

# 베이지안 회귀모델을 활용한 5G 사물인터넷 수요 예측

박경진\* · 김태한\*\*

## Forecasting Demand of 5G Internet of things based on Bayesian Regression Model

Kyung Jin Park\* · Taehan Kim\*\*

### Abstract

In 2019, 5G mobile communication technology will be commercialized. From the viewpoint of technological innovation, 5G service can be applied to other industries or developed further. Therefore, it is important to measure the demand of the Internet of things (IoT) because it is predicted to be commercialized widely in the 5G era and its demand hugely effects on the economic value of 5G industry. In this paper, we applied Bayesian method on regression model to find out the demand of 5G IoT service, wearable service in particular. As a result, we confirmed that the Bayesian regression model is closer to the actual value than the existing regression model. These findings can be utilized for predicting future demand of new industries.

Keywords : 5G, Bayes' Theorem, Regression, Demand Forecast, IoT

Received : 2019. 02. 27.    Revised : 2019. 04. 19.    Final Acceptance : 2019. 04. 26.

※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00066, Development of time-space based spectrum engineering technologies for the preemptive using of frequency).

\* ETRI & UST, Graduate Student, e-mail : kjins90@etri.re.kr

\*\* Corresponding Author, ETRI & UST, Principal Researcher & Professor, 218, Gajeong-ro, Yuseong-gu, Daejeon, 34129, Republic of Korea, Tel : +82-42-860-5889, e-mail : taehan@etri.re.kr

## 1. 서 론

통신 기술의 기술생애주기에 따르면 새로운 통신 기술의 수명은 평균 10년이며, 이는 새로운 통신 기술이 향후 10년간의 통신 산업을 지배함을 의미한다. 이러한 통신 기술의 발전은 단순히 통신 산업에 국한되지 않고 새로운 산업을 지속적으로 파생시켜 왔다. 지금까지의 통신 기술의 핵심은 빠른 통신 서비스를 사용자에게 제공하는 것이었다. 그러나 5G 기술은 단순한 속도의 증가 외에 다른 기술적 활용도에서 주목받고 있다. 즉 5G 기술은 데이터 전송속도뿐만 아니라 다수 기기의 동시 연결, 전송 지연의 최소화 등 다른 기술적 특성이 부각되고 있다. 이는 이동통신기술에 있어서 속도 외의 특성을 활용한 산업이 발전되었으며, 5G 기술의 도입으로 인하여 기술의 활용 가능성이 증가하였기 때문이다. 5G 기술의 예로는 평창올림픽에서 보여준 자율주행자동차 시범 서비스, 경기를 더 실감나게 보여준 가상현실시스템(AR/VR), 개막식을 장식한 드론 등이 있다.

5G 기술의 활용성이 더욱 주목받는 이유는 기술이 출현한지 얼마되지 않았으며, 아직 상용화되지 않았기 때문이다. 국내 통신사에서 발표한 Kim[2018]의 연구결과에 따르면 5G 기술은 자동차, 제조, 의료 등 기존과 다른 산업 분야에서 주로 활용될 것으로 예측하였으며, 2030년에는 약 47.8조 원 정도의 사회경제적 가치를 창출할 것으로 보고 있다. 이뿐만 아니라 전체적인 산업분야에서 최소 통신 분야 외 10개의 산업분야에서 약 42.3조 원의 경제적 가치를 창출할 것으로 예측하였다. 국외에서 발표한 GSMA[2017]의 자료에 따르면 2025년에는 전 세계의 인구 중 34%가 5G 서비스를 사용할 것이며 기존의 기술보다 더 빠르게 수용될 것이라고 예상하고 있다. 이렇듯 엄청난 경제적 효과를 창출할 5G 기술 활용 산업이 성공하기 위해서는 산업에 대한 수요를 사전에 파악하는 것이 중요하다. 본 연구의 목적은 5G 기술이 활용될 것으로 예상되는 사물인터넷(Internet of things: IoT)에 대한 수요를 측정하기 위한 방법을 제시하는 것이다. An[2009]의 연구에 따르면 어떤 산업의 미래 수요에는 초기의 수요가 큰 영향을 미치며, 특히 수요가 안정적으로 형성되어 고정적인 수익을 형성하기 이전까지의 단계에서는 그 초기 수요가 산업의 발전을 좌우할 수 있으므로 이에 대한 예측이 중요하다. 따라서 본 연구에서는

수요를 예측하는 방법 중에서도 5G 이동통신과 같이 아직 산업이 본격적으로 형성되지 않은 단계에서의 수요를 측정할 수 있는 방법을 검토하고자 한다. 5G 기술을 활용한 초기 산업 중 사물인터넷 산업이 향후 큰 비중을 담당할 것으로 예측되기 때문에 사물인터넷의 수요를 측정하고자 하였다. 하지만 아직 5G 기술을 이용한 사물인터넷 산업은 상용화되지 않았으므로, 5G 서비스의 초기 수요에 대한 참고자료가 부족하여 기존 수요 예측방식을 사용하는 데에는 한계가 있다.

따라서 기존 방식을 대체할 수 있는 방안으로 확률분포를 사용하는 베이지안 모델을 활용한 베이지안 회귀모델을 사용하였다. 베이지안 모델을 사용하는 이유는 기존 방법론과 다르게 적은 자료로도 추론이 가능하다는 장점이 있기 때문이다. 분석을 위해 과기정통부에서 발표하는 유·무선가입자수 자료 중 이동통신망을 통해 제공되고 있는 사물인터넷의 가입자수 자료를 활용하였으며, 특히 사물인터넷 중 신규 산업과 유사한 특징을 보여주는 웨어러블 서비스의 가입자수에 대해 분석하였다. 이를 통해 단순회귀모델과 베이지안 회귀모델의 적합성을 비교해 보고, 적합한 방법을 선택하여 미래의 가입자 수에 대해 예측해 보았다.

## 2. 선행 연구

### 2.1 회귀 및 시계열 예측 모델

기존의 수요 예측 방법으로는 인과관계를 통한 회귀 분석방법과 시계열 분석방법이 있으며, 신상품의 시장 확산 추세를 통해 분석한 확산수요 예측 등이 있다. 이 방법론은 과거 데이터를 기반으로 종속변수와 독립변수들 사이의 관계성을 분석하는 방법이다. Lim[1992]에 따르면 기존의 방법론들은 동일 서비스의 과거 자료가 존재할 때 사용하며 예측기간에 따라 구분할 수 있는데, 단기간을 예측하기 위해서는 시계열 모형을 사용하였으며, 장기간을 예측하기 위해서는 확산모형을 주로 사용해왔다.

단기간 예측을 위한 회귀분석의 특징은 과거 데이터에서 나타나는 추세 혹은 특징이 미래에도 지속된다고 가정하는 것이다. 따라서 변수를 정확하게 정의해 주어야 하며, 독립변수와 종속변수간의 관계를 파악하는 것이 중요하다. Jung[2014]에서는 회귀모형을 이용하여 주당 최대 전력수요 예측을 하였다. 전력수요

의 영향을 주는 변수가 많으므로 단순 회귀가 아닌 다중회귀분석을 사용하였으며, 주 최대 전력수요로 변수를 제한하여 외부적으로 영향을 미칠 수 있는 요인을 제어하였다. 특히 다량의 과거 데이터를 통해 변수간의 관계를 설명하고 이 추세에 따라 예측값을 구했다는 점에서 의의가 있다.

시계열 분석은 회귀분석과 마찬가지로 과거 데이터를 기반으로 하나, 과거의 모든 자료를 동등 비중으로 이용하는 회귀분석보다는 상대적으로 최근 자료에 큰 비중을 두는 방법이다. 예측하고자 하는 변수에 대한 하나의 시계열, 즉 시간의 인과성이 존재할 경우 사용할 수 있다. 대표적인 방법으로는 단순이동평균법, 가중이동평균법, 자기회귀 누적이동평균모형, 추세반영법 등이 있다. Hwang[2015]에서는 독도관광 수요 예측을 위해 일별량 시계열 모형을 사용하였다. 이 연구는 다양한 예측 방법을 비교분석하여 시간적 특징이 있는 독도관광 수요를 예측하기 위해서는 시계열 모형이 최적이라는 것을 밝혀냈다는 점에서 의의가 있다.

## 2.2 확산수요 예측 모델

신상품이 출현하면 시장에 도입되어 확산되는 과정을 설명 혹은 예측하는 방법이 필요하고, 이를 설명하기 위해 쓰이는 모형이 성장곡선모형이다. 이를 토대로 가장 많이 사용하고 있는 신상품 수요 예측방법은 로지스틱 모형, 콤펙트 모형, Bass 모형 등이 있다. Lim[1992]과 Park[2018]의 연구에 따르면 Bass 모형은 로지스틱, 콤펙트 모형과 달리 수요를 측정하기 위해 내·외부 요인을 모두 고려하는 방법이다. 신기술의 수요자를 대중매체에 영향을 받는 외부집단인 혁신자와 구전에 의해 영향을 받는 내부집단인 모방자로 구분하였다. Bass 모형은 확률함수로부터 유도되며, 혁신계수, 모방계수, 잠재시장에 대한 추정치를 구하여 인자로 사용한다. 특히 확산 모형을 통해 기존의 회귀, 시계열 분석과는 달리 장기적인 수요를 예측하는 방안으로 사용한다. 하지만 변수인 구전계수, 마케팅지수는 상용화 이후의 실자료가 필요하다는 점에서 상용화 이전의 수요를 예측하는데 한계가 있다. 또 실자료가 있더라도 그 수가 충분하지 않다면 모수 추정치 안정성에 있어서 다른 모형보다 단점이 많아질 수 있다는 한계점을 지닌다. 또 Lim[1992]에 따르면 주관적 예측 방법으로 집단 토의법, 델파이법, AHP

등을 소개하였으나 설문조사자나 연구자의 주관으로 인해 쉽게 편향될 수 있다는 한계를 지닌다.

## 2.3 베이지안 모델

베이지안 모델을 이용한 예측 연구는 주로 인공지능, 인지과학, 기계학습 등에서 주로 시스템 예측에 사용되어 왔으며, 사회과학 분야에서는 행동 및 수요 예측에 제한적으로 사용되어 왔다. 그 이유는 베이지안 모델에서는 사전확률분포를 구하는 것이 중요한 부분을 차지하나 사회과학 연구에서는 사전확률분포를 수치로 정의하는 것이 힘들기 때문이다. 따라서 Lee[2006]에서는 베이지안 모델을 사회과학적 모델에서 사용하기 위해, 데이터 대신 설문조사를 통해 사전확률을 수치화하였다. 그러나 설문조사를 통한 사전확률분포는 설문지라는 변수에 사전확률분포가 영향을 받을 수 있으며, 설문지라는 수단 자체가 연구자의 주관에 치우칠 수 있는 한계를 지니고 있다.

Lee[2006]의 연구와는 달리 Jeon[1991], An[2007]의 연구에서는 사전정보로 설문조사를 통한 사전확률분포를 구하지 않고 수치 자료를 사용하여 수요를 예측하였다. Jeon[1991]은 일정 기간 동안 사전 총 누적수요 데이터와 수요 패턴을 활용하여 베이지안 모델을 통해 철도 이용 승객수를 예측하였다. 이 논문은 베이지안 모델을 I 과 II로 구분지어 측정했다는 점에서 의의가 있다. 모델 I에서는 외부요인을 고려하지 않은 단순 확률 베이지안 모델을 사용했으나 모델 II에서는 외부적 환경요인을 측정했다. An[2007]은 계절성에 의해 수요 변동이 큰 패션 상품의 수요를 예측하였다. 유행상품은 유행 기간이 짧으므로 충분한 데이터를 수집할 수 없어 기존의 회귀분석 방법은 적용이 어렵다고 보았다. 즉 다년간의 과거 자료가 없기 때문에 일정 기간의 판매비율을 근거로 총 수요를 예측하는 회귀모형을 제시하였으며, 누적판매의 비율을 구하는 데 있어서 누적확률분포를 사용함으로써 베이지안 모델에 회귀분석을 응용했다는 점에서 의의가 있다.

## 3. 이론적 배경 및 활용

### 3.1 베이지안 정리와 추론

베이지안 추론을 적용하기 위해서는 베이지안 정리를

전체로 한다. 베이저안 정리는 주어진 데이터를 가지고 가정된 모수를 추론하여 미래를 예측하는 기존의 예측 방법론과 달리, 확률 분포를 사용하여 사전확률 분포와 주어진 정보를 바탕으로 사후확률분포를 추정하는 방식이다[Downey, 2014]. 사전확률분포의 모수에 대한 설정은 연구자가 가지고 있는 사전정보를 바탕으로 지정할 수 있다. 이는 베이저안 정리의 한계점으로 자주 거론되지만, MCMC(Markov Chain Monte Carlo)와 같은 보완책이 존재하며, 신규 자료가 증가하면서 최초의 사전확률분포의 영향이 줄어들기 때문에 이러한 한계를 극복할 수 있다. 사후확률 분포란 사전확률분포와 주어진 데이터를 결합하여 모수에 대한 정보를 업데이트하여 추론한 확률분포이다.

기본적으로 베이저안 정리는 확률분포를 기반으로 추정치를 구하며, 기본식은 다음과 같다[Kang, 2005].

$$p(\theta|E) = \frac{p(E|\theta)p(\theta)}{p(E)} \quad (1)$$

여기서  $\theta$ 는 추정하고자 하는 모수이며,  $p(\theta)$ 는  $\theta$ 의 사전확률이다.  $E$ 는 새로 주어지는 데이터이며,  $p(E)$ 는 상수의 역할을 한다.  $p(E|\theta)$ 는  $\theta$ 에 대한 우도함수라고 한다. 이를 활용하여 사전정보와 새로운 데이터를 통하여 사후확률분포를 추론하는 방법이 베이저안 추론이다. 우도를 사용한 기본식은 "사후확률  $\approx$  우도  $\times$  사전확률"이다. 이를 수식으로 표현하면  $p(\theta|E) \propto p(E|\theta)p(\theta)$  이 된다. 우도함수를 통해 사후확률분포를 구하는 식은 다음과 같다.

$$p(x|E) = \int_{\theta} p(x|\theta)p(\theta|E)d\theta \quad (2)$$

식 (2)를 사용하여 새로운  $x$ 가 발생할 확률분포를 구한다. 즉 베이저안 추론에서 우도는  $\theta$ 라는 사건이 발생하였을 때  $x$ 가 일어날 확률값을 의미한다. 이때  $x$ 는 주어진 데이터에 따라 계속해서 변경된다. 이러한 특징으로 인해서 베이저안 추론은 최신 자료를 반영시킬 수 있어 일부 분야에서 기존의 예측 방법들을 대신해서 사용되고 있다.

### 3.2 베이저안 회귀모형

베이저안 회귀모형은 회귀분석에 베이저안 확률을

적용한 추론 모델이다[Aldrich, 2008]. 이 모델을 사용하기 위해서는 먼저 회귀분석에 대해 살펴본다. 회귀 분석은 주어진 변수들에 대해 변수 사이의 관계를 구하고, 서로 간의 적합도를 측정하는 방법이다. 특히 시간에 따른 변화나 변수의 영향력 혹은 인과 관계를 추측하고 예측하는데 주로 이용되며 선형회귀방정식을 사용한다. 회귀분석의 기본 식은 다음과 같다.

$$Y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n + \epsilon \quad (3)$$

하나의 종속변수  $Y$ 를 설명하는 독립변수  $x$ 의 개수에 따라, 하나의 독립변수  $x_1$ 만을 사용한다면 단순회귀 모델이 되며, 독립변수가 두 개 이상( $x_i \geq x_2$ )이라면 다중회귀모델이 된다. 본 논문에서는 다른 변수들의 영향력이 없다는 가정하에 단순회귀모델을 적용하였다. 또 회귀방정식의 변수를 베이저안 정리를 통해 확률변수로 변경하는 베이저안 회귀모델을 사용하였다. 베이저안 회귀모델의 기본 식은 다음과 같다[Jun, 2017].

$$Y[n] = a + bx[n] + \epsilon[n] \quad (4)$$

이 식에서 변수  $Y$ ,  $x$ ,  $\epsilon$ 를 확률변수로 취급한다. 종속 변수인  $Y$ 는 결과값을 의미하며 독립변수인  $x$ 는 예측변수라고 부른다.  $\epsilon$ 의 경우 데이터에서 발생하는 이상치를 나타낸다. 단순회귀모델과의 차이점은, 베이저안 회귀모델을 사용하기 위해 다음에 논의할 MCMC 방법을 적용하는 것이다.

### 3.3 MCMC(Markov Chain Monte Carlo) 방법

베이저안 모델을 사용하는데 있어서 현실적으로 복잡한 모형에 대해서는 추적 불가능한 적분식을 사용하게 된다. 이를 해결하기 위한 방법이 MCMC(Markov Chain Monte Carlo) 방법이다. 이 방법은 Monte Carlo 방식을 통해 Markov Chain을 구축함을 의미한다. Monte Carlo 방식은 어떤 분포에서 무수히 많은 무작위 값을 추출하는 방법이다[KB Financial Group, 2017]. 이를 통해 생성된 집단에서 각 단계가 현재 위치 이전의 단계들과 완전히 독립적으로 추출된다. MCMC 방법은, 먼저 파라미터의 대표성을 찾고, 이를 검증하며, MCMC 방법을 사용하기 위한 효율적인 방안을 찾는 순서로 적용할 수 있다.

대표성을 측정하는 방법으로는 시각적인 방법이 있다. 이는 Trace plot을 통해 파라미터의 대표성을 측정하는 것이다. Trace plot은 세로축을 파라미터 값으로 가지며, 가로축이 반복 횟수를 의미하는 연쇄 단계 그래프이다. 그래프 내에서 변수간의 중첩정도가 격리되는 지점 없이 부드럽게 중첩된다면 수렴함을 의미한다. 시각적인 방법 외에도 수치로 확인하는 방법 역시 존재하며, 이는 수렴을 나타내는 과동 내 분산에 대하여 과동간의 상대적 분산을 측정하는 방법이다. 모든 과정이 대표 표본에 귀속되어 있다면, Markov chain으로 형성되는 구간의 평균 차가 Monte Carlo에서 형성되는 파라미터들의 평균 차와 같아야 한다는 개념이다. 분석의 결과인 Gelman-Rubin 통계량 혹은 Brooks-Gelman-Rubin 통계량을 통해 수렴성을 확인할 수 있다. 결과가 완전히 수렴하면 통계량이 1.0이 되나, 1.1 이상의 값이 나온다면 수렴하지 않는다.

MCMC 방법의 결과값이 대표성을 가진다고 판단되면 대표값의 정확성을 검증해야 한다. 따라서 표본이 얼마나 독립적인지를 측정하기 위해서 자기상관성을 측정해야 한다. 자기상관성을 측정하기 위해서는 중첩된 단계에 대하여 자기상관함수(ACF: Auto-correlation Function)를 사용한다. ACF는 그래프로 표현할 수 있는데 이 그래프를 통해 독립되지 않는 결과값이 있는지 파악할 수 있다. 자기상관 내에서 독립적인 정보가 얼마나 형성되어 있는지를 파악하는 방법은 유효 표본크기(ESS: Effective Sample Size)를 측정하는 것이다. MCMC 방법의 실제 실행 횟수를 N으로 설정한다면 ESS를 구하는 식은 (5)와 같다.

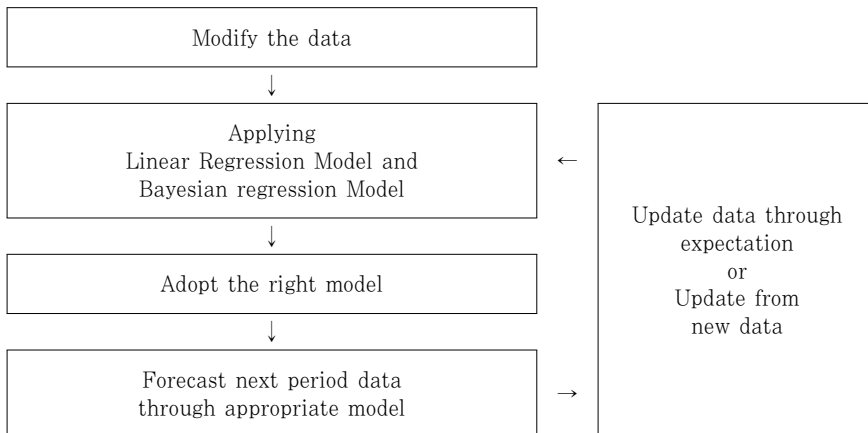
$$ESS = N / (1 + 2 \sum_{k=1}^{\infty} ACF(k)) \quad (5)$$

식 (5)에서 k는 중첩 단계가 k인 상태를 의미하며, ACF(k)는 중첩이 k인 자기상관을 의미한다. Kruschke [2014]에 따르면 이 식을 활용하여 정확하고 안정적인 결과값을 구하기 위해 일반적으로 ESS로 10000을 사용한다. 마지막으로 MCMC 방법 적용의 효율성을 측정하기 위해서는 모형을 변경하거나 측정 방법(R, SPSS, JAGS 등)을 변경하는 방법이 있다.

### 3.4 연구 절차

앞서 설명한 베이지안 모델과 MCMC 방법을 활용하기 위한 연구 절차는 <Figure 1>과 같다.

먼저 분석할 자료로 과학기술정보통신부에서 발표한 유무선 가입자 수 자료 중 사물인터넷 자료를 사용하였다. 사물인터넷의 분류는 과학기술정보통신부의 기준에 따라 차량관제, 원격관제, 무선결제, 태블릿 PC, 웨어러블, 기타 사물 인터넷 등 6가지로 분류된다. 각 산업별 자료는 2014년 10월부터 2018년 8월까지 월별로 표기된 47개의 자료로 구성되어 있다. 자료 분석을 위해 명목형 변수인 각 기간을 서열변수로 변경하였다. 예를 들어 최초 값인 2014년 10월을 1기로 선정하였으며 마지막 자료인 2018년 8월을 47기로 선정하였다. 이를 통해 각 산업별 가입자 수 분포를 측정하였으며 그 중 본 연구의 신사업 수요 예측에 알맞은 특성을 지닌 웨어러블 가입자 수 데이터를 분석대상으로 사용하였다.



<Figure 1> Process of Using Bayesian Regression Model

베이지안 회귀모델의 기본식인  $Y[n] = a + bx[n] + \epsilon[n]$ 에서, 종속변수  $Y$ 는 사물인터넷의 각 세부 수요량을 의미하며, 독립변수  $x$ 는 시기 즉 관측 시점을 의미한다. 베이지안 회귀모델을 적용하기 위해 R 통계 패키지내의 Stan 모듈을 사용하였다. Stan에서는 MCMC 방법 중 HMC(Hamiltonian Monte Carlo) 방법을 사용하고 있다. HMC 방법은 MCMC 방법 중 하나인 Metropolis algorithm의 일부를 차용한다. 본래의 Metropolis algorithm은 파라미터에 대한 공간을  $[0, 1]$ 로 지정하여 random walk를 실시하며, 사후확률이 상대적으로 높은 파라미터값을 선호하며, 현재 중심점을 기점으로 파라미터 공간 내에서 random walk가 독립적으로 시행된다. 반면 HMC에서는 Metropolis algorithm와 같은 과정을 거치지만 중심점이 변경 가능하며, 중심점에 따라서 우도가 변한다. 특히 HMC는 중심점이 변경하기 때문에 중심점의 방향을 나타내는 경사도(gradient)가 존재한다. 경사도란 사후분포가 경사가 만들어진 방향으로 증가하는 경향이 있다는 것을 암시한다. 따라서 본 논문에서는 베이지안 회귀모델을 사용하는데 있어 본래의 Metropolis algorithm 보다는 더 적합한 HMC 방법을 사용하여 MCMC 방법을 적용하였다.

### 3.5 독립 변수 설정

독립 변수 설정을 위해 웨어러블 가입자 수에 영향을 줄 수 있는 변수를 찾고, 이에 대한 단순회귀분석과 다중회귀분석을 통해 예측에 적합한 변수를 찾아보았다. 웨어러블 가입자 수에 영향을 미칠 것으로 판단되어 분석에서 사용할 변수로는 기간, 경제활동인구, 취업자 수, 이동통신 가입자 수가 선정되었다.

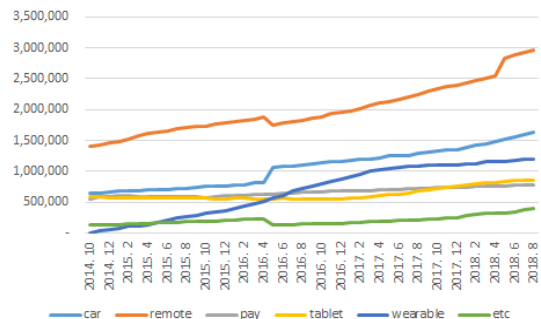
이와 같이 변수를 설정한 이유로는 기술적 요인과 소득 요인의 두 가지를 들 수 있다. 먼저 기술적 요인으로는 기간과 이동통신 가입자 수를 설정하였다. 기간이 증가할수록 기술적 진보가 발생할 가능성이 높으며, 기술의 진보는 소비자가 해당 재화를 소비할 가능성을 높여준다. 또한 기술의 진보는 Bass 모형에서 언급한 바와 같이 기술에 대한 구전효과로 인해 기간이 증가함에 따라 소비자들의 소비가 증가할 가능성을 높여준다. 따라서 기간을 변수 중 하나로 선정하였다. 이동통신 가입자 수의 경우 현재 웨어러블의 기술적 특징 상 단독 사용자 보다 이동통신과 연계하여 사용

하는 비중이 많은 것으로 추정되어 영향력 있는 변수로 설정하였다. 소득 요인의 경우 취업자 수와 경제활동인구를 설정하였다. 취업자의 경우 실질적 소득이 있기 때문에 웨어러블 기기 등의 착용 가능성이 높으며, 경제활동 인구 역시 현재 소득자와 차후 소득을 획득하는 사람으로 구성되어 있기 때문에 웨어러블 가입자 수에 영향력을 미칠 수 있는 변수로 설정하였다.

## 4. 연구 결과

### 4.1 기초 자료 분석

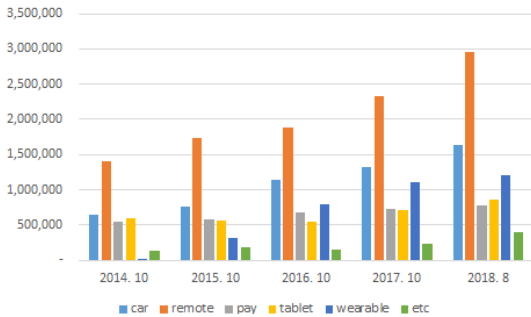
사물인터넷의 세부 분야 중 연구에 적합한 분야를 선정하기 위해 각 세부 산업별 분포 추세를 알아보았다. 각 산업별 분포는 세 가지 방식으로 분석하였다. 첫째, 각 산업별 시기별 변동 추이를 알아보았다. 둘째, 매 해 10월을 기준으로 4년간의 10월 데이터를 비교해 보았다. 단 연구 당시 2018년 자료는 8월이 마지막 관측치 자료이므로 8월의 자료를 사용하였다. 마지막으로 가장 최근 자료인 2018년 8월의 각 산업별 가입자 수 비중을 알아보았다. 각 내용은 다음 그림과 같다.



〈Figure 2〉 Trends of IoT

〈Figure 2〉은 과기정통부의 2014년 10월부터 18년 8월까지의 47개월간의 사물인터넷의 각 산업별 월별 추세를 보여준다. Car는 차량관제를, Remote는 원격관제를, Pay는 무선결제, Tablet은 태블릿 PC를, Wearable은 웨어러블 가입자 수를 의미하며, etc는 기타 산업의 가입자 수를 의미한다. 사물인터넷 산업 중 가장 많은 분포를 차지하고 있는 산업은 원격 관제이며, 가장 적은 분포를 차지한 산업은 기타 산업이다. 주목할 점은 웨어러블 산업의 수요가 급격하게

증가했다는 점이다. 반면 차량관제의 경우 2016년 4월부터 6월까지의 급격한 증가세를 제외하고 일정한 증가 추세를 보인다. 초기 웨어러블 가입자 수의 경우 산업 분포에서 가장 낮은 분포를 차지하고 있었지만 2016년 6월 무선결제 가입자 수를 초과하였다.



<Figure 3> Subscriber Distribution of Each Year

<Figure 3>은 1년을 주기로 각 연도별 10월의 산업별 가입자 수 분포를 나타낸 것이다. <Figure 2>에서 확인한 바와 같이 차량관제와 원격관제의 가입자 수는 꾸준히 증가하는 추세를 보이고, 웨어러블 가입자 수의 경우 2015년 급격한 성장을 보여 2016년부터 무선결제 가입자 수를 넘어서는 모습을 보여준다.

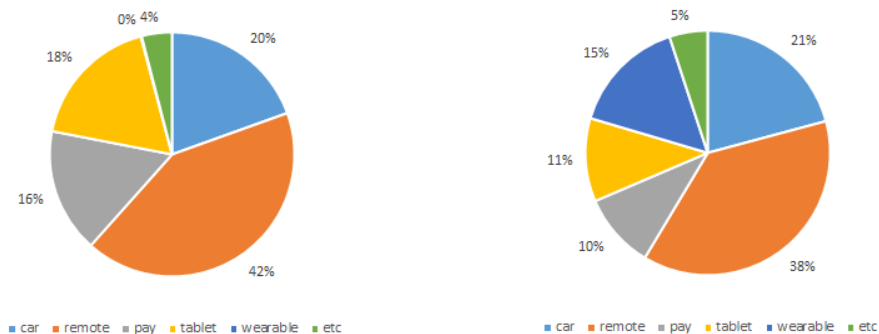
마지막으로 <Figure 4>에서 최초 자료인 2014년 10월과 마지막 자료인 2018년 8월의 가입자 수 분포를 비교해 보았다. 주목할 점은 최초 시기에 0%를 차지하고 있던 웨어러블 가입자 수가 15%까지 증가했으며, 최대치를 차지하고 있던 원격관제의 경우 <Figure 2>과 <Figure 3>에서 보았듯이 절대적인 수치는 여전히 가장 크지만 구성 비율이 감소했다는 점이다. 차량관제의 경우 가장 안정적인 비율인 20%를 유지하는

모습을 보여준다. 위의 자료분석을 통해, 가장 안정성을 보이는 자료는 차량관제 가입자 수이며, 반대로 특이성을 보이는 자료는 웨어러블 가입자 수임을 알 수 있었다. 특히 웨어러블 가입자 수는 초기에 급격하게 증가하는 추세를 볼 수 있는데 이는 초기 산업의 수요와 유사하기 때문에 대표 자료로 설정하여 연구를 진행하였다.

4.2 독립 변수 검증 결과

웨어러블 가입자수에 영향을 미칠 수 있는 것으로 설정된 각 변수에 대하여 단순회귀분석 및 다중회귀분석을 실시하였다. 각 변수의 이름으로 기간은 Num, 이동통신 이용자는 LTE, 취업자 수는 Job, 경제활동인구 수는 Econ으로 설정하여 분석을 진행하였으며 그 결과는 <Table 1>과 같다.

<Table 1>의 number of x는 변수의 개수를 의미한다. 'x = 1'은 단순회귀를 의미하며 'x > 2'는 다중회귀를 수행했음을 의미한다. 다중회귀는 4개의 변수의 모든 부분집합에 대해서 실시하였다. 따라서 Name of x는 각 부분집합에 속해 있는 원소들을 보여준다. 먼저 단순회귀분석 중 조정된 결정계수( $R^2$ )를 통해 회귀식의 적합성을 판단하였다. 분석 결과 기간 변수와 경제활동인구 변수만이 0.5 이상의 수치를 보이고 있기 때문에 이 두 변수가 유의미한 설명력을 가지고 있다고 판단할 수 있다. 하지만 경제활동인구 변수는 유의확률이 0.48로, 변수로서 웨어러블 수요와의 관계성을 설명하기 어렵다. 다중회귀분석 결과 기간변수를 포함한 다중회귀식만이 유의미한 결과를 보이고 있다. 또 F검정을 통해 알 수 있듯이 모든 전체 회귀식은 유의미 하지만 기간 변수를 제외한 각



<Figure 4> Subscriber Distribution of October 2014 and August 2018

〈Table 1〉 Results of Regression

Number of x	1							
Name of x	Num	LTE		Job	Econ			
Adj R <sup>2</sup>	0.97	0.01		0.32	0.54			
P-value	2e-16	0.18		2e-05	0.48			
Number of x	2							
Name of x	Num & LTE	Num & Job	Num & Econ	LTE & Job	LTE & Econ	Job & Econ		
Adj R <sup>2</sup>	0.96	0.96	0.96	0.33	0.57	0.55		
F-test	4.12e-33	5.42e-33	4.38e-33	4.75e-05	3.02e-09	7.6E-09		
P-value	Num		LTE		Job			
	1.2e-33	5.3e-30	2.6e-26	0.16	-0.00	0.15		
	LTE	Job	Econ	Job	Econ	Econ		
	0.36	0.59	0.40	2.3e-05	0.49	1.2e-05		
Number of x	3							
Name of x	Num & LTE & Job		Num & LTE & Econ		Num & Job & Econ		LTE & Job & Econ	
Adj R <sup>2</sup>	0.96		0.96		0.96		0.59	
F-test	9.58e-32		6.88e-32		1.21e-31		4.95E-09	
P-value	Num			LTE				
	3.92e-29		3.5e-25		2.5e-25		0.03	
	LTE	0.56	LTE	0.28	Job	0.80	Job	0.09
	Job	0.35	Econ	0.30	Econ	0.50	Econ	3.66e-06
Number of x	4							
Name of x	Num & LTE & Job & Econ							
Adj R <sup>2</sup>	0.96							
F-test	1.41E-30							
P-value	Num		LTE		Job		Econ	
	4.47e-24		0.26		0.67		0.35	

변수들은 의미를 갖지 못함을 알 수 있었다. 따라서, 기간변수를 사용하여 웨어러블 가입자 수에 대한 베이지안 회귀모형을 적용하였다.

#### 4.3 웨어러블 가입자 수 분석결과

본 연구에서는 초기 산업의 급격한 수요 증가의 특이점을 보여주는 자료인 웨어러블 가입자 수에 대해 베이지안 회귀모형을 적용하였다. 독립변수는 앞 절에서 논의한 바와 같이 기간으로 설정하였으며, 종속변수로 웨어러블 가입자 수를 설정하였다. 독립변수의 경우 기간은 명목척도이기 때문에 서열척도로 변경하였다. 베이지안 회귀모형의 적합성을 측정하기 위해 실험을 3회 실시하였다. 이를 위해 45(2018년 6월)기, 46(2018년 7월)기, 47(2018년 8월)기의 데이터를 제거한 후 각각의 데이터에 대하여 베이지안 회귀

모형과 단순선형회귀분석을 통해 45, 46, 47기의 데이터를 예측하였으며, 이를 실제값과 비교하였다. 먼저 회귀분석을 실시한 결과 〈Table 2〉와 같이 도출되었다.

회귀분석 결과 기간인 변수  $x$ 가 44, 45, 46일 때  $x$ 가 1증가할 때마다 웨어러블의 수요는 각각 약 30,865, 30,365, 29,920이 증가하였다. 회귀분석의 설명력을 보여주는 조정된 결정계수( $R^2$ ) 역시 모든 기간에서 약 0.97로, 이는 전체 웨어러블 가입자 수 변동을 설명하는데 있어서 기간 변수가 97%의 설명력을 가지고 있다는 것을 의미한다. 유의 확률 역시 모든 기간에서 0.05 이하인  $2e^{-16}$ 가 도출되었으므로 유의하다고 볼 수 있다. 따라서 앞 절의 변수에 대한 검증결과와 〈Table 2〉를 토대로 기간 변수는 웨어러블 수요를 파악하는데 유의미한 결과를 보임을 알 수 있다.

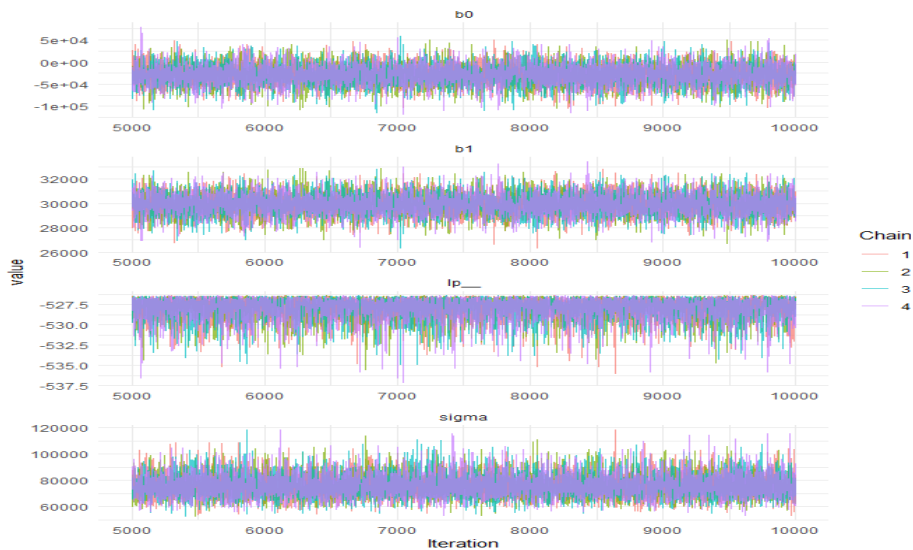


<Table 2> Regression Analysis Value of Wearable Device

x			44	45	46		
Residuals	Min		-149911	-157859	-146607		
	1Q		-41873	-43327	-46789		
	Median		-7496	-11734	-14907		
	3Q		63300	66930	67744		
	Max		115357	122699	129067		
Coefficients	Num						
		estimate Std.	30865.1	30364.5	29920.2		
		Error	795.5	813.2	819.1		
		t value	38.798	37.337	36.530		
	Pr(> t )		< 2e-16***	< 2e-16***	< 2e-16***		
Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1							
Multiple R-squared	0.973	0.970	0.968	Adjusted R-squared	0.972	0.969	0.967

베이지안 회귀모델과 회귀분석을 비교하기 위하여 동일한 독립변수와 종속변수를 사용하여 분석을 실시하였다. 베이지안 회귀모델을 적용하기 위해 MCMC 방법을 활용하여 각 변수들의 수렴성을 측정했고, 각 변수들이 베이지안 회귀모델에서 일정범위 안에 수렴함을 볼 수 있었다. 수렴성을 검증한 결과의 예시로, 46기에 대한 수렴성을 <Figure 5>에 나타내었다. 이와 같은 과정을 통해 상기 변수들을 베이지안 회귀모델에서도 사용할 수 있음을 알 수 있다. 따라서 각 변수들을 R의 Stan 모듈을 활용하여 베이지안 회귀모델을 적용하였다.

<Table 3>은 베이지안 회귀모델을 적용한 결과이다. 10,000개의 ESS를 사용하였으며, HMC가 시행된 random walk 결과 값은 20,000단계를 형성하고 있다. <Table 3>의 b0, b1, sigma, lp\_는 모수의 이름이며, lp\_는 사후확률이 높은 지점 즉 HMC의 경사도를 찾기 위한 값으로, 기존 베이지안 회귀모델 식에 로그를 취하여 사후확률을 구한 값이다. 즉, 베이지안 회귀모델 식을  $\log p(\theta|Y) = \log p(Y|\theta) + \log p(\theta) + const$ 로 변경하여 모수  $\theta$ 에 대하여 편미분 한 값이다. Rhat는 MCMC 방법의 정확도를 수치로 표현한 값으로, 이론적 배경에서 언급한 바와 같이 연쇄 값이 1.1 이상이면 수렴하지



<Figure 5> Bayes Trace Plot of Period 46

〈Table 3〉 Result of Bayes Regression Model of Period 44, 45, 46

4 chains, each wit iter = 10000; warmup = 5000; thin = 1				
post-warm up draws per chain			5000	
total post-warmup draws			20000	
	b0	b1	sigma	lp_
x = 44				
mean	-43562.29	30871.56	69128.35	-500.40
se_mean	240.88	9.32	80.04	0.02
sd	21393.81	828.20	8006.06	1.28
2.5%	-85504.91	29215.25	55644.83	-503.79
50%	-57833.15	30322.31	63505.76	-500.97
97.5%	-1166.10	32472.42	87009.71	-498.96
Rhat	1	1	1	1
x = 45				
mean	-35791.29	30364.05	72964.44	-514.44
se_mean	242.19	9.10	80.50	0.01
sd	22081.40	835.61	8164.79	1.25
2.5%	-79261.63	28724.12	59012.40	-517.62
50%	-36014.64	30366.37	72265.19	-514.12
97.5%	7807.37	31995.22	90735.47	-513.01
Rhat	1	1	1	1
x = 46				
mean	-29089.18	29937.84	75829.95	-527.92
se_mean	248.96	9.41	83.62	0.02
sd	22773.80	844.87	8459.32	1.27
2.5%	-74290.35	28270.48	61642.20	-531.17
50%	-28925.79	29933.05	74989.96	-527.60
97.5%	15028.23	31600.93	94587.22	-526.49
Rhat	1	1	1	1

〈Table 4〉 Predicted Value of Period 45, 46, 47 by Regression Model and Bayes Regression Model

	Min	2.5%	Forecasted value		97.5%	Max
x = 45						
Regression result	1381465	-	Fit : 1510783		-	1640101
Bayes result	1070832	1197654	Median	Mean	1489204	1633060
			1345002	1346320		
x = 46						
Regression result	1408474	-	Fit : 1537071		-	1665668
Bayes result	1070986	1220288	Median	Mean	1510118	1620457
			1362898	1363623		
x = 47						
Regression result	1435626	-	Fit : 1563635		-	1691645
Bayes result	965977	1200571	Median	Mean	1524148	1737575
			1361990	1362608		

않으나, 분석결과 모든 모수들의 Rhat 값이 1.1보다 작기 때문에 유의하다고 볼 수 있다. 분석결과 도출된 예측모델을 활용하여 2018년 6월, 2018년 7월, 2018년 8월인 45, 46, 47기의 데이터를 예측해 본 결과는

〈Table 4〉과 같다. 〈Table 2〉과 〈Table 3〉을 통해서 회귀분석과 베이저안 회귀모델의 기울기(b1)를 비교하면 각 기간마다 약 5, 0.5, 17만큼 차이가 나며, 이는 〈Table 4〉에 나타난 예측값의 차이에 반영되었다.

〈Table 5〉 Predicted Value of Period 48 by Bayes Regression Model

Bayes result	Min	2.5%	Median	Mean	97.5%	Max
	1021049	1230772	1392480	1392708	1555440	1811548

〈Table 4〉의 단순회귀분석 결과에 따르면 회귀분석 예측 적합값인 Fit은 각 기수별 1,510,783, 1,537,071, 1,563,635임을 확인할 수 있으며 오차범위는 1,381,465~1,640,101, 1,408,474~1,665,668, 1,435,626~1,691,645로 나타났다. 각 기수별로 예측값의 최소값인 1,381,465, 1,408,474, 1,435,626은 실제 관측치인 1,172,819, 1,200,960, 1,203,479와 모두 약 200,000정도의 차이를 보여주었다. 반면 베이지안 회귀분석 결과에 따르면 적절한 값인 Mean은 1,346,320, 1,363,623, 1,362,608임을 알 수 있다. 각 기수의 실측치는 앞서 언급한 바와 같이 1,172,819, 1,200,960, 1,203,479인데, 회귀분석과 달리 베이지안 회귀모델의 경우 5%의 유의확률, 즉 최고밀도구간(Highest Density Region: HDI)인 1,197,654~1,489,204, 1,220,288~1,510,118, 1,200,571~1,524,148 안에 실제 관측치가 속해 있음을 알 수 있다. 따라서 웨어러블 가입자 수 예측에 보다 적합한 모형은 단순회귀분석보다는 베이지안 회귀모델임을 확인할 수 있다. 이를 근거로 베이지안 회귀모델을 활용하여 자료에 주어지지 않은 48기(2018년 9월)에 대하여 예측해 보았다. 새로운 관측치인 47기의 데이터를 베이지안 회귀모델에 업데이트한 후 48기의 데이터를 예측한 결과는 〈Table 5〉와 같다.

〈Table 5〉에 따르면 48기의 웨어러블의 수요의 Mean은 1,392,708이며, 유의확률 5% 구간은 1,230,772~1,555,440으로 나타났다. 따라서 47기에 비해 48기에서는 약 30,201~31,292명 정도 증가할 것으로 예측되었다.

## 5. 결론 및 향후 연구 방향

2018년 평창 동계 올림픽을 기점으로 5G를 활용한 서비스는 점차적으로 주목받고 있다. 평창올림픽의 시작을 알리는 드론, 경기를 새로운 시각에서 보여준 AR/VR 등 사물인터넷 서비스에 대한 관심 역시 높아졌다. 특히 5G 기술을 활용한 서비스 혹은 제품에 지배적 제품이 없다는 점은 기술혁신이론(Utterback, 1996)에 따르면 무한한 성장 가능성을 지니고 있으

며, 지배제품을 선점하면 First mover로서 경제적 이익을 취할 수 있다. 따라서 5G 기술의 현 위치를 파악하여 미래를 대비해야 앞으로의 통신 서비스 사업에서 생존할 수 있을 뿐만 아니라 기술 선도국으로서의 입지를 확고히 할 수 있다. 따라서 본 연구에서는 다채롭게 등장할 5G 기술의 응용 분야 중 사물인터넷 산업의 수요 변화를 예측하기 위한 방법에 대해 연구하였다. 사물인터넷을 선정한 이유는 다른 5G 응용 산업과는 달리 이미 어느 정도 산업이 형성되어 있으며, 지배제품이 출시될 확률이 크기 때문이다. 수요를 예측하는 이유는 앞으로 등장할 신산업의 경우 경쟁에서 살아남기 위해서 초반 수요를 예측하는 것이 중요하다고 판단했기 때문이다.

기존의 자료가 필요한 선행 수요 예측 방법론은 자료의 양적인 한계로 인하여 적용하는데 제한이 존재하기 때문에, 자료의 양적 한계를 극복할 수 있는 베이지안 회귀모델을 활용하여 수요를 예측하였다. 5G 기술이 상용화되지 않았기 때문에, 공식적으로 입수할 수 있는 자료 중 과기정통부에서 발표하는 사물인터넷 가입자 수 자료를 활용하였다. 또한 그 중 새로운 산업이 출현했을 때의 특성과 유사한 행태를 보이는 웨어러블 산업의 가입자 수에 관하여 분석하였다. 베이지안 회귀모델 분석에 앞서 다중회귀분석을 통해 사전에 선정한 변수들 중에서 기간변수만이 유의미하다는 결과를 얻을 수 있어 기간변수를 사용하여 베이지안 회귀모델을 시행하였다. 분석결과 베이지안 회귀모델을 활용한 결과값이 단순회귀분석을 통한 결과값보다 더 실측치에 가까웠으며, 단순회귀분석의 결과와 달리 HDI 즉 유의확률 5%안에 실측치가 관찰되었다. 이러한 결과는 베이지안 회귀모델이 기존의 모델인 단순회귀분석보다 정밀한 결과를 도출할 수 있음을 의미한다. 또 베이지안 회귀모델의 경우 변동성이 큰 자료에서도 사용이 가능함을 살펴볼 수 있었다. 또한 기존연구들과는 달리 사전확률분포를 구함에 있어 설문 사용하지 않아 연구자의 임의성을 최대한 배제하고 객관성을 확보했다는 점에서도 의의가 있다.

하지만 베이지안 회귀모델을 활용한 결과값이 단순회귀분석의 결과값보다 실측치에 가까웠음에도 여전히

히 어느 정도의 차이가 발생하기 때문에 완벽하게 예측할 수 있다고 판단할 수는 없다. 또 분석에 있어서 비록 다른 변수에 관해서 검증은 하였지만 베이지안 회귀모델을 사용함에 있어서 수요의 증가 원인인 독립 변수를 단순히 기간 변수 하나로 상정했다는 점에서도 한계가 존재한다. 마지막으로 수요가 급격하게 증가하는 특이 자료를 사용했다는 점에서 일관성을 보이는 자료에서 베이지안이 모든 기업 수요에 정밀한 결과를 가진다고 말할 수 없다. 이러한 한계점에도 불구하고, 기존 산업에 대한 자료도 계속해서 예측 모델에 추가할 수 있으며, 5G가 본격적으로 상용화되는 시점부터는 새로운 5G 응용 산업의 실제 가입자 수를 모델에 추가할 수 있으므로 베이지안 회귀모델이 웨어러블 같이 급격하게 증가하는 특이점을 보이는 산업 초기의 수요 예측에 있어서 예측방안으로서 더 적합할 것으로 보인다. 또 5G를 응용한 산업뿐만 아니라 다른 분야에서 출현하는 수요가 급격하게 증가하는 특이점을 보이는 새로운 산업의 수요를 측정하는 데 있어서도 더욱 정확성을 높이는데 기여할 수 있다고 판단된다.

## References

- [1] Aldrich, J., "R. A. Fisher on Bayes and Bayes' Theorem", *Bayesian Analysis*, Vol. 3, No. 1, 2008, pp. 161-170.
- [2] An, B. and Youn, K., "Sales Forecast Model for Stylish Goods", *Business Management Review*, Vol. 40, No. 1, 2007, pp. 41-52.
- [3] An, B., "Empirical Bayesian Demand Forecasting Model", *Business Management Review*, Vol. 42, No. 2, 2009, pp. 77-89.
- [4] Downey, A. B., *Think Bayes*, O'Reilly, 2014.
- [5] GSMA, *The Mobile Economy 2017*.
- [6] Hwang, K., Kim, W., and Jeong, C., "Demand Forecasting of Dok-do Tourism using Comparison of Univariate Time Series", *Journal of Tourism & Leisure Research*, Vol. 27, No. 2, 2015, pp. 59-77.
- [7] Jeon, C. and Lim, H., "Forecasting Seasonal Demands Using Bayesian Approach", *Korean Management Science Review*, Vol. 8, No. 2, 1991, pp. 25-35.
- [8] Jun, S., "Technology Forecasting using Bayesian Discrete Model", *Journal of Korean Institute of Intelligent Systems*, Vol. 27, No. 2, 2017, pp. 179-186.
- [9] Jung, H., Kim, S., and Song, K., "Weekly Maximum Electric Load Forecasting Method for 104 Weeks Using Multiple Regression Models", *The Transactions of the Korean Institute of Electrical Engineers*, Vol. 63, No. 9, 2014, pp. 1186-1191.
- [10] Kang, K., *Bayesian Statistics*, Free Academy Press, 2005.
- [11] KB Financial Group Management Research Institute, "Introduction of Monte-Carlo Method and Industrial and Financial Applications", *KB Knowledge Vitamin*, Vol. 17-31, 2017.
- [12] Kim, J., Cho, I., Kim, S., Lee, C., Kim, J., and Kim, H., *5G Socioeconomic Impact Analysis*, KT Institute of Management Economics, 2018.
- [13] Kruschke, J. K., *Doing Bayesian Data Analysis : A Tutorial with R, JAGS, and Stan*, (2nd ed.), Academic Press, 2014.
- [14] Lee, K. and Jang, W., "Predicting Financial Success of a Movie Using Bayesian Choice Model", *Proceedings of Spring Joint Conference of the Korean Institute of Industrial Engineers*, 2006, pp. 1428-1433.
- [15] Lim, J. and Oh, H., "A Study on New Product Forecasting Methodology", *Journal of the Korean Institute of Industrial Engineers*, Vol. 18, No. 2, 1992, pp. 51-63.
- [16] Park, K. and Kim, T., "A Study on the Diffusion Model of 5G Mobile Communication", *JCCI 2018 Proceedings*, 2018, pp. 33-34.
- [17] Utterback, J. M., *Mastering the Dynamics of Innovation*, Harvard Business Press, 1996.

## ■ 저자소개



**박 경 진**

중앙대학교에서 경제학 학사학위를 취득하고, 과학기술연합대학원대학교 ETRI 스쿨에서 과학기술경영정책 전공 석사과정을 수료하였다. 관심 분야는 5G 및 응용산업, IoT, 수요 예측 및 통계 등이며, 한국정보기술응용학회 학술대회 등에서 5G 및 수요 예측에 관한 연구 결과를 발표하였다.



**김 태 한**

서울대학교에서 경영학 학사학위를 취득하고, 한국과학기술원 산업공학과에서 박사학위를 취득하였다. 현재 한국전자통신연구원 미래전략연구소 책임연구원 및 과학기술연합대학원대학교 과학기술경영정책전공 교수로 재직중이다. 관심 분야는 전파자원 이용정책, 5G 등 전파이용 산업, 정보통신기술 비즈니스 모델 등이다.