Journal of the Korean Data & Information Science Society
2017, **28**(3), 547–557

http://dx.doi.org/10.7465/jkdi.2017.28.3.547
한국데이터정보과학회지

베타회귀분석 방법을 이용한 건강 관련 삶의 질 자료 분석[†]

장은진[1]

[1]안동대학교 정보통계학과

접수 2017년 4월 18일, 수정 2017년 5월 14일, 게재확정 2017년 5월 17일

요 약

건강 관련 삶의 질 자료는 정규분포를 따르지 않고 치우친 분포를 보이며, 등분산 가정을 만족하지 않는 경우가 대부분이다. 또한 건강 관련 삶의 질 자료는 범위가 정해져 있는 자료이며, 건강한 상태를 나타내는 경우 최대값을 가지는 천장효과가 있는 자료이다. 본 연구에서는 건강 관련 삶의 질 자료인 EQ-5D에 대해 선형회귀모형과 베타회귀모형, 그리고 평균과 정밀도에 대한 하위모형을 가지고 있는 확장된 베타회귀모형을 이용하여 예측모형을 개발하고 모형의 예측 정확도를 비교하였다. 선형회귀모형에 비해 확장된 베타회귀모형의 예측 정확도가 높기는 하지만 신뢰구간이 겹치고 있기 때문에 확장된 베타회귀모형의 정확도가 더 높다고 할 수는 없다. 하지만 확장된 베타회귀모형은 공변량에 따라 분산이 달라지는 부분을 설명할 수 있으며 선형회귀모형이 제한된 범위를 벗어난 값을 예측하는 부분을 개선할 수 있다. 따라서 범위가 제한되고 이분산이 있는 치우친 자료에 대해 공변량들이 평균 및 정밀도에 영향을 주는 정도를 동시에 고려하는 확장된 베타회귀모형은 건강 관련 삶의 질 자료인 EQ-5D를 분석하는 방법으로 적절하다고 할 수 있다.

주요용어: 건강 관련 삶의 질, 베타회귀모형, 이분산, 한국의료패널조사.

1. 서론

건강 관련 삶의 질 (health-related quality of life; HRQOL) 자료는 임상시험 또는 환자등록자료 구축을 통한 의학연구에서 중요한 결과변수로 최근 중요성이 증가하고 있으며, 특히 비용-효용 분석과 같이 치료법의 경제성을 평가하는 연구분야에서는 특히 중요하다 (Longworth와 Rowen, 2013). 비용-효용 분석에서는 삶의 질 보정 생존년수 계산을 위해 특정한 인구집단 또는 질병의 상태에 대한 효용을 단일 지표를 이용하여 건강 관련 삶의 질을 나타내는데 (Drummond 등, 2005), 선호 기반의 건강 관련 삶의 질 도구인 EQ-5D (Dolan, 1997)와 SF-6D (Brazier 등, 2002; Brazier와 Roberts, 2004)와 같은 도구를 많이 사용한다. 특히 영국 국립보건임상연구원은 치료법에 대한 경제성 평가시 EQ-5D를 이용하여 삶의 질 보정 생존년수를 계산할 것을 권고하고 있다 (Longworth와 Rowen, 2013).

EQ-5D는 일반적인 건강상태를 측정하기 위해 EuroQoL 그룹에 의해 개발되었으며 (EuroQoL group, 1990), 운동능력 (morbility), 자기관리 (self-care), 일상활동 (usual activities), 통증/불편감 (pain/ discomfort), 불안/우울감 (anxiety/depression)의 5개 차원으로 구성되어 있으며 (Bang, 2016), 3가지 척도 (어려움 없음, 약간 어려움, 어려움 많음)로 응답하도록 구성이 되어 있는 도구이다. EQ-5D의 건강상태에 대해서는 TTO (time trade off) 방법을 사용하여 각각의 선호점수를 산출할 수 있다 (Lee 등, 2009; Jo 등, 2008).

[†] 이 논문은 2014학년도 안동대학교 학술연구조성사업에 의하여 연구되었음.
[1] (36729) 경상북도 안동시 경동로 1375, 안동대학교 정보통계학과, 조교수. E-mail: ejjang@anu.ac.kr

만일 효용을 추정하기 위해 EQ-5D와 같은 건강 관련 삶의 질 자료가 필요한데 자료에서 조사되지 않았을 경우, 건강 관련 삶의 질을 나타내는 다른 변수들을 이용하여 건강 관련 삶의 질을 추정하는 방법을 사용한다 (Longworth와 Rowen, 2013). 이를 위해 최소제곱법을 이용한 선형회귀모형이 가장 일반적으로 사용되는데, 선형회귀모형은 오차가 정규분포를 따르며 분산이 동일하다는 가정하에 수행될 수 있다. 하지만 일반적으로 건강 관련 삶의 질 자료는 오차가 정규분포를 따르지 않고 치우친 분포를 보이며, 등분산 가정을 만족하지 않는 경우가 대부분이다. 또한 건강 관련 삶의 질 자료는 범위가 정해져 있는 자료이며, 완전한 건강상태를 나타내는 경우 최대값을 가지는 천장효과 (ceiling effect)가 있는 자료이다. 따라서 선형회귀모형을 이용하여 건강 관련 삶의 질을 추정하는 방법은 부정확한 예측값을 추정하거나 예측변수의 영향 추정시 바이어스가 발생할 가능성이 있다고 할 수 있다.

Austin (2002)은 Health Utilities Index Mark 3 도구를 이용하여 측정된 삶의 질 자료를 이용하여 선형회귀모형, Tobit 모형 (Tobin, 1958), 중도절단최소절대편차 (censored least absolute deviation, CLAD) 모형 (Powell, 1984)의 예측 정확도를 비교했는데, CLAD 모형의 예측오차가 가장 작은 것으로 나타났다. Huang 등 (2008)은 천장효과를 고려하여 EQ-5D를 추정하기 위해 선형회귀모형, CLAD 모형, Two-part 모형, 잠재계급 (latent class) 모형 등의 예측 정확도를 비교했는데, 잠재계급 모형과 로그변환한 값을 이용한 Two-part 모형의 예측 정확도가 다른 모형에 비해 다소 높은 것으로 나타났다.

베타회귀분석은 최근 의학연구에서 범위가 정해져 있는 자료를 분석하기 위해 많이 이용되고 있는데, 특히 건강 관련 삶의 질 자료를 분석하기 위해 많이 이용되고 있다 (Conrado 등, 2014; Tutoglu 등, 2014; Gheorghe 등, 2015; Kent 등, 2015). Hunger 등 (2011)은 SF-6D 자료를 이용하여 선형회귀모형과 베타회귀모형의 예측 정확도를 비교하였는데, 평균과 정밀도에 대한 하위모형을 가지고 있는 확장된 베타회귀모형의 예측 정확도가 선형회귀모형 보다 다소 높은 것으로 나타났다.

본 연구에서는 2013년도 한국의료패널 자료를 이용하여 18세 이상 성인에서 건강 관련 삶의 질을 나타내는 변수들을 이용하여 선형회귀모형과 베타회귀모형을 이용하여 EQ-5D 예측모형을 개발하고, 모형들의 예측 정확도를 비교하고자 한다.

2. 연구방법

2.1. 자료원

한국의료패널조사는 한국보건사회연구원과 국민건강보험공단 컨소시엄에서 보건의료이용실태와 의료비 지출수준, 건강수준 및 건강행태 등에 관한 기초자료를 생산하기 위하여 2008년부터 매년 실시되고 있다. 한국의료패널은 조사목적상 전국규모의 대표성을 유지하기 위해 2005년 인구주택총조사 90% 전수 자료를 추출 틀로 하고 있으며, 표본가구 선정은 1단계로 표본조사구 (집락)를 추출하고, 2단계에서는 표본조사구 내 표본가구를 추출하는 방식으로, 2단계 확률비례 층화집락추출 방법을 통해 결정하였다. 한국의료패널조사에서는 가구단위의 사회경제적 특성, 생활비 지출, 의약품 구매 및 개인단위의 인구사회경제적 특성, 경제활동 및 소득, 건강수준, 의약품 복용 행태, 삶의 질, 일자리 등을 조사하고 있다. 2013년에는 총 5,200가구가 조사되었으며, 가구원 14,823명에 대한 조사가 수행되었으며, 최근 한국의료패널조사 자료를 이용한 다양한 연구들이 수행되고 있다 (Jeong 등, 2016; Han과 Park, 2017).

본 연구에서는 EQ-5D를 예측하는 모형을 개발하기 위하여 2013년도 한국의료패널 가구원 중 18세 이상 성인들을 우선 선정하고, 반응변수와 설명변수들이 결측치가 있는 대상자를 제외한 후 10,513명의 대상자를 선정하였다. EQ-5D는 질병관리본부에서 표준안으로 제시하고 있는 Lee 등 (2009)의 질 가중치를 고려하여 계산하였으며, 예측모형의 정확도를 비교하기 위한 교차-검증 방법을 사용하기 위하여 전체 대상자를 50%씩 임의로 나누어 모형개발 자료와 모형검증 자료를 생성하였다. 모형개발 자료와 모

형검증 자료의 EQ-5D의 분포를 살펴 보면 완전한 건강상태를 나타내는 1인 값을 가지는 경우가 많은 왼쪽으로 꼬리가 긴 형태의 치우친 분포를 보이는 것을 알 수 있다 (Figure 2.1). 모형개발 자료의 경우 전체 5,256명 중 3,291명 (62.6%)이 EQ-5D가 1인 값을 가지는 것으로 나타났으며, 모형검증 자료의 경우 전체 5,257명 중 3,297명 (62.7%)이 EQ-5D가 1인 값을 가지는 것으로 나타났다.

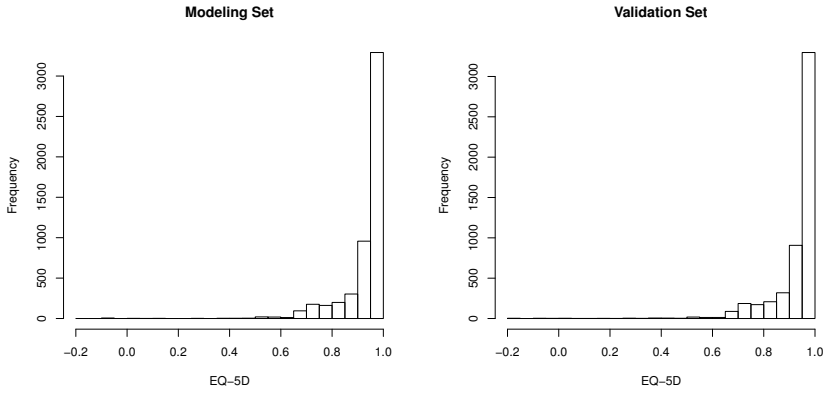

Figure 2.1 Distribution of EQ-5D

2.2. 베타회귀모형

일반적으로 연속형 반응변수에 대한 선형회귀모형은 오차들이 독립이며 분산이 동일한 정규분포를 따른다고 가정한다. 만일 연속형 반응변수의 분포가 치우쳐 있고 이분산 (heteroskedasticity)이 있으며, 0과 1 사이의 단위구간으로 주어지는 경우 베타분포를 이용한 베타회귀모형을 적용할 수 있다.

베타분포를 따르는 확률변수 Y에 대한 확률밀도함수는

$$f(y|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1},\ 0<y<1, \tag{2.1}$$

와 같으며, 여기서 형태 모수는 $a>0$, $b>0$이고 $\Gamma(\cdot)$은 감마함수를 나타낸다. 베타분포는 형태모수에 따라 분포의 형태를 다양하게 나타낼 수 있는 확률분포로 평균 $E(Y)=a/(a+b)$, 분산 $Var(Y)=ab/\left\{(a+b)^2(a+b+1)\right\}$을 가진다.

회귀분석에서는 일반적으로 평균에 대해 모형화를 하는 것이 더 유용하므로 (2.1)식에서 회귀모형의 구조를 평균과 정밀도로 나타내기 위해 $\mu = a/(a+b)$, $\phi = a+b$로 두면, $E(Y)=\mu$, $Var(Y)=\mu(1-\mu)/(1+\phi)$이 되고, 베타분포 $Y \sim Beta(\mu\phi,(1-\mu)\phi)$의 확률밀도함수는 다음과 같이 된다 (Paolino, 2001).

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\ 0<y<1. \tag{2.2}$$

여기서 $0<\mu<1$, $\phi>0$이며, ϕ는 분산의 역수의 함수로 정밀도를 나타내는 모수라고 할 수 있다. 만일 반응변수의 범위가 (a,b)로 주어지는 경우 $y'=(y-a)/(b-a)$와 같이 변환하여 $(0,1)$ 범위의 자료로 변환한 후 베타회귀모형을 적용할 수 있다.

$y_1, \cdots, y_n$은 $y_i \sim Beta(\mu_i\phi, (1-\mu_i)\phi)$로 부터의 확률표본이라고 하자. 정밀도 모수 ϕ가 상수라고 가정할 경우 평균에 대한 베타회귀모형은 반응변수의 범위가 $(0,1)$이므로 로짓 연결함수를 이용하여 다음과 같이 정의할 수 있다 (Ferrari와 Cribari-Neto, 2004).

$$\log \frac{\mu_i}{1-\mu_i} = \mathbf{x}_i^T \beta. \tag{2.3}$$

여기서 $\beta = (\beta_0, \beta_1, \cdots, \beta_k)$와 $\mathbf{x}_i$는 평균에 대한 베타회귀모형의 회귀계수벡터와 설명변수들의 벡터를 나타낸다. 여기서 회귀계수는 최대우도법을 이용하여 추정할 수 있으며, 회귀계수는 오즈비와 같이 해석할 수 있다.

Smithson과 Verkuilen (2006)는 이분산을 고려한 모형으로, 평균에 대한 하위모형과 정밀도에 대한 하위모형을 가지는 확장된 베타회귀모형을 제안하였다. 이때 $y_1, \cdots, y_n$은 $y_i \sim Beta(\mu_i\phi_i, (1-\mu_i)\phi_i)$이며, 정밀도에 대해서는 로그 연결함수를 고려하였다.

$$\log \frac{\mu_i}{1-\mu_i} = \mathbf{x}_i^T \beta, \quad \log \phi_i = \mathbf{z}_i^T \gamma. \tag{2.4}$$

여기서 $\beta = (\beta_0, \beta_1, \cdots, \beta_k)$와 $\gamma = (\gamma_0, \gamma_1, \cdots, \gamma_{k'})$는 평균에 대한 하위모형과 정밀도에 대한 하위모형의 회귀계수벡터를 나타내며, $\mathbf{x}_i$와 $\mathbf{z}_i$는 평균에 대한 하위모형과 정밀도에 대한 하위모형의 설명변수들의 벡터를 나타낸다. 여기서 회귀계수는 최대우도법을 이용하여 추정할 수 있으며, 평균에 대한 하위모형의 회귀계수는 오즈비와 같이 해석할 수 있다.

베타회귀모형은 반응변수가 $(0,1)$ 구간인 경우 적용 가능하므로, 0 또는 1의 반응변수값을 가지는 경우 반응변수의 값을 다음과 같이 변환하여 $(0,1)$ 구간으로 변환하는 것이 필요하다 (Lui와 Eugenio, 2016).

$$y'' = (y'(n-1) + 0.5)/n. \tag{2.5}$$

따라서 본 연구에서 반응변수로 사용한 EQ-5D는 이론적인 최소값이 -0.1771이며, 최대값이 1이므로 일차적으로 $y' = (y-a)/(b-a)$를 이용하여 $[0,1]$ 범위의 자료로 변환한 후, 식 (2.5)를 이용하여 $(0,1)$ 범위의 자료로 변환하였다. 따라서 모형개발 자료에서 완전한 건강상태를 나타내는 EQ-5D가 1인 값은 0.9999로 변환되었다.

2.3. 예측모형 개발 및 정확도 비교

EQ-5D를 예측하는 모형의 설명변수로는 일반적으로 건강 관련 삶의 질에 영향을 미치는 변수로 알려져 있는 성별, 연령, 교육수준, 소득수준, 의료보험종류, 고용상태, 체질량지수, 스트레스, 우울감, 주관적 건강상태 및 만성질환여부를 고려하였다 (Lee와 Han, 2015; Song 등, 2015). 모형개발 자료와 모형검증 자료에서 연속형 설명변수에 대해 평균, 표준편차를 구하고, 범주형 설명변수에 대해서는 빈도와 백분율을 제시하였으며, 설명변수 수준에 따른 EQ-5D의 중위수와 분산을 비교하기 위하여 크루스칼-왈리스 검정 및 프리그너-킬른 (Fligner-Killen) 검정 (Conover 등, 1981)을 실시하였다.

EQ-5D를 예측하는 모형으로 선형회귀모형을 적합하고, (2.3)식의 평균에 대한 베타회귀모형 및 (2.4)식의 평균과 정밀도에 대한 확장된 베타회귀모형을 적합하고, 3가지 모형에 대한 예측 정확도를 비교하기 위하여 모형검증 자료를 이용하여 예측모형의 절대오차와 제곱오차의 비율을 나타내는 R^1 계수과 R^2 계수를 다음과 같이 계산하였다 (Huang 등, 2008; Hunger 등, 2011). 모든 통계분석의 유의성은 유의수준 0.05하에서 판단하였다.

$$R^1 = 1 - \frac{\sum |Y - \hat{Y}|}{\sum |Y - \bar{Y}|}, \tag{2.6}$$

$$R^2 = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}. \tag{2.7}$$

여기서 Y는 관찰된 EQ-5D이며, $\hat{Y}$는 모형에 의해 추정된 EQ-5D, $\bar{Y}$는 관찰된 EQ-5D의 평균이다. 모형검증 자료를 대상으로 1,000번의 붓스트랩 표본을 생성하여 R^1 계수과 R^2 계수를 구하고, 95% 백분위수 신뢰구간을 구하여 모형을 비교하였다. 본 연구에서는 베타회귀모형 적합을 위하여 R의 betareg 패키지를 이용하였다 (Cribari-Neto와 Zeileis, 2010).

3. 연구결과

모형개발 자료 5,256명 중 남성이 2,378명 (45.24%), 여성이 2,878명 (54.76%)였으며, 연령 평균은 51.79 (표준편차=16.93)세로 나타났다. 교육수준은 초등학교 졸업 미만이 1,199명 (22.81%)였으며, 초등학교 졸업부터 고등학교 졸업이 2,194명 (41.74%), 전문대학 이상이 1,863명 (35.45%)로 나타났으며, 소득수준은 소득수준이 가장 낮은 제1오분위수 범위가 810명 (15.41%), 제2오분위수 범위가 992명 (18.87%), 제3오분위수 범위가 1,102명 (20.97%), 제4오분위수 범위가 1,162명 (22.64%), 소득 수준이 가장 높은 제5오분위수 범위가 1,190명 (22.64%)로 나타났다. 건강보험을 가지고 있는 사람은 5,050명 (96.08%), 의료급여 대상자는 206명 (3.92%)였으며, 일을 하고 있는 사람은 3,167명 (60.25%)로 나타났다. 체질량지수 (body mass index)는 저체중 (< 18.5)인 경우가 289명 (5.50%), 정상 (18.5-22.9)인 경우가 2,352명 (44.75%), 과체중 (23-24.9)인 경우가 1,304명 (24.81%), 비만 (≥ 25)인 경우가 1,311명 (24.94%)로 나타났다. 정신적, 신체적 스트레스를 전혀 받지 않는 대상자는 2,320명 (44.14%), 가끔 스트레스를 받는 대상자는 1,919명 (36.51%), 자주 스트레스를 받는 대상자는 1,217명 (23.15%)였으며, 우울감을 전혀 느끼지 않는 대상자는 4,833명 (91.95%), 우울감을 느끼는 대상자는 423명 (8.05%)로 나타났다. 본인이 느끼는 건강상태 점수는 평균이 69.43 (표준편차=16.06)으로 나타났으며, 골관절염을 가지고 있는 대상자는 788명 (14.99%), 당뇨병을 가지고 있는 대상자는 476명 (9.06%), 암환자는 190명 (3.61%), 고혈압을 가지고 있는 대상자는 1,276명 (24.28%), 심장병을 가진 대상자는 181명 (3.44%), 뇌혈관질환을 가진 대상자는 155명 (2.95%), 신장병을 가진 대상자는 21명 (0.40%)로 나타났다. 모형검증 자료의 대상자들의 기저특성도 모형개발 자료의 대상자와 유사하게 나타났다 (Table 3.1).

공변량의 수준에 따른 EQ-5D의 평균과 분산을 살펴 보면, 여성의 경우 남성 보다 EQ-5D의 평균이 낮으나 분산은 큰 것으로 나타났으며, 연령대가 높아 질수록 EQ-5D의 평균은 낮아지며, 분산은 증가하는 것으로 나타났다. 교육 수준이 높을 수록 EQ-5D의 평균이 높아지며, 분산이 감소하는 것으로 나타났으며, 소득 수준이 높을 수록 EQ-5D의 평균이 높아지고, 분산도 감소하는 것으로 나타났다. 의료보장 형태에 따라 살펴 보면, 건강보험 대상자에 비해 의료급여 대상자의 EQ-5D의 평균이 낮고 분산이 큰 것으로 나타났으며, 일을 하고 있는 사람이 EQ-5D의 평균이 높고 분산도 작은 것으로 나타났다. 체질량 지수의 경우 정상인 대상자들의 EQ-5D의 평균이 가장 높고 분산도 가장 작은 것으로 나타났으며, 스트레스가 많을 수록 EQ-5D의 평균이 낮고 분산이 커지는 경향이 있었으며, 우울감이 있을 경우 EQ-5D의 평균이 낮고 분산이 큰 것으로 나타났다. 건강상태가 좋다고 느낄 수록 EQ-5D의 평균이 커지고 분산이 작은 경향이 있었으며, 골관절염, 당뇨병, 암, 고혈압, 심장병, 뇌혈관 질환, 신장병이 있는 대상자들이 전반적으로 EQ-5D의 평균이 낮고 분산이 큰 것으로 나타났다 (Table 3.2, Figure 3.1).

Table 3.3에서 선형회귀모형과 베타회귀모형의 추정된 회귀계수를 살펴 보면, 교육수준이 높거나 소득이 많거나 본인이 건강하다고 느낄 수록 EQ-5D의 평균이 통계적으로 유의하게 증가하는 것으로 나타났으며, 연령이 높아지고 의료급여 대상자이거나 일을 하지 않거나 스트레스 또는 우울감이 있거나 골관절염, 뇌혈관질환, 신장병이 있는 경우 EQ-5D의 평균이 통계적으로 유의하게 감소하는 것으로 나

Table 3.1 Baseline characteristics

Variables	Total (N=10,513)		Modeling set (N=5,256)		Validation set (N=5,257)	
	n	%	n	%	n	%
Gender						
Male	4,738	45.07%	2,378	45.24%	2,360	44.89%
Female	5,775	54.93%	2,878	54.76%	2,897	55.11%
Age (years), mean±SD	52±16.67		51.79±16.93		51.31±17.02	
Education level						
Less than elementary school	2,286	21.74%	1,199	22.81%	1,087	20.68%
High school	4,500	42.80%	2,194	41.74%	2,306	43.87%
Greater than college	3,727	35.45%	1,863	35.45%	1,864	35.46%
Income level						
Q1 (poorest)	1,592	15.14%	810	15.41%	782	14.88%
Q2	1,967	18.71%	992	18.87%	975	18.55%
Q3	2,246	21.36%	1,102	20.97%	1,144	21.76%
Q4	2,326	22.12%	1,162	22.11%	1,164	22.14%
Q5 (richest)	2,382	22.66%	1,190	22.64%	1,192	22.67%
Type of medical insurance						
National health insurance	10,108	96.15%	5,050	96.08%	5,058	96.21%
Medical aid beneficiaries	405	3.85%	206	3.92%	199	3.79%
Employment status						
Working	6,318	60.10%	3,167	60.25%	3,151	59.94%
Not working	4,195	39.90%	2,089	39.75%	2,106	40.06%
Boby mass index						
< 18.5	553	5.26%	289	5.50%	264	5.02%
18.5-22.9	4,724	44.93%	2,352	44.75%	2,372	45.12%
23-24.9	2,597	24.70%	1,304	24.81%	1,293	24.60%
≥ 25	2,639	25.10%	1,311	24.94%	1,328	25.26%
Stress						
None	4,676	44.48%	2,320	44.14%	2,356	44.82%
Sometimes	3,444	32.76%	1,919	36.51%	1,725	32.81%
More than often	2,393	22.76%	1,217	23.15%	1,176	22.37%
Depression						
Not feel	9,664	91.92%	4,833	91.95%	4,831	91.90%
Feel	849	8.08%	423	8.05%	426	8.10%
Health status (0-100), mean±SD	70±16.01		69.43±16.06		70.17±15.95	
Osteoarthritis	1,531	14.56%	788	14.99%	743	14.13%
Diabetes	970	9.23%	476	9.06%	494	9.40%
Cancer	376	3.58%	190	3.61%	186	3.54%
Hypertension	2,529	24.06%	1,276	24.28%	1,253	23.83%
Heart disease	342	3.25%	181	3.44%	161	3.06%
Cerebrovascular disease	323	3.07%	155	2.95%	168	3.20%
Renal disease	45	0.43%	21	0.40%	24	0.46%

타났다. 평균에 대한 베타회귀모형과 확장된 베타회귀모형의 평균에 대한 하위모형의 회귀계수를 살펴보면, 두 모형에서 모두 연령이 증가하거나 일을 하지 않거나 스트레스 또는 우울감이 있을 수록 EQ-5D의 평균이 유의하게 감소하는 것으로 나타났으나, 확장된 베타회귀모형에서의 회귀계수의 크기가 더 크게 추정되는 경향이 있었다. 또한 교육수준이 높거나 소득이 많거나 본인이 건강하다고 느낄 수록 EQ-5D의 평균이 통계적으로 유의하게 증가하는 것으로 나타났으나, 확장된 베타회귀모형에서의 회귀계수의 크기가 더 크게 추정되는 경향이 있었다. 베타회귀모형에서 암환자의 경우 회귀계수가 통계적으로 유의하지 않았으나 확장된 베타회귀모형의 평균 하위모형에서는 암환자일수록 EQ-5D가 통계적으로 유의하게 감소하는 것으로 나타났으며, 골관절염이 있을 경우 EQ-5D의 평균이 통계적으로 유의하게 많이 감소하는 것으로 나타났다.

따라서 확장된 베타회귀모형의 평균 하위모형에 따라 연령이 0.1세 증가할수록 EQ-5D의 평균은 exp(-0.200)=0.819배 증가하며, 초등학교 졸업 미만에 비해 고등학교 졸업인 경우 1.314배, 전문대학 이상인 경우 1.498배 증가하며, 일을 하는 경우에 비해 일을 하지 않는 경우 0.867배 증가하며, 스트레스를 전혀 받지 않는 경우에 비해 가끔 스트레스를 받는 경우 0.708배, 자주 스트레스를 받는 경우 0.543배 증가하는 것으로 나타났다. 또한 우울감을 느끼는 경우 우울감을 느끼지 않는 경우에 비해 EQ-

Table 3.2 Empirical means and variances of the EQ-5D in differenct covariates subgroup in modeling set

Variables	Means	P-value of Kruskal-Wallis test	Variance	P-value of Fligner-Killeen Test
Gender		<0.001		<0.001
Male	0.954		0.009	
Female	0.927		0.012	
Age (years)		<0.001		<0.001
18-39	0.978		0.004	
40-49	0.969		0.005	
50-59	0.949		0.007	
60-69	0.918		0.013	
$\geq$ 70	0.863		0.020	
Education level		<0.001		<0.001
Less than elementary school	0.871		0.019	
High school	0.949		0.009	
Greater than college	0.972		0.005	
Income level		<0.001		<0.001
Q1 (poorest)	0.873		0.021	
Q2	0.926		0.012	
Q3	0.948		0.008	
Q4	0.959		0.006	
Q5 (richest)	0.968		0.006	
Type of medical insurance		<0.001		<0.001
National health insurance	0.943		0.010	
Medical aid beneficiaries	0.839		0.025	
Employment status		<0.001		<0.001
Working	0.956		0.006	
Not working	0.914		0.017	
Boby mass index		0.007		0.002
$<$ 18.5	0.919		0.020	
18.5-22.9	0.944		0.010	
23-24.9	0.935		0.011	
$\geq$ 25	0.939		0.011	
Stress		<0.001		<0.001
None	0.959		0.008	
Sometimes	0.945		0.008	
More than often	0.893		0.019	
Depression		<0.001		<0.001
Not feel	0.946		0.009	
Feel	0.857		0.023	
Health status		<0.001		<0.001
0-50	0.860		0.025	
51-70	0.936		0.009	
71-80	0.972		0.003	
80-100	0.979		0.004	
Osteoarthritis		<0.001		<0.001
No	0.955		0.008	
Yes	0.851		0.021	
Diabetes		<0.001		<0.001
No	0.946		0.009	
Yes	0.875		0.024	
Cancer		<0.001		<0.001
No	0.941		0.010	
Yes	0.892		0.024	
Hypertension		<0.001		<0.001
No	0.954		0.008	
Yes	0.893		0.016	
Heart disease		<0.001		<0.001
No	0.942		0.010	
Yes	0.868		0.021	
Cerebrovascular disease		<0.001		<0.001
No	0.942		0.010	
Yes	0.836		0.036	
Renal disease		<0.001		<0.001
No	0.940		0.011	
Yes	0.801		0.043	

5D의 평균은 0.629배 증가하며, 본인이 느끼는 건강상태 점수가 0.01점 증가할수록 10.67배 증가하며, 골관절염이 있는 경우 0.696배, 암환자인 경우 0.744배 증가하는 것으로 나타났다. 확장된 베타회귀 모형에서 정밀도 하위모형을 살펴 보면 연령이 증가하거나 스트레스가 증가하거나 암환자인 경우 EQ-5D의 정밀도가 통계적으로 유의하게 감소하는 것으로 나타났으며, 교육수준이 높거나 본인이 건강하다고 느낄 수록 EQ-5D의 정밀도가 통계적으로 유의하게 증가하는 것으로 나타났다 (Table 3.3).

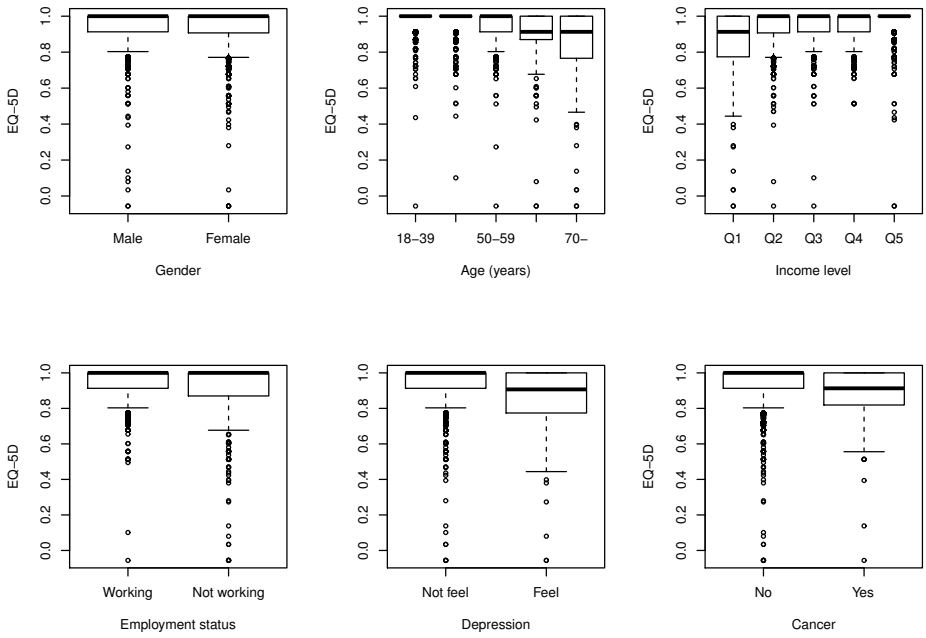

Figure 3.1 Distribution of the EQ-5D in differenct covariates subgroup in modeling set

선형회귀모형과 베타회귀모형, 확장된 베타회귀모형의 예측 정확도를 비교해 보면, 베타회귀모형이 세 모형 중 정확도가 가장 낮으며, 평균과 정밀도에 대한 하위모형을 가지고 있는 확장된 베타회귀모형이 R^1 계수와 R^2 계수가 가장 높은 것으로 나타났다 (Table 3.4).

4. 결론 및 고찰

본 연구에서는 건강 관련 삶의 질 자료인 EQ-5D에 대한 예측모형을 개발하는 방법으로 선형회귀모형과 베타회귀모형, 그리고 평균과 정밀도에 대한 하위모형을 가지고 있는 확장된 베타회귀모형을 비교하였다. 건강 관련 삶의 질 자료는 일반적으로 범위가 제한되어 있으며 왼쪽으로 꼬리가 긴 형태의 치우친 분포로 완전한 건강상태를 나타내는 1인 값을 가지는 경우가 많다. 선형회귀모형에 비해 확장된 베타회귀모형의 R^1 계수와 R^2 계수가 높기는 하지만 신뢰구간이 겹치고 있기 때문에 확장된 베타회귀모형의 정확도가 더 높다고 할 수는 없다. 하지만 선형회귀모형은 공변량에 따라 분산이 달라지는 부분을 설명할 수 없으며, 확장된 베타회귀모형은 선형회귀모형에 비해 빈도가 낮은 EQ-5D가 낮은 부분을 좀 더 잘 예측하는 것으로 나타났으며, 선형회귀모형이 제한된 범위를 벗어난 값을 예측하는 부분을 개선할 수 있다. 이는 Hunger 등 (2011)이 SF-6D 자료를 이용하여 선형회귀모형과 베타회귀모형의 예측 정확도를 비교한 결과와 유사하다. 따라서 범위가 제한되고 이분산이 있는 치우친 자료에 대해 공변량들이 평균 및 정밀도에 영향을 주는 정도를 동시에 고려하는 확장된 베타회귀모형은 건강 관련 삶의 질 자료인 EQ-5D를 분석하는 방법으로 적절하다고 할 수 있다.

Table 3.3 Parameter estimates of regression model

Variables	Linear regression			Beta regression			Extended beta regressiion					
							Mean submodel			Precision submodel		
	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
Intercept	0.885	0.010	<0.001	2.802	0.123	<0.001	2.934	0.218	<0.001	1.696	0.239	<0.001
Gender												
Male	Ref.			Ref.			Ref.			Ref.		
Female	-0.004	0.003	0.145	-0.077	0.030	0.011	-0.103	0.057	0.073	-0.031	0.063	0.620
Age/10 (years)	-0.009	0.001	<0.001	-0.086	0.012	<0.001	-0.200	0.024	<0.001	-0.147	0.026	<0.001
Education level												
Less than elementary school	Ref.			Ref.			Ref.			Ref.		
High school	0.024	0.004	<0.001	0.232	0.042	<0.001	0.273	0.072	<0.001	0.066	0.079	0.403
Greater than college	0.020	0.004	<0.001	0.190	0.053	<0.001	0.404	0.095	<0.001	0.254	0.105	0.015
Income level												
Q1 (poorest)	Ref.			Ref.			Ref.			Ref.		
Q2	0.009	0.004	0.033	0.035	0.049	0.476	0.080	0.082	0.330	0.076	0.090	0.400
Q3	0.014	0.004	0.001	0.088	0.051	0.088	0.133	0.088	0.129	0.080	0.097	0.407
Q4	0.012	0.004	0.007	0.078	0.052	0.135	0.130	0.092	0.156	0.088	0.101	0.382
Q5 (richest)	0.013	0.005	0.003	0.119	0.053	0.026	0.201	0.096	0.036	0.120	0.105	0.254
Type of medical insurance												
National health insurance	Ref.			Ref.			Ref.			Ref.		
Medical aid beneficiaries	-0.031	0.006	<0.001	-0.218	0.074	0.003	-0.201	0.110	0.068	-0.011	0.123	0.931
Employment status												
Working	Ref.			Ref.			Ref.			Ref.		
Not working	-0.013	0.003	<0.001	-0.071	0.031	0.021	-0.143	0.056	0.011	-0.112	0.062	0.069
Boby mass index												
< 18.5	-0.011	0.005	0.044	-0.028	0.062	0.656	-0.104	0.112	0.354	-0.140	0.122	0.249
18.5-22.9	Ref.			Ref.			Ref.			Ref.		
23-24.9	-0.002	0.003	0.590	-0.033	0.035	0.341	-0.071	0.064	0.263	-0.047	0.070	0.500
≥ 25	0.003	0.003	0.264	0.037	0.035	0.293	0.009	0.065	0.894	-0.034	0.072	0.632
Stress												
None	Ref.			Ref.			Ref.			Ref.		
Sometimes	-0.012	0.003	<0.001	-0.188	0.032	<0.001	-0.345	0.063	<0.001	-0.209	0.069	0.002
More than often	-0.035	0.003	<0.001	-0.417	0.038	<0.001	-0.611	0.067	<0.001	-0.285	0.074	<0.001
Depression												
Not feel	Ref.			Ref.			Ref.			Ref.		
Feel	-0.040	0.005	<0.001	-0.382	0.052	<0.001	-0.464	0.079	<0.001	-0.122	0.089	0.172
Health status/100	0.155	0.008	<0.001	1.207	0.094	<0.001	2.370	0.160	<0.001	1.625	0.174	<0.001
Osteoarthritis	-0.044	0.004	<0.001	-0.438	0.045	<0.001	-0.363	0.071	<0.001	0.019	0.079	0.805
Diabetes	-0.015	0.004	0.001	-0.078	0.051	0.127	-0.156	0.082	0.059	-0.131	0.090	0.148
Cancer	-0.011	0.006	0.071	0.003	0.074	0.967	-0.296	0.126	0.019	-0.396	0.134	0.003
Hypertension	-0.002	0.003	0.535	-0.042	0.039	0.284	-0.020	0.066	0.766	0.032	0.073	0.664
Heart disease	-0.007	0.007	0.292	-0.009	0.076	0.909	-0.037	0.119	0.754	-0.021	0.132	0.871
Cerebrovascular disease	-0.042	0.007	< 0.001	-0.243	0.081	0.003	-0.237	0.125	0.057	-0.131	0.136	0.335
Renal disease	-0.064	0.019	0.001	-0.674	0.204	0.001	-0.396	0.261	0.128	0.330	0.324	0.310

Table 3.4 Predictive accuracy

Statistical models	R^1 (95% CI)	R^2 (95% CI)
Linear regression	0.291 (0.275, 0.307)	0.359 (0.331, 0.386)
Beta regression (Mean model)	0.225 (0.212, 0.237)	0.320 (0.291, 0.347)
Extended beta regression (Mean + precision model)	0.308 (0.291, 0.324)	0.386 (0.344, 0.426)

References

Austin P. C. (2002). A comparison of methods for analyzing health-related quality-of-life measures. *Value Health*, **5**, 329-337.

Bang, S. Y. (2016). Quality of life and its related factors in patients with Korean chronic obstructive pulmonary disease. *Journal of the Korean Data & Information Science Society*, **27**, 1349-1360.

Brazier, J. E., Roberts, J. and Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*, **21**, 271-292.

Brazier, J. E. and Roberts, J. (2004). The estimation of a preference-based measure of health from the SF-12. *Medical Care*, **42**, 851-859.

Conover, W. J., Johnson, M. E. and Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, **23**, 351-361.

Conrado, D. J., Denney, W. S. and Chen, D. (2014). An updated Alzheimer's disease progression model: Incorporating non-linearity, beta regression, and a third-level random effect in NONMEM. *Journal of Pharmacokinetics and Pharmacodynamics*, **41**, 581-598.

Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, **34**, 1-24.

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical Care*, **35**, 1095-108.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J. and Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes*, 3d ed., Oxford University Press, New York.

EuroQol Group (1990). EuroQol-a new facility for the measurement of health-related quality of life. *Health Policy*, **16**, 199-208.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, **31**, 799-815.

Gheorghe, M., Brouwer, W. and van Baal, P. (2015). Did the health of the Dutch population improve between 2001 and 2008? investigating age- and gender-specific trends in quality of life. *The European Journal of Health Economics*, **16**, 801-811.

Han J. Y. and Park H. S. (2017). Factors influencing quality of health care: Based on the Korea health panel data. *Journal of the Korean Data & Information Science Society*, **28**, 195-206.

Huang, I. C., Frangakis, C., Atkinson, M. J., Willke, R. J., Leite, W. L., Vogel, W. B. and Wu, A. W. (2008). Addressing ceiling effects in health status measures: A comparison of techniques applied to measures for people with HIV disease. *Health Services Research*, **43**, 327-339.

Hunger, M., Baumert, J. and Holle, R. (2011). Analysis of SF-6D index data: Is beta regression appropriate? *Value In Health*, **4**, 759-767.

Jeong, S. R., Doo, Y. T., and Lee, W. K. (2016). Effect on ambulatory dental visitation frequency according to pack-years of smoking. *Journal of the Korean Data & Information Science Society*, **27**, 419-427.

Jo, M. W., Yun, S. C. and Lee, S. I. (2008). Estimating quality weights for EQ-5D health states with the time trade-off method in South Korea. *Value In Health*, **11**, 1186-1189.

Kent, S., Gray, A. and Schlackow, I. (2015). Mapping from the Parkinson's disease questionnaire PDQ-39 to the generic EuroQol EQ-5D-3L: The value of mixture models. *Medical Decision Making*, **35**, 902-911.

Lee, K. E. and Han, S. H. (2015). Factors affecting the health-related quality of life among male elders. *International Journal of Bio-Science and Bio-Technology*, **7**, 65-74.

Lee, Y. K., Nam, H. S., Chuang, L. H., Kim, K. Y., Yang, H. K., Kwon, I. S., Kind, P., Kweon, S. S. and Kim, Y. T. (2009). South Korean time trade-off values for EQ-5D health states: Modeling with observed values for 101 health states. *Value In Health*, **12**, 1187-1193.

Longworth, L. and Rowen D. (2013). Mapping to obtain EQ-5D utility-values for use in NICE health technology assessments. *Value In Health*, **16**, 202-210.

Lui, F. and Eugenio, E. C. (2016). A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. *Statistical Methods in Medical Research*, Epub ahead of print.

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, **9**, 325-346.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, **25**, 303-325.

Smithson, M. and Verkuilen, J. (2006). Better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, **11**, 54-71.

Song, T., Ding, Y., Sun, Y., He, Y. N., Qi, D., Wu, Y., Wu, B., Lang, L., Yu, K., Zhao, X., Zhu, L., Wang, S. and Yu, X. S. (2015). A population-based study on health-related quality of life among urban community residents in Shenyang, Northeast of China. *BMC Public Health*, **15**, 921-932.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24-36.

Tutoglu, A., Boyaci, A. and Koca, I. (2014). Quality of life, depression, and sexual dysfunction in spouses of female patients with fibromyalgia. *Rheumatology International*, **34**, 1079-1084.

Journal of the Korean Data & Information Science Society
2017, **28**(3), 547–557

http://dx.doi.org/10.7465/jkdi.2017.28.3.547
한국데이터정보과학회지

Analysis of health-related quality of life using Beta regression[†]

Eun Jin Jang[1]

[1]Department of Information Statistics, Andong National University

Received 18 April 2017, revised 14 May 2017, accepted 17 May 2017

Abstract

The health-related quality of life data are commonly skewed and bounded with spike at the perfect health status, and the variance tended to be heteroscedastic. In this study, we have developed a prediction model for EQ-5D using linear regression model, beta regression model, and extended beta regression model with mean and precision submodel, and also compared the predictive accuracy. The extended beta regression model allows to model skewness and differences in dispersion related to covariates. Although the extended beta regression model has higher prediction accuracy than the linear regression model, the overlapped confidence intervals suggested that the extended beta regression model was superior to the linear regression model. However, the expended beta regression model could explain the heteroscedasticity and predict within the bounded range. Therefore, the expended beta regression model are appropriate for fitting the health-related quality of life data such as EQ-5D.

Keywords: Beta regression, health-related quality of life, heteroskedasticity, Korea health panel survey.

[†] This research was supported by Andong National University Research Grant 2014.
[1] Assistant professor, Department of Information Statistics, Andong National University, Gyeongbuk 36729, Republic of Korea. E-mail: ejjang@anu.ac.kr