Journal of the Korean Data & Information Science Society
2014, **25**(2), 305–315

http://dx.doi.org/10.7465/jkdi.2014.25.2.305
한국데이터정보과학회지

단순회귀모형에서 선형성 검정통계량[†]

박천건[1] · 이경은[2]

[1]경기대학교 수학과 · [2]경북대학교 통계학과

접수 2014년 1월 20일, 수정 2014년 1월 28일, 게재확정 2014년 2월 10일

요 약

전통적으로 단순선형회귀모형에서 설명변수와 반응변수의 선형성 평가는 산점도로 쉽게 파악되었다. 보통 반복수가 존재하는 자료에서 적합결여검정은 선형성을 평가하는데 사용되었다. 하지만 반복수가 오직 하나인 경우에 선형성 검정이 수월하지 않다. 본 연구에서는 반복수가 오직 하나인 단순선형회귀모형의 선형성을 검정하는 통계량을 제안하고 모의실험 및 실증연구를 통하여 신뢰성을 파악한다.

주요용어: 단순회귀모형, 적합결여검정, 평균기울기.

1. 머리말

단순선형회귀모형은 회귀분석에서 사용되는 가장 단순한 모형으로 한 개의 반응변수와 한 개의 설명변수와의 관계를 설명하는데 사용되며 다음과 같이 정의된다.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \cdots, n \tag{1.1}$$

여기서 회귀계수 β_0와 β_1은 각각 y절편과 기울기를 나타내는 미지의 모수이며, ϵ_i은 $E(\epsilon_i) = 0$, $Cov(\epsilon_i, \epsilon_j) = 0$, $i \neq j$이고 $Var(\epsilon_i) = \sigma^2$인 오차를 나타내는 확률변수로 일반적으로 정규분포를 따른다고 가정한다.

회귀분석은 회귀계수의 추정 및 검정을 하는 추론 부분과 사용할 회귀모형 및 가정의 타당성 (모형진단), 관측값이 모형 및 가정에 미치는 영향 (자료진단) 등을 확인하는 회귀진단 부분으로 나눌 수 있다. 모형진단에서는 반응변수와 설명변수의 선형 관계성 (선형성), 오차의 등분산성, 정규성, 독립성을 진단한다 (Seo와 Yoon, 2013). 만약 이러한 진단에서 어느 하나라도 만족하지 못하면 주어진 자료가 본 모형에 적용될 수 없게 된다 (Hocking, 2003).

주어진 자료가 제시된 모형에 적합하지 못한 원인은 무수히 많다. 설명변수와 반응변수의 비선형성이 적합성을 방해하는 가장 큰 원인 중에 하나이다 (Belsley 등, 1980; Park, 2013). 선형성을 만족한다고 가정하면, 이상값의 존재와 오차에 대한 가정의 위배 등이 본 모형의 적합성을 저해한다 (Barnett와 Lewis, 1984).

모형진단은 그래프적 기법과 가설검정을 통해 이루어진다. 그래프적 기법은 모형 적합 이전과 이후로 구분되어 수행하여 선형회귀모형의 적합성을 진단한다. 또한 오차항에 대한 가정의 진단은 오차의

[†] 이 논문은 2012학년도 경북대학교 학술연구비에 의하여 수행되었음.
[1] (443-760) 경기도 수원시 영통구 이의동 산 94-6, 경기대학교 수학과, 조교수.
[2] 교신저자: (702-701) 대구 광역시 북구 대학로 80, 경북대학교 통계학과, 조교수. Email: artlee@knu.ac.kr.

등분산성, 독립성 등을 그래프적 기법과 가설검정으로 평가한다 (Weisberg, 1985; Chatterjee와 Hadi, 2012). 본 연구에서는 단순회귀모형에서 모형진단에 관련하여 그래프적 기법에 대해서는 다루지 않고 선형성에 대한 가설검정에 초점을 맞추어 새로운 검정통계량을 제시한다.

제시할 검정통계량은 평균기울기의 개념을 이용하여 고안되었다. 단순선형회귀모형이 직선방정식과 동일하고 평균기울기가 설명변수 (x)의 변화량에 따른 반응변수 (y)의 변화량의 비율이 된다 (Stewart, 2007). 이러한 원리를 오차항이 없는 단순회귀모형에 적용하면, 평균기울기는 설명변수의 변화량의 크기와 상관없이 항상 회귀계수인 기울기이다. 또한 추정된 회귀식도 평균기울기의 원리가 동일하게 적용될 수 있다.따라서 반응변수와 추정 회귀식의 평균변화율을 기반을 한 비율은 일정한 값을 갖는다. 만약 추정된 회귀식이 잘 적합된 것이면 대응되는 반응변수와 그 비율은 1에 가까운 값이 된다. 이러한 이유로 임의의 서로 다른 두 점에 대한 반응변수와 추정된 회귀식의 평균기울기의 비율로 검정통계량을 설정한다.

본 연구의 구성은 다음과 같다. 2절에서는 선형성 측정에 대한 검토를 하고 3절은 선형성 검정을 위한 검정통계량을 제시하고 그 원리와 해석을 제시한다. 4절에서는 모의실험 및 실증분석을 통하여 제시된 검정통계량의 장점과 단점을 보여준다. 마지막으로 5절에서는 결론과 추후 연구를 제시한다.

2. 선형성의 강도

단순선형회귀모형에서는 반응변수와 설명변수의 선형적 관계 (선형성), 오차의 등분산성, 독립성, 정규성을 가정한다. 선형회귀모형의 적합성을 검정하기 위해서는 설명변수 x의 수준에서 반복측정이 있는 경우, 적합결여검정 (lack of fit test; LOF test)을 이용할 수 있으며, 반복측정이 오직 하나인 경우, 설명변수와 반응변수 또는 추정된 반응값과 잔차의 산점도를 통해 시각적으로 확인 할 수 있다.

2.1. 선형성의 강도

반응변수와 설명변수의 선형적 관계를 만족한다는 대가정하에서, 선형성의 강도는 다음의 여러 방법으로 확인할 수 있다. 첫 번째로, 반응변수 (y)와 설명변수 (x)의 상관계수를 통해 알 수 있는 데, 상관계수의 절대값이 1에 가까울수록, 반응변수와 설명변수간의 선형성의 강도는 크다. 두 번째로, 반응변수 y와 적합된 $\widehat{y}$ 사이의 상관계수를 통해서도 알 수 있다. 실질적으로 두 상관계수는 다음과 같은 관계에 있다.

$$|corr\,(x,y)| = corr\,(y,\widehat{y})\,.$$

세 번째로, 결정계수를 통해서도 알 수 있다. y의 제곱편차의 총합 $\left(=\sum_{i=1}^{n}(y_i-\bar{y})^2\right)$을 SST (total sum of squared deviations), 회귀에 기인한 제곱합 $\left(=\sum_{i=1}^{n}(\widehat{y_i}-\bar{y})^2\right)$을 SSR (sum of squares due to regression), 잔차제곱합 $\left(=\sum_{i=1}^{n}(y_i-\widehat{y_i})^2\right)$을 SSE (sum of squared errors)이라고 하면 $SST = SSR + SSE$의 관계가 성립한다. 결정계수 (R^2)는 SSR/SST으로 반응변수의 변동 중에서 회귀모형에 의해서 설명되는 변동을 나타내므로 1에 가까울수록 반응변수와 설명변수사이에 강한 선형관계가 있음을 알 수 있다.

2.2. 기울기의 가설검정

단순선형회귀모형에서 선형관계의 타당성을 보일 수 있는 방법 중 하나가 기울기 β_1에 대한 가설 검정을 통해서이다. 오차의 독립성, 정규성, 등분산성 가정 하에서 β_1의 최소제곱추정량 $\widehat{\beta}_1$의 형태와 분

포는 다음과 같다.

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \quad \widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right).$$

잔차제곱합의 분포는 카이제곱분포를 따르고 $\left(\frac{SSE}{\sigma^2} \sim \chi^2_{n-2}\right)$, $\widehat{\beta}_1$와 SSE가 독립이므로 기울기 $\beta_1 = 0$에 대한 가설검정의 검정통계량의 분포는 귀무가설하에서 다음과 같게 된다.

$$t = \frac{\widehat{\beta}_1}{\widehat{\sigma}/\sqrt{\sum(x_i-\bar{x})^2}} \sim t_{n-2}.$$

여기서 $\widehat{\sigma} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$이다. 귀무가설 ($\beta_1 = 0$)이 의미하는 것은 반응변수 ($y$)와 설명변수 ($x$)에 선형적인 관계가 없다는 것이다. 귀무가설을 기각하는 것은 y와 x사이에 선형적인 관계가 있다는 것을 의미하는 것이지 반드시 y와 x사이에 선형적인 관계가 강하다고 단정 지을 수 없다.

2.3. 적합결여검정

선형성의 적합성 검정을 하기 위해서는 설명변수 x의 수준 $x_1, x_2, ..., x_m$에서 y값들이 반복적으로 관측된 경우에만 가능하다. x의 각 수준인 x_i에서 관측값의 수를 n_i라고 하자 (모든 n_i의 값이 1보다 클 필요는 없다). 잔차제곱합을 순수오차분에 해당하는 제곱합 (sum of squares due to pure error; SS_{PE})과 적합결여에 따른 제곱합 (sum of squares due to lack of fit; SS_{LOF})으로 나누어 볼 수 있다.

$$SSE = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\widehat{y}_{ij})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_i-\widehat{y}_{ij})^2 = SS_{PE} + SS_{LOF}.$$

귀무가설 $H_0 : E(y_{ij}) = \beta_0 + \beta_1 x_i,\ i = 1, 2, ..., m, j = 1, ..., n_i$에 대한 적합결여검정통계량과 귀무가설하에서의 분포는 다음과 같게 된다.

$$F_{LOF} = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} \sim F_{m-2,n-m}. \tag{2.1}$$

F_{LOF}가 $F_\alpha(m-2, n-m)$보다 크게 되면 귀무가설을 기각, 유의수준 α에서 선형성 가정이 타당하다고 결론지을 수 없다.

3. 단순회귀모형의 선형성에 관한 가설검정

3.1. 평균 기울기

식 (1.1)의 단순선형회귀모형은 설명변수 (x)의 일차함수와 평균이 영이고 분산이 존재하는 오차항의 합으로 구성된 모형이고, 주요 관심인 회귀계수를 파악하는데 있다. 특히, 기울기 (β_1)는 오차항이 없는 임의의 서로 다른 두 x의 값과 대응되는 평균 회귀값 ($z = \beta_0 + \beta_1 x$)의 평균 기울기로 구할 수 있다. 또한 상수항 (β_0)도 직선의 방정식으로 구할 수 있다 (Stewart, 2007).

$$\widehat{\beta}_{(i,j)} = \frac{z_{(i)} - z_{(j)}}{x_{(i)} - x_{(j)}}, \quad \text{단 } i \neq j.$$

여기서 설명변수의 순서는 $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$이고 평균 회귀값의 순서도 $z_{(1)} < z_{(2)} < \cdots < z_{(n)}$ (또는 $z_{(1)} > z_{(2)} > \cdots > z_{(n)}$)이라 가정하자.

그러나 일반적인 다중회귀모형에서 설명변수 또는 평균 회귀값의 순서는 제공되지 않는다. 그러나 단순회귀모형에서는 설명변수와 평균 회귀값의 순서를 모른다고 가정해도, 평균 회귀값의 순서가 추정된 회귀값 ($\widehat{y}$)의 오름차순 (또는 내림차순)으로 쉽게 파악된다. 이 순서를 이용하여 추정된 기울기 ($\widehat{\beta}_1$)도 서로 다른 두 설명변수의 값과 대응되는 추정된 회귀값 ($\widehat{y}$)의 평균 기울기로 나타낼 수 있다.

$$\widehat{\beta}_1 = \frac{\widehat{y}_{(i)} - \widehat{y}_{(j)}}{x_{(i)} - x_{(j)}}, \quad i \neq j. \tag{3.1}$$

여기서 추정된 회귀값의 순서는 $\widehat{y}_{(1)} < \widehat{y}_{(2)} < \cdots < \widehat{y}_{(n)}$ (또는 $\widehat{y}_{(1)} > \widehat{y}_{(2)} > \cdots > \widehat{y}_{(n)}$)이고 이에 대응되는 설명변수의 순서를 $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$라 하고 $x_{(i)}$에 대응되는 y값을 $y_{(i)}$라고 하자 (설명변수 $x_{(i)}$에서 반복 측정된 경우는 $x_{(i)}$에 대응되는 y들의 평균값을 $y_{(i)}$라고 하자). 반응변수에 대한 평균 기울기는 기울기 (β_1)와 (i, j)번째 오차의 평균기울기의 합으로 표현된다.

$$\widetilde{\beta}_{1_{(i,j)}} = \frac{y_{(i)} - y_{(j)}}{x_{(i)} - x_{(j)}} = \beta_1 + E_{(i,j)}, \quad \text{단 } i \neq j. \tag{3.2}$$

여기서 $E_{(i,j)} = \left(\epsilon_{(i)} - \epsilon_{(j)}\right) / \left(x_{(i)} - x_{(j)}\right)$.

$\widetilde{\beta}_{1_{(i,j)}}$의 통계적 성질을 살펴보면, $E(\widetilde{\beta}_{1_{(i,j)}})=\beta_1$인 불편추정량이고 $Var(\widetilde{\beta}_{1_{(i,j)}})=2\sigma^2/(x_{(i)}-x_{(j)})^2$으로 설명변수의 서로 다른 두 값의 간격이 벌어지면 질수록 분산은 작아진다.

3.2. 검정통계량

식 (3.1)과 식 (3.2)로부터 단순선형회귀모형의 선형성을 평가하기 위해서 관찰값의 평균기울기와 추정된 회귀값의 평균기울기의 비가 적용된다.

$$\text{Ratio} = \frac{\widehat{y}_{(i)} - \widehat{y}_{(j)}}{y_{(i)} - y_{(j)}}, \left(\text{또는 } \frac{y_{(i)} - y_{(j)}}{\widehat{y}_{(i)} - \widehat{y}_{(j)}}\right). \tag{3.3}$$

여기서 i와 j는 임의의 다른 추정값 및 대응되는 관찰값이다.

만약 추정된 회귀식이 관찰된 자료를 잘 설명한다면, 식 (3.3)의 비 ($Ratio$)은 1에 가까운 값을 가질 것이고 $\widehat{\beta_1} \approx \widetilde{\beta}_{1(i,j)}$인 것을 쉽게 추론할 수 있다. 따라서 $y_{(i)} - y_{(j)} \approx \widehat{y}_{(i)} - \widehat{y}_{(j)}$, $i \neq j$로 표현될 수 있다. 결론적으로 단순선형회귀모형의 선형성에 대한 평가를 순서로 정렬된 관찰값과 추정값의 차 (또는 합)를 이용하려고 한다.

$$M_{(i,j)-} = (\widehat{y}_{(i)} - y_{(i)}) - (\widehat{y}_{(j)} - y_{(j)}) \text{ 또는 } M_{(i,j)+} = (\widehat{y}_{(i)} - y_{(i)}) + (\widehat{y}_{(j)} - y_{(j)}). \tag{3.4}$$

식 (3.4)에서 제시된 통계량은 i-번째와 j-번째의 잔차로 표현했지만 본 논문에서 적용되는 순서는 첫 번째와 마지막 위치에 있는 값의 차를 이용한 범위로 변형하면 다음과 같은 통계량으로 표현된다.

$$M_{(1,n)-} = (\widehat{y}_{(1)} - y_{(1)}) - (\widehat{y}_{(n)} - y_{(n)}) \text{ 또는 } M_{(1,n)+} = (\widehat{y}_{(1)} - y_{(1)}) + (\widehat{y}_{(n)} - y_{(n)}). \tag{3.5}$$

식 (3.5)에서 제시된 통계량에 대한 가설은 선형성의 적합성을 결정하는데 적용될 수 있다.

$$\begin{aligned} &\mathrm{H}_0 : E(M_{(1,n)-}) = 0 \Leftrightarrow \text{ 선형} \qquad \mathrm{H}_0 : E(M_{(1,n)+}) = 0 \Leftrightarrow \text{ 선형} \\ &\text{또는} \\ &\mathrm{H}_\mathrm{A} : E(M_{(1,n)-}) \neq 0 \Leftrightarrow \text{ 선형이 아님} \quad \mathrm{H}_\mathrm{A} : E(M_{(1,n)+}) \neq 0 \Leftrightarrow \text{ 선형이 아님} \end{aligned} \tag{3.6}$$

또한 식 (3.6)에서 제시된 통계량을 행렬의 형태로 변환하면 정규분포를 따름을 쉽게 알 수 있다. 식 (1.1)로부터 $\boldsymbol{Y} \sim N\left(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}\right)$이고 최소제곱법으로 추정된 회귀값의 오름차순에 대응되는 관찰값의 분포를 $\boldsymbol{Y_o} \sim N\left(\boldsymbol{X_o\beta}, \sigma^2\boldsymbol{I}\right)$라고 하자.

$$M_{(1,n)-} = d_1^T(I - H_o)Y_o \sim N(0, \sigma^2 d_1^T(I - H_o)d_1)$$
또는
$$M_{(1,n)+} = d_2^T(I - H_o)Y_o \sim N(0, \sigma^2 d_2^T(I - H_o)d_2). \tag{3.7}$$

여기서 $\boldsymbol{Y} = (y_1, ..., y_n)^T$ (또는 $\boldsymbol{Y_o} = (y_{(1)}, ..., y_{(n)})^T$)인 $n \times 1$벡터, $\mathbf{1} = (1, ..., 1)^T n \times 1$벡터와 $X = (x_1, ..., x_n)^T$ (또는 $X_o = (x_{(1)}, ..., x_{(n)})^T$) $n \times 1$벡터로 구성된 $\boldsymbol{X} = (\mathbf{1}, X)$ (또는 $\boldsymbol{X_o} = (\mathbf{1}, X_o)$)인 $n \times 2$행렬 그리고 $d_1 = (-1, 0, 0, \cdots, 0, 0, 1)^T$ (또는 $d_2 = (1, 0, 0, \cdots, 0, 0, 1)^T$)라고 하자. 또한 $H_o = \boldsymbol{X_o}(\boldsymbol{X_o}^{\boldsymbol{T}}\boldsymbol{X_o})^{-1}\boldsymbol{X_o^T}$이다.

식 (3.6)과 식 (3.7)로부터 제시된 검정통계량은 다음과 같다.

$$T_- = \frac{M_{(1,n)-}}{s.e.(M_{(1,n)-})} \text{ 또는 } T_+ = \frac{M_{(1,n)+}}{s.e.(M_{(1,n)+})}. \tag{3.8}$$

식 (3.7)의 가설은 개별적으로 가설검정을 수행한 것으로 반대의 결과가 나올 수 있어 동시에 검정할 수 있는 검정통계량과 가설은 설정하면 다음과 같다.

$$\boldsymbol{M} = \begin{pmatrix} M_{(1,n)-}2 \\ M_{(1,n)+} \end{pmatrix} = \boldsymbol{D}^T(\boldsymbol{I} - \boldsymbol{H}_0)\boldsymbol{Y}_0 \sim N(\mathbf{0}, \sigma^2\boldsymbol{D}^T(i - \boldsymbol{H}_0)\boldsymbol{D}) \tag{3.9}$$

$$\mathrm{H_0} : E(\boldsymbol{M}) = \mathbf{0} \Leftrightarrow \text{선형}$$
$$\mathrm{H_A} : E(\boldsymbol{M}) \neq \mathbf{0} \Leftrightarrow \text{선형이 아님.}$$

여기서 $\boldsymbol{D} = (d_1, d_2)$이며, 귀무가설하에서 식 (3.9)의 검정통계량은 다음과 같다.

$$F = \frac{\boldsymbol{M}^T(\boldsymbol{D}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{D})^{-1}\boldsymbol{M}/2}{\sigma^2} \tag{3.10}$$

단, 설명변수에 반복이 있는 자료인 경우에는 d의 행벡터에 반복수만큼 나눔으로 y값 대신 반복된 x값에 대응되는 y값들의 평균을 사용하여 각 검정통계량을 구한다.

식 (3.8)과 식 (3.10)에서의 검정통계량의 분포는 σ^2의 추정량으로 앞 절에서처럼 MSE를 이용하게 되면 분자와 분모가 독립이 아니어서 t분포나 F분포를 따르지 않는다. 정확한 분포를 구하는 것이 쉽지 않으므로 붓스트랩 (bootstrap) 방법을 이용하여 유의확률을 다음의 절차를 이용하여 구할 수 있다.

1. 원자료 $(x_1, y_1), ..., (x_n, y_n)$을 이용하여 회귀계수를 추정, 적합된 회귀선과 잔차를 구한다.
$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$
$$e_i = y_i - \widehat{y}_i$$
2. 잔차들의 붓스트랩된 샘플, $\boldsymbol{e}_b^* = (e_{b1}^*, e_{b2}^*, \ldots, e_{bn}^*)^T$을 이용하여 붓스트랩된 y값, $\boldsymbol{y}_b^* = \widehat{\boldsymbol{y}} + \boldsymbol{e}_b^*$을 생성한다.
3. $\boldsymbol{y}_b^*$와 고정된 x을 이용하여 회귀선을 적합하여 검정통계량 $T_{+b}^*, T_{-b}^*, F_b^*$를 계산한다.

4. 절차 1-3을 B번 반복한 후 유의확률을 구한다.

$$P_{T+} = \frac{1}{B}\sum_{b=1}^{B}(|T^*_{+b}| > T_+)$$
$$P_{T-} = \frac{1}{B}\sum_{b=1}^{B}(|T^*_{-b}| > T_-)$$
$$P_F = \frac{1}{B}\sum_{b=1}^{B}(|F^*_b| > F).$$

4. 모의실험 및 실증연구

이번 절에서는 모의실험과 Anscombe (1973)의 4개의 인공자료 분석을 통하여 앞에서 제안한 검정방법의 타당성을 보이려고 한다.

먼저, 모의실험에서는 다음의 4가지 상황을 고려하여 보았다: 전형적인 단순선형회귀모형 (Model 1), 선형성 가정이 위배되는 이차모형 (Model 2), 선형성은 만족되나 이상점을 가지는 모형 (Model 3), 선형성을 만족하나 등분산성이 위배되는 이분산모형 (Model 4). 구체적인 모형식은 다음과 같다.

$$Model1 : y_{ij} = 3 + 0.5x_i + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0,1)$$
$$Model2 : y_{ij} = -0.2\,(x_i - 9)^2 + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0,1)$$
$$Model3 : y_{ij} = 3 + 0.5x_i + d_{ij} + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0,1)$$
$$Model4 : y_{ij} = 3 + 0.5x_i + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0, w_i^2)$$

여기서 $i = 1, 2, ..., 13,\ j = 1, 2, 3$이다. 설명변수 x_i는 $x_i = i + 3$로 두고, 각 수준 x_i에서, $N(0,1)$을 따르는 오차들을 3개씩 생성, 모회귀식 (population linear regression line)에 더하여서, 3개의 관측값 (y_{i1}, y_{i2}, y_{i3})들을 생성하였다. Model 3은 평균 이상점 모형 (mean shifted outlier model)으로 이상점 변수 d_{ij}가 0인 경우는 (i,j)번 째 관측값이 이상점이 아닌 경우이고 d_{ij}가 0아닌 다른 값을 가진 경우는 관측값이 이상점인 경우이다. 편의상 $d_{ij} = 0$ 또는 5라고 정하고, 이상점 수는 3개로 고정시켰다. 이상점 위치는 39개 중에서 램덤하게 선택하였다. Model 4에서 이분산과 관련된 w_i는 $w_i = -0.25x_i + 4.25$로 두었다. Figure 4.1은 설정된 모형들 가정 하에 생성된 자료들의 산점도이다.

각 모형을 1000번 반복하여 자료를 생성하였다. 반복관측이 있는 경우 적합결여검정으로 모형의 선형성을 검정할 수 있으므로, 앞에서 제안한 세 가지 검정 방법과의 비교를 위해 적합결여검정도 수행하였다. 식 (2.1), (3.8) 그리고 (3.10)로부터 네 가지 검정방법의 통계량들 (F, T_-, T_+, F_{LOF})의 P-값들과 결정계수를 계산하여 평균과 표준편차를 Table 4.1에서 정리하였고 Figure 4.2는 네 가지 검정방법들에 대한 P-값들의 상자그림이다. P-값은 귀무가설하 (선형성 만족)에서는 균등분포를 따르므로 앞에서 설정한 모형하에서는 Model 1, 3, 4에서 P-값들은 구간 $(0,1)$에서 랜덤하게 분포할 것을 예상할 수 있으며 Table 4.1과 Figure 4.2에서 확인할 수 있다. 유의수준 $\alpha = 0.05$하에서, Model 1, 3, 4는 선형성을 만족하므로 제 1종 오류율을, Model 2는 선형성을 만족하지 않으므로 제 2종 오류율을 Table 4.2에서 정리하였다. Table 4.2에서 보듯 우리가 제안한 세 가지 검정통계량들의 오류율은 Model 1, 2, 4에서는 적합결여 검정보다 낮은 제 1종 오류율을 보이고 있다. 선형성 가정이 위배된 Model 2에서는 네 가지 검정법 모두 제 2종 오류율을 0으로 보이고 있다.

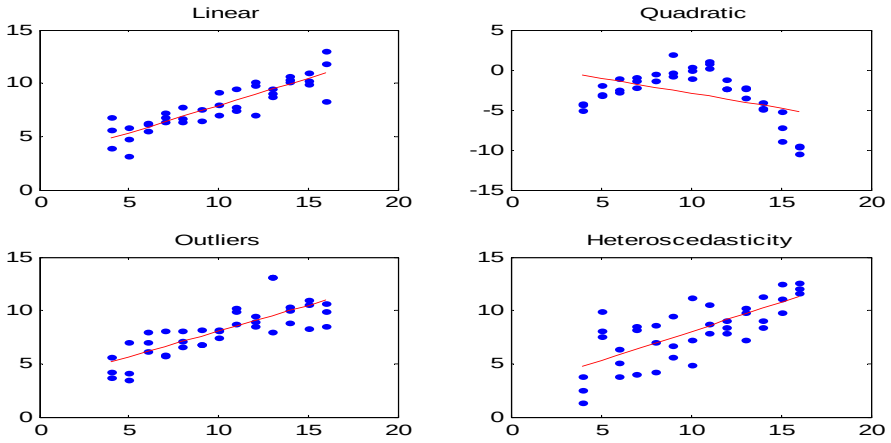

Figure 4.1 Scatter plots of an explanatory variable and a response variable for four type data sets

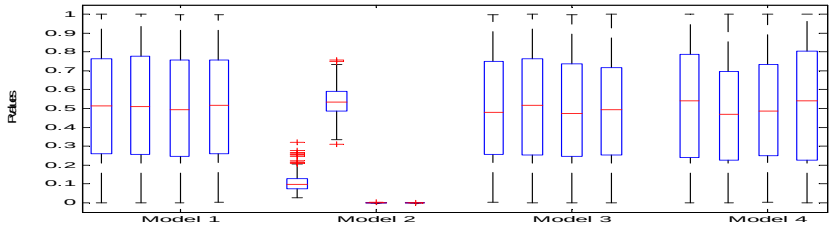

Figure 4.2 Boxplots of P-values of F, T_-, T_+ and F_{LOF} for four models

Table 4.1 Mean and standard deviations of P-values and R^2

Test Statistic	Model							
	1		2		3		4	
F	0.508	(0.293)	0.103	(0.073)	0.497	(0.290)	0.514	(0.283)
T_-	0.501	(0.293)	0.000	(0.040)	0.485	(0.288)	0.490	(0.305)
T_+	0.514	(0.291)	0.538	(0.000)	0.508	(0.288)	0.469	(0.286)
F_{LOF}	0.507	(0.291)	0.000	(0.000)	0.516	(0.263)	0.501	(0.311)
R^2	0.788	(0.046)	0.241	(0.042)	0.478	(0.154)	0.353	(0.120)

Table 4.2 Ratios of Type I (or II) error for four test statistics ($\alpha = 0.05$)

Test Statistic	Model			
	1	2	3	4
	Type I error	Type II error	Type I error	Type I error
F	0.050	0.941	0.046	0.084
T_-	0.045	1.000	0.044	0.063
T_+	0.047	0.000	0.062	0.052
F_{LOF}	0.052	0.000	0.048	0.092

앞의 모의실험은 반복이 존재하는 자료에 대해서 선형성의 가설검정을 살펴보았고, Table 4.3으로부터 반복수가 오직 하나인 경우에 Anscombe (1973)는 결정계수, 추정된 회귀선 등 주요 기초 통계량들이 같은 네 가지 특성의 자료를 제시하였다. 이들 자료는 단순히 수치적인 접근으로 단순선형회귀모형

의 평가를 하면 동일한 결론이 도출되어 정확한 회귀진단은 시각적인 기법이 적용되어야만 한다. 본 논문에서는 앞에서 제안한 검정 방법을 이들 자료에 적용하여 선형성을 검정해 보이려고 한다.

Table 4.3과 Figure 4.3으로부터 Data 1은 전형적인 선형모형이고, Data 2는 이차함수관계가 있고, Data 3은 한 개의 이상점이 있는 모형 그리고 Data 4는 지렛점 (leverage point)이 존재하는 모형이다. 이들 자료들의 회귀분석 결과인 추정된 회귀선 ($\widehat{y} = 3 + 0.5x$), 결정계수 ($R^2 = 0.67$), 회귀제곱합 (SSR=27.47 27.51), 잔차제곱합 (SSE=13.74 13.78), 기울기 추정량의 표준오차 ($s.e.(\widehat{\beta}_1)$=0.0139) 등이 거의 동일하다.

Table 4.3 Four types of data sets

Data 1		Data 2		Data 3		Data 4	
x1	y1	x2	y2	x3	y3	x4	y4
4	4.26	4	3.1	4	5.39	8	5.25
5	5.68	5	4.74	5	5.73	8	5.56
6	7.24	6	6.13	6	6.08	8	5.76
7	4.82	7	7.26	7	6.42	8	6.58
8	6.95	8	8.14	8	6.77	8	6.89
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	7.71
11	8.33	11	9.26	11	7.81	8	7.91
12	10.84	12	9.13	12	8.15	8	8.47
13	7.58	13	8.74	13	12.74	8	8.84
14	9.96	14	8.1	14	8.84	19	12.5

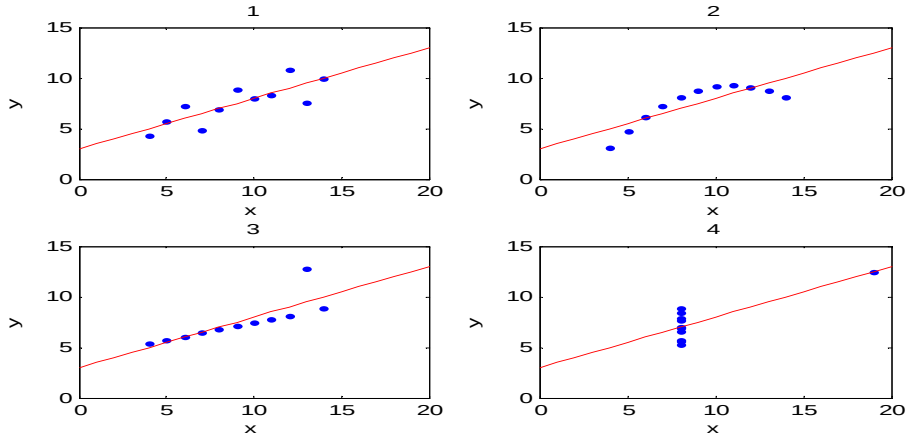

Figure 4.3 Scatter plots of x versus y for four types

Table 4.4은 제시된 모형들의 선형성을 알아보는 검정통계량의 P-값들이 제시되었다. Data 1에 대해서 모든 검정통계량의 P-값들이 0.05보다 상당히 큰 값으로 계산되어 선형성이 존재한다는 옳은 결론에 도달했고, 비선형인 Data 2에서는 T_-만이 P-값이 1로 강한 선형관계의 잘못된 결론을 내렸고, 한 개의 이상치가 존재하는 Data 3에서는 유의수준 0.05하에서 모든 검정통계량이 선형성이 유지되는 결론에 도달했다. Data 4에서 모든 통계량의 P-값들이 매우 크며 매우 강한 선형성이 존재하는 결론에 도달했다.

Table 4.4 P-values of three test statistics

Test Statistic	Data			
	1	2	3	4
F	0.789	0.024	0.434	0.542
T_-	0.605	1	0.266	0.694
T_+	0.682	0.007	0.697	0.702

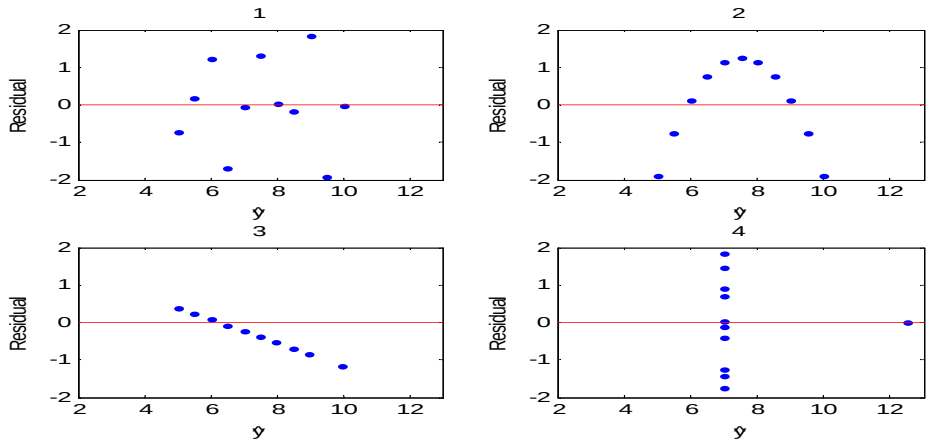

Figure 4.4 Residual plots for four type datasets

5. 결론 및 추가연구

선형회귀모형에서 설명변수와 반응변수의 선형관계에 대한 적합성 평가는 매우 중요하고 기본적인 모형진단 중에 하나이다. 전통적인 진단 방법은 시각적인 접근과 수치적인 접근으로 구분한다. 단순선형회귀모형에서 시각적인 접근은 설명변수와 반응변수의 산점도로 쉽게 선형성을 파악할 수 있다. 하지만 다중선형회귀모형에서는 산점도만으로 설명변수들과 반응변수의 선형관계를 파악하기가 쉽지 않다. 또한 반응값과 추정된 회귀값의 상관계수나 회귀계수의 가설검정만으로 그 선형성의 적합성을 판단하기란 쉽지 않다. 그래서 잔차를 이용한 여러 절차를 통하여 선형성 및 기타 오차항의 가정을 진단하는 것이 전통적인 방식이다.

본 연구는 단순선형회귀모형의 선형성을 평가하는데 평균기울기의 원리를 적용하여 시각적인 관점에서 벗어나 수치적인 정보를 제공할 수 있는 검정통계량을 개발하고 그 신뢰성을 제공하였다. 식 (3.4)의 비율을 토대로 개발된 세 가지 검정통계량은 설명변수에서 얻어진 순서의 정보를 통해서 최소 반응값과 최대 반응값에 대응되는 잔차들로 구성되었다. 식 (3.8)은 최소 반응값과 최대 반응값에 대응되는 잔차들의 차 또는 합으로 고안된 검정통계량이고 식 (3.10)은 식 (3.8)의 두 검정통계량을 동시에 수행하는 검정통계량이다. 반복이 있는 단순선형회귀모형에 대한 모의실험 결과인 Table 4.1, Table 4.2 그리고 Table 4.4로부터 기존의 적합결여검정 (LOF test)가 새롭게 제안한 세 검정통계량과 비슷한 것으로 나타났다. 또한, Anscombe (1973)의 인공적인 4가지 자료들에 적용하여 보았는데, 이 자료들은 모두 설명변수 각 수준에서 반복이 오직 하나인 경우로써 T_+가 옳은 결정을 보이고 있다.

평균기울기는 선형관계를 가지고 있는 두 변수의 기울기를 나타내는데 매우 효율적인 측정도구이다. 하지만 전통적으로 선형모형에서 그 선형성을 검정하는데 잔차에 대한 시각적인 표현에 많이 의존해 왔다. 따라서 본 연구에서 제안된 세 검정통계량이 선형회귀모형에서 선형성을 평가하는 수치적인 정보로 매우 유용하며, 앞으로 다중선형회귀모형의 선형성 평가에도 적용될 수 있는 연구가 필요하다.

References

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, **27**, 17-21.

Barnett, V. and Lewis, T. (1984). *Outliers in statistical data*, 2nd ed., Wiley, New York.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*, Wiley, New York.

Chatterjee, S. and Hadi, A. S. (2012). *Regression analysis by example*, Wiley, New York.

Hocking, R. R. (2003). *Methods and applications of linear models: Regression and the analysis of variance*, Wiley, New York.

Lee, H. Y. (2013). Goodness-of-fit tests for a proportional odds model. *Journal of the Korean Data & Information Science Society*, **24**, 1465-1475.

Park, H. C. (2013). Non-linear regression model considering all association thresholds for decision of association rule numbers. *Journal of the Korean Data & Information Science Society*, **24**, 267-275.

Seo, H. S. and Yoon, M. (2013). Regression diagnostics for response transformations in a partial linear model. *Journal of the Korean Data & Information Science Society*, **24**, 33-39.

Stewart, J. (2007). *Calculus*, 6th ed., Cengage Learning, Stamford.

Weisberg, S. (1985). *Applied linear regression*, Wiley, New York.

Journal of the Korean Data & Information Science Society
2014, **25**(2), 305–315

http://dx.doi.org/10.7465/jkdi.2014.25.2.305
한국데이터정보과학회지

A linearity test statistic in a simple linear regression[†]

Chun Gun Park[1] · Kyeong Eun Lee[2]

[1]Department of Mathematics, Kyonggi University
[2]Department of Statistics, Kyungpook National University

Received 20 January 2014, revised 28 January 2014, accepted 10 February 2014

Abstract

In a simple linear regression, a linear relationship between an explanatory variable and a response variable can be easily recognized in the scatter plot of them. The lack of fit test for the replicated data is commonly used for testing the linearity but it is not easy to test the linearity when the explanatory variable is not replicated. In this paper, we propose three new test statistics for testing the linearity regardless of replication using the principle of average slope and validate them through several simulations and empirical studies.

Keywords: Average slope, lack of fit test, simple linear regression.

[†] This research was supported by Kyongpook National University Research Grant 2012.
[1] Assistant professor, Department of Mathematics, Kyonggi University, Gyeonggi-do 443-760, Korea.
[2] Corresponding author: Assistant professor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail: artlee@knu.ac.kr