만(예: 성별) 가능하다. 즉 설명변수 이외에 분류형 변수(이를 지시변수라 한다. 다음에 다루기로 한다)가 있을 때만 그린다.

⑤**잔차에 대한 일변량 분석**: Stem and Leaf plot과 Shapiro-Wilks W-통계량(정규성), Box-Whisker plot(정규성, 이상치 혹은 영향치) 이상치나 영향치는 ②의 그래프에서 진단되므로 정규성만 검정하면 된다. 그러나 일반적으로 앞에서 언급하였듯이 회귀모형에서 비정규성은 큰 문제는 아니다.

그러므로 실제 잔차분석은 ②의 그래프만 이용하여 실시한다.



3.3.2 통계소프트웨어 이용하기

EXAMPLE 3-1 잔차분석

AD.xls (엑셀 데이터)

잔차분석을 위한 ⑴잔차와 종속변수 추정치 산점도 ⑵잔차의 정규성 검정을 실시하시오.

원 데이터 20개 모두 사용하여 회귀계수를 추정하고 이상치로 판단되면 제외하고 회귀모형을 다시 추정하여 추정하시오.

SAS

```
proc reg data=ad;
    model rate=spend/p r;
    plot rate*spend;
    output out=out1 predicted=yhat residual=res student=sres;
run;
```

MODEL 문장의 옵션 중 P, R을 사용하면 회귀계수 추정 결과와 함께 예측치, 잔차(스튜던트 잔차도 출력)가 출력된다. 이것을 이용하며 어느 관측치의 스튜던트 잔차가 2 이상인지 발견할 수 있다.

Output Statistics

Obs	Dep Var rate	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	99.6000	49.4857	5.8736	50.1143	23.335	2.148	|****|	0.146
2	11.7000	29.5862	6.2068	-17.8862	23.249	-0.769	| * |	0.021
3	21.9000	30.8935	6.0364	-8.9935	23.294	-0.386	| |	0.005
4	60.8000	52.4997	6.2504	8.3003	23.237	0.357	| |	0.005
5	78.6000	37.1393	5.4775	41.4607	23.431	1.769	|***|	0.086
6	92.4000	90.0836	14.5074	2.3164	19.198	0.121	| |	0.004
7	50.7000	32.3460	5.8668	18.3540	23.337	0.786	| * |	0.020
8	21.4000	29.9857	6.1530	-8.5857	23.263	-0.369	| |	0.005
9	40.1000	82.9300	12.7090	-42.8300	20.433	-2.096	|****|	0.850
10	40.8000	32.3823	5.8628	8.4177	23.338	0.361	| |	0.004
11	10.4000	39.1365	5.4019	-28.7365	23.449	-1.225	| ** |	0.040
12	88.9000	78.8266	11.7006	10.0734	21.027	0.479	| |	0.036
13	12.0000	24.3935	7.0247	-12.3935	23.015	-0.538	| * |	0.014
14	29.2000	40.6254	5.3811	-11.4254	23.454	-0.487	| |	0.006
15	38.0000	32.3460	5.8668	5.6540	23.337	0.242	| |	0.002
16	10.0000	24.6477	6.9802	-14.6477	23.028	-0.636	| * |	0.019
17	12.3000	25.3376	6.8614	-13.0376	23.064	-0.565	| * |	0.014
18	23.4000	25.9186	6.7638	-2.5186	23.093	-0.109	| |	0.001
19	71.1000	34.3432	5.6710	36.7568	23.385	1.572	|***|	0.073
20	4.4000	24.7929	6.9549	-20.3929	23.036	-0.885	| * |	0.036

OUTPUT 문장은 분석 결과를 SAS 데이터로 만드는데 사용된다. “OUT=”에는 만들어지는 SAS 데이터 이름을 지정한다. “PREDICTED=” (혹은 “P=”)는 종속변수 예측치의 변수이름 지정, “RESIDUAL=” (혹은 “R=”) 잔차, “STUDENT=”는 스튜던트 잔차 변수 이름을 지정하게 된다. 스튜던트 제외 잔차는 “RSTUDENT=” 사용한다.

만들어진 데이터 안에 어떤 변수와 관측치가 있는지 출력해 보자. 이 과정은 필요 없다. 단지 확인 상 보는 것이다.

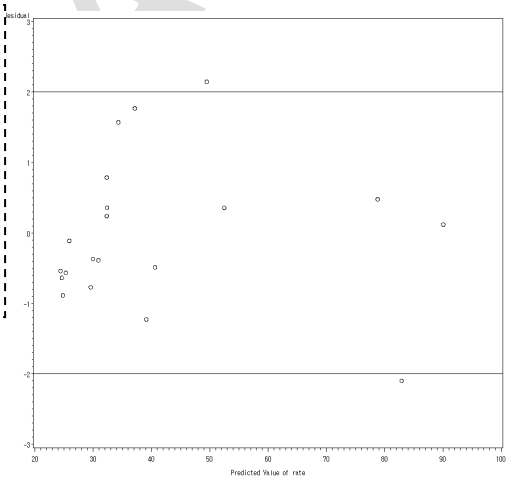
```
proc print data=out1;
run;
```

Obs	name	spend	rate	group	yhat	res	sres
1	PEPSI	74.1	99.6	1	49.4857	50.1143	2.14759
2	STROH'S	19.3	11.7	2	29.5862	-17.8862	-0.76934
3	FED'L EX	22.9	21.9	2	30.8935	-8.9935	-0.38609
4	BURGER K	82.4	60.8	3	52.4997	8.3003	0.35720
5	COCO-COL	40.1	78.6	1	37.1393	41.4607	1.76946
6	MC DONAL	185.9	92.4	3	90.0836	2.3164	0.12066

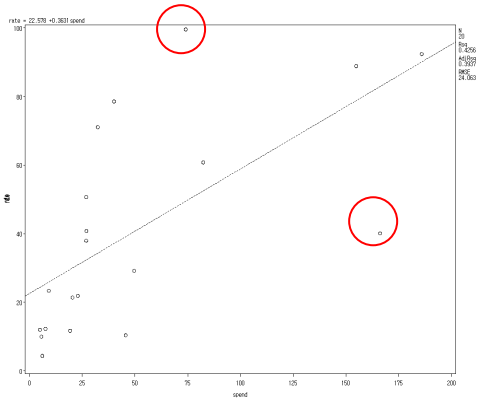
스튜던트 잔차와 종속변수의 예측치의 산점도를 그려보자. VREF 옵션은 수평 참조선을 긋는 옵션이다. 스튜던트 잔차가 ±2을 넘는 관측치가 2개이므로 이를 제외하자. 제외하고 회귀분석 다시 실시하려면 OUPUT 문장에 의해 저장하여 PROC PRINT 사용하는 것보다는 MODEL에서 옵션 사용하여 회귀계수 추정 결과와 함께 출력하는 것이 편리하다. 산점도에는 특별한 형태가 없으므로(무작위) 이상치 문제만 해결하면 된다.

잔차의 형태는 무작위로 보인다.
이상치가 +2보다 큰 것 하나, -2보다 작은 것 하나, 두 개의 이상치가 존재한다.
이 산점도는 이상치 판단의 통계량을 나타낸 것이므로 참조선을 벗어난 관측치는 이상치로 판단하면 된다.

```
proc gplot data=out1;
    symbol v=circle;
    plot sres*yhat/vref=2 -2;
run;
```



페이지 53, MODEL 문장의 R 옵션에 의해 출력된 결과를 보면 관측치1, 9는 이상치에 해당된다. PLOT 문장에 의해 그려진 산점도를 보자.



스튜던트 잔차에 의해 원의 두 관측치가 이상치로 밝혀졌다. 이를 제외하고 다시 회귀계수를 추정해 보자. REWEIGHT 문장은 관측치를 제외하고 분석하라는 의미이다.

```
proc reg data=ad;
     model rate=spend/p r;
     reweight obs.=1; reweight obs.=9;
     output out=out1 predicted=yhat residual=res student=sres;
run;
proc gplot data=out1;
     symbol v=circle;
     plot sres*yhat/vref=2 -2;
run;
```

산점도를 보면 다시 2개(관측치 5, 19)가 이상치이다. 이를 제외하고 다시 회귀분석 하면 1개 관측기치 이상치(관측치 7), 이를 제외하고 재분석 하면 관측치11가 이상치이다. 이제 더 이상 이상치는 없다.

```
proc reg data=ad;
     model rate=spend/p r;
     reweight obs.=1; reweight obs.=9;
     reweight obs.=5; reweight obs.=19;
     reweight obs.=7;
     reweight obs.=11;
     output out=out1 predicted=yhat residual=res student=sres;
run;
proc gplot data=out1;
     symbol v=circle;
     plot sres*yhat/vref=2 -2;
run;

proc univariate data=out1 normal;
     var res;
run;
```

이상치가 존재하지 않는 것이 확인되면 오차의 정규성을 잔차에 대한 UNIVARIATE 분석을 통하여 이용하여 검정한다. Shapiro-Wilk 정규성 검정통계량의 유의확률을 보면 0.17 이므로 귀무가설이 채택되어 정규분포를 따른다고 할 수 있다. 정규성 검정 통과

정규성 검정

검정		----통계량----		-------p-값-------
Shapiro-Wilk	W	0.932641	Pr < W	0.1736
Kolmogorov-Smirnov	D	0.160231	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.115378	Pr > W-Sq	0.0676
Anderson-Darling	A-Sq	0.667695	Pr > A-Sq	0.0729

20개 관측치 중 6개가 이상치로 제외되었으므로 14개의 관측치만 이용하여 최종 회귀모형을 얻었다. 20개 모두 사용하여 얻은 추정 기울기가 0.86이었는데, 이상치를 제외한 후 추정된 회귀모형은 기울기는 0.47로 변했음을 알 수 있다. 광고비를 1단위 증가시키면 평가도는 0.47만큼 증가한다.

최종 회귀모형: $\underset{(t=3.97,\ p=0.0015)}{\text{평가} = 12.34 + 0.47 * \text{광고비}}$ (페이지 44와 비교)

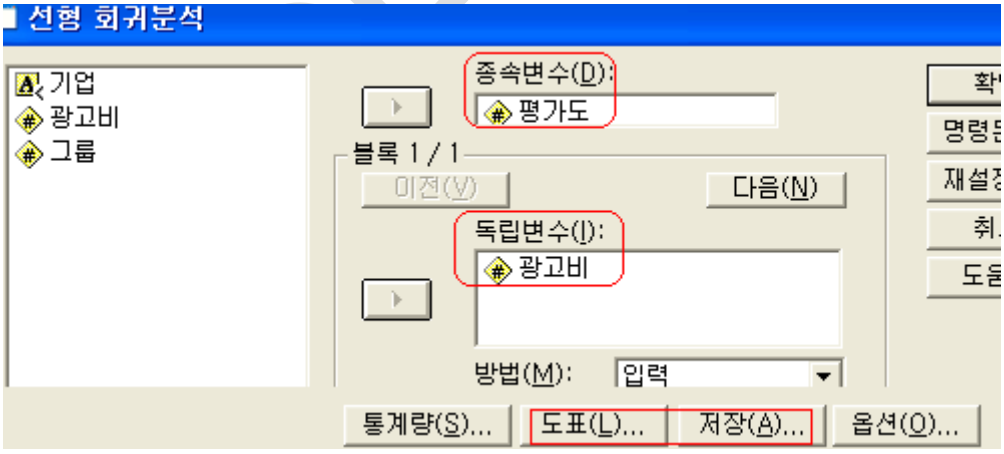
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9618.53063	9618.53063	124.66	<.0001
Error	12	925.89795	77.15816		
Corrected Total	13	10544			

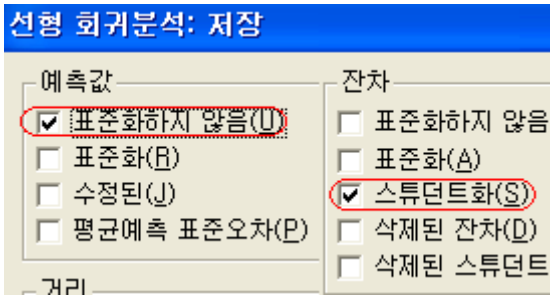
Parameter Estimates

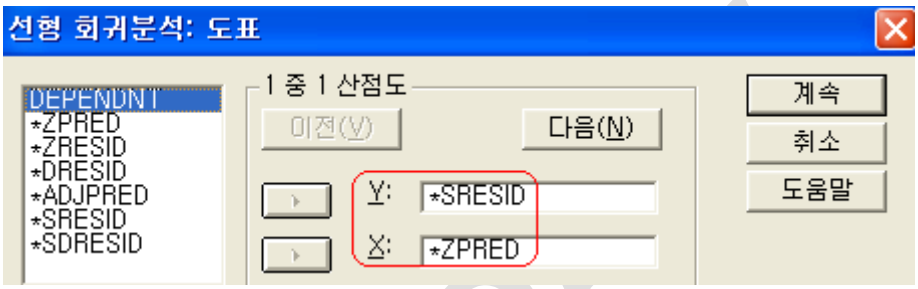
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.34738	3.00949	4.10	0.0015
spend	1	0.47245	0.04231	11.17	<.0001

SPSS 분석(A) ▶ 회귀분석(R) ▶ 선형(L)... 메뉴를 선택하고 나타난 "선형 회귀분석" 창에서 종속변수와 독립변수를 선택한다.

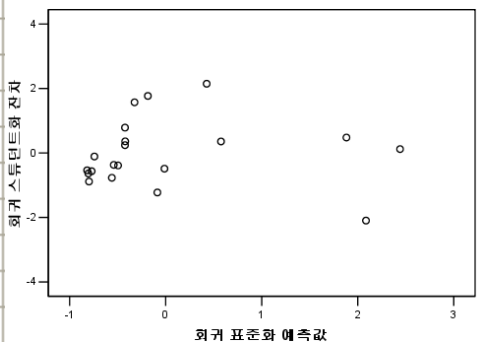


저장(A)... 옵션에서는 잔차와 예측치를 변수로 저장하고 도표(L)...에서는 산점도를 그린다. X-축의 변수로는 예측치를 사용하나 SPSS에서는 표준화된 종속변수 예측치만 있어 이것을 대신 사용하였다. 상관 없다.

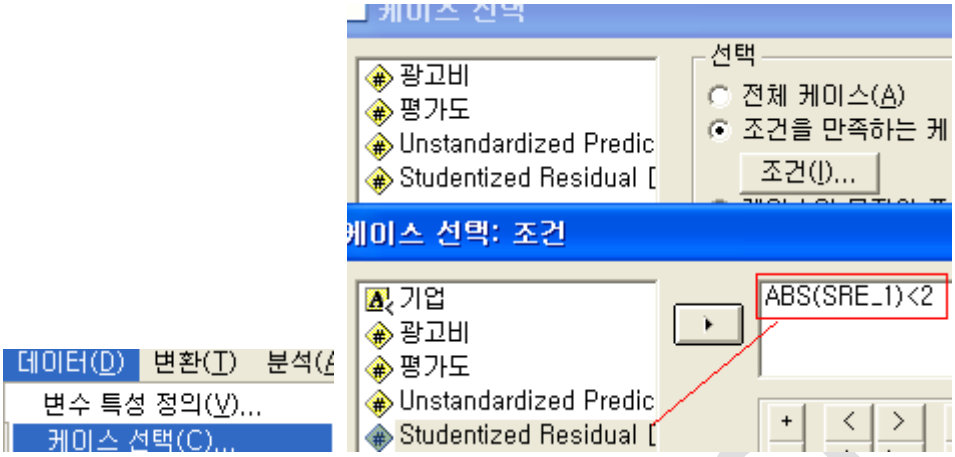




기업	광고비	평가도	PRE_1	SRE_1
PEPSI	74.10	99.60	49.4857	2.14759
STROH'S	19.30	11.70	29.5862	-.76934
FED'L EX	22.90	21.90	30.8935	-.38609
BURGER K	82.40	60.80	52.4997	.35720
COCO-COL	40.10	78.60	37.1393	1.76946
MC DONAL	185.90	92.40	90.0836	.12066
MCI	26.90	50.70	32.3460	.78648
DIET COL	20.40	21.40	29.9857	-.36907
FORD	166.20	40.10	82.9300	-2.0961
LEVI'S	27.00	40.80	32.3823	.36069
BUD LITE	45.60	10.40	39.1365	-1.2255
ATT/BELL	154.90	88.90	78.8266	.47908
CALVIN K	5.00	12.00	24.3935	-.53850
WENDY'S	49.70	29.20	40.6254	-.48715
POLAROID	26.90	38.00	32.3460	.24228
SHASTA	5.70	10.00	24.6477	-.63607
MEOW MIX	7.60	12.30	25.3376	-.56528
OSCAR ME	9.20	23.40	25.9186	-.10906
CREST	32.40	71.10	34.3432	1.57180
KIBBLES	6.10	4.40	24.7929	-.88526



산점도가 특별한 형태를 띄지 않으므로 이상치 문제만 해결하면 된다. 스튜던트 잔차의 절대값이 ±2 이상인 관측치가 이상치로 판단하므로 이를 제외하자. ABS는 절대값을 구하는 함수이다.



관측치 1, 9는 이제부터 제외된다. 다시 회귀모형을 추정하면 SAS 결과에서 본 것처럼 (5, 19), 7, 11이 차례로 제외된다. 최종 결과는 SAS와 동일하다.

	기업	광고비	평가도	PRE_1	SRE_1	filter_$
1	PEPSI	74.10	99.60	49.4857	2.14759	0
2	STROH'S	19.30	11.70	29.5862	-.76934	1
3	FED'L EX	22.90	21.90	30.8935	-.38609	1
4	BURGER K	82.40	60.80	52.4997	.35720	1
5	COCO-COL	40.10	78.60	37.1393	1.76946	1
6	MC DONAL	185.90	92.40	90.0836	.12066	1
7	MCI	26.90	50.70	32.3460	.78648	1
8	DIET COL	20.40	21.40	29.9857	-.36907	1
9	FORD	166.20	40.10	82.9300	-2.0961	0
10	LEVI'S	27.00	40.80	32.3823	.36069	1
11	BUD LITE	45.60	10.40	39.1365	-1.2255	1

HOMEWORK #4-1 DUE 3월 30일(수)

CANCER.txt (텍스트 데이터)

연 평균 온도(F: Fahrenheit, 설명변수)가 여성 종양 사망지수(mortality index)에 영향을 미치는지 알아보기 위하여 유럽 몇 지역을 대상으로 조사한 자료이다. **SPSS 이용하기**

HOMEWORK#3 산점도만 보고 관측치 하나만 제외했는데 여기서는 스튜던트 잔차를 이용하여 이상치를 판단하고 결과를 비교하시오.

3.3.4 정규성 검정

진단방법 ①오차의 추정치 잔차의 정규성 검정 실시: 그래프 stem and leaf, Q-Q plot 등을 을 이용하기도 하지만 S-W통계량이나 K-S통계량을 이용한다.
②잔차와 예측치 산점도, 이차 함수 형태

해결방법 ①종속변수변환, $Ln(y)$ 변환이나 y^2 변환이 가장 일반적이다.

앞에서 언급하였듯이 정규성 파괴는 그렇게 큰 문제가 아니다. 데이터 크기가 20개 이상이면 정규성 검정은 생략해도 문제가 되지 않는다.

EXAMPLE 3-6 정규성 검정

3.3.3절 예제 데이터의 경우 종속변수를 LOG 변환하여 $\ln(y)$ 와 설명변수(X)의 선형회귀모형을 적합을 시켜 유의함을 알았다. 잔차의 정규성을 검정해 보자.

```
proc reg data=two;
    model ly=x;
    plot ly*x;
    plot student.*predicted./vref=-2 vref=2;
    output out=out1 r=res;
run;

proc univariate data=out1 normal plot;
    var res;
run;
```

아래 출력 결과를 보면 유의확률이 0.31이므로 잔차는 정규분포를 따른다는 가정을 기각하지 못한다.(오차에 대한 정규성 검정 만족) 줄기-잎 그림이나 나무-상자 그림(이상치, 치우침)은 참고 그래프로 이용할 수 있으나 최종 판단은 S-W 검정통계량을 이용한다.

정규성 검정

검정		----통계량----		-------p-값-------
Shapiro-Wilk	W	0.915164	Pr < W	0.3184
Kolmogorov-Smirnov	D	0.172388	Pr > D	>0.1500

```
줄기 잎                    #
   2 028                   3
   1
   0 456                   3
  -0 97                    2
  -1
  -2
  -3 54                    2
     ----+----+----+----+
        값 : (줄기.잎)*10**-1
```

HOMEWORK #5-2 DUE 4월 6일(수)

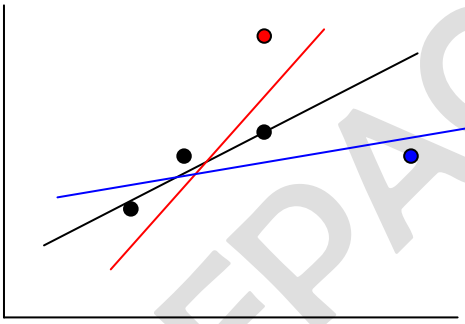
HOMEWORK#4-1, #4-2에서 정규성 검정을 실시하고 문제가 있으면 해결하시오.

3.3.5 영향치나 이상치 존재 여부

진단방법 ①표준화 잔차(잔차를 표준오차 $\sqrt{MSE}$ 로 나눈 값)와 예측치의 산점도

해결방법 ①영향치(influential obs.) 는 영향치를 포함하고 회귀모형을 추정하고 제외하고 추정하여 모두 제시한다.

②이상치(outlier)는 제외하고 모형을 추정한다.



영향치나 이상치는 모두 다른 관측치에 비해 오차(잔차)가 큰 관측치라는 점에서 공통점이 있으나 이상치(빨강 관측치)는 비교할 대상이(설명 변수 관계 속에서) 있어 그 값들에 비해 값이 매우 크거나 작아 회귀 계수 추정 값을 변화시킨다. 한편 영향치(파랑 관측치)는 회귀 계수 추정 값을 변화시키지만 비교 대상이 되는 관측치가 없으므로 이상치인지 판단할 수 없는 경우에 해당된다. 영향치가 존재하는 경우에는 (1)영향치를 제외하고 회귀모형을 추정하고 (2)포함하여 회귀 모형을 추정한 두 가지 모두를 제시하는 것이 옳다.

이상치도 정보를 가진 관측치이다. 회귀모형 적합을 위해서는 제외하지만 왜 이 관측치가 다른 관측치에 비해(설명변수 기준) 종속변수의 값이 크거나 작은지 파악하여 정보를 얻고 이를 연구 결과나 의사결정에 반영할 필요가 있다.

EXAMPLE 3-7 이상치 문제 해결

NFL 데이터의 경우 종속변수를 LOG 변환하여 $\ln(y)$ 와 설명변수(X)의 선형회귀모형을 적합을 시켰고 유의함을 알았다. 또한 잔차 분석 결과 모든 것이 유의하였다. 이제 영향치나 이상치가 있으면 이를 제외하고 최종 회귀 모형을 얻어 보자.

종속변수 SALARY를 로그변환 하자.

```
data nfl1;
    set nfl;
    log_sa=log(salary);
run;
```

LOG(salary)를 종속변수, draft를 설명변수로 하여 회귀모형을 적합을 하자. 표준화 잔차를 볼 수 있는 방법은 MODEL에서 R을 사용하거나 표준화 잔차와 예측치 산점도를 이용하면 된다. 참조선을 ±2로 하여 그릴 수 있다.

```
proc reg data=nfl1;
    model log_sa=draft/r;
    plot student.*predicted./vref=2 vref=-2;
run;
```

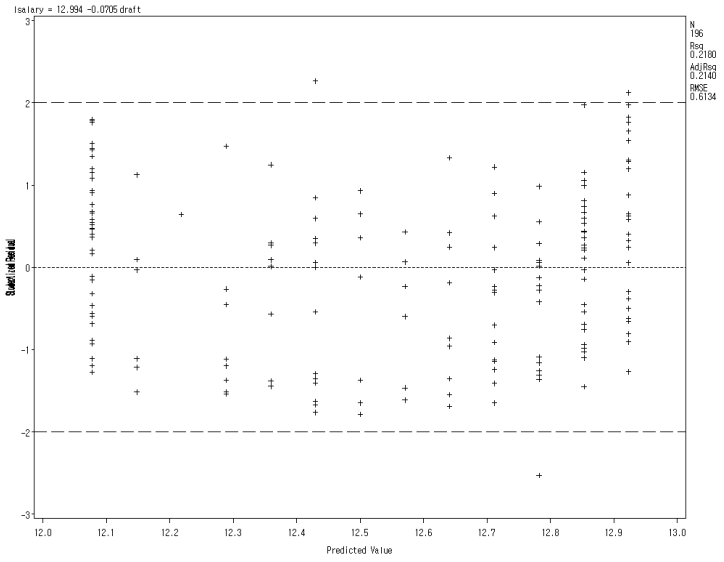
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	20.34977	20.34977	54.09	<.0001
Error	194	72.99331	0.37625		
Corrected Total	195	93.34309			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.99407	0.07534	172.46	<.0001
draft	1	-0.07051	0.00959	-7.35	<.0001

Obs	Dep Var lsalary	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2
172	12.0137	12.2889	0.0558	-0.2752	0.611	-0.451	
173	12.8347	12.7825	0.0546	0.0522	0.611	0.0854	
174	14.0387	12.9236	0.0678	1.1151	0.610	1.829	***
175	12.3673	12.3594	0.0504	0.007893	0.611	0.0129	
176	12.7068	12.7825	0.0546	-0.0757	0.611	-0.124	
177	11.6082	12.6415	0.0458	-1.0333	0.612	-1.689	***
178	11.3621	12.2889	0.0558	-0.9268	0.611	-1.517	***
179	12.8347	12.8530	0.0608	-0.0184	0.610	-0.0301	

Plot of Stud.Res*Yhat



이상치가 3개 나타났다. STUDENT Res.에서 절대값이 2 이상인 관측치를 찾으면 12, 44, 162번째 관측치다.

REWEIGHT 문을 사용하면 그 관측치들은 제외하고 회귀모형을 추정한다.

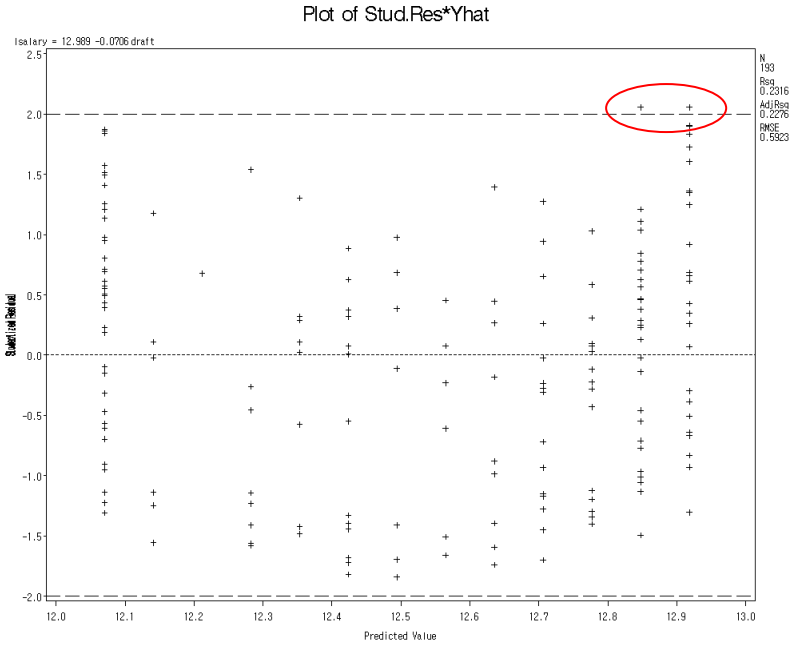
```
proc reg data=nfl1;
    model lsalary=draft/r;
    reweight obs.=12;
    reweight obs.=44;
    reweight obs.=162;
    plot student.*predicted./vref=2 vref=-2;
run;
```

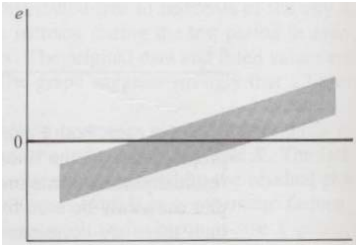
분산분석을 보면 총변동의 자유도는 192(총 관측치 수는 193)로 이전 페이지보다 3개 줄었다. 이는 관측치 3개가 제외되었기 때문이다.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	20.19393	20.19393	57.57	<.0001
Error	191	66.99944	0.35078		
Corrected Total	192	87.19337			

이상치를 제외하고 표준화 잔차를 재계산 하면 다시 이상치가 생길 가능성이 있다. (아래 산점도에서 타원 부분) 이상치가 없을 때까지 위의 방법으로 계속 이상치를 제거하면 된다. 이상치 제거가 계속되면 어디에서 멈출까? 기준을 높여 ± 2.5 수준으로 결정하시오.



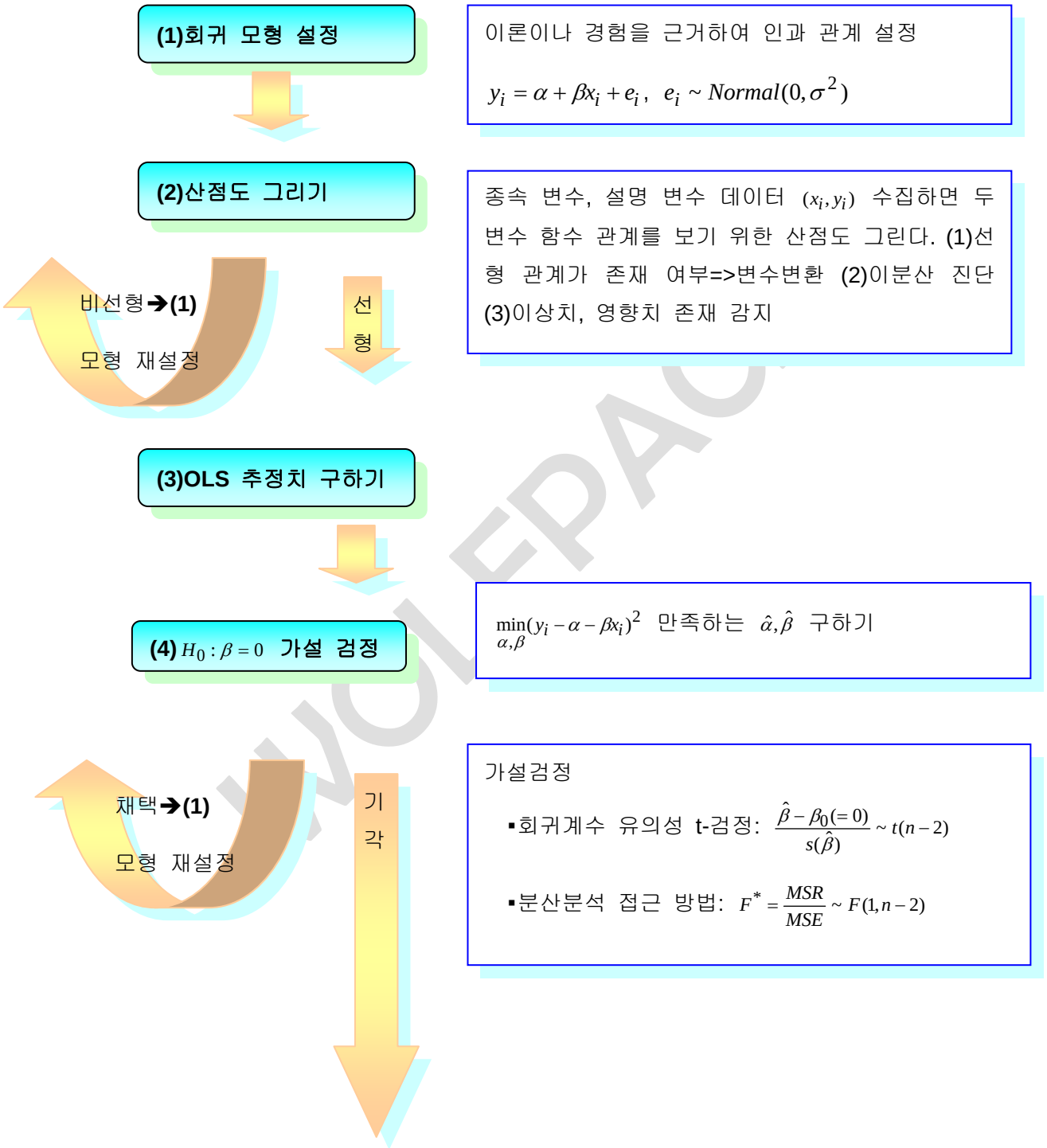


3.3.6 설명변수 누락

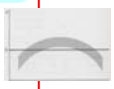
고려된 설명 변수 이외에 종속 변수에 영향을 미치는 설명 변수가 존재한다면 잔차와 추정치 산점도는 오른쪽과 같다. 이런 경우 이론이나 경험을 바탕으로 새로운 변수를 회귀모형에 고려해야만 한다.

3.4 회귀 분석 절차

3.4.1 단순회귀분석 절차



(5)선형성 진단



도구: 잔차와 예측치($\hat{Y}$) 산점도
진단: 포물선이나 3 차식 형태
해결책: 설명변수 이차항(X^2)이나 삼차항(X^3)

Go to (3)

(6)등분산성 검정

도구: 잔차와 예측치($\hat{Y}$) 산점도
진단: 나팔 모양
해결책: WLS($1/x_i^2, 1/\hat{y}_i^2$) 혹은 종속변수 변환

Go to (1)

(7)주요 변수 존재

도구: 잔차와 예측치($\hat{Y}$) 산점도
진단: 직선 경향
해결책: 새로운 설명변수 고려해야 한다.

Go to (3)

(8)이상치 진단

도구: 표준화 잔차와 예측치 산점도
진단: ± 2 넘는 관측치, 이상치나 영향치 구별
해결책: 이상치 제거, 영향치 진단

(9)정규성 검정

도구: 잔차에 대한 일변량 분석
진단: 줄기-잎 그림, Shapiro-Wilk W-통계량
해결책: 적절한 종속변수 변환

(10)최종 회귀 모형

(1)추정 회귀식 제시(유의확률과 함께) $\underset{(p=...)}{\hat{y} = \hat{\alpha} + \hat{\beta}x}$, 결정계수 R^2

(2)선형 회귀 모형에서 기울기 $\hat{\beta}$에 대한 해석

(3)관심 있는 설명변수 값에 대한 종속변수 예측치와 신뢰구간

3.4.2 CANCER 데이터 회귀 분석하기

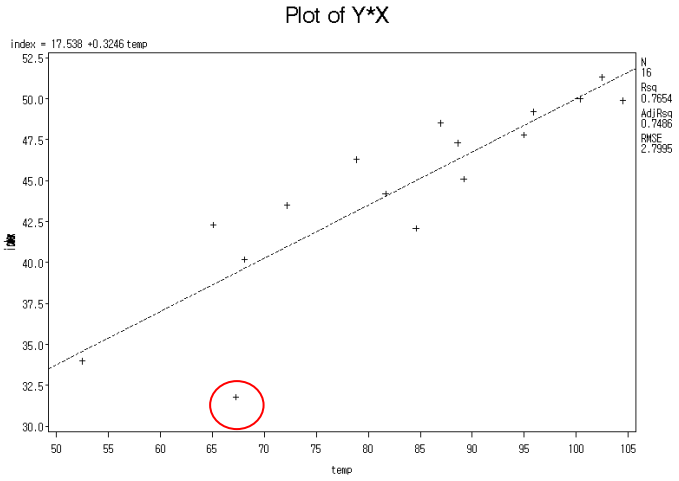
EXAMPLE 3-8 회귀분석 하기

연 평균 온도가 암 사망지수에 영향을 미치는지 알아보자. CANCER.txt

순서(1)-(2) 종속변수와 설명변수 산점도

위의 산점도를 살펴본 결과 (1)선형 관계는 존재하는 것 같다. 만약 산점도의 형태가 변수변환이 필요하다고 판단되면 (종속)변수 변환 후 회귀분석을 실시하면 된다. 변수변환을 해야 하는지 여부는 잔차 분석에서도 판단되지만 산점도에 의해 변수변환이 필요하다고 판단되면 변환을 하는 것이 좋다. (2)이분산 문제는 없어 보인다. (3)영향치는 없으나 이상치가 존재하는 것 같다.

(1)선형 관계
(2)등분산 가정 만족
(3)이상치 존재




```
proc reg data=cancer;
    title 'Plot of Y*X';
    model index=temp;
    plot index*temp;
run;
```

순서(3)-(4) 회귀계수 유의성 검정

회귀계수 유의성($H_0 : \beta = 0, H_a : \beta \neq 0$) 검정이나 설명변수의 유의성(설명변수가 종속변수를 선형적으로 설명한다) 검정은 동일하다. 선형 회귀의 경우 설명변수의 유의성 검정을 위한 t-검정이나 분산분석적 유의성 검정F-검정은 동일하다.

귀무가설 $H_0 : \beta = 0$ 에 대한 가설 검정 결과 p-값이 0.05보다 작으므로 통계적으로 유의하다. 즉 온도는 사망 지수에 양의 영향을 미치고 1도 올라가면 사망 지수가 0.325 높아진다. 잔차 분석 실시하기 전에 잠정적으로 다음 회귀 모형은 적절하다고 할 수 있다.

$$Index = 17.58 + 0.32 * Temp$$
$$(t = 4.32, p = 0.001) \quad (t = 6.76, p < 0.0001)$$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	357.92491	357.92491	45.67	<.0001
Error	14	109.72447	7.83746		
Corrected Total	15	467.64938			

Root MSE	2.79955	R-Square	0.7654
Dependent Mean	44.59375	Adj R-Sq	0.7486
Coeff Var	6.27789		

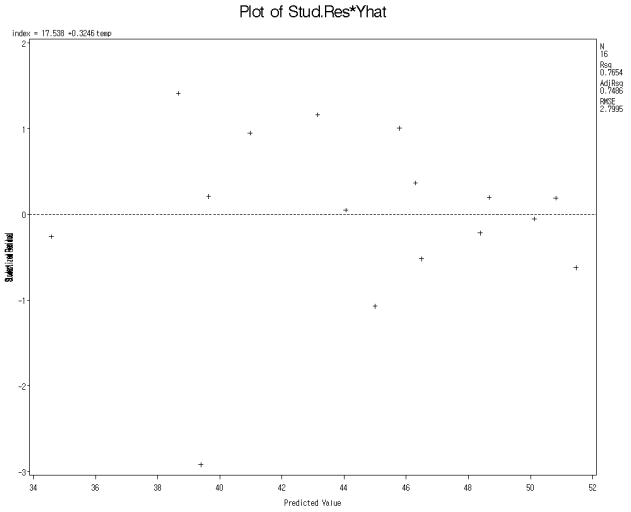
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.53816	4.06430	4.32	0.0007
temp	1	0.32463	0.04804	6.76	<.0001

만약 회귀계수가 유의하지 않으면 다음 순서로 갈 필요는 없다. 아직 추정 회귀모형을 발표하기에는 이르다. 모형에 대한 가정 진단 및 이상치 발견을 위한 잔차 분석 과정이 남아 있기 때문이다.

순서(5)-(7) 잔차분석

```
proc reg data=cancer;
     title 'Plot of Stud.Res*Yhat';
     model index=temp;
     plot student.*predicted.;
run;
```

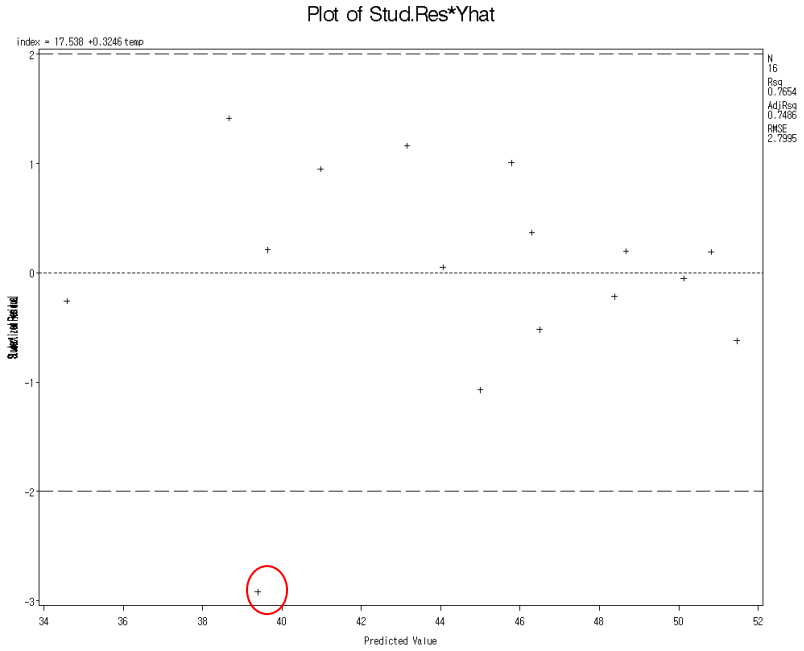


선형성, 등분산성이 무너질만한 특별한 패턴이 존재하지 않고, 변수변환이 필요한 것 같지도 않다. 회귀모형의 선형성과 오차의 등분산 가정이 성립한다.

순서(8) 이상치 진단

이상치 존재여부는 스튜던트 잔차가 ±2 이상인지에 대한 판단으로 결정한다. 다음 프로그램을 실행하자. P는 종속변수 예측치 R은 잔차, 표준화 잔차, 스튜던트 잔차 등을 출력 창에 출력하라는 명령이고 VREF, HREF 옵션은 산점도에 참조선을 긋는 옵션이다.

```
proc reg data=cancer;
    model index=temp/r p;
    plot student.*predicted./vref=2 href=2;
run;
```



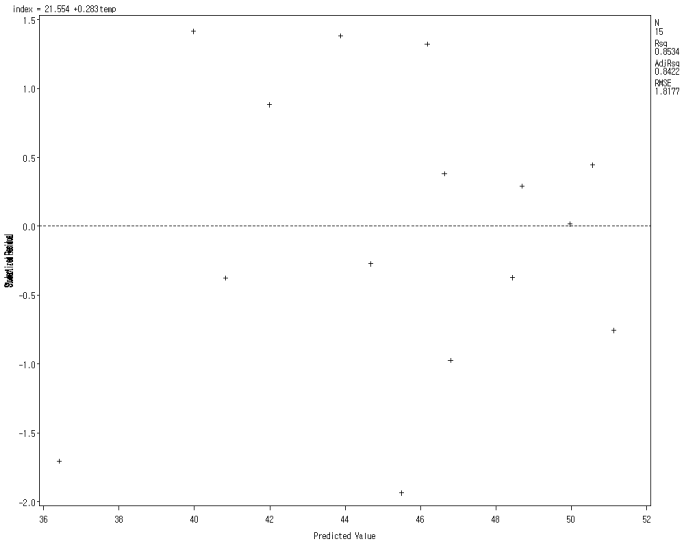
위의 산점도를 보면 이상치가 하나 존재한다. 출력 창의 잔차 출력 결과를 보면 15번째 관측치가 이상치이다. 이는 순서(2) 산점도(종속변수와 설명변수)에 의해서도 이미 예상된 결과이다. 이상치인지 영향치인지 여부는 종속변수와 설명변수의 산점도에 의해 결정된다.

Output Statistics

Obs	Dep	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	51	0.4876	2.550	0.191		0.004
13	42	3.6287	2.565	1.415	**	0.191
14	40	0.5548	2.610	0.213		0.003
15	31	-7.5855	2.599	-2.919	*****	0.684
16	34	-0.5811	2.270	-0.256		0.017

다음은 관측치 15번째를 제외하고 회귀분석을 실시한 결과이다. 스튜던트 잔차가 ±2 이상인 관측치는 더 이상 존재하지 않는다. 이상치는 없다고 할 수 있다.

```
proc reg data=cancer;
    title 'Plot of Stud.Res*Yhat(reweight)';
    model index=temp/r;
    reweight obs.=15;
    plot student.*predicted./vref=-2 vref=2;
run;
```



Residual	Std Error Residual	Student Residual
0.7340	1.655	0.444
-1.2321	1.630	-0.756
0.0284	1.677	0.0169
0.5021	1.716	0.293
2.3212	1.754	1.323
-0.6432	1.722	-0.374
0.6683	1.751	0.382
-1.7015	1.749	-0.973
2.4139	1.747	1.382
-3.3995	1.756	-1.936
-0.4787	1.754	-0.273
1.5103	1.710	0.883
2.3199	1.640	1.415
-0.6293	1.674	-0.376
-8.8028	.	.
-2.4138	1.416	-1.704

순서(9) 정규성 검정

오차의 정규성을 검정하기 위하여 OUTPUT 옵션에 의해 스튜던트 잔차를 SRES 변수명으로, 잔차는 RES 변수명으로 SAS 데이터 OUT1에 저장하였다. 일반적으로 잔차를 이용하여 오차의 정규성을 검정하지만 SAS의 경우 REWEIGHT에 의해 이상치를 제외한 경우 제외된 이상치라도 출력 결과 창에는 결측치로 나오나 OUT1에는 잔차가 계산되어 저장되어 있다. 그러므로 스튜던트 잔차로 해야 된다. (차이는 무시할 만 하다) 물론 SPSS에서는 "표준화 하지 않은" 잔차를 이용하면 된다.

Obs	temp	index	res	sres
1	102.5	51.3	0.73400	0.44363
2	104.5	49.9	-1.23209	-0.75592
12	102.2	50.9	0.02829	0.01694
13	65.1	42.3	2.31987	1.41475
14	68.1	40.2	-0.62926	-0.37592
15	67.3	31.8	-8.80283	.
16	52.5	34.0	-2.41377	-1.70433

잔차에 대한 정규성 검정을 위하여 NORMAL(통계량 출력), PLOT(줄기-잎 그림, 상자 수염 그림)을 그리게 했다.

```
proc reg data=cancer;
     model index=temp/r p;
     reweight obs.=15;
     plot student.*predicted./vref=2 href=2;
     output out=out1 student=sres r=res;
run;
proc univariate data=out1 normal plot;
     var sres;
run;
```

시각적 도움은 얻을 수 있으나 줄기-잎 그림이나 상자-수염 그림에 의해서는 정규 분포를 따르는지(아니 적어도 좌우 대칭이 되는지) 알 수 없을 뿐 아니라 정규분포를 따르는지에 대한 유의성을 판단할 수 없다. 유의확률이 0.05보다 크므로 귀무가설(정규분포를 따른다)이 채택되어 정규성 가정은 만족함을 알 수 있다. S-W W-검정통계량과 K-S D-검정통계량(Goodness of fits 검정) 중 어는 것을 사용할 것인가? 일반적으로 W-검정통계량을 사용한다.

정규성 검정

검정		----통계량----		-------p-값-------
Shapiro-Wilk	W	0.952303	Pr < W	0.5614
Kolmogorov-Smirnov	D	0.099418	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.025602	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.226219	Pr > A-Sq	>0.2500

```
줄기 잎                    #        상자그림
   1 344                   3           |
   0 9                     1        +-----+
   0 0344                  4        +-----+
  -0 443                   3        |  +  |
  -0 8                     1        +-----+
  -1 0                     1           |
  -1 97                    2           |
     ----+----+----+----+
```

순서(10)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	250.10693	250.10693	75.70	<.0001
Error	13	42.95040	3.30388		
Corrected Total	14	293.05733			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21.55392	2.78591	7.74	<.0001
temp	1	0.28304	0.03253	8.70	<.0001

관측치 개수는 16개였으나 1개가 이상치로 판단되어 15개 데이터만 이용하여 회귀분석을 실시하여 다음 결과를 얻었다. 평균 온도는 암 사망 지수에 양의 영향을 미치고 온도가 1도 올라가면 사망지수가 0.283만큼 높아진다.

$$Index = 21.55 + 0.283 * Temp \quad (t = 8.7,\ p < 0.0001), \quad R^2 = 0.85$$

만약 온도가 90도인 국가의 암 사망지수를 예측하고 95%신뢰구간을 구하려면 다음과 같이 하면 된다. P는 예측치, CLM은 평균에 대한 신뢰구간을 출력하라는 명령이다.

```
52.5    34
90 .
run;
```

```
proc reg data=cancer;
    model index=temp/ p clm;
    reweight obs.=15;
run;
```

데이터 마지막 라인에 예측을 원하는 설명변수 값을 지정하고 종속변수는 결측치를 의미하는 . 을 찍는다.
그러면 이 관측치는 회귀모형 추정에 사용되지는 않고 예측 결과만 출력된다.
위의 회귀모형에서 TEMP에 90을 넣으면 47.3이 나오는데 이는 아래 결과와 일치한다.

Output Statistics

Obs	Weight Variable	Dep Var index	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	1.0000	51.3000	50.5660	0.7526	48.9400	52.1920
2	1.0000	49.9000	51.1321	0.8045	49.3940	52.8702
12	1.0000	43.5000	41.9897	0.6149	40.6613	43.3182
13	1.0000	42.3000	39.9801	0.7842	38.2859	41.6743
14	1.0000	40.2000	40.8293	0.7084	39.2988	42.3598
15	0	31.8000	40.6028	0.7281	39.0298	42.1759
16	1.0000	34.0000	36.4138	1.1393	33.9524	38.8752
17	0	.	47.0279	0.5033	45.9407	48.1152

HOMEWORK #5-3 DUE 4월 6일(수)

3.4.2절 CANCER 데이터 회귀분석 작업을 SPSS로 시행하시오.