# 정량적, 정성적 회귀분석의 오적용과 이해
# Understanding of the Misuse Cases of Quantitative and Qualitative Regression Analysis

최 성 운*

## Abstract

The research shows misuse cases of quantitative regression analysis used in QC circle activity and six sigma movement which presents guidelines of correct use for quality practitioners. Additionally, the qualitative regression analysis that responses nonconforming ratio of variable y, is reviewed based on misuse cases for proper use by practitioners in the field. In most cases, there are frequent errors that involve the correlation analysis or ANOVA, regardless of using quantitative regression analysis. In addition, qualitative regression analysis for the nonconforming ratio that has dependent variable of discrete and categorical data, is often applied with quantitative regression and result in ineffective quality improvement.

Keywords: Quantitative and Qualitative Regression Analysis, Misuse Cases, Understanding, Correct Use, Correlation Analysis, ANOVA, Categorical Data

## 1. 서 론

분석적, 해석적 방법(Analytical Method)은 전체를 부분으로 분해하는 시스템 분석(System Analysis)과 원인과 결과로 분해하는 인과관계 분석(Causal Analysis)으로 구분된다. 인과관계의 대표적인 정성적인(Qualitative) 분석기법으로는 특성요인도(Cause and Effect Diagram)가 있으며 정량적인(Quantitative) 분석기법으로는 회귀분석(Regression Analysis)이 있다.

* 경원대학교 산업공학과

정량적 회귀분석[5,10]은 설명변수(원인, 투입, 독립변수) $x$와 반응변수(결과, 산출, 종속변수) $y$가 모두 계량연속형(Continuous, By Variable) 데이터로 $x$가 다변량일 경우 $x$간의 독립성을 요구한다. 정량적 회귀분석은 $y=f(x)$의 형태로 $x$에 대한 $y$의 수학함수(Mathematical Function) 회귀식의 인과관계를 도출하여 주어지는 $x$값으로(Fixed Variable), 랜덤한 $y$를 (Random Variable) 예측(Forecasting, Prediction)하는 적극적인 방법이다. 이와 다르게 상관분석(Correlation Analysis)은 $x$원인간의 관계, $y$결과간의 관계를 수학의 정비례, 반비례 개념의 양의 상관관계, 음의 상관관계로 파악하는 소극적 방법이다.

그러나 품질분임조의 QC Story 15단계와 식스시그마의 DMAIC(Define, Measure, Analyze, Improve, Control) 5단계에서 사용되는 회귀분석 방법은 상관분석, 분산분석(ANOVA)의 기법과 혼용되어 잘못 사용되고 있다 .특히 정량적 회귀분석의 경우 $x$가 정성적인 데이터인 경우도 가변수(Dummy Variable) 처리 없이 정량적인 데이터와 같이 그대로 대입하는 오류를 범하고 있다. 또한 $y$가 불량갯수, 결점수인 계수이산(Discrete, By Attribute) 범주형(Categorical) 데이터인 정성적 회귀분석[1-9, 11, 12]의 경우 정량적 회귀분석을 잘못 적용하여 기업의 품질관리 활동에 큰 지장을 초래한다.

따라서 본 연구에서는 $y$가 계량연속 데이터인 정량적 회귀분석의 품질활동에서의 오적용사례와 적용방안을 제시한다. 또한 $y$가 불량률과 같은 계수이산 범주형 데이터인 정성적 회귀분석의 오적용 사례와 올바른 이해를 위해 로지스틱(Logistic) 회귀함수의 원리와 특징을 고찰하고자 한다.

본 연구의 차별성은 품질분임조, 식스시그마 활동에서의 회귀분석에 대한 실제 오적용사례를 제시하고 적용 가이드라인을 제시하는데 있다.

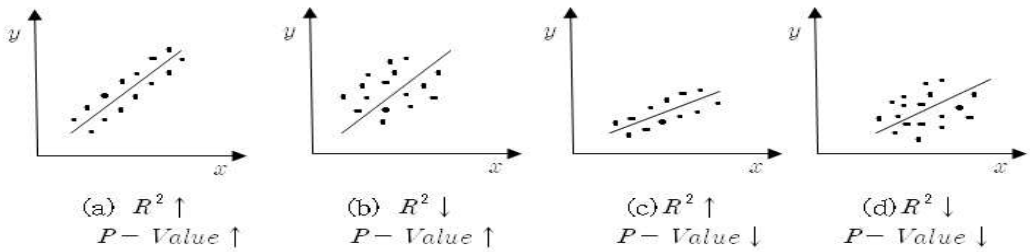
## 2. 정량적 회귀분석의 특징 및 오적용

### 2.1 정량적 회귀분석의 종류 및 검추정

$y$가 계량연속형 데이터인 정량적 회귀분석의 종류로는 $x$변수가 1개 선형(직선)인 단순(Simple)회귀, $x$가 1개 2차이상인 곡선(Curvilinear)회귀, $x$가 여러개 선형인 다중(Multiple)회귀, $x$가 여러개 2차이상인 다항(Polynomial)회귀가 있다. 2차 다항회귀는 $x$의 독립성과 직교성을 유지하기 위해 CCD(Central Composite Design)를 Rotation하는 방법을 사용하는 RSM(Response Surface Methodology)의 DOE(Design of Experiment)가 많이 사용된다. 다중회귀에서 $x$간의 상관관계가 존재하는 경우 공분산을 이용한 다변량분석을 실시한다.

정량적 회귀분석의 계수는 LSM(Least Square Method)에 의해 추정(Estimation)되며 회귀계수의 의미는 검정(Test)으로 유의성을 판정한다. 가설검정의 방법으로는 결정계수(Coefficient of Determination) $R^2$가 80%이거나, ANOVA(Analysis of Variance)의 $F_0$값에 대한 유의확률(Significance Probability) P-Value가 유의수준(Significance Level) $\alpha$보다 작을 경우 대립가설 $H_1$인 회귀계수는 의미가 있다라고 판정을 한다. 단순직선회

귀의 경우 회귀계수의 유의성검정은 산점도(Scatter Diagram)의 점들이 회귀식위에 몰려 있으며, 회귀식 기울기가 가장 크게 기울어진 경우가 [그림 1.a] 같이 이상적인 회귀식으로 판정된다. 전자의 경우는 $R^2$와 P-Value에 의해 나오는 회귀식의 적합형태는 [그림 1]과 같으며 (d)의 경우는 무의미한 회귀계수로 판정되나 (b),(c)의 경우는 $R^2$, P-Value 모두를 고려하여 사용여부를 신중히 고려하여야 한다.




[그림 1] $R^2$와 $P-Value$에 의한 단순회귀식

## 2.2 정량적 회귀분석의 오적용

정량적 회귀분석에서 첫 번째 오적용 사례로 설명변수 $x$가 계량연속형 데이터가 아닌 기계의 종류, 원료의 종류 등과 같은 계수이산형 데이터인 경우는 그대로 명목형 숫자를 대입해서는 안되고 가변수(Dummy Variable)처리를 해주어야 한다. 만약 3개의 범주형 질적인자 $x$가 있을 경우 가변수처리는 $(D_1, D_2) = (1,0), (0,1), (0,0)$로 2개가 필요하다.

두 번째 오적용 사례는 상관분석과 회귀분석을 구별못하는 데 있다. 회귀식을 구하고 회귀분석 ANOVA의 P-Value로는 상관관계가 있다라는 그릇된 결론을 맺는다. 이 이유는 2가지로 추측되는데 첫째는 상관은 두 변수가 관계가 있다라고 쉽게 단어를 이해할 수 있는 반면에 회귀는 돌아간다(Regression)라는 뜻이 몸에 와닿지 않기 때문이다. 둘째는 회귀식의 유의성 검정시 상관계수 $R$의 제곱인 기여율 $R^2$를 사용하므로 상관과 회귀는 같은 것으로 오인하기 때문이다. 회귀라고 사용하는 이유는 기준점을 $y$의 평균으로($\because x$는 고정변수, $y$는 확률변수) 회귀식을 뺑뺑 돌리다가 기울기가 있으면 $H_1$ 채택, 평균인 수평선으로 돌아오면 기울기가 없으므로 $H_0$채택을 판정하는 데서 유래한다. 상관계수 $r$이 두 변수의 관계정도를 1점 만점으로 소극적으로 평가하는 반면, 회귀식 $y = \beta_0 + \beta_1 x$는 함수의 적극적 관계로 $x$값에 의해 $y$값을 예측하므로 품질개선활동에서는 당연히 상관보다는 회귀를 사용해야 한다. 그러나 $x$값이 제어가능하고 쉽게 구할 수 있다는 가정이 요구되며 품질현장에서 $x$는 온도, 압력등의 생산기술조건, $y$는 치수, 중량 등의 제품스펙인 경우 회귀분석의 적용이 가능하다.

세 번째 오적용사례로는 DOE의 분산분석(ANOVA)과 회귀분석의 용도를 구별못하는 데 있다. 실험계획법의 분산분석은 특정인자(Factor)의 수준처리(Level Treatment) $x$에 의한 반복 특성치(Characteristice Value) $y$가 영향을 주었는가를 판정하는 방법

이다. DOE의 ANOVA는 수평선인 총평균, 인자수준간 평균을 사용하므로 기하학적 관점으로 수평선을 기준으로 점의 높낮이가 있는 경우는 $H_1$채택으로 판정한다. 따라서 수학함수로 표시되는 회귀분석보다 DOE의 분산분석은 적극적인 정보를 주지 못하기 때문에 $x$에 의해 $y$를 예측하는 회귀분석을 사용하는 것이 바람직하다. 부수적인 실수로 DOE의 분산분석의 데이터로 회귀분석을 적용하는 MINITAB분석의 경우 제한된 인자수준 $x$로 인해 회귀분석의 타당성이 문제가 되므로 이 경우는 적어도 $n \geq 30$이 되는 인자수준 $x$값과 반복없는 $y$값을 사용해야 한다.

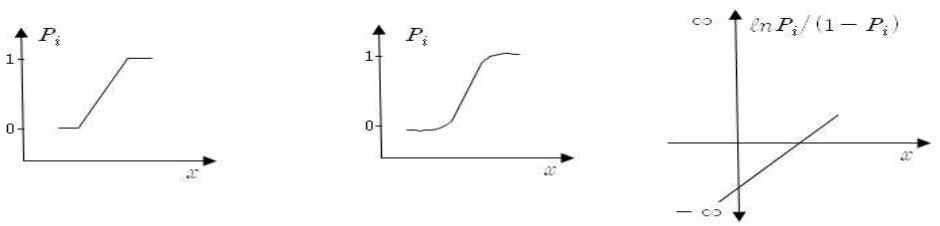
DOE의 ANOVA를 분석할 경우 회귀식 모양의 GLM(Generalized Linear Model)을 사용하는 것은 변수처리의 효율성 관점이지 두 방법이 동일한 용도의 분석을 실시하는 것은 아니라는 점에 유의해야 한다.

# 3. 정성적 회귀분석의 이해와 오적용

## 3.1 정성적 회귀분석의 이해

$y$가 계수이산 범주형 데이터이고 $x$가 계량연속데이터인 경우는 Logistic Regression 분석을, $x$가 계수이산 범주형 데이터인 경우는 Log Linear분석을 실시한다.

Logistic Regression 분석은 [그림 2]와 같이 이분형(Dichotomous) 선형회귀함수에 대한 $y$ 또는 $P_i$의 0~1 값을 $-\infty \sim \infty$로 변환해주기 위한 방법이다. <그림2.6b>의 로지스틱 함수 $P_i = \exp(\beta_o + \beta_1 x)/(1+\exp(\beta_0 + \beta_1 x)) = 1/(1-\exp(\beta_0 + \beta_1 x))$이고 다음과 같은 과정을 거쳐 [그림 2.c]와 같은 Logistic Regression 함수가 된다. $P_i = 1/(1-\exp(\beta_0 + \beta_1 x))$의 $y$ Range를 $0 \sim \infty$로 바꿔주고 Odds Ratio를 구하기 위해서 $(1-P_i)$로 나누면 $P_i/(1-P_i) = \exp(\beta_0 + \beta_1 x))$이다. $y$값의 Range를 $-\infty \sim \infty$로 바꿔주기 위해 $\ell n$ 변환하면 $\ell n P_i/(1-P_i) = \beta_0 + \beta_1 x$가 된다. Odds Ratio는 Relative Ratio에 비해 두 집단의 비를 구할 경우 샘플의 크기에 영향을 받지 않고 구할 수 있는 장점이 있다. Logistic 함수를 사용하는 Logit 변환대신에 누적정규분포함수를 사용하는 Probit 변환이 있는데 $P_i = 1/\sqrt{2\pi}\int_{-\infty}^{x} \exp(-x^2/2)dx$이다.




(a) 이분형 선형 회귀함수 (b) 로지스틱 함수 (c) Logistic Regression 함수

[그림 2] Logistic Regression 함수

## 3.2 정성적 회귀분석의 오적용

정성적 회귀분석의 첫 번째 오적용 사례는 $y$가 불량률같이 계수이산 범주형 데이터를 2절과 같은 $y$가 계량연속형 데이터인 정량적 회귀분석을 잘못 적용하는 것이다. 이 경우는 $y$의 Binary, Ordinal, Nominal 범주형 데이터의 형태에 따라 Logisitc Regression 분석을 적용해야 한다.

두 번째 오적용 사례는 Logistic Regression 결과를 잘못 해석하는 것이다. $\ell n(P_i/(1-P_i)) = \beta_0 + \beta_1 x$의 결과를 $y = \beta_o + \beta_1 x$으로 오인하여 $y$대신에 $P_i$를 대입하는 경우이다. 이는 정성적 회귀분석의 원리에 대한 이해부족에 기인한다.

# 4. 결 론

본 연구에서는 품질분임조와 식스시그마혁신활동에서 가장 많이 사용되는 회귀분석에 관한 오적용사례와 적용방안을 제시하였다. $y$가 계량연속형 데이터인 정량적 회귀분석의 경우 $x$의 가변수 처리, 상관분석과 DOE의 ANOVA와의 차별적인 사용을 제시하였으며 $y$가 계수이산 명목형 데이터인 정성적 회귀분석의 경우 올바른 적용과 해석을 위한 Logistic Regression 함수의 원리와 특징을 고찰하였다.

# 5. 참 고 문 헌

[1] Agresti A., 정광모외 역, 범주형 자료 분석 개론 : SAS의 응용 및 해석, 자유아카데미, 2009.
[2] Hosmer D.W., Lemeshow S., Applied Logistic Regression, 2 Edition, Wiley, 2000.
[3] Kleinbaum D.G., Logistic Regression : A Self-Learning Text, 3 Edition, Springer, 2010.
[4] Menard S.W., Logistic Regression : From Introductory to Advanced Concepts and Applications, Sage Publications, 2009.
[5] 김두섭외, 회귀분석, 나남, 2008.
[6] 김순귀외, SPSS를 활용한 로지스틱 회귀모형의 이해와 응용, 자유아카데미, 2003.
[7] 남궁평외, 범주형 자료의 통계분석, 자유아카데미, 1991.
[8] 노형진, SPSS에 의한 범주형 데이터 분석, 효산, 2007.
[9] 박상언, 판별분석, 로지스틱 회귀모형, 민영사, 2002.
[10] 성웅현, 응용 로지스틱 회귀분석, 탐진, 2001.
[11] 최성운, “식스시그마 품질개선단계에서 GLM 회귀분석의 이해와 적용”, 대한안전경영과학회 춘계학술대회 발표문집, (2010):539-550.
[12] 최성운, “MRA에서 특성값의 측정단위와 수치형태에 따른 종합 만족도 산출 방법”, 대한안전경영과학회 추계학술대회 발표문집, (2009):565-572.