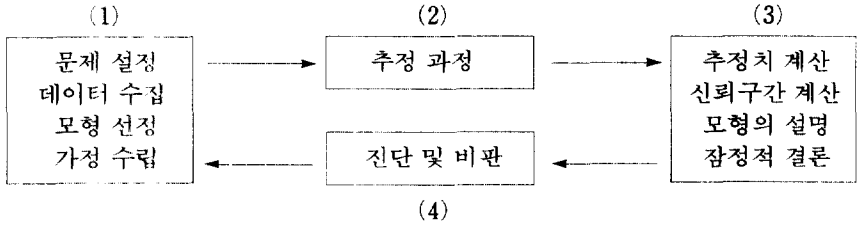
회귀 모형 진단

강창욱
한양대학교 산업공학과, 본지 편집위원

Regression Diagnostics

Chang Wook Kang
Dept. of Industrial Engineering, Hanyang University

통계 분석 기법 중에서 많이 활용되고 있는 회귀 분석(Regression Analysis)은 어떤 현상을 규명하기 위하여 종속 변수와 독립 변수들 간의 관계를 연구하는 것이다. Box (1980)는 통계적 모형 구축 과정을 다음과 같은 계통을 세워 설명하고 있다.



위에서 (1), (2), (3) 만의 분석 과정은 선정된 모형과 가정이 모두 맞다는 전제하에서 이루어진다. 그러나 실제 상황에서는 그러한 경우가 아주 드물기 때문에 (4) 과정을 거치지 않는 경우에는 모든 분석 결과에 대해서 신뢰할 수 없게 된다. 그러므로 회귀 진단을 실시할 때에는 첫째, 각각의 케이스가 분석 결과에 어느 정도의 영향을 미치는가, 둘째, 선정된 모형이 현재의 데이터를 잘 반영하고 있는가 하는 질문에 대한 해답을 구하려고 노력해야 한다.

1. 회귀 모형 진단(Regression Diagnostics)의 정의

회귀 모형 진단이란 주어진 데이터로 부터 추정된 회귀 모형과 그 데이터 간에 서로 일치되지 않는 사항들을 찾아내는 과정을 말한다. 이러한 과정은 크게 두 부분으로 나눌 수가 있는데 첫째는 데이터에 관한 진단 (data criticism) 이고 둘째는 모형에 관한 진단 (model criticism) 이다.

1.1 데이터에 관한 진단 (data criticism)

데이터 진단은 각각의 케이스(case)에 어떤 변형을 가했을 때 추정된 모형과 그로부터 계산된 추정치 및 다른 분석 결과들에 얼마만큼의 영향을 미치는 지를 알아보는 과정이다. 이때 우리는 추정된 모형이 맞다는 가정하에 데이터 진단을 실시한다. 데이터 진단은 다음과 같은 사항을 각각의 케이스에 대해 구하고 그 영향도를 판단한다.

(1) outlier
(2) leverage
(3) influence

이러한 데이터 진단 방법은 다음과 같은 원칙을 갖는 것이 좋다.

가. 케이스 변형 방법이 잘 정의되어야 한다.
나. 데이터 진단은 문제의 특정 부분에 관한 것이어야 한다.
다. 데이터 진단은 반드시 주어진 데이터만 가지고 계산되어야 한다.
라. 데이터 진단은 케이스간에 상호 비교될 수 있어야 한다.

1.2 모형에 관한 진단 (model criticism)

일반적으로 회귀 모형을 사용하여 종속 변수와 독립 변수의 함수관계를 구할 때의 모형은 $y=\beta_0+\beta_1 x_1+\cdots+\varepsilon,\ \varepsilon\sim iid\ N(0,\sigma^2)$ 이다. 즉 그 기본 가정들은 다음과 같다.

(1) 선형성(linearity)
(2) 등분산성(constant variance or homoscedasticity)
(3) 정규성(normality)
(4) 독립성(independence)

모형 진단은 추정된 모형이 이러한 가정들을 과연 잘 만족하고 있는지를 판단하기 위한 과정이다. 이러한 모형 진단 방법은 다음과 같은 원칙을 갖는 것이 좋다.

가. 진단방법은 가능한한 정확히 알려져야만 한다.
나. 진단방법은 모수적 방법으로 유도되도록 한다.
다. 진단방법은 계산이 간단해야 한다.
라. 진단방법은 그림형태나 그림으로 나타낼 수 있어야 한다.
마. 진단방법은 조치를 취할수 있는 방안을 제시해야 한다.

2. 데이터 진단

2.1 Outlier

어떤 데이터도 연구 대상이 되는 현상을 아주 명백하게 나타낸다는 보장을 할 수가 없다. 직관적으로 생각해 볼 때 관측치들에 대한 신뢰성은 유사한 조건하에서 얻은 다른 관측치들과의 관계에 의해 나타난다. 관측자가 생각하기에 대부분의 데이터들과 멀리 떨어져 있는 관측치들은 다음과 같은 용어들로 많은 문헌에서 사용되었다 : "outliers", "discordant observations", 또는 "contaminants". 이때 discordant observations란 관

측자가 전혀 예상하지 못한 경우의 관측값을 말하고, contaminants란 추정되는 분포에서는 실현 가능성이 없는 관측값을 말하고, outliers란 위의 두 경우를 총칭하여 말한다. 따라서 우리는 문제가 되는 관측값을 outlier 라고 한다.

데이터의 수가 적을 때는 그림을 이용하든지 아니면 직접 숫자들을 육안으로 파악하여 outlier 들을 검사할 수도 있겠으나 데이터의 수가 많다든지 또는 문제가 복잡한 경우(예를 들면, 중회귀, 다변량 샘플, 실험 계획법 등)에는 단순히 데이터를 육안으로만 검사하려고 한다면 이것은 거의 불가능할 것이며 혹 할 수 있다해도 이는 주관적 판단에 불과한 것이다. 그러므로 그와 같은 데이터를 다루는 객관적인 검사 방법을 모색하여야 한다.

이러한 outlier 들이 만약 데이터내에 존재한다면, 그 원인들을 살펴보면 크게 세가지로 나눌 수가 있다.

(가) global model weakness : 현재의 모형을 변형시키든지 아니면 완전히 새로운 모형을 설정해야할 만큼 치명적인 원인들을 의미한다. 예를 들면, 잘못된 척도로 종속변수를 측정한 경우 또는 알려진 상황하에서 outlier 들이 빈번하게 생기는 경우.

(나) local model weakness : 모형 전체에는 영향을 미치지 않으나 문제가 되는 관측치에만 해당하는 원인들을 의미한다. 예를 들면, 다른 관측치들과 차이가 많은 경우, 기록상의 오류를 범한 경우, 설명 변수의 공간에서 멀리 떨어져 큰 영향을 미치는 관측치가 있는 경우 등이다.

(다) Natural Variability : 모형이 잘못 선정되었다든지 또는 다른 요인에 의해서가 아니라 현상 자체가 갖고있는 변동성에 해당하는 원인들을 의미한다. 예를 들면, 포유류의 몸무게 중 코끼리의 경우.

위에서 설명한 원인들에 의해서 생겨난 outlier 들을 찾아내는 객관적인 방법들 중에서 가장 많이 사용되는 기법 두가지를 설명하고자 한다.

(가) 표준화 잔차(studentized residuals)를 이용하는 방법

만약 i번째 케이스가 outlier 이다라고 의심이 간다면 다음과 같은 순서로 확인을 한다.

(1) 원 데이터에서 i번째 케이스를 제외시킨다. 그러면 이제 $n-1$개의 케이스만 있다.

(2) $n-1$ 개의 데이터만을 가지고 회귀 분석을 실시한다. 이 경우에 추정된 회귀 계수와 잔차의 분산을 각기 $\beta_{(i)}$와 $\hat{\sigma}^2_{(i)}$로 표기한다. 첨자 (i)는 i번째 케이스는 분석에서 제외되었음을 알려주기 위한 것이다. 따라서 $\hat{\sigma}^2_{(i)}$는 $n-p'-1$의 자유도를 가진다. p'은 회귀계수 모수의 수이다.

(3) 제외된 i번째 케이스에 대해 회귀값을 추정한다. $\tilde{y}_i = x_i^T \beta_{(i)}$. 그런데 i번째 케이스는 추정 과정에서 사용되지 않았기 때문에 y_i와 $\tilde{y}_i$는 독립적이다. 따라서 $y_i - \tilde{y}_i$의 분산을 구하면

$$Var(y_i - \tilde{y}_i) = \sigma^2 + \sigma^2 x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i$$

여기서 $X_{(i)}$는 i번째 행이 빠진 X행렬을 의미한다. 그리고 이 분산을 추정할 때는 σ^2 대신에 $\hat{\sigma}^2_{(i)}$을 대입하면 된다.

(4) 만약에 y_i가 outlier가 아니라면 $E(y_i-\tilde{y}_i)=0$ 일 것이다. 오차항의 정규성 가정하에 $E(y_i-\tilde{y}_i)=0$ 이라는 가설을 검정하기 위한 t검정 통계량은

$$t_i = \frac{y_i - y_i}{\hat{\sigma}_{(i)}\sqrt{1+x_i^T(X_{(i)}^TX_{(i)})^{-1}x_i}}$$

이 검정 통계량은 자유도가 $n-p'-1$인 t분포를 따른다. 이때의 t_i를 우리는 externally studentized residuals이라 하고 이는 internally studentized residual로 우리는 r_i와 다음과 같은 관계에 있다.

$$t_i = r_i\left(\frac{n-p'-1}{n-p'-r_i^2}\right)^{\frac{1}{2}} = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

(나) Mean-shift 모형을 이용하는 방법

Mean-shift 모형은 주로 이상이 있는 속 변수의 값을 찾는데 사용되지만 이는 X의 i번째 행의 이상 유무를 확인하는 것과 같다. 만약 i번째 케이스가 의문시 된다면 Mean-shift 모형을 이용하는 방법은 다음과 같이 설명할 수가 있다. Mean-shift 모형은

$$Y = X\beta + d_i\phi + \varepsilon \tag{2.1}$$

여기서 d_i는 i번째 원소는 1 이고 나머지 원소는 0인 n-벡터 이다. ϕ 값이 0 이 아니면 i번째 케이스는 outlier 이다. 다음에는 X행렬의 i번째 행이 모르는 값 δ_i만큼 오차가 있다고 하자. 그러면

$$Y = [X + \begin{matrix}0\\ \vdots\\ \delta_i^T\\ \vdots\\ 0\end{matrix}]\beta + \varepsilon$$

이고 $\phi_i = \delta_i^T\beta$ 이다. 만일 (2.1) 식을 X의 열들에 직교하는 변수를 첨가하여 다시 쓰면

$$Y = X\beta^* + (I-H)d_i\phi + \varepsilon$$

여기서 β^*는 (2.1)식의 β와 같지 않으나 ϕ 는 같다. 변수들의 직교성 때문에 다음과 같은 분석을 할 수가 있다.

(1) 종속 변수는 Y, 독립 변수는 X인 회귀 분석을 한다. 이 경우의 잔차 $e=(I-H)Y$를 계산하여 다른 변수로 저장해 둔다.

(2) 또다른 회귀 분석을 하는데 이번에는 종속 변수는 (1)의 잔차가 되고 독립 변수는 첨가된 변수 $(I-H)d_i$ 이다. 이 경우의 회귀 계수 추정치는 $\phi = e_i/(1-h_{ii})$ 이다.

(3) $\phi=0$ 이라는 가설에 대한 검정을 실시하기 전에 전체 제곱합을 요소별로 나누어 본다.

전체 제곱합 $= Y^{T}Y$

X에 의한 회귀 제곱합 $= Y^{T}HY$

첨가 변수에 의한 회귀 제곱합 $= \hat{\phi}^{2}(d_i^{T}(I-H)^{2}d_i)$

$= e_i^{2}/(1-h_{ii})$

잔차 제곱합 $= Y^{T}(I-H)Y - e_i^{2}/(1-h_{ii})$

따라서 정규성의 가정하에 $\phi = 0$ 가설에 대한 t검정 통계량은

$$t_i = [e_i^{2}/(1-h_{ii})]^{0.5}/[RSS/(n-p'-1)]^{0.5} \sim t_{(n-p'-1)}$$

t는 귀무 가설 하에서 자유도가 $n-p'-1$인 t분포를 따른다. 이 t_i는 앞에서 정의된 externally studentized residuals t_i와 동일하다.

결국 표준화 잔차를 이용하는 방법이나 Mean-shift 모형을 이용하는 방법은 같은 결론에 도달하게 된다. 실제 우리는 어떤 특정한 케이스에 대해서만 outlier 검정을 실시하지 않고 모든 케이스에 대해서 실시하기 때문에 판정을 내릴 때는 Bonferroni's critical values를 이용한다.

2.2 Leverage

Hoaglin과 Welsch(1978)는 Hat matrix를 이용하여 leverage에 관하여 설명하고 있다. 최소자승법에 의한 회귀 분석에서 각각의 데이터 y_j가 fitted values $\hat{y}_i$에 얼마 만큼의 영향력을 행사하는 지를 알아보는 것도 매우 중요하다. 왜냐하면 y_i와 $\hat{y}_i$간의 관계는 아주 쉽게 알 수 있으므로 독립 변수의 값 중에서 어려운 문제점을 파악할 수도 있기 때문이다. 관측값 y_j의 선형 결합으로 $\hat{y}_i$을 설명할 수 있는데, $\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_i$, 그 선형 결합 계수가 바로 hat matrix의 i번째 행의 원소들이다. 이러한 hat matrix의 성질에 대해서 살펴보면 다음과 같다.

hat matrix는 $H = X(X^{T}X)^{-1}X^{T}$로 정의된다.

(1) H는 symmetric 하다 ($H = H^{T}$).

(2) H는 idempotent 하다 ($H = H^{2}$).

(3) (1)과 (2) 때문에 H는 projection matrix 이다.

(4) $0 \leq h_{ii} \leq 1$

(5) $rank(H) = p$

(6) $tr(H) = p$

(7) $\sum_j h_{ij} = \sum_i h_{ij} = 1$

i번째 잔차의 분산은 $\hat{\sigma}^{2}(1-h_{ii})$ 이다. 여기서 우리는 h_{ii}가 큰 케이스 일수록 그 케이스의 잔차의 분산은 작아진다. 즉 h_{ii}가 1 에 접근하면 i번째 잔차의 분산은 0 으로 접근한다. 다음의 관계식을 살펴보자.

$$\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j$$

즉 h_{ii}가 1 에 접근하면 분산의 값은 0 에 접근한다. 그 말은 h_{ii}가 1 에 접근하면 $\hat{y}_i$은

y_i가 되므로 이 케이스는 구하고자 하는 회귀 직선이 자기를 반드시 거쳐가게 영향력을 행사하는 것이 된다.

(6) 에 의해 h_{ii}의 합은 p이며 h_{ii}의 평균 값은 p/n 이다. Hoaglin과 Welsch는 경험상 평균치보다 두 배나 큰 h_{ii}값은 흔하지 않으므로 이러한 크기의 h_{ii}값을 갖는 케이스를 영향력이 큰 케이스(high-leverage points)라고 했다. 따라서 우리는 h_{ii}의 값을 보아 $2p/n$보다 크면 그 케이스에 대한 면밀한 조사를 해봐야 한다.

2.3 Influence of cases

데이터 진단에 있어서의 또 다른 영역은 모형을 주어진 데이터에 대해 적합을 시킬 때, 각 케이스들이 분석 결과에 어떤 영향을 미치는 지에 대해서 알아 보는 것이다. 이러한 연구를 각 케이스의 influence 분석이라고 한다. 즉 그 기본 개념은 데이터에 어떠한 형태로든 약간의 변화를 주었을 때 분석의 특정 부분이 어떻게 변화하는가를 알아 보는 것이다. 이러한 influence 분석은 모형이 맞다고 가정을 하고 실시한다. 가장 많이 사용되는 데이터 변형 방법은 한번에 케이스를 하나씩 제외시키는 방법이다. 그런 다음에 모든 데이터에 의한 분석 결과와 한개의 케이스가 제외된 경우의 분석 결과를 비교, 검토하여 차이가 크게 나는 지를 알아 본다. 이때 차이를 크게 하는 케이스를 우리는 influential case라고 한다.

우리가 데이터를 분석할 때에 influential case를 찾아 낼 수 있다면 대략 두가지 측면에서 유리할 것같다. 첫째, 분석 결론에 대한 신뢰성과 가정된 모형과의 연관성에 대한 정보를 얻을 수가 있고, 둘째, 멀리 떨어져 있는 케이스들이 비교적 분석 결론에 많은 영향을 끼치는 경향이 있음을 알게 된다. 따라서 influential case들이 대부분의 케이스들 보다 더 중요한 정보를 제공할 때도 종종 있다. 대부분의 경우 influence 분석은 influential case들을 찾아내는 데 중점을 둔다. 왜냐하면 처리 방법은 문제의 성격에 따라 달라질 수도 있기 때문에 특별한 방안을 제시한다는 것 자체가 불가능할 지도 모른다. 그러나 일반적으로 influential case 를 일단 발견했다면 어떠한 가능한 조치를 하는 것이 바람직하다. 다음과 같은 조치를 취하는 것도 괜찮을 것이다.

(1) 측정이나 입력을 잘못한 경우 – 데이터 군에서 제외시킨다.

(2) 실험 조건이 부적절한 경우 – 실험 조건을 수정한다.

(3) 원인을 알수 없거나 모형이 맞다는 확신이 있는 경우 – outlier test 를 실시한다.

(4) 데이터의 수가 적어서 판단이 곤란한 경우 – 새로운 데이터를 추가한다.

(5) 위의 경우에 해당하지 않을 때는 두 종류의 분석 보고서를 작성하여 제출한다. 하나는 문제가 있는 케이스들을 포함한 것이고, 다른 하나는 그 케이스들을 제외시킨 것이다.

Influence 분석에서 가장 많이 사용되는 진단 통계량은 Cook's Distance D_i 통계량이다. Cook(1977) 은 D_i를 다음과 같이 정의했다.

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T (X^T X)(\hat{\beta}_{(i)} - \hat{\beta})}{p \hat{\sigma}^2}$$

D_i 통계량의 특징을 살펴보면

(1) 일정한 D_i 값의 등고선은 타원체이다(비교가능).

(2) 등고선의 간격은 모수 공간에서 $\hat{\beta}_{(i)}$과 $\hat{\beta}$사이의 거리로 보아도 된다.

(3) D_i는 변수 변환에 영향을 받지 않는다.

D_i 통계량은 다른 회귀 진단 통계량과도 밀접한 관계를 갖는다. 즉,

$$D_i = \frac{1}{p} r_i^2 \left(\frac{h_{ii}}{1-h_{ii}} \right)$$

D_i 통계량의 값이 크면 우리는 i번째 케이스가 influential 하다고 한다. 어느 정도 큰 값을 크다고 하겠는가는 문제마다 달라질 수도 있으나, 대략 D_i값이 1 보다 크면 두 분석 사이에 모수의 추정치가 상당히 차이가 있다고 보아도 좋을 것이다. 그러나 D_i값이 1보다는 작지만 다른 D_i값들에 비해 특히 큰 경우에는 면밀히 검토해 볼 필요가 있다.

3. 모형 진단

3.1 산점도(Scatter plot)

변수가 X, Y 두개인 경우에 X와 Y의 관계를 단순히 그래프에 나타낸 것을 산점도 라고 한다. 회귀 분석에서도 종속 변수와 독립 변수의 산점도는 개략적이지만 어느정도의 관계는 보여 주기 때문에 분석을 시작할 때 항상 이용된다. 특히 종속 변수와 독립 변수가 직선 관계에 있지 않을 때는 산점도를 통해 쉽게 확인할 수가 있다. 또한 산점도를 통해 outlier일 가능성이 큰 케이스를 미리 엿볼수도 있고 다른 영향력을 갖고 있을지도 모르는 케이스도 예감할 수도 있다. 그러나 요즈음은 3 차원 그래프가 이용되고 있지만 중회귀의 경우 변수의 전체 수가 4 이상인 경우에는 하나의 산점도를 통해 각 변수들 사이의 관계를 알아 볼 수는 없다. 따라서 matrix plot을 사용하기도 하는데 마치 상관 행렬이 모든 변수들의 상관 관계 계수를 하나의 행렬에 표기한 것처럼 이는 모든 두 변수 사이의 산점도를 한 면에 나타낸 것을 말한다.

그리고 index plot이 있는데 이것은 어떤 통계량의 값을 각 케이스 일련 번호에 맞춰 나타낸 그림이다. 예를 들면 D_i 통계량 값을 y축에 그리고 케이스 번호 i를 x축에 나타내면 이것이 바로 index plot이다. 이러한 index plot은 큰 값을 찾는 데 아주 수월해진다. 왜냐하면 숫자들의 군에서 큰 숫자를 찾아내는 것 보다는 그림을 통해 찾는 것이 쉽고 틀릴 염려가 없다.

3.2 Residual plot

이러한 산점도 중에서 특히 잔차와 fitted values사이의 산점도를 우리는 residual plot이라고 한다. 이 residual plot은 회귀 분석에서 아주 중요한 위치를 점하고 있다. 위에서 살펴본 바와 같이 잔차는 관측할 수 없는 오차항의 대용으로 많이 사용되고 있으며 또한 오차항에 대한 정보를 많이 갖고 있다. 따라서 잔차들이 랜덤하게 흩어져 있는가, 잔차들의 흩어진 폭이 fitted values의 증감에 따라 변화하고 있는가, 또는 잔차들의 흩어짐

이 곡선형의 경향은 없는가를 조사하는 것은 아주 중요하다. 아래에 구체적으로 설명할 모형 진단 방법들은 바로 이 residual plot으로부터 시작한다. 또한 위에서 설명한 outlier 검사도 모두 잔차를 이용하고 있기 때문에 residual plot으로부터 아주 중요한 정보를 입수한다.

그런데 residual plot은 주로 ordinary residuals vs fitted values, ordinary residuals vs predictors, studentized residuals vs fitted values, studentized residuals vs predictors이다. 독립 변수와의 residual plot은 모형이 각각의 독립 변수들과 잘 부합하고 있는 지를 알려주고 또한 polynomial 회귀 분석의 필요성 여부 및 만약 있다면 어떠한 조치가 필요한지에 대한 사전 정보를 줄 수도 있기 때문에 중요하다. residuals vs y values 그림은 그리지 않는데 그 이유는 residuals 과 y값의 공분산이 0 이 아니기 때문이다. 즉, 그림상에 어떠한 패턴이 발견되었을 경우 이것이 과연 가정상의 문제점인지 아니면 두 변량의 상관 관계 때문인지 구별하기가 어렵다. ordinary residuals와 fitted values의 공분산은 0 이고 studentized residuals와 fitted values의 공분산은 거의 0 이며 residuals와 독립 변수들의 공분산은 0 이다.

3.3 Nonlinearity 문제

회귀 분석에서는 일반적으로 선형 모형이 많이 사용되고 있다. 여기서 말하는 선형 모형이란 모형에 사용되는 모수들 즉, β들이 선형 결합을 하고 있다는 것을 의미한다. 선형 모형인 경우, 분석이 용이하며 모형의 해석도 비선형의 경우 보다 수월하다. 그래서 때로는 비선형 모형도 적절한 변수 변환을 통해 선형 모형화 해서 분석을 하기도 한다. 이러한 모형의 선형성에 대한 가정은 그림을 통해 체크할 수가 있다. 독립 변수가 하나인 경우의 단순 선형 회귀 분석에서는 residual plot이 아주 유용한 자료이며 독립 변수가 두개 이상인 중회귀 분석에서는 residual plot과 added variable plot이 모형의 선형성을 체크하는 데 중요한 역할을 한다. 따라서, 그림에 점들이 곡선 형태로 나타나면 변수 변환의 필요성을 나타낸다. 이러한 비선형의 경우 종속 변수의 변환, 독립 변수의 고차항 추가, 또는 독립 변수들의 cross product 항을 추가함으로서 문제를 해결해 나갈 수가 있다. 그런데 종속 변수를 변환시켜야 할 경우와 독립 변수의 고차항을 추가시켜야 할 경우가 비슷하기 때문에 어떤 조치를 취할지는 경험에 의한 판단에 의존할 수 밖에 없다.

변수 변환을 통하여 다음과 같은 효과를 얻을 수 있다.

(1) 비선형성 또는 비등분산성을 해결한다.

(2) 변환된 모형에서 오차항의 정규성이 더욱 만족된다.

(3) 데이터를 설명하기에 더욱 간단한 모형을 얻는다.

3.3.1 종속 변수의 변환 방법

Box 와 Cox(1964) 는 모형을 확장하는 기법을 이용하여 종속 변수의 변환을 새로운 모수의 추정 문제로 유도하는 객관적인 방법을 제시하였다. 모형 확장 기법은 다음과 같다.

(1) 모든 y값들이 양수임을 가정한다.

(2) λ를 변환 차수로 하고, 만약 $\lambda=0$이면 $y_i^{\lambda}=\ln(y_i)$로 변환하고, $\lambda=1$이면 종속 변수를 변환시키지 않는다.

(3) 그외에는 새로운 모형

$$Y^{\lambda} = X\beta + \varepsilon, \quad Var(\varepsilon) = \sigma^2 I \tag{3.1}$$

을 검토한다. 즉, 새로운 모수 λ와 (β, σ^2)을 동시에 추정하는 문제이다. 그런데 (3.1) 모형은 선형 모형이 아니기 때문에 최소 자승법이 아닌 다른 추정 방법이 필요하다.

그러므로 λ를 추정하는 방법이 곧 종속 변수의 변환 방법이 되는데 특히 많이 사용되는 두가지 기법을 소개하기로 한다.

(가) Box-Cox 의 방법

만약에 우리가 변환 차수 λ의 값을 안다면, β의 추정이나 잔차 제곱합(RSS)의 계산은 다음과 같이 간단하다.

$$\hat{\beta} = (X^{\top}X)^{-1}X^{\top}Y^{\lambda}$$
$$RSS(\lambda) = (Y^{\lambda})^{\top}(I-H)Y^{\lambda} \tag{3.2}$$

즉, 종속 변수를 Y^{λ}로 하고 독립 변수들에 대해 회귀 분석을 실시하면 된다. 그러나, λ의 값을 모르기 때문에 각각의 λ값에 대해서 분석을 해보아야 한다. λ값의 적절한 범위는 -2에서 $+2$ 사이이며 이 범위 밖의 λ값에 대해서는 Box-Cox방법의 효력이 좋지 않다. 그리고 다양한 λ값에 대해서 $RSS(\lambda)$를 직접 비교할 수는 없다. 왜냐하면 각각의 λ값에 따라 $RSS(\lambda)$의 단위가 서로 다르기 때문이다. 그래서 $RSS(\lambda)$ 대신에 log-likelihood function을 구하여 서로 비교한다.

확장 모형에서 λ가 0 이 아닌 경우의 log-likelihood function은

$$L(\lambda) = n\ln(|\lambda|) - \frac{n}{2}\ln(RSS(\lambda)) + n(\lambda-1)\ln(GM(y)) \tag{3.3}$$

여기서 $GM(y)$는 y의 기하 평균을 의미한다. 만약 λ가 0 이면

$$L(0) = -\frac{n}{2}\ln(RSS(0)) - n\ln(GM(y)) \tag{3.4}$$

이때 $L(\lambda)$를 최대화 시키는 λ값이 바로 λ의 추정치가 되는데, $L(\lambda)$와 λ의 그래프를 그려서 최고점 주위에 있는 일반적인 값(즉 0, -1, 0.5 등)을 택하여 추정치로 정하면 된다.

위의 방법과 동일하지만 다른 확장 모형을 사용하는 방법은 다음과 같다. 즉,

$$z_i^{\lambda} = \frac{y_i^{\lambda}-1}{\lambda[GM(y)]^{\lambda-1}}, \quad \lambda \neq 0$$
$$= GM(y)\ln(y_i), \quad \lambda = 0$$

위의 y_i^{λ}는 $\lambda=0$에서 연속이 아니지만 z_i^{λ}는 $\lambda=0$에서도 연속이다. 새 모형은

$$Z^{\lambda} = X\beta + \varepsilon \tag{3.5}$$

이다. 이번에는 종속 변수를 z_i^{λ}로 하고 모든 독립 변수에 대해 회귀 분석을 한다. 각

각의 λ값에 대해서 $RSS_z(\lambda)$는 같은 단위이기때문에 상호 비교가 가능하다. $RSS_z(\lambda)$가 최소로 되는 λ를 추정치로 한다.

(나) Atkinson 의 방법

Box-Cox방법은 계산량이 많고 또 log-likelihood function을 유도해야 하는 불편함이 뒤따른다. 그래서 Atkinson(1973)은 $\lambda=1$의 가설에 대한 간단한 검정 방법을 제시했고 이어서 Atkinson(1981)은 간단하게 λ값을 추정할 수 있는 방법을 제시했는데, 이 방법을 설명하기로 한다. 이 방법은 $\lambda = 1$에서 Z^{λ}를 Taylor expansion한 뒤 처음 두 항으로 근사시킨 것을 이용한 것이다. 즉,

$$Z^{\lambda} \cong Z^{1} + (\lambda-1)\frac{\partial Z^{\lambda}}{\partial \lambda}\Big|_{\lambda=1}$$
$$= Z^{1} + (\lambda-1)G \qquad (3.6)$$

여기서 $g_i = y_i\{\ln[y_i/GM(y)]-1\} + \{\ln[GM(y)]+1\}$

식(3.6)을 식(3.5)에 대입하면

$$Z^{1} \cong X\beta + (1-\lambda)G + \varepsilon$$

그런데 $Z^{1}=Y-1$이기 때문에 $\phi = 1-\lambda$로 놓고 Z^{1}을 Y로 대체하면

$$Y = X\beta + \phi G + \varepsilon \qquad (3.7)$$

이 된다. 그러므로 $\phi = 0$에 대한 검정이나 $\lambda=1$에 대한 검정이 같다고 할 수 있다. $\phi = 0$에 대한 검정 통계량은 Y의 X와 G에 관한 회귀 분석에서 G에 대한 t검정 통계량이다. Atkinson은 이 검정 통계량의 p-value의 근사값을 구하기 위해서는 표준 정규 분포표의 사용을 권하고 있다. 모형 (3.7)에서의 ϕ 의 추정치를 $\hat{\phi}$ 라고 한다면 우리가 구하고자 하는 λ의 추정치는 $\hat{\lambda}=1-\hat{\phi}$ 가 된다. 새로운 변수 G에 대한 added variable plot도 좋은 진단 방법이 된다.

3.3.2 독립 변수의 변환 방법

종속 변수의 변환으로도 비선형성이 해결되지 않는다면 다음으로 독립 변수의 변환을 생각해 보아야 할 것이다. 독립 변수의 변환 방법은 첫째 Polynomial regression으로의 전환, 둘째는 특정 독립 변수의 변환 차수를 추정하는 것이다. 여기서는 Box와 Tidwell (1962)이 제안한 변환 차수를 추정하는 방법을 간단히 소개하기로 한다.

이 방법도 모델 확장 개념을 이용한다. 독립 변수의 수가 p개인 중회귀에서 X_1을 먼저 변환시켜야 하는 경우를 가정한다. 선형 모형

$$Y = \beta_0 + \sum_{j=1}^{p}\beta_j X_j + \varepsilon \qquad (3.8)$$

을 다음과 같이 확장한다.

$$Y = \beta_0 + \beta_1 X_1^{\alpha_1} + \sum_{j=2}^{p}\beta_j X_j + \varepsilon \qquad (3.9)$$

전과 같이 $\alpha_1=0$이면 $X_1^{\alpha_1}=\ln(X_1)$으로 한다. 식 (3. 9)를 $\alpha_1=1$에서 $X_1^{\alpha_1}$을 Taylor expansion하여 다시 쓰면

$$X_1^{\alpha_1} \cong X_1 + (\alpha_1 - 1)\frac{\partial}{\partial \alpha_1} X_1^{\alpha_1}\Big|_{\alpha_1=1}$$
$$\cong X_1 + (\alpha_1 - 1) X_1 \ln(X_1) \qquad (3.10)$$

식 (3.10)을 식 (3.9)에 대입하면

$$T \cong \beta_0 + \beta_1[X_1 + (\alpha_1 - 1)X_1\ln(X_1)] + \sum_{j=1}^{p}\beta_j X_j + \varepsilon$$
$$= \beta_0 + \sum_{j=1}^{p}\beta_j X_j + \beta_1(\alpha_1 - 1)X_1\ln(X_1) + \varepsilon$$
$$= \beta_0 + \sum_{j=1}^{p}\beta_j X_j + \eta X_1\ln(X_1) + \varepsilon \qquad (3.11)$$

여기서 $\eta=\beta_1(\alpha_1-1)$이다. $\eta=0$는 $\beta_1=0$이거나 $\alpha_1=1$과 같으므로 $\eta=0$에 대한 검정이 결국 변환의 필요성에 대한 검정이다. η는 역시 독립 변수들과 추가되는 변수 $X_1\ln(X_1)$에 대한 회귀 분석에서 추정할 수가 있다. 그리고 t검정 통계량은 자유도가 $n-p'-1$인 t분포를 따른다. $X_1\ln(X_1)$에 대한 added variable plot을 진단 방법으로 사용할 수도 있다. 그러므로 우리가 구하고자 하는 α_1의 추정치는

$$\hat{\alpha}_1 = \frac{\hat{\eta}}{\hat{\beta}_1} + 1 \qquad (3.12)$$

인데 $\hat{\eta}$은 (3. 11)로 부터 추정하고 $\hat{\beta}_1$은 (3.8)로 부터 추정한 것이다.

그러나 우리는 어떤 독립 변수 부터 변환의 필요성에 대한 검정과 변환 차수의 추정을 해야 할지 알 수가 없다. 그렇다고 모든 독립 변수에 대해서 실시한다면 계산량이 무척 많아질 수도 있다. 다음과 같은 방안이 제시되었다 (Weisberg,1985).

(1) 각 독립 변수에 대해 최대값과 최소값의 비율이 10 보다 큰 경우, 그 독립 변수는 로그 변환을 시킨다.
(2) 그리고 Box-Cox방법이나 Atkinson방법에의해 종속변수의 변환을 실시한다.
(3) 새로운 회귀 분석을 통해 t값이 큰 독립 변수에 대해서 Box-Tidwell방법을 사용한다.

3.2 Nonconstant Variance (또는 heteroscedasticity) 문제

회귀 분석에서의 또 하나의 가정은 모든 데이터에 대해서 오차항의 등분산성이다. 실제 문제에 있어서는 많은 경우에 이 등분산성의 가정이 만족되지 않는다. 오차항의 분산이 종속 변수나 독립 변수의 값에 영향을 받기도 하고 때로는 시간이나 인위적인 순서에 의해 영향을 받기도 한다. 만약 이분산성이 진단되었으나 그 분산의 값을 알 수 없을 때는 두 가지의 조치를 취할 수가 있을 것이다.

(1) 경험에 의한 가중치를 주어 Weighted least squares regression을 사용한다.

(2) 종속 변수에 변환을 가하여 등분산이 되도록 유도한다. 주로 사용되는 변환은 log Y, $\sqrt{Y}$, y^{-1}등이며, 이러한 변환을 variance stablizing transformation이라고 한다.

Cook 과 Weisberg(1983) 은 등분산성 가정을 체크할 수 있는 진단 방법을 제시하였다 그 기본 개념은 하나의 새로운 모수에 대한 검정을 하는 것이 곧 등분산성의 가정에 대한 검정이 되도록 한 것이다. 지금 부터 이 진단 방법을 설명하기로 한다. 우선 $Var(\varepsilon_i)$이 모르는 모수 벡터 λ와 알 수 있는 벡터 z_i에 의존한다고 가정을 한다. 이때의 z_i는 $\hat{y}_i$이나 X_i 또는 다른 알려져 있는 벡터를 의미한다. z_i가 주어졌을 때 다음을 가정한다.

$$Var(\varepsilon_i) = \sigma^2 [\exp(\lambda^T z_i)] \tag{3.13}$$

(1) 모든 z_i에 대해서, $Var(\varepsilon_i) > 0$.

(2) 분산이 z_i와 λ에 의존하지만 반드시 $\lambda^T z_i$ 관계로서 의존한다.

(3) $Var(\varepsilon_i)$은 z_i에 비례한다.

(4) 만약 $\lambda=0$이면 $Var(\varepsilon_i)=\sigma^2$ 이다.

ε_i가 독립적이고 정규 분포를 따른다면 $\lambda=0$에 관한 검정은 간단히 할 수 있는데 그 과정은 다음과 같다.

(1) 원래의 종속 변수와 모든 독립 변수에 대한 회귀 분석으로부터 잔차를 계산하여 저장한다.

(2) scaled squared residuals, $u_i = \hat{e}_i^2/\tilde{\sigma}^2$를 계산한다. $\tilde{\sigma}^2 = \sum \hat{e}_i^2/n$이다.

(3) 종속 변수를 u_i, 독립 변수를 z_i로 한 회귀 분석에서 regression 제곱합을 계산한다.

(4) 검정 통계량은 $s = SSreg/2$이며 p-value를 구하기 위해서는 귀무 가설 하에서 s의 점근 분포인 $\chi^2(q)$을 이용한다. 이때 q는 z_i에 포함된 변수의 갯수이다.

λ이 0이 아닌 경우에는 s의 값이 아주 크기 때문에 s의 값이 큰 경우에 우리는 이분산성이 존재한다고 보는 것이다.

위의 검정 방법과 동일하지만 그림을 이용하는 진단 방법은 다음과 같다.

(1) $q=1$인 경우, 위의 (1)에서 internally studentized residuals r_i의 제곱을 계산한다.

(2) r_i^2을 y축에, $(1-h_{ii})z_i$를 x축에 놓고 그린 그림에서 쐐기(wedge) 형태가 되면 이것은 이분산성임을 나타낸다. 단 q가 2이상일 때에는 x축에 $(1-h_{ii})\hat{\lambda}^T z_i$을 놓는데 $\hat{\lambda}^T z_i$은 위의 (3)에서의 fitted values 이다.

가장 좋은 방법은 위에서 설명한 두 방법을 모두 동원하는 것이다. 왜냐하면 그림에서 x축의 값들이 고르게 흩어져 있지 않으면 과연 쐐기 모양인지 판단하기가 애매모호할 경우가 있을 수도 있기 때문이다. 한편 그림에 의한 방법은 통계량에 의한 결론을 뒷받침해 준다.

3.3 Nonnormality 문제

지금까지 설명한 회귀 분석에서의 추정이나 검정등 모든 결과들은 오차항이 정규 분포를 따른다는 가정 하에서 얻어진 것이다. 따라서 이 가정이 만족되지 않는다면 분석 결론을 크게 신뢰하지 못할 것이다. 데이터의 수가 적은 경우에는 오차항의 비정규성을 잔차 분석 만으로는 진단하기가 곤란하다. 선형 모형

$$Y = X\beta + \varepsilon, \qquad Var(\varepsilon) = \sigma^2 I$$

에서 비정규성을 체크한다고 하자. 그러나 오차항은 관측될 수 없는 변량이므로 그 대신에 잔차에 의거한 검정 방법을 모색해야 한다.

$$\begin{aligned} e &= (I-H)Y \\ &= (I-H)X\beta + (I-H)\varepsilon \\ &= (I-H)\varepsilon \end{aligned} \tag{3.14}$$

$$e_i = \varepsilon_i - \left(\sum_{j=1}^{n} h_{ij}\varepsilon_j \right) \tag{3.15}$$

그러므로 e_i는 ε_i에서 나머지 잔차들의 가중치 합을 뺀 것이다. 만약에 데이터의 수가 적고 어떤 h_{ii}의 값이 크다면, e_i의 분포를 파악하는 데 있어 ε_i보다는 $(\sum_{j=1}^{n} h_{ij}\varepsilon_i)$이 더 중요한 역할을 할 수도 있다. 그리고 오차항이 정규 분포를 따르지 않더라도 $(\sum_{j=1}^{n} h_{ij}\varepsilon_i)$은 중심 극한 정리에 의해 정규 분포를 따른다. 그러므로 데이터의 수가 적을 때는 잔차에 의한 오차항의 정규성 검정이 어렵다는 것을 알 수가 있다. 반면에 데이터의 수를 증가시키면 h_{ij}는 0에 가까와질 것이고 따라서 (3.15)에서 $(\sum_{j=1}^{n} h_{ij}\varepsilon_i)$은 중요하지 않게되므로 ε에 적용하는 방법이나 e에 적용하는 방법이나 얻는 정보의 양은 거의 비슷하게 될 것이다.

Wilk와 Gnanadesikan(1968)은 오차항의 정규성을 진단하기위한 Normal Probability Plot을 제시하였다. 이것을 간단히 설명하면 다음과 같다.

우선 n개의 데이터 $z_1, z_2, \cdots, z_n$이 있는데 이 데이터가 과연 평균이 μ이고 분산이 σ^2인 정규 분포를 따르고 있는 지를 확인하고자 한다. 그렇게 하기 위해서는

(1) 데이터 z_i를 크기 순으로 z의 순위 통계량 $z_{(1)}, z_{(2)}, \cdots, z_{(n)}$을 구한다.

(2) n개의 정규 순위 값, $u_{(1)}, u_{(2)}, \cdots, u_{(n)}$을 구한다. 이때의 정규 순위 값이란 우리가 표준 정규 분포로 부터 임의로 n개의 값을 취했을 때 기대되는 값을 말한다. 즉 우리가 표준 정규 분포로 부터 n개의 값을 취한 뒤 이들의 순위 통계량을 구한다. 계속 반복하여 과정을 되풀이 했을 때 얻어진 각 순위 통계량의 평균을 취하면 이것이 바로 정규 순위 값이다.

(3) 만약에 z가 평균이 μ이고 분산이 σ^2인 정규 분포를 따른다면

$$E(z_{(i)}) = \mu + \sigma u_{(i)}$$

의 관계가 성립될 것이다. 따라서 $z_{(i)}$와 $u_{(i)}$를 산점도로 나타내면 점들은 직선 상에 놓일 것이다. 그러나 만약에 데이터가 정규 분포를 따르지 않는다면 당연히 산점도에서 점들이 직선 상에 놓이지 않을 것이다. 흔히 볼 수 있는 Normal Probability Plot은 네가지 형태가 있다.

(i) 데이터들이 평균 주위에만 있는 경우 :
(ii) 데이터들이 평균 주위 보다는 주로 멀리 떨어져 있는 경우 :
(iii) 데이터들이 평균의 왼쪽으로 치우친 경우 :
(iv) 데이터들이 평균의 오른쪽으로 치우친 경우 :

Normal Probability Plot에서 점들이 대략 직선 형태를하면 정규성의 가정은 잘 만족하고 있다고 볼 수 있다.

Shapiro와 Wilk(1965)는 데이터의 정규성을 검정하는 W통계량을 제안하였는데 많이 활용되고 있다. 이 W통계량은 Normal Probability Plot에서의 $z_{(i)}$와 $u_{(i)}$의 상관 계수의 제곱 값이다. p-value를 구하기 위해서는 Shapiro와 Wilk에의한 표에서 찾아야 하지만 일반적으로 W통계량의 값이 너무 작으면 정규성 가설은 기각 된다.

3.4 Correlated Errors 문제

오차항의 또 다른 가정은 서로 독립적인 확률 변수라는 것이다. 확률 변수인 오차항의 독립성도 회귀 분석을 하는데 있어서 중요한 역할을 하는 것이 사실이다. 그러나 데이터가 시간적 또는 공간적 순으로 수집되었다면 인접하고 있는 케이스들은 서로 영향을 미칠 것이다. 따라서 회귀 분석 결과를 뒷받침 하기 위해서는 데이터들의 독립성을 입증하는 것이 필요하다. 그중에서 Durbin-Watson통계량이 많이 사용되는데 이것은 데이터가 일정한 시간 간격으로 수집되었을 때 인접한 케이스들의 상관 관계를 검정하는 것이다. Durbin-Watson의 d통계량은 다음과 같이 정의된다.

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{\sum_{i=2}^{n} e_i^2 + \sum_{i=2}^{n} e_{i-1}^2 - 2\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}$$

그런데, 여기서 만약 n이 크다면 $\sum_{i=1}^{n} e_i^2 = \sum_{i=2}^{n} e_i^2 = \sum_{i=2}^{n} e_{i-1}^2$이라고 할 수 있다. 따라서,

$$d \simeq \frac{2\sum e_i^2 - 2\sum e_i e_{i-1}}{\sum e_i^2} = 2(1-r)$$

r는 e_i와 e_{i-1}의 상관 계수로서,

$$r = \frac{\sum e_i e_{i-1}}{\sqrt{\sum e_i^2 \sum e_{i-1}^2}}$$

그러므로 r이 0 즉, e_i와 e_{i-1}이 무상관이면 d의 값은 2가 된다. r이 1에 접근 하면 d는 0에 접근하게 되고 r이 -1에 접근하면 d는 4에 접근한다. d의 값이 2 주위에 있으면 독립성의 가정은 만족한다고 할 수 있다.

데이터의 독립성에대한 진단은 일반적으로 어려우며 Durbin-Watson통계량 처럼 특수한 상황하에서 가능하다. 따라서, 가장 좋은 방법은 데이터의 발생 과정을 주의 깊게 살펴 보는 것이다.

4. 결론

지금까지 설명한 회귀 진단 방법들은 가장 기본적인 것들이다. 반복되는 표현이지만 최고의 또는 최선의 진단 방법은 없다. 회귀 진단이 다만 문제를 해결하는 데 있어 결과에 대한 신뢰성을 높여 주는 것이지 그것이 바로 답은 아니기 때문이다. 우리가 주어진 데이터를 가지고 회귀 분석을 한다고 하지만 우리의 회귀 모형 자체가 현상을 대략적으로 설명하는 것이다. 목적의 현상 규명을 위하여 종속 변수는 제대로 선택하였는지 또는 중요한 독립 변수가 실제로 빠져 있는지는 않은지 등등 회귀 진단이외에도 고려해야 할 사항들이 많을 것이다. 그리고 회귀 진단 방법을 선택할 때는 문제의 성격에 맞는 특수한 방법들도 고려해야 한다.

참고문헌

[1] Atkinson, A. C. (1973), "Testing Transformations to Normality," *Journal of Royal Statistical Society, Ser. B*, Vol. 35, pp. 473-479.

[2] Atkinson, A. C. (1981), "Robustness, Transformations and Two Graphical Displays for Outlying and Influential Observations in Regression," *Biometrika*, Vol. 68, pp. 13-20.

[3] Beckman, R. and R. D. Cook(1983), "Outlier...s," *Technometrics*, Vol. 25, pp. 119-149.

[4] Besley, D. A., E. Kuh, and R. E. Welsch(1980), *Regression Diagnostics*, New York: Wiley.

[5] Box, G. E. P. and D. R. Cox(1964), "An Analysis of Transformations(with discussion)," *Journal of Royal Statistical Society, Ser. B*, Vol. 26, pp. 211-246.

[6] Box, G. E. P. and P. W. Tidwell(1962), "Transformations of The Independent Variables," *Technometrics*, Vol. 4, pp. 531-550.

[7] Cook, R. D.(1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, Vol. 19, pp. 15-18.

[8] Cook, R. D. and S. Weisberg(1982), *Residuals and Influence in Regression*, London: Chapman Hall.

[9] Cook, R. D. and S. Weisberg(1983), "Diagnostics for Heteroscedasticity in Regression," *Biometrika*, Vol. 70, pp. 1-10.

[10] Durbin, J. and G. S. Watson(1950), "Testing for Serial Correlation in Least Squares Regression I," *Biometrika*, Vol. 58, pp. 1-19.

[11] Hoaglin, D. C. and R. E. Welsch(1978), "The Hat Matrix in Regression and ANOVA," *American Statistician*, Vol. 32, pp. 17-22.

[12] Shapiro, S. S. and M. B. Wilk(1965), "An Analysis of Variance Test for Normality(Complete Samples)," *Biometrika*, Vol. 52, pp. 591−611.

[13] Weisberg S.(1983), "Principles for Regression Diagnostics and Influence Analysis, discussion of a paper by R. R. Hocking," *Technometrics*, Vol. 25, pp. 240−244.

[14] Weisberg S.(1985), *Applied Linear Regression, 2nd ed.*, New York: Wiley.