Study on R^2 for no-intercept Model

Jong-Doo Do[1], Gyu-Moon Song[2], Tae-Yoon Kim[3]

Abstract

There have been some controversies on the use of the coefficient of determination for linear no-intercept model. One definition of the coefficient of determination, $R^2 = \sum \hat{y}^2 / \sum y^2$, is being widely accepted only for linear no-intercept models though Kvalseth(1985) demonstrated some possible pitfalls in using such R^2. Main objective of this article is to provide a cautionary notice for use of the R^2 by pointing out its tricky aspects by means of empirical simulations.

Keywords : Coefficient of Determination, Empirical Simulations, no-intercept models.

1. 서 론

아래와 같은 식 (1)의 간단한 선형관계를 가정하자.

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + \varepsilon_i , \quad i = 1, \cdots, n \tag{1}$$

여기에서 ε_i는 iid인 오차항이고, $\beta_o, \cdots, \beta_k$는 회귀모수들이다. 자료 분석에 있어서 식 (1)을 사용할 때 결정계수 R^2은 적합도로서 가장 널리 사용되는 측도의 하나이며 R^2에는 다음과 같은 다양한 종류들이 있다:

$$R_1^2 = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2}$$

$$R_2^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3}$$

$$R_3^2 = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n} y_i^2 \tag{4}$$

$$R_4^2 = \sum_{i=1}^{n} \hat{y}_i^2 / \sum_{i=1}^{n} y_i^2 \tag{5}$$

여기에서 $\hat{y}_i$ 은 $x_{i1}, \cdots, x_{ik}$ 자료에 의한 y의 적합치이며 $\bar{y}$ 는 자료 y_i들의 표본

---

1) First Author : Invitation Full-time Instructor, Department of Statistics, Keimyung University, Taegu, 704-701, Korea.
E-mail : jddo@kmu.ac.kr
2) Professor, Department of Statistics, Keimyung University, Taegu, 704-701, Korea.
3) Professor, Department of Statistics, Keimyung University, Taegu 704-701, Korea.

평균이다.

절편항 (식 (1)에서 $\beta_0 \neq 0$) 이 있는 선형모형의 R^2 선택은 간단한 문제로써 R_1^2 이나 R_2^2을 결정계수로 사용하며 이 경우 식 (6)과 같은 등식관계가 성립하는 것으로 알려져 있다.

$$R_1^2 = R_2^2 \tag{6}$$

하지만 $\beta_0 = 0$(절편항이 없는 모형)에서 R_2^2이 1을 초과한다든지 또는 R_1^2이 음수인 경우가 관찰되기도 하며, 잘 알려진 식 (6)의 등식이 성립되지 않기도 하기 때문에 R^2사용에 관하여 논란이 있었다. 이에 따라 R_3^2과 R_4^2은 절편항이 없는 모형에서 주로 사용되어 지고 있으며 이 경우 $R^2_3 = R^2_4$인 것을 쉽게 확인할 수 있다 (Uyar and Erderm (1990) 참조). 여기서 주목할 점은 R_1^2또는 R_2^2이 $\bar{y}$ 주위에 있는 y의 총변동에 대한 회귀변동의 비율인 반면 R_3^2과 R_4^2은 0 주위에 있는 y의 총 변동 비율에 대한 회귀 변동의 비율이라는 사실이다.

Kvalseth (1985)는 절편항이 있는 모형과 없는 모형 사이에서 모형 선택을 할 때 절편항이 없는 선형모형에서 R_3^2 (또한 R_4^2)를 배타적으로 사용하는 것을 경고하고 있다. 즉 그는 적절한 모형 선택을 위해서는 비교되는 모형 간에 동일한 R^2을 사용해야 한다고 주장하며 절편항이 없는 선형모형에 대한 R^2_3의 배타적 사용이 가져올 수 있는 문제를 간단한 예를 통해 지적하고 있다. 결과적으로 그는 절편항이 없는 선형모형에 조차 R_1^2을 일관성 있게 사용할 것을 권하였다. 그렇지만 그의 이러한 관점은 그가 만든 예가 다소 가상적이고 설득력이 부족한 까닭으로 일반 통계 분석자들의 많은 관심을 얻지 못하고 있는 것으로 보인다 (2절의 논의 참조). 본 논문에서는 R_3^2 또는 R_4^2의 문제점을 좀 더 설득력 있는 새로운 예를 통해 지적함으로써 Kvalseth가 주장하고 있는 관점을 재조명하고자 한다. 이 논문은 다음과 같이 구성된다. 2절은 시뮬레이션 예들을 통해 R_3^2 사용상의 문제점들을 논의하며 3절은 결론을 다룬다.

2. 시뮬레이션 예와 R_3^2의 문제점

먼저 우리는 Kvalseth (1985)가 만든 가상의 예제를 다루어 "적합한 모형 선택을 위해 비교되는 모형 간에는 동일한 R^2을 사용해야한다."는 그의 주장을 살펴보기로 한다. 그의 자료는 (x, y) = (1, 15), (2, 37), (3, 52), (4, 59), (5, 83), (6, 92)로 구성되어 있는데, 이것들은 [그림 1]에 나타나있다. [그림 2]와 [그림 3]에는 절편항이 있는 것과 없는 상황 각각의 선형모형을 위한 추정 선을 나타내었으며 <표 1>은 두 모형을 사용하였을 때 요약 통계이다.

[표 1] 절편항이 있는 것과 없는 것의 선형모형에서 Kvalseth 자료에 대한 모수 추정치와 관련 통계량.

Parameters Statictics	$\hat{y} = b_0 + b_1 x$	$\hat{y} = b_1 x$
b_0	3.3333	-
b_1	15.1429	15.9121
R_1^2	0.9808	0.9777
R_2^2	0.9808	1.0836
R_3^2	-	0.9961
R_4^2	-	0.9961
MSE	19.6191	18.2593

[표 1]에서의 마지막 통계량은 오차평균제곱(mean squared error : MSE)이며 아래 식 (7)로 정의된다.

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{(n-k)} \tag{7}$$

이 표로부터 Kvalseth는 다음과 같은 사실에 주목하였다.

(i) R_2^2 은 $\beta_0 = 0$ 일 때 1을 초과하기 때문에 부적절하다.

(ii) R_1^2과 R_3^2(0.9808 < 0.9961)에 각각 근거하여 절편항이 없는 모형과 절편항이 있는 모형을 선택한다면 절편항이 없는 모형을 선택할 수 있을 것이다. 반면에, R_1^2의 일관된 사용은 절편항이 있는 모형을 선택하게 된다(즉, 0.9808 > 0.9777).

이외의 다른 유사한 예제(Kvalseth (1985)의 다변량 데이터 경우)를 통해, 그는 비교 모형 간에서 적절한 모형 선택을 하기 위해 R_1^2을 일관적으로 사용할 것을 추천하였다. 그러나 Kvalseth에 의해 제시된 위의 간단한 예는 절편항이 없는 모형에 대해 R_3^2(그리고 R_4^2)만을 고집하여 사용하고 있는 자료 분석자들에게 그 위험성을 충분히 이해시키지 못했다. 무엇보다도 그의 예제에서는 절편항이 있는 것과 절편항이 없는 모형간의 선택이 그다지 중요한 문제인 것처럼 보이지 않는다. 실제로 [그림 1], [그림 2], [그림 3]을 통해 쉽게 확인할 수 있는 것은 절편항이 없는 모형과 있는 모형 둘 다 주어진 자료에 대해 적절한 모형으로 보인다는 점이며 이는 [표 1]에서 기울기(b1)의 추정치를 포함한 요약 통계에서도 명확히 드러나고 있다. 따라서 그의 예제는 절편이 없는 모형에서 R_3^2을 배타적으로 사용했을 때 생기는 문제들을 효과적으로 보여주고 있지 못하다.

본 논문에서 R_3^2 의 배타적 사용의 문제점을 효과적으로 설명하기 위해 다음과 같

은 예를 고려한다. $i=1,\cdots,n$에 대해

$$y_i=\mu+\varepsilon_i \tag{8}$$

이다. 여기서 ε_i는 $iid\ N(0,\sigma^2)$인 오차항이며, μ는 상수이다.

식 (8)은 R_1^2 (또는 R_2^2)이 가정된 모형의 적합성을 말해 주기 위해 사용되는 것과 어긋나는 모형이다. 즉, R_1^2 (또는 R_2^2)은 $\overline{y}$ 주위에 있는 y의 총 변동에 대한 $\overline{y}$ 주위에 있는 $\hat{y}$의 변동의 비율로 제공(회귀에 의해 설명되는 변동의 비율)되는 것으로서, 자료가 식 (8)에 의해서 이루어진다면 대략적으로 $\hat{y}=\overline{y}$이기 때문에 R_1^2(또는 R_2^2)은 0에 가깝다는 사실을 알 수 있다. 또한 R_3^2(또는 R_4^2)은 0 주위에 있는 y의 총 변동에 대한 0 주위에 있는 $\hat{y}$의 총 변동의 비율로 제공되는 것으로서, 식 (8)의 $\mu\neq 0$일 경우 회귀모형의 적합성을 나타낸다면 $\overline{y}$는 0과 상당히 차이를 나타낼 것이며, 절편항이 없는 모형에서 적합성만을 확인하는데 쓰여 질 경우 R_3^2은 쉽게 분석자들을 현혹시킬 것이다.

이 문제를 알아보기 위해 시뮬레이션 예제를 통하여 나타내면 다음과 같다. 식 (8)에서 $\mu=0,1,3$이고, σ^2=1 일 때 보면, 먼저 $\mu=0$ 인 경우의 자료는 [그림 4]와 같으며 각각의 모형(절편을 가진 모형과 절편이 없는 모형)을 위한 적합선은 [그림 5]와 [그림 6]과 같다. 또한 두 모형의 모수 추정치와 요약 통계량은 [표 2]에서 나타난 것과 같이 모든 R_i^2들의 값은 0에 가깝다는 것을 알 수 있으며, 이는 두 모형이 다 부적절함을 나타내고 있음을 알 수 있다.

[표 2] $\mu=0$인 다른 모형을 위한 대체 통계량의 모수 추청치

Parameters Statictics	$\hat{y}=b_0+b_1x$	$\hat{y}=b_1x$
b_0	-0.4071	-
b_1	0.01739	0.00267
R_1^2	0.02834	0.0016
R_2^2	0.02834	0.00711
R_3^2	-	0.00278
R_4^2	-	0.00278
MSE	1.52645	1.52923

이와 같이, $\mu=1$일 때 [표 3] 및 [그림 7, 8, 9]와 $\mu=3$일 때 [표 4] 및 [그림 10, 11, 12]의 결과이며, $\mu=1$, 3에서 절편항이 있는 모형에 대해서 R_1^2이 0에 가까운 반면 절편항이 없는 모형에 대해서는 R_3^2은 0과 상당한 차이를 나타냄을 알 수 있다. 따라서 R_3^2을 독점적으로 사용할 때 절편항이 있는 모형보다 절편항이 없는 모형을

선호할 것이다. 그러나 이러한 경우 $y=0+\varepsilon$ 모형과 $y=\beta_1 x+\varepsilon$모형에서 상대적인 R_3^2 값의 비교밖에 할 수 없을 것이다. 즉, R_3^2은 제로모형($y=0+\varepsilon$)이 제거($\overline{y}$가 0과 차이가 있는 경우)된 경우와 제거되지 않은 모형간의 뚜렷한 비교를 위해 제공된 자료로서, 제로모형 일 때 보다 상대적으로 높다고 해서 절편항이 없는 모형에서 적합성만을 확인하는데 사용할 경우 분석자들을 속일 수 있다는 것을 보여주는 것이다.

또한 관찰되어진 예제에서 나타났듯이 절편항이 없는 모형에 대한 R_1^2이 사용될 때 $\overline{y}$가 0에서 상당히 떨어져 있을 때 R_1^2이 음수이거나 또는 1을 초과하게 되는 문제가 있음을 알 수 있었다. R_1^2과 R_3^2이 가진 이 문제는 0에서부터 μ의 증가할 때 나타난 결과로서 [표 3]과 [표 4]에서 알 수 있었으며, 절편항이 없는 모형에서 R_1^2의 사용이나, 0에서 상당히 떨어진 자료에 대해서 R_3^2을 사용하는 것은 좋지 않음을 알 수 있다.

[표 3] $\mu=1$인 다른 모형을 위한 대체 통계량의 모수 추청치

Parameters Statictics	$\hat{y}=b_0+b_1x$	$\hat{y}=b_1x$
b_0	0.5205	-
b_1	0.0159	0.0347
R_1^2	0.0399	-0.0336
R_2^2	0.0399	0.2077
R_3^2	-	0.4328
R_4^2	-	0.4328
MSE	0.8969	0.9414

[표 4] $\mu=3$인 다른 모형을 위한 대체 통계량의 모수 추청치

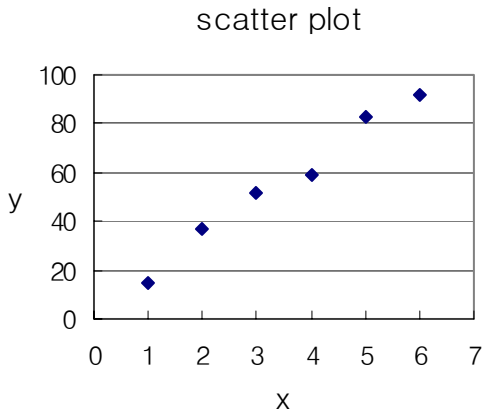
Parameters Statistics	$\hat{y}=b_0+b_1x$	$\hat{y}=b_1x$
b_0	2.63755	-
b_1	0.01435	0.10968
R_1^2	0.0253	-1.4490
R_2^2	0.0253	1.83695
R_3^2	-	0.71514
R_4^2	-	0.71514
MSE	1.16479	2.85390

3. 결 론

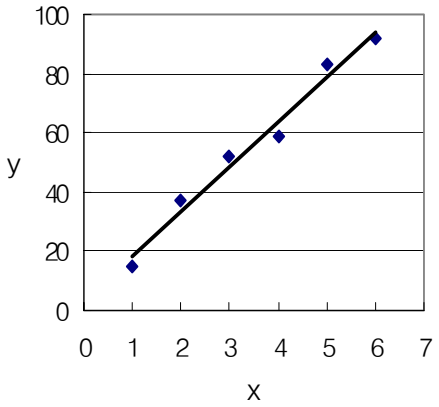
본 논문은 절편항이 없는 모형에 대한 결정계수의 사용상의 문제점에 관한 다루기 힘든 면을 논의하였으며, 시뮬레이션 연구를 통하여 R_1^2과 R_3^2의사용에 대한 있을지도 모를 함정을 보여주었다. 또한 절편항이 없는 모형에 대한 R_1^2의사용이나, 0에서 상당히 떨어진 자료에 대해서 R_3^2의사용을 제안하지 않으며, 특히 이 문제는 R_3^2(또는 R_4^2)이 0과 상당히 떨어진 자료를 사용할 때 집중되었음을 알 수 있다. 결론적으로 본 연구의 내용은 Kvalseth의 결론적 소견인 "절편항이 없는 모형에서는 이론적인 정당화와 경험적인 자료 분석 둘 다 적절하게 사용되어야만 한다." 그리고 "결정계수에 대한 유일한 신뢰는 중요한 데이터의 특성과 부적절한 모형을 드러내지 못하는 경우에만 사용되어져야 한다."라는 점을 강조한다.

참 고 문 헌

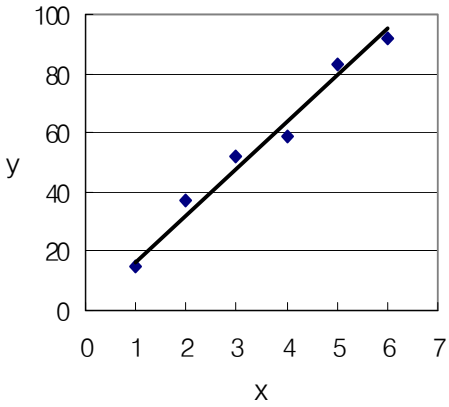
1. Kvalseth, T.O.(1985). Cautionary Note about R^2, *The American Statistician*, 39, 279-285.
2. Uyar, B. and Erdem, O. (1990). Regression Procedures in SAS: Problems?, *The American Statistician*, 44, 296-301.
3. Willet, J. B. and Singer, J. D. (1988). Another Cautionary Note about R^2: Its use in Weighted Least-Squares Regression Analysis, *The American Statistician*, 42, 236-238.


scatter plot
100
80
60
40
20
0
y
0 1 2 3 4 5 6 7
x

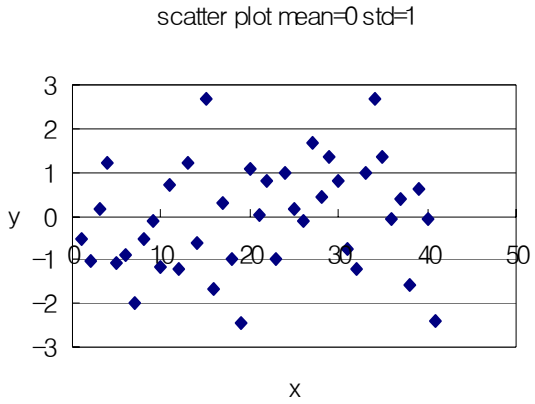
[그림 1]


100
80
60
40
20
0
y
0 1 2 3 4 5 6 7
x

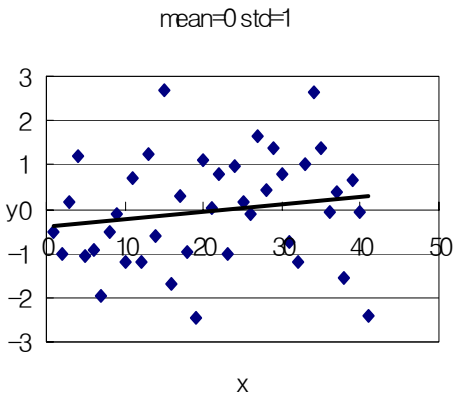
[그림 2]


100
80
60
40
20
0
y
0 1 2 3 4 5 6 7
x

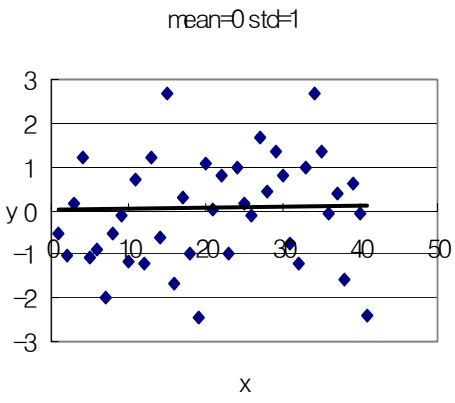
[그림 3]


scatter plot mean=0 std=1
3
2
1
y 0
−1
−2
−3
0
10
20
30
40
50
x

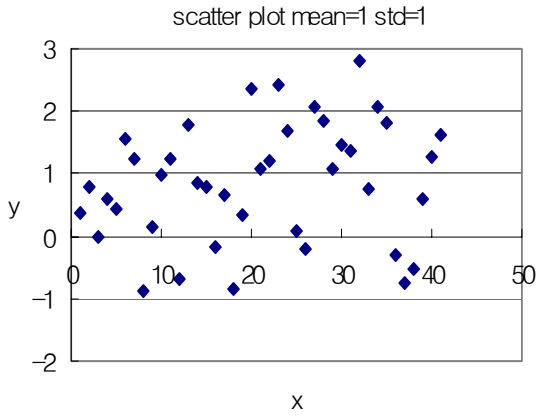
[그림 4]


mean=0 std=1
3
2
1
y0
−1
−2
−3
0
10
20
30
40
50
x

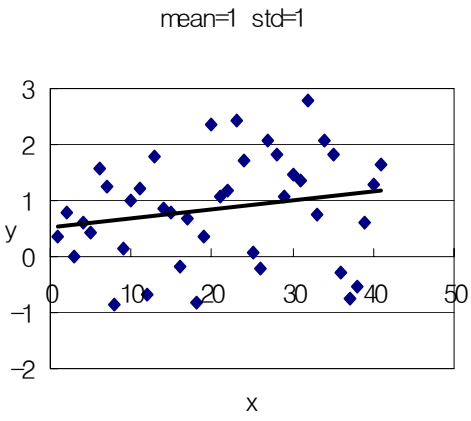
[그림 5]


mean=0 std=1
3
2
1
y 0
−1
−2
−3
0
10
20
30
40
50
x

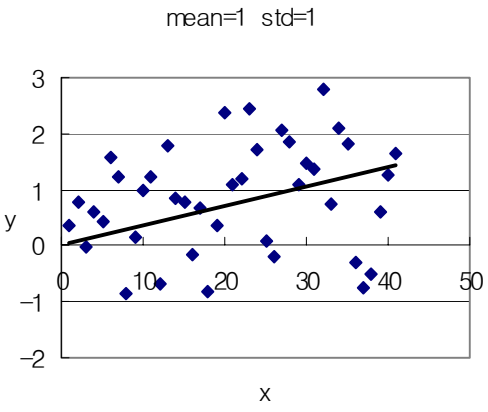
[그림 6]


scatter plot mean=1 std=1
3
2
1
y
0
−1
−2
0
10
20
30
40
50
x

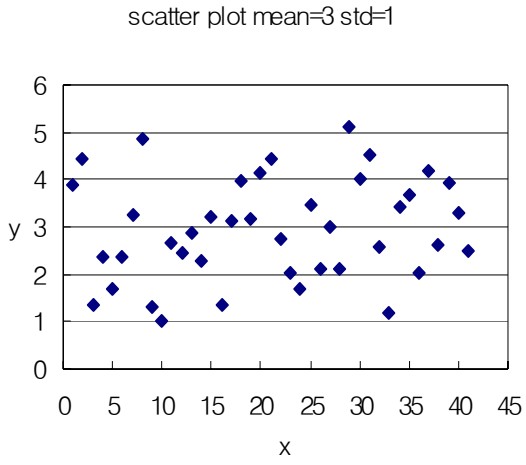
[그림 7]


mean=1 std=1
3
2
1
y
0
−1
−2
0
10
20
30
40
50
x

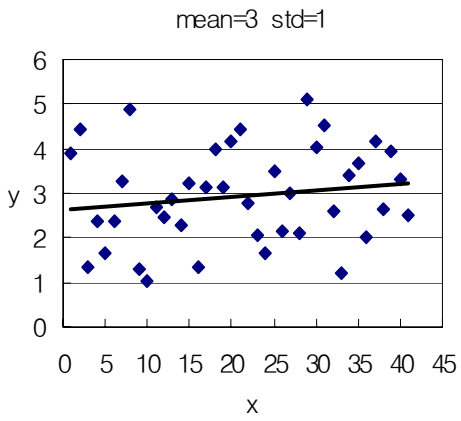
[그림 8]


mean=1 std=1
3
2
1
y
0
−1
−2
0
10
20
30
40
50
x

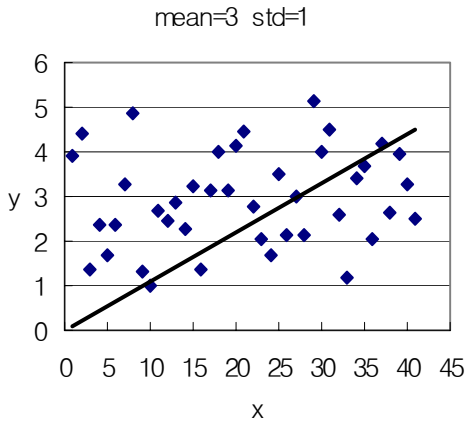
[그림 9]



[그림 10]



[그림 11]



[그림 12]