응용통계연구 (2011)
24(4), 567–573
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.4.567

가정용수의 수요량 예측을 위한 통계적 모형 비교

명성민[1] · 이두진[2] · 김화수[3] · 조진남[4]

[1]중원대학교 의료보건학부, [2]한국수자원공사, [3]도화엔지니어링, [4]동덕여자대학교 정보통계학과

(2011년 3월 접수, 2011년 6월 채택)

요 약

본 연구는 3년간 가정용수의 실측사용량 자료를 바탕으로 표본가구의 가구특성, 주택특성, 월 특성을 나타내는 항목들을 조사하여 가정 용수 수요예측모형을 개발하는 것이다. 그러나 가정용수 사용량의 분포가 왼쪽으로 치우쳐져 있는 형태를 가지므로 정규분포를 따르지 않는다. 따라서 반응변수가 정규분포를 가정하는 다중회귀모형 적용 시 추정치가 편의 되며, 모형의 설명력이 매우 낮은 결과를 초래한다. 그리고 자료의 대용량화로 인하여 오차분산이 매우 작아지므로 분산분석표에 나타나는 설명변수들의 검정 시 항상 유의하게 나타나는 결과를 초래한다. 이에 대한 대안으로 와이블 회귀모형 및 대수정규 회귀모형을 이용하여 가정 용수 수요량 예측 모형을 통계적으로 분석하고자 한다. 분석결과를 토대로 가정용수의 수요예측, 수요관리 정책수립, 수도 관련 기자재 및 시설 규격결정 등에 기초자료로 활용될 수 있을 것으로 기대된다.

주요용어: 대수정규회귀, 와이블 회귀, 예측모형, 물 사용경향.

1. 서론

우리나라 상수도 통계 (환경부, 2006)에 따르면, 생활용수 사용량은 유수수량 기준으로 가정용이 2,939백만m^2(65.6%)로 가장 많고, 다음은 영업용 826백만m^2(18.4%), 업무용 612백만m^2(13.7%), 욕탕용 107백만m^2(2.4%) 순이다.

생활용수 중 가장 많은 비중을 차지하는 가정용수의 경우 주거형태, 거주자 생활양식, 주택구조 등의 내부적인 요인과 온도, 날씨, 수도요금 등과 같은 다양한 외부요인들이 영향을 미치게 된다. 그러나 가정용수에 대하여 여러 영향 인자에 따른 사용량, 사용패턴 등에 대하여 조사된 사례는 많지 않다.

미국, 영국 등의 선진국에서는 용수의 수요관리, 수도시설의 적정규모 산정 등을 목적으로 가정용수의 용도별 사용량에 대한 모니터링을 지속적으로 시행하고 있으며, 실측조사의 결과를 바탕으로 정책시행의 효과를 검증하고 향후 방향 설정 등에 활용하고 있다.

2005년 미국의 덴버 수도국에서는 과거 10년간 가정용수의 수요 패턴이 어떻게 변화했는지를 평가하기 위하여 단독주택 용수사용량을 조사하였으며, 특히 2002년부터 2004년 사이에 극심한 가뭄으로 인하여 강력한 절수정책을 시행하고 그 효과를 검증하였다 (Denver Water, 2006).

용수사용량에 영향을 미치는 인자에 대한 분석은 수요량 예측모형을 개발하는 기초가 된다. Cochran과 Cotton (1985)은 물 소비에 영향을 미치는 사회, 환경, 경제적 요인을 고려하여 정책입안자들에게 도시의 장기 생활용수 수요를 예측할 수 있는 방법론을 제시하였다.

[4]교신저자: (136-714) 서울시 성북구 하월곡동 23-1, 동덕여자대학교 정보통계학과, 교수.
E-mail: jinnam@dongduk.ac.kr

우리나라에서도 용도별 물 사용량에 대한 기초자료를 확보하고 사용패턴을 분석하고자 2002년부터 2006년 사이에 전국 140여 표본가구를 대상으로 가정용수 용도별 사용량을 측정하여 가정용수의 수요예측모형을 다중회귀분석을 이용하여 제시하였다 (한국수자원공사, 2006).

그러나, 측정된 자료는 왼쪽으로 치우쳐 있는 형태(left-skewed)로 나타나 정규분포를 따르지 않는다. 또한 다중회귀모형 적용 시 추정치가 편의하며, 모형의 설명력이 매우 낮게 나타났다. 그리고 자료의 대용량화로 인하여 오차분산이 매우 작아지므로 분산분석표에 나타나는 설명변수들의 검정시 매우 유의하게 나타난다.

따라서 본 연구에서는 가정용수의 1인당 물 사용량(liter per capita per day; ℓpcd)에 대한 가장 적절한 회귀모형을 찾고자 한다.

첫째, 다중회귀모형 대신 와이블 회귀모형 혹은 대수정규 회귀모형을 고려한다. 그 이유는 용수자료가 왼쪽으로 치우쳐 있는 형태를 가지므로 정규분포를 따르지 않는다고 판단되기 때문이다. 이에 대한 적합도를 확인하기 위해 로그우도함수(log-likelihood)와 AIC 및 척도모수를 이용하여 최적의 분포모형을 결정한다.

둘째, 용수자료의 대용량화로 인한 오차분산이 너무 작아지는 것을 방지하기 위하여 가구별 일일 자료의 형태를 가구별 월별 자료의 형태로 변환시킨다.

제시된 모형을 바탕으로 본 연구는 다음과 같이 구성된다. 용수자료에 대한 취득방법 및 이상치 제거방법, 와이블 회귀모형 및 로그정규회귀모형에 관하여 2장에서 설명하고, 3장에서는 가구별 월별 자료를 대상으로 와이블 회귀모형을 적용한 결과를 살펴보고, 4장에서는 결론 및 본 연구의 성과에 대해서 설명한다.

2. 자료소개 및 분석방법

2.1. 자료수집

본 연구에서 사용된 자료는 가정용수의 용도별 유량조사를 위한 55개 시·군에서 표본 선정된 140가구 중에서 설문을 실시하지 못한 4가구를 제외한 136가구에서 조사하였다 (한국수자원공사, 2001). 표본가구에서 용도별 유량자료를 취득하기 위해서 전자식 유량계에 데이터 저장과 무선전송이 가능한 로거(logger)를 부착한 유량모니터링 시스템을 제작하여 각각의 수도꼭지에 설치하였으며, 유량 실측기간은 2002년∼2006년이다.

2.2. 이상점의 제거 및 자료의 변환

용도별 용수 측정 자료에서 이상점이 존재하는 경우는 측정기계의 오류 등으로 인해 관찰치가 다른 관찰치 들과 크게 다른 경우라 판단하였다. 이상점의 제거 방법은 탐색적 데이터 분석(exploratory data analysis)에서 이용되는 외부상한(upper outer fence)을 이용하였다 (Tukey, 1977). 외부상한은 $Q3 + 3 * \text{IQR}$로 정의하는데 Q3은 3사분위수로서 자료의 75%에 해당되는 값이며, IQR은 사분위 편차(inter quantile range)로서 3사분위수와 1사분위수의 차이를 의미한다. 본 연구에서는 다음과 같은 절차로 이상점을 제거하였다.

첫 번째 단계로 기계의 오차로 인하여 나타난 음수 값을 제거하였으며, 두 번째 단계로 1인당 용도별 용수 사용총량에 대하여 가구별 외부상한을 초과하는 사용량을 제거하였다. 마지막으로 두 번째 단계가 적용된 자료에서 한번 더 가구별 외부상한 초과 사용량을 제거하였다. 그 이유는 과다하게 나타난 자료를 엄격하게 제거하기 위해서이다. 이상점 제거 후 가정용수의 1인당 물 사용량을 계산하였다. 여기서,

가구별 일일자료의 형태를 가구별 월별 자료의 형태로 변환하였는데, 이는 자료에 대한 안정성을 확보하며, 또한 오차분산이 작게 나타나는 형태를 보완하기 위한 것이다.

2.3. 와이블 회귀모형의 설정

양의 값을 취하는 용수사용량 t는 와이블 분포(weibull distribution) 또는 대수정규분포(lognormal distribution)를 따른다고 가정한다. 용수사용량 t가 와이블 분포를 취할 때의 밀도함수는 다음과 같다 (조진남과 백재욱, 2002; Zhou 등, 2000; Lloyd, 1993 참조).

$$f(t) = \frac{\varphi}{\lambda}\left(\frac{t}{\lambda}\right)^{\varphi-1}\exp\left[-\left(\frac{t}{\lambda}\right)^{\varphi}\right], \quad t>0, \tag{2.1}$$

여기서 λ는 척도모수(scale parameter), φ는 형상모수(shape parameter)이다. 용수사용량 t를 $y=\ln(t)$로 변환시킬 때 y는 다음과 같은 극단값 분포(extreme-value distribution)의 밀도함수를 가진다.

$$f(y) = \frac{1}{\sigma}\exp\left\{\frac{y-\mu}{\sigma} - \exp\left(\frac{y-\mu}{\sigma}\right)\right\}. \tag{2.2}$$

이 때 $\mu = -\ln\lambda$와 $\sigma = \varphi^{-1}$는 각각 위치모수와 척도모수가 된다. 용수사용량 t가 와이블 분포를 따르고, 공변수(covariate)에 의하여 영향을 받을 때 다음과 같은 와이블 회귀모형(weibull regression model)을 설정할 수 있다.

$$Y_i = \ln(t_i) = \underline{x_i^T\beta} + \sigma\epsilon_i, \quad i=1,2,\ldots,n, \tag{2.3}$$

여기서 $\{t_i\}$는 용수사용량, $x_i^T=(1,x_{i1},x_{i2},\ldots,x_{ip})$는 설명변수들의 벡터, $\underline{\beta}=(\beta_0,\beta_1,\ldots,\beta_p)$는 회귀계수들의 벡터이다. 오차 $\{\epsilon_i\}$는 서로 독립이며 같은 분포를 따른다. t_i가 와이블 분포라면, 오차 ϵ_i는 표준 극단값 분포(standard extreme-value distribution)이다. 와이블 회귀 모형에서 회귀모수들을 추정하기 위하여 데이터가 주어졌을 때의 우도함수(maximum likelihood function)는 아래와 같다.

$$L\left(\underline{\beta},\sigma^2\right) = \prod\left(\frac{1}{\sigma}\right)\exp\left\{\left[\left(y_i-\underline{x_i^T\beta}\right)/\sigma\right] - \exp\left[\left(y_i-\underline{x_i^T\beta}\right)/\sigma\right]\right\}. \tag{2.4}$$

우도함수를 최대로 해주는 회귀계수의 최우추정치(MLE; Maximum Likelihood Estimator)를 구할 수 있다. θ를 $(\underline{\beta},\sigma^2)$라 가정하고, 우도비 검정에서 $H_0:\theta_i=0$에 대한 가설검정에서 검정 통계량

$$\Gamma = -2\ln\frac{L(\hat{\theta}_{(-i)})}{L(\hat{\theta})} \tag{2.5}$$

을 사용하며, $L(\hat{\theta}_{(-i)})$는 i번째 설명변수가 빠졌을 때의 우도함수이다. 귀무가설이 사실이라는 조건하에서 Γ는 자유도 1인 점근적 χ^2분포(asymptotic χ^2 distribution)를 따른다.

2.4. 대수정규 회귀모형의 설정

용수사용량 t가 대수정규분포를 할 때의 밀도함수는 다음과 같다 (Wu와 Hamada, 2000; Meeker와 Escobar, 1998 참조).

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}}t^{-1}\exp\left(\frac{-(\ln t-\mu)^2}{2\sigma^2}\right), \quad at>0, \tag{2.6}$$

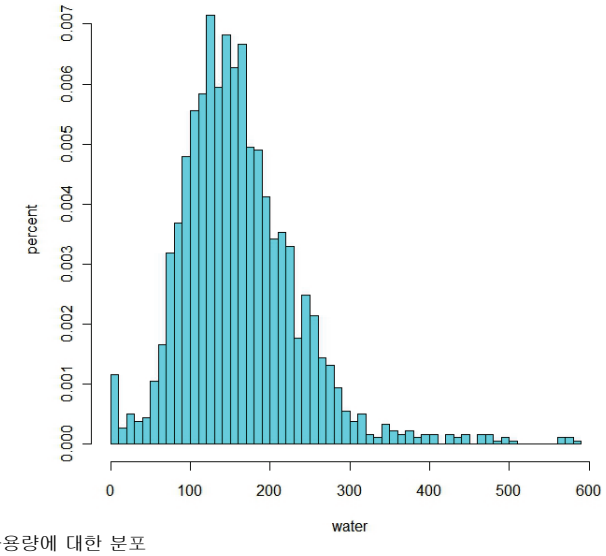

그림 **3.1.** 1인당 물 사용량에 대한 분포

여기서 μ와 σ^2은 용수사용량 t를 $y = \ln(t)$로 변환시킬 때 평균과 분산이 되므로, $y = \ln(t)$는 평균 μ 분산 σ^2인 정규분포를 따른다. 따라서 $t \sim LN(\mu, \sigma^2)$으로 표시할 수 있다. 용수사용량 t가 대수정규분포를 따르고, 설명변수들에 의하여 영향을 받을 때 다음과 같은 대수정규 회귀모형(lognormal regression model)을 설정한다.

$$y_i = \ln(t_i) = \underline{x_i^T \beta} + \sigma\epsilon_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \sigma\epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.7}$$

t_i가 대수정규분포를 따르면, 오차 ϵ_i는 평균 0, 분산 1인 표준정규분포를 취한다. t_i가 대수정규분포를 따를 때의 대수정규회귀모형의 우도함수는 아래 식과 같다.

$$L(\underline{\beta}, \sigma^2) = \prod (2\pi\sigma^2)^{\frac{1}{2}} \exp\left\{\left(-\frac{1}{2}\right)\left[\left(y_i - \underline{x_i^T \beta}\right)/\sigma\right]^2\right\} \tag{2.8}$$

와이블 회귀모형과 마찬가지로 우도함수를 최대로 해주는 회귀계수의 최우추정치를 구할 수 있으며, 우도비 검정 역시 와이블 회귀모형과 같은 방법으로 회귀계수들을 검정할 수 있다.

3. 분석결과

본 절에서는 가정용수의 1인당 물 사용량에 대한 분포의 적합도를 판정하고, 가정한 회귀모형들 중 가장 적합한 모형의 회귀모형을 찾아 그 결과를 제시한다. 자료에 대한 분석은 SAS V9.2의 LIFEREG 프로시저를 이용하였다.

3.1. 1인당 물 사용량에 대한 분포

1인당 물 사용량에 대하여 분포는 그림 3.1과 같다. 분포의 모양이 왼쪽으로 치우쳐 있어 정규분포를 따르지 않는다고 판단된다. 따라서 본 연구에서는 자료의 분포모형을 와이블분포 또는 대수정규분포를 가정하여 다음 절에서 최적의 분포모형을 결정한다.

표 **3.1.** 모형적합성에 대한 기준값

분포	로그우도함수값	AIC	척도모수
와이블	−960.341	−1956.682	0.3604
대수정규	−1093.490	−2222.981	0.4444
정규	−10256.505	−20549.010	68.4298

표 **3.2.** 와이블 회귀모형 결과

변수		추정치	표준오차	카이제곱값	유의확률
절편		5.4520	0.0609	8012.71	<.0001
실거주인원		−0.1976	0.0091	476.27	<.0001
건평		0.0065	0.0012	29.40	<.0001
방의 개수		0.0790	0.0218	13.07	0.0003
주택유형	단독	−0.0480	0.0426	1.27	0.2604
	아파트	0.0488	0.0350	1.94	0.1635
	공동주택	0.0000	·	·	·
월	1	−0.0303	0.0375	0.65	0.4186
	2	0.0106	0.0377	0.08	0.7796
	3	−0.0203	0.0426	0.23	0.6334
	4	0.0382	0.0428	0.79	0.3732
	5	0.0924	0.0431	4.60	0.0319
	6	0.1201	0.0426	7.95	0.0048
	7	0.1594	0.0410	15.09	0.0001
	8	0.1181	0.0399	8.74	0.0031
	9	0.0779	0.0395	3.90	0.0484
	10	0.0528	0.0381	1.92	0.1661
	11	0.0173	0.0376	0.21	0.6461
	12	0.0000	·	·	·
척도모수		0.3604	0.0064		

3.2. 회귀모형에 대한 비교

분포별 회귀모형을 설정하기 위하여 독립변수는 실거주인원, 건평, 방의 개수, 주택유형, 월별 사용량으로, 종속변수는 1인당 물 사용량으로 하였다. 독립변수의 설정은 Foster과 Beattie (1979)의 연구와 Cochran과 Cotto (1985)에서 제안되었던 물소비에 영향을 미치는 사회, 환경적 요인을 고려하여 제안되었던 변수들을 선별하였다. 사회적 요인으로서는 주택특성(건평, 방의 개수)과 가구특성(실거주인원, 주택유형)이며, 환경적요인은 월 효과를 의미한다. 모형의 비교를 위한 기준값으로 로그우도함수 및 AIC, 척도모수를 이용하였다. 일반적으로 관찰된 자료에 대해서 로그우도함수와 AIC 값은 크면 클수록, 척도모수는 작을수록 모형이 적합하다고 판정한다 (Cantor, 2003). 이 기준에 따르면 세 개의 분포를 가정한 회귀모형 중 와이블 회귀모형이 적절하다고 판단하였다 (표 3.1).

3.3. 와이블 회귀모형의 추정결과

5개 변수를 독립변수로, 1인당 물 사용량(ℓpcd)를 종속변수로 하여 와이블 회귀분석을 실시하여 물 사용량에 영향을 미치는 변수를 추출하였다. 실거주인원, 건평, 방의 개수, 12월에 비하여 5월∼9월의 항목이 유의하게 나타났다 (표 3.2). 실거주인원의 경우 1인 물 사용량의 로그값이 -0.1976ℓpcd만큼 감소하는 경향으로 나타났는데, 이는 향후 기본 물 사용량 관리 차원에서 거주 인구분포에 대한 면밀한 관

칠이 필요하다는 것으로 판단된다. 건평과 방의 개수의 경우 양의 관계가 나타났다. 건평의 경우, 평수가 1단위 증가하면 물 사용량의 로그값이 0.0065ℓpcd만큼 증가하며, 방의 개수가 증가할수록 물 사용량의 로그값이 0.079ℓpcd만큼 증가함을 확인하였다. 주택유형의 경우 유의하지는 않지만 아파트, 공동주택, 단독주택 순으로 사용량이 많았다. 월별 사용량에 대한 결과를 확인하여 보면 12월에 비하여 기온이 높아지는 여름철(5월∼9월)에 사용량이 증가하며, 겨울철(1월∼3월)에는 12월과 동일한 수준의 사용량을 보여주고 있다.

4. 결론

본 연구에서는 가정용수의 실측사용량 자료를 바탕으로 사회/환경적 요인(표본가구의 가구특성, 주택특성, 월효과)을 나타내는 항목들을 조사하여 가정용수 수요예측모형을 개발하였다. 가정용수 사용량의 분포를 그림 확인한 결과 왼쪽으로 치우쳐 있는 형태로서, 정규분포를 따르지 않는다는 사실을 확인하였다. 회귀모형에 대한 비교를 위해서 로그우도함수, AIC, 척도모수를 사용한 결과 와이블 회귀모형이 가장 적합한 모형임을 밝혔다. 와이블 회귀분석 결과 실거주인원, 건평, 방의 개수, 12월에 비하여 5월∼9월이 1인당 물 사용량에 영향을 미치는 주요 요인임을 알 수 있었다. 본 연구의 결과를 토대로 가정용수의 수요예측, 수요관리 정책수립, 수도관련 기자재 및 시설 규격결정 등에 기초자료로 활용될 수 있을 것으로 기대된다.

참고문헌

조진남, 백재욱 (2002). 신뢰성 향상을 위한 실험설계 및 분석, <신뢰성 응용연구>, **2**, 47–61.

한국수자원공사 (2001). <용도별 유량계 설치를 위한 표본선정 용역 수립 보고서>, 한국수자원공사.

한국수자원공사 (2006). <가정용수의 수요량 예측모델 개발 연구>, 한국수자원공사.

환경부 (2006). <상수도 통계>, 환경부.

Cantor, A. B. (2003). *SAS Survival Analysis Techniques for Medical Research*, 2nd ed., SAS Institute Inc.

Cochran, R. and Cotton, A. W. (1985). Municipal water demand study, Oklahoma city and tulas, Oklahoma, *Water Resources Research*, **21**, 941–943.

Denver Water (2006). *Post Drought Changes in Residential Water Use*, Denver Water.

Foster, H. S. and Beattie, B. R. (1979). Urban residential demand for water in the united states, *Journal of Hydrology*, **55**, 43–58.

Lloyd, W. C. (1993). *Reliability Improvement with Design of Experiments*, Marcel Dekker.

Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*, John Wiley & Sons, Inc, New York.

Neter, Kutner, Nachtsheim, Wasserman (1990). *Applied Linear Statistical Models*, 4th ad., Irwin.

Tukey, J. W. (1977). *Exploratory Data Anlysis*, Addison-Wesley.

Wu, C. F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis and Parameter Design Optimization*, John Wiley & Sons, New York.

Zhou, S. L., McMahon, T. A., Walton, A. and Lewis, J. (2000). Forcasting daily water demand: A case study of Melboune, *Journal of Hydrology*, **236**, 153–164.

A Comparison of Statistical Prediction Models in Household Water End-Uses

Sungmin Myoung[1] · Doo-jin Lee[2] · Hwa Soo Kim[3] · Jinnam Jo[4]

[1]Faculty of Health Science,Jungwon University; [2]Korea Institute of Water and Environment
[3]Dohwa Engineering Co.; [4]Department of Information and Statistics, Dongduk Women's University

(Received March 2011; accepted June 2011)

Abstract

This study develops a predictive model for household water end-uses based on data that have measured household characteristics, housing characteristics and other items, surveyed over 3 years in Korea. However, the measured data was left-skewed and it was not fitted to normal distribution. The parameter estimates were biased when using a multiple regression model. In addition, the results of the testing for the model were usually of significance due to the tiny residual from a large number of observations. In order to solve the problem, we suggested log-normal regression model and Weibull regression model as alternatives. The results of this study can be utilized in the planning stages of water and waste water facilities.

Keywords: Log-normal regression, Weibull regression, prediction model, water use pattern.

[4]Corresponding author: Professor, Department of Information and statistics, Dongduk Women's University, Seoul 136-714, Korea. E-mail: jinnam@dongduk.ac.kr