The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.2.227
27(2), 227–237

DD-Plot for ANCOVA Models

Dae-Heung Jang[a,1]

[a]Department of Statistics, Pukyong National University

(Received December 2, 2013; Revised March 13, 2014; Accepted April 05, 2014)

Abstract

We use the regression model with the indicator variables in the case that we use qualitative variables as some predictor variables in regression analysis. We use the ANCOVA(Analysis of Covariance) model when comparing the response variable among groups while statistically controlling for variation in the response variable caused by a variation in the covariate. *DD*-plot can be used as a graphical exploratory data analysis tool before the confirmatory data analysis. With the *DD*-plot, we can discriminate the difference of groups in the regression model with the indicator variables or the ANCOVA model at a glance. Making *DD*-plot does not demand the statistical model assumption about error terms in regression model. Several examples show the usefulness of *DD*-plots as a graphical exploratory data analysis tool for the regression analysis.

Keywords: Data depth, *DD*-plot, ANCOVA.

1. 서론

우리는 회귀분석에서 설명변수들 중 일부가 질적 변수인 경우 지시변수(indicator variable, 또는 가변수(dummy variable)라 함)를 사용한다. 지시변수 사용 회귀모형에서는 집단을 구분하여 주는 질적변수를 지시변수로 표현할 수 있다. 예로 성별을 구분하는 질적변수인 경우 남자를 0, 여자를 1로 하거나 남자를 1, 여자를 0로 하여 지시변수를 지정한다. 공분산분석모형은 분산분석모형과 회귀분석모형이 합쳐진 혼합모형으로서 관심인자의 효과에 대한 유의성 검정시 연속변수인 공변수(covariate)로 주어지는 방해인자를 미리 회귀분석으로 제거하는 방법이다. 이러한 지시변수 사용 회귀모형이나 공분산분석모형에 대한 확증적 자료분석 전에 탐색적 자료분석의 한 수단으로서 자료깊이에 근거한 *DD*-plot을 이용하면 집단 간의 차이를 쉽게 알아볼 수 있다. 이 방법은 오차항의 통계모형을 가정하지 않으므로 유용한 탐색적 방법이 될 수 있다.

자료깊이(data depth)는 지난 20여 년동안 다변량통계의 다양한 영역에서 강력한 도구로 사용되어 오고 있다. 그 이유는 이 방법이 분포의 중심성(centrality)을 알기 위한 비모수적이고 강건한 통계량을 제시하기 때문이다. 자료깊이에는 simplicial depth, Oja depth, Mahalanobis depth, half-space depth, projection depth, convex hull peeling depth, majority depth, likelihood depth 등과 같은 다양한 종류가 있으며 자료깊이의 개념에 대해서는 Liu 등 (1999)와 Li 등 (2012)에 잘 제시되어 있다. *DD*-plot이란 두 개의 다변량분포를 비교하기 위하여 이 두 개의 다변량분포에서 추출된 두 개의

This work was supported by a Research Grant of Pukyong National University(2013Year).
[1]Department of Statistics, Pukyong National University, 45 Yongso-Ro, Nam-Gu, Pusan 608-707, Korea.
E-mail: dhjang@pknu.ac.kr

표본을 대상으로 서로의 자료깊이를 계산하여 산점도 형태로 그린 그림이다. 하나의 확률분포 F로부터 나온 확률표본을 $\mathbf{X} = \{X_1, X_2, \ldots, X_l\}$이라 하고 또 하나의 확률분포 G로부터 나온 확률표본을 $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_m\}$이라 하면 DD-plot은 다음과 같이 정의된다.

$$DD(F_l, G_m) = \{(D_{F_l}(x), D_{G_m}(x)), x \in \mathbf{X} \cup \mathbf{Y}\},$$

여기서 $D(\cdot)$는 어파인-불변인 깊이를 나타낸다. 두 개의 분포 F와 G가 같으면 DD-plot은 45도 각도의 직선 주위에 점들이 나열될 것이고 두 개의 분포 F와 G가 서로 다르다면 45도 각도의 직선에서부터 서로 멀리 떨어져 있게 될 것이다. Li 등 (2012)은 시뮬레이션을 통하여 자료깊이를 기반으로 하는 이 DD-plot이 비모수적 분류자로서 탁월한 성능이 있음을 밝혔다. 또한 자료의 일부가 오염되어 있거나 특이값이 존재해도 강건함을 보였다. DD-plot은 다변량분포를 특정분포(예로 다변량정규분포)로 가정할 필요가 없는 분포무관방법이므로 모든 다변량자료의 판별에 사용할 수 있다. Liu와 Singh (1993)은 두 개의 다변량분포를 비교하기 위하여 이 두 개의 다변량분포에서 추출한 두 개의 표본을 대상으로 서로의 자료깊이를 계산하여 얻어지는 통계량인 품질지수(quality index)를 제시하고 이 통계량을 이용한 비모수검정을 제안하였다. 품질지수는 다음과 같이 정의된다.

$$Q(F, G) = P(\{D(F; \mathbf{X}) \leq D(F; \mathbf{Y}) | \mathbf{X} \sim F, \mathbf{Y} \sim G\}).$$

두 개의 분포 F와 G가 같으면 품질지수 값은 0.5가 될 것이고 품질지수 값이 0.5보다 작다면 분포 F의 $F-$깊이 50%이상이 분포 G보다 더 깊게 된다. 즉, 두 개의 분포 F와 G가 서로 다르면 다를수록 품질지수 값은 0.5보다 점점 더 작아질 것이다.

본 논문을 통하여 우리는 DD-plot이 지시변수 사용 회귀모형이나 공분산분석모형을 위한 그래픽 탐색적 자료분석방법으로서 유용함을 보이고자 한다. 2절에서 지시변수 사용 회귀모형이나 공분산분석모형에 대한 몇 가지 사례 및 시뮬레이션을 통하여 DD-plot이 지시변수 사용 회귀모형이나 공분산분석모형을 위한 그래픽 탐색적 자료분석방법으로서 유용함을 보였다. 3절에서 결론으로 마무리하였다.

2. 지시변수 사용 회귀분석과 공분산분석을 위한 *DD*-plot

2.1. 지시변수가 하나인 경우

지시변수 사용 회귀모형에서는 집단을 구분하여 주는 질적변수를 지시변수로 표현할 수 있다. 지시변수가 하나인 경우 그룹은 두 개로 나누어진다. 첫 번째 사례로서 Kang 등 (1996)에 나와 있는 당뇨병관련 쥐실험 자료는 당뇨병을 갖고 있는 쥐 9마리와 정상 쥐 25마리에 대하여 몸무게(body weight, 단위: gram)와 신장무게(kidney weight, 단위: mg)를 측정한 자료이다. 이 자료에 대하여 산점도를 그려보면 Figure 2.1과 같다. 당뇨병을 갖고 있는 쥐 그룹과 정상 쥐 그룹 사이에 몸무게와 신장무게 사이의 회귀식이 다르게 나타날 수 있음을 암시하고 있다. 이 경우 통상 우리는 당뇨병 여부를 지시변수로 정하여 당뇨병이 있는 쥐는 지시변수 값은 1로, 정상 쥐는 지시변수 값을 0으로 설정하여 다음과 같은 지시변수가 포함된 회귀분석모형을 우선 고려하게 된다. 이러한 회귀분석모형을 위한 검정을 위해서 우리는 오차항에 대한 가정을 하게 되고 오차항에 대한 가정이 맞는 지 틀리는 지를 알기 위하여 잔차를 통한 모형적합성 검토를 필히 실시하여야 한다.

$$y = \beta_0 + \beta_1 x + \beta_2 z + \epsilon, \tag{2.1}$$

여기서 y는 신장무게를 나타내는 반응변수, x는 몸무게를 나타내는 설명변수, z는 지시변수(당뇨병이 있는 쥐는 1, 정상 쥐는 0), ϵ은 오차항이다.

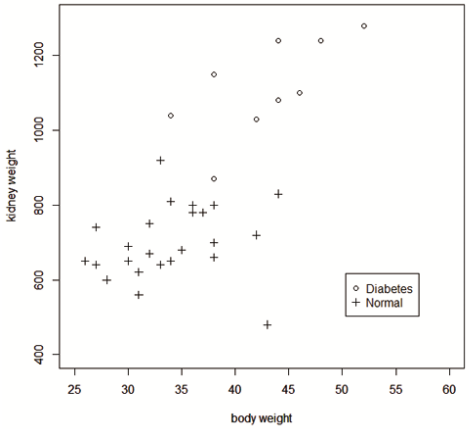

Figure 2.1. Scatterplot for diabetes data

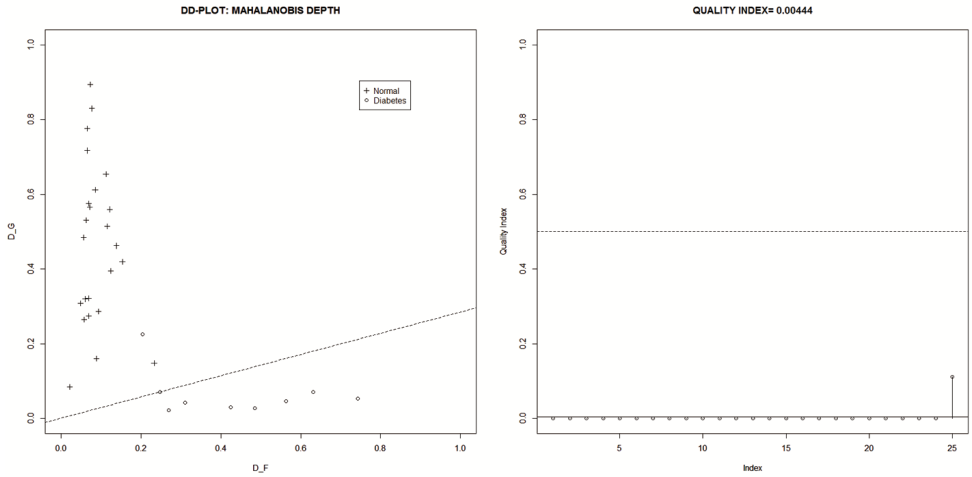

Figure 2.2. *DD*-plot and quality index plot for diabetes data

본격적인 회귀분석에 들어가기 전에 우리는 Figure 2.2와 같은 *DD*-plot을 이용하여 당뇨병을 갖고 있는 쥐 그룹과 정상 쥐 그룹이 다른 그룹인지 아닌 지를 비모수적 방법으로 알 수 있게 된다. *DD*-plot에서 점선으로 나타낸 separating line을 이용하여 두 그룹을 구분하면 오분류율이 0.0294임을 알 수 있고 품질지수값이 0.0044이어서 유의수준 0.05에서 두 그룹은 다른 그룹임을 알 수 있다. 두 그룹이 다른 그룹이라는 사전 정보를 얻게 된다. 이 방법은 오차항의 통계모형을 가정하지 않으므로 유용한 탐색적 방법이 될 수 있다. separating line은 *DD*-plot 상의 원점 $(0, 0)$을 지나는 다항식 중 일차식에서 시작하여 점차 차수를 늘리며 두 그룹을 잘 나누는 다항식을 찾으면 된다. 본 예제에서는 일차식으로 충분하다. 자료깊이로는 편의상 Mahalanobis depth를 사용하였으나 어떤 깊이를 사용하더라도 비슷한 결과를 나타낸다.

우리가 지시변수 사용 회귀모형분석을 행하여 보면 *DD*-plot에서 우리가 얻었던 사실(당뇨병을 갖고 있는 쥐 그룹과 정상 쥐 그룹은 다른 그룹임)을 확인할 수 있다. 최종적으로 회귀분석을 통하여 두 그룹의

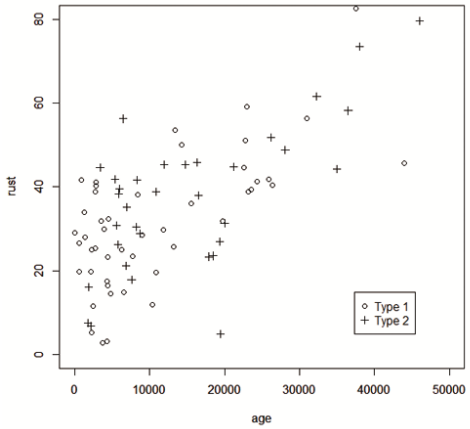

Figure 2.3. Scatterplot for rust data

추정회귀식을 찾으면(교호작용을 고려한) 다음과 같다.

$$\hat{y}_1 = 431.286 + 15.929x, \qquad \hat{y}_2 = 699.600, \tag{2.2}$$

여기서 $\hat{y}_1$은 당뇨병을 갖고 있는 쥐 그룹에 대한 추정회귀식, $\hat{y}_2$는 정상 쥐 그룹에 대한 추정회귀식이고, x는 몸무게를 나타내는 설명변수이다. 특이한 사항은 정상적인 쥐 그룹에서는 몸무게와 상관없이 일정한 신장의 무게를 갖는다는 것이다.

두 번째 사례(http://statmaster.sdu.dk/maskel/docs/sample/ST111/data에서 발췌함)로서 콘드라이트(chondrite) 운석의 두 가지 타입에 대하여 녹슨 정도(%)가 다른 지를 알고자 하는 공분산분석에서 공변수로 콘드라이트의 연령(year)을 고려하고자 한다. 이 자료에 대하여 산점도를 그려보면 Figure 2.3과 같다. type 1 그룹과 type 2 그룹 사이에 연령과 녹슨 정도 사이의 회귀식이 같게 나타날 수 있음을 암시하고 있다. 공분산분석모형을 위한 검정을 위해서 우리는 오차항에 대한 가정을 하게 되고 오차항에 대한 가정이 맞는 지 틀리는 지를 알기 위하여 잔차를 통한 모형적합성 검토를 필히 실시하여야 한다.

본격적인 공분산분석에 들어가기 전에 우리는 Figure 2.4와 같은 *DD*-plot을 이용하여 type 1 그룹과 type 2 그룹이 다른 그룹인지 아닌 지를 비모수적 방법으로 알 수 있게 된다. *DD*-plot에서 점선으로 나타낸 45도 각도를 이루는 직선을 중심으로 두 그룹에 속한 각각의 점들이 뒤섞여 두 그룹을 구분하기 어려움을 알 수 있고 품질지수값이 0.5086이어서 유의수준 0.05에서 두 그룹은 같은 그룹임을 알 수 있다. 두 그룹이 같은 그룹이라는 사전 정보를 얻게 된다. 이 방법은 오차항의 통계모형을 가정하지 않으므로 유용한 탐색적 방법이 될 수 있다.

우리가 공분산분석을 행하면 상호작용효과 age∗type에 대한 p-값이 0.6018이어서 상호작용효과가 없어 각 타입별 기울기들이 모두 일치함으로 공통 기울기를 가정할 수 있다. 제 3종 제곱합 SS(age|type), 즉 모형에 type이 들어가 있는 상태에서 모형에 새로이 age가 추가되었을 때 추가되는 제곱합에 대한 p-값이 0.0001보다 작아 공변수와 반응변수 간에 뚜렷한 회귀 관계가 존재함을 알 수 있는 반면 SS(type|age) 에 대한 p-값이 0.9337이어서 요인수준간 차이가 존재하지 않음을 알 수 있다. 제 3종 제곱합은 비가중가설을 검증하는 데 사용하는 제곱합으로서 분산분석에서 제일 흔하게 사용하는 제곱합이다. *DD*-plot에서 우리가 얻었던 사실(type 1 그룹과 type 2 그룹은 같은 그룹임)과 일치함을 알 수

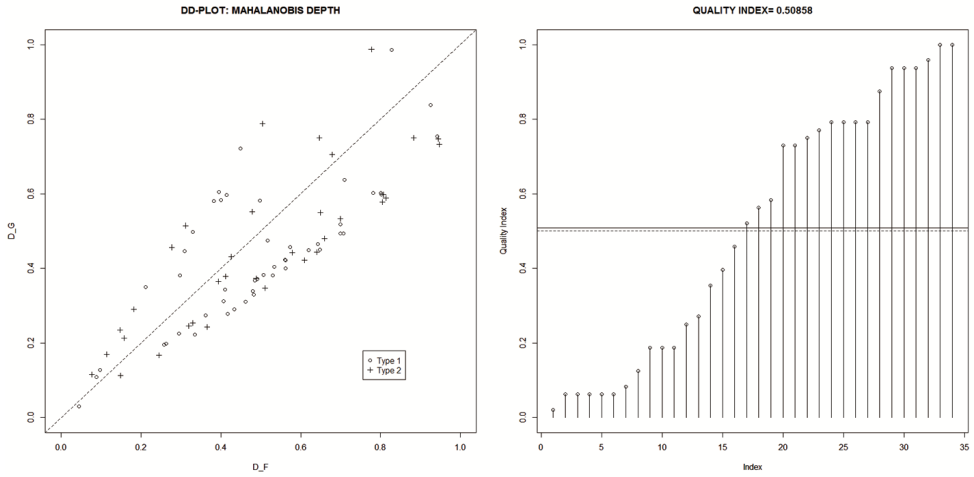

Figure 2.4. *DD*-plot and quality index plot for rust data

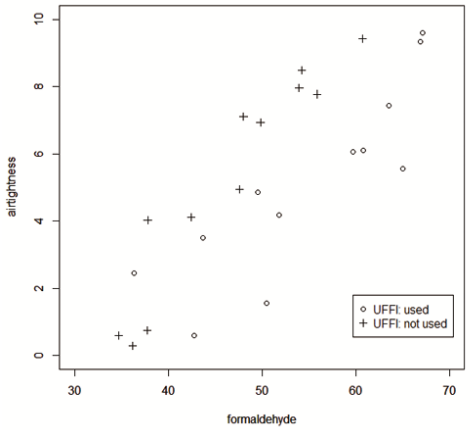

Figure 2.5. Scatterplot for UFFI data

있다. 회귀분석을 통하여 공통의 추정회귀식을 찾으면 다음과 같다.

$$\hat{y} = 23.52 + 0.0008919x, \tag{2.3}$$

여기서 $\hat{y}$은 추정회귀식이고, x는 연령을 나타내는 설명변수이다.

세번째 사례로서 Jørgensen (1993)에 나와 있는 UFFI(urea formaldehyde foam insulation) 자료에서 UFFI 사용 여부에 따라 기밀성(airtightness)에 차이가 생기는 지를 알고자 하는 공분산분석에서 공변수로 일주일동안 측정한 formaldehyde 평균농도(단위: ppb)를 고려하고자 한다. 이 자료에 대하여 산점도를 그려보면 Figure 2.5와 같다. UFFI 사용 여부에 따라 formaldehyde 평균농도와 기밀성 사이의 회귀식이 다르게 나타날 수 있음을 암시하고 있다. 공분산분석모형을 위한 검정을 위해서 우리는 오차항에 대한 가정을 하게 되고 오차항에 대한 가정이 맞는 지 틀리는 지를 알기 위하여 잔차를 통한 모형 적합성 검토를 필히 실시하여야 한다.

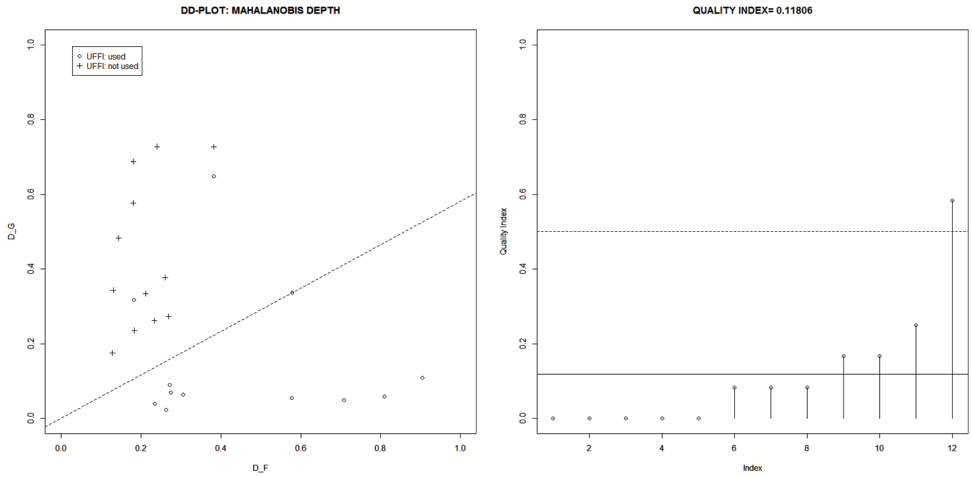

Figure 2.6. *DD*-plot and quality index plot for UFFI data

본격적인 공분산분석에 들어가기 전에 우리는 Figure 2.6과 같은 *DD*-plot을 이용하여 UFFI 사용 그룹과 UFFI 미사용 그룹이 다른 그룹인지 아닌 지를 비모수적 방법으로 알 수 있게 된다. *DD*-plot에서 점선으로 나타낸 separating line을 이용하여 두 그룹을 구분하면 오분류율이 0.0833임을 알 수 있고 품질지수값이 0.1181이어서 유의수준 0.10에서 대략 두 그룹은 다른 그룹임을 알 수 있다. 두 그룹이 다른 그룹이라는 사전 정보를 얻게 된다. 이 방법은 오차항의 통계모형을 가정하지 않으므로 유용한 탐색적 방법이 될 수 있다.

우리가 공분산분석을 행하면 상호작용효과 formaldehyde∗uffi에 대한 p-값이 0.0536이어서 상호작용효과가 있다고 할 수 있다. *DD*-plot에서 우리가 얻었던 사실(UFFI 사용 그룹과 UFFI 미사용 그룹이 다른 그룹임)과 일치함을 알 수 있다. 회귀분석을 통하여 각각의 추정회귀식을 찾으면 다음과 같다.

$$\hat{y}_1 = -7.7676 + 0.2348x, \qquad \hat{y}_2 = -11.3803 + 0.3558x, \tag{2.4}$$

여기서 $\hat{y}_1$은 UFFI 사용시의 추정회귀식, $\hat{y}_2$는 UFFI 미사용시의 추정회귀식이고, x는 formaldehyde 평균농도를 나타내는 설명변수이다.

우리는 지시변수가 하나인 경우(모형식: $y = \beta_0 + \beta_1 x + \beta_2 z + \epsilon$)를 가정하고 다음과 같이 시뮬레이션($y$-절편은 0이고 동일 기울기 가정)을 행하여 볼 수 있다.

첫 번째 그룹(지시변수 값이 0): $y = x + \epsilon,\ \epsilon \sim N(0, 0.05^2)$,

두 번째 그룹(지시변수 값이 1): $y = x + \beta + \epsilon,\ \epsilon \sim N(0, 0.05^2)$.

x는 0과 1사이의 난수로 발생시키고 지시변수에 대응되는 회귀계수 β를 0에서 0.5까지 변화시켜가며 산점도의 변하는 모습과 이에 대응되는 동적 *DD*-plot의 변하는 모습을 서로 비교하여 보면 동적으로 *DD*-plot의 유용성을 확인 할 수 있다. Figure 2.7–2.9는 β 값이 각각 0, 0.136, 0.264에 대응되는 두 개의 그룹에 대한 산점도 및 동적 *DD*-plot를 나타낸다. 각 그림에서 십자 표시는 첫 번째 그룹을 나타내고 동그라미 표시는 두 번째 그룹을 나타낸다. β 값이 0일 때는 *DD*-plot에서 점선으로 나타낸 45도 각도를 이루는 직선을 중심으로 두 그룹에 속한 각각의 점들이 뒤섞여 두 그룹을 구분하기 어려움을 알 수 있고 품질지수값이 0.556이어서 유의수준 0.05에서 두 그룹은 같은 그룹임을 알 수 있다. β 값이 0.136일 때는 품질지수값이 0.0352여서 유의수준 0.05에서 두 그룹이 다른 그룹임을 알 수 있다. β 값

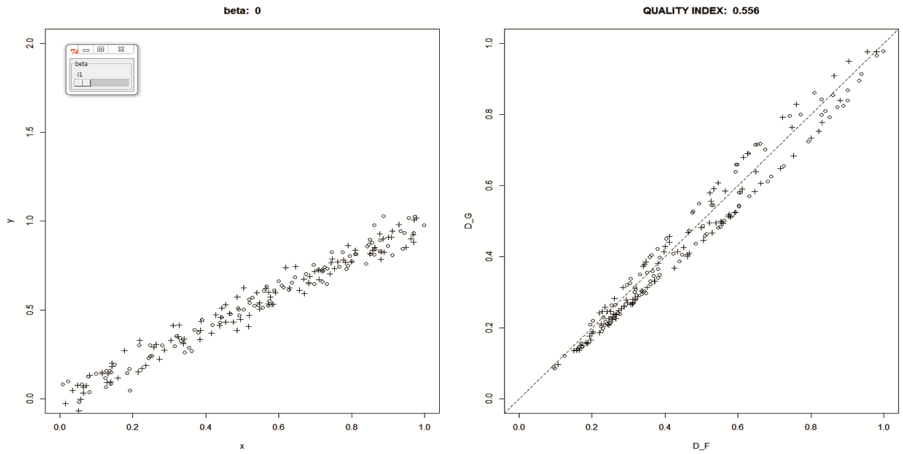

Figure 2.7. Dynamic scatterplot and dynamic DD-plot with $\beta = 0$

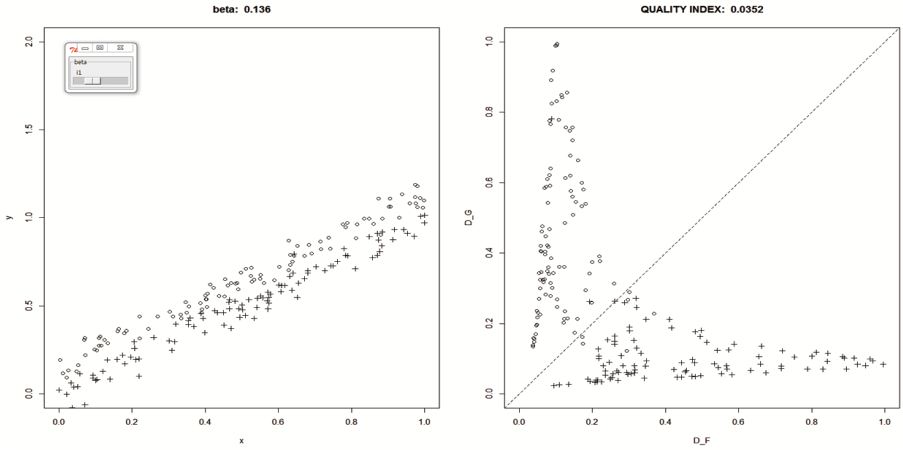

Figure 2.8. Dynamic scatterplot and dynamic DD-plot with $\beta = 0.136$

이 0.264일 때는 품질지수값이 0이어서 유의수준 0.05에서 두 그룹이 확실하게 다른 그룹임을 알 수 있고 오분류율도 0이 됨을 알 수 있다. 슬라이드바를 좌우로 움직이며 β를 0에서 0.5까지 변화시키면서 산점도의 변하는 모습과 이에 대응되는 동적 DD-plot의 변하는 모습을 서로 비교하여 볼 수 있다.

2.2. 지시변수가 두 개 이상인 경우

지시변수 사용 회귀모형에서 지시변수가 두 개 이상인 경우 그룹은 세 개 이상이 되므로 더 이상 DD-plot을 사용할 수 없다. 그러므로 우리는 DD-plot의 확장을 고려하여야 한다. 예로 지시변수가 두 개인 경우 그룹이 세 개가 되므로 DD-plot의 확장으로서 우리는 DDD-plot을 제안할 수 있다. 세 개의 확률분포 F, G, H로부터 나온 확률표본을 각각 $\mathbf{X} = \{X_1, X_2, \ldots, X_l\}$, $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_m\}$, $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_n\}$이라 하면 DDD-plot은 다음과 같이 정의된다.

$$DDD(F_l, G_m, H_n) = \{(D_{F_l}(x), D_{G_m}(x), D_{H_n}(x)), x \in \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}\}.$$

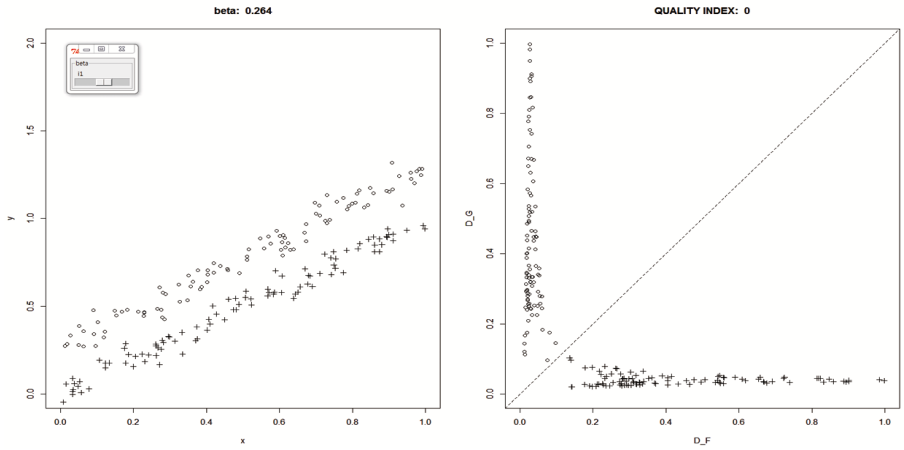

Figure 2.9. Dynamic scatterplot and dynamic DD-plot with $\beta = 0.264$

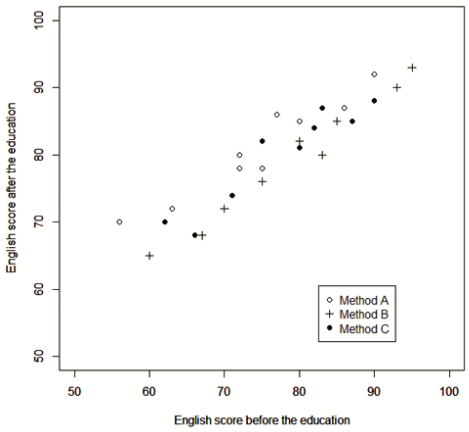

Figure 2.10. Scatterplot for English score data

이러한 DDD-plot은 삼차원 그림이 된다. $F = G = H$이면 DDD-plot은 45도 각도의 삼차원 직선 주위에 점들이 나열될 것이고 세 개의 분포들이 서로 다르다면 45도 각도의 삼차원 직선에서부터 서로 멀리 떨어져 있게 될 것이다.

네 번째 사례로서 Kang과 Kim (2013)에 나오는 영어교육방법 비교 자료를 보면 최근에 제안된 영어교육방법 세 가지(A, B, C)를 비교하기 위하여 중학교 2학년 남학생을 대상으로 6개월 동안 각 영어교육방법으로 교육을 실시한 후 영어시험성적을 측정하였다. 교육대상자의 IQ는 랜덤화로 평준화시키고 공변량으로서 교육전 영어시험성적을 이용하였다. Figure 2.10은 교육전 영어시험성적과 교육후 영어시험성적을 영어교육방법별로 구분한 산점도이다.

본격적인 공분산분석에 들어가기 전에 Figure 2.11과 같은 DDD-plot를 이용하여 세 가지 영어교육방법들에 대하여 비모수적인 방법으로 구분하여 볼 수 있다. DDD-plot은 삼차원 그림이므로 컴퓨터 화면 상에서 여러 방향으로 회전시켜 가며 확인하여 볼 수 있다. 교육방법 A는 교육 방법 B와 C와는 구

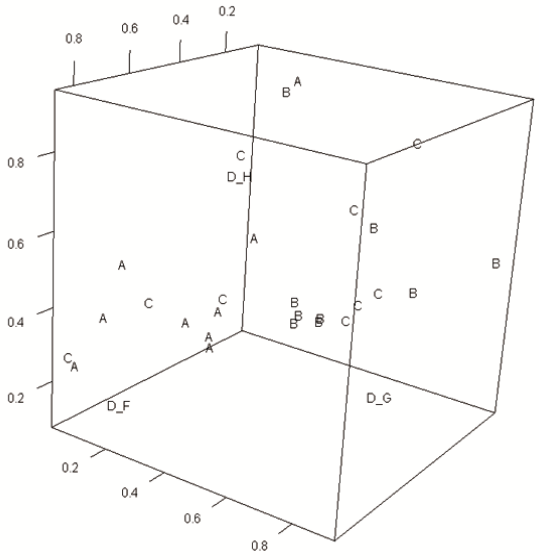

Figure 2.11. *DDD*-plot for English score data

별되며 교육 방법 B와 C는 구별되지 않음을 알 수 있다. 이 방법은 오차항의 통계모형을 가정하지 않으므로 유용한 탐색적 방법이 될 수 있다.

영어교육방법이 세 가지이므로 교육방법 C를 기준범주로 설정하고 다음과 같이 두 개의 지시변수 Z_1(교육방법이 A이면 1, 아니면 0)과 Z_2(교육방법이 B이면 1, 아니면 0)를 정의하고 공변량 x(교육전 영어시험성적)을 사용하여 회귀식을 찾으면(동일한 기울기 가정) 다음과 같다.

$$\hat{y} = 22.677 + 3.055Z_1 - 1.875Z_2 + 0.740x, \tag{2.5}$$

여기서 $\hat{y}$은 반응변수(교육후 영어시험성적)에 대한 추정회귀식이다. 4개의 회귀계수 각각에 대한 양측 검정결과 y-절편과 공변량에 대응하는 회귀계수들에 대한 p-값은 10^{-3}보다 작고 Z_1에 대응되는 회귀계수에 대한 p-값은 0.0059인데 반해 Z_2에 대응되는 회귀계수만이 유의수준 5%에서 통계적으로 유의하지 않으므로(p-값 = 0.0741) 교육방법 A와 C간에는 유의한 차이가 있고 교육방법 B와 C간에는 유의한 차이가 없음을 알 수 있다. *DDD*-plot에서 우리가 얻었던 사실과 일치함을 알 수 있다. 그러나 공분산분석모형을 위한 검정을 위해서 우리는 오차항에 대한 가정을 하게 되고 오차항에 대한 가정이 맞는지 틀리는 지를 알기 위하여 잔차를 통한 모형적합성 검토를 필히 실시하여야 한다. 지시변수 사용 회귀모형에서 지시변수가 세 개 이상인 경우는 *DD*-plot의 확장으로서 더 이상 *DDD*-plot을 사용할 수 없으므로 *DD*-plot matrix같은 그래픽도구를 사용하여야 한다. *DD*-plot matrix에서 각 패널은 대응되는 두 개의 그룹 사이의 *DD*-plot이 된다.

3. 결론

우리는 지시변수 사용 회귀모형이나 공분산분석모형에 대한 확증적 자료분석 전에 탐색적 자료분석의 한 수단으로서 자료깊이에 근거한 *DD*-plot을 이용하면 집단 간의 차이를 알아볼 수 있다. 이 방법은 오차항의 통계모형을 가정하지 않으므로 유용한 탐색적 방법이 될 수 있다.

References

Jørgensen, B. (1993). *The Theory of Linear Models*, Chapman & Hall, New York.

Kang, G. S. and Kim, C. R. (2013). *Linear Regression Analysis*, Kyowoosa, Seoul.

Kang, M. O., Kim, Y. I., Ahn, C. H. and Lee, Y. G. (1996). *Regression Analysis*, Yoolgok Pub, Seoul.

Li, J., Cuesta-Albertos, J. A. and Liu, R. (2012). *DD*-Classifier: Nonparametric classification procedure based on *DD*-plot, *Journal of the American Statistical Association*, **107**, 737–753.

Liu, R., Parelius, J. M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference, *The Annals of Statistics*, **27**, 783–858.

Liu, R. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests, *Journal of the American Statistical Association*, **88**, 252–260.

http://statmaster.sdu.dk/maskel/docs/sample/ST111/data

ANCOVA 모형을 위한 *DD*-plot

장대흥[a,1]

[a]부경대학교 통계학과

(2013년 12월 2일 접수, 2014년 03월 13일 수정, 2014년 04월 05일 채택)

요 약

우리는 회귀분석에서 설명변수들 중 일부가 질적 변수인 경우 지시변수를 사용한다. 또한 공분산분석모형에서는 관심인자의 효과에 대한 유의성 검정시 연속변수인 공변수로 주어지는 방해인자를 미리 회귀분석으로 제거한다. 지시변수 사용 회귀모형이나 공분산분석모형을 위한 확증적 자료분석 전에 탐색적 자료분석의 한 수단으로서 자료깊이에 근거한 *DD*-plot을 이용하면 집단 간의 차이를 쉽게 알아볼 수 있다. 이 방법은 오차항의 통계모형을 가정하지 않으므로 유용한 탐색적 방법이 될 수 있다. 몇 가지 사례들을 통하여 *DD*-plot이 지시변수 사용 회귀모형이나 공분산분석모형을 위한 그래픽 탐색적 자료분석방법으로서 유용함을 보였다.

주요용어: 자료깊이, *DD*-plot, ANCOVA.

이 논문은 부경대학교 자율창의학술연구비(2013년)에 의하여 연구되었음.

[1](608-737) 부산광역시 남구 용소로 45, 부경대학교 통계학과. E-mail: dhjang@pknu.ac.kr.