# 식스시그마 품질개선 단계에서 GLM 회귀분석의 이해와 적용

# Application and Understanding of Regression Analysis in the Quality Improvement Activities

최 성 운*
Sung-woon Choi*

## Abstract

The study presents the application strategy and understanding of regression analysis with GLM(Generalized Linear Model) unifying with other statistical techniques such as correlation analysis and design of experiment(DOE). The quidelines proposed in this paper can be used for practioners to implement GLM and ANOVA(Analysis of Variance) for the DMAIC 5 steps of six sigma breakthrough.

**Keywords: Regression Analysis, Guidelines, GLM, DOE, ANOVA, DMAIC**

* 경원대학교 산업공학과

# 1. 서 론

식스시그마 프로젝트(Project)와 품질분임조 과제(Task) 수행 시 객관적 데이터를 효율적으로 표현하는 과학적인 도구가 통계적 기법이다. 대표적인 통계적 기법으로는 가설검정, 추정(Estimation), 관리도(Control Chart), 상관분석(Correlation Analysis), 회귀분석(Regression Analysis), 분산분석(ANOVA, Analysis of Variance), 직교계획(Orthogonal Design) 등이 있다. 여기서 기술통계량과 확률분포(Probability Distribution)를 기초로 한 검정, 추정, 샘플링검사(Acceptance Sampling), 상관, 회귀분석을 통계적 품질관리(SQC, Statistical Quality Control)라 하고 관리도, 공정능력분석(Process Capability Analysis), 측정시스템분석(MSA, Measurement System Analysis)을 통계적 공정관리(SPC, Statistical Process Control)라 한다. 분산분석과 $k^n$ 요인 및 부분계획(Factorial and Fractional Factorial Design), 반응표면분석(Response Surface Analysis), 혼합물 계획(Mixture Design), 타구치 로버스트 계획(Taguchi's Robust Design)을 실험계획법(DOE, Design of Experiment)이라 한다. 이 외에 표나 그림에 의해 수치 및 언어 데이터를 효율적으로 표현하는 QC 7가지 도구(QC 7 Tools), 신 QC 7가지 도구(New QC 7 Tools)가 있으며 FMEA(Failure Mode Effect Analysis), Why-Why 등의 설비 신뢰성(Reliability) 기법이 있다.

식스시그마 DMAIC 혁신 프로세스 중 Define 단계에서는 관리도, 공정능력분석, Measure 단계에서는 MSA, 관리도, 공정능력분석, 특성요인도, 회귀분석[1, 4-7, 9], Analyze 단계에서는 검정, 추정, 분산분석[5, 8], 상관분석, 회귀분석, Improve 단계에서는 분산분석, 회귀분석 직교계획[3], Control 단계에서는 관리도 등의 통계적 기법[2]을 프로젝트의 성격에 따라 적절히 선택하여 사용한다.

이러한 통계적 기법 중 회귀분석은 개선 프로세스의 여러 단계에서 사용되는 기본적인 도구인 데도 불구하고 일부 전문가와 기업 실무자의 이해 부족으로 잘못 적용되는 경우가 빈번하게 발생한다.

따라서 본 연구에서는 분산분석과 회귀분석의 차이점과 보완적 설계에 의한 분석방법, 상관분석과 회귀분석의 적용에서의 차이점, 단계별(Stepwise) 회귀분석과 범주형(Categorical) 회귀분석의 적용방법 등을 제시한다. 또한 식스시그마 DMAIC 혁신(Innovation) 5단계 중 Analyze, Improve 단계, 품질분임조 QC Story 개선(Improvement) 15단계 중 현상파악, 대책실시 단계 등에서 실무자를 위한 적용가이드라인을 제안한다.

# 2. 분산분석과 회귀분석의 차이점과 보완적 설계

## 2.1 분산분석과 회귀분석의 차이점

분산분석(Analysis of Variance)은 고객이 요구하는 스펙에 대한 특성치 데이터가 만족하는 제품기술 및 생산기술 인자 수준(Factor Treatment)의 영향여부를 검정하고

최적 수준을 추정하는 방법이다. 이 분석은 인자수준내의 오차는 작고 수준간의 평균 차가 클 경우 인자가 특성치에 영향을 주었다는 유의적인 $H_1$가설의 채택을 하게 된다. 1원배치법의 Effect Model $y_{ij}=\mu+I_i+e_{ij}$에서 $\mu$는 Overall Mean, $I_i$는 $i$th Treatment Effect 일 경우 SS(Sum of Squares)의 독립성을 이용한 Cochran의 정리를 사용하여 분석한다. $H_0$ : $I_i=0$, $H_1$ : $I_i\neq 0$의 가설에서 $H_0$채택일 경우 $Y_{ij}=\mu+e_{ij}$로 수평선에 가까우며 $H_1$채택일 경우 $y_{ij}=\mu+I_i+e_{ij}$로 $I_i$의 크기만큼 수평선을 기준으로 높낮이가 형성된다. 즉 분산분석은 기하학적으로 수평선을 기준으로 높낮이의 여부를 알아보는 통계적 방법이다.

회귀분석(Regression Analysis)은 기하학적 관점에서 수평선의 높낮이를 보는 소극적 분산분석과 다르게 특성치 데이터($y$)와 인자수준($x$)간에 직선, 포물선 또는 일정한 그림의 함수관계(Functional Relationship)가 성립한다.

따라서 분산분석의 그래프에서 함수관계가 성립할 경우 이는 회귀분석으로 실시해야 한다. 이 경우 대부분의 전문가와 실무자는 분산분석을 위해 <표1>과 같이 특정 인자수준 내에서 반복된 평균 데이터만을 이용하기 때문에 $(x,y)$의 함수관계 파악을 위한 회귀분석 시 데이터의 수는 수준수 $\ell$개밖에 되지 않아 유의성 판정의 신뢰성에 의문이 생길 수 있다. 일부 교재에서는 <표1>과 같은 데이터 배열에서 분산분석과 회귀분석을 동시에 실시하기를 권고하고 있고 일부 실무자가 이를 식스시그마의 Analyze, Improve 단계에서 잘못 적용하고 있다. 회귀분석의 경우 반복을 취하지 않은 적어도 30개 이상의 수준 $x$와 특성값 $y$의 대응 데이터가 준비되어 있어야 한다.

<표1> 1원배치법 분산분석과 회귀분석 데이터 배열

| 인자수준 / 반복 | $A_1$ | $A_2$ | … | $A_\ell$ | |
|---|---|---|---|---|---|
| 1<br>2<br>⋮<br>$m$ | $y_{ij}(i=1,2,\cdots,\ell, j=1,2,\cdots,m)$ | | | | |
| 평균 | $\overline{y_{i.}}$ | | | | $\overline{\overline{y}}$ |
| 회귀 | $\overline{\hat{y}_{i.}}$ | | | | |

분산분석은 $y_{ij}=\mu+I_i+e_{ij}$에서 Cochran의 정리에 의하여 *Total SS = Error SS + Model SS*로 나타내며 $\sum_i\sum_j(y_{ij}-\overline{\overline{y}})^2=\sum_i\sum_j(y_{ij}-\overline{y_{i.}})^2+\sum_i\sum_j(\overline{y_{i.}}-\overline{\overline{y}})^2$이고 자유도는 각각 $(\ell m-1)$, $(\ell m-\ell)$, $(\ell-1)$이 된다.

회귀분석은 1차 회귀식 $\overline{\hat{y}_{ij}}=\hat{\beta}_i x_i+e_{ij}$에서 *Total SS = Residual SS +Regression SS*로 나타내며 이를 식으로 표현하면 $\sum_i\sum_j(y_{ij}-\overline{\overline{y}})^2=\sum_i\sum_j(y_{ij}-\overline{\hat{y}_{i.}})^2+\sum_i\sum_j(\overline{\hat{y}_{i.}}-\overline{\overline{y}})^2$

이고 자유도는 각각 $(\ell m-1)$, $(\ell m-2)$, $(2-1=1)$이다. 여기서 $Residual\ SS = Pure\ Error\ SS + Lack\ of\ Fit\ SS$이고 $\sum_i\sum_j(y_{ij}-\widehat{y_{i.}})^2=\sum_i\sum_j(y_{ij}-\overline{y_{i.}})^2+\sum_i\sum_j(\overline{y_{i.}}-\widehat{y_{i.}})^2$이며 자유도는 $(\ell m-2)$, $(\ell m-\ell)$, $(\ell-2)$이다. 분산분석을 위한 <표1>과 같은 상태하에서 회귀분석을 실시할 경우 장점은 잔차오차를 순수한 오차와 적합결여 오차로 구분해서 볼 수 있다는 것이고 단점은 $x$의 수준수가 $\ell$개밖에 되지 않아 회귀분석을 위한 최소 데이터의 요건이 성립되지 않는다는 것과 $x$ 행렬의 독립성과 직교성의 결여로 회귀계수의 신뢰성에 문제가 있을 수 있다.

따라서 이러한 단점을 피하기 위해 $k^n$요인 배치법, RSM(Response Surface Method) 등의 실험계획(DOE : Design of Experiment)에서는 인자수준 $x$ 행렬에 대해 직교계획의 성질을 활용하여 2.2절과 같이 분산분석과 회귀분석의 보완적 분석을 통해 최적 설계 방법을 추구한다.

## 2.2 분산분석과 회귀분석의 보완적 설계

DOE에서는 ANOVA 분산분석과 GLM 회귀분석을 보완적 관점에서 동시에 사용하는데 ANOVA는 인자의 유의성을 판정하는 경우 유용하며 GLM은 1차 $x_1$, $x_2$는 선형직선, 2차 교호작용 $x_1x_2$는 틀어진 곡률(Distorted, Twisted Curvature), 2차 $x_1^2, x_2^2$은 봉우리(Peak) 등을 나타내며 최적 인자수준의 선정 시 2차원 등고선(Contour Plot)과 3차원 표면도(Surface Plot) 등의 기하학적 시각을 이용할 수 있다.

DOE에서는 ANOVA와 GLM 분석을 위한 $x$의 Design Matrix의 실험설계가 매우 중요하며 효율적인 분석을 위해 Low Level, Hight Level의 2수준 실험을 실시하며 2차항 이상의 분석이 필요한 경우 중심점(Center Point : CT) 또는 축점(Axial Point)을 추가한다.

$x$의 직교성(Orthogonality)에 의해 GLM 회귀분석 시 $(x'x)^{-1}$의 대각요소 $D_{kk}$를 이용한 $t_0=\widehat{\beta_k}/\sqrt{\sigma^2 D_{kk}}$검정시 분산을 작게 하고 $x$의 독립성을 유지하여 회귀계수 $\widehat{\beta_j}$의 추정치의 BLUE(Best Linear Unbiased Estimator) 성질을 유지하게 한다.

직교성을 이용한 DOE에는 $2^k$ FD(Factorial Design), $2^{k-p}$ FFD(Fractional Factorial Design), RSM의 CCD(Central Composite Design), BBD(Box-Bhenken Design), FCD(Face-Centered Design), PBD(Placket-Burman Design), MD(Mixture Design) 등이 있다.

본 연구에서는 직교계획 관점에서의 Design Matrix와 기하학적 좌표를 ANOVA, GLM의 대수적 분석 방법과 연계하여 실무자 적용 가이드라인을 제시한다.

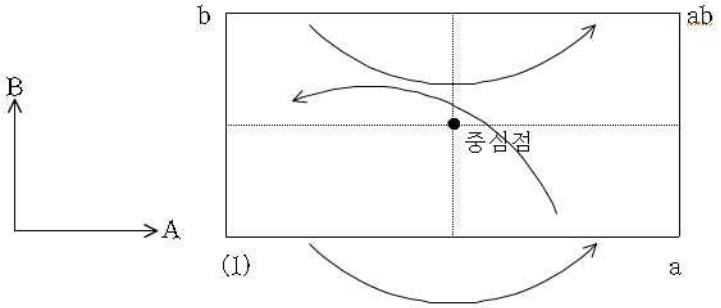
### 2.2.1 ANOVA와 GLM 동시 분석을 위한 직교 Design Matrix

#### 1) $2^2$ FD Design Matrix

MINITAB에서 Treatment Combinations에 의한 실험 순서(Run, Standard Order,

Yate's Order)는 <표1>과 같은 Geometrical View를 가리며 Design Matrix는 <표2>와 같다. <표2>에서 $I$는 $\beta_0$를 위한 것으로 인자의 총 평균(Grand Average of Factor)을 나타내며 -는 Low Level, +는 High Level이고 열간 직교성을 지닌다.

<표1> $2^2$FD Geometrcal View



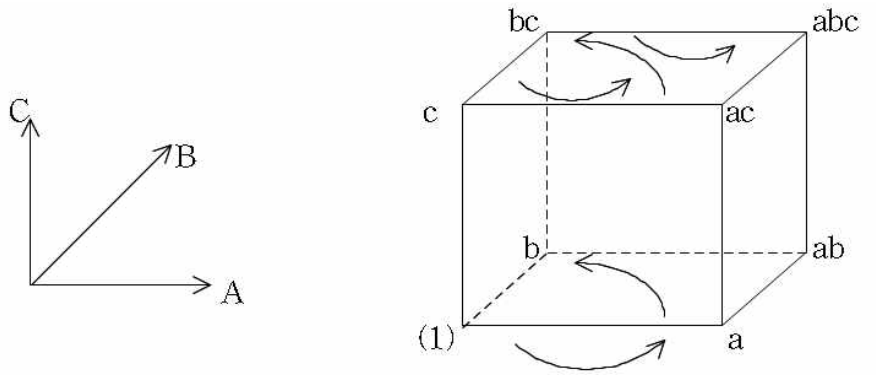

<표2> $2^2$FD Design Matrix

| 실험순서 | ANOVA | 인자효과 | $I$ | $A$ | $B$ | $AB$ | 중심점 | 반복(Replicate) |
|---|---|---|---|---|---|---|---|---|
| | GLM | 회귀계수 | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_{12}$ | | |
| | | $x$ | | $x_1$ | $x_2$ | $x_1x_2$ | | 데이터 $y$ |
| (1) | | | + | − | − | + | 0 | |
| a | | | + | + | − | − | 0 | |
| b | | | + | − | + | − | 0 | |
| ab | | | + | + | + | + | 0 | |

### 2) $2^3$FD Design Matrix

MINITAB에서 Treatment Combination에 의한 Geometrical View와 Design Matrix는 <표3><표4>와 같다.

<표3> $2^2$FD Geometrical View



<표4> $2^3$FD Design Matrix

| 실험 순서 | ANOVA | 인자효과 | $I$ | $A$ | $B$ | $AB$ | $C$ | $AC$ | $BC$ | $ABC$ | 중심점 | 반 복 데이터 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GLM | 회귀계수 | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_{12}$ | $\beta_3$ | $\beta_{13}$ | $\beta_{23}$ | $\beta_{123}$ | | |
| | | $x$ | | $x_1$ | $x_2$ | $x_1x_2$ | $x_3$ | $x_1x_3$ | $x_2x_3$ | $x_1x_2x_3$ | | $y$ |
| (1) | | | + | − | − | + | − | + | + | − | 0 | |
| a | | | + | + | − | − | − | + | + | + | 0 | |
| b | | | + | − | + | − | − | + | − | + | 0 | |
| ab | | | + | + | + | + | − | − | − | − | 0 | |
| c | | | + | − | − | + | + | − | − | + | 0 | |
| ac | | | + | + | − | − | + | + | − | − | 0 | |
| bc | | | + | − | + | − | + | − | + | − | 0 | |
| abc | | | + | + | + | + | + | + | + | + | 0 | |

### 3) $2^{k-p}$ FDD Design Matrix

<표4>의 $2^3$FD에서 Alias 정의대비 $I=ABC$에 의해 2개로 Block화된 $2^{3-1}$FDD는 Block 1이 $ABC$열에서 부호가 음수인 4개의 처리조합 (1), ab, ac, bc 이고 Block 2는 $A=BC$, $B=AC$, $C=AB$ 등의 Alias의 관계를 이루어 Design Matrix는 I, $A(BC)$, $B(AC)$, $C=AB$의 4개 열로 구성되어 1차 직선 선형 주효과만이 검출되고 2차 교호작용은 교락된다. $C=AB$는 Design Matrix 생성 시 $A$, $B$ 기본 열로 $C$는 $A\times B$방법으로 열을 생성하라는 의미이다.

### 4) $PB_N(k)$ PBD

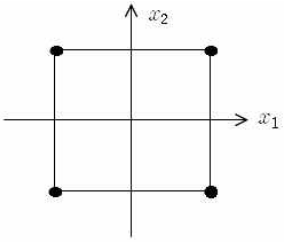
$N$은 실험횟수로 4의 배수 12, 20, 24, 36으로 $k=N-1$로 배치가능한 인자의 수를 나타내며 한 열의 마지막을 다음열의 첫 번째 인자수준으로 Design Matrix를 설계한다. 이 방법은 직교성을 이용하지 않은 간편한 설계의 이점이 있는 반면에 주효과만을 검출할 경우에 국한하여 사용해야 한다는 단점이 있다. 따라서 많은 잠재인자 중 치명인자를 선별하는 식스시그마 Analyze 단계에서는 Taguchi 방법과 같이 교호작용을 무시하고 주요한 인자를 선정하는 경우 사용하면 효율적이다. 대신 교호작용의 여부는 $2^k$FD, $RSM$ 등의 DOE를 이용하여 Improve단계에서 확인한 후 최적화를 추구한다.

### 5) $RSM$

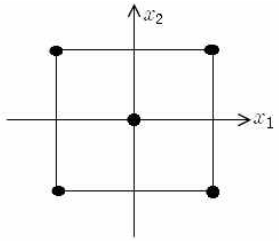
$RSM$에서 사용되는 CCD, BBD, FCD의 Design Matrix는 중심점, 요인점, 축점 등을 <표5>와 같이 사용하여 설계한다. CCD는 $x_1$, $x_2$ 인 경우 2차원의 원(Circle) 형태로 $x_1$, $x_2$, $x_3$인 경우 3차원인 구(Spherical) 형태로 Uniform Variance, Rotation, Block Orthogonality의 성질을 유지한다. FCD는 CCD에서 축점의 실험이 시간과 비용의 관점에서 비효율적인 경우 2차원의 경우 사용되며 Square로. 3차원의 경우 Cuboidal의 기하학적 좌표로 끌어 들여 축소된 Design Matrix를 설계한다. BB 역시 꼭지점의 실험이 비효율적인 경우 생략하는 방법이다.

<표5> *RSM* Design Matrix

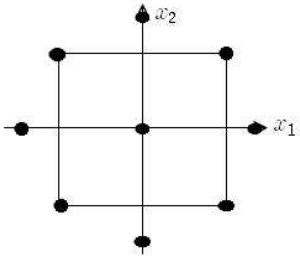
(a) $2^2$FDD



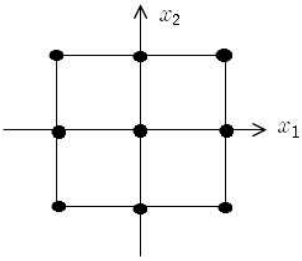

(b) 중심점이 있는 $2^2$FDD



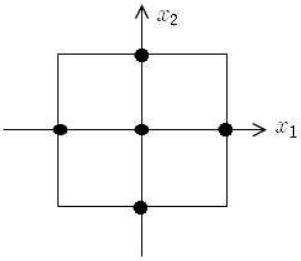

(c) CCD




(d) FCD




(e) BB

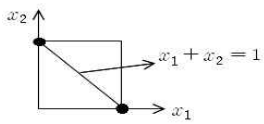



6) *MD*

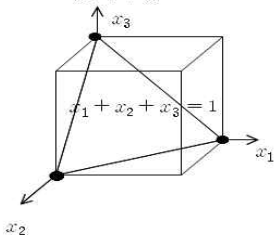
혼합물의 경우 $\sum_k x_k = 1$ 제약조건으로 <표6>과 같이 2차원의 경우 1차원 직선, 3차원의 경우 2차원 평면의 Simplex Design Matrix로 설계된다. {p, m} Lattice Design Matrix는 p차원을 m등분하는 설계로 중심점(Centroid)이 있는 {3,2} Lattice Design <표6>과 같다.

<표6> MD Design Matrix

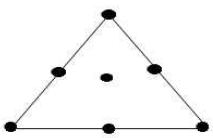
(a) 2차원




(b) 3차원




(c) 중심점이 있는 {3,2} Lattice Design



### 2.2.2 GLM 회귀분석과 ANOVA 분산분석의 모형

<표4>에서 $2^3$FDD의 GLM $\hat{y}=\hat{\beta_0}+\hat{\beta_1}x_1+\hat{\beta_2}x_2+\hat{\beta_3}x_3+\hat{\beta_{12}}x_1x_2+\hat{\beta_{23}}x_2x_3$이고 3차 교호작용은 반복을 수행하지 않아 오차항에 교락되며 중심점(ct pt)을 포함한다.

MINITAB에서 $x$ Design Matrix는 Natural(Uncoded) Variable을 $x_1=(x'_1-(x'_{max}+x'_{min})/2)/(x'_{max}-x'_{min})/2)$의 Coded Variable로 변수변환하면 $x_1=+1,-1$의 값을 갖는다.

GLM에서는 Coded Variable 형태의 상수, $x_1$, $x_2$, $x_3$, $x_1*x_2$, ct pt의 효과, 계수, 계수SE, $T$, $P$, VIF(혼합물의 경우 1에 가까우면 안정)를 구하고 ANOVA에서는 주효과, 2차 교호작용, 곡면성, 잔차오차, 적합성 결여, 순수오차, 총계에 대한 DF, SS, MS, F, P를 구한다. 끝으로 GLM에서는 Uncoded Variable의 형태로 $\hat{y}$의 계수에 의한 회귀식으로 최적화를 추구한다.

2차 RSM의 GLM과 ANOVA는 같으며 $\hat{y}=\hat{\beta_0}+\hat{\beta_1}x_1+\hat{\beta_2}x_2+\hat{\beta_{12}}x_1x_2+\hat{\beta_{11}}x_1^2+\hat{\beta_{22}}x_2^2$이고 회귀분산분석은 선형, 제곱, 교호작용, 잔차오차, 적합성결여, 순수오차, 전체 등의 DF, SS, MS, F, P를 구한다.

MD는 $\sum_k x_k=1$의 제약조건으로 $\beta_0$와 $x^2$이 없는 Special, Full 2차, 3차 회귀식을 구한다.

### 2.2.3 GLM과 ANOVA 분석의 대수적 표현

2.2.2절에서의 효과, $\beta$계수, SS, DF 등을, 2.2.1절의 Design Matrix를 통해 구하는 대수적 공식은 다음과 같다.

GLM에서 $\beta$계수는 LSE(Least Square Estimation)에 의해 $\hat{\beta}=(x'x)^{-1}x'y$를 통해 구할수 있고 $T$와 계수 SE는 2.2절의 $t_0=\hat{\beta_k}/\sqrt{\sigma^2 D_{kk}}$에서 $t_0$와 분모항 $\sqrt{\sigma^2 D_{kk}}$에 해당한다. $\beta$계수는 효과/2를 통해 구하며 효과(Effect)는 두 처리 조합 Block의 평균차를 의미하며 Contrast/(전체개수/2)로 구한다.

ANOVA에서 DF는 2.2.1절에 제시된 Design Matrix의 열배치가 된 인자의 자유도로 모두 1이다. $SS=Contrast^2/$전체개수 로 간단히 구한다. <표2>와 <표4>와 같이 중심점이 있는 경우 곡률(Curvature)이 발생하며 $SS_{곡률}=n_F n_c(\overline{y_F}-\overline{y_c})/(n_F+n_c)$로 $n_F$는 요인점 개수(Factorial Runs), $n_c$는 중심점 개수(Center Runs)로 $\overline{y_F}-\overline{y_c}$가 작을수록 곡률효과는 없다고 판정되며 자유도는 1이다.

2.2.2절에서 GLM, ANOVA의 잔차오차, 적합성결여, 순수오차는 <표7>의 기호를 사용해서 구한다. 참고로 적합성 결여 또는 곡면성의 P-Value가 $\alpha$=5%, 1%보다 클 경우 $H_0$ : 적합성 결여 없음, $H_0$ : 곡면성 없음으로 판정한다. 이는 정규성 검정의 원리와 같다.

<표7> 잔차 오차 Design Matrix

| 실험 순서 | ANOVA | 인자효과 | $I$ | $A$ | $B$ | $AB$ | $r_i$ | $\overline{y_{i.}}$ | $\hat{y_{i.}}$ |
|---|---|---|---|---|---|---|---|---|---|
| | GLM | 회귀계수 | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_{12}$ | | | |
| | | $x$ | | $x_1$ | $x_2$ | $x_1x_2$ | | | |
| 1<br>2<br>3<br>⋮<br>$s$ | | | | | | | | | |
| | | | | | | | $\sum_{i=1}^{s} r_i = 3$ | $\overline{\overline{y}}$ | |

<표7>에서 실험순서 $i=1, 2, \cdots, s$이고 반복은 $j=1, 2, \cdots, r_i$인 데이터는 $y_{ij}$이고 $\hat{y_{i.}} = \hat{\beta_0} + \hat{\beta_1}x_1 + \cdots + \hat{\beta_k}x_k$ ($k$=1, 2, ⋯, $p$)의 GLM 모형이다.

GLM의 *Total SS*= *Residual Error SS* + *Regression Model SS*에서 $\sum_i\sum_j (y_{ij} - \overline{\overline{y}})^2 = \sum_i\sum_j (y_{ij} - \hat{y_{y.}})^2 + \sum_i\sum_j (\hat{y_{i.}} - \overline{\overline{y}})^2$이고 자유도는 각각 $(n-1)$, $(n-(p+1))$,$((p+1)-1)$이다. 여기서 *Residual Error SS* = *Pure Error SS* + *Lack of Fit Error SS* 에서 $\sum_i\sum_j (y_{ij} - \hat{y_{i.}})^2 = \sum_i\sum_j (y_{ij} - \overline{y_{i.}})^2 + \sum_i\sum_j (\overline{y_{i.}} - \hat{y_{i.}})^2$이고 자유도는 각각 $(n-(p+1))$, $(n-s)$, $(s-(p+1))$이다. Regression MS의 유의성은 $R^2$=1-$MS_{Regression}$/ $MS_{Total}$, $i$번째 데이터를 제외하고 새로운 데이터로 예측하는 *PRESS SS*= $\sum_{i=1}^{s} (e_{(i)}/(1-D_{ij}))^2$을 이용한 P*rediction* $R^2$=1−$PRESS/SS_{Total}$로 판정한다.

ANOVA에서 *Total SS*=*Error SS* + *Effects Model SS*에서 $\sum_i\sum_j (y_{ij} - \overline{\overline{y}})^2 = \sum_i\sum_j (y_{ij} - \overline{y_{i.}})^2 + \sum_i\sum_j (\overline{y_{i.}} - \overline{\overline{y}})^2$로 자유도는 각각 $(n-1)$, $(n-s)$, $(s-1)$이다.

# 3. 상관분석과 회귀분석의 차이점과 보완적 분석

## 3.1 상관분석과 회귀분석의 차이점

식스시그마 5단계 중 치명인자 설정을 위한 Analyze 단계와 이 단계에서 설정된 인자에 대한 최적 수준을 결정하는 Impreve 단계를 명확히 구분하지 못하여 대부분의 실무자가 통계적 기법의 선정 시 실수를 하게 된다. 이러한 오류 중 하나가 Analyze 단계에서는 상관분석(Correlation Analysis)을 Improve 단계에서는 회귀분석(Regression Analysis)을 실시한다.

상관분석은 정비례($r$=1.0), 반비례($r$=-1.0)의 수학적 직선의 엄격한 관계를 오차를 고려하여 상관계수 $r$로 표현하는 방법이다. 따라서 모든 대응되는 ($x$,$y$)데이터가 직선 위로 올라가 있을 경우 상관계수 $r$은 1.0 또는 -1.0이 되고 멀어질수록 Zero에 가까워진다. 이 방법은 $x$와 $y$의 소극적 관계를 직선의 정비례, 반비례 관점에서 해석하는 방법이다.

이와 다르게 회귀분석은 ($x$,$y$)의 함수관계를 회귀식으로 나타내며 $x$값으로 $y$의 예측값 $\hat{y}$을 적극적으로 알아 볼 수 있는 방법이다. 통상 품질 분임조에서 $x$는 생산기술 조건의 인자수준, $y$는 특성치로 설정하기 때문에 소극적 해석방법인 상관분석보다 적극적 함수 관계에 의한 회귀분석을 사용해야 한다. 회귀분석 사용의 장점은 목표설정된 $y$에 대해 최적 인자수준 $x$를 회귀식을 통해 쉽게 구할 수 있다는데 있다.

## 3.2 상관분석과 회귀분석의 보완적 분석

회귀분석에서 회귀계수 $\hat{\beta}=(x'x)^{-1}x'y$에서 분산은 $(x'x)^{-1}$의 대각행렬로 결정되기 때문에 이를 크지 않게 하려면 $x$ 인자수준 간에 독립성 또는 직교성이 성립되는 것이 요구된다. 2.2절과 같이 $x$의 직교계획에 의한 회귀분석을 실시할 경우 더 이상의 상관분석은 불필요하나 그렇지 않은 경우 회귀분석 전 $x$의 독립성을 검토하기 위해 상관계수에 의한 분석을 실시하는 것은 보완적 관점에서 바람직한 방법이다.

# 4. 단계별 회귀분석과 범주형 회귀분석의 적용

## 4.1 다중 회귀분석과 단계별 회귀분석의 차이

식스시그마 Analyze 단계에서 많은 잠재인자 중 치명인자를 선별할 경우 사용되는 방법이 단계별(Stepwise) 회귀분석이다. 이는 특성치 $y$에 설명이 많이 되는 인자수준 $x$를 선정해 주는 효율적인 통계적 기법이다. 그러나 만약 실무자가 설정한 모든 인자수준 $x$에 대해 다중(Multiple) 회귀분석을 실시할 경우 유의하지 않은 회귀계수까지도 포함하기 때문에 목표설정된 $y$값을 달성하는 데 불필요한 현장노력을 기울여야 한다.

## 4.2 범주형 반응변수 회귀분석의 적용

식스시그마 Improve 단계, QC Story 대책실시 단계 등에서 주로 사용하는 회귀분석과 분산분석은 모두 $y$ 즉 특성치가 계량연속형 데이터를 전제로 한다. 그러나 일부 실무자는 두가지 분석기법을 사용할 시 특성치 $y$를 부적합품(불량)의 계수이산형 데이터로 취하는 실수를 범한다.

그러나 분산분석 시 계수이산형 데이터($\chi^2$분할표 검정)를 사용할 경우는 계량연속형

데이터의 방법을 그대로 적용해서는 안되며 설령 적용하더라도 정규근사 조건 $np \geq 5$의 만족여부를 먼저 검토하여야 한다. 계수이산형 데이터는 적어도 50번 이상의 반복을 취해야 하며 불량개수가 5이하인 경우 정규근사 조건을 만족하지 못하여 적용할 수 없다.

마찬가지로 회귀분석에서도 특성치 $y$가 범주형 반응변수인 경우 Binary(양품, 불량품) Regression, Ordinal(1등급, 2등급, 3등급) Regression, Nominal(A형, B형, C형) Regression 등을 사용해야 한다.

## 5. 결 론

식스시그마 Analyze와 Improve 단계에서 적용되는 회귀분석의 실무자 가이드라인은 다음과 같다.

1) 1원배치법의 분산분석의 데이터 배열로 회귀분석을 실시할 경우 제한된 수준의 수로 인한 데이터의 결여로 회귀식의 효과성에 문제가 될 수 있으므로 30개 이상의 수준에 대해 반복을 취하지 않은 데이터로 재배열하여 회귀분석을 실시한다.
2) 본 연구에서 제안한 GLM 회귀분석과 ANOVA 분산분석의 보완적 방법을 사용하고 회귀계수의 정밀도를 향상하기 위해 인자수준의 Design Matrix의 직교성을 이해하고 용도에 맞게 활용한다.
3) 생산기술조건에 따른 스펙 데이터의 관계는 소극적 상관계수보다 적극적인 회귀식에 의해 목표설정을 달성할 수 있는 최적조건을 설정한다.
4) 상관분석은 회귀분석의 인자수준 Design Matrix의 독립성을 검토하기 위한 보조도구로 활용한다.
5) Analyze 단계에서 많은 잠재인자 중 소수의 치명적인 핵심인자를 선별할 시 단계별 회귀분석 또는 교호작용이 무시되는 $2^{k-p}$부분배치법, Placket-Burman법을 사용한 후 Improve 단계에서 교호작용이 고려된 $2^k$요인배치법, 반응표면분석으로 최적 인자수준을 선정한다.
6) 회귀분석, 분산분석의 특성값 반응변수는 계량연속형 데이터를 활용해야하며 부적합품의 계수이산형 데이터일 경우 범주형 회귀분석을 활용한다.

## 6. 참 고 문 헌

[1] 김두섭 외, 회귀분석, 나남, 2008.
[2] 새 MINITAB 실무완성, 이레테크, 2009.
[3] 최성운 외, “안전 및 환경적용을 위한 최소 실험계획”, 대한안전경영과학회지, 7(5) (2005) : 69-84
[4] Chatterjee S., Price B., Regression Analysis by Example, John Wiley, 1977.

[5] Cox. D.R., Reid N., Theory of the Design of Experiments, Chapman and Hall, 2000.
[6] Draper N., Smith H., Applied Regression Analysis, 2nd Edition, John Wiley, 1981.
[7] Gunst R.F., Mason R.V., Regression Analysis and Its Application : A Data-Oriented Approach, Marcel Dekker, 1980.
[8] Montgomery D.C., Design and Analysis of Experiments, 6th Edition, John Wiley & Son, 2005.
[9] Mosteller F., Tukey J.W., Data Analysis and Regression, Addison-Wesley, 1977.