Journal of Engineering Technology and Applied Sciences, 2017 *e-ISSN:* 2548-0391
Received 13 June 2017, Accepted 3 October 2017 *Vol. 2, No. 2, 101-111*
Published online 9 October 2017 *doi: 10.30931/jetas.321165*

Citation: Unlu, H., Aktas, S., "Beta Regression for the Indicator Values of Well-Being Index for Provinces in Turkey". Journal of Engineering Technology and Applied Sciences 2 (2) 2017 : 101-111.

BETA REGRESSION FOR THE INDICATOR VALUES OF WELL-BEING INDEX FOR PROVINCES IN TURKEY

Hande Ünlü[a], Serpil Aktaş[b*]

[a] *Institue of Public Health, Hacettepe University, Ankara, Turkey*
handekonsuk@gmail.com

[b*] *Department of Statistics, Hacettepe University, Ankara, Turkey*
spxl@hacettepe.edu.tr (corresponding author)

Abstract

Beta regression assumes that the dependent variable follows a beta distribution and that its mean is related to a set of exploratory variables through a linear predictor with unknown coefficients and a link function. The model also includes a dispersion parameter. This paper describes the beta regression model along with its properties. The application of the model is made on well-being index data of Turkey 2015, which comprise the dimensions of housing, work, income and wealth, health, education, environment, safety, civic participation and access to infrastructure services and social life. As the life satisfaction index lies between 0 and 1 and the values close to 1 refers to a better level of life. Beta regression fits the data well and the regression parameters are well interpreted in terms of the mean of the response variable.

Keywords: beta distribution, beta regression, dispersion parameter, life satisfaction survey

1. Introduction

The linear regression model, in particular, is commonly used in many field of applied statistics. In classical linear regression model, the normality assumption is required in order to conduct hypothesis testing, particularly if the sample size is small. The ordinary least square regression is not appropriate for regression with a bounded dependent variable, such as index variables. If the response variable is restricted to the interval (0, 1), the fitted values for the variable of interest might exceed its lower or upper bounds. Hence, inference based on the normality assumption can be misleading. Ferrari and Cribari-Neto [1] proposed a regression model where the dependent variable is measured continuously on the standard unit interval, i.e. (0, 1) such as rates and proportions. The proposed model is based on the assumption that the response variable follows a beta distribution. The beta distribution is very flexible for modeling proportions since its density has different shapes depending on the values of the two parameters that index the distribution. The beta density is given by

$$\pi(y,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \;, 0 < y < 1 \tag{1}$$

where $\Gamma(.)$ is the gamma function, $\alpha > 0$, $\beta > 0$. The mean and variance are defined as:

$$E(y) = \frac{\alpha}{(\alpha+\beta)} \tag{2}$$

$$Var(y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{3}$$

As the Beta distribution is characterized by two shape parameters, a simple algebraic transformation of these parameters defines the Beta distribution in terms of its mean and a scaling, or precision, parameter [2][3][4]. Therefore, the Beta Regression accommodates the dependent variable's mean and/or precision as a function of explanatory variables

2. Beta Regression

We will define a regression model for beta distributed random variables. In order to obtain a regression structure for the mean of the response along with a dispersion parameter, Ferrari and Cribari-Neto [1] proposed a different parameterization by setting $\mu = \frac{\alpha}{(\alpha+\beta)}$ and $\phi = \alpha + \beta$.

The density of y can be written, in the new parameterization,

$$f(y,\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \quad 0 < y < 1\;, \tag{4}$$

with $0 < \mu < 1$ and $\phi > 0$. It follows from (2) and (3) that $E(y) = \mu$ and $Var(y) = \frac{V(\mu)}{1+\phi}$. where, $V(\mu) = \mu(1-\mu)$ so that μ is the mean of the response variable and ϕ can be interpreted as a dispersion parameter, for fixed μ. The larger the value of ϕ, the smaller the variance of y.

Let $y_1, \ldots, y_n$ be independent random variables, each y_t follows the density in (1) with mean μ_t and dispersion parameter ϕ. The regression model can be written as

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i \tag{5}$$

where, $\beta = (\beta_1, \ldots, \beta_k)^T$ is a vector of unknown regression parameters, $x_{t1}, \ldots, x_{tk}$ are the fixed covariates; $g(.)$ is a monotonic and double differentiable link function over (0,1).

The beta coefficients give the additional increase or decrease in the log-odds of the response variable. Beta Regression provides more accurate and efficient parameter estimates than ordinary least squares regression when the dependent variable follows a skewed distribution [5] and when there is heteroskedasticity [6].

The usual practice if the response variable lies on the interval (0,1), is to transform y into, $log\left(\frac{y}{(1-y)}\right)$ and to apply a standard linear regression analysis. But this approach has several shortcomings. Such as, the regression parameters are interpretable in terms of the mean of transformed y, and not in terms of the mean of y. The unit interval such as rates and proportions are heteroskedastic and asymmetric. The classical approach to fit a beta regression model is to use maximum likelihood estimation with AIC-based variable selection [5] [7]. In this paper, beta regression is applied to a real data set in which the response variable is proportion. The standard linear regression is also applied to the transformed y to see the accuracy of the beta regression results.

3. Well-Being Index for Provinces Data in Turkey

The data from Turkish Statistics Association are analyzed by the beta regression model. Data in Table 1 include the indicator values of well-being index for provinces, 2015 (http://www.turkstat.gov.tr/PreHaberBultenleri.do?id=24561). Living index in the provinces consists of eleven dimensions which are housing, work, income and wealth, health, education, environment, safety, civic participation and access to infrastructure services and social life. The dimensions are represented by 41 indicators. Level of happiness (y) is measured as a life satisfaction index that lies between 0 and 1 and the values close to 1 refers to a better level of life.

Table 1. Indicator values of well-being index for provinces, 2015.

DIMENSIONS	VARIABLES
Housing	Number of rooms per person (x_1), Toilet presence percentage in dwellings (x_2), Percentage of households having problems with quality of dwellings (x_3)
Work Life	Percentage of households having problems with quality of dwellings (x_4), Employment rate (x_5), Average daily earnings (x_6), Job satisfaction rate(x_7)
Income and Wealth	Savings deposit per capita (x_8), Percentage of households in middle or higher income groups (x_9), Percentage of households declaring to fail on meeting basic needs (x_{10}),
Health	Infant mortality rate (x_{11}), Life expectancy at birth (Years) (x_{12}), Number of applications per doctor (x_{13}), Satisfaction rate with health status (x_{14}), Satisfaction rate with public health services (x_{15})
Education	Net schooling ratio of pre-primary education between the ages of 3 and 5 (x_{16}), Average point of placement basic scores of the system for Transition to Secondary Education from Basic Education (points) (x_{17}), Average points of the Transition to Higher Education Examination (points) (x_{18}), Percentage of higher education graduates (x_{19}), Satisfaction rate with public education services (x_{20})
Environment	Average of PM10 values of the stations (air pollution) (μg/m^3) (x_{21}), Forest area per km^2 (%)(x_{22}),Percentage of population receiving waste services (%)(x_{23}), Percentage of households having noise problems from the streets (%)(x_{24}), Satisfaction rate with municipal cleaning services (%)(x_{25})
Safety	Murder rate (per million people) (x_{26}), Number of traffic accidents involving death or injury (per thousand people) (x_{27}), Percentage of

	people feeling safe when walking alone at night (%) (x_{28}),Satisfaction rate with public safety services (%)(x_{29})
Civic engagement	Voter turnout at local administrations (%)(x_{30}), Rate of membership to political parties (%)(x_{31}), Percentage of persons interested in union/association activities (%)(x_{32})
Access to infrastructure services	Number of internet subscriptions (per hundred persons) (x_{33}), Access rate of population to sewerage and pipe system (%)(x_{34}), Access rate to airport (x_{35}), Satisfaction rate with municipal public transport services (%)(x_{36})
Social Life	Number of cinema and theatre audience (per hundred persons) (x_{37}), Shopping mall area per thousand people (m^2) (x_{38}), Satisfaction rate with social relations (%)(x_{39}),Satisfaction rate with social life (%)(x_{40})
Life satisfaction	Level of happiness (%) (y)

Well-being index allows us to compare well-being across provinces, based on eleven dimensions in the areas of material living conditions and quality of life. The data report that the province with the highest living index is Sinop while the lowest value is Tunceli.
All computations are carried out using betareg package in R Project ver. 3.3.1 [8] and IBM SPSS ver. 22.0 [9].

The arguments of betareg() are:
betareg(formula, data, subset, na.action, weights, offset, link = c("logit", "probit", "cloglog", "cauchit", "log", "loglog"), link.phi = NULL, type = c("ML", "BC", "BR"), control = betareg.control(...), model = TRUE, y = TRUE, x = FALSE, ...).

The formulation of type y ~ x1+...+x40 describes y and x_i for the mean equation of the beta regression.

4. Results

Beta regression model is applied to the data taking the level of happiness as a response variable. Four link functions, logit, loglog, probit and cauchit are used, AIC values and R square values are given in Table 2. Results support that probit link provides a slightly better fit. Selection of an appropriate link function can greatly improve the model fit [10], hence the following results will be over the probit link.

Table 2. AIC and R^2 values of the beta regression models

Link function	Df	AIC	Pseudo R^2
Logit	42	-251.2053	0.8318
Loglog	42	-249.4899	0.8247
Probit	42	**-251.4519**	**0.8335**
Cauchit	42	-249.4178	0.8125

Coefficients of beta regression and their standard errors (Table 3) includes the significance of the independent variables x1-x40. The variables x2, x11, x17, x23, x26, x29 and x36 are found to be statistically significant.

Table 3. Beta regression coefficients (mean model with probit link)

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	1.501e+00	2.625e+00	0.572	0.56751
x1	9.191e-02	1.451e-01	0.633	0.52647
x2	6.366e-03	2.433e-03	2.616	0.00889 **
x3	3.739e-03	4.764e-03	0.785	0.43255
x4	-4.900e-03	3.859e-03	-1.270	0.20423
x5	-8.423e-03	6.061e-03	-1.390	0.16461
x6	-1.151e-03	2.840e-03	-0.405	0.68533
x7	6.582e-03	4.326e-03	1.522	0.12811
x8	6.419e-07	1.136e-05	0.056	0.95495
x9	-4.985e-03	3.452e-03	-1.444	0.14862
x10	-2.669e-03	3.011e-03	-0.886	0.37539
x11	-1.233e-02	6.115e-03	-2.016	0.04380 *
x12	-1.625e-02	2.174e-02	-0.747	0.45478
x13	2.294e-05	1.289e-05	1.780	0.07512
x14	3.860e-03	4.869e-03	0.793	0.42785
x15	8.992e-03	5.508e-03	1.633	0.10257
x16	-1.223e-03	2.583e-03	-0.474	0.63583
x17	-4.854e-03	1.640e-03	-2.959	0.00309 **
x18	-3.183e-03	4.044e-03	-0.787	0.43131
x19	1.110e-02	1.116e-02	0.994	0.32026
x20	3.963e-03	4.256e-03	0.931	0.35175
x21	1.250e-04	7.394e-04	0.169	0.86574
x22	-1.324e-03	1.109e-03	-1.194	0.23234
x23	-4.898e-03	2.195e-03	-2.231	0.02568 *
x24	-7.182e-03	4.085e-03	-1.758	0.07867
x25	-1.834e-03	2.008e-03	-0.914	0.36095
x26	-3.809e-03	1.267e-03	-3.006	0.00265 **
x27	5.041e-02	2.855e-02	1.766	0.07743 ·
x28	4.009e-03	2.374e-03	1.688	0.09134 ·
x29	-1.269e-02	4.780e-03	-2.655	0.00792 **
x30	1.167e-02	7.115e-03	1.640	0.10098
x31	-4.239e-03	3.176e-03	-1.335	0.18202
x32	7.504e-03	6.667e-03	1.126	0.26037
x33	-4.067e-03	1.011e-02	-0.402	0.68756
x34	4.018e-03	2.486e-03	1.616	0.10602
x35	1.726e-05	1.197e-05	1.442	0.14927
x36	-5.541e-03	1.905e-03	-2.908	0.00363 **
x37	7.316e-04	1.089e-03	0.672	0.50175
x38	4.572e-04	3.119e-04	1.466	0.14272
x39	4.842e-03	5.434e-03	0.891	0.37289
x40	2.426e-03	1.814e-03	1.337	0.18118

Signif. codes: ** 0.01; * 0.05 ; ·0.10

Backward stepwise selection method is applied to the model to find the best model (Table 4). Log-likelihood value is 161.2 on 24 degrees of freedom (P<2.2e-16 ***). We can see from the results that the twenty two independent variables among forty have significant effect on level of happiness. While some of the variables have negative effect, some have positive.

Table 4. Model selection (beta regression with probit link function)

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-1,23E+02	7,75E+02	-0.159	0.87370
x2	4,65E+00	1,74E+00	2.673	0.00751 **
x4	-7,85E+00	3,32E+00	-2.363	0.01811 *
x5	-1,41E+01	5,81E+00	-2.426	0.01526 *
x7	7,27E+00	3,53E+00	2.063	0.03915 *
x9	-3,60E+00	2,47E+00	-1.456	0.14527
x11	-7,84E+00	5,34E+00	-1469	0.14176
x15	1,19E+01	3,68E+00	3.239	0.00120 **
x17	-4,86E+00	9,05E-01	-5.363	8.17e-08 ***
x22	-1,17E+00	7,47E-01	-1.567	0.11716
x23	-2,14E+00	1,09E+00	-1.959	0.05014 .
x24	-8,20E+00	3,42E+00	-2.400	0.01640 *
x26	-5,00E+00	9,95E-01	-5.024	5.06e-07 ***
x27	4,64E+01	2,22E+01	2.094	0.03624 *
x28	3,54E+00	2,16E+00	1.644	0.10010
x29	-1,55E+01	3,65E+00	-4.229	2.34e-05 ***
x30	1,38E+01	5,89E+00	2.350	0.01876 *
x32	1,27E+01	5,05E+00	2.512	0.01201 *
x35	1,85E-02	8,99E-03	2.062	0.03919 *
x36	-5,69E+00	1,31E+00	-4.351	1.36e-05 ***
x38	5,67E-01	2,35E-01	2.409	0.01599 *
x39	9,01E+00	4,63E+00	1.946	0.05170 .
x40	4,09E+00	1,64E+00	2.490	0.01278 *
AIC=--274.0317 , Pseudo R^2=0.8043				

Signif. codes: *** 0.001;** 0.01; * 0.05 ; .0.10

The precision parameter ϕ varies through a linear regression structure (Table 5). For fixedμ, the larger ϕ the smaller the variance of y. As the precision parameter estimate is statistically significant, this means that the change in the variance of the dependent variable represents a unit change in the explanatory variables.

Table 5. Phi coefficients (precision model with identity link)

	Estimate	Std. Error	Z-value	Pr(>|z|)
Phi	210.25	32.96	6.378	1.79e-10***

***P< 0.001

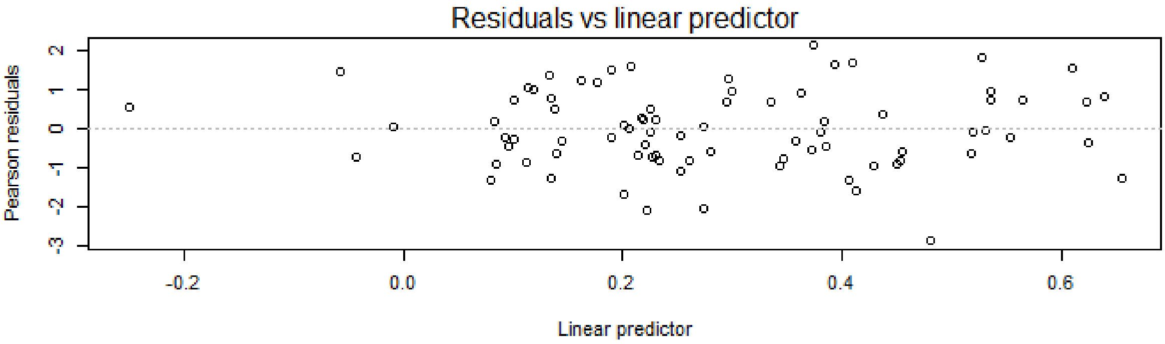

Figure 1. Residual plots for beta regression model

The residuals plot in Figure 1 diagnoses how well our model fits the data. The residuals plots seem to be randomly distributed approximately around zero and show no specific pattern. It seems that there is no outlier since the residuals are bounded (-3, 3).
The statistically significant variables with their associated dimensions are summarized in Table 6.

Table 6. Significant variables on level of happiness for beta regression

DIMENSIONS	**SIGNIFICANT VARIABLES**
Housing	Toilet presence percentage in dwellings (x_2)
Work Life	Percentage of households having problems with quality of dwellings (x_4), Employment rate (x_5), Job satisfaction rate(x_7)
Income and Wealth	Percentage of households in middle or higher income groups (x_9)
Health	Infant mortality rate (x_{11}), Satisfaction rate with public health services (x_{15})
Education	Average point of placement basic scores of the system for Transition to Secondary Education from Basic Education (points) (x_{17})
Environment	Forest area per km^2 (%)(x_{22}),Percentage of population receiving waste services (%)(x_{23}), Percentage of households having noise problems from the streets (%)(x_{24})
Safety	Murder rate (per million people) (x_{26}), Number of traffic accidents involving death or injury (per thousand people) (x_{27}), Percentage of people feeling safe when walking alone at night (%) (x_{28}), Satisfaction rate with public safety services (%)(x_{29})
Civic engagement	Voter turnout at local administrations (%)(x_{30}), Percentage of persons interested in union/association activities (%)(x_{32})
Access to	Access rate to airport (x_{35}), Satisfaction rate with municipal public transport

infrastructure services	services (%)(x_{36})
Social Life	Shopping mall area per thousand people (m^2) (x_{38}), Satisfaction rate with social relations (%)(x_{39}), Satisfaction rate with social life (%)(x_{40})
Life satisfaction	Level of happiness (%) (y)

We also consider the standard linear regression model using the $\tilde{y}= log\left(\frac{y}{(1-y)}\right)$ transformation (Table 7-8). As the data consist of ten dimensions, the variables under the dimensions are likely to have multicollinearity problem. For the full model, VIF values refer to a situation in which two or more explanatory variables in a multiple regression model are highly correlated.

Table 7. Backward model selection summary of standard linear regression model

Model Summary[b]					
R	R Square	Adjusted R Square	Std. Error of the Estimate	F	Sig.
0.878[a]	0.771	0.718	0.17494	14.386	<0.0001

a. Predictors: (Constant), x2, x4, x5, x11, x15, x17, x24, x26, x27, x28, x29, x35, x36, x38, x40

b. Dependent Variable: $\tilde{y}= log\left(\frac{y}{(1-y)}\right)$

Table 8. Standard linear regression results

	B	Std. Error	t	Sig.
(Constant)	3.391	0.745	4.553	<0.001
x2	0.009	0.003	3.387	.001
x4	-0.011	0.006	-1.891	.063
x5	-0.024	0.011	-2.274	.026
x11	-0.026	0.011	-2.393	.020
x15	0.021	0.007	3.138	.003
x17	-0.010	0.002	-5.819	<0.001
x24	-0.020	0.006	-3.362	.001
x26	-0.008	0.002	-4.418	<0.001
x27	0.101	0.039	2.554	.013
x28	0.010	0.004	2.365	.021
x29	-0.020	0.007	-3.028	.004
x35	2.956E-05	.000018	1.673	.099
x36	-0.011	0.002	-4.628	<0.001
x38	0.001	0.00004	1.873	.066
x40	0.009	0.003	3.561	.001
a. Dependent Variable:$\tilde{y}= log\left(\frac{y}{(1-y)}\right)$				

The model selection procedure is performed for the standard linear regression. This procedure gives only the variables that are statistically significant. The best subset includes the variables x2, x4, x5, x11, x15, x17, x24, x26, x27, x28, x29, x35, x36, x38 and x40 with larger coefficients and standard errors when compared to the coefficients of beta regression. As a result of linear regression and beta regression, some common variables are found as follows x2, x4, x5, x11, x15, x17, x24, x26, x27, x28, x29, x35, x36, x38. These common parameters have the same effect i.e. positive or negative on response variable similarly in linear regression and beta regression. No overlapping occurs between standard linear regression and beta regression results.

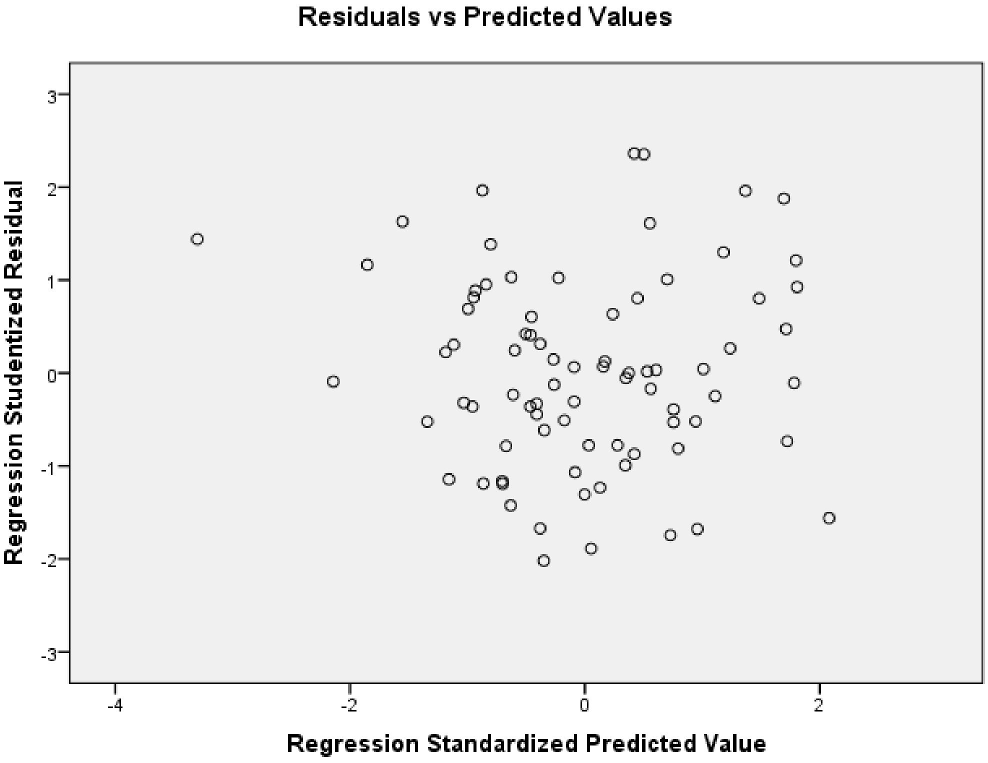

Figure 2. Residual plots for linear regression

The residuals plot in Figure 2 diagnose how well our model fits the data for linear regression. The residuals plot seem to be randomly distributed approximately around zero and shows no specific pattern. It seems that there is no outlier since the residuals are bounded (-3, 3).

Table 9. Summary table of Beta regression and liner regression

Link function	AIC	R^2	P
Probit	**-274.0317**	**0.8043**	**< 2.2e-16** ***
Linear Regresion	-35.754	0.771	<0.00001

***P< 0.001

AIC and R^2 values of beta regression with probit link function and linear regression are given in Table 9. Both of the models are found as statistically significant. In beta regression model, AIC values are lower and R^2 are greater than in linear regression. It is concluded that beta regression model fits the data best.

5. Conclusions

Real data often fail to hold the normality assumption. The flexibility of the Beta distribution enables to accommodate the skew data. If the multicolinarity problem exists when two or more of the predictors in a regression model are correlated, the precision of the estimated regression coefficients are questionable. Beta regression provides more accurate and efficient parameter estimates than ordinary least squares regression when the dependent variable follows a skewed underlying distribution or when there is a heteroskedasticity. The examined model here is useful for situations where the response variable is continuous and restricted to the interval (0,1) [1]. Moreover, the Pearson residuals are not necessarily expected to be normal and centered around zero. In beta regression, we are basically using the results and inferences from standard general linear modeling. This paper uses a regression model where the response is beta distributed using a parameterization of the beta distribution that is shaped by mean and dispersion parameters. The results show that the 22 independent variables in Table 6 among 40 have significant effect on level of happiness.

References

[1] Ferrari S., Cribari-Neto, F. “Beta Regression for Modelling Rates and Proportions.” Journal of Applied Statistics 31(7) (2004): 799-815.

[2] Swearingen, C.J., Castro, M.S.M. & Bursac, Z. Modeling percentage outcomes: the %Beta Regression macro. SAS Global Forum 2011, (2011) Paper 335-2011. Available at http://support.sas.com/resources/papers/proceedings11/335-2011.pdf.

[3] Smithson M, Verkuilen J. “A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables.” Psychol Methods 11(1) (2006):54–71.

[4] Ospina, R., Cribari-Neto, F., and Vasconcellos, K.LP. “Improved point and interval estimation for a beta regression model.” Computational Statistics & Data Analysis 51(2) (2006):960–981

[5] Paolino P. “Maximum likelihood estimation of models with beta-distributed dependent variables.” Political Analysis, 9(2011):325-346.

[6] Kieschnick R, McCullough BD. “Regression analysis of variates observed on (0,1): percentages, proportions and fractions.” Statistical Modelling 3(2003):193-213.

[7] Rocha AV, Simas AB.. “Influence diagnostics in a general class of beta regression models. Test, epub 23.Swearingen CJ, Melguizo castro MS, and Bursac Z. Modeling percentage outcomes: The %Beta_Regression macro. SAS® Global Forum Proceedings “(2011); Paper 335:1–12.

[8] Cribari-Neto F, Zeileis A. “Beta Regression in R.” Journal of Statistical Software, 34(2) (2010) 1–24.

[9] IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

[10] McCullagh, P. and Nelder, J. A. Generalized Linear Models. Chapman & Hall, London, 2nd edition (1989).