제3장

그래픽 내검기법을 이용한 내검 효율성 제고

이 의 규 · 심 규 호

제1절 서 론

1. 연구의 배경 및 필요성

특이치(exceptional data)는 조사의 전체 결과에 큰 영향을 준다거나 오류인 까닭으로 검토되어야 한다. 일반적으로 조사업무 담당자는 자료 수집과정에서 나타날 수 있는 이러한 특이치를 검색한다. 일단 의심스러운 건수가 검출되면 원 조사표를 검토하거나 응답자를 재접촉하여 자료가 정확한 것인지를 확인한다. 주어진 시한 내에 가능한 많은 오류를 수정하기 위해서는 가장 오류일 가능성이 높은 건수를 정확하게 식별하는 것이 필요한데 이 과정이 효율적 내검(editing)의 중요한 부분이다.

자료의 이상치, 패턴, 자료 속에 내재된 관계는 순수하게 수치·분석적인 방법만으로 찾아내기 쉽지 않다. 그래픽 내검(graphical editing)은 사람의 시각적 인지력을 이용하여 이들을 쉽게 검출하는 방법이다. 이러한 방법은 그동안 많은 국가 통계 기관에서 통계조사의 비용을 줄이는 효율적인 도구로 사용되어 왔다. 이는 그래프를 이용하는 방법이 내검 과정을 개선하고 관리하는 데 도움이 되기 때문일 것이다.

따라서 내검작업 수행 시 수치자료를 분석·검토하는 불편함과 지루하고 반복되는 내검작업을 개선할 여지가 있다. 또한 내검 기준 설정 시 주어진 자료의 정보가 충실히 반영될 수 있는 방법을 이용하는 것이 필요하다. 대칭적 증감률의 내검 기준보다는 자료의 중심과 산포 등의 정

보를 반영하여 이상치를 분별할 수 있도록 하는 것이 바람직하다.

현재 광업・제조업 조사에서는 주요항목별 전년도 대비 증감률에 대한 임의조건을 설정함으로써 지역별로 내검량 조정이 가능하게 되었다. 그럼에도 불구하고 내검요원은 여러 번 검색조건을 반복하여 결정하는 번거로움이 있을 뿐 아니라 전체자료의 움직임이나 패턴을 이해할 수 없다는 단점이 있다. 더욱이 한 번에 하나의 레코드를 검토하는 것은 관련된 레코드 전체의 움직임을 크게 보기 어렵다. 여러 레코드를 종합적으로 검토하는 것은 총 결과에 미치는 개별 레코드의 영향을 가늠하는 데 있어서도 중요하다.

2. 연구의 목적 및 내용

본 연구는 이상치 탐색 방법인 Hidiroglou-Berthelot(H-B) 방법과 기초적인 탐색적 자료 분석(Exploratory Data Analysis; EDA) 방법을 소개하고 사업체 대상 조사에 이들을 적용함으로써 종합적이고 시각적으로 이상치를 탐색하고자 한다. 즉 H-B방법에 의한 특이값 구분의 기준 설정과 산점도에 의한 이상치나 패턴 등의 감지를 통해 조사의 내검 효율성을 제고하고자 한다.

먼저 해외의 그래픽 내검 사례를 살펴본다. 미국, 뉴질랜드, 스웨덴을 중심으로 예제 그래프와 함께 간략히 특징만을 살펴보기로 한다. 이후 간단한 EDA 기법과 H-B 기법을 소개하고, 광업・제조업 조사 자료에 탐색기법을 적용한다. 상자그림, 산점도 등 매우 간단한 그래픽 기법과 함께 Hidiroglou-Berthelot에 의해 고안된 방법을 통해 이상치를 검출하고 이에 따른 적정 내검량 설정 방안을 검토한다.

제2절 해외의 그래픽 내검 사례

해외의 그래픽 내검은 대부분 자체 그래픽 분석 프로그램이나 시스템을 보유하고 있지 않고 SAS/INSIGHT나 SAS JMP 등의 프로그램을

나름대로 이용하고 있는 실정이다. 그럼에도 불구하고 미국을 포함한 유럽의 여러 나라의 통계 조사 기관들은 매크로 데이터의 내검을 위해 조사에 적합한 그래픽 방법을 활용하고 연구하고 있다. 그래픽 내검은 간단하게는 단변량 자료에 대한 산점도, 히스토그램, 시계열 그림 등을 통한 EDA(Exploratory Data Analysis)와 Hidiroglou와 Berthelot가 제안한 이상치 검출 기법, 스코어 함수 내검(score function edit)기능이 포함되어 있다.

1. 미국

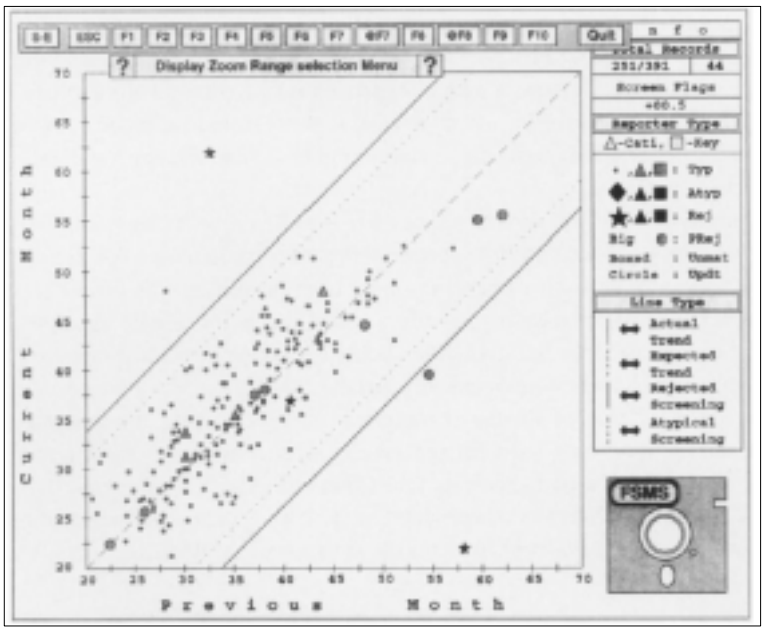
미국은 그래픽 내검의 활용이 가장 활발한 나라 중 하나이다. BLS(Bureau of Labor Statistics)와 U.S Census Bureau, NASS(National Agricultural Statistics Service), EIA(Energy Information Administration) 등에서 각 기관의 조사 자료에 대하여 그래픽 내검을 적용하고 시스템 개발・연구를 하고 있다.

가. ARIES

대규모의 주기적인 경제・사회 조사에서 생산되는 다량의 자료는 자료의 품질 향상을 위한 기법들을 필요로 하였다. BLS(Bureau of Labor Statistics)는 새로운 시스템을 개발하기 위해 다음의 세 가지 주요한 동기를 제시하였다. 첫째는 내검요원에 의해 보고되는 조사표의 재점검을 줄이거나 제거하는 것이다. 두 번째는 내검요원으로 하여금 전체적으로 조사 자료를 파악할 수 있도록 함과 동시에 적은 긴장감으로 이상점(outlier)을 찾게 하는 것이다. 마지막으로는 시간에 종속되는 추정값(estimates)과 세부 그룹 간의 하부 추정값(subestimates)에 대한 광범위한 현황을 제공하는 것이다(Esposito, 1994).

[그림 3-1]의 산점도에서 가로축은 이전 달의 값이고 세로축은 현재의 값을 의미한다. 가운데 점선은 이 두 값이 정확히 같은 값으로 일치할 때 이 선상에 나타남을 의미하고 이 선에서 벗어나면 벗어날수록 불일치함을 나타낸다. 또한 하한과 상한을 줌으로써 이상값을 구별하고

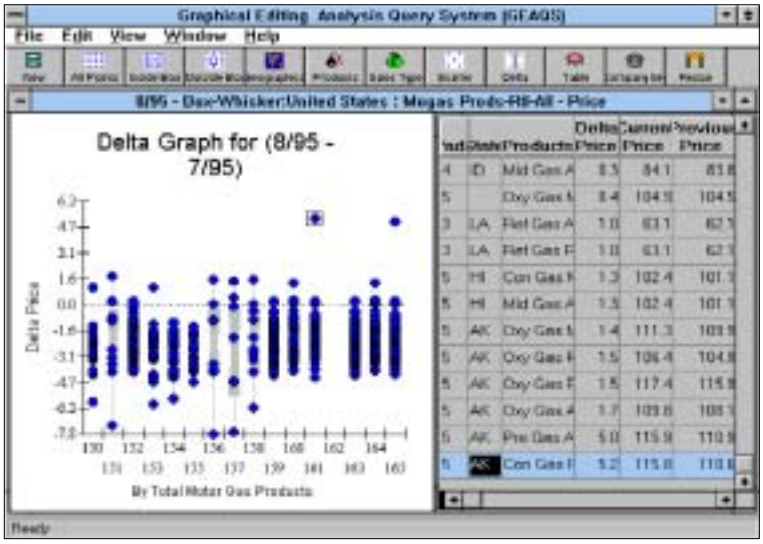
있으며 각 점들은 색깔과 모양으로 그 관측대상을 구분하여 추가정보를 나타내고 있다.



[그림 3-1] ARIES의 산점도 화면

나. GEAQS

GEAQS(Graphical Editing Analysis Query System)는 EIA(Energy Information Administration)에서 개발한 그래픽 내검 시스템이다. 이 시스템은 ARIES System, Bureau of Census Working Group Prototype, Graphical Macro-Editing Application (Statistics Sweden), Distributed EDDS Editing Project(DEEP; Federal Reserve Board)의 네 개의 시스템에서 주요 기능을 참고하여 개발되었다. 이 시스템은 상자그림(Box-whisker Plot), 산점도(Scatter Plot) 등의 William Cleveland가 제안한 많은 그래픽 자료 표현 기법을 포함하고 있으며 EDA(Exploratory Data Analysis)기법을 이용하여 정상적이지 않은 석유 가격을 찾아내는 데 도움을 준다(Dale, 2000).



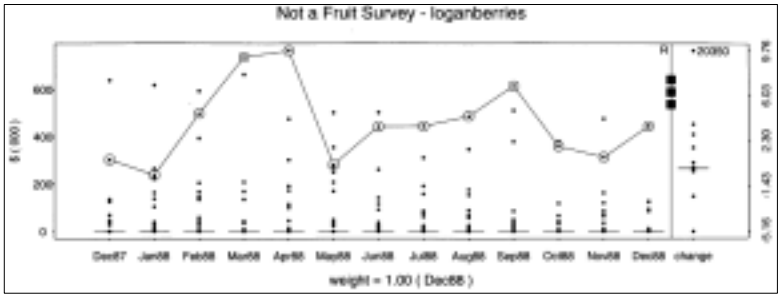
[그림 3-2] GEAQS의 상자그림

[그림 3-2]에서 세로축은 95년 8월의 휘발유 가격에서 7월의 가격을 뺀 가격이다. 여러 주의 가격 차이 값이 각 생산품별로 표시되고 있다. 따라서 0을 중심으로 하는 수평선에서 벗어날수록 당월의 가격이 전월의 가격과 변동을 보이고 있음을 나타낸다. 특히 네모로 선택된 점은 알라스카(AK)주의 해당 생산품 가격 차이 값이 5.2로 전월보다 크게 증가하고 있어 다른 자료와 확연히 구분되고 있다. 이 점에 대한 정보가 옆의 화면에 나타나고 있음을 볼 수 있다.

그 밖에 DEEP(Distributed EDDS Editing Program)는 연방준비위원회(Federal Reserve)에 의해 금융 정보를 분석하기 위해 사용되며 ARIES와 GEAQS와 많은 부분 유사하게 디자인되었다(Dale, 2000).

2. 뉴질랜드

NZDos(New Zealand Department of Statistics)에서는 "GRED"라는 상호적인(interactive) 그래픽 에디팅 시스템을 개발하였다(Houston, 1993). GRED는 조사과정 중 매크로 내검(macro-editing) 단계에서 사용되도록 설계되었으며 경제 조사의 내검에 사용되었다. [그림 3-3]에서 볼 수 있듯이 GRED는 다음과 같은 특징을 가지고 있다.



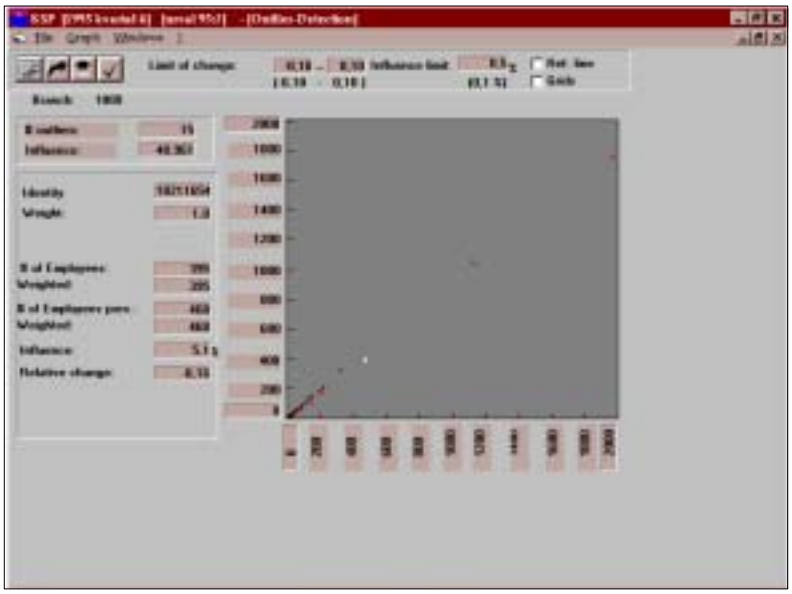
[그림 3-3] GRED의 시계열 화면

- Boxplot을 시간의 경과에 따라 도식함으로써 시점에 걸쳐 이상치를 식별할 뿐만 아니라 자료의 추세를 보여주고 있다.
- 그래프 상에 출력된 자료값은 마우스 동작에 민감하여(mouse-sensitive) 사용자로 하여금 그래프상의 이상점에 대한 정보를 빠르게 볼 수 있게 한다.
- 다중화면 보기(multiple view)는 사용자에게 그래프와 통계값 등을 동시에 보여줌으로써 거시자료(macro)와 미시자료(micro)를 함께 볼 수 있게 한다.
- 컬러와 다른 강조 기술의 사용은 중요한 정보를 분리하여 준다.

3. 스웨덴

스웨덴 통계청에서는 단기고용조사(Short Periodic Employment Survey)에 적용하기 위해 Graphical Macro-editing PC Application을 개발하였다. 이 프로그램은 Visual Basic을 기반으로 개발되었으며 Microsoft SQL Server를 Database로 사용하고 있다(Engström, 2005).

[그림 3-4]에서 보면 현재와 과거의 종사자 수의 변화를 나타내고 있으며 이에 따른 이상치를 도식화하고 있다. 또한 점의 색깔을 달리하여 그림에서 표시되도록 하고 있다. 의심스러운 추정 값을 마우스로 클릭했을 때 해당되는 관측 값이 보여 이에 대한 조치를 취하고 있다.



[그림 3-4] 스웨덴의 그래픽 내검 화면

제3절 EDA와 Hidiroglou-Berthelot 방법

탐색적 자료 분석(Exploratory Data Analysis; EDA) 기법은 조사자로 하여금 자료의 패턴을 볼 수 있도록 하며 우선적으로 추적해야할 자료를 결정하는 데 중요한 정보를 제공할 수 있다. 그리고 Hidiroglou와 Berthelot에 의해 고안된 방법은 주기적 조사 자료에 많이 쓰이는 이상치 검출 방법이다. 본 연구에서는 사용자가 이해하기 쉽고 사용하기 편리하도록 대표적인 방법을 소개하였다. 특히 여기서 소개된 의심스러운 자료의 검색방법은 사업체대상 조사의 종합내검과 연계하여 사용될 수 있다.

1. EDA 방법

탐색적 자료 분석(Exploratory Data Analysis) 기법은 자료에서 탐색해내기 힘든 것을 용이하게 찾는 방법으로 흔히 표현되곤 한다. 특히 대부분의 자료와 구별되는 자료를 결정하는 방법을 제공하고 이를 도식화한다. EDA에서 가장 기본적인 그래프는 산점도와 상자그림이다.

산점도(scatter plot)는 두 변수 간 관계를 살피고자 할 때 흔히 사용되는 방법이다. 이는 간단하지만 매우 유용한 도구이다. 특히 현재와 이전의 자료에 대한 산점도의 경우, 기울기가 1인 직선에 근접한 자료는 현재의 자료가 과거 자료와 일치되는 사업체를 의미하며 이 선에서 멀어질수록 불일치하다는 것을 나타내므로 사업체조사에서 유용하게 쓰일 수 있다.

상자그림(box plot)은 자료를 요약하거나 이상치를 파악하고자 할 때 유용하게 사용된다. 이는 자료의 중심과 퍼짐, 그리고 형태를 간결하게 나타내는 도구이다. 상자그림은 기본적으로 사분위수에 근거하여 작성된다. 상자(box)의 양쪽 끝 위치는 제1사분위수(Q_1)와 제 3사분위수(Q_3)로 자료의 50%가 이 상자 구간에 있음을 나타내고 상자에 달린 수염은 최소값과 최대값에 선으로 연결한다. 그러나 상자 양 끝에서 보통 상자 길이의 1.5배수를 빼거나 더한 값, 즉 $Q_1 - 1.5(Q_3 - Q_1)$ 또는 $Q_3 + 1.5(Q_3 - Q_1)$의 범위 안에서 가장 작거나 큰 값은 인접값에 연결

되고 이 범위를 벗어난 자료는 개별 점으로 표시하여 보통 이상치(mild outlier)로 간주한다. $Q_1 - 3(Q_3 - Q_1)$ 또는 $Q_3 + 3(Q_3 - Q_1)$을 벗어나는 자료는 극단 이상값(extreme outlier)으로 구분하기도 한다. 자료가 대칭적이라면 이러한 이상치의 구분은 검토를 필요로 하는 자료를 정의하는데 좋은 도구가 된다.

그러나 경제 자료는 일반적으로 오른쪽으로 길게 늘어진다. 만약 자료가 이와 같이 한 쪽으로 치우치는 속성을 갖는다면, 꼬리 쪽의 많은 자료가 실제로는 그 분포에서 일어날 수 있는 값이지만 이상치로 표시된다. 따라서 이러한 경우에 원래의 값에 대해 상자그림을 이용하여 이상치를 발견하는 것은 그리 유의하지 않을 수 있다.

한편, 현재년도 출하액과 과거년도 출하액의 비(ratio)에 대한 상자그림을 고려하여 보자. 비의 중위수에서 멀어질수록 이들 자료는 검토되어야 마땅할 것이므로 비에 대한 상자그림이 유용할 것이다. 그런데 비의 중위수가 0.8이라 하자. 만약 일반적으로 비의 중심이 1이라 생각하고 0.8과 1.2로 설정하여 검색한다면 이는 더 많은 자료가 검출될 뿐 아니라 이상치를 구별하는 데 도움이 되지 못한다. 따라서 상자그림에서와 같이 자료의 분포를 반영하는 검색기준이 필요하다. 다음 절에서는 이상치 탐색방법의 하나인 Hidiroglou와 Berthelot에 의해 고안된 방법을 좀 더 자세히 살펴보기로 한다.

2. Hidiroglou-Berthelot 방법

이상치의 검출은 무응답 대체나 오류자료의 수정에 앞서 필요한 절차로 중요하다. 연간조사나 월간조사와 같이 연속적인 조사에서 이상치를 검출하기 위한 방법으로 Hidiroglou와 Berthelot(1986)에 의해 고안된 방법을 대표적으로 사용한다. 이 방법은 현재의 자료에서 한 항목과 그와 관련된 항목의 비(ratio)나, 한 항목의 현재자료와 과거자료와의 비가 일정 범위 안에 포함되는지를 판단하는 방법이다. 범위는 변환 비의 대푯값을 중심으로 상한과 하한 값으로 정하여진다. 이 한계값을 결정하는 데 있어서 사용자가 설정하는 승수가 포함되어 주관이 완전히 배제되지 않는다는 약점이 있다.

한계값의 결정은 이상치 검출 건수에 영향을 주며 이에 대한 재조사를 고려하게 되면 내검량과 연관된다. 이때 그래프를 이용하여 현재와 과거자료의 비(ratio)를 도식화하면, 시각적으로 중심에서 벗어난 자료를 확인할 수 있을 뿐 아니라 내검량의 기준이나 내검규칙을 설정하는 데 도움이 될 것이다.

캐나다 통계청의 이상치 검출 과정은 두 가지 형태의 값으로 식별한다(Banff Support Team, 2007). 하나는 대체 이상치(Outlier Detection Imputation: ODI)라 하는데 이는 나중에 대체되어야할 정도로 나머지 값과는 매우 다른 값이다. 두 번째는 사용배제 이상치(Outlier Detection Exclusion ODE)이다. 정상적이지는 않지만 대체할 정도는 아닌 값으로 나중에 대체과정에서 이 값이 쓰이지 않도록 하기 위해 구분한다. 즉 Donor 임퓨테이션 절차에서 이들을 제공하고 싶지 않거나 임퓨테이션 추정 시에 배제되는 값이다. 이들은 이상치 검출을 위해 Hidiroglou와 Berthelot에 의해 고안된 방법으로 한계값을 정의하고 있다. 한계값은 고정적이지 않으며 사용자가 설정하는 값과 자료로부터 얻어지는 값에 의해 정의된다.

Hidiroglou와 Berthelot에 의해 고안된 절차는 비를 이용하는 방법(ratio method), 과거자료를 이용하는 방법(historical trend method), 현 자료를 이용하는 방법(current method)으로 구분하여 사용될 수 있다. 만약 신뢰할 수 있는 보조정보가 있다면 비를 이용하는 방법(ratio method)을 사용하고, 과거자료를 이용하는 방법(historical trend method)은 비를 이용하는 방법의 특별한 경우로서 과거의 자료가 보조정보로서 취해질 때 사용하게 된다. 어떤 보조정보나 과거자료가 없는 경우에는 현 자료를 이용하는 방법(current method)을 이용한다. 이 세 가지 방법 중 적절한 선택은 자료에 따라 결정된다. 이상치를 결정하는 방법은 상자그림에서 이상치를 판단하는 것과 유사하다.

가. 현 자료를 이용하는 방법(current method)

현 자료를 이용하는 방법(current method)은 선택된 각 변수에 대해 다음과 같이 수행된다.

- 제1사분위수 Q_1, 중위수 M, 그리고 제3사분위수 Q_3를 계산한다.
- D_L, D_U를 계산한다.
  이들은 각각 중위수로부터 제1사분위수와 제3사분위수의 거리이다(각 사분위수가 너무 가까우면 사용자가 설정한 중위수의 일정 배수를 채택하게 함).

$$D_L = Max(M - Q_1, |K|)$$
$$D_U = Max(M - Q_3, |K|)$$

- 대체 이상치와 사용배제 이상치를 결정하는 구간을 결정한다. 식별 구간은 D_L, D_U와 사용자 설정 승수의 함수이다. 사용자 설정 승수 c_1 과 c_0를 설정시 둘 중 하나 또는 두 개 다 설정할 수 있으며 두 개를 설정한다면 c_1 은 c_0 보다 커야한다(예를 들면 c_1 = 6, c_0= 3). 극단적 이상치를 식별하기 위한 범위는 아래와 같다.

$$x_i < M - c_1 D_L \text{ 또는 } x_i > M + c_1 D_U$$

사용배제 이상치(ODE)를 식별하기 위한 범위는 아래와 같다.

$$M - c_1 D_L \le x_i < M - c_0 D_L \text{ 또는 } M + c_0 D_U < x_i \le M + c_1 D_U$$

나. 비를 이용하는 방법(ratio method)

비를 이용하는 방법(ratio method)은 각 선택된 변수에 대해 다음과 같이 수행된다.

- $r_i = \frac{x_i}{y_i}$ 를 계산한다.
  x_i : i번째 레코드의 선택되어진 항목 값 $(x_i > 0)$
  y_i : i번째 레코드의 보조 항목 값 $(y_i > 0)$
- s_i 값으로 변환한다.

$$s_i = \begin{cases} 1 - \frac{r_m}{r_i} \ , \ 0 < r_i < r_m \\ \frac{r_i}{r_m} - 1 \ , \ r_i \ge r_m \end{cases}$$

여기서 r_m은 r_i의 중위수(median)이다. s_i 는 분포의 양쪽 부분에서 이상치가 동등하게 검출될 수 있도록 해 준다.

- 효과(effect) $e_i = s_i[\max(x_i, y_i)]^u$ 를 계산한다.
  사용자가 u값을 정할 수 있는데 $u=0$은 관측값이 크거나 작거나 똑같이 취급하고(이 경우 e_i 는 s_i 값과 같음), $u=1$은 큰 단위의 작은 변동에 더 큰 의미를 부여한다.
- 이 e_i 의 제1사분위수 e_{q1} , 중위수 e_m , 그리고 제3사분위수 e_{q3} 를 계산하여 앞의 current method에서와 같은 방법으로 범위를 결정한다. 즉, 대체 이상치(ODI)를 식별하는 범위는 아래와 같다.

$$e_i < e_m - c_1 e_{D_L} \text{ 또는 } e_i > e_m + c_1 e_{D_U}$$

그리고 사용배제 이상치(ODE)를 식별하는 범위는 아래와 같다.

$$e_m - c_1 e_{D_L} \le e_i < e_m - c_0 e_{D_L} \text{ 또는 } e_m + c_0 e_{D_U} < e_i \le e_m + c_1 e_{D_U}$$

다. 과거자료를 이용하는 방법(historical method)

과거자료를 이용하는 방법(historical method)은 비를 이용하는 방법(ratio method)의 보조 변수 y_i 대신 x_i 의 과거자료를 사용하는 특수한 경우이다. 즉,

$$r_i = \frac{x_{i,t}}{x_{i,t-1}}$$

식을 계산한 후 앞의 비를 이용하는 방법(ratio method)과 같은 절차를 따라 식별 범위를 구한다.

현 자료를 이용하는 방법(current method)은 선택된 항목 값의 전체 분포를 고려하여 예외적으로 작은 값이나 큰 값을 탐지하는 반면 과거자료를 이용하는 방법(historical method)은 전체 레코드에 걸쳐 현재자료와 과거자료와의 상대적인 차이의 분포를 고려하여 예외적인 값을 탐색하는 것이 다르다.

광업·제조업 조사는 매년 실시되는 조사로서 새로 생기거나 사멸하는 사업체를 제외하고 과거자료가 존재하므로 과거자료를 이용하는 방

법(historical method)을 사용하는 것이 효과적일 것이다. 다음 절에서는 지금까지 살펴본 과거자료와 현재자료를 이용하는 H-B방법과 산점도를 광업·제조업 조사자료에 적용한다.

제4절 광업·제조업 조사에 적용

광업·제조업 조사는 조사의 정도를 높이기 위해 종합내검을 실시하고 있다. 종합내검이란 주요항목의 전년대비 증감률이 큰 사업체를 대상으로 올해 조사된 내용이 정확하게 조사되었는지를 재검토하는 작업을 말한다(통계청, 2008). 2007년 기준 광업·제조업 조사의 입력시스템 이용자 지침서를 살펴보면, 종합내검 검색조건은 기본고정조건과 임의조건 두 가지로 구성되어 있다.

기본고정조건은 과거자료 분석 후 합리적이라 판단되는 조건을 시스템에 고정으로 설정된 조건이다. 원하는 항목을 선택하여 사업체를 검색할 수 있다. 종사자, 출하액, 유형자산, 급여액, 주요 비용 등 각 규모별로 내검기준이 설정되어 있다(<표 3-1> 참조). 임의조건은 고정된 조건을 이용하여 검색하였을 경우 너무 많은 사업체가 검색되는 지역에서 기본고정조건을 변경하여 사용하는 조건이다.

<표 3-1>에서 해당항목의 지정된 규모의 사업체가 규모에 대응되는 증감률을 벗어날 경우 오류가 있으면 수정하고, 없으면 그 사유를 입력시스템에 입력한다. 예를 들면 종사자 수 500명 이상인 사업체가 전년대비 종사자 수 ±20%를 벗어날 경우 그 내용을 검토하는 것이다. 그러나 이러한 기준은 매 조사마다 변화가 없음을 가정한 것이다. 따라서 조사의 변화를 적절히 반영하는 내검기준의 설정이 필요하다. 한편 고정조건으로 검색 시 너무 많은 사업체가 검색될 수 있어 임의로 조건을 변경할 수 있게 시스템을 구성하고 있으나 이 시스템도 반복을 통해 그 결과를 확인하고 결정해야 하므로 좀 더 자료에 의해 판단할 수 있는 방안이 필요할 것으로 본다.

〈표 3-1〉 주요항목별 증감률 범위

항목	규모	내검기준 (증감률)	항목	규모	내검기준 (증감률)
종사자	500명 이상	±20%	급여액	500억 이상	±20%
	50명～500명	±30%		10억～500억	±30%
	30명～50명	±40%		3억～10억	±50%
	20명～30명	±50%		3억 미만	±500%
	10명～20명	±70%			
	10명 미만	±200%			
출하액	1조 이상	±20%	주요비용	1조 이상	±20%
	100억～1조	±30%		100억～1조	±30%
	50억～100억	±50%		50억～100억	±50%
	10억~50억	±100%		10억～50억	±100%
유형 자산	1,000억 이상	±20%			
	100억～1,000억	±30%			
	30억～100억	±40%			
	10억～30억	±50%			
	5억～10억	+100%			
	1억～5억	+500%			
출하액 대비 주요생산비	출하액 ±30% AND 주요생산비 ∓30%		종사자 수 대비 급여	종사자 수 +30% AND 급여 -30%	

자료: 2007년 기준 광업·제조업 조사 입력시스템 이용자 지침서

이 절에서는 2005년과 2006년 기준 광업·제조업 조사 자료에 EDA와 H-B기법을 적용하여 그 결과를 분석하고자 한다. 적용 자료는 이미 내검이 완료된 과거자료를 사용하였으므로 그 의미와 해석은 제한적이다. 그러나 이 자료의 분석결과를 토대로 적용기법의 유용성을 살피고자 한다. 적용될 주요항목으로는 종사자 수, 급여액, 출하액, 주요비용, 유형자산이다.

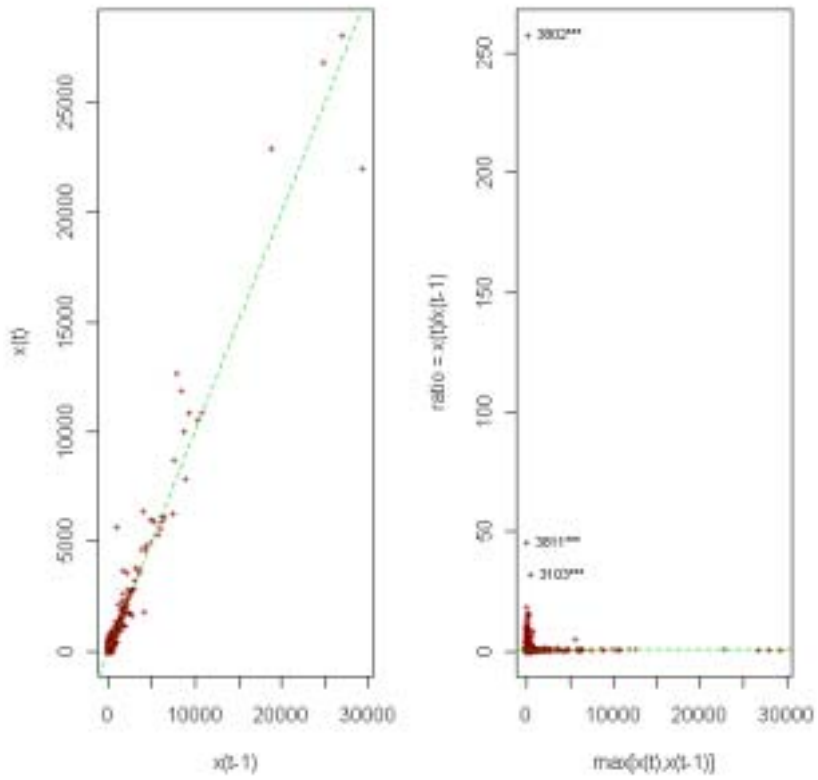
2006년과 2007년도의 광업·제조업 조사 자료에서 사업체 고유번호가 두 조사에 존재하는 경우에 자료를 병합하였다. 이 결과 102,406건의 사업체가 매칭되었다. 이 자료에서 종사자 수, 급여액, 출하액, 주요비용, 유형자산, 출하액과 주요생산비, 종사자 수와 급여액 항목의 두 시점 간 자료에 대해 각각 H-B기법과 그래픽 방법을 적용하였다. 본 연구의 분석은 통계 프로그램 R을 이용하여 작성하였다. R은 인터넷에서 무료로 내려 받아 사용할 수 있다(http://www.r-project.org/).

먼저 주요항목별로 현재년도와 과거년도의 두 자료 간 산점도와 비를 살펴본다. 종사자 수, 급여액, 출하액, 주요비용, 유형자산에 대해 차례로 분석한다. 각 항목에서 규모가 가장 큰 범위에 속하는 사업체에 대해 다시 한 번 산점도와 비 그래프를 통하여 특이치를 파악한다. 이후 특정 행정구역별 자료로 구분하여 살펴본다. 여기서는 ○○시 ○○구 사업체의 각 규모별로 주요항목에 대해 산점도를 작성하고 H-B 한계값을 통해 특이치를 검출한다.

1. 주요항목별 현·전 시점 자료 간 산점도와 비

가. 종사자 수

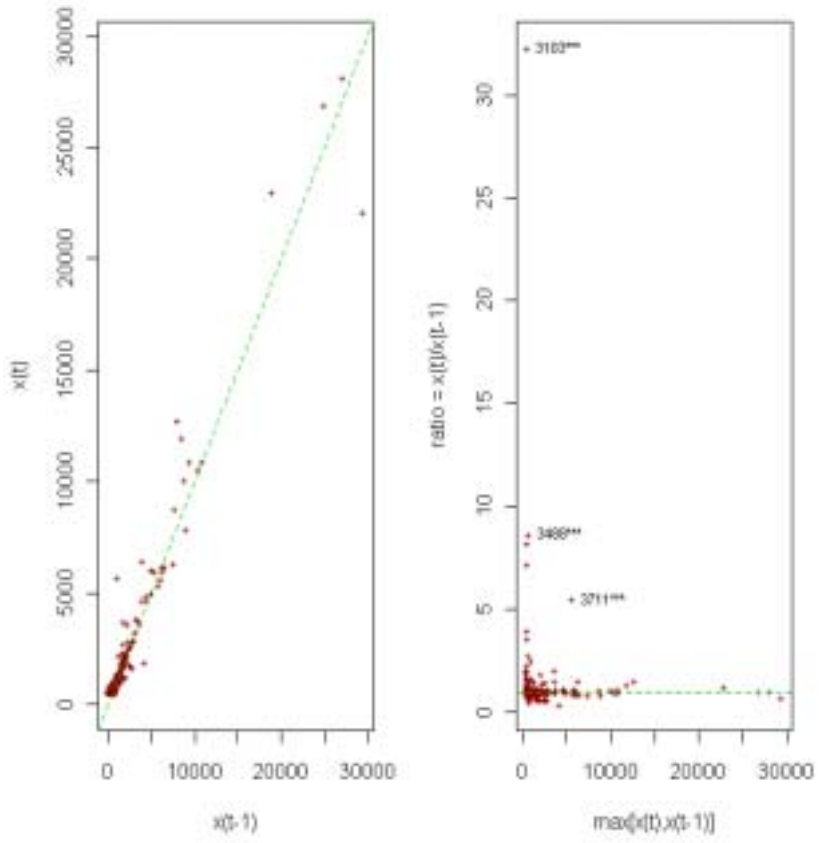
[그림 3-5]은 2005년도와 2006년도의 광업·제조업 조사에서 나타난 종사자 수의 산점도와 그 두 값의 비를 도식화한 것이다. [그림 3-5]의 왼쪽 그림이 2005년도와 2006년도의 종사자 수의 산점도이다. 그림에서 점들이 직선 근처에 흩어져 있어 2006년도의 종사자 수가 2005년도와 비교하여 그리 큰 변동이 없음을 나타내고 있다. 그러나 오른쪽 그림에서 보면 그들의 비가 매우 큰 점 들이 발견되고 있다. 이러한 비의 그림을 통해 보면 산점도에서 잘 볼 수 없는 특이점을 쉽게 발견할 수 있다.



[그림 3-5] 종사자 수 간의 산점도와 비

사업체고유번호가 3802***, 3811***, 3103***인 사업체는 종사자 수가 1년 사이에 큰 변화를 보이고 있다. 3802*** 사업체는 2005년도 종사자 수가 1명이었으나 그 다음해 258명으로 증가한 것으로 나타났고, 3811*** 사업체는 1명에서 46명으로, 3103*** 사업체는 17명에서 549명으로 각각 증가한 것으로 나타났다.

[그림 3-6]은 2006년도에 500명 이상 종사자 규모를 갖는 사업체로 국한한 그림이다. 왼쪽 그림은 2006년도의 종사자 수를 전년도의 종사자 수와 대응시킨 것이며 오른쪽 그림은 이들 규모에서의 사업체의 전년 비를 나타낸다.

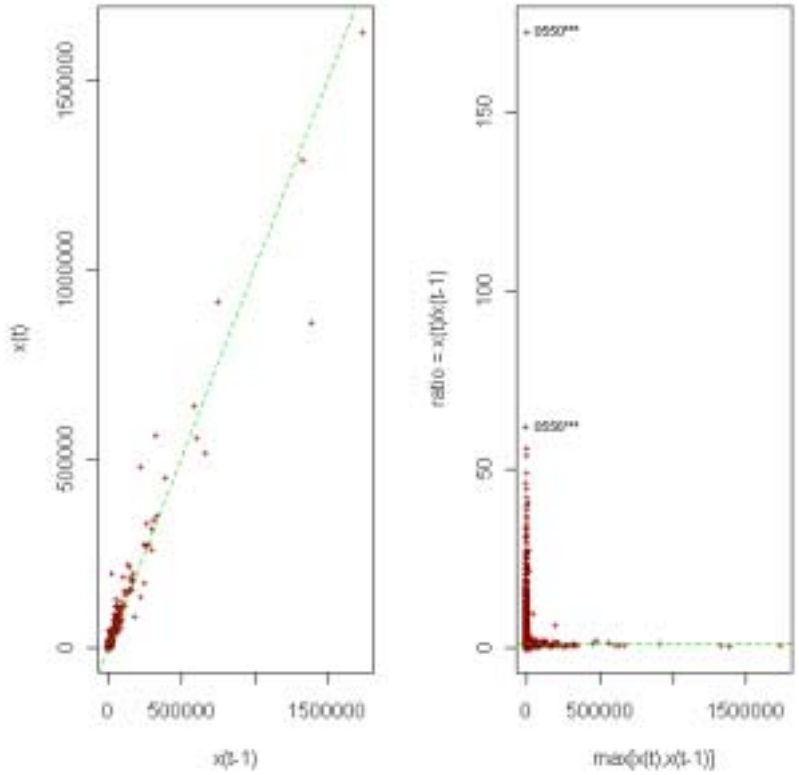


[그림 3-6] 종사자 수 간의 산점도와 비(2006년도 종사자 500명 이상)

3103*** 사업체 외 몇 개의 사업체가 전년도보다 두드러지게 많은 종사자 수가 집계되었음을 보여준다. 3103*** 사업체는 그림에서도 나타나고 있듯이 2006년도 종사자 수가 500명이 넘는 사업체로서 2005년도에 비해 30배를 초과한다. 3488*** 사업체는 약 9배의 증가를 보이고 있으며 3711*** 사업체는 5000명이 넘는 사업체로서 5배 이상의 증가를 나타내고 있다. 이와 같이 종사자 수를 규모별로 나누면서 그래프를 확인하면 다른 사업체와 확연히 구분되는 이상점들을 검색할 수 있다.

나. 급여액

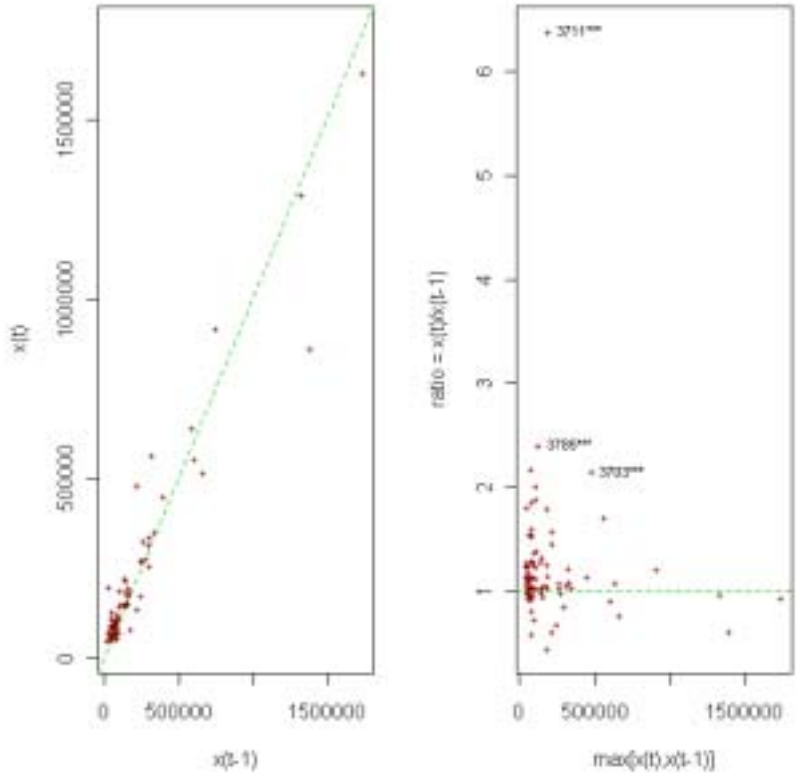
2005년도와 2006년도의 급여액수의 산점도와 두 값의 비를 [그림 3-7]에 나타내었다. 앞에서와 같이 [그림 3-7]의 왼쪽 그림은 급여액 간 산점도인데 대부분의 점들이 기울기가 1인 점선 근처에 나타나고 있다.



[그림 3-7] 급여액 간의 산점도와 비

한편, 오른쪽 그림에서 사업체고유번호 0550***인 사업체는 5백만 원에서 8억 6천4백만 원으로 급여액수가 1년 사이 약 170배보다 크게 증가한 것으로 나타났다. 또한 0556*** 사업체는 2백만 원에서 1억 2천 4백만 원으로 62배로 증가하고 있다.

[그림 3-8]의 좌측 그림은 2006년도에 500억 이상 급여액을 갖는 사업체들 중에서 2005년도와 2006년도 급여액의 산점도이다. 앞에서와 마찬가지로 우측 그림은 이들 사업체의 2005년도 급여액수와 2006년도 급여액수와의 비를 나타낸다.

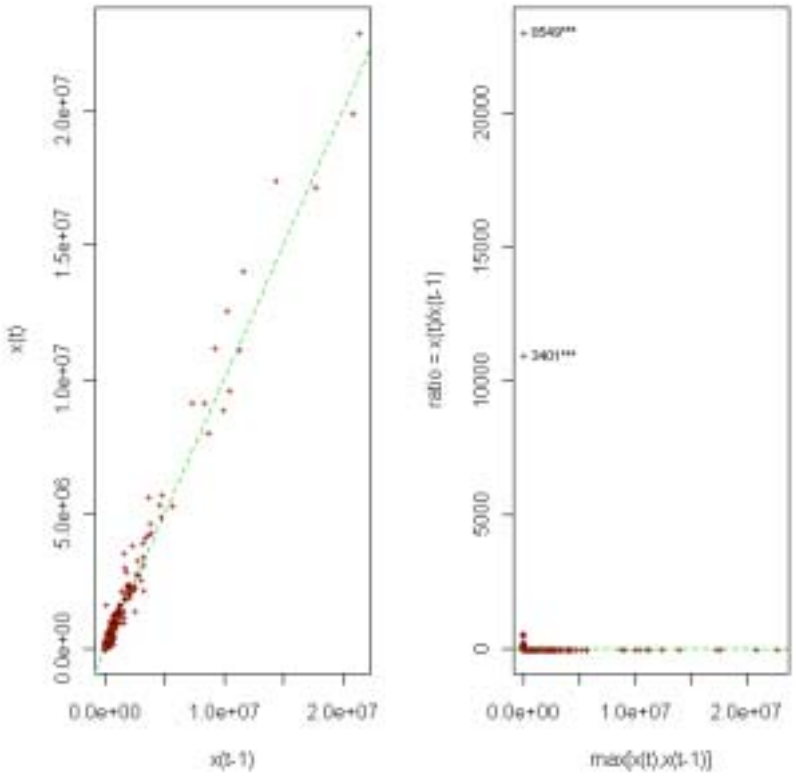


[그림 3-8] 급여액 간의 산점도와 비(2006년도 급여액 500억 이상)

이 중 3711*** 사업체가 2005년도에 31,247백만 원에서 2006년도에는 199,710백만 원으로 전년도보다 6배 많은 총 급여액을 보이고 있다. 한편 이 사업체는 앞에서 종사자 수가 5배 이상의 증가를 보이고 있으므로 급여액의 증가와 일치되고 있다.

다. 출하액

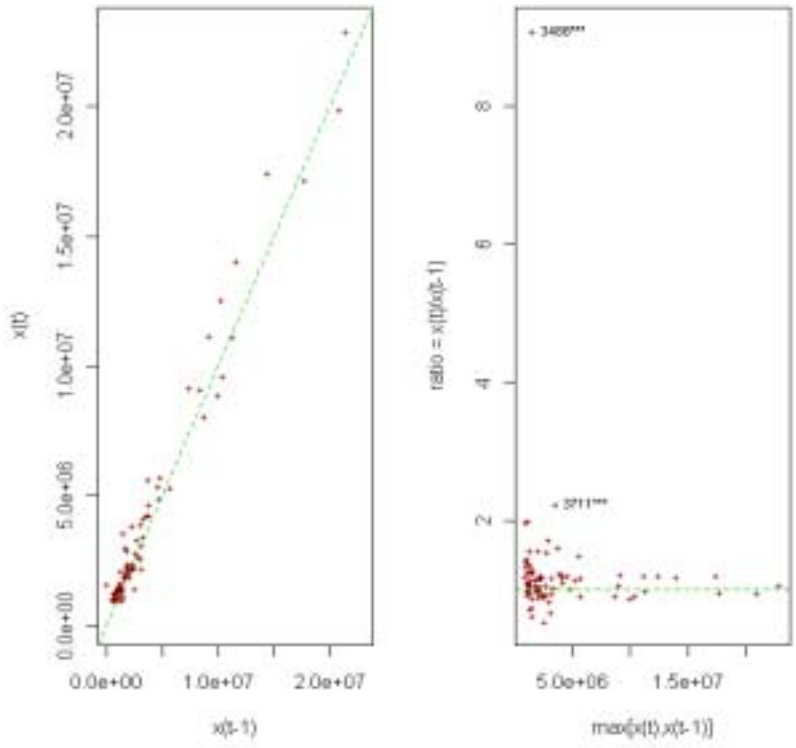
[그림 3-9]는 2006년도와 2005년도 출하액의 산점도와 두 값의 비에 대한 그림이다. 좌측 그림은 출하액 간 산점도로서 앞에서와 마찬가지로 점들이 기울기가 1인 점선 주위에 나타나고 있다.



[그림 3-9] 출하액 간의 산점도와 비

그러나 우측 그림에서 보듯이 사업체 고유번호 0549***과 3401*** 사업체는 각각 1백만 원에서 23,039백만 원으로, 1백만 원에서 10,966백만 원으로 전년대비 10,000배 이상의 증가를 보이고 있다. 이 두 사업체의 극심한 변화로 인해 나머지 사업체의 변화가 잘 나타나고 있지 않다.

[그림 3-10]은 2006년도에 1조 이상 출하액을 갖는 사업체를 대상으로 한 결과이다. 좌측이 출하액 간 산점도이며 우측이 과거 출하액과의 비를 보여준다.

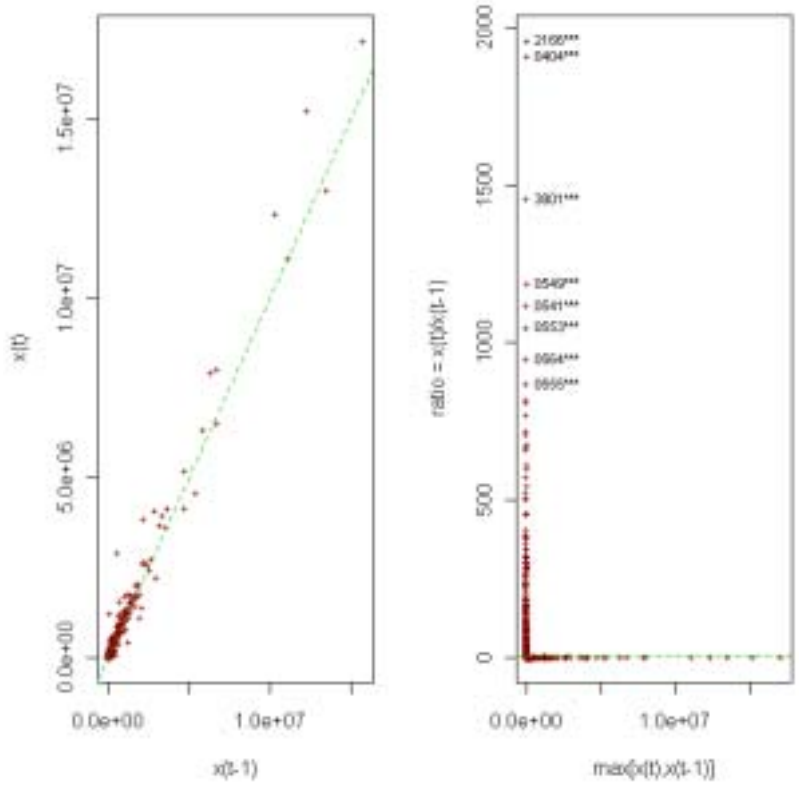


[그림 3-10] 출하액 간의 산점도와 비(2006년도 출하액 1조 이상)

특히 사업체 중 3488*** 사업체는 다른 사업체와 달리 2005년도에 비해 약 9배의 출하액(180,258→1,637,383백만 원)을 보이고 있다. 그러나 이 사업체는 앞에서 종사자 수가 약 9배의 증가를 보이고 있어 일관되어 보인다. 또한 3711*** 사업체는 약 2배의 출하액 증가를 보이고 있으나 이 사업체 역시 앞에서 종사자 수는 5배, 급여액은 6배 이상의 증가를 나타낸 사업체로서 2배의 출하액은 예상할 수 있는 증가로 보인다.

라. 주요비용

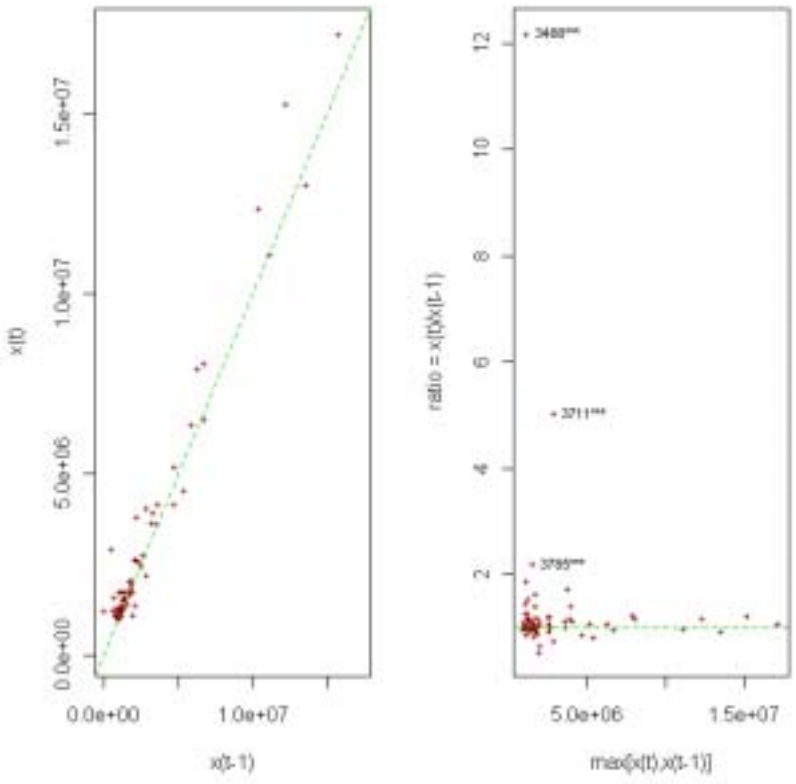
[그림 3-11]은 주요비용 항목에 대한 도표이다. [그림 3-8]의 좌측 산점도를 보면 주요비용도 전년도와 비교할 때 거의 비슷한 양상을 보인다. 그러나 우측 그림에서 보면 주요비용이 작은 규모의 사업체에서 전년도에 비해 상당히 많은 변동을 보이고 있음을 알 수 있다.



[그림 3-11] 주요비용 간의 산점도와 비

따라서 주요비용이 작은 규모의 사업체에서 좀 더 세세하게 살펴볼 필요가 있을 것이다. 그러나 본 절에서는 각 항목에서 가장 영향력이 큰 규모의 사업체에 대해 살펴보기로 한다.

[그림 3-12]는 주요비용이 1조 이상인 사업체를 대상으로 하여 그린 산점도와 과거 주요비용과의 비를 보여준다.

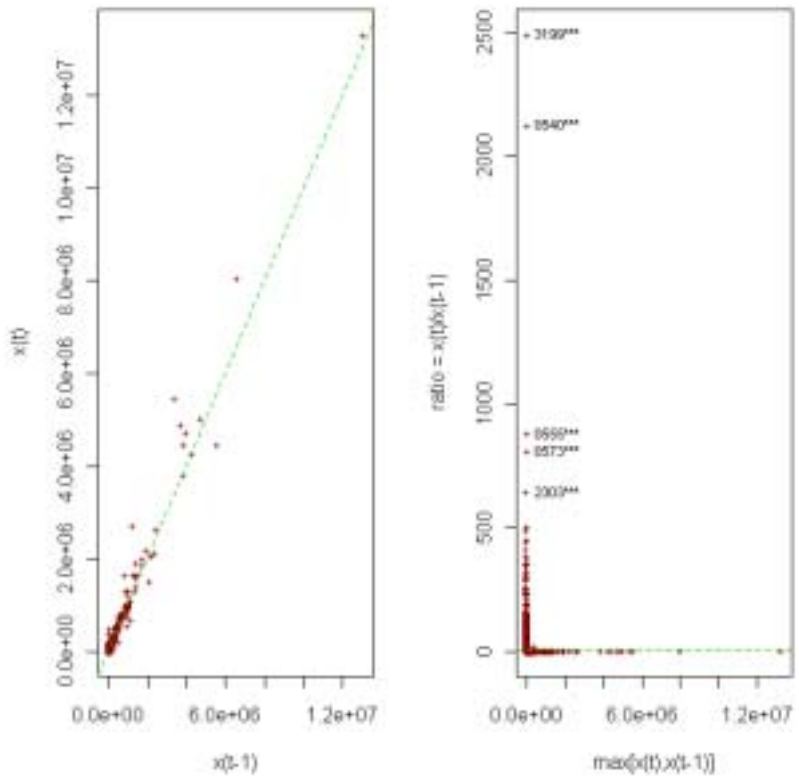


[그림 3-12] 주요비용 간의 산점도와 비(2006년도 주요비용 1조 이상)

2006년도에 주요비용이 1조 이상으로 조사된 사업체의 전년도 출하액과의 비를 보면 3488*** 사업체가 다른 사업체와는 매우 다르게 2005년도에 비해 약 12배의 주요비용의 증가를 보이고 있으며 3711*** 사업체도 약 5배의 증가를 보이고 있다. 그러나 앞에서 3488*** 사업체와 3711*** 사업체는 출하액에서도 그 증가가 두드러지게 나타난 사업체로서 주요비용의 증가가 예상되는 사업체이다. 따라서 두 개의 사업체는 모두 항목 간 일치된 결과를 보이고 있다.

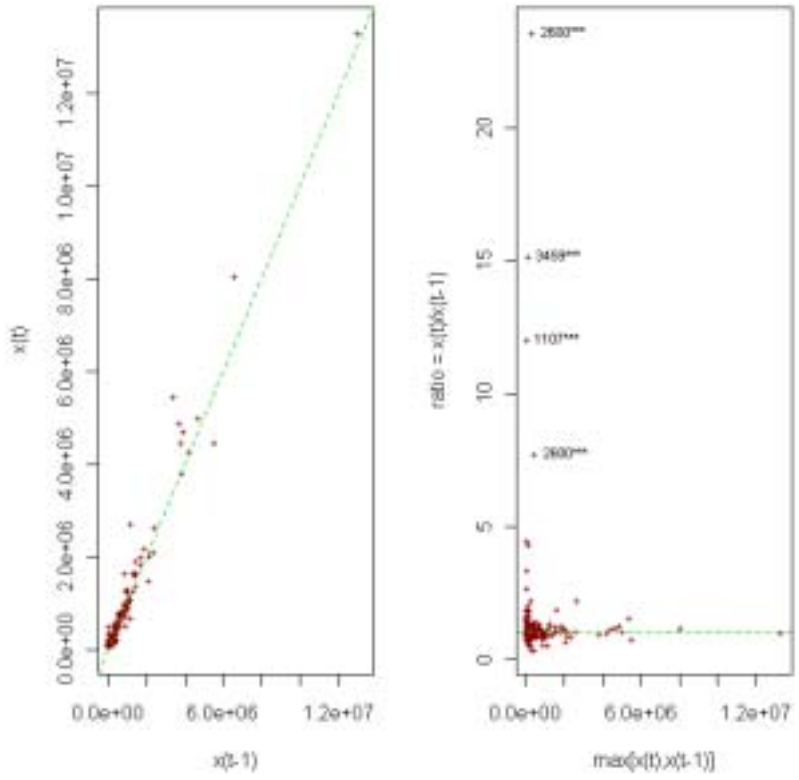
마. 유형자산

[그림 3-13]의 산점도에서 보면 유형자산도 전년도와 비슷한 값이 조사되고 있음이 확인된다. 그러나 그들의 비를 나타낸 그림을 보면, 3199***와 0540*** 사업체는 각각 2백만 원에서 4,991백만 원으로, 1백만 원에서 2,130백만 원으로 유형자산이 크게 증가하였음을 보여주고 있다. 이는 작은 규모의 유형자산을 보유한 사업체에서보다 큰 변화들이 존재하고 있음을 나타낸다. 따라서 유형자산을 규모별로 도식화하여 살펴볼 필요가 있다.



[그림 3-13] 유형자산 간의 산점도와 비

여기서는 2006년도에 유형자산이 1,000억 이상을 보유한 사업체를 살펴보기로 한다. [그림 3-14]는 2005년도와 2006년도 유형자산 간 산점도와 그들의 비를 나타낸다.



[그림 3-14] 유형자산 간의 산점도와 비(2006년도 유형자산 1,000억 이상)

2006년도에 유형자산이 1,000억 이상으로 조사된 사업체의 과거 유형자산과의 비를 보면 2600***, 3459***, 1107***, 2600*** 사업체가 2005년도에 비해 약 5배 이상 증가된 유형자산액을 보이고 있어 다른 사업체와는 확연히 구분되고 있다. 이들 중 2600*** 사업체는 16,170백만 원에서 381,489백만 원으로 약 24배의 증가를 나타내고 있으며 1107*** 사업체는 9,643에서 116,373백만 원으로 약 12배 증가하고 있다.

2. 특정지역의 현·전 시점자료 간 산점도와 특이치 검출

앞에서는 전국자료의 각 주요항목에 대하여 과거년도와의 비교를 통해 특이점을 검색하였으나 여기서는 특정 행정구역별 자료를 선택하여 살펴보고자 한다. 주요항목에 대한 산점도와 비에 대한 도표는 각 행정구역별로 적용될 수 있다.

하나의 예로 ○○시 ○○구 사업체를 각 규모별로 주요항목에 대해 산점도를 작성하고 앞 절에서 소개된 H-B 한계값을 그림 위에 제시하였다. H-B 한계값을 계산 시 사용자가 설정하는 승수 c_1 과 c_0는 각각 6과 3으로 고정하였으며 이는 각 규모에 따라 조정할 수 있다.

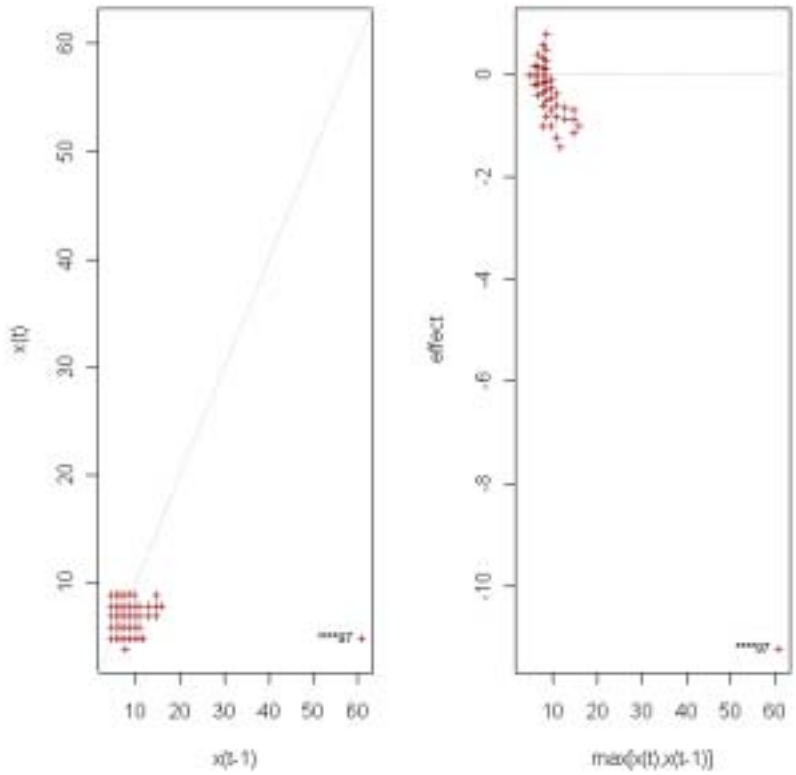
앞의 전국자료에서는 현·전 시점 자료의 비를 도식화하였으나 여기서는 효과 $e_i = s_i[\max(x_i, y_i)]^u$ 의 값을 세로축으로 (현·전 시점 자료 중 큰 값을 가로축으로) 하여 도식화하였다. 이는 비(ratio)의 값이 1보다 크면 과거와 비교해 증가한 것으로, 1보다 작으면 감소한 것을 의미하나 1을 중심으로 대칭이 되지 않는다. 예를 들어 과거자료의 값이 150이고 현재자료의 값이 50이라면 비는 0.333이고 현재자료의 값이 150이고 과거자료의 값이 50이라면 비는 3이다. 즉 같은 100의 변화를 갖더라도 비의 값은 1을 중심으로 대칭이 되지 않는다. 그러나 변화된 비인 효과 e_i는 0에 대해 좌우 대칭이 되도록 한다.

효과 $e_i = s_i[\max(x_i, y_i)]^u$ 에서 $u = 0$일 경우는 $e_i = s_i$가 된다. 본 절에서는 변환된 비의 값으로부터 H-B 한계값을 계산하여 그림 위에 제시하였으며, 필요시 $u = 1$인 경우, 즉 s_i 값에 $\max(x_i, y_i)$를 곱한 값들로부터 구한 H-B 한계선을 제시하였다. 앞에서도 언급한 바와 같이 $u = 1$인 경우에는 해당 항목이 규모가 큰 값일수록 상대적으로 작은 변화에도 특이값으로 검색되도록 하는 효과를 갖는다.

다시 한 번 언급하지만, 본 절의 적용자료는 이미 내검이 완료된 자료이므로 자료 자체에서 그 의미를 찾기보다는 H-B 기법에 의한 한계값의 설정과 산점도 등의 그래픽을 이용한 내검기법의 유용성 측면에서 바라보아야 할 것이다.

가. 종사자 수

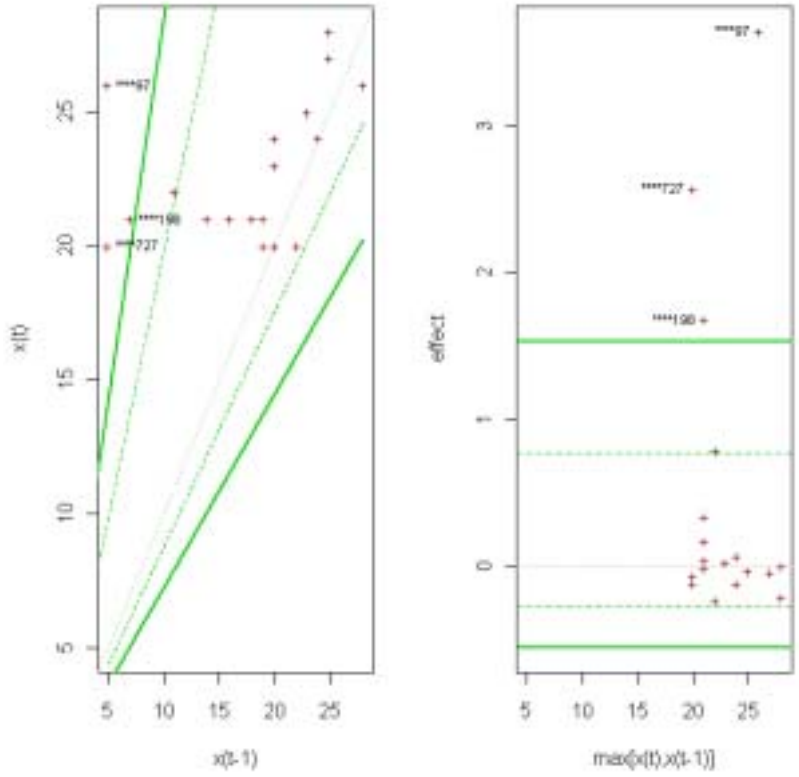
[그림 3-15]는 ○○시 ○○구에서 2006년도 종사자 10명 미만 사업체 중 2005년 종사자 수와의 산점도와 그 비(ratio)의 변형된 값들을 도식화 한 것이다. 그림에서 쉽게 확인할 수 있듯이 ****97 사업체는 다른 사업체와 달리 매우 동떨어져 있음을 알 수 있다. 이 사업체는 2005년도에 61명에서 2006년도에 5명으로 줄어든 경우이다.



[그림 3-15] 특정지역 종사자 수간 산점도와 변환비
(2006년도 종사자 수 10명 미만)

한편, 종사자 수가 10명 미만인 사업체의 경우 전년도와 비교하여 종사자가 줄어든 경우가 증가한 경우보다 다소 많이 나타나고 있다.

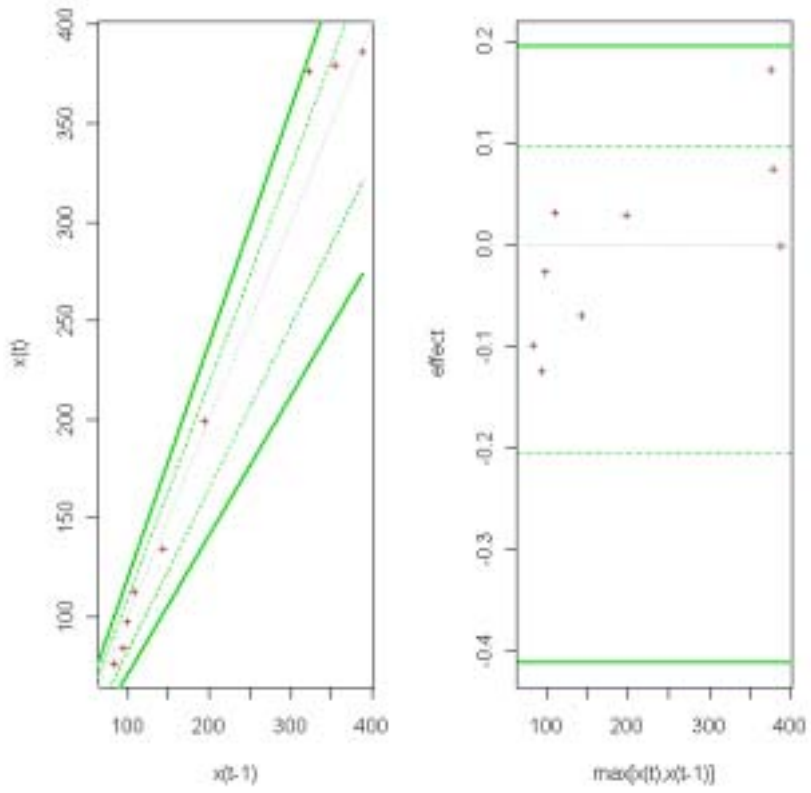
[그림 3-16]은 20명 이상 30명 미만 사업체의 경우이다. 몇 개의 사업체가 한계선 밖에 위치하고 있어 검토가 필요한 사업체로 볼 수 있다. 나머지 사업체는 한계선 내에서 같은 수준이거나 다소 증가한 것으로 나타나고 있다. 이 사업체 규모에서는 10명 미만인 경우와는 달리 대부분 증가한 것으로 나타나고 있다.



[그림 3-16] 특정지역 종사자 수간 산점도와 변환비
(2006년도 종사자 수 20명 이상 30명 미만)

특히 전년도에 비해 3배 이상 증가한 사업체(****97, ****727, ****198)가 상한선 밖에 존재하고 있다. 이 중 ****97 사업체는 5에서 26명으로, ****727사업체는 5에서 20명으로 증가한 사업체이다.

[그림 3-17]은 ○○시 ○○구에서 종사자가 50명 이상 500명 미만의 사업체의 경우이다. [그림 3-17]에서는 점들이 모두 굵은 한계선 내에 포함되고 있어 대부분의 사업체와 뚜렷하게 구분되는 관측값은 없는 것으로 보인다.

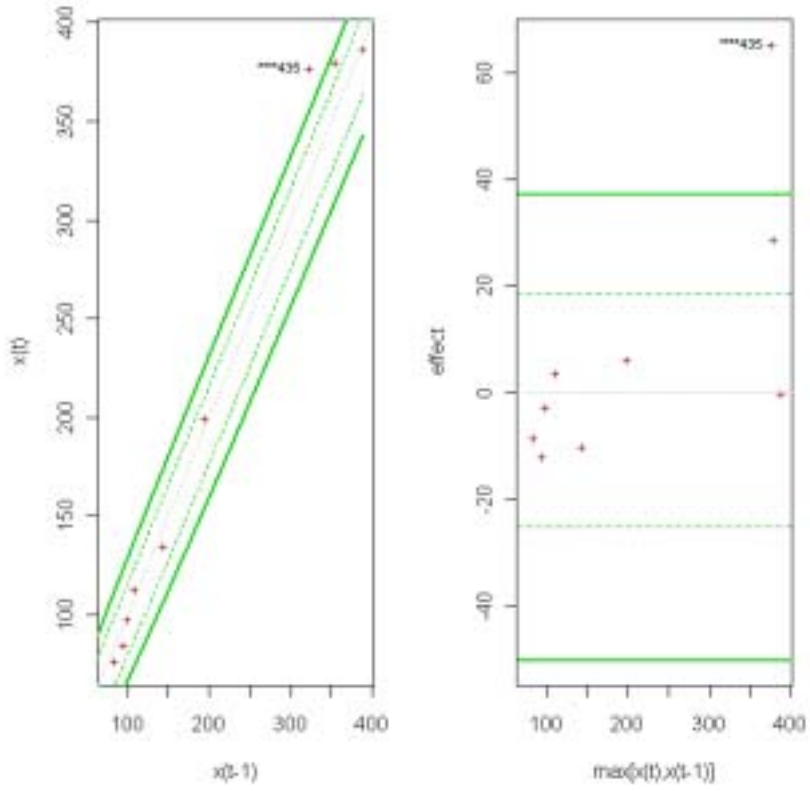


[그림 3-17] 특정지역 종사자 수간 산점도와 변환비
(2006년도 종사자 수 50명 이상 500명 미만)

여기서 제시된 [그림 3-17]에서의 한계선은 $e_i = s_i[\max(x_i, y_i)]^u$ 에서 $u = 0$일 경우로부터 구해진 H-B 한계선이다. [그림 3-18]에는 $u = 1$인 경우의 한계선을 표시하였다. 앞에서도 언급한 바와 같이 $u = 1$인 경우에는 규모가 큰 값일수록 작은 변화에도 특이값으로 검색되도록 하

는 효과를 주게 된다.

종사자 규모를 고려한 [그림 3-18]의 한계선에서는 종사자 규모가 큰 ****435 사업체가 한계선 밖으로 나가 있어 종사자 규모가 큰 경우에는 특이치를 민감하게 검출할 수 있음을 알 수 있다. 즉 앞에서 검출되지 않은 사업체가 규모를 반영하여 검출된 결과를 나타낸다. 이 사업체는 323명의 사업체가 377명으로 증가한 경우이다.

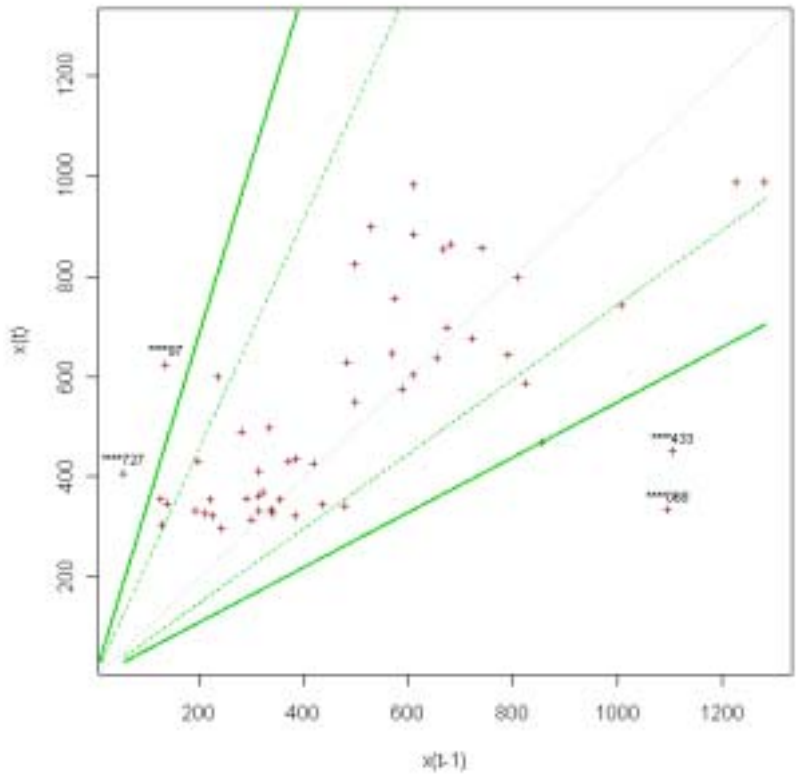


[그림 3-18] 특정지역 종사자 수간 산점도와 변환비($u = 1$인 경우)
(2006년도 종사자 수 50명 이상 500명 미만)

나. 급여액

[그림 3-19]는 ○○시 ○○구 광업・제조업 사업체 중 2006년도 급여

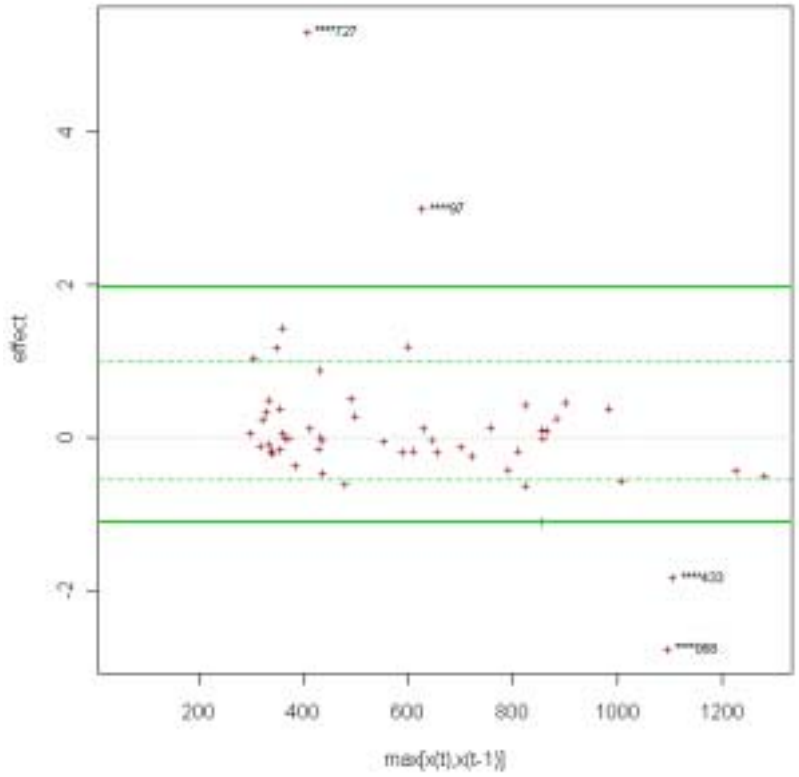
액이 3억 이상 10억 미만인 사업체의 경우 현·전 급여액 간 산점도와 그 위에 H-B 한계선을 제시한 그림이다.



[그림 3-19] 특정지역 급여액 간 산점도와 한계선
(2006년도 급여액 3억 이상 10억 미만)

그림에서 보듯이 ****727 사업체는 56백만 원에서 408백만 원으로, ****97 사업체는 136백만 원에서 627백만 원으로 각각 상한선 밖에 있다. 이들 사업체는 앞에서 종사자 수가 5~6배로 증가한 경우로 서로 일치하고 있다. 한편 ****433 사업체는 1,107백만 원에서 452백만 원으로, ****068 사업체는 1,096백만 원에서 336백만 원으로 감소하여 하한선 밖에 나타나고 있다.

[그림 3-20]은 $e_i = s_i[\max(x_i, y_i)]^0 = s_i$으로 규모를 반영하지 않은 해당비의 변환 값을 도식화한 것이다. 2005년도와 2006년도의 2개의 급여액 중 가장 큰 값만을 가로축에 나타내어 하나의 값이 생략되었지만 [그림 3-19] 보다 특이치를 검출하기 용이하다.

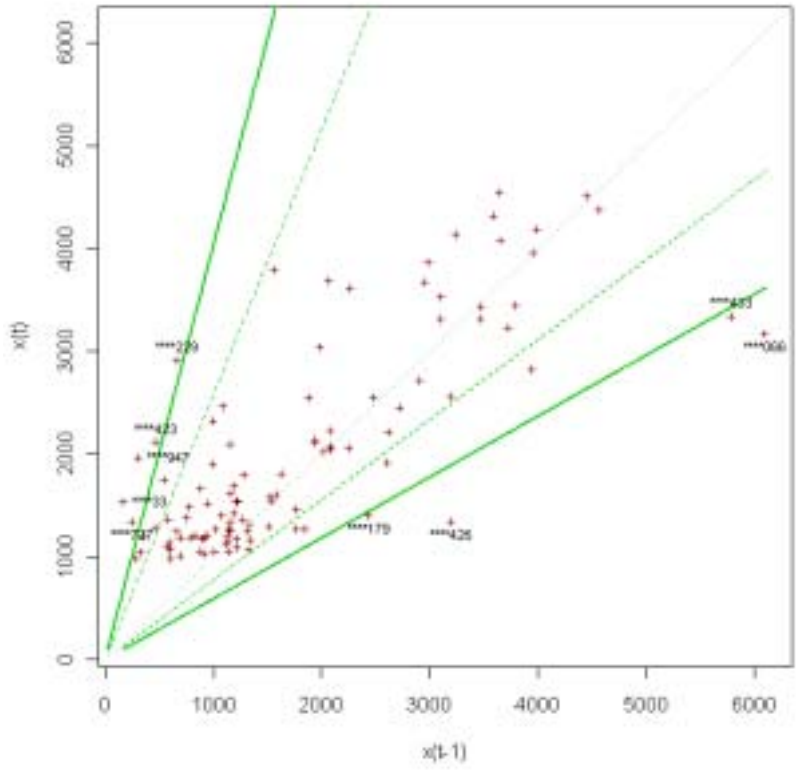


[그림 3-20] 특정지역 급여액 간 변환비
(2006년도 급여액 3억 이상 10억 미만)

역시 ****727 사업체가 가장 큰 차이를 보이고 있으며, ****97 사업체도 과거에 비해 특이하게 증가하였음을 나타내고 있다. ****433 사업체와 ****068 사업체는 2005년도에 비해 다른 사업체와는 구분될 정도로 감소하여 하한선 밖에 나타나고 있다.

다. 출하액

[그림 3-21]은 2006년도 출하액이 10억 이상 50억 미만인 사업체에서의 출하액 간 산점도와 H-B 한계선을 나타낸다.

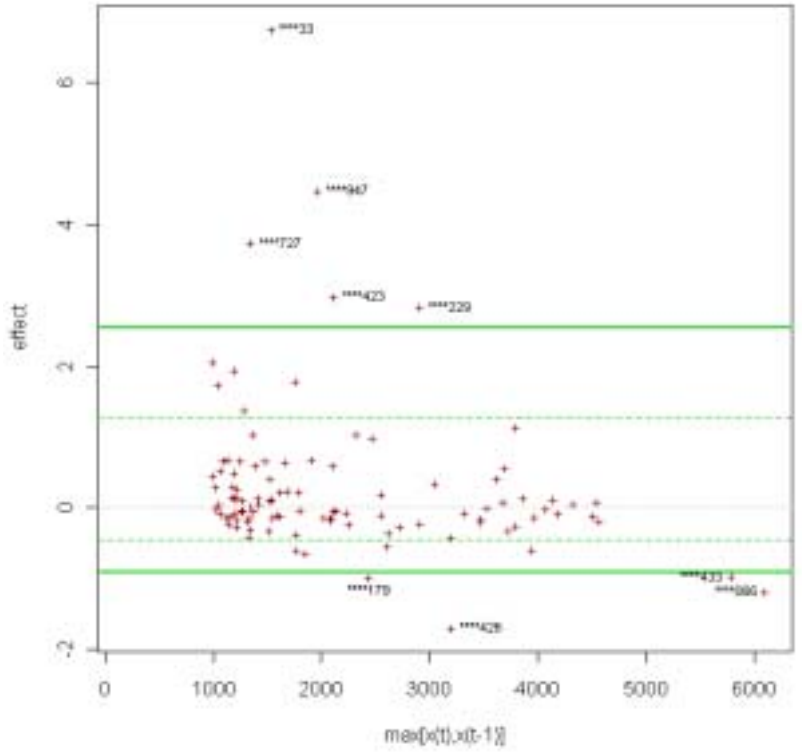


[그림 3-21] 특정지역 출하액 간 산점도와 한계
(2006년도 출하액 10억 이상 50억 미만)

앞의 그림과 유사하게 5개의 사업체(****33, ****947, ****727, ****423, ****229)가 상한선 밖에 있다. ****33 사업체는 175백만 원에서 1,541백만 원으로 증가한 것으로 나타났다. 사업체 ****727은 종사자 수, 급여액의 증가와 더불어 출하액도 증가한 것으로 나타나 일관성을 띠고 있다. 또 4개의 사업체(****433, ****066, ****179, ****426)가 하한선 밖

에 있으며 ****433 업체는 급여액에서도 급감을 보이고 있는 사업체이다.

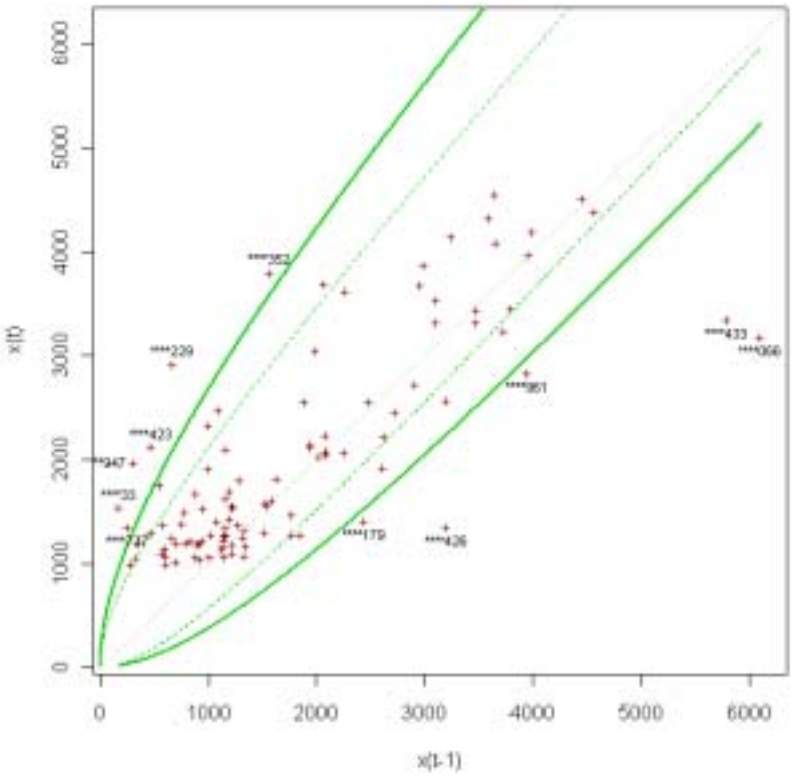
[그림 3-22]는 해당되는 비의 변환 값($e_i = s_i$)을 도식화한 것이다. 변환 비는 0을 중심으로 좌우대칭이 되도록 한 값이다. 역시 앞의 그림에서보다도 특이치가 쉽게 눈에 띈다.



[그림 3-22] 특정지역 출하액 간 변환비
(2006년도 출하액 10억 이상 50억 미만)

그림에서 ****33, ****947, ****727, ****423, ****229의 5개 사업체가 상한선 밖에 있어 과거년도에 비해 다른 사업체와는 확연히 다르게 증가한 것을 쉽게 확인할 수 있으며 ****179, ****433, ****066, ****426 4개의 사업체가 다른 사업체와 구분되어 하한선 밖에 있어 감소하고 있음을 나타낸다.

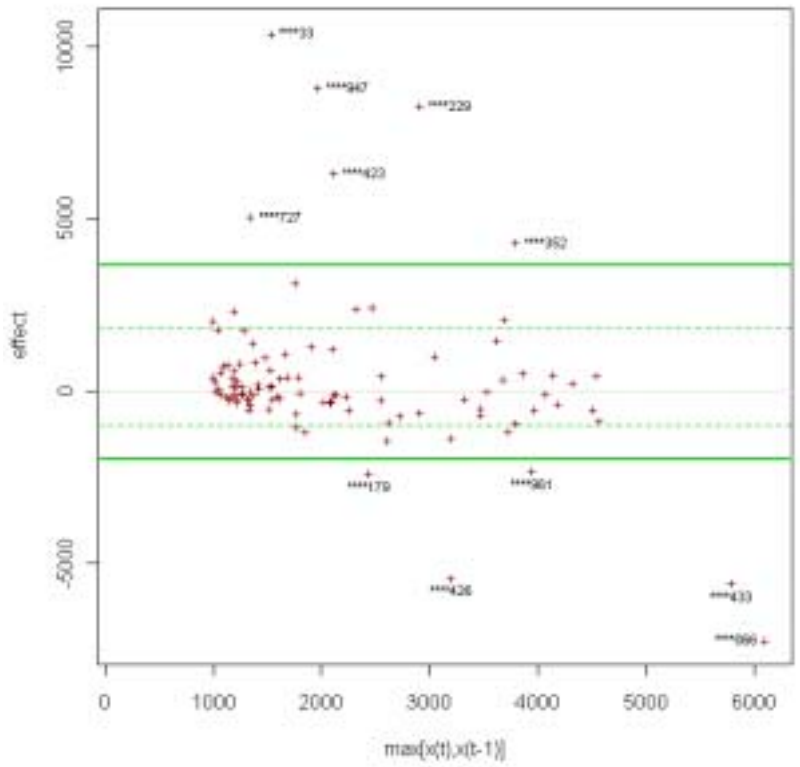
[그림 3-23]은 [그림 3-21]의 산점도 위에 $u=0$인 경우로부터 계산된 H-B 한계선 대신에 $u=1$인 경우의 한계선을 적용한 결과이다. 한계선은 더 이상 직선이 아닌 곡선으로 표현되며 규모가 커질 때 하한과 상한을 좁혀 줌으로써 이상치가 민감하게 검출되도록 한다.



[그림 3-23] 특정지역 출하액 간 산점도와 한계선($u=1$인 경우)
(2006년도 출하액 10억 이상 50억 미만)

이 결과로 ****33, ****947, ****727, ****423, ****229 사업체외 ****352가 상한선 밖에 위치해 특이치로 구분된다. 반면 ****179, ****433, ****066, ****426 사업체 외에 ****961 사업체가 하한선 밖에 있어 특이치로 구별되고 있다. 따라서 규모를 고려하여 좀 더 민감하게 특이치를 구별하고자 하는 경우에 이 방법을 사용할 수 있을 것이다.

[그림 3-24]는 $u=1$인 경우로서 $e_i = s_i[\max(x_i, y_i)]$을 세로축으로 하고 있다. 이는 규모를 반영한 해당비의 변환 값을 도식화한 것이다. 변환 비가 0을 중심으로 좌우대칭이 되도록 한 것은 앞에서와 같으나 변환된 비의 값에 출하액 중 최대값이 곱해져서 규모가 클수록 큰 효과 값을 가지게 된다.

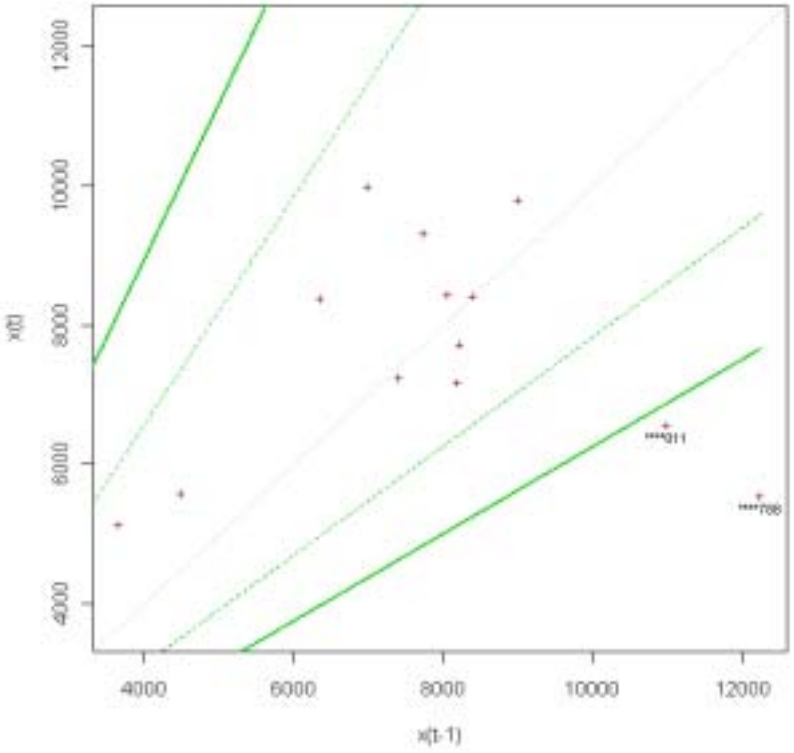


[그림 3-24] 특정지역 출하액 간 변환비($u=1$인 경우)
(2006년도 출하액 10억 이상 50억 미만)

[그림 3-24]은 [그림 3-22]과 유사하게 나타나고 있지만 ****33, ****947, ****727, ****423, ****229의 5개 사업체 외에 ****352 사업체가 상한선 위로 존재하고 ****179, ****433, ****066, ****426 사업체 외에

****961 사업체가 하한선 아래에 있음을 좀 더 쉽게 확인할 수 있다.

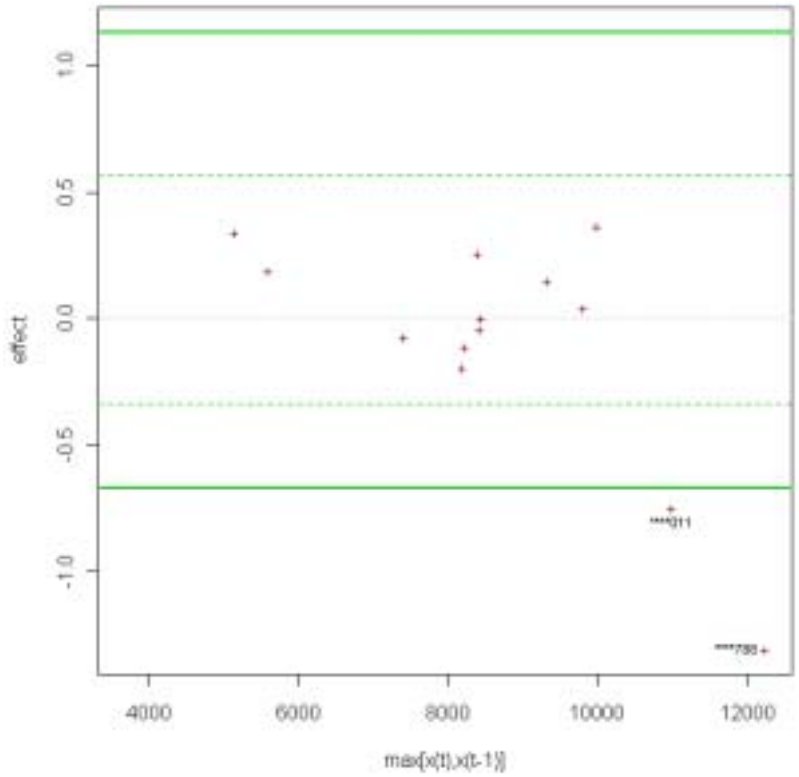
[그림 3-25]는 앞과 같은 항목인 출하액에 대한 도표이다. 이는 2006년도 출하액이 50억 이상 100억 미만인 사업체에 대하여 출하액 간 산점도와 H-B 한계선을 그 위에 제시한 것이다.



[그림 3-25] 특정지역 출하액 간 산점도와 한계선
(2006년도 출하액 50억 이상 100억 미만)

그림에서 대부분의 사업체가 가운데 점선 근처에 흩어져 있으며 한계선 내에 있어 전년도에 비해 큰 변동이 없는 것으로 보인다. 그러나 그림에서 볼 수 있듯이 두 개의 사업체(****011, ****788)는 하한선 밖에 놓여 있으며 이들은 전년도에 100억 이상이었던 출하액이 절반으로 줄어든 것을 확인할 수가 있다.

[그림 3-26] 역시 전년도와의 비를 변환하여 0을 중심으로 대칭되도록 나타낸 것이다. 비(ratio)들의 중위수로부터 상대적인 차이를 나타내는 것으로 효과 값 0은 비의 중위수를 나타낸다.

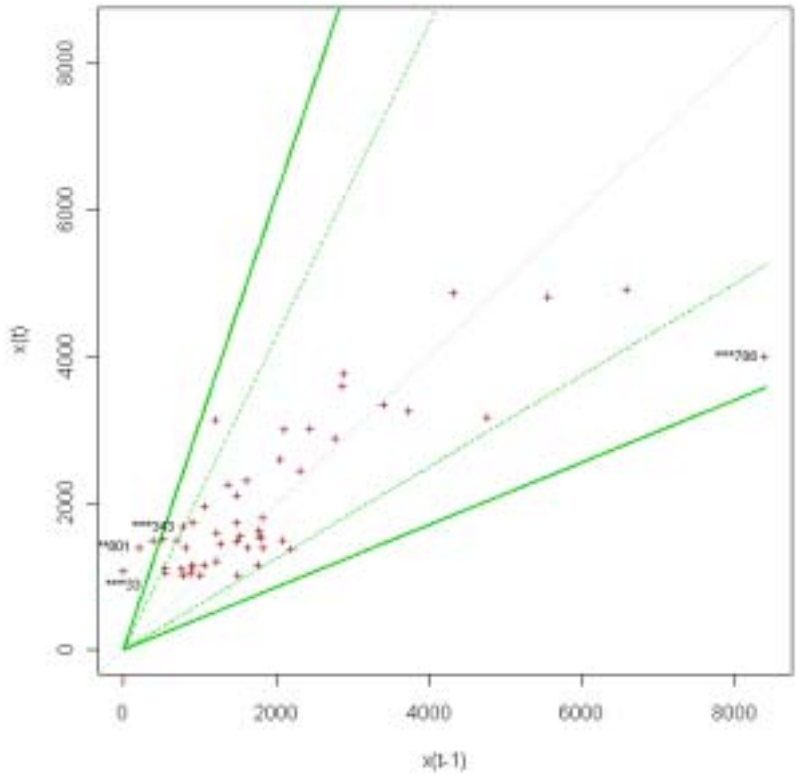


[그림 3-26] 특정지역 출하액 간 변환비
(2006년도 출하액 50억 이상 100억 미만)

대부분의 사업체가 유사한 패턴을 보이나 [그림 3-26]에서와 같이 두 사업체(****011, ****788)가 다른 사업체와는 달리 떨어져 하한선 밖에 위치하고 있다.

라. 주요비용

[그림 3-27]은 주요비용 항목에 대한 도표로 2006년도 주요비용이 10억 이상 50억 미만인 경우를 대상으로 하여 2006년도와 2005년도의 주요비용 간 산점도와 H-B 한계선을 표시하고 있다. 대부분의 사업체가 주요비용에서 전년도와 큰 차이를 보이고 있지는 않으나 몇 개의 사업체에서는 전년도와 큰 차이를 보이고 있다

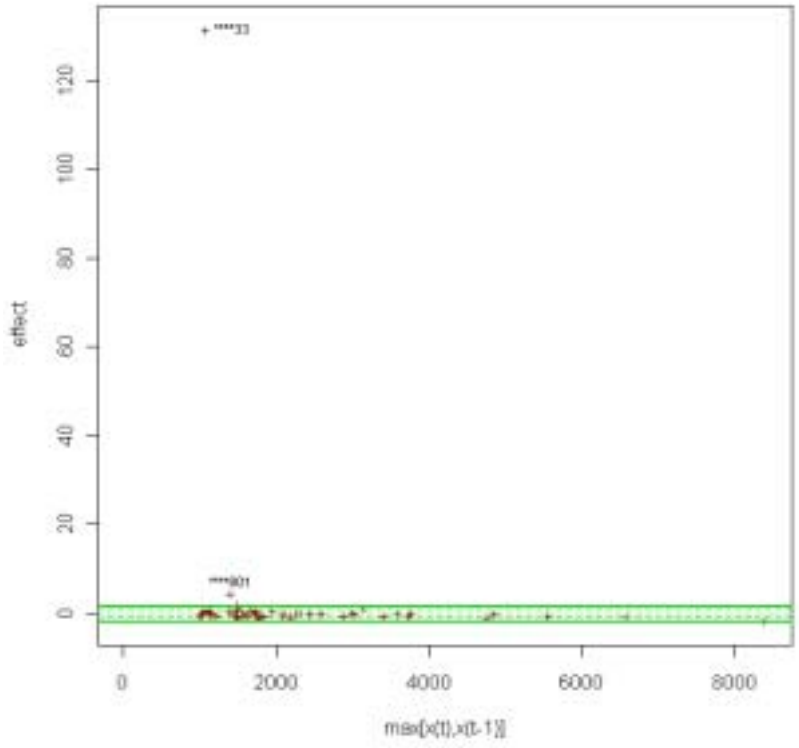


[그림 3-27] 특정지역 주요비용 간 산점도와 한계선
(2006년도 주요비용 10억 이상 50억 미만)

특히 ****733 사업체는 주요비용이 약 10억으로 전년도 7백만 원에 비해 150배가량 증가한 액수이다. 한편 이 사업체의 출하액은 1억 7천만 원에서 15억 원으로 증가한 것으로 조사되어 전년도의 주요비용에

대한 검토가 필요할 것으로 보인다.

[그림 3-28]은 $u=0$인 경우, 즉 규모를 고려하지 않은 효과 값을 세로축에 나타내어 변화가 상대적으로 큰 값을 검출하고 있다. 대부분의 사업체가 주요비용에서 전년도와 큰 차이를 보이고 있지는 않으나 몇 개의 사업체에서는 전년도와 큰 차이를 보이고 있다.

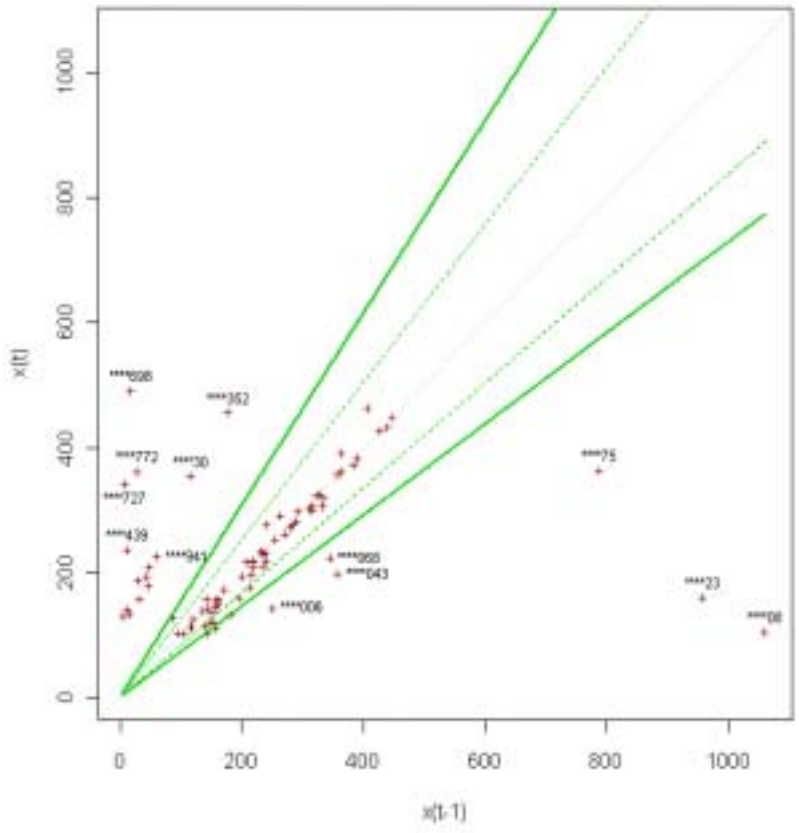


[그림 3-28] 특정지역 주요비용 간 변환비
(2006년도 주요비용 10억 이상 50억 미만)

****33 사업체는 앞의 그림에서와는 달리 뚜렷하게 특이치로 구분되고 있으며 ****801 사업체도 검출되고 있다. 그러나 앞의 그림에서 ****343 사업체 역시 특이치로 구분되나 ****33 사업체의 변화가 너무 커서 이 사업체는 표현되지 못하고 있다.

마. 유형자산

[그림 3-29]는 유형자산 항목에 대한 도표이다. 특정지역의 2006년도 유형자산이 1억 이상 5억 미만인 경우에만 국한하여 2006년도와 2005년도의 유형자산 간 산점도와 한계선을 나타내었다.

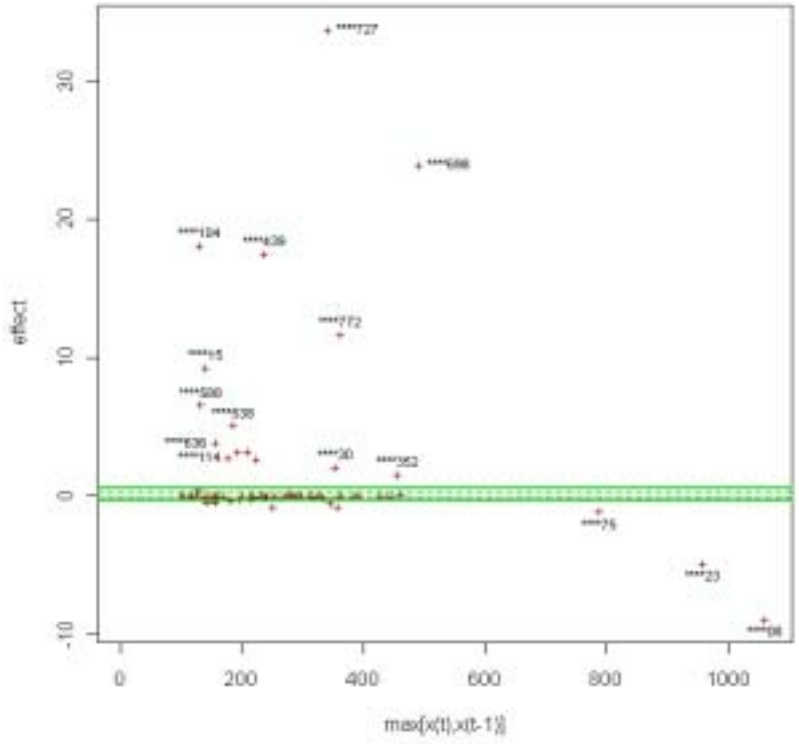


[그림 3-29] 특정지역 유형자산 간 산점도와 한계선
(2006년도 유형자산 1억 이상 5억 미만)

대부분의 사업체가 이전년도와 비교하여 안정적으로 움직이고 있는 반면 유형자산에 있어서 큰 변동을 보이고 있는 사업체가 확연히 구분

되어 포착되고 있다.

[그림 3-30]은 효과값을 세로축으로 하여 H-B 한계선을 나타낸 그림으로서 앞의 그림보다 특이치를 간결하게 구분하고 있다. 이 규모의 유형자산에 있어서는 다소 많은 특이치가 존재하고 있으며 과거에 비해 증가한 경우가 감소한 경우보다 많이 나타나고 있음을 알 수 있다.



[그림 3-30] 특정지역 유형자산 간 변환비
(2006년도 유형자산 1억 이상 5억 미만)

이상에서 우리는 특정지역 하나를 선택하여 주요항목별로 살펴보았다. 특히 관심 있는 규모에 대해 각 사업체의 특이치를 산점도와 H-B 방법에 근거한 한계선을 이용하여 검출하고자 하였다. 따라서 그 밖의 지역에 대해서도 유사한 방법으로 이를 적용할 수 있을 것이다.

제5절 결론

이상에서 살펴본 H-B기법에 의한 한계값 설정은 사용자가 경험에 의해 설정하는 승수가 포함되어 있어 역시 주관적이기는 하나 자료의 분포와 자료에 내재된 움직임을 함께 반영한 결과로서 중요한 의미를 갖는다. 즉 경기변동이나 특수 지역, 품목 상황이 검색대상 자료로부터 이상치 탐색에 반영될 수 있다. 따라서 내검의 판단기준이 증감률 ±50%와 같이 더 이상 대칭적으로 나타나지 않는다.

H-B 방법에서 규모가 커질수록 작은 변화를 감지할 수 있도록 하였던 한계선 조정방법($u=1$)은 각 규모별로 나누어 보는 경우에는 규모를 고려하지 않은 방법($u=0$)과 그다지 큰 차이가 없는 것으로 나타났다. 그러나 구분된 각 규모 안에서도 주어진 자료의 크기가 크면 작은 변동에도 좀 더 잘 검출되도록 하고자 할 때 이 방법을 사용할 수 있을 것이다.

H-B 기법에 의한 특이값 구분의 단점은 살펴본 바와 같이 도표를 통해 보완될 수 있다. 즉 한계값을 가이드라인으로 삼고 도표를 통해 나타난 시각적 인지력을 통해 자료들이 어떻게 산포되어 있는지를 동시에 고려하면 어떤 자료를 얼마나 내검해야 할지를 결정하는 데 유용할 것으로 본다. 더 나아가 화면에 출력된 이상치를 클릭할 경우 바로 그 레코드의 구체적인 값들을 볼 수 있도록 구성한다면 더욱 손쉬운 작업이 진행될 수 있을 것이다.

한편, 본 연구에서 적용된 자료는 이미 내검이 완료된 자료이므로 적용방법의 의미가 잘 드러나지 않을 수도 있다. 예를 들면 내검전 자료 적용 시 단위착오로 인해 유사한 패턴이 다른 값과 동떨어져 나타날 수 있음을 시각적으로 확인할 수도 있을 것이다. 따라서 내검 전 자료에 이상의 방법을 적용하여 오류패턴을 감지하고 내검량이 적정하게 설정될 수 있도록 한계값을 설정할 수 있다.

이상과 같이 EDA와 H-B 기법을 통해 이상치 검출을 용이하게 할 수 있으며 좀 더 능동적으로 내검규칙을 설정할 수 있을 것이다. 또한 시각적인 내검을 통해 내검작업의 효율성을 높일 수 있을 것으로 기대한다. 이 방법은 통계청에서 주기적으로 실시되고 있는 다른 사업체 연간조사

에도 유사하게 적용할 수 있을 것이다.

본 연구에서 작성된 프로그램은 단기적으로는 산업통계과에서 종합내검 시스템 운영 시 전체 또는 시도별 광업・제조업 사업체에 대한 내검에 참고적으로 사용할 수 있을 것이다. 프로그램 R은 PC로 내려 받아 이미 작성된 프로그램의 명령어를 입력하여 간단하게 사용할 수 있다. 장기적으로는 그래픽내검을 기존 종합내검시스템과 연계하여 보다 효율적인 종합내검 시스템을 구축할 필요가 있다고 판단된다. 즉 그래픽에서 패턴과 움직임을 파악하고 이상치를 클릭 시 기존 종합내검정보가 나타날 수 있도록 구성된다면 본청 뿐 아니라 지방청에서도 매우 유용하게 사용될 것이다. 따라서 기존 시스템을 연계하고 화면분할, 마우스 작동, 컬러링, 디자인 등 이용자가 편리하게 사용할 수 있는 기능의 추가와 관리를 위해서는 전문적인 시스템 개발자의 참여가 필요하다.

한편, 지역별, 규모별 외에 산업분류별, 주요품목별로 이상치를 검색한다면 품목특성이나 산업특성에 따라 그들의 패턴과 움직임을 이해할 수 있고 이를 근거로 내검량을 설정할 수 있을 것이다. 특히 월간 동향조사에서는 자료가 시계열적으로 나타나고 있으므로 시계열 변동을 고려하여야 하는 바, 과거의 시계열 움직임을 통해 이상치 점검을 할 수 있을 것으로 보인다. 또한 여러 개의 항목들을 동시에 고려하여 이상치를 검출하는 다변량자료의 이상치 검출방법에 관한 연구와 정상적인 범위 안에 있으나 자료가 다른 대부분의 자료와 달리 동떨어진 경향을 보이는 인라이어(inlier) 자료의 검출에 관한 연구도 추후 연구해야할 과제 중 하나이다.

참고문헌

통계청(2008), 「2007년 기준 광업·제조업 통계조사 조사지침서」.

통계청(2008), 「2007년 기준 광업·제조업 통계조사 입력시스템 이용자 지침서」.

Atkinson, D.(2000), "Developing a State-of-the-art Editing and Imputation System for NASS' Agricultural Censuses and Sample Surveys", UN/ECE Work Session on Statistical Data Editing, (Cardiff, United Kingdom, 18-20 October 2000), Topic III: New techniques and tools for editing imputation.

Banff Support Team(2007), "Functional Description of the Banff System for Edit and Imputation", Generalized System Methods Section, Business Survey Methods Division.

Bienias, L. Julia, D.M. Lassman, S.A. Scheleur, and H. Hogan(1994), "Improving Outlier Detection in Two Establishment Surveys", UN/ECE Work Session on Statistical Data Editing.

Desjardins, D. and P. Liars(2001), "New Graphical EDA+(EDA Plus) Techniques for Understanding Data", In Proc. SAS User's Group International Conference (SUGI26), Long Beach, CA.

Engström, P. and C. Ängsved(2005), "A Graphical Macro-Editing Application", UN/ECE Work Session on Statistical Data Editing.

Esposito, R., J.K. Fox, D. Lin, and K. Tidemann(1994), "ARIES: A Visual Path in the Investigation of Statistical Data", Journal of Computational and Graphical Statistics, Vol. 3, No. 2, pp. 113-125.

Hidiroglou, M.A. and J.M. Berthelot(1986), "Statistical Editing and Imputation for Periodic Business Surveys", Survey Methodology, 12, pp.73-84.

Houston, G. and A.G. Bruce(1993), "gred : Interactive graphical editing for business surveys", journal of official statistics vol.9.no.1, 1993, pp. 81-90.

Todaro, T. and K. Perritt(1999), "www.fcsm.gov/99papers/todaro.html",

National Agricultural Statistics Services.
Weir, P.(1994), “The Graphical Editing Analysis Query System”, UN/ECE Work Session on Statistical Data Editing.