응용통계연구
제 4 권 제 2 호, 1991년

尤度거리에 의한 결정계수 R^2에의 통합적 접근

허명회* · 이종한** · 정진환***

요 약

결정계수 R^2은 회귀분석에서 실제적으로는 매우 이용도가 높은 기술 측도라고 하겠으나, 회귀모형이 절편항을 포함하는 표준적인 선형회귀모형 이외인 경우에는 결정계수의 정의에 관하여 여러 논란이 있어 왔다. 절편항이 없는 선형회귀모형에서와 가중선형회귀모형, 로버스트 선형회귀모형에서의 결정계수의 적절한 정의와 용법이 대표적인 문제라고 하겠다. 기존의 여러 연구, 예를 들어 Kvalseth(1985)나 Willet and Singer(1988)에서는 이러한 각 경우에 각기 적용될 수 있는 결정계수의 여러 변형들을 제안·비교하고 있다. 그러나 이런 기존의 연구들이 일반적인 원칙이 없이 경우별로 단편적으로 대응하고 있을 뿐더러 약간의 오류를 포함하고 있어 오히려 통계전문가가 아닌 통계 이용자들에게 혼란을 불러 일으킬 염려가 있다.

따라서 결정계수의 일반적 정의를 제안한 본 연구는 현재와 같은 결정계수의 여러 변종의 범람으로 인한 혼란을 없애는 데 기여하리라고 생각된다. 이 통합결정계수는 尤度거리(likelihood distance)를 이용하여 정의되는데, 선형회귀모형 이외에도 비선형 회귀모형과 일반화 선형모형에 일관되게 적용 가능하다는 장점을 갖는다.

1. 서 론

결정계수 R^2은 모형의 적합정도를 알아보는 측도의 하나로서 종속변수 Y가 연속형인 선형회귀모형에서 널리 사용되고 있다. 선형회귀모형이 절편항을 포함하는 경우, R^2의 정의에 관하여 異論이 없는 것은 아니지만 여러 동치적인 표현이 가능하다(Kvalseth, 1985). 예를 들자면 선형회귀모형이 절편항을 포함하고 최소제곱법으로 적합값을 구할 때

$$R_1^2 = 1-\Sigma(y_i-\hat{y}_i)^2/\Sigma(y_i-\bar{y})^2.$$

$$R_2^2 = \Sigma(\hat{y}_i-\bar{y})^2/\Sigma(y_i-\bar{y})^2$$

* (136-701) 서울특별시 성북구 안암동 고려대학교 통계학과 교수
** (136-701) 서울특별시 성북구 안암동 고려대학교 통계학과 박사과정
*** (136-701) 서울특별시 성북구 안암동 고려대학교 통계학과 석사과정 졸업

은 수학적으로 완전히 동치이다. 그러나 절편항을 갖지 않는 선형회귀모형의 경우에는 R_1^2과 R_2^2이 다를 수 있으며 더우기 0과 1 사이 구간을 벗어나는 값을 취할 수 있기 때문에 결정계수로서의 최소한의 자질도 의심스럽다고 하겠다. 이런 경우에는 앞의 R_1^2과 R_2^2에서 $\bar{y}$를 없앤 새로운 정의가 유용한 것으로 알려져 있지만 이 새로운 정의에 대한 명확한 해석은 알려져 있지 않다.

Kvalseth(1985)는 이 새로운 정의를 언급하고 있으나 절편항이 없는 경우에도 이것 보다 오히려 R_1^2을 사용할 것을 추천하고 있다. 또한 Uyar and Erdem(1990)도 SAS(Statistical Analysis System)의 회귀모형 분석과정의 문제점을 지적하면서 절편항이 없는 회귀모형 경우에 결정계수 R_1^2을 사용해야 한다고 주장하였는데 이와 같은 주장들은 그들이 새로운 정의의 의미를 적절히 해석할 수 없었기 때문인 것으로 생각된다.

이와 유사한 맥락에서 가중회귀모형(weighted regression model)에서도 결정계수의 수정에 관하여 논란이 있어 왔다. 예컨데 Willett and Singer(1988)는 Kvalseth(1985)의 관점에 의한 R_{WLS}^2을 pseudo R_{WLS}^2으로 수정 · 제안하였다.

본 연구에서는 결정계수의 이러한 여러 변종의 범람으로 인한 개념적 혼란과 오용을 타파하기 위하여 모든 선형회귀모형에 적합측도로서 적용할 수 있는 새로운 개념의 통합결정계수(unified coefficient of determination) R_U^2을 尤度거리(likelihood distance)를 이용하여 정의할 것이다. 그런데 이 통합결정계수는 개념상으로 쉽게 비선형회귀모형, 로지스틱회귀모형 등의 일반화 선형모형(generalized linear model)에 일관되게 적용 가능하다는 利點이 있다.

본고의 2절에서는 Kvalseth(1985)와 Willett and Singer(1988)를 소개하고 이들 내용의 오류를 바로 잡았으며, 3절에서는 통합결정계수 R_U^2을 일반화 선형모형의 맥락에서 정의하였고 이 통계량이 갖고 있는 성질들을 살펴 보았다. 마지막으로 4절에서는 비선형 회귀모형, 로지스틱 회귀모형 그리고 로그 선형모형의 통합 결정계수를 실제 분석사례와 함께 제시하였다.

2. 기존의 결정계수에 대한 비판적 고찰

다음과 같은 절편항을 포함하는 단순 선형회귀모형을 생각해 보자.

$i=1, \cdots, n$에 대하여

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{2.1}$$

여기서 ε_i는 평균이 0이고 분산이 σ^2인 정규분포를 따른다고 가정한다. 만약 설명변수 X가 없다면, 즉 $\beta_1=0$라면 모든 i번째 적합값 $\hat{y}_i$은 최소제곱법에 의해 $\bar{y}$로 주어지고, 이 때 오차제곱합은 $\Sigma(y_i-\bar{y})^2$이 된다. 그러나 β_1에 어떤 제한을 두지 않으면 $\hat{y}_i$은 $\hat{\beta}_0+\hat{\beta}_1 x_i$로 계산되고(여기서 $\hat{\beta}_0$와 $\hat{\beta}_1$은 β_0와 β_1의 최소제곱 추정치이다), 이때의 오차제곱합은 $\Sigma(y_i-\hat{y}_i)^2$이 된다. 이제 결정계수 R^2을 이 두 오차제곱합의 비율에 의해 다음과 같이 정의해 보기로 하자.

$$R^2 = 1 - \frac{\beta_1\text{에 어떤 제한도 두지 않았을 때의 오차제곱합}}{\beta_1\text{을 0으로 두었을 때의 오차제곱합}}$$

위의 정의를 따를 경우 절편항을 포함하는 단순 선형회귀모형의 결정계수는

$$R^3 = 1-\Sigma(y_i-\hat{y}_i)^2/\Sigma(y_i-\bar{y})^2 \qquad (2.2)$$

과 같이 되고 이것은 Kvalseth(1985)가 열거한 R_1^2과 같다.

이제 다음의 (2.3)와 같은 절편항을 갖지 않는 단순 선형회귀모형을 생각해 보자.

$$y_i = \beta_i\ x_i+\varepsilon_i \qquad (2.3)$$

설명변수 X가 없다면, 즉 β_1이 0이라면, 모든 i번째 적합값 $\hat{y}_i$은 0이 될것이며, 이때 오차제곱합은 $\Sigma\ y_i^2$이 된다. Kvalseth(1985)는 설명변수가 없는 경우에도 모든 i번째 적합값이 $\bar{y}$가 된다고 주장하였는데 이것은 모형(2.3)에서 설명변수 x_i들이 모두 1의 값을 갖는 특수한 경우의 적합값이다. 그리고 β_1에 어떤 제한을 두지 않은 경우에는 $\hat{y}_i$ 은 $\hat{\beta}_1\ x_i$ 가 된다(여기서 $\hat{\beta}_1$은 β_1의 최소제곱 추정치다). 그리고 이 때의 오차제곱합은 $\Sigma(y_i-\hat{y}_i)^2$이 되므로 결정계수는

$$R_0^2 = 1-\Sigma(y_i-\hat{y}_i)^2/\Sigma\ y_i^2 = \Sigma\ \hat{y}_i^2/\Sigma\ y_i^2 \qquad (2.4)$$

으로 정의하는 것이 합리적이며 이것은 Kvalseth(1985)가 열거한 결정계수 R_7^2, R_8^2과 같다. Kvalseth(1985)는 (2.3)의 모형에서도 (2.2)가 결정계수로서 타당하다고 주장하였으나 이 경우에 (2.2)는 0과 1 사이의 범위를 벗어날 뿐만 아니라 그 기본 아이디어도 이해하기 어렵다.

다음으로는 (2.5)과 같은 가중 선형회귀모형을 생각해 보자.

$$y_i = \beta_0+\beta_1\ x_i+\varepsilon_i. \qquad (2.5)$$

여기서 ε_i는 평균이 0이고 분산이 σ^2/w_i인 정규분포를 따르며 가중치 w_i는 알려져 있다고 가정한다. (2.5)의 양변에 $w_i^{1/2}$를 곱하면 아래와 같이 된다.

$$y_i^* = \beta_0\ w_i^{1/2}+\beta_1\ x_i^*+\varepsilon_i^* \qquad (2.6)$$

여기서 $y_i^*=w_i^{1/2}$, $x_i^*=w_i^{1/2}x_i$, $\varepsilon_i^*=w_i^{1/2}\varepsilon_i$이며 이때 ε_i^*의 평균과 분산은 각각 0과 σ^2이 된다. (2.5)가 (2.6)과 같이 재표현되므로, Kvalseth(1985)의 관점에서 본 가중선형회귀모형에서의 결정계수는 다음과 같이 된다.

$$R_{WLS}^2 = 1-\Sigma(y_i^*-\hat{y}_i^*)^2/\Sigma(y_i^*-\bar{y}^*)^2$$

여기서 $\hat{y}_i^*=\hat{\beta}_{0w}\ w_i^{1/2}+\hat{\beta}_{1w}x_i^*$이고 $\hat{\beta}_{0w}$과 $\hat{\beta}_{1w}$은 가중최소제곱 추정치다. 그러나 (2.6)에서 β_0 다음의 설명항이 1이 아니라 $w_i^{1/2}$이기 때문에 β_1을 0으로 두었을 때의 y_i^*의 적합값 $\hat{y}_i^*$은 $\bar{y}^*$가 아니라 $w_i^{1/2}(\Sigma w_i\cdot y_i)/\Sigma\ w_i$와 같이 되어야 한다. 한편 Willet and Singer(1988)는 가중선형회귀모형에서의 결정계수로

$$\text{pseudo } R_{WLS}^2 = 1-\Sigma(y_i-\hat{y}_{iw})^2/\Sigma(y_i-\bar{y})^2$$

을 제안하였는데 여기서 $\hat{y}_{iw}=\hat{\beta}_{0w}+\hat{\beta}_{1w}\ x_i$이고 $\hat{\beta}_{0w}$과 $\hat{\beta}_{1w}$은 가중최소제곱 추정치이다. 그런데 (2.5)에서 y_i의 분산이 σ^2/w_i이므로 β_1을 0으로 두었을 때의 y_i의 적합값 $\hat{y}_i$은 $\bar{y}$가 아니고 $\Sigma\ w_i\ y_i/\Sigma\ w_i$이 된다. 따라서 pseudo R_{WLS}^2 역시 가중선형회귀모형에 타당한 적합도의 측도로 생각하기 어렵다.

우리는 이제 앞에서와 일관되게 가중선형회귀모형에서의 결정계수를 정의해 보기로 하자. 모형

(2.5)에서 β_1이 0인 경우에는 모든 i번째 적합치 $\hat{y}_i$은 $\bar{y}_w = \Sigma\, w_i\, y_i / \Sigma\, w_i$이고 이때 가중오차 제곱합은 $\Sigma\, w_i(y_i - \bar{y}_w)^2$이 된다. 그리고 β_1에 어떤 제한도 두지 않을 때의 y_i에 대한 적합값은 $\hat{y}_i = \hat{\beta}_{0w} + \hat{\beta}_{1w}\, x_i$ (여기서 $\hat{\beta}_{0w}$과 $\hat{\beta}_{1w}$은 가중최소제곱 추정치이다)이고 가중오차제곱합은 $\Sigma\, w_i(y_i - \hat{y}_{iw})^2$이므로 가중 선형회귀모형에서의 결정계수는

$$R_w^2 = 1 - \Sigma\, w_i(y_i - \hat{y}_{iw})^2 / \Sigma\, w_i(y_i - \bar{y}_w)^2 \qquad (2.7)$$

과 같이 정의할 수 있다.

이제까지의 모든 논의에서는 편의상 설명변수가 하나 뿐인 선형회귀모형으로 국한되어 있었으나 (2.2)와 (2.4) 그리고 (2.7)의 결정계수 정의에 관한 한 설명변수의 갯수가 아무 문제가 되지 않음은 분명하다.

로지스틱 회귀모형에서의 결정계수로 Magee(1990)는 우도함수를 이용한 결정계수 $R^2_{LR} = 1 - \exp(-LR/n)$를 제안하고 정규분포를 가정한 표준적인 회귀모형에서 뿐만아니라 일반적인 모형에서도 유용한 적합측도가 될 수 있다고 주장하였다(여기서 $LR = 2(L_p - L_0)$이며 L_0은 절편항만을 갖는 모형의 최대로그우도함수이고 L_p는 상수항과 p개의 설명변수를 갖는 모형의 최대로그우도함수이다). 그러나 R^2_{LR}은 모형에 따라서는 1의 값을 갖지 못하며 이 통계량이 갖는 의미가 분명하지 않다.

한편 David and Lemeshow(1989)는 로지스틱 회귀모형의 결정계수로서 $R_L^2 = (L_0 - L_p)/(L_0 - L_s)$을 제안하였다(여기서 L_0는 절평항만 있는 로지스틱 모형의 최대로그우도함수이고, L_p는 절편항과 p개의 설명변수가 있는 모형의 최대로그우도함수이며, L_s는 포화모형의 최대로그우도함수이다). 그런데 David and Lemeshow(1989)는 R_L^2이 모형 적합의 측도이기 보다는 p개의 설명변수에 대한 우도비 검정의 다른 표현에 불과하다고 하였는데, 이것은 R_L^2과 Kvalseth(1985)가 열거한 기존의 결정계수들간의 관계를 적절히 규명하지 못했기 때문인 것으로 생각된다. 또한 로지스틱 회귀모형에서의 R_L^2은 3절에서 일반화 선형모형의 관점에서 정의하게 될 우리의 R_u^2과 동치적 관계가 성립하게 된다.

지금까지는 주로 선형회귀모형에서 적합도의 측도로서의 결정계수에 대하여 고찰해 보았으나 좀 더 시야를 넓혀 3절에서는 로그 선형모형의 등의 일반화 선형모형(generalized linear model)에서의 적합도의 측도를 설정하는 문제를 다루기로 한다. 이 일반적인 "결정계수"는 절편항을 갖는 선형회귀모형에서의 R_l^2을 포괄하고 있으며 지금까지 다룬 특수한 모형인 절편항이 없는 경우와 가중치가 있는 경우까지 일관된 형식으로 포괄할 수 있기 때문에 우리는 이 적합도 계수를 "통합결정계수(unified coefficient of determination)"로 부르기로 한다.

3. 尤度거리를 이용한 통합결정계수

Cook and Weisberg(1982)는 회귀진단의 맥락에서 尤度거리(likelihood distance)라는 용어를 쓴 바 있다. 우리는 여기서 이 용어를 다음과 같이 정의하기로 한다.

〈정의 1〉 모형 $\mathcal{M}_t$가 모형 $\mathcal{M}_s$의 특수한 경우라고 하자. 이러한 위치적 관계를 $\mathcal{M}_t \subset \mathcal{M}_s$로 나

타내기로 한다.

이 때 모형 $\mathcal{M}_t$와 $\mathcal{M}_s$의 尤度거리를

$$LD(\mathcal{M}_t, \mathcal{M}_s) = \sup_{\theta} 2\ L(\theta\ ;\ \mathcal{M}_s) - \sup_{\theta} 2\ L(\theta\ ;\ \mathcal{M}_t)$$

로 정의한다. 여기서 $L(\theta\ ;\ \mathcal{M})$는 모형 $\mathcal{M}$에서의 로그尤度함수를 말한다.

이제 尤度거리를 이용하여 통합결정계수 R_U^2을 다음과 같이 정의하기로 한다.

〈정의 2〉 현재모형(current model) $\mathcal{M}_c$의 零모형(null model) $\mathcal{M}_0$와 또 이것의 포화모형(saturated model) $\mathcal{M}_1$에 견주어 현재모형 $\mathcal{M}_c$의 통합결정 계수를

$$\begin{aligned} R_U^2 &= 1 - LD(\mathcal{M}_c, \mathcal{M}_1) / LD(\mathcal{M}_o, \mathcal{M}_1) \\ &= LD(\mathcal{M}_o, \mathcal{M}_c) / LD(\mathcal{M}_0, \mathcal{M}_1) \end{aligned}$$

로 정의한다. 여기에서 고려되는 세 모형은 $\mathcal{M}_0 \subset \mathcal{M}_c \subset \mathcal{M}_1$의 관계에 있음을 유의하자.

결정계수의 정의가 가장 확고하게 정립되어 있는 절편항과 설명변수 $x_1, x_1, x_2, \cdots x_p$를 갖는 선형 회귀모형의 경우에서 통합결정계수를 구하여 보자. 이 경우에 모형 (3.1)

$$y_i = \mu_i + \varepsilon_i \tag{3.1}$$

에서 ε_i가 평균 0과 알려진 분산 σ^2인 정규분포를 따른다고 하면 零모형, 현재모형, 포화모형이 각기

$$\begin{aligned} &\mathcal{M}_0 : \ \mu_i = \mu \\ &\mathcal{M}_c : \ \mu_i = \beta_0 + \beta_1\ x_{1i} + \beta_2\ x_{2i} + \cdots + \beta_p\ x_{pi} \\ &\mathcal{M}_1 : \ \mu_i \text{에 아무 제한이 없음} \end{aligned}$$

이 된다. 이 경우는 별로 어렵지 않게 $R_U^2 = R_1^2 = R_2^2$임을 알 수 있다. 그리고 가중치가 있는 경우에는 R_U^2은 2절에서 정의된 R_W^2임을 확인할 수 있다.

절편항을 포함하지 않는 선형 회귀모형인 경우에는 (3.1)의 모형하의 零모형, 현재모형, 포화모형이 각기

$$\begin{aligned} &\mathcal{M}_0 : \ \mu_i = 0 \\ &\mathcal{M}_c : \ \mu_i = \beta_1\ x_{1i} + \beta_2\ x_{2i} + \cdots \beta_p\ x_{pi} \\ &\mathcal{M}_1 : \ \mu_i \text{에 아무 제한이 없음} \end{aligned}$$

이 된다. 이 경우에도 별로 어렵지 않게 R_U^2은 2절에서 정의한 R_0^2임을 알 수 있다.

통합결정계수가 항상 0과 1 사이의 범주에 놓인다는 것은 확실하다. 왜냐하면

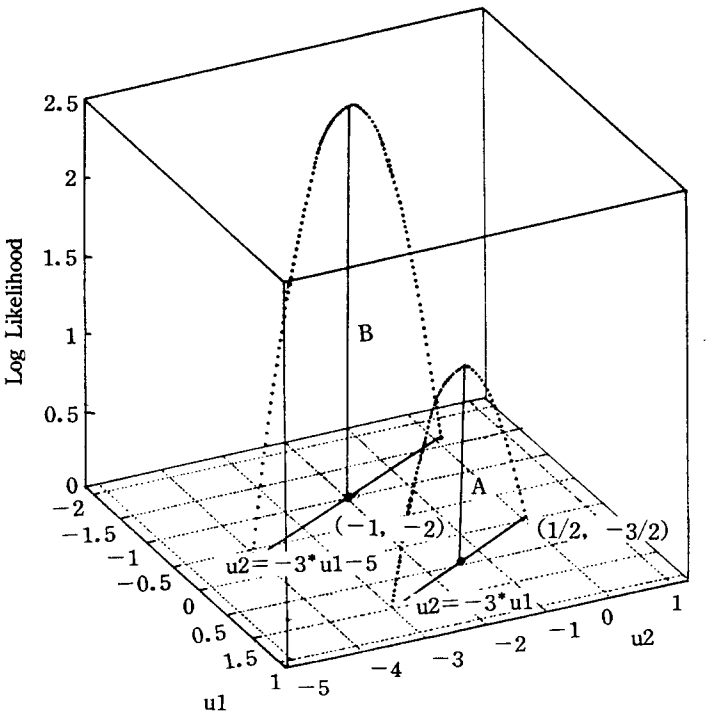
$$LD(\mathcal{M}_0, \mathcal{M}_1) = LD(\mathcal{M}_0, \mathcal{M}_c) + LD(\mathcal{M}_c, \mathcal{M}_1) \tag{3.2}$$

이며, $LD(\mathcal{M}_0, \mathcal{M}_c)$와 $LD(\mathcal{M}_c, \mathcal{M}_1)$ 모두 비음이기 때문이다. 따라서 통합결정 계수가 의미하는 것은 〈정의 2〉에서와 (3.2)에서 볼 수 있는 그대로이다. 즉, 零모형과 포화모형 사이의 尤度거리를

1로 놓았을 때, 현재모형이 얼마만큼에 위치하고 있는가를 말해주고 있다. 이와 같은 통합결정계수의 의미를 가상의 데이타 {(x, y) : (1, −1), (−3, −2)}와 다음과 같이 설정된 모형 (3.3)을 통하여 살펴보기로 하자.

$$
\begin{aligned}
& y_i = \mu_i + \varepsilon_i \\
& \mathcal{M}_0 : \ \mu_i = 0, \\
& \mathcal{M}_c : \ \mu_i = \beta_1 x_i, \ i = 1, 2 \\
& \mathcal{M}_1 : \ \mu_i \text{에 아무 제한이 없음}
\end{aligned}
\tag{3.3}
$$

여기서 ε_i는 N(0, 1)을 따르는 확률변수라고 가정한다. 이 예에서 零모형의 최대로그우도함수값은 $\sup_\theta L(\theta ; \mathcal{M}_0) = \log 2\pi - 5/2$이고, 현재모형의 최대로그우도함수값은 $\mu_1 = 1/2$, $\mu_2 = -3/2$일때 $\sup_\theta L(\theta ; \mathcal{M}_c) = \log 2\pi - 5/4$가 되며 포화모형의 최대로그우도함수값은 $\mu_1 = -1$, $\mu_2 = -2$ 일때 $\sup_\theta L(\theta ; \mathcal{M}_s) = \log 2\pi$가 된다. 따라서 이때의 통합결정 계수는 위의 〈정의 2〉에 의해 $R_u^2 = (5/4)/(5/2) = 1/2$이 된다. 〈그림 3−1〉은 각 우도함수를 零모형의 최대로그우도함수값으로 감하여 그래프로 나타낸 것인데 여기서 통합결정 계수는 A/B로 나타내진다.



〈그림 3.1〉 가상 데이타의 로그우도함수

그런데 회귀모형에서 기존에 정의된 수정결정계수(adjusted coefficient of determination)와 부분결정계수(partial coefficient of determination ; Helland, 1987)도 尤度거리를 이용하여 일반화 선형모형에서 각각 다음과 같이 정의할 수 있다.

〈정의 3〉 현재모형(current model) $\mathcal{M}_c$의 零모형(null model) $\mathcal{M}_0$와 또 이것의 포화모형 $\mathcal{M}_1$에 견주어 현재모형 $\mathcal{M}_c$의 통합수정결정계수(unified adjusted coefficient of determination)를

$$R_{U.adj}^2 = 1-(df_1/df_2)\times LD(\mathcal{M}_c, \mathcal{M}_1)/LD(\mathcal{M}_0, \mathcal{M}_1)$$

로 정의한다. 여기에서 df_1='모형 $\mathcal{M}_1$의 모수의 수－모형 $\mathcal{M}_0$의 모수의 수'이고 df_2='모형 $\mathcal{M}_1$의 모수의 수－모형 $\mathcal{M}_c$의 모수의 수'이다.

〈정의 4〉 처음의 현재모형을 $\mathcal{M}_{c1}$, 여기에 추정해야 할 모수가 추가된 두번째 현재모형을 $\mathcal{M}_{c2}$, 포화모형을 $\mathcal{M}_1$이라고 할 때 모형 $\mathcal{M}_{c1}$에 견주어 모형 $\mathcal{M}_{c2}$의 통합부분결정계수(unified partial coefficient of determination)를

$$R_{U2.1}^2 = LD(\mathcal{M}_{c1}, \mathcal{M}_{c2})/LD(\mathcal{M}_{c1}, \mathcal{M}_1)$$

로 정의한다. 여기에서 $\mathcal{M}_{c1}=\mathcal{M}_0$ 일때는 $R_{U2.1}^2=R_U^2$의 관계가 성립한다.

이와같이 정의된 통합결정계수들은 선형 회귀모형 이외에도 비선형 회귀모형 그리고 로지스틱 회귀모형과 로그 선형모형같은 일반화 선형모형 등에 적용될 수 있겠다. 이에 대한 개별사례는 다음의 4절에서 보이기로 한다.

4. 개별 적용사례

여기에서는 비선형 회귀모형, 로지스틱 회귀모형 그리고 로그 선형모형의 실제 사례분석을 통하여 앞에서 정의한 통합결정계수를 구해 보기로 하겠다.

4.1 비선형 회귀모형

자본과 노동의 대체생산고정탄력성(constant elasticity of substitution production ; CES) 함수의 적합을 위한 데이타(SAS Institute, 1985, p.591)를 이용하여 설정된 모형(4.1)의 통합결정계수를 구해보기로, 하자.

$$\log Q = \beta_0+\beta_1 \log(\beta_2 L^{\beta_3}+(1-\beta_2)K^{\beta_3})+\varepsilon \qquad (4.1)$$

여기서 Q는 생산량, K는 투입된 자본, L은 투입된 노동을 나타내며 ε는 $N(0, \sigma^2)$을 따르는 확률변수이다. 먼저 Gauss Newton 방법에 의해 위의 모형을 적합시키면 다음과 같이 된다.

$$\log Q = 0.12-0.34 \log(0.34L^{-3.01}+0.66K^{-3.01})$$

다음으로 (4.1)과 같은 절편항이 있는 비선형 회귀모형에서는 통합결정계수가 $R_U^2=R_1^2$임을 쉽게 확인할 수 있고 위의 예에서 이것을 구해보면 $R_U^2=0.971$이 된다.

4.2 로지스틱 회귀모형

백혈구의 수와 AG의 반응에 따른 백혈병 환자의 생존에 관한 데이타(Cook and Weisberg, 1982, p.193)를 이용하여 로지스틱 회귀모형의 통합결정계수를 구하여 보자. 이 데이타의 적합을 위한 다음의 로지스틱 회귀모형을 생각해 보자.

$$\log(p_i/(1-p_i)) = \beta_0+\beta_1 \log(\mathrm{WBC})+\beta_2\ \mathrm{AG} \qquad (4.2)$$

여기서 WBC는 백혈구의 수, AG는 반응이 양성이면 1, 음성이면 0인 변수, p_i는 i번째 범주에 속한 환자가 생존하게 될 확률이다. SAS의 PROC CATMOD를 수행하여 모형(4.2)의 모수를 최우추정법으로 구해 보면 다음과 같다.

$$\log(p_i/(1-p_i)) = 8.096-1.089 \log(\mathrm{WBC})+2.520\ \mathrm{AG}$$

또한 SAS의 PROC CATMOD 출력결과에서 현재모형의 최대로그우도함수값 $\sup_\theta L(\theta\ ;\ \mathcal{M}_c)=-13.416$과 완전모형의 최대로그우도함수값 $\sup_\theta L(\theta\ ;\ \mathcal{M}_1)=-1.911$을 구할 수 있다. 그런데 SAS의 PROC CATMOD 출력결과에서는 절편항이 없는 영모형의 최대로그우도함수값 만을 제공하고 있으므로 다음의 식

$$\sup_\theta L(\theta\ ;\ \mathcal{M}_0) = \Sigma\{y_i \log[(\Sigma\ y_i/\Sigma\ c_i)/(1-\Sigma\ y_i/\Sigma\ c_i)] - c_i \log[1/(1-\Sigma\ y_i/\Sigma\ c_i)\} \qquad (4.3)$$

(여기서 c_i는 i번째 범주에 속한 환자의 수이고 y_i는 그중에 생존한 환자의 수이다)에 의해 절편항이 있는 영모형의 최대로그우도함수값은 -21.005임을 알 수 있다. 이제 〈정의 2〉에 의해 로지스틱 회귀모형(4.3)의 통합결정계수 $R_u^2=0.398$를 구할 수 있다.

한편 Harrell(1986)의 PROC LOGIT에서는 로지스틱 회귀모형의 결정계수를 다음과 같이 정의하고 있다.

$$R^2 = [LD(\mathcal{M}_0, \mathcal{M}_c)-2p]/-2\sup_\theta L(\theta\ ;\ \mathcal{M}_0),\ \text{if } LD(\mathcal{M}_0, \mathcal{M}_c)\geq 2p$$
$$= 0,\ \text{if } LD(\mathcal{M}_0, \mathcal{M}_c)<2p$$

여기서 p는 '현재모형에서 절편항을 제외한 모수의 수'이다. 그러나 이 정의에서는 적합도가 낮은 모형들의 결정계수가 아무 구별없이 모두 0의 값을 가질 수 있으며 포화모형의 결정계수도 1의 값을 갖지 못하는 제한점을 갖고 있다.

4.3 로그 선형모형

2차 세계대전중 미국 공군 지원자들이 치룬 적성검사 결과와 10년 후에 추적조사를 통하여 조사된 당시 지원자들의 그 후 교육수준과 직업에 관한 NBER(National Bureau of Economic Research) 데이타(Fienberg, 1980, p.45)를 이용하여 로그 선형모형을 분석해 보기로 하자. 여기서 적성은 5가지 수준 그리고 교육과 직업은 4가지 수준으로 각각 분류되었다. 먼저 다음과 같은 위계적 모형(hierachical model)들을 생각해 보자.

$\mathcal{M}_{AEO}$: $\log y_{ijk} = \mu + u_{1(i)} + u_{2(j)} + u_{3(k)}$

$\mathcal{M}_{A|EO}$: $\log y_{ijk} = \mu + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}$

$\mathcal{M}_{A|OE}$: $\log y_{ijk} = \mu + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)}$

$\mathcal{M}_{E|OA}$: $\log y_{ijk} = \mu + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}$

$\mathcal{M}_{A|EA|O}$: $\log y_{ijk} = \mu + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)}$

$\mathcal{M}_{A|EE|O}$: $\log y_{ijk} = \mu + u_{1(j)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)}$

$\mathcal{M}_{E|OA|O}$: $\log y_{ijk} = \mu + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}$

$\mathcal{M}_{A|EA|OE|O}$: $\log y_{ijk} = \mu + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$

$\mathcal{M}_{A|E|O}$: $\log y_{ijk} = \mu + \mu_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$

여기서 y_{ijk}는 i번째 적성, j번째 교육, k번째 직업에 속한 사람의 수를 나타낸다. 위의 로그선형모형에서 주효과만을 고려한 모형 $\mathcal{M}_{AEO}$를 영모형으로 생각하기로 하자. 이제 SAS의 PROC CATMOD를 수행하여 위 모형들의 최대로그우도함수값과 통합결정계수, 그리고 통합수정결정계수를 구하여 보면 표 4.5와 같이 된다. 그런데 수정결정계수는 자유도에 의해 조정되기 때문에 음의 값을 가질 수도 있다.

〈표 4.5〉 로그 선형모형의 최대로그우도함수값, 통합결정계수 그리고 통합수정결정계수(ADD-CELL=10^{-4})

모 형	최대로그 우도함수	통합결정계수	통합수정 결정계수
$\mathcal{M}_{AEO}$	−16477.80	0.000	0.000
$\mathcal{M}_{A|EO}$	−16389.15	0.117	−0.068
$\mathcal{M}_{A|OE}$	−16459.10	0.028	−0.177
$\mathcal{M}_{E|OA}$	−15913.45	0.832	0.807
$\mathcal{M}_{A|EA|O}$	−16370.45	0.158	−0.291
$\mathcal{M}_{A|EE|O}$	−15824.75	0.963	0.946
$\mathcal{M}_{A|OE|O}$	−15894.75	0.859	0.799
$\mathcal{M}_{A|EA|O'E|O}$	−15811.85	0.982	0.964
$\mathcal{M}_{A|E|O}$	−15799.35	1.000	1.000

5. 맺 음 말

로버스트 회귀모형의 경우에도 Huber, Andrews, Tukey 등이 제안한 $\Sigma\ \rho[(y_i - x_i'\beta)/\sigma]$를 로그우도함수로 생각하여 앞에서와 마찬가지로 통합결정계수를 정의할 수 있을 것이다. 그러나

장애모수 σ를 미리 추정해야 하는 것과 $\rho[(y-x_i'\beta)/\sigma]$의 적절한 변형을 확률밀도함수로 간주하는 것 자체가 타당한가 하는 문제점이 있다. 따라서 가중회귀모형의 통합결정계수를 구하는 방법을 준용하여 가중치 w_i를

$$\Psi[(y_i-x_i'\hat{\beta}_{MC})/\hat{\sigma}]/[(y_i-x_i'\hat{\beta}_{Mc})/\hat{\sigma}]$$

로 하여(여기서 $\hat{\beta}_{Mc}$는 현재모형에서 구한 β의 로버스트 추정치이고 마찬가지로 $\hat{\sigma}$도 현재모형에서 구한 장애모수 σ의 추정치이다. 또한 $\Psi(t)=\rho'(t)$이다) 다음과 같이 로버스트 회귀모형에서의 통합결정계수를 정의할 수 있을 것이다.

$$R_{U.M}^2 = 1-\Sigma\, w_i(y_i-x_i'\hat{\beta}_{Mc})^2/\Sigma\, w_i(y_i-x_i'\hat{\beta}_{M0})^2$$

여기서 $\hat{\beta}_{Mc}$과 $\hat{\beta}_{M0}$는 각각 현재모형과 영모형에서 구한 β의 로버스트 추정치이다.

참 고 문 헌

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

[2] David, W. and Lemeshow, S.(1989), *Applied Logistic Regression*. John Wiley & Sons, N: New York.

[3] Fienberg, S.E.(1980), *The Analysis of Cross-Classified Categorical Data*. 2nd Edition, MIT Press, MA: Cambridge.

[4] Harrell, F.E.(1986), "The LOGIST Procedure," *SAS Supplemental Library User's Guide*. Version 5 Edition. Cary, NC: SAS Institute, Inc.

[5] Helland, I.S.(1987), "On the Interpretation and Use of R^2 in Regresion Analysis," *Biometrics*, **43**, 61-69.

[6] Kvalseth, T.O.(1985), "Cautionary Note about R^2," *The American Statistician*, **39**, 279-285.

[7] Magee, L.(1990), "R^2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests," *The American Statistican*, **44**, 250-253.

[8] SAS Institute(1985), *SAS User's Guide: Statistics*, Version 5 Edition. Cary, NC: SAS Institute Inc.

[9] Uyar, B. and Erdem, O.(1990). "Regression Procedures in SAS: Problems," *The American Statistican*, **44**, 296-301.

[10] Willett, J.B. and Singer, J.D.(1988). "Another Cautionary Note about R^2: Its Use in Weighted Least-Squares Regression Analysis," *The American Statistician*, **42**, 236-238.

Unified Approach to Coefficient of Determination R^2 Using Likelihood Distance

Myung-Hoe Huh*, Jong-Han Lee* and Jin-Whan Jung*

ABSTRACT

Coefficient of determination R^2 is most frequently used descriptive measure in practical use of linear regression analysis. But there have been controversies on defining this measure in the cases of linear regression without the intercept, weighted linear regression and robust linear regression. Several authors such as Kvalseth(1985) and Willet and Singer(1988) proposed many variations of R^2 to meet the situations. However, theire measures are not satisfactory due to the lack of a universal principle.

In this study, we propose a unfied approach to defining the coefficient of determination R^2 using the concept of likelihood distance. This new measure is in good accordance with typical R^2 in linear regression and, moreover, can be applied to nonlinear regression models and generalized linear models such as logit and log−linear models.

* Department of Statistics, Korea University, Anamdong Seongbukgu Seoul 136−701, Korea.