VALUE IN HEALTH 14 (2011) 759–767

available at www.sciencedirect.com

SCIENCE DIRECT®

journal homepage: www.elsevier.com/locate/jval

Analysis of SF-6D Index Data: Is Beta Regression Appropriate?

Matthias Hunger, MSc[1,*], *Jens Baumert, PhD*[2], *Rolf Holle, PhD*[1]

[1]*Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Health Economics and Health Care Management, Neuherberg, Germany;* [2]*Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Epidemiology II, Neuherberg, Germany*

ABSTRACT

Background: Preference-weighted index scores of health-related quality of life are commonly skewed to the left and bounded at one. Beta regression is used in various disciplines to address the specific features of bounded outcome variables such as heteroscedasticity, but has rarely been used in the context of health-related quality of life measures. We aimed to examine if beta regression is appropriate for analyzing the relationship between subject characteristics and SF-6D index scores. **Methods:** We used data from the population-based German KORA F4 study. Besides classical beta regression, we also fitted extended beta regression models by allowing a regression structure on the precision parameter. Regression coefficients and predictive accuracy of the models were compared to those from a linear regression model with model-based and robust standard errors. **Results:** The beta distribution fitted the empirical distribution of the SF-6D index better than the normal distribution. Extended beta regression performed best in terms of predictive accuracy but confidence intervals of the fit measures suggested that no model was superior to the others. Age had a significant negative effect on the precision parameter indicating higher variation of health utilities in older age groups. The observations reporting perfect health had a high influence on model results. **Conclusions:** Beta regression, especially with precision covariates is a possible supplement to the methods currently used in the analysis of health utility data. In particular, it accounted for the boundedness and heteroscedasticity of the SF-6D index. A pitfall of the beta regression is that it does not work well in handling one-valued observations.

Keywords: beta regression, dispersion covariates, health-related quality of life, heteroscedasticity, SF-6D.

Introduction

Health-related quality of life (HRQL) is an important outcome parameter in health economic evaluation. For the calculation of quality-adjusted life years, it is required that HRQL can be expressed by a generic single index that reflects preferences or utilities chosen by individuals or by the society for specific health states [1]. Besides the health-utility index [2] and the EQ-5D [3], the SF-6D [4,5] is one of the most popular preference-based HRQL instruments.

Regression models are frequently used to model preference-weighted index scores as a function of individual characteristics [6]. However, health utility data exhibit specific characteristics so that the choice of regression methods is not straightforward [7]. A common feature of health utility data is that their distribution is skewed to the left and truncated at one. As a consequence, it is important to check if the distributional assumptions of the regression methods are met by the data [8].

To our knowledge, all studies that modelled SF-6D index scores as a function of person-level characteristics relied on linear regression using ordinary least-squares estimation (OLS) [9–11]. However, linear regression assumes that error terms are homoscedastic, and this assumption is unlikely to be met for bounded variables where the variability of scores declines as the mean approaches the bounds [7,8]. Also, the additional assumption that the error terms are normally distributed may not hold. To account for heteroscedasticity in the linear model, the calculation of robust standard errors has been advocated [8,12].

Alternative regression methods that have been used in literature to address the idiosyncrasies of HRQL data are censored least absolute deviation models [6,13,14], Tobit models [6,13,14], latent class models [13], two-part models (TPM) [12,13] and median regression [6,12]. Censored least absolute deviation models and Tobit approaches model an underlying latent variable censored at one. However, models that allow for censoring are not appropriate for preference-weighted index data because preferences are measured on a scale where values by definition cannot exceed perfect health [8]. Latent class models and TPM address the non-normality of the data by assuming a mixed distribution for the health utility scores. They have shown advantages in handling pronounced ceiling effects like that of the EQ-5D index [12,13], but may be less relevant for modelling SF-6D data where the percentage of one-valued observations was low [12,15,16]. Median regression provides a robust alternative to OLS regression for non-normal data and has shown good properties in predicting index scores for multiattribute health state classifiers [17]. However, if preference-weighted index scores are regressed on individual covariates to estimate quality-adjusted life years in an economic evaluation, regression must focus on the mean, not on the median [8,18].

For bounded outcomes, the mean must be a nonlinear function of the covariates and the variance must be heteroscedastic [7].

[*] *Address correspondence to:* Matthias Hunger, Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Health Economics and Health Care Management, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany.

Email: matthias.hunger@helmholtz-muenchen.de.

1098-3015/$36.00 – see front matter Copyright © 2011, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

doi:10.1016/j.jval.2010.12.009

Beta regression, introduced by Ferrari and Cribari-Neto [19], fulfils both requirements and has been used in various disciplines to model variables observed on the standard unit interval. By variation of mean and precision parameter, the beta distribution provides a variety of shapes. Whereas in the classical beta regression model only the mean parameter is modelled as a function of explanatory variables, an extended beta regression model as formally introduced by Simas et al. [20] allows a regression structure on both mean and precision parameter. In that way, extended beta regression allows to naturally model skewness and differences in dispersion related to covariates [21].

Two recent studies used a (classical) beta regression model to analyze EQ-5D index data and visual analogue scale data, respectively [22,23]. Pullenayegum et al. [8] moreover argued that beta regression was an appealing regression method for health utility data because of the boundary at one, but pointed out that the lower boundary at zero may restrict its applicability to populations where negative utilities are not observed. There is first evidence that in some situations, beta regression may be superior to traditional OLS methods when regressing HRQL measures on covariates [24]; however, the comparison of the beta regression model with competing models has not been sufficiently examined. Beta regression with dispersion covariates has been used by Cheung et al. [25] to model the scores of SF-36 domains.

The purpose of this study was to examine the applicability of classical and extended beta regression to analyze the relationship between the SF-6D index score and subject characteristics, and to compare estimated effects and predictive accuracy of these methods to the traditionally used linear regression model. We used data from the population-based German KORA-F4 study for the analysis.

Methods

Data sources

The data used for analysis were taken from the German population-based KORA F4 study (2006–2008), which is a follow-up study of the KORA S4 survey conducted in 1999–2001. The study region comprises the city of Augsburg and its two surrounding counties in southern Germany. A detailed description of the study design, sampling method, data collection, and response rate can be found elsewhere [26–28]. In brief, 6640 individuals aged 25 to 74 years were randomly selected from population registries for the baseline survey (S4) in 1999–2001, of which 4261 participated. In 2006–2008 these subjects were reinvited for follow-up examination of whom 3080 (72%) were investigated. Subjects who died during follow-up, moved too far outside the study area, or were lost to follow-up were excluded from follow-up.

The SF-6D is a multidimensional health state classification derived from responses to the SF-12 health survey. It comprises six dimensions (physical functioning, role limitations, social functioning, pain, mental health, and vitality), each of which consists of three up to five levels, yielding a total of 7500 different health states. These health states can be combined into a single index score using an algorithm that is based on valuations of a representative UK general population sample [5]. The SF-6D index scores range from 0.345 to 1. It should be noted that there also exists a version of the SF-6D derived from 11 items of the SF-36 [4], which differs from SF-6D that we used. Participants self-administered the SF-12 questionnaire at the study center, and completeness of the questionnaire was checked by the study personnel.

Covariates in the analyses were age, sex, body mass index (BMI), marital status (single or living with partner), education (primary/secondary education and tertiary education), smoking status (current smoker, ex-smoker, or never smoker) as well as history of diabetes, cancer, a cardiovascular event (myocardial infarction or angina pectoris), or stroke (Table 1). For our descriptive analyses, participants were classified into BMI categories according to the World Health Organization guidelines [29], and age was categorized into five 10-year groups.

Table 1 – Sociodemographic and clinical characteristics of the study population.

Variable	Mean ± SD or %
Age (y)	55.90 ± 13.21
BMI	27.57 ± 4.75
Sex	
Man	48.7%
Woman	51.3%
Marital status	
Single	24.4%
With partner	75.6%
Smoking status	
Never smoker	44.2%
Ex-smoker	37.9%
Current smoker	17.9%
Education	
Primary/secondary	44.0%
Tertiary	56.0%
Diabetes mellitus	6.8%
Stroke	1.9%
Cancer	7.9%
Cardiovascular event	7.0%

BMI, Body mass index; SD, standard deviation.

Descriptive analyses

In our descriptive analyses, we examined how mean, variance, and shape of the SF-6D index differ across various subgroups defined by the covariates. We calculated sample means in the different subgroups and compared them using the Kruskall-Wallis test. To examine heteroscedasticity, we investigated how the dispersion of the SF-6D index differs between the covariate subgroups by calculating sample variances and using Levene's test for equality of variances.

We further show histograms of the SF-6D index score by age groups. This helps to examine how the shape of the conditional distribution (and not only its first central moments) changes with increasing age. We added two curves to each histogram, representing the density functions of the beta and the normal distribution that were fitted to the data.

Statistical models

Beta regression

Beta regression is a fully parametric approach, assuming that the dependent variable follows a beta distribution with density function

$$f(y;\,\mu,\,\varphi)=\frac{\Gamma(\varphi)}{\Gamma(\mu\varphi)\Gamma((1-\mu)\varphi)}y^{\mu\varphi-1}(1-y)^{(1-\mu)\varphi-1},\quad 0<y<1, \tag{1}$$

where $\Gamma(.)$ denotes the gamma function [30]. The parameter μ denotes the expected value of Y; that is, $E(Y) = \mu$. The parameter φ fulfils the definition of a precision parameter because for fixed μ the greater the value of φ, the smaller the variance of the dependent variable. More specifically,

$$\mathrm{Var}(Y)=\frac{V(\mu)}{1+\varphi}, \tag{2}$$

where $V(\mu) = \mu\,(1-\mu)$.

In the classical beta regression model, as in the generalized linear model framework, only the mean parameter μ of the beta distribution is expressed as a function of covariates, whereas the precision parameter φ is treated as nuisance. To map the linear predictor into the space of observed values on the unit interval, the logit link has been proposed as the link of choice [21]:

$$\log\frac{\mu_i}{1-\mu_i} = x_i^T\beta, \tag{3}$$

where x_i^T denotes the vector of covariates, and β refers to the vector of regression coefficients.

The extended beta regression model relates both parameters to covariates through distinct linear predictors [20,21]. With the precision parameter φ being an inverse measure of dispersion, this approach is similar to a generalized linear model with dispersion covariates [31]. It reflects the idea that the φ is of interest on its own and that in many situations covariates have an effect on the variation of the dependent variable, thus involving heteroscedasticity [21,32]. Extended beta regression comprises two submodels, one for the mean and one for the precision parameter. The submodel for the mean is identical to that given by equation (3), whereas the precision submodel uses a log link to guarantee that φ is always positive:

$$\log(\varphi_i) = w_i^T\delta, \tag{4}$$

where w_i^T denotes the vector of precision covariates and δ refers to the vector of the respective regression coefficients.

The beta distribution has a nonzero support only on the open unit interval. If ones and zeros occur in the data, it has been suggested to minimally compress the range of observed values, applying the transformation

$$Y^* = [Y(N-1) + 0.5]/N \tag{5}$$

where Y^* is the transformed and Y is the untransformed dependent variable [21]. This transformation has already been used in several studies [21,23,32], but may bias results if the number of boundary values is large. Therefore, it has been proposed to experiment with different endpoint handling schemes and to examine if this has an effect on the parameter estimates [21]. Following (5), the one-valued observations in our sample were transformed to $Y^* = 0.99983$.

Linear regression

The linear regression model is given by

$$Y_i = x_i^T\beta + \epsilon_i,$$

where Y_i denotes the health utility for individual i, and the ε_i are uncorrelated random variables with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$ for any i. In particular, error terms are homoscedastic because their variance is constant irrespective of i.

If the assumptions on ε_i hold, then the OLS estimator $\hat{\beta}$ is the best linear unbiased estimator according to Gauss-Markov theorem. If one additionally assumes normality of the error terms, then Y_i given x_i is also normal, and maximum likelihood estimation of $\hat{\beta}$ coincides with OLS. It also follows that $\hat{\beta}$ is normal, allowing confidence intervals and P values to be calculated.

If the assumption of homoscedasticity is violated, then $\hat{\beta}$ remains unbiased and consistent, but standard errors are biased so that hypothesis tests are no longer valid. Linear regression with robust standard errors is based on the calculation of a heteroscedasticity-corrected covariance matrix (HCCM) of the OLS estimate, also known as sandwich matrix. There exist several versions of HCCMs of which we used version HC3 in our analysis [8,33].

Model estimation

We fitted classical and extended beta regression models to estimate the effects of the covariates on the conditional mean and the conditional distribution of the SF-6D index scores. We compared the estimates to those of the linear regression model with model-based and robust standard errors in terms of significance and direction of effects. We included all available covariates in the mean (sub) models. To avoid overcomplex models the precision submodel of the extended beta regression only comprised significant covariates.

We examined the predictive distributions of the two methods by comparing estimated regression quantiles.

We used a cross-validation method to determine the predictive accuracy of the competing methods. We randomly partitioned the data into training (90% of the data) and validation set (10% of the data). We estimated the level of fit by calculating R^1 and R^2 coefficients as well as logarithmic scores (LogS) [13,34]. R^1 and R^2 coefficients measure the proportion of absolute and square error that was predicted by the respective regression methods, while the LogS is a standard measure of the accuracy of probabilistic forecasts that assesses how well the predictive distribution corresponds to the observed values in the validation set. R^1, R^2, and LogS are calculated as follows:

$$R^1 = 1 - \frac{\sum_i |Y_i - \hat{Y}_i|}{\sum_i |Y_i - \bar{Y}|}$$

$$R^2 = 1 - \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y})^2}$$

$$LogS = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\widehat{\mu_i}, \hat{\sigma}_i)$$

where Y_i denotes the observed SF-6D index score of individual i, $\hat{Y}_i$ is the predicted SF-6D index score of individual i, and $\bar{Y}$ is the mean of the observed SF-6D index scores in the validation sample.

The above cross-validation process was repeated in 1000 bootstrap samples, and mean and 95% percentile intervals of the three considered predictive accuracy measures were computed across bootstrap samples.

We conducted model diagnostic analyses calculating generalized leverages for each observation [35]. Leverage is one of the central components of influence diagnosis in regression models. It measures the importance of individual observations and reflects their influence on the model fit [36]. High leverage points are characterized by a high instantaneous rate of change in their predicted value with respect to their response value [35]. For the exact calculation, see Rocha and Simas [35].

For statistical testing, the level of significance was set at $\alpha = .05$. All calculations were carried out using the statistical software R 2.10.1, including the add-on package betareg [30,37].

Results

Descriptive analyses

Excluding participants with missing data reduced the final sample size from 3080 to 2933. The sociodemographic and clinical characteristics of the final sample population are summarized in Table 1. The mean SF-6D index in the sample was 0.793 ± 0.126 and values ranged from 0.345 to 1. A number of 98 individuals (3.3%) had an SF-6D index value of 1.

The bivariate analyses between SF-6D index and the covariates revealed that utilities were lower in older age groups and that the variation of SF-6D index values increased with growing age, as shown in Table 2. Significant associations with the mean SF-6D score were further detected for all other covariates with the exception of smoking status. Levene's test showed that the empirical variances of the SF-6D index differed significantly across the covariate subgroups, with higher mean scores involving reduced dispersion.

Figure 1 shows that the distribution of the SF-6D index score by age groups and for the entire sample is skewed to the left. Also, not

Table 2 – Empirical means and variances of the SF-6D index score in different covariate subpopulations.

Covariate	N	Mean SF-6D score	P (Kruskal-Wallis test)	Empirical variance of SF-6D scores	P (Levene's test)
Age (y)			<0.001		<0.001
31-39	380	0.824		0.010	
40-49	678	0.813		0.012	
50-59	672	0.784		0.016	
60-69	625	0.790		0.017	
70-82	578	0.763		0.021	
BMI			<0.001		<0.001
<18.5	10	0.788		0.014	
18.5-25	926	0.805		0.014	
25-30	1227	0.798		0.016	
30-35	556	0.781		0.017	
35-40	154	0.747		0.017	
> 40	60	0.743		0.019	
Sex			<0.001		0.004
Man	1429	0.811		0.015	
Woman	1504	0.777		0.016	
Marital status			<0.001		<0.001
Single	715	0.764		0.017	
With partner	2218	0.803		0.015	
Smoking status			0.840		0.242
Never smoker	1297	0.791		0.015	
Ex-smoker	1112	0.795		0.017	
Current smoker	524	0.795		0.015	
Education			<0.001		<0.001
Primary/secondary	1290	0.778		0.018	
Tertiary	1643	0.805		0.014	
Diabetes mellitus			<0.001		<0.001
No	2735	0.797		0.015	
Yes	198	0.742		0.020	
Stroke			<0.001		<0.001
No	2876	0.795		0.015	
Yes	57	0.696		0.030	
Cancer			<0.001		<0.001
No	2701	0.796		0.015	
Yes	232	0.759		0.020	
Cardiovascular event			<0.001		0.005
No	2728	0.799		0.015	
Yes	205	0.715		0.019	

BMI, Body mass index.

only the mean of the SF-6D index score but also the shape of its distribution changes across age groups: As age increases, the distribution gets broader and the skewness is reduced. The estimated curves suggest that the beta distribution (solid curve) fits the data better than the normal distribution (dashed curve). However, the beta distribution performed poorly in the oldest age group.

Regression models

The parameter estimates of classical and extended beta regression model are shown in Table 3. In both models, high BMI, being a woman, living alone, and all comorbidities were significantly associated with lower mean SF-6D index scores. Age had a small positive but insignificant effect on the mean SF-6D index score in the classical beta model, whilst in the extended beta model this effect was significant and negative. Significant covariates in the precision submodel were age, smoking status, education, and stroke. The results show that the precision parameter φ decreased by a factor of $\exp(-0.02) = 0.98$ for every additional year. According to (2), this effect refers to the additional change in dispersion that is beyond the change already implied by the fact that increasing age is related with lower SF-6D scores and hence higher dispersion.

In the linear model (Table 4), significant predictors of reduced HRQL were the same as in the mean submodel of the extended beta regression. In particular, age had a negative effect on the mean SF-6D index score. Differences between model-based and robust standard errors were only small except for stroke, where the robust standard was about 1.5 times higher.

Figure 2 compares the estimated age effect of the extended beta regression model and the linear regression with normality assumption. The displayed curves represent the mean as well as the 5%, 25%, 75%, and 95% quantiles of the estimated predictive distributions for a typical healthy individual where all covariates with the exception of age are fixed (i.e., male sex, mean BMI, with partner, never smoker, higher education level, and without comorbidities). The left plot in Figure 2 refers to the extended beta regression model and expresses that the distribution of the SF-6D index is skewed to the left because the 75% quantile and 95% quantile are close together whereas the corresponding 5% and 25% quantiles are shifted downward. The graph indicates that not only the mean declines with advancing age but that older age is also associated with a bigger left tail in the conditional distribution. In contrast, linear regression cannot provide such information about

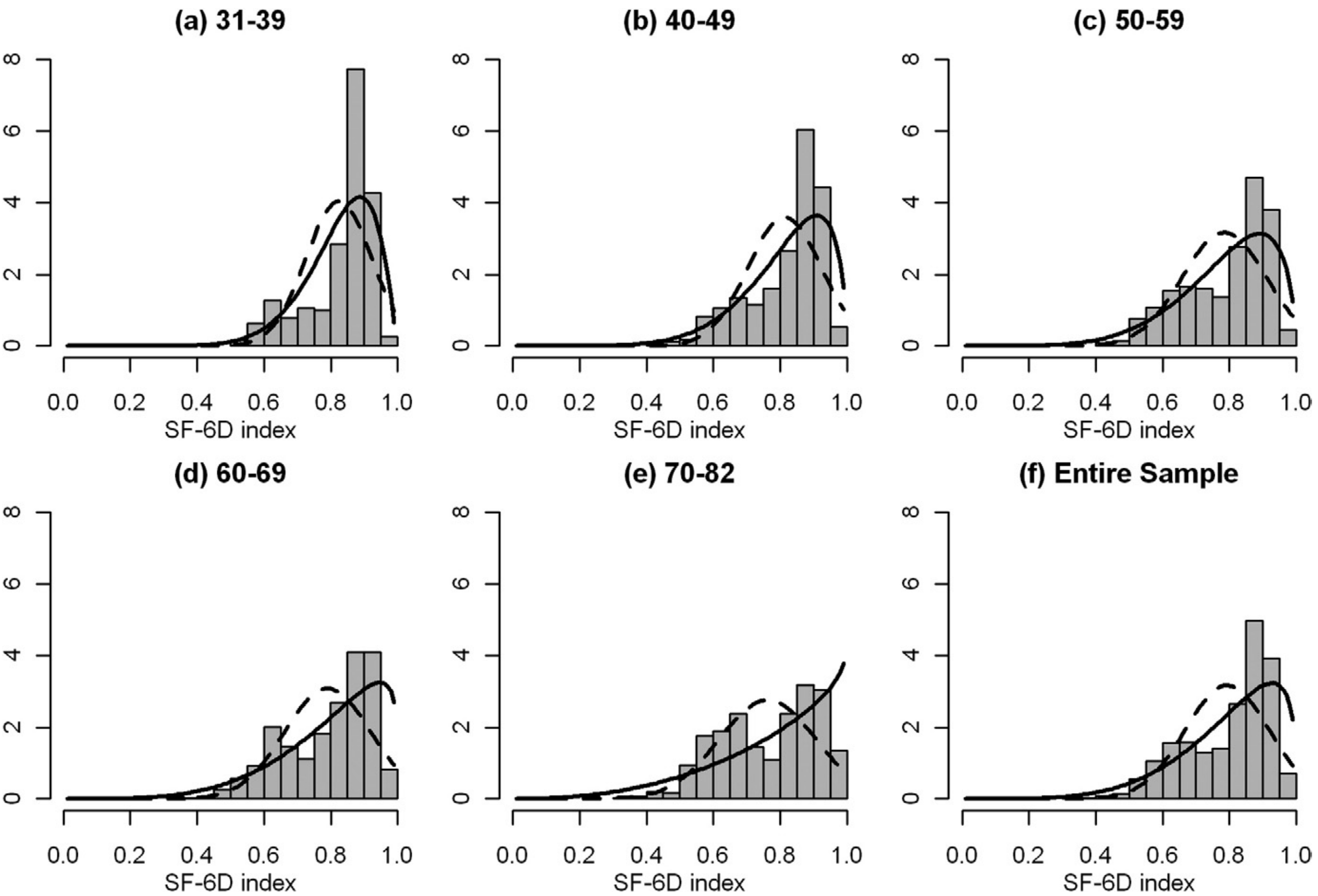

Fig. 1 – Distribution of the SF-6D index score by age groups (a-e) and for the entire sample (f). The curves represent the estimated density functions of the beta (solid) and the normal (dashed) distribution.

changes in both location and shape. Assuming homoscedasticity and normality of the error terms, the linear regression model predicts a symmetric conditional distribution with the quantiles lying on equidistant curves, as shown in the right plot of Figure 2. Furthermore, beta regression model respects the restricted range of the SF-6D index, whereas in the linear model parts of the estimated distribution are lying above one.

The predictive accuracy of each model is summarized in Table 5. The classical beta regression model performed worst on all three predictive accuracy measures. Extended beta regression performed best on R^1 and LogS, whereas the linear model had the highest R^2. However, confidence intervals were overlapping across all models, suggesting that no model was superior to the others.

Model diagnostic analyses revealed high generalized leverages for individuals with perfect health (results not shown). Using the transformation given in formula (5), these subjects had an SF-6D index score of 0.99983. It is likely that their high influence is related to the proximity to the right boundary of the beta distribution range. To test this supposition, we successively transformed the value of perfect health to slightly lower values in a range between 0.99 and 0.99983 and refitted the beta regression models. The different transformations only slightly changed the coefficient estimates in the mean submodels (an exception was the age effect in the classical beta model that became negative for smaller boundary values). Precision, however, was affected remarkably: The asymptotic standard errors in both models decreased on average by a factor of about 0.87 if perfect health was assigned the value 0.99 instead of 0.99983, and still by a factor of about 0.9 if perfect health was assigned the value 0.995. In the latter case, the high influence of these health states was also reduced, resulting in lower generalized leverages (results not shown).

Discussion

The distribution of preference-based health indices is commonly non-normal, exhibiting skewness to the left and a boundary at one [8]. This study examined the applicability of classical and extended beta regression to address these features using the example of studying the relationship between SF-6D index score and subject characteristics in a population-based German health study. Our results show that the predictive accuracy of the beta regression was similar to those of linear regression, but that extended beta regression had some advantages in estimating predictive distributions. However, beta regression has been shown to have limitations in handling one values.

For bounded response variables, mean, and variance are not independent. In our bivariate analyses this effect was considerable. As one would expect, we observed reduced dispersion in subgroups with higher mean scores, indicating potential heteroscedasticity related to the covariates. Examining the distribution of SF-6D index values in different age subgroups, we observed that the beta distribution addresses the skewness of the data better than the normal distribution.

For bounded variables like the SF-6D index, the mean must be a nonlinear function of the covariates and the variance must be heteroscedastic since the variability of scores decreases as the mean approaches the boundary points [7]. Both requirements are fulfilled

Table 3 – Regression coefficients of the classical and extended beta regression model.

Covariate	Beta regression (mean covariates only)			Extended beta regression (mean and precision covariates)		
	Estimate	Standard error	P	Estimate	Standard error	P
Mean submodel (logit link)						
Intercept	1.6473	0.1169	<0.0001	1.9090	0.1183	<0.0001
Age (y)	0.0015	0.0013	0.2424	−0.0037	0.0014	0.0065
BMI	−0.0098	0.0033	0.0030	−0.0085	0.0032	0.0080
Sex[†]						
Woman	−0.2426	0.0311	<0.0001	−0.2141	0.0300	<0.0001
Marital status[‡]						
With partner	0.1376	0.0348	<0.0001	0.1566	0.0339	<0.0001
Smoking status[§]						
Ex-smoker	0.0373	0.0339	0.2718	−0.0177	0.0361	0.6246
Current smoker	−0.0710	0.0431	0.0998	−0.0638	0.0427	0.1352
Education[¶]						
Primary/secondary	−0.0146	0.0316	0.6448	−0.0283	0.0335	0.3975
Diabetes mellitus	−0.1977	0.0602	0.0010	−0.1796	0.0636	0.0047
Stroke	−0.2556	0.1052	0.0151	−0.3543	0.1385	0.0105
Cancer	−0.1429	0.0551	0.0095	−0.1485	0.0568	0.0089
Cardiovascular event	−0.3378	0.0575	<0.0001	−0.3447	0.0606	<0.0001
Precision submodel (log link; after variable selection[*])						
Intercept	1.9488	0.0256	<0.0001	3.2792	0.1211	<0.0001
Age (y)				−0.0204	0.0021	<0.0001
Smoking status[§]						
Exsmoker				−0.1857	0.0565	0.0010
Current smoker				−0.0195	0.0734	0.7902
Education[¶]						
Primary/secondary				−0.1178	0.0532	0.0267
Stroke				−0.4378	0.1778	0.0138

BMI, Body mass index.

[*] For beta regression with mean covariates only, the precision submodel only consists of an intercept term; for the extended beta regression, only significant covariates were included in the precision submodel (at the 5% level).

[†] Reference category: Man.

[‡] Reference category: Single.

[§] Reference category: Never smoker.

[¶] Reference category: Tertiary education level.

by the beta regression model. In contrast, linear regression contravenes these conditions because it assumes homoscedasticity and a linear expectation function. As a consequence, the linear model may produce biased standard errors and out-of-range predictions.

When making inference about the mean SF-6D index scores only; however, there was no substantial difference between the linear model and beta regression in our study: As confidence intervals of the predictive accuracy measures were overlapping, no model was superior to the others. Also, the linear model predicted no values superior to one, and there were only slight differences between model-based and robust standard errors.

Our analyses show that extended beta regression is a useful supplement to currently used methods when focus is not only on the mean but also on the predictive distribution of the utility index: It performed best in terms of the logarithmic score and respected the boundary at one, whereas in the linear model a part of the estimated distribution was lying above one. Also, by modelling dispersion in terms of covariates, extended beta regression provided information about the shape of the distribution that is not available in other methods. This is important when the objective is for example to derive population-based reference values for HRQL in form of centile curves. Such reference values are important for decision-analytic models in cost-effectiveness analyses.

In the extended beta regression model, age, smoking status, education and stroke were related to changes in precision: The results suggest that growing age is not only associated with a diminishing mean, but also with an increased variation of the SF-6D index score. This finding might be explained by the fact that age is considered as a proxy of comorbitity representing diseases with various effects on HRQL [38]. A similar result is given by Li and Fu [12] who reported that the distribution of the EQ-5D in a subpopulation with several comorbidities had a bigger left tail than the distribution of individuals with no comorbidities. Exsmokers showed higher variability of scores than never-smokers, indicating heterogeneity that could be explained by differences in time since smoking cessation. Increased variation for patients with a stroke history could indicate differences in disease severity. Extended beta regression have previously been used in economic and psychological applications [21,32,39], and we know of one study that used a beta regression model with dispersion covariates to analyze subscale scores of the SF-36 [25].

Because the beta distribution is only defined on the open unit interval, we used a transformation suggested in literature to slightly compress the range of observed values. As a consequence, the perfect health observations were assigned the value 0.99983 instead of 1. However, our model diagnostics indicated that these observations near the boundary led to increased asymptotic standard errors so that the parameter coefficients were estimated less precisely. In our sample, this effect could be lessened if perfect health was assigned a value of 0.995 rather than 0.99983, suggest-

Table 4 – Regression coefficients of the linear regression model with model-based and robust standard errors.

Covariate	Estimate	Model-based		Robust	
		Standard error	*P*	Standard error	*P*
Intercept	0.8890	0.0174	<0.0001	0.0173	<0.0001
Age (y)	−0.0008	0.0002	0.0001	0.0002	0.0001
BMI	−0.0015	0.0005	0.0032	0.0005	0.0031
Sex*					
Woman	−0.0351	0.0046	<0.0001	0.0045	<0.0001
Marital status†					
With partner	0.0293	0.0052	<0.0001	0.0055	<0.0001
Smoking status‡					
Exsmoker	−0.0022	0.0050	0.6683	0.0051	0.6682
Current smoker	−0.0115	0.0064	0.0733	0.0062	0.0732
Education§					
Primary/secondary	−0.0080	0.0047	0.0881	0.0048	0.0880
Diabetes mellitus	−0.0318	0.0092	0.0006	0.0103	0.0006
Stroke	−0.0779	0.0163	<0.0001	0.0240	<0.0001
Cancer	−0.0211	0.0084	0.0116	0.0092	0.0115
Cardiovascular event	−0.0623	0.0090	<0.0001	0.0101	<0.0001

BMI, Body mass index.
* Reference category: Man.
† Reference category: Single.
‡ Reference category: Never smoker.
§ Reference category: Tertiary education level.

ing that the proposed transformation does not work well in large samples where the difference between the transformed value and one is extremely small. As a consequence, we recommend to try different transformation techniques and to carefully examine their effects on parameter estimates and standard errors. Nevertheless, any approach to handle one values is arbitrary, revealing a major drawback of beta regression compared to other approaches.

An alternative approach to cope with the one-valued observations would be the use of a TPM that explicitly models the probability of one value through a separate logistic regression model [8,12,13]. Such a TPM has also been presented under the name of one-inflated beta regression [40]. However, we decided against the TPM for the following reason: The TPM assumes that data come from two different data-generating processes, the first concerning individuals who tend to have lower utilities and the second concerning individuals with perfect health. This assumption may be appropriate for utility indices with pronounced ceiling effects; however, such a distinction was difficult to justify in our data where the percentage of one-valued observations was very small (3.3%) and where the gap to the next lower value (0.958) was small as well. This point of view is also supported by Li and Fu [12] who stated that one may prefer not to use a TPM if the amount of perfect health states comprises less than 5% of the data.

One may argue that assuming the lower bound at zero is questionable since the smallest possible SF-6D index value is 0.345, a fact related to the well-acknowledged floor effect of the SF-6D [5]. However, as shown in Figure 1, the beta density function on the unit interval fitted well to the observed data. More-

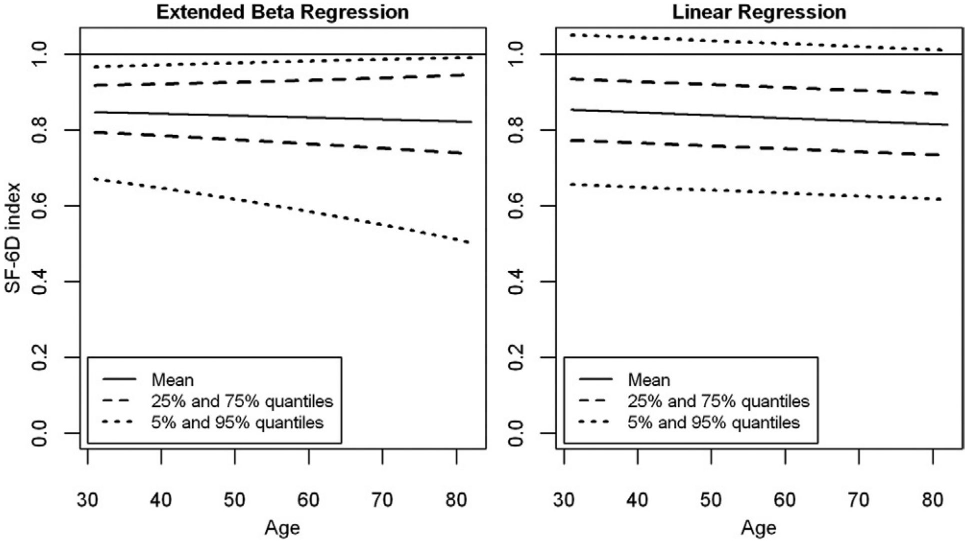

Fig. 2 – Age-specific reference curves for SF-6D index score (male sex, mean BMI, with partner, never smoker, higher education level, and without comorbidities) estimated from the extended beta (left plot) and the linear regression model (right plot).

Table 5 – Predictive accuracy of the beta regression and the linear regression models.

Model	R^1 (95% CI)	R^2 (95% CI)	LogS (95% CI)
Classical beta	0.049 (0.000–0.092)	0.063 (0.000–0.130)	0.694 (0.611–0.769)
Linear regression*	0.061 (0.008–0.108)	0.087 (0.021–0.159)	0.699 (0.616–0.777)
Extended beta	0.065 (0.018–0.111)	0.082 (0.018–0.150)	0.715 (0.620–0.802)

R^1, proportion of the absolute error predicted by the model; R^2, proportion of the square error predicted by the model; LogS, Logarithmic Score, mean log-density of the observations under the model; 95% CI, 95% percentile bootstrap interval.

* Normality assumption of error terms for the calculation of predictive densities in the logarithmic score.

over, the lower boundary at zero also has a theoretical justification due to its equality with death [5].

A surprising result was that age showed a positive (but insignificant) effect on the mean SF-6D index score in the classical beta regression model, and a significant negative effect in the extended beta regression model. Possibly, this finding reflects another manifestation of the high influence that values near the boundary have on estimation. Contrary to what one would expect, there was a relative high proportion of older individuals among the perfect health observations, and it seems likely that these observations influenced the estimation of the overall age effect. This is supported by the fact that the age effect became negative when other transformations of one values were used. The age effect in the mean submodel of the extended beta regression model was negative, indicating that the high proportion of elderly individuals among the perfect health observations was an effect of dispersion rather than of the mean.

In our sample a lot of individuals shared the same SF-6D index value, making our dependent variable in some intervals quasidiscrete. For example, the three most frequent index values—0.863, 0.922, and 0.8—together accounted for almost 50% of respondents. The beta distribution was found to be generally robust against violations of continuity assumptions [41], but further research would be helpful to more thoroughly examine how the shape of the outcome distributions affects the estimation.

Our statistical analyses were carried out using the package betareg in the software R where the extended beta regression model is implemented. Beta regression with precision covariates can also easily be fitted using the module betafit in STATA. In SPSS and SAS, fitting extended beta regression models requires the log-likelihood function to be constructed manually within the NLR procedure and PROC NLMIXED, respectively.

A limitation of our study was that we did not comprehensively study the health status and its relationship to sociodemographic characteristics in the target population. For example we neither explored if our sample differed from the population of interest nor examined interactions between the covariates. The purpose of this study was to compare the performance of beta and linear regression in modelling the SF-6D index as a function of commonly used covariates. Therefore, the KORA F4 data set only served as an example data set.

Another limitation is that the variance function of the beta regression model is inverse U-shaped with its maximum at 0.5, although studies have shown that the variation of health utility scores increases with deteriorating health. Only 40 observations in our data set were lying below 0.5. Thus, as the inverse U-shape implicates monotonicity between 0.5 and 1, the variance function of the beta model does not stand in opposition to the aforementioned findings.

Conclusions

We note that beta regression, especially extended beta regression, is a possible supplement to the currently used methods in the analysis of health utility data. In particular, extended beta regression accounted for the fact that the SF-6D index is bounded at one, and provided information about how covariates change its distribution that is not available in other methods. When making inferences about the mean health utility only, we observed no substantial differences in the predictive accuracy between beta regression and the linear regression model with robust standard errors. A pitfall of the beta regression is that it does not work well in handling one-valued observations, and we suggest to carefully examine their effect on model estimation. In further research, it would be interesting to explore how the beta regression model performs for other HRQL measures. However, application in instruments where utilities equal or inferior to zero are possible (such as EQ-5D index or HUI) is not straightforward.

Source of financial support: The KORA research platform (KORA, Cooperative Research in the Region of Augsburg) was initiated and financed by the Helmholtz Zentrum München - German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research and by the State of Bavaria.

REFERENCES

[1] Drummond MF, Sculpher MJ, Torrance GW, et al. Methods for the Economic Evaluation of Health Care Programmes. 3d ed. Oxford: Oxford University Press; 2005.

[2] Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002;40:113–28.

[3] Dolan P. Modeling valuations for EuroQol health states. Med Care 1997; 35:1095–108.

[4] Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271–92.

[5] Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care 2004;42:851–9.

[6] Austin PC. A comparison of methods for analyzing health-related quality-of-life measures. Value Health 2002;5:329–37.

[7] Kieschnick R, McCollough B. Regression analysis of variates observed on (0,1): percentages, proportions and fractions. Stat Model 2003;3: 193–213.

[8] Pullenayegum EM, Tarride JE, Xie F, et al. Analysis of health utility data when some subjects attain the upper bound of 1: are Tobit and CLAD models appropriate? Value Health 2010;13:487–94.

[9] Barton GR, Sach TH, Doherty M, et al. An assessment of the discriminative ability of the EQ-5Dindex, SF-6D, and EQ VAS, using sociodemographic factors and clinical conditions. Eur J Health Econ 2008;9:237–49.

[10] Dan AA, Kallman JB, Srivastava R, et al. Impact of chronic liver disease and cirrhosis on health utilities using SF-6D and the health utility index. Liver Transpl 2008;14:321–6.

[11] Wee HL, Cheung YB, Loke WC, et al. The association of body mass index with health-related quality of life: an exploratory study in a multiethnic Asian population. Value Health 2008;11(Suppl. 1):S105–14.

[12] Li L, Fu AZ. Some methodological issues with the analysis of preference-based EQ-5D index score. Health Serv Outcomes Res Methodol 2009;9:162–76.

[13] Huang IC, Frangakis C, Atkinson MJ, et al. Addressing ceiling effects in health status measures: a comparison of techniques applied to measures for people with HIV disease. Health Serv Res 2008;43:327–39.

[14] Sullivan PW, Ghushchyan V. Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample. Med Decis Making 2006;26:401–9.

[15] Bharmal M, Thomas J 3rd. Comparing the EQ-5D and the SF-6D descriptive systems to assess their ceiling effects in the US general population. Value Health 2006;9:262–71.

[16] Hanmer J. Predicting an SF-6D preference-based score using MCS and PCS scores from the SF-12 or SF-36. Value Health 2009;12:958–66.

[17] Shaw JW, Pickard AS, Yu S, et al. A median model for predicting United States population-based EQ-5D health state preferences. Value Health 2010;13:278–88.

[18] Thompson SG, Barber JA. How should cost data in pragmatic randomised trials be analysed? BMJ 2000;320:1197–200.

[19] Ferrari SLP, Cribari-Neto F. Beta regression for modeling rates and proportions. J Appl Stat 2004;31:799–815.

[20] Simas AB, Barreto-Souza W, Rocha AV. Improved estimators for a general class of beta regression models. Comput Stat Data Anal 2010;54:348–66.

[21] Smithson M, Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychol Methods 2006;11:54–71.

[22] Hubben GA, Bishai D, Pechlivanoglou P, et al. The societal burden of HIV/AIDS in Northern Italy: an analysis of costs and quality of life. AIDS Care 2008;20:449–55.

[23] Moberg C, Alderling M, Meding B. Hand eczema and quality of life: a population-based study. Br J Dermatol 2009;161:397–403.

[24] Basu A, Manca A. Regression estimators for quality of life and quality-adjusted life years (QALYs). Value Health 2009;12:A28.

[25] Cheung YB, Thumboo J, Machin D, et al. Modelling variability of quality of life scores: a study of questionnaire version and bilingualism. Qual Life Res 2004;13:897–906.

[26] Rathmann W, Haastert B, Icks A, et al. High prevalence of undiagnosed diabetes mellitus in Southern Germany: target populations for efficient screening. The KORA survey 2000. Diabetologia 2003;46:182–9.

[27] Meisinger C, Strassburger K, Heier M, et al. Prevalence of undiagnosed diabetes and impaired glucose regulation in 35-59-year-old individuals in Southern Germany: the KORA F4 Study. Diabet Med 2010;27:360–62.

[28] Holle R, Happich M, Lowel H, et al. KORA–a research platform for population based health research. Gesundheitswesen 2005;67(Suppl. 1):S19–25.

[29] World Health Organization Expert Committee on Physical Status. The Use and Interpretation of Anthropometry. Geneva: World Health Organization; 1995.

[30] Cribari-Neto F, Zeileis A. Beta regression in R. J Stat Softw 2010;34: 1–24.

[31] McCullough P, Nelder JA. Generalized Linear Models. New York: Chapman and Hall; 1989.

[32] Zimprich D. Modeling change in skewed variables using mixed beta regression models. Res Hum Dev 2010;7:9–26.

[33] Long JS, Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. Am Stat 2000;54:217–24.

[34] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction and estimation. J Am Stat Assoc 2007;102:359–78.

[35] Rocha A, Simas A. Influence diagnostics in a general class of beta regression models. Test. [Epub ahead of print] 23 MAR 2010.

[36] Wei BC, Hu YQ, Fung WK. Generalized leverage and its applications. Scand J Stat 1998;25:25–37.

[37] R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.

[38] Michelson H, Bolund C, Brandberg Y. Multiple chronic health problems are negatively associated with health related quality of life (HRQoL) irrespective of age. Qual Life Res 2000;9:1093–104.

[39] Bruche M, González-Aguado C. Recovery rates, default probabilities and the credit cycle. J Banking Finance 2010;34:754–64.

[40] Stasinopoulos DM, Rigby RA. Generalized Additive Models for Location Scale and Shape (GAMLSS). R. J Stat Softw 2007;23:1–46.

[41] Tamhane A, Ankenman B, Yang Y. The beta distribution as a latent response model for ordinal data (I): estimation of location and dispersion parameters. J Stat Comput Sim 2002;72:473–94.