The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.2.237
26(2), 237–252

Estimating Automobile Insurance Premiums Based on Time Series Regression

Yeong-Hwa Kim[a,1] · Wonseo Park[b]

[a]Department of Applied Statistics, Chung-Ang University
[b]Department of Statistics, Graduate School of Chung-Ang University

(Received October 12, 2012; Revised January 14, 2013; Accepted February 11, 2013)

Abstract

An estimation model for premiums and components is essential to determine reasonable insurance premiums. In this study, we introduce diverse models for the estimation of property damage premiums(premium, depth and frequency) that include a regression model using a dummy variable, additive independent variable model, autoregressive error model, seasonal ARIMA model and intervention model. In addition, the actual property damage premium data was used to estimate the premium, depth and frequency for each model. The estimation results of the models are comparatively examined by comparing the RMSE(Root Mean Squared Errors) of estimates and actual data. Based on real data analysis, we found that the autoregressive error model showed the best performance.

Keywords: Depth, Durbin-Watson, frequency, insurance, premium, regression, time series.

1. 서론

현대인에게 자동차가 생활필수품이 되면서 자연스럽게 자동차 보험료 또한 가계지출 항목의 필수항목이 되었다. 자동차 보험계약은 보험사업자와 보험계약자 사이에 이루어진 계약으로서, 피보험자의 위험에 대하여 미리 정한 조건하에서 보험금을 계약자에게 지급하고 그 대가로 보험계약자는 본인의 상황에 맞는 보험료를 납입하는 것을 그 내용으로 한다. 이때 보험료란 보험회사에서 판매하는 '보험'이라는 상품의 가격이다. 보험료가 너무 낮게 산정되면 보험회사가 손해를 보게 되고, 반대로 너무 높게 산정되면 보험사 또는 전체 보험업계에 지나치게 많은 이익이 발생하고 전체 보험료의 인상은 물가인상에 영향을 미치게 되어 보험 감독기관에서도 보험료 수준에 특별한 관심을 가지고 있기 때문에 적정한 보험료를 책정하는 것이 매우 중요하다. 보험료는 심도(depth)와 빈도(frequency)의 곱으로 나타낼 수 있으며, 여기서 심도란 사고건당 피해액을 말하며 빈도란 사건발생률로 기간에 유효한 계약건수 중 사고건수의 비율을 말한다. 자동차 대물 보험료 산정에 있어 보험회사에 판단 구분에 따라 심도 및 빈도의 추정 방식이 달라지는데, 자료를 적합시키는 모형에 따라 그 추정값에 차이게 있게 되므로 이에 따라 추정되는 보험료 역시 달라진다.

본 연구에서는 보험료 추정에 사용할 수 있는 몇 가지 통계적 모형을 소개하고 실제 국내 보험사 자료에

This research was supported by 2011 Research Grant for graduate student of Chung-Ang University.
[1]Corresponding author: Professor, Department of Applied Statistics, Chung-Ang University, 221 Heuksuk-Dong, Dongjak-Gu, Seoul 156-756, Korea. E-mail: gogators@cau.ac.kr

적합시켜 보고자한다. 현재 업계에서 주로 사용하고 있는 방법으로 가변수를 이용한 시계열 회귀모형이 있는데 이 모형에서 문제가 되는 것은 잔차의 자기상관 존재여부이다. 이러한 문제를 고려한 모형으로 독립변수 추가모형과 자기회귀 오차모형이 있다. 또한 자료의 특성상 계절형 ARIMA 모형과 개입모형 등을 고려해 볼 수 있다. Cummins와 Powell (1980)의 연구와 같이 기존 연구에서는 계절적요인 등 이상요인을 제거하는 OLS 모형 및 인플레이션 효과 등을 제외하는 경제적 모형 등이 제안되었으며, 이는 예측에 있어 편의(bias)가 발생되는 요인을 제거하면서 예측력을 높이는 모형에 관련된 연구이다. 본 논문은 이러한 변동요인을 모형내에 반영하는 시계열 분석 방법의 제안으로서 예측력에 있어 정교함을 향상시키는 방법을 제시하고자 하였으며, 본 논문은 비정상성을 갖는 자동차 대물 보험료의 실제 데이터를 이용하여 가변수를 이용한 시계열회귀모형, 독립변수 추가모형, 자기회귀 오차모형, 계절형 ARIMA 모형 및 개입모형에 자료를 적합시켜 얻는 예측치들과 RMSE(root mean squared error)를 통해 예측력을 비교해 봄으로써 가장 우수한 모형에 대해 논의해 보고자 한다.

2. 시계열 회귀모형과 개입모형

이 장에서는 보험업계에서 실제 사용되고 있는 모형인 시계열 회귀모형에 대해 알아보고 이러한 모형에서 발생할 수 있는 오류와 이를 고려한 새로운 모형을 고려해 보고자 한다. 시계열 자료를 이용한 예측방법의 하나로 회귀분석을 사용할 수 있는 데 회귀분석을 이용하여 미래에 대한 예측치를 얻어내는 데는 시간의 흐름에 따라서 변동을 갖는 시계열 자료의 모형화에 제한적이라는 점과 미래에 대한 예측이 독립변수의 예측이 선행되어야 한다는 점에서 현실적으로 어려움이 있다. 또한 시계열 자료에는 본질적으로 자기상관이 존재하므로 회귀분석에서 오차항의 가정에 위배된다. 따라서 이러한 문제를 어떻게 효과적으로 해결하느냐에 관한 것이 중요한 문제가 된다. 이 장의 이론적인 내용은 Park과 Kim (2003), Cho와 Sohn (2009), Park과 Song (1998)의 내용을 참조하였다.

2.1. 추세를 이용한 시계열 회귀모형

시계열 z_t는 추세만을 이용하여 $z_t = T_t + \epsilon_t$로 표현할 수 있다. 단, 여기서 z_t는 시점 t에서의 관측값, T_t는 시점 t에서의 추세, ϵ_t는 시점 t에서의 오차항이다. 이 모형은 시계열 관측값들이 시계열의 추세를 중심으로 임의의 변동을 가질 때 의미를 갖게 된다. 실제 시계열 자료에는 다양한 추세가 존재하며 이를 다항 추세모형(polynominal trend model)을 이용하여 다음과 같이 확장할 수 있다.

$$z_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p + \epsilon_t. \tag{2.1}$$

이러한 다항 추세모형 (2.1)에서 모수의 추정은 최소제곱추정법이 사용되고 오차항에 대한 가정도 회귀분석에서 사용하는 것과 동일하게 적용된다. 따라서 잔차를 이용하여 모형의 가정에 대한 점검이 요구된다. 잔차분석을 통해 오차항의 독립성 가정을 만족시키지 못할 경우에는 잔차가 아직 데이터에 대해 설명할 정보를 가지고 있다는 것을 의미하므로 추가적인 잔차의 모형화를 통해 보다 높은 설명력을 가지고 있는 모형을 유도할 수 있다.

2.2. 가변수를 이용한 시계열 회귀모형

주어진 데이터가 연별자료일 경우 계절변동을 찾아보기 힘들지만 월별 또는 분기별로 측정되는 시계열 자료는 많은 경우 계절변동을 뚜렷이 확인할 수 있다. 따라서 시계열 분석에서 계절변동은 반드시 고려되어야할 요인이 된다. 시계열 분석에서 계절요인을 분석 모형 내에 반영하기 위하여 계절요인의 변동을 시간의 변화에도 일정하게 변동량을 유지하는 고정 계절변동과 시간이 변화함에 따라 증가하는 확산

계절변동의 두 가지 형태로 가정한다. 이러한 계절요인을 반영하는 시계열 회귀분석의 방법으로 가변수를 이용하는 방법이 있다. 시계열 z_t가 고정계절변동의 형태를 가지고 있는 경우 다음과 같은 형태의 모형을 생각할 수 있다.

$$z_t = T_t + S_t + \epsilon_t, \tag{2.2}$$

여기서 z_t는 시점 t에서의 시계열 관측값, T_t는 시점 t에서의 시계열 추세, S_t는 시점 t에서의 계절요인, ϵ_t는 시점 t에서의 오차항을 의미한다. 모형 (2.2)는 시계열 z_t의 평균이 시간에 의존하여 $\mu_t = T_t + S_t$로 구성되고 시간에 따른 시계열의 변동은 오차항 ϵ_t에 의해서 나타난다고 보는 것이다. 오차항에 대하여 일반적인 회귀분석과 동일한 가정을 하고 가변수(dummy variable)를 이용하여 계절요인을 반영한 시계열 회귀모형은 다음과 같다.

$$z_t = \beta_0 + \beta_1 t + \beta_{D_1} D_{1,t} + \beta_{D_2} D_{2,t} + \cdots + \beta_{D_{L-1}} D_{L-1,t} + \epsilon_t, \tag{2.3}$$

여기서 $D_{i,t}$는 시점 t에서 계절이 i인 경우($i = 1, 2, \ldots, L-1$)에만 '1'이 되는 가변수이다.

모형 (2.3)에서 계절요인이 $L-1$번째까지 정의되어 있는 것은 가변수를 포함하는 회귀모형에서 최소제곱법으로 모수를 추정하는 과정에서 발생하는 가변수 함정을 피하기 위한 것으로 최소제곱 추정법 적용 시 역행렬을 구할 수 없게 되는 문제와 계절요인의 효과를 나타내는 $\beta_{D_1}, \beta_{D_2}, \ldots, \beta_{D_{L-1}}$의 추정치 해석에 영향을 주게 되므로 주의해야 할 부분이다. 예를 들어, 관측된 시계열 자료가 분기별로 측정된 자료이고($L = 4$) 시점 t에서 3분기가 되는 경우, 모형은 $z_t = \beta_0 + \beta_1 t + \beta_{D_1}(0) + \beta_{D_2}(0) + \beta_{D_3}(1) + \epsilon_t = \beta_0 + \beta_1 t + \beta_{D_3}(1) + \epsilon_t$가 된다. 가변수를 이용한 시계열 회귀모형은 모형에서도 알 수 있듯이 시계열 자료가 고정 계절변동을 갖는 경우 적합한 방법이다. 만약 시계열 자료가 확산 계절변동을 갖는 경우 가변수를 이용한 시계열 회귀모형을 적용하려면 시계열 자료를 적절한 변환(제곱근변환, 로그변환)을 통하여 확산 계절변동을 고정 계절변동으로 변환한 후 모형을 적합하여야 한다. 물론 모형에 의하여 얻어지는 예측치들은 역변환을 적용하여 원래의 시계열의 값과 동일한 단위로 바꾸어주어야 한다.

2.3. 독립변수 추가모형과 자기회귀 오차모형

시계열 자료는 항상 자기상관이 존재한다고 보아도 무방할 것이다. 자기상관은 시계열 자료가 시간의 흐름에 따라 계속적으로 측정된 자료이기 때문에 발생하는 것으로 현재 측정된 값은 과거의 측정된 값의 영향을 받게 된다는 것을 의미하며 잔차에 자기상관이 존재한다는 것은 오차항의 독립성 가정을 만족하지 못하는 것으로 잔차가 아직 모집단을 설명할 정보를 가지고 있다는 것을 의미한다. 그러나 회귀모형을 이용하여 자료를 분석할 때는 일반적으로 오차항이 서로 독립이라는 가정을 전제로 한다. 오차의 가정이 지켜지지 않고 자기상관이 존재할 경우 새로운 독립변수를 추가하거나 또는 오차항에 추가적인 모형을 설정해줌으로써 이를 해결할 수 있다.

독립변수를 새롭게 추가한다는 것은 많은 예산과 시간이 필요로 될 수 있는 일이다. 그러나 시계열 자료의 경우 현재의 값은 과거의 영향을 받아 실현된다는 기본적인 가정에 따라 전 시점의 값을 독립변수로 고려해볼 수 있다. 오차항의 추가적인 모형으로 AR(p) 모형을 고려해볼 수 있다. 오차항에 대한 일반적인 가정은 오차항이 ARMA(p, q) 모형을 따른다는 것이나, MA 모형은 AR 모형에 의해 충분히 잘 설명될 수 있기 때문에 오차항이 AR 모형을 따른다는 가정은 큰 무리가 없다는 것이 많은 연구에 의해 밝혀져 있다. 오차항이 AR(p)를 따르는 경우의 k차 자기회귀 오차모형(autoregressive error model)은 다음과 같다.

$$\begin{aligned} z_t &= \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_k X_{tk} + \epsilon_t, \\ \epsilon_t &= \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots + \phi_p \epsilon_{t-p} + a_t, \end{aligned}$$

여기서 a_t들은 서로 독립이다. 자기회귀 모형과 다른 점은 오차항인 ϵ_t들이 회귀모형에서처럼 서로 독립이 아니고 자기상관관계를 가지며, 특히 k차 자기회귀과정 AR(p)를 따른다는 것이다

2.4. 개입모형

개입이란 시계열의 변화를 발생시키는 외부의 영향(충격)을 말하며 개입의 발생으로 인한 효과는 다양한 형태로 나타난다. 이러한 개입에 의해 시계열에 변동이 생기게 되면 기존에 설명하였던 시계열 분석 방법들로는 만족할 만한 결과를 얻기가 쉽지 않다. 개입변수들은 일반적인 시계열변수와 달리 어떤 사건의 발생이 지속되는 기간에 따라 T시점에서 발생하여 개입효과가 T시점에만 영향을 미치는 지시함수(indicator function)와 T시점에서 발생하여 개입효과가 발생시점 이후로 지속적으로 영향을 미치는 계단함수(step function)의 두 가지 형태를 따른다 (Box와 Tiao, 1975). 이러한 개입효과를 입력변수로 반영시킨 모형을 개입모형이라 하며 일반적인 모형은 다중 전이함수모형의 특수한 형태로 다음과 같이 정의한다.

$$Y_t = \nu_1(B)X_{1t} + \nu_2(B)X_{2t} + \cdots + \nu_k(B)X_{kt} + N_t,$$
$$X_{jt} = P_t(T) \quad \text{또는} \quad S_t,\ N_t = \frac{\theta(B)\Theta(B^s)}{\phi(B)\Phi(B^s)}a_t.$$

개입모형 분석에서는 개입의 효과가 어떻게 시계열 자료에 반영될지에 따라 개입모형의 형태를 결정한다. 개입변수의 형태는 다음과 같다.

$$\frac{w(B)}{\delta(B)}B^b X_t,$$

여기서 X_t는 입력변수, $w(B) = w_0 - w_1B - \cdots - w_sB^s$, $\delta(B) = \delta_0 - \delta_1B - \cdots - \delta_sB^s$이고, b는 개입효과가 발생시점부터 즉시 반영되지 않고 b 시차만큼 지연되어 반영됨을 나타내는 값이며, w_i는 개입의 초기 기대효과를 의미하고 δ_i는 개입의 영속적인 효과를 나타낸다.

2.5. 예측력 비교

모형간의 예측력은 RMSE를 단순 비교하여 더 작은 RMSE값을 갖는 모형을 우수한 모형으로 판단하는 방법과 각 모형의 MSE의 차이가 통계적으로 유의미한가를 검정하는 Diebold-Mariano 검정이 있다. Diebold-Mariano 검정의 영가설은 '기준모형의 평균제곱 예측오차가 비교모형의 평균제곱 예측오차와 같다'이며, 다음과 같은 통계량을 t-검정하여 모형간 예측력을 비교한다.

$$S_{DM} = \left[\hat{V}\left(\bar{d}\right)\right]^{-\frac{1}{2}}\bar{d},$$
$$\hat{V}\left(\bar{d}\right) = T^{-1}\left(\hat{\gamma_0} + 2\sum_{k=1}^{h-1}\hat{\gamma_k}\right), \quad \hat{\gamma_k} = T^{-1}\sum_{t=k+1}^{T}\left(d_t + \bar{d}\right)\left(d_k + \bar{d}\right),$$

여기서 T는 예측기간의 길이, h는 예측시계(horizon)이며 $\bar{d} = T^{-1}\sum_{t=1}^{T}(e_{a,t}^2 - e_{b,t}^2)$는 기준모형의 예측오차 $e_{b,t}$와 비교모형의 예측오차 $e_{a,t}$의 MSE의 차이이다.

3. 실증 데이터 분석

국내 자동차 보험회사인 S사의 실제 자료에 대하여 가변수를 이용한 시계열 회귀모형과 독립변수 추가모형, 자기회귀 오차모형, 계절형모형 및 개입모형을 통해 자료를 적합시키고 이를 통해 추정치를 구해

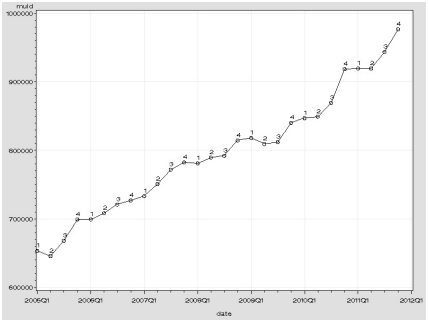

Figure 3.1. Time series plot for depth

본다. 또한 최종적으로 각각의 추정치들과 실제 자료를 비교해 보고 각 추정치의 RMSE를 통하여 예측력을 비교한다. 자기상관관계에 대한 검정으로는 Durbin과 Watson (1950, 1951, 1971)이 제안한 더빈-왓슨 검정과 Ljung과 Box (1979)의 포트맨토 검정(Portmanteau test) 등을 고려하였으며 단위근 검정으로는 Dickey와 Fuller (1979)의 Dickey-Fuller 단위근 검정을 고려하였다. 또한 예측력 비교를 위한 검정으로는 Diebold와 Mariano (1995)의 Diebold-Mariano 검정을 사용하였다.

3.1. 자료의 구성

분석에 사용된 자료는 국내 자동차 보험사인 S사의 2005년~2011년 실제 자동차 대물 보험료 자료로서 분기별 심도, 빈도, 보험료와 2005년~2010년 사이의 자동차 보험료에 영향을 미칠 수 있는 외부변수 등으로 구성된 자료를 모형화에 사용하였고 2011년도 자료를 예측력 비교 자료로 사용하였다. 외부변수와 변수명, 분기를 고려한 변수값은 다음과 같다.

D_1: 1분기의 계절효과, D_2: 2분기의 계절효과, D_3: 3분기의 계절효과

X_1: 2005년 7월 주 5일제 확대실시, $X_1 = I_{\{t\geq 3\}}$

X_2: 2006년 6월 주 5일제 전면실시, $X_2 = I_{\{t\geq 7\}}$

X_3: 2005년 11월 정비수가 인상, $X_3 = I_{\{t\geq 4\}}$

X_4: 2006년 4월 현대 모비스 부품값 인상, $X_4 = I_{\{t\geq 6\}}$

X_5: 2010년 1월 물적할증기준 선택제, $X_5 = I_{\{t\geq 21\}}$

X_6: 2010년 9월 현대 모비스 부품값 인상, $X_6 = I_{\{t\geq 23\}}$

여기서 $I_{\{\cdot\}}$는 지시함수, 심도는 사고건당 피해액, 빈도는 사고건수/유효대수, 유효대수는 해당기간에 유효한 계약건수, 정비수가는 사고차량 수리 시 보험처리를 위하여 표준작업시간과 시간당 공임을 지정하여 정비업체에서 받는 비용, 물적할증기준은 대물이나 자기차량 손해보험 처리 시 처리금액에 따라 보험료가 할증되는 기준을 의미한다. Figure 3.1, Figure 3.2, Figure 3.3은 각각 심도, 빈도, 보험료의 시도표이다.

각 변수의 시도표를 살펴보면 심도의 경우 상대적으로 계절효과가 작아 보이나 매년 3분기에서 4분기로 높게 상승하는 것을 볼 수 있으며 빈도와 보험료의 경우 계절적인 효과가 심도보다 더 확연하게 나타나고 있다. 또한 각 변수에 영향을 미치는 외부변수를 살펴보면 심도의 경우 사고발생시 차량 수리비에 영향을 줄 수 있는 정비수가 인상(X_3) 및 부품값 인상(X_4, X_6)이고 빈도의 경우 사고발생률에 영향을 미

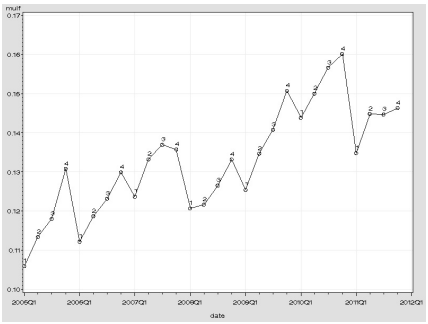

Figure 3.2. Time series plot for frequency

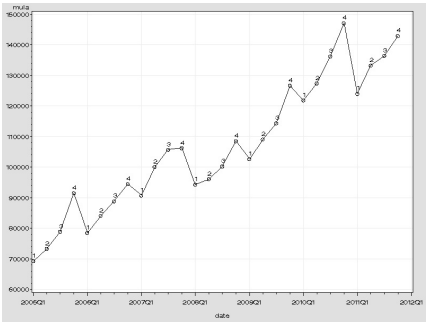

Figure 3.3. Time series plot for premium

치는 주 5일제 확대실시(X_1) 및 전면실시(X_2)와 물적할증기준 선택제(X_5)이며. 보험료의 경우 심도와 빈도에 영향을 미치는 모든 요인($X_1, X_2, X_3, X_4, X_5, X_6$)을 모두 고려해야 한다.

3.2. 가변수를 이용한 시계열회귀모형에 의한 추정

3.2.1. 심도 심도 자료에 대하여 가변수를 이용한 시계열 회귀모형에 적합시킨 결과, 모형화에 사용된 외부변수는 정비수가 인상과 부품값 인상에 관련된 X_3, X_4, X_6이며 최종적으로 추정된 모형은 다음과 같다.

$$\hat{z_t} = 648996 + 8795.07t - 9916.22D_1 - 14271D_2 - 13229D_3 + 22390X_3 + 22556X_6.$$

또한 각 모수의 추정치는 다음 Table 3.1과 같다.

모형의 RMSE는 9838.73149, R^2 값은 0.9849, Adj-R^2 값은 0.9792로 나타났다. 또한, 더빈-왓슨 검정결과 더빈-왓슨 D통계량 값이 1.492, 유의확률값이 0.0563으로 1차 자기상관은 존재하지 않는 것으로 나타났다.

3.2.2. 빈도 빈도 자료에 대하여 가변수를 이용한 시계열 회귀모형에 적합시킨 결과, 모형화에 사용된 외부변수는 주 5일제 실시와 물적할증기준 선택제와 관련된 X_1, X_2, X_5이며 최종적으로 추정된 모형은 다음과 같다.

$$\hat{z_t} = 0.122 + 0.001t - 0.014D_1 - 0.009D_2 - 0.005D_3 + 0.012X_5$$

Table 3.1. Parameter estimates using dummy variables for depth

Variable	*d.f*	Parameter estimate	Standard error	t-value	P-value
Intercept	1	648996.00	8094.88	80.17	< 0.0001
t	1	8795.07	434.36	20.25	< 0.0001
D_1	1	-9916.22	6119.35	-1.62	0.1247
D_2	1	-14271.00	5876.67	-2.43	0.0273
D_3	1	-13229.00	5742.66	-2.30	0.0350
X_3	1	22390.00	9105.09	2.46	0.0257
X_6	1	22556.00	8916.79	2.53	0.0223

Table 3.2. Parameter estimates using dummy variables for frequency

Variable	*d.f*	Parameter estimate	Standard error	t-value	P-value
Intercept	1	0.12242	0.00324	37.81	< 0.0001
t	1	0.00111	0.00021	5.38	< 0.0001
D_1	1	-0.01420	0.00299	-4.75	0.0002
D_2	1	-0.00923	0.00287	-3.22	0.0050
D_3	1	-0.00528	0.00284	-1.86	0.0806
X_5	1	0.01234	0.00361	3.42	0.0032

Table 3.3. Parameter estimates using dummy variables for premium

Variable	*d.f*	Parameter estimate	Standard error	t-value	P-value
Intercept	1	81355.00	2747.80	29.61	< 0.0001
t	1	2068.67	175.76	11.76	< 0.0001
D_1	1	-13161.00	2536.83	-5.19	< 0.0001
D_2	1	-9938.66	2432.06	-4.09	0.0008
D_3	1	-6295.99	2412.93	-2.61	0.0183
X_5	1	12637.00	3059.58	4.13	0.0007

또한 각 모수의 추정치는 Table 3.2와 같다.

모형의 RMSE는 0.00491, R^2 값은 0.8951, Adj-R^2 값은 0.8642로 나타났다. 또한, 더빈-왓슨 검정 결과 D통계량 값이 0.782, 유의확률값이 0.0003으로 1차 양의 자기상관은 존재하는 것으로 확인되었다.

3.2.3. 보험료 보험료 자료에 대하여 가변수를 이용한 시계열 회귀모형에 적합시킨 결과, 모형화에 외부변수는 모든 변수를 사용하였으며 최종적으로 추정된 모형은 다음과 같고

$$\hat{z}_t = 81355 + 2068.67t - 13161D_1 - 9938.66D_2 - 6296D_3 - 12637X_5.$$

각 모수의 추정치는 다음 Table 3.3과 같다.

모형의 RMSE는 4168.22, R^2 값은 0.9626, Adj-R^2 값은 0.9516으로 나타났다. 또한, 더빈-왓슨 검정 결과 D통계량 값이 0.942, 유의확률값이 0.0019로 1차 양의 자기상관은 존재하는 것으로 나타났다.

3.3. 독립변수 추가모형에 의한 적합

잔차에 1차 자기상관이 존재하는 빈도와 보험료 자료의 경우에 생각할 수 있는 방법 중 하나로 독립변수의 추가를 고려해 볼 수 있다. 새로운 독립변수를 추가하는 것은 많은 비용과 시간이 필요하지만 본 절

Table 3.4. Parameter estimates via the model with additional independent variables for frequency

Variable	*d.f*	Parameter estimate	Standard error	t-value	P-value
Intercept	1	0.0060	0.0009	6.29	< 0.0001
D_1	1	-0.0169	0.0020	-8.48	< 0.0001

Table 3.5. Parameter estimates via the model with additional independent variables for premium

Variable	*d.f*	Parameter estimate	Standard error	t-value	P-value
Intercept	1	9472.19	2321.21	4.08	0.0009
D_1	1	-15324.00	2003.85	-7.65	< 0.0001
D_3	1	-3128.85	1707.41	-1.83	0.0856
$\bigtriangledown z_{t-1}$	1	0.25	0.12	2.13	0.0493
X_3	1	-10312.00	4167.02	-2.47	0.0249
X_4	1	7911.30	3505.25	2.26	0.0383

에서는 현재의 값은 과거의 영향을 받아 실현된다는 시계열 자료의 기본적인 가정에 따라 이전 시점의 반응변수를 독립변수로 고려하여 분석한 결과 빈도자료와 보험료자료에서 모두 단위근이 존재하는 것으로 확인되어 각 자료의 차분된 시계열을 반응변수로 하여 분석을 실시하였다.

3.3.1. 빈도 빈도 자료에 대하여 독립변수 추가모형을 적합시킨 결과, 모형화에 사용된 변수 $t, D_1, D_2, D_3, z_{t-1}, X_1, X_2, X_5$ 가운데 중 계절효과를 반영하는 D_1과 이전 시점의 반응변수인 z_{t-1}이 유의미한 변수로 나타났다. 그러나 z_{t-1}의 계수가 0.9738로 1에 가까운 값을 갖는 것으로 나타났으며 단위근 검정 결과 유의확률이 0.6349로 단위근이 존재하는 것을 확인할 수 있었다. 이에 따라 차분한 빈도자료를 반응변수로 하고 시간(t), 계절효과(D_1, D_2, D_3), 차분된 이전 시점의 반응변수($\bigtriangledown z_{t-1}$)와 빈도에 영향을 미칠 것으로 생각되는 외부변수(X_1, X_2, X_5)를 독립변수로 하여 분석한 결과가 다음과 같다.

$$\widehat{\bigtriangledown z_t} = 0.0060 - 0.0169 D_1, \qquad \widehat{\bigtriangledown z_t} = z_t - z_{t-1}.$$

추정된 모형의 더빈-왓슨 D통계량 값은 0.1447, 유의확률 값이 0.2482로 1차 자기상관은 없는 것으로 확인되었으나 R^2 값은 0.7518로 가변수를 활용한 시계열 회귀모형보다 작게 나타났으며 1분기 계절효과를 반영하는 변수 D_1을 제외한 시간(t), 2분기 및 3분기 계절효과(D_2, D_3) 차분 이전 시점의 반응변수($\bigtriangledown z_{t-1}$), 외부변수(X_1, X_2, X_5)는 모두 유의미하지 않아 모형에서 제거된 것을 확인할 수 있다. 따라서 이전 시점의 반응변수는 빈도변수를 설명하는데 유의미하지 않은 것으로 보인다.

3.3.2. 보험료 보험료 자료에 대하여 독립변수 추가모형을 적합시킨 결과, 모형화에 사용된 변수 $t, D_1, D_2, D_3, z_{t-1}, X_1, X_2, X_3, X_4, X_5, X_6$ 중에서 계절효과를 반영하는 D_1과 이전 시점의 반응변수인 z_{t-1}이 유의미한 변수로 확인되었다. 그러나, z_{t-1}의 계수가 1.0557로 1에 가까운 값을 갖는 것으로 나타났으며, 단위근 검정 결과 유의확률이 0.9449로 단위근이 존재하는 것을 확인할 수 있었다. 이에 따라 차분한 빈 도자료를 반응변수로 하고 시간(t), 계절효과(D_1, D_2, D_3), 차분된 이전 시점의 반응변수($\bigtriangledown z_{t-1}$)과 빈도에 영향을 미칠 것으로 생각되는 외부변수(X_1, X_2, X_5)를 독립변수로 하여 분석한 결과가 다음과 같다.

$$\widehat{\bigtriangledown z_t} = 9472.19 - 15324 D_1 - 3128.85 + 0.25 \bigtriangledown z_{t-1} - 10312 X_3 + 7911.3 X_4$$
$$\widehat{\bigtriangledown z_t} = z_t - z_{t-1}, \quad \bigtriangledown z_{t-1} = z_{t-1} - z_{t-2}.$$

Table 3.6. Dubin-Watson statistics for frequency

Order	DW	Pr < DW	Pr > DW
1	0.7818	0.0003	0.9997
2	1.6201	0.1790	0.8210
3	2.4436	0.9103	0.0897
4	3.0545	0.9944	0.0056
5	2.8726	0.9988	0.0012

Table 3.7. Backward elimination of autoregressive parameters for frequency

Lag	Estimate	t-value	P-value
2	0.0102	0.03	0.9746
5	−0.0306	−0.11	0.9130
3	0.1474	0.60	0.5597

Table 3.8. Estimates of autoregressive parameters for frequency

Lag	Coefficient	Standard error	t-value
1	−0.4881	0.1806	−2.70
4	0.4406	0.1806	2.44

추정된 모형식을 살펴보면 더빈-왓슨 D통계량 값이 1.342, 유의확률이 0.0698로 1차 자기상관은 존재하지 않는 것으로 확인되었으나 R^2 값은 0.8584로 가변수를 활용한 시계열 회귀모형보다 낮게 나타났으며 시간 t, 2분기 계절효과를 반영하는 D_2와 외부변수 가운데 X_1, X_2, X_5, X_6는 유의미하지 않아 모형에서 제거된 것을 확인할 수 있다.

3.4. 자기회귀 오차모형

잔차에 1차 자기상관이 존재하는 경우에 생각할 수 있는 또 하나의 방법으로 오차항에 추가적인 모형을 설정해주는 것을 고려해 볼 수 있다.

3.4.1. 빈도 빈도 자료에 대하여 자기회귀 오차모형을 적합시킨 결과는 다음과 같다.

사용한 프로시져는 SAS9.2의 PROC AUTOREG를 사용하였으며 분기자료이므로 $\epsilon_t \sim$ AR(5)의 가정하에서 더빈-왓슨 검정법을 사용하였고, ϵ_t의 모형을 BACKSTEP 옵션을 통해 추정하였으며, 모수추정방법은 SAS 9.2 PROC AUTOREG에서 디폴트로 제공하고 있는 FGLS 추정법이다.

Table 3.6은 ϵ_t의 자기상관검정에 대한 결과이며 시차 1, 시차 4, 시차 5에서 자기상관이 존재하는 것을 볼 수 있다.

Table 3.7과 Table 3.8은 가정한 AR(5) 모형 중 유의미한 시차만을 출력해주는 BACKSTEP 옵션에 의한 결과이다. 앞에서 알아본 Dubin-Watson 검정결과에서 자기상관이 나타났던 시차가 출력된 것을 확인할 수 있다. 모형화에 사용된 변수는 $t, D_1, D_2, D_3, X_1, X_2, X_5$이며 최종적으로 추정된 모형은 다음과 같고 각 FGLS모수의 추정치는 Table 3.9와 같다.

$$\hat{z_t} = 0.1218 + 0.0012t - 0.0014D_1 - 0.0088D_2 - 0.0051D_3 + 0.0104X_5 + e_t,$$
$$e_t = 0.488e_{t-1} - 0.441e_{t-4}, \quad e_t = \hat{\epsilon_t}, \quad \widehat{\mathrm{Var}(a_t)} = 0.0000118.$$

가변수를 이용한 시계열 회귀모형에 의한 적합과 비교해 보면 RMSE는 0.00491에서 0.00344로 줄어든

Table 3.9. Parameter estimates via the autoregreesive model for frequency

Variable	*d.f*	Parameter estimate	Standard error	t-value	P-value
Intercept	1	0.1218	0.0025	48.42	< 0.0001
t	1	0.0012	0.0025	6.42	< 0.0001
D_1	1	-0.0014	0.0014	-10.08	< 0.0001
D_2	1	-0.0088	0.0014	-6.20	< 0.0001
D_3	1	-0.0051	0.0012	-4.10	0.0009
X_5	1	0.0104	0.0038	2.75	0.0150

Table 3.10. Dubin-Watson statistics for premium

Order	DW	Pr < DW	Pr > DW
1	0.9400	0.0019	0.9981
2	1.8797	0.3980	0.6020
3	2.4096	0.8967	0.1033
4	2.5668	0.8698	0.1302
5	2.4672	0.9747	0.0253

Table 3.11. Backward elimination of autoregressive parameters for premium

Lag	Estimate	t-value	P-value
4	0.0604	0.19	0.8559
3	0.1502	0.56	0.5851
5	0.2027	0.88	0.3926
2	0.3082	1.25	0.2287

Table 3.12. Estimates of autoregressive parameters for premium

Lag	Coefficient	Standard error	t-value
1	-0.5004	0.2164	-2.31

것을 확인할 수 있고, R^2 값은 0.8951에서 0.9546로 증가한 것을 볼 수 있다. 또한, 더빈-왓슨 검정 결과 D통계량 값이 1.7206으로 1차 자기상관은 존재하지 않는 것으로 나타났다.

3.4.2. 보험료 보험료 자료에 대하여 자기회귀 오차모형을 적합시킨 결과는 다음과 같다.

Table 3.10은 ϵ_t의 자기상관검정에 대한 결과이며, 시차 1에서 자기 상관이 존재하는 것을 확인할 수 있다. 또한 Table 3.11과 Table 3.12는 가정한 모형 중 유의미한 시차만을 출력해주는 BACKSTEP 옵션에 의한 결과이며, 앞서 수행한 더빈-왓슨 검정결과에서 자기상관이 나타났던 시차가 출력된 것을 확인할 수 있었다.

모형화에 사용된 변수는 $t, D_1, D_2, D_3, X_1, X_2, X_3, X_4, X_5, X_6$이며 최종적으로 추정된 모형은 다음과 같고 각 모수의 추정치는 Table 3.13과 같다.

$$\hat{z}_t = 79097 + 2269t - 12490D_1 - 9354D_2 - 6022D_3 + 9173X_5 + e_t,$$
$$e_t = 0.500e_{t-1} - 0.441e_{t-4}, \quad e_t = \hat{\epsilon}_t, \quad \widehat{\mathrm{Var}(a_t)} = 12776586.$$

가변수를 이용한 시계열회귀 모형에 의한 적합과 비교해 보면 모형의 RMSE는 4168.21813에서 3574로 감소하였고, R^2 값은 0.9626에서 0.9741로 증가한 것으로 나타났다. 또한, 더빈-왓슨 검정 결과 D통계

Table 3.13. Parameter estimates via the autoregressive model for premium

Variable	*d.f*	Parameter estimate	Standard error	*t*-value	*P*-value
Intercept	1	79097.00	3561.00	22.21	< 0.0001
t	1	2269.00	250.65	9.05	< 0.0001
D_1	1	-12490.00	1804.00	-8.93	< 0.0001
D_2	1	-9354.00	1923.00	-4.86	0.0020
D_3	1	-6022.00	1647.00	-3.66	0.0021
X_5	1	9173.00	3707.00	2.47	0.0249

Table 3.14. Maximum likelihood estimates of depth

Parameter	Estimate	Standard error	*t*-value	*P*-value	Lag
MA1,1	-0.7947	0.1821	-4.37	< 0.0001	1

Table 3.15. Result of Portmanteau test for depth

Lag	Chi-square	*d.f*	*P*-value	Autocorrelations					
6	4.27	5	0.511	-0.238	-0.054	0.120	-0.171	0.181	-0.168
12	12.17	11	0.351	0.183	-0.081	-0.316	0.227	-0.047	-0.054
18	15.99	17	0.524	0.108	-0.006	-0.024	-0.086	-0.030	0.068

량 값이 1.7206로 1차 자기상관은 존재하지 않는 것으로 나타났다. 추정된 모형을 보면 빈도와 보험료 모두 기존의 모형에서 오차항이 AR(5) 모형을 따른다는 가정하에 유의미한 시차만을 남기고 제거한 모형이다. 두 모형 모두 모형은 다소 복잡해졌으나 RMSE는 감소하고 R^2 값은 증가하였으며 오차항의 자기상관이 존재하지 않는 것을 확인할 수 있다.

3.5. 계절형 ARIMA 모형 및 개입모형에 의한 적합

3.5.1. 심도 심도 자료에 대하여 계절형 ARIMA 모형을 적합시킨 결과, 모형화에 사용한 개입변수는 시계열 회귀분석과 같이 정비수가 인상과 부품값 인상과 관련된 변수를 사용하였으나 개입모형 적합결과 모두 유의미하지 않았기 때문에 최종적으로 ARIMA$(0,1,1)\times(0,1,0)_4$ 모형에 적합시켜 추정된 모형은 다음과 같고 각 모수의 추정치는 Table 3.14와 같다. 또한 모형적합 후 잔차의 자기상관성을 점검하는 포트맨토 검정결과는 Table 3.15와 같다.

$$z_t = \frac{1+0.7947B}{(1-B)(1-B^4)}\,a_t.$$

모형을 추정하기 위하여 심도 변수에 추세차분 및 계절차분을 취한 후 추정하였으며 포트맨토 검정에서의 유의확률을 보면 6, 12, 18시차에서 모두 0.05보다 높으므로 잔차에 자기상관이 존재하지 않음을 확인할 수 있었다.

3.5.2. 빈도 빈도 자료에 대하여 계절형 ARIMA 모형을 적합시킨 결과, 모형화에 사용한 개입변수는 시계열 회귀분석과 같이 주 5일제 실시와 물적할증기준 선택제와 관련된 변수를 사용하였으나 개입모형 적합결과 모두 유의미하지 않은 것으로 판단되어 최종적으로 ARIMA$(0,1,0)\times(2,1,0)_4$ 모형에 적합시켜 추정된 모형은 다음과 같고 각 모수의 추정치는 Table 3.16과 같다.

$$z_t = \frac{1}{(1+0.8754B^4+0.5319B^8)(1-B)(1-B^4)}\,a_t.$$

Table 3.16. Maximum likelihood estimates of frequency

Parameter	Estimate	Standard error	t-value	P-value	Lag
AR1,1	−0.8754	0.2130	−4.11	< 0.0001	4
AR1,2	−0.5319	0.2009	−2.65	0.0081	8

Table 3.17. Result of Portmanteau test for frequency

Lag	Chi-square	$d.f$	P-value	Autocorrelations					
6	6.37	4	0.1730	0.118	0.220	−0.211	0.058	−0.357	−0.058
12	8.36	10	0.5934	−0.161	0.036	−0.132	0.041	0.068	−0.038
18	8.72	16	0.9244	−0.035	0.054	−0.018	−0.007	−0.004	−0.001

Table 3.18. Maximum likelihood estimates of premium

Parameter	Estimate	Standard error	t-value	P-value	Lag	Variable
AR1,1	1.105	0.178	6.2	< 0.0001	1	Premium
AR1,2	−0.668	0.178	−3.75	0.0002	2	Premium
NUM1	11528.000	1751.800	6.58	< 0.0001	0	X_3
NUM2	4003.500	1495.400	2.68	0.0074	0	X_4
NUM3	5234.300	1921.500	2.72	0.0064	0	X_6

Table 3.19. Result of Portmanteau test for premium

Lag	Chi-square	$d.f$	P-value	Autocorrelations					
6	5.50	4	0.2396	−0.121	0.138	−0.320	0.150	−0.224	−0.071
12	8.93	10	0.5391	0.079	0.058	−0.070	0.085	0.213	0.073
18	9.37	16	0.8974	−0.036	−0.058	0.028	−0.009	0.000	0.001

Table 3.17은 포트맨토 검정의 결과이다. 모형을 추정하기 위하여 빈도 변수에 추세차분 및 계절차분을 취한 후 추정하였으며 포트맨토 검정에서의 유의확률을 보면 6, 12, 18시차에서 모두 0.05보다 크므로 잔차에 자기상관이 존재하지 않음을 확인할 수 있다.

3.5.3. 보험료 보험료 자료에 대하여 계절형 ARIMA 모형을 적합시킨 결과, 모형화에 사용한 개입변수는 시계열 회귀분석과 같이 외부변수 모두를 개입변수로 사용하였으며, 그 가운데 정비수가 증가 및 부품값 인상과 관련된 X_3, X_4, X_6가 최종모형 적합에 사용되었다. 최종적으로 적합된 모형은 다음과 같은 ARIMA$(2,1,0)\times(0,1,0)_4$ 모형이며, Table 3.18과 같다.

$$z_t = 11528.000X_3 + 4003.500X_4 + 5234.300X_6 + \frac{1}{(1-1.105B+0.668B^2)(1-B)(1-B^4)}\,a_t.$$

Table 3.19는 포트맨토 검정의 결과이다. 모형을 추정하기 위하여 자료에 대하여 선형추세차분 및 계절차분을 취하였고, 변환 후 정상시계열이 된 것을 확인할 수 있었다. 개입모형 적합 시 개입변수의 형태로는 개입의 효과가 b시차 후에 반영되며 그 크기가 w로 일정한 경우로 모형에 반영하였고 각 모형 적합 후 포트맨토 검정에서의 유의확률을 보면 6, 12, 18시차에서 모두 0.05보다 크므로 잔차에 자기상관이 존재하지 않음을 확인할 수 있었다.

3.6. 추정치 및 예측력 비교 결과

3.6.1. 심도 심도 자료의 추정에 대한 RMSE와 각 분기별 예측치는 다음과 같다.

Table 3.20. RMSE & Estimates for depth

	심도(원자료)	시계열 회귀모형	개입모형(계절형 ARIMA)
RMSE		13596.73	13985.80
2011년 1분기	919583	903902.50	935447.00
2011년 2분기	919643	908342.80	937734.00
2011년 3분기	943317	918179.90	957854.00
2011년 4분기	976646	940204.00	1006869.00

Table 3.21. Result of Mariano test for depth

시계열 회귀모형	Diebold-Mariano 검정	
	t-value	P-value
ARIMA$(0,1,1)\times(0,1,0)_4$	-0.5172	0.3025

Table 3.22. RMSE & Estimates for frequency

	빈도(원자료)	시계열 회귀모형	자기회귀 오차모형	개입모형(계절형 ARIMA)
RMSE		0.00786	0.00623	0.00789
2011년 1분기	0.13482	0.14843	0.14692	0.14861
2011년 2분기	0.14486	0.15452	0.15294	0.15308
2011년 3분기	0.14466	0.15958	0.15698	0.15859
2011년 4분기	0.14634	0.16597	0.16403	0.16597

심도 자료의 추정을 위한 모형 중 시계열 회귀모형의 경우 각 분기의 효과 D_1, D_2, D_3와 정비수가 및 부품값 인상과 관련된 외부요인 X_3, X_4, X_6 가운데 유의미한 외부요인 X_3, X_6를 가변수로 활용하여 분석하였으며, 개입모형의 경우 정비수가 및 부품값 인상과 관련된 외부요인 중 유의미한 개입변수가 존재하지 않아 외부요인을 활용하지 않은 계절형 ARIMA 모형을 사용하였다. 추정된 심도 자료의 RMSE와 추정치 비교 결과, 가변수를 이용한 시계열 회귀모형이 계절형 ARIMA 모형보다 상대적으로 낮은 RMSE를 갖는 것을 볼 수 있으나 모든 분기에서 추정치가 낮게 추정되는 현상을 보인다. 이를 보다 자세하게 알아보기 위해 Diebold-Mariano 검정을 수행하였다. Table 3.21은 기준모형과 비교모형의 두 평균제곱 예측오차 간 유의미한 차이가 있는지를 검정해주는 Diebold-Mariano 검정 결과이다. 기준모형은 가변수를 활용한 시계열 회귀모형이며 비교모형은 ARIMA$(0,1,1)\times(0,1,0)_4$ 모형이다. 여기서 사용한 Diebold-Mariano 검정의 영가설은 '기준모형의 평균제곱 예측오차가 비교모형의 평균제곱 예측오차보다 크거나 같다'이며 검정에 사용한 프로그램은 R의 dm.test 함수를 이용하였다. 검정결과 유의확률이 0.3025로 영가설을 기각하지 못하므로 두 모형간의 유의미한 차이가 있다고 할 수 없다.

3.6.2. 빈도 빈도 자료의 추정에 대한 RMSE와 각 분기별 예측치는 다음과 같다.

빈도 자료의 추정을 위한 모형 중 시계열 회귀모형의 경우 각 분기의 효과 D_1, D_2, D_3와 주 5일제 실시 및 물적할증기준과 관련된 외부요인 X_1, X_2, X_5 가운데 유의미한 외부요인 X_5을 가변수로 활용하여 분석하였으며, 독립변수 추가모형에서는 전 시점의 관측값을 독립변수로 추가하여 분석하였으나 원자료를 그대로 사용할 경우 단위근이 존재하는 것을 확인하였다. 이에 따라 차분된 시계열의 적합결과 추가한 독립변수가 유의미하지 않은 것으로 나타났다. 또한, 자기회귀 오차모형에서는 오차항이 AR(5)를 따른다는 가정하에 분석하였으며 개입모형에서는 주 5일제 실시와 물적할증기준과 관련된 외부요인 중 유의미한 개입변수가 존재하지 않아 외부요인을 활용하지 않은 계절형 ARIMA 모형을 사용하였다. 추정된

Table 3.23. Result of Mariano test for depth

자기회귀 오차모형	Diebold-Mariano 검정	
	t-value	P-value
시계열 회귀모형	-2.4637	0.0069
ARIMA$(0,1,0)\times(2,1,0)_4$	-3.4477	0.0003

Table 3.24. RMSE & Estimates for premium

	보험료 (원자료)	시계열 회귀모형	독립변수 추가모형	자기회귀 오차모형	개입모형 (계절형 ARIMA)
RMSE		4967.73	6573.28	4742.35	8694.73
2011년 1분기	123977	132497.78	139066.82	134695.03	141922.46
2011년 2분기	133219	137787.19	142877.93	139012.15	148006.82
2011년 3분기	136458	143496.53	150185.17	144068.30	152368.20
2011년 4분기	142922	151859.19	157194.65	152087.06	163833.05

Table 3.25. Result of Mariano test for premium

자기회귀 오차모형	Diebold-Mariano 검정	
	t-value	P-value
시계열 회귀모형	-0.2622	0.3966
독립변수 추가모형	-0.2154	0.4147
개입모형	-1.1656	0.1341

빈도변수의 RMSE와 추정치 비교 결과, 모든 모형이 원자료보다 큰 값으로 추정되었으나 자기회귀 오차모형이 상대적으로 가장 작은 RMSE를 갖는 것을 볼 수 있다.

Table 3.23은 자기회귀 오차모형을 기준모형으로 하고, 가변수 시계열 회귀모형, ARIMA$(0,1,0)\times(2,1,0)_4$ 모형을 비교모형으로 설정하고 각각 Diebold-Mariano 검정을 수행한 결과이다. 검정결과 자기회귀 오차모형이 가변수를 이용한 시계열 회귀모형, ARIMA$(0,1,0)\times(2,1,0)_4$ 모형과 비교하여 모두 영가설을 기각할 수 있었다. 또한 빈도 자료에서는 자기회귀 오차모형이 가장 우수한 모형으로 확인되었다.

3.6.3. 보험료 보험료 자료의 추정에 대한 RMSE와 각 분기별 예측치는 다음과 같다.

보험료 자료의 추정을 위한 모형 중 시계열 회귀모형의 경우 각 분기의 효과 D_1, D_2, D_3와 심도와 빈도에 영향을 미치는 모든 외부요인 $X_1, X_2, X_3, X_4, X_5, X_6$ 가운데 유의미한 외부요인 X_5을 가변수로 활용하여 분석하였으며, 독립변수 추가모형에서는 전 시점의 관측값을 독립변수로 추가하여 분석하였고, 자기회귀 오차모형에서는 오차항이 AR(5)를 따른다는 가정하에 분석하였다. 또한, 개입모형에서는 외부요인 중 유의미한 개입변수 X_3, X_4, X_6를 활용하여 개입모형을 사용하였다. 추정된 보험료변수의 RMSE와 추정치의 비교결과 심도에서와 같이 자기회귀 오차모형이 가장 작은 RMSE를 보였고, 시계열 회귀모형, 독립변수 추가모형, 개입모형의 순서로 RMSE가 작았으며, 모든 모형에서 원자료보다 큰 값으로 추정된 것을 Table 3.24에서 확인할 수 있다.

Table 3.25는 자기회귀 오차모형을 기준모형으로, 가변수를 이용한 시계열 회귀모형, 독립변수 추가모형, 개입모형을 비교모형으로 설정하고 각각 Diebold-Mariano 검정을 수행한 결과이다. 검정결과 자기회귀 오차모형이 가변수를 이용한 시계열 회귀모형, 독립변수 추가모형, 개입모형과 비교하여 영가설을 기각할 수 없었다.

4. 결론

보험료를 추정하는 방법으로 심도 및 빈도의 추정을 통해 두 추정치의 가중평균을 이용하여 보험료를 추정하는 방법과 실제 관측된 보험료를 통해 보험료를 직접 추정하는 방법이 있다.

본 연구에서는 적정한 보험료 산정을 위해 보험료, 심도, 빈도를 추정하는 몇 가지 모형들을 소개하고 실제 자동차 대물 보험료 자료를 이용하여 각 모형들에 적합시켜 보았다. 자료를 각 모형에 적합시킨 결과, 현재 업계에서 주로 사용되는 가변수를 이용한 시계열 회귀모형의 경우 빈도와 보험료에서 잔차에 자기상관이 존재하는 것을 확인할 수 있었다.

이러한 문제를 해결할 수 있는 방안으로 고려한 독립변수 추가모형과 자기회귀 오차모형에서 잔차에 자기상관이 더 이상 존재하지 않음을 확인할 수 있었으나, 독립변수 추가모형에서는 원자료를 그대로 이용할 경우 빈도 자료와 보험료 자료 모두 단위근이 존재하는 것을 알 수 있었고 차분된 자료를 이용할 경우 빈도 자료에서는 추가적으로 고려한 독립변수가 적합하지 않은 것으로 나타났다. 또한 계절형 데이터 적합에 널리 사용되고 있는 계절형 ARIMA 모형 및 개입모형의 적합 결과, 심도 및 빈도 자료에서는 개입변수들의 효과가 나타나지 않았으며 보험료 자료에서는 정비수가 인상 및 부품값 인상이 영향을 미치는 것으로 나타났으나, 최종적으로 추정된 예측치 비교결과 모든 모형에서 실제 관측값보다 큰 값을 보였으며 개입모형을 제외하고는 모형 간 추정치의 차이가 크지 않았다.

RMSE 비교결과, 자기회귀 오차모형 가장 작은 RMSE를 갖는 것으로 나타났으나 Diebold-Mariano 검정결과 심도 자료와 보험료 자료에서는 모형간 유의미한 차이를 보이지 않았으며 빈도 자료의 경우 자기회귀 오차모형이 가장 우수한 모형으로 판단되었다.

References

Box, G. E. P and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems, *Journal of the American Statistical Association*, **65**, 70–79.

Cho, S. and Sohn (2009). *Time Series Analysis using SAS/ETS*, Yulgok Press, Seoul.

Cummins, J. A. and Powell, A. (1980). The performance of alternative models for forecasting automobile insurance paid claim costs, *ASTIN Bulletin*, **11**, 91–106.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy, *Journal of Business and Economic Statistics*, **13**, 253–263.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association*, **74**, 427–431.

Durbin, J. and Watson, G. (1950). Testing for serial correlation in lest squared regression I, *Biometrika*, **37**, 409–428.

Durbin, J. and Watson, G. (1951). Testing for serial correlation in lest squared regression II, *Biometrika*, **38**, 159–178.

Durbin, J. and Watson, G. (1971). Testing for serial correlation in lest squared regression III, *Biometrika* **58**, 1–19.

Ljung, G. and Box, G. (1979). On a measure of lack of fit in time series models, *Biometrika*, **66**, 265–270.

Park, Y. and Kim, K. (2003). *Time Series Analysis I using SAS/ETS*, Free Academy, Seoul.

Park, Y. and Song, S. (1998). *Economic Data Analysis using SAS/ETS*, Jeong-Il press, Seoul.

시계열 회귀모형에 근거한 자동차 보험료 추정

김영화[a,1] · 박원서[b]

[a]중앙대학교 응용통계학과, [b]중앙대학교 대학원 통계학과

(2012년 10월 12일 접수, 2013년 1월 14일 수정, 2013년 2월 11일 채택)

요 약

보험료 및 보험료 구성요소에 대한 예측모형은 합리적인 보험료 결정에 필수적이다. 본 연구에서는 가변수 회귀모형, 독립변수 추가모형, 자기회귀 오차모형, 계절형 ARIMA 모형, 개입모형 등 적정한 자동차 대물 손해보험료 추정에 사용되는 다양한 모형을 소개하였다. 또한 실제 자동차 대물 보험료 자료를 이용하여 각 모형을 이용하여 보험료, 심도, 빈도 등을 추정하였으며, 모형의 추정결과는 추정치와 실제 자료값의 차이에 근거한 RMSE(Root Mean Squared Errors) 값을 통해 비교하였다. 실제 자료 분석 결과, 자기회귀 오차모형이 가장 좋은 성능을 보여주는 것을 알 수 있었다.

주요용어: 더빈-왓슨 통계량, 보험료, 빈도, 시계열, 심도, 회귀모형.

이 논문은 2011년도 중앙대학교 연구장학기금 지원에 의한 것임.

[1]교신저자: (156-756) 서울특별시 동작구 흑석동 221, 중앙대학교 경영경제대학 응용통계학과, 교수.
E-mail: gogators@cau.ac.kr