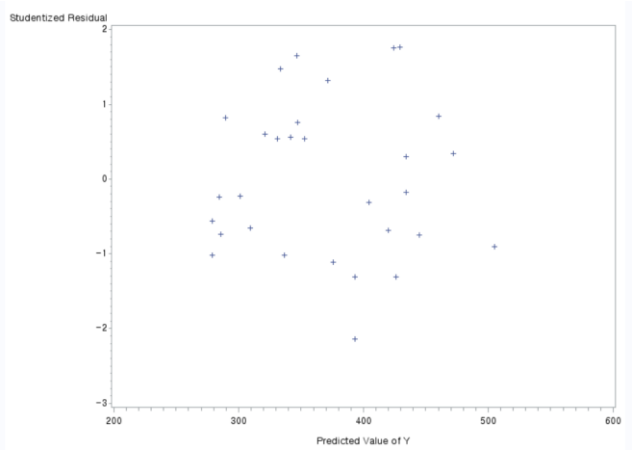
오유진(2015325) 회귀분석론 과제1

5.2

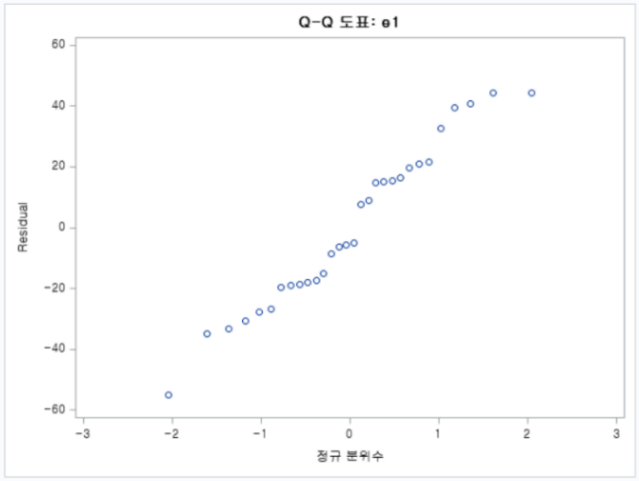
5.2.1

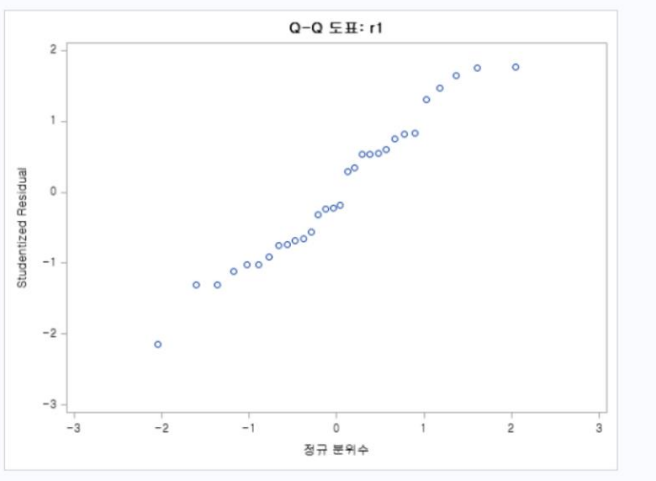
헬스클럽자료 중 Y를 설명변수, X1,X2,X3,X4를 설명변수라고 두고 스튜던트화잔차와 적합값의 잔차산점도는 아래와 같다.



이는 회귀모형이 적절하게 적합되었을 때 보여지는 잔차산점도라고 볼 수 있다.

5.2.2 정규확률그림





두 그림 모두 직선에서 크게 벗어났다고 볼 수 없다.

5.2.3

* 오차의 등분산성

스코어검정을 통해 오차의 등분산을 알아보고자한다.

σ^2=SSE/표본의수 = 20551/30=685.03333 이라는 것을 구해서, ei^2/685.03333를 새로운 변수인 u로 두고 u를 각각의 설명변수 x1, x2, x3, x4 로 회귀하였고 Y와 모든 설명변수의 등분산성을 보기위해 u를 4개의 설명변수 (x1, x2, x3, x4)로 회귀하여 진행했다.

아래 사진들은 u를 x1, x2, x3, x4, 4개 모두(x1 x2 x3 x4)로 회귀시켰을 때 결과이다.

The REG Procedure
Model: MODEL1
Dependent Variable: u

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.02726	0.02726	0.02	0.8804
Error	28	33.08005	1.18143		
Corrected Total	29	33.10731			

The REG Procedure
Model: MODEL2
Dependent Variable: u

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.48178	0.48178	0.41	0.5254
Error	28	32.62553	1.16520		
Corrected Total	29	33.10731			

The REG Procedure
Model: MODEL3
Dependent Variable: u

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.31474	1.31474	1.16	0.2911
Error	28	31.79256	1.13545		
Corrected Total	29	33.10731			

The REG Procedure
Model: MODEL4
Dependent Variable: u

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.05400	0.05400	0.05	0.8322
Error	28	33.05330	1.18048		
Corrected Total	29	33.10731			

The REG Procedure
Model: MODEL5
Dependent Variable: u

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6.94314	1.73579	1.66	0.1911
Error	25	26.16416	1.04657		
Corrected Total	29	33.10731			

각각의 설명변수 스코어 검정 통계량은 x1, x2, x3, x4, 그리고 전체 변수 x1 x2 x3 x4 순으로 각각 0.02726/2=0.01363, 0.48178/2=0.24089, 1.31474/2=0.65737, 0.05400/2=0.027 이다. 그리고 p-값이 각각 0.8804, 0.5254, 0.2911, 0.8322, 0.1911 로 0.05보다 크므로 귀무가설을 기각하지 못한다. 이때 x1과 x3, x4는 매우 크므로 이분산의 현상은 몸무게(x1), 근력(x3), 1/4마일 시험주행속도(x4) 과는 무관하다. 그리고 x2를 보면 분산은 분당 정지 맥박수의 함수는 아니라는 것을 알 수 있고, x1, x3, x4와는 무관하므로 x2의 영향을 받는 다는 것을 알 수 있다. 4개 전체를 회귀한 것을 보면 분산이 어느정도 4개의 함수라는 것을 알 수 있다.

*모형의 선형성

잔차산점도를 통해 모형의 선형성을 진단할 수 있다.

위 5.2.1의 잔차산점도를 보면, 스튜던트화잔차값이 곡선의 형태를 보이지 않기 때문에 비선형성을 나타낸다고 볼 수 있다.

*오차의 정규성

UNIVARIATE 프로시저
변수: e1 (Residual)

적률			
N	30	가중합	30
평균	0	관측값 합	0
표준 편차	26.6208121	분산	708.667637
왜도	0.05541453	첨도	-0.8358721
제곱합	20551.3615	수정 제곱합	20551.3615
변동계수	.	평균의 표준 오차	4.8602731

정규성 검정				
검정	통계량		p 값	
Shapiro-Wilk	W	0.960525	Pr < W	0.3195
Kolmogorov-Smirnov	D	0.114788	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.074372	Pr > W-Sq	0.2405
Anderson-Darling	A-Sq	0.437994	Pr > A-Sq	>0.2500

UNIVARIATE 프로시저
변수: r1 (Studentized Residual)

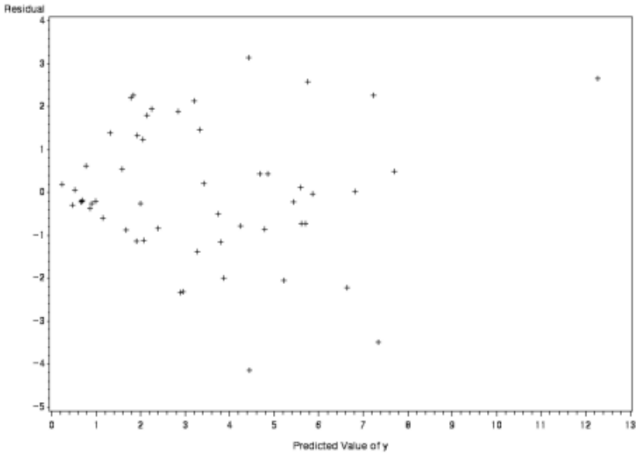
적률			
N	30	가중합	30
평균	0.00281541	관측값 합	0.08446243
표준 편차	1.03290819	분산	1.06689933
왜도	0.09192792	첨도	-0.7849332
제곱합	30.9403183	수정 제곱합	30.9400805
변동계수	36687.6086	평균의 표준 오차	0.18858237

정규성 검정				
검정	통계량		p 값	
Shapiro-Wilk	W	0.962364	Pr < W	0.3555
Kolmogorov-Smirnov	D	0.107803	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.067761	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.41039	Pr > A-Sq	>0.2500

w-통계량의 값은 각각 0.9605, 0.9624이고 이때 p value가 각각 0.3195, 0.3555 로 크게나와서 귀무가설을 기각하지 못하므로 정규분포를 따른다고 할 수 있다. 또한 5.2.2의 정규확률그림을 통해서도 정규분포를 따른다고도 할 수 있다.

5.5

5.5.1



The REG Procedure
Model: MODEL1
Dependent Variable: u

Number of Observations Read	53
Number of Observations Used	53

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	13.66027	13.66027	7.51	0.0084
Error	51	92.79105	1.81943		
Corrected Total	52	106.45132			

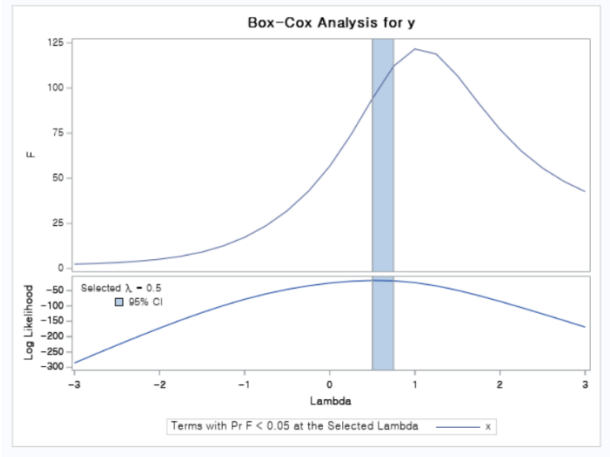
위의 Y와 x의 잔차산점도를 보면, 이는 적합값의 증가에 따라 잔차가 증가했다가 감소한다. 즉 이 잔차산점도를 통해 분산이 일정하지않다는 것을 알 수 있다. 소비자가 쓰는 전기수요를 계수(count)로 본다면 이는 포아송분포로 볼 수 있기 때문에 제곱근변환을 통해 분산을 안정화시킬 수 있다. 스코어 검정은 σ^2=SSE/표본의 수=126.86602/53=2.393698을 통해, ei^2/2.393698를 새로운 변수인 u로 두고 u를 설명변수 x로 회귀시켰다.

스코어 검정 통계량은 13.66027/2=6.83 이고 p-값은 0.0084로 0.05보다 작으므로 귀무가설을 기

각할 수 있다.

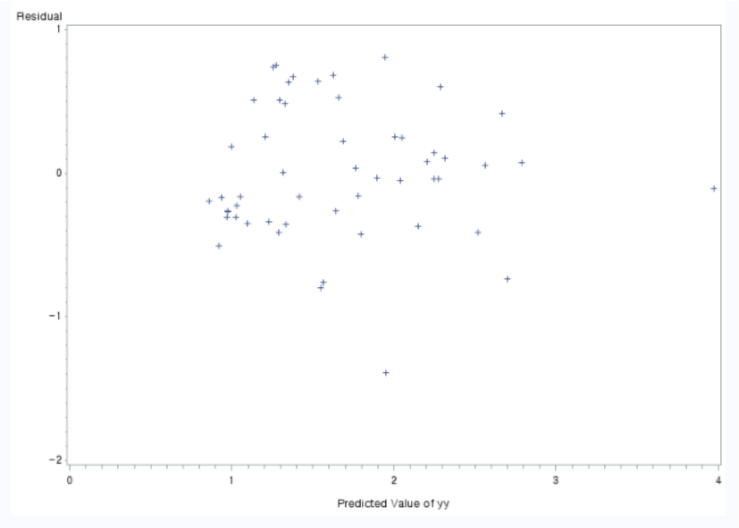
5.5.2

제곱근변환이 적절한 변환인지 파악하기 위해 Box-Cox 변환방법을 이용하여 알아볼 수있다. 표 5.15에 따르면 λ는 [-1,1]범위 내에서 약 0.6이므로 λhat을 0.5로 어림하여 제곱근변화에 해당하는 것을 알 수 있다. 아래그림에서 파란색 범위는 신뢰구간을 뜻한다. 이 신뢰구간을 1을 포함하지않고 있으므로 변환이 요구된다는 의미로 해석할 수 있다.



5.53

√(y)를 yy라는 새로운 설정변수을 생성하여 추가하여 √(y)로 제곱근 변환한 모형에 대해 잔차산점도는 다음과 같다.



이는 오른쪽에 동떨어진 한 개의 점을 제외하고는 회귀모형이 적절하게 적합되었을 때 보여지는 잔차산점도로, 골고루 잘 퍼져있는 것을 볼 수 있다. 또한 이 분포는 정규분포를 따른다고 할 수 있다. 따라서 제곱근변환과 같은 변수변환은 분산을 안정화한다.