International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2015): 6.391

Modelling the US Diabetes Mortality Rates via Generalized Linear Model with the Tweedie Distribution

Oznur Ozaltin[1], Neslihan Iyit[2, 3]

[1]Ataturk University, Faculty of Science, Department of Mathematics, Erzurum, Turkey
[2]Selcuk University, Faculty of Science, Department of Statistics, Alaeddin Keykubat Campus, Konya, Turkey
[3]Corresponding Author E-mail: niyit@selcuk.edu.tr

Abstract*: In this study, we are interested in modelling the response variable as the US diabetes mortality rate in the aspect of different types of neoplasms, endocrine, nutritional and metabolic diseases, musculoskeletal system diseases, obesity, sugar intake, and alcohol use disorder via generalized linear model (GLM) with the Tweedie distribution. In this study, firstly, we will focus on the effects of changing the variance power parameter and the index of the power link function on the AIC goodness-of-fit test statistic and also Pearson chi-square and deviance statistics for the dispersion parameter and the residuals in the GLMs with the Tweedie distribution for the US diabetes mortality data. The best link function is determined as "identity" with the variance power parameter "1.9" and the link function power "1" belonging to the Tweedie distribution in the GLM for the US diabetes mortality data. Secondly, the importance of model diagnostic plots based on the residuals, Cook's distance and leverage is emphasized to determine the extreme observations that may cause some problems for parameter estimations, hypothesis tests, and statistical inferences in the GLM for the US diabetes mortality data from the Tweedie distribution.*

Keywords: Diabetes mortality, generalized linear model, Tweedie distribution, link function

1. Introduction

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. World Health Organization (WHO) projects that diabetes will be the seventh leading cause of death in 2030[44].

In many studies, diabetes mortality has been investigated in different aspects to enable a better understanding of the risk factors affecting it. Especially diabetes mortality is one of the most important causes of early mortality in the *US* than many European countries [37].In the literature, Goodkin (1975), Fuller et al. (1983), Williamson et al. (2000), Cifuentes et al. (2000), Kaati et al. (2002), Wen et al. (2005), Franco et al. (2007) and Secrest et al. (2014) investigated diabetes mortality risk factors in different aspects. A large number of studies on diabetes mortality have been included in the literature.

In recent years, generalized linear models (*GLMs*) including regression models based on the exponential family of distributions with the flexibility of modelling the probability distribution of the continuous and discrete type response variables have gained popularity. In the literature, Nelder and Wedderburn (1972), Cameron and Trivedi (1986), McCullagh and Nelder (1989), Firth (1991), Liao (1994), Blough et al. (1999), Lindsey (2000), Diggle (2002), Renshaw and Haberman (2003), Dobson and Barnett (2008), Fitzmaurice et al. (2012), Grover et al. (2013), Agresti(2015) and Iyit et al. (2016) investigated *GLMs* approach in details.

GLM with the response variable coming from the Tweedie distribution as a member of the class of mixed distributions known as the Tweedie family has attracted great interest in statistical modelling especially in actuarial sciences and risk modelling. Jorgensen and Paes De Souza (1994), Dunn and Smyth (2001), Smyth and Jorgensen (2002), Wuthrich (2003), Candy (2004), Dunn (2004), Dunn and Smyth (2005), Kaas (2005), Dunn and Smyth (2008), Shono (2008), Brown and Dunn (2011), Zhang (2013) and Simsekli et al. (2015) are good and exciting references for the *GLM* with the Tweedie distribution in the literature.

In this study, we will focus on statistical modelling of the response variable as the *US* diabetes mortality rate in the aspect of different types of neoplasms, endocrine, nutritional and metabolic diseases, musculoskeletal system diseases, obesity, sugar intake, and alcohol use disorder via *GLM* with the Tweedie distribution. With the feature of the attractiveness of the Tweedie distribution still not well-known and not widely used, this study has not been done before for the diabetes mortality data in the literature.

2. Materials and Method

Generalized linear models (*GLMs*) include regression models based on the exponential family of distributions. In *GLMs*, the distribution of the response variable comes from the exponential family. A monotonic and differentiable link function as $g(\mu)$ exists that linearizes the relationship between the linear predictor $\eta = X\beta$ and the mean of the response variable $E(Y) = \mu$ [21].

The Tweedie [41] distribution is useful to model a response variable $y \geq 0$. The response variable having the Tweedie distribution belonging to the exponential family can be modelled by using *GLMs* approach. The variance of the Tweedie distributed response variable is;

$$V(Y) = \phi\mu^p \quad (1)$$

where $E(Y)=\mu$ is the mean of the Tweedie distributed response variable, ϕ is the dispersion parameter and p is an extra variance power parameter over the mean in the variance of the response variable, respectively.

The Tweedie family is known as the class of mixed distributions [36], [40].The Tweedie family of distributions for specific values of the variance power parameter p in Eq.(1) are given in Table 1.

Table 1: Many important known distributions as special cases ofthe Tweedie family

p values	Distributions
p=0	Normal
p=1	Poisson
1<p<2	Tweedie
p=2	Gamma
p=3	Inverse Gaussian

The Tweedie distribution for 1<p<2 is a special case of the Tweedie family and less well known from other distributions [10]-[13]. For 1<p<2, the Tweedie distribution is a compound Poisson-gamma mixture distribution, which is the distribution of S defined as [36];

$$S=\sum_{i=1}^{N} Y_i \qquad (2)$$

where $N \sim Poisson(\lambda)$ is independently and identically distributed (i.i.d) Poisson random variable with λ parameter and $Y_i \sim Gamma(\alpha,\theta)$ is i.i.d gamma random variables with the shape parameter α and the scale parameter θ [6], [13], [14], [36] .

For p>1, the Tweedie family has the following form;

$$f(y;\mu,\phi,p)=a(y,\phi)\exp\left[\frac{1}{\phi}\left(\frac{y\mu^{1-p}}{1-p}-\kappa(\mu,p)\right)\right] \qquad (3)$$

where $\kappa(\mu,p)=\mu^{2-p}/2-p$ for $p\neq 2$ and $\kappa(\mu,p)=\log(\mu)$ for $p=2$, respectively. The function $a(y,\phi)$ does not have an analytical expression. It is usually evaluated by using the series expansion methods described in [13], [36].

Parameter estimates in *GLM* for the Tweedie distribution are obtained traditionally by using iteratively reweighted least squares (*IRLS*) method. *IRLS* method both linearizes the relationship between the model linear predictor $\eta=X\beta$ and the fitted value for the expected value of the response variable $E(Y)=\mu$, and provides a robust way to estimate model parameters. Deviance statistic is used as the convergence criteria in *IRLS* method and also used as the goodness-of-fit test statistic in *GLMs*. In *IRLS* method, as an advantage, convergence is achieved in a few iterations when the change in deviance statistic values between two iterations is below a specified level of tolerance generally taken as 10^{-6} in the *R version 3.4.0*[21].

In this study, it is interested in the situation where the variance power parameter p is between 1 and 2.

Some possible link functions in *GLMs* for the Tweedie distribution are given in Table 2 where "link power" indicates index of the power link function. In *GLMs*, generally index of the power link function takes values in the range from -3 to +3 [13], [21]. In this study, we are interested in the indexes of the power link function from -1 to +1.

Table 2: Relationships between some link powers and link functions in *GLMs* for the Tweedie distribution [21]

Link power	Link function
3	*cubic*
2	*quadratic*
1	*identity*
0.5	*square root*
0	*log*
-0.5	*inverse square root*
-1	*inverse*
-2	*inverse quadratic*
-3	*inverse cubic*

Determining the best link function in *GLMs* for the Tweedie distribution among possible link functions given in Table 2 via goodness-of-fit test statistics is an important problem for obtaining accurate interpretations of the model parameter estimates. The best link function in *GLMs* for the Tweedie distribution is the one with the smallest value of Akaike Information Criterion (AIC) [2];

$$AIC=-2(L-k) \qquad (4)$$

as the goodness-of-fit test statistic. In Eq.(4), L indicates the value of the model log-likelihood and k specifies the number of independent variables in the model, including the intercept term [21].In *GLMs*, deviance measures how closely the model-based fitted values of the response variable approximate to the observed values of the response variable [8], [22].

The Pearson dispersion is the ratio of the model Pearson chi-square dispersion statistic to the model degrees of freedom defined as the number of observations in the *GLM* minus the number of independent variables in the model, including the intercept term [21].

In the analysis of residuals, Pearson residuals and deviance residuals are important model diagnostic tools for *GLMs*. The Pearson residual is defined as;

$$R_P=\frac{y-\hat{\mu}}{\sqrt{V(\hat{\mu})}} \qquad (5)$$

The sum of squared Pearson residuals is the Pearson chi-square residual statistic. The deviance residual is defined as;

$$R_D=\mathrm{sgn}(y-\hat{\mu})\sqrt{deviance} \qquad (6)$$

The deviance residual statistic is then defined as the sum of squared deviance residuals [21].

Visually, *GLM* diagnostic plots can be drawn by using each observation's Cook's distances and leverage values besides the residual values to detect influential outliers that negatively affect the model.

MAE; mean absolute error, *RMSE*; root mean square error, *NRMSE*; normalized root mean square error, *RSR*; ratio of *RMSE* to the standard deviation of the observations, *mNSE*;

modified Nash-Sutcliffe efficiency, *md*; modified index of agreement, and *VE*; volumetric efficiency between predicted values from the *GLM* and observed values from the dataset are the alternative goodness-of-fit measures in *GLM* validation [28], [27], [31], [47].

In this study, by using *R version 3.4.0*, simulations are done to investigate the effect of different values of the variance power parameter (*p)* (1<p<2) on the Tweedie distribution when the value of the dispersion parameter ϕ and the mean value of the Tweedie distributed response variable μ are taken as 1. Initially, 50 random numbers, in the next case 100 random numbers, and finally 1000 random numbers are generated from the Tweedie distribution as given in Figure 1, Figure 2, and Figure 3, respectively

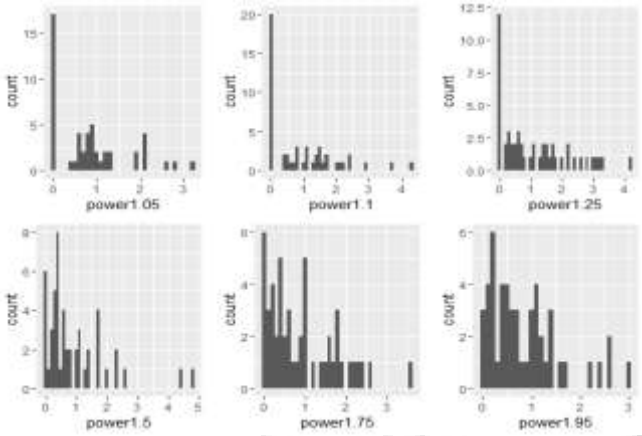

Figure 1: Histograms of 50 random numbers generated from the Tweedie distribution for different *p* values (1<p<2)

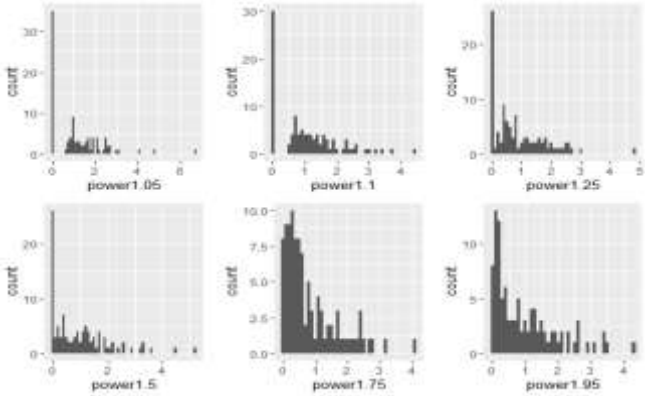

Figure 2: Histograms of 100 random numbers generated from the Tweedie distribution for different *p* values (1<p<2)

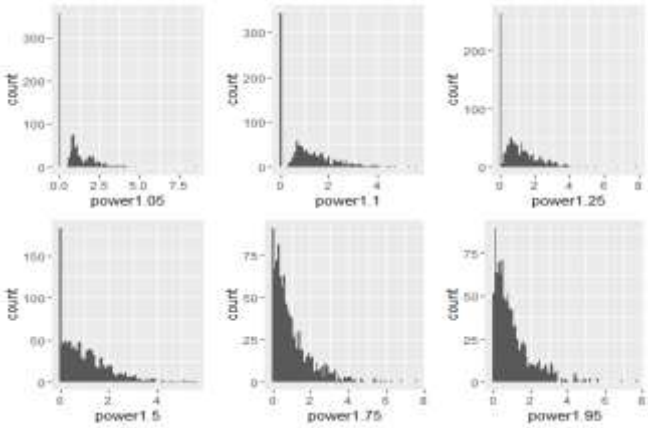

Figure 3: Histograms of 1000 random numbers generated from the Tweedie distribution for different *p* values (1<p<2)

As seen from Figure 1 to Figure 3, as the generated random numbers go up to 1000 from 50, and the variance power parameter (*p)* value from 1.05 to 1.95, it can be easily observed that the shape of the distribution comes closer to Gamma distribution from Poisson distribution as indicated in Table 1.

3. Generalized Linear Model for the US Diabetes Mortality Data from the Tweedie Distribution

In this study, by using the *R version 3.4.0*and *IBM SPSS 23.0*, an application of the Tweedie distribution on diabetes mortality data for the *US* is investigated by using *GLM* approach. The data used in this study are taken from the OECD Database [34].

In this study, sugar intake, certain infectious and parasitic diseases, malignant neoplasms, diseases of the blood; endocrine, nutritional and metabolic diseases, mental and behavioral disorders, diseases of the nervous system, and circulatory system; ischemic heart diseases, diseases of the respiratory system, digestive system, skin, and musculoskeletal system; and also obesity are taken as candidate independent variables to the *GLM* for the *US* diabetes mortality data. Response variable is taken as "diabetes mortality rate" of the*US* between 1960 and 2010 to the *GLM* with the Tweedie distribution by using different link functions. The link functions are taken as identity, log, inverse square root and inverse link functions. Also, *AIC* is used as goodness-of-fit test statistic for comparing these different link functions and different values of the variance power parameter (*p*) for the Tweedie distribution in the *GLM* approach for the same data. Furthermore, the dispersion parameter and the residuals are examined via the deviance and Pearson chi-square statistics by using *R version 3.4.0*.

4. Results and Discussion

In Table 3, according to the different values of the variance power parameter as "var.power (*p*)" between 1 and 2 given in Table 1, and the indexes of the power link function between -1 and 1 given in Table 2, 36 different *GLM*s for the US diabetes mortality data from the Tweedie distribution are constituted and compared by using *AIC* goodness-of-fit test statistic. And also, Pearson chi-square and deviance statistics for the dispersion parameter and the residuals are obtained in Table 3.

Table 3: Results for comparing *GLMs* for the *US* diabetes mortality data in the cases of different values of the variance power parameter (p) and link function powers

Candidate GLMs	*Var. power (p)*	*Link power*	*AIC*	*Pearson chi-square residual statistic*	*Deviance residual statistic*	*Pearson chi-square dispersion statistic*	*Deviance dispersion statistic*
GLM 1	1.1	1	62.70377	0.06669965	0.06674653	0.00476426	0.004767609
GLM 2	1.2	1	62.65253	0.048293689	0.04833094	0.003449549	0.00345221
GLM 3	1.3	1	62.60121	0.034966839	0.034996	0.002497631	0.00249973
GLM 4	1.4	1	62.54975	0.025317509	0.025341	0.001808393	0.001810039
GLM 5	1.5	1	62.49808	0.018330892	0.01834885	0.001309349	0.001310632
GLM 6	1.6	1	62.44614	0.0132722258	0.01328616	0.0009480161	0.0009490117
GLM 7	1.7	1	62.39385	0.00960950	0.009620275	0.000686393	0.0006871625
GLM 8	1.8	1	62.34117	0.0069575194	0.006965817	0.0004969657	0.0004975583
GLM 9*	**1.9**	**1**	**62.28802**	**0.0050373702**	**0.005043739**	**0.0003598122**	**0.0003602671**
GLM 10	1.1	0	63.19473	0.067329455	0.06739216	0.004809247	0.004813726
GLM 11	1.2	0	63.15664	0.048761107	0.04881102	0.003482936	0.003486502
GLM 12	1.3	0	63.11897	0.035313869	0.03535331	0.002522419	0.002525236
GLM 13	1.4	0	63.08168	0.025575253	0.02560622	0.001826804	0.001829016
GLM 14	1.5	0	63.0447	0.018522390	0.01854657	0.001323028	0.001324755
GLM 15	1.6	0	63.00798	0.0134145528	0.01343334	0.0009581823	0.0009595242
GLM 16	1.7	0	62.97148	0.0097153181	0.009729853	0.0006939513	0.0006949895
GLM 17	1.8	0	62.93514	0.0070362151	0.007047417	0.0005025868	0.000503387
GLM 18	1.9	0	62.89891	0.0050959134	0.005104518	0.0003639938	0.0003646084
GLM 19	1.1	-0.5	64.56295	0.06915174	0.0692246	0.00493941	0.004944614
GLM 20	1.2	-0.5	64.53523	0.050090225	0.05014842	0.003577873	0.00358203
GLM 21	1.3	-0.5	64.50801	0.036283269	0.03632941	0.002591662	0.002594958
GLM 22	1.4	-0.5	64.48121	0.026282277	0.02631862	0.001877305	0.001879901
GLM 23	1.5	-0.5	64.45479	0.01903804	0.01906651	0.00135986	0.001361893
GLM 24	1.6	-0.5	64.42869	0.0137906210	0.01381281	0.0009850444	0.0009866291
GLM 25	1.7	-0.5	64.40285	0.0099895805	0.0100068	0.0007135415	0.0007147712
GLM 26	1.8	-0.5	64.37723	0.0072362261	0.007249533	0.0005168733	0.0005178238
GLM 27	1.9	-0.5	64.35176	0.0052417711	0.00525202	0.0003744122	0.0003751443
GLM 28	1.1	-1	66.60132	0.07196462	0.07204736	0.00514033	0.00514624
GLM 29	1.2	-1	66.58842	0.052142109	0.05220848	0.003724436	0.003729177
GLM 30	1.3	-1	66.57613	0.037780045	0.03783288	0.002698575	0.002702348
GLM 31	1.4	-1	66.56441	0.027374116	0.0274159	0.001955294	0.001958278
GLM 32	1.5	-1	66.55317	0.019834487	0.01986734	0.001416749	0.001419096
GLM 33	1.6	-1	66.54234	0.014371581	0.01439729	0.001026542	0.001028378
GLM 34	1.7	-1	66.53186	0.0104133463	0.01043337	0.0007438105	0.0007452406
GLM 35	1.8	-1	66.52166	0.0075453216	0.007560854	0.0005389515	0.000540061
GLM 36	1.9	-1	66.51168	0.0054672179	0.005479224	0.0003905156	0.0003913732

**The smallest values for AIC, Pearson chi-square and the deviance statistics for the residuals and the dispersion parameter indicate that the best link function is "identity" with the variance power parameter "1.9" and the link function power "1"belonging to the Tweedie distribution in the GLM for the US diabetes mortality data.*

Parameter estimates, standard errors of the parameter estimates, *z*-test statistic values with the corresponding significance values for 4 statistically significant independent variables with 2interaction terms in the best *GLM* constituted for the *US* diabetes mortality data are given in Table 4, by using *R version 3.4.0*at $\alpha = 0.05$ significance level.

In Table 4, parameter estimates in the *GLM* for the *US* diabetes mortality data coming from the Tweedie distribution are obtained by using iteratively reweighted least squares (*IRLS*) method in *R version 3.4.0*.

Table 4: Results of the best *GLM* for the *US* diabetes mortality data from the Tweedie distribution in the case of the variance power parameter=1.9 and the link function power=1

Independentvariables	$\hat{\beta}$	s.e$(\hat{\beta})$	Hypothesis Test		95% Confidence Interval	
			zvalue	Pr(>|z|)	Lower	Upper
Intercept	-37.44900	17.28953	-2.166	0.030312	-71.3364788	-3.5615212
malignantneoplasms	-0.68858	0.34877	-1.974	0.048348	-1.3721692	-0.0049908
malignantneoplasms of trachea, bronchusandlung	2.32193	0.89398	2.597	0.009396	0.5697292	4.0741308
malignantmelanoma of skin	1.12913	0.39406	2.865	0.004165	0.3567724	1.9014876
obesity	48.42683	15.15225	3.196	0.001393	18.72842	78.12524
sugarintake × endocrine, nutritionalandmetabolicdiseases	2.58751	0.85443	3.028	0.002459	0.9128272	4.2621928
alcohol × obesity	-5.02772	1.49233	-3.369	0.000754	-7.9526868	-2.1027532

In the final stage, residuals, Cook's distance and leverage plots as the model diagnostics are investigated for the best *GLM* belonging to the *US* diabetes mortality data from the Tweedie distribution in Figure 4.

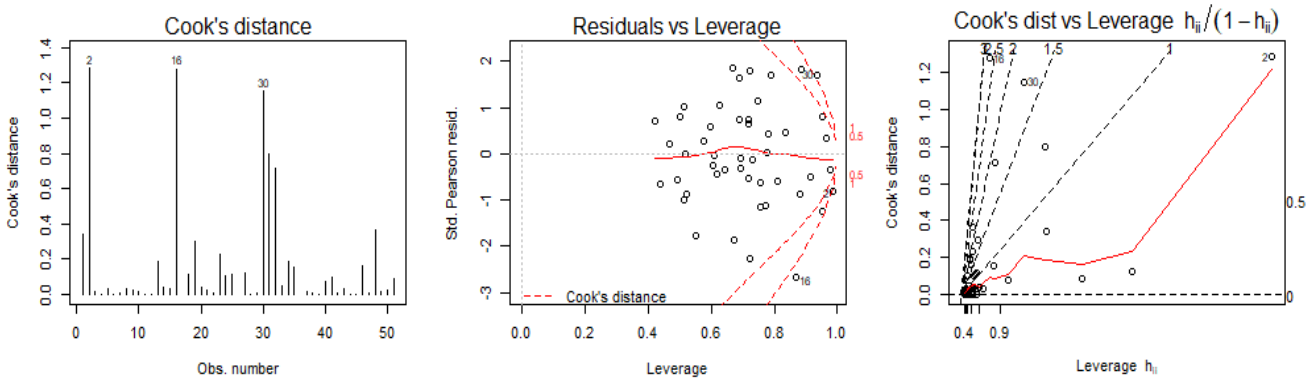

Figure 4: *GLM* diagnostic plots for the *US* diabetes mortality data from the Tweedie distribution

As seen from Figure 4, model diagnostic plots based on the values of the residuals, Cook's distance and leverage indicate that 2, 16., and 30. Observations seem as extreme values that may cause some problems for parameter estimations, hypothesis tests, and statistical inferences in the *GLM* for the *US* diabetes mortality data from the Tweedie distribution. There are lots of discussions about what to do on problems occurring of these extreme observations in *GLM*. In our opinion, when such a situation is encountered, the best thing to do is to omit them. And then the *GLM* for the *US* diabetes mortality data from the Tweedie distribution must be reconstructed again under the same conditions after ignoring these observations.

Parameter estimates, standard errors of the parameter estimates, *z*-test statistic values with the corresponding significance values for 6 statistically significant independent variables with 3 interaction terms in the reconstructed *GLM* constituted for the *US* diabetes mortality data from the Tweedie distribution are given in Table 5, by using the *IRLS* method in *R*version 3.4.0 at $\alpha = 0.05$ significance level.

Table 5: Results of the reconstructed *GLM* for the *US* diabetes mortality data from the Tweedie distribution in the case of variance power parameter=1.9 and the link function power=1

Independent variables	$\hat{\beta}$	s.e$(\hat{\beta})$	Hypothesis Test		95% Confidence Interval	
			z value	Pr(>|z|)	Lower	Upper
malignant neoplasms	-1.449627	0.338299	-4.285	1.83e-05	-2.11269304	-0.78656096
malignant neoplasms of trachea, bronchus and lung	-2.217560	0.690512	-3.211	0.00132	-3.57096352	-0.86415648
malignant melanoma of skin	1.102750	0.273787	4.028	5.63e-05	0.56612748	1.63937252
endocrine, nutritional and metabolic diseases	-3.871801	1.802597	-2.148	0.03172	-7.40489112	-0.33871088
diseases of the musculoskeletal system	2.260208	0.854840	2.644	0.00819	0.5847216	3.9356944
Obesity	38.944957	13.140215	2.964	0.00304	13.1901356	64.6997784
sugar intake × endocrine, nutritional and metabolic diseases	4.914137	1.949977	2.520	0.01173	1.09218208	8.73609192
sugar intake × malignant neoplasms	1.040975	0.431513	2.412	0.01585	0.19520952	1.88674048
alcohol × obesity	-3.939097	1.322117	-2.979	0.00289	-6.53044632	-1.34774768

Useful alternative goodness-of-fit measures to compare fitted *GLMs* validations for the *US* diabetes mortality data from the Tweedie distribution are given in Table 6.

As seen from Table 6, after omitting extreme observations from the dataset as a result of examining the residuals, Cook's distance and leverage, the new fitted *GLM* for the *US* diabetes mortality data from the Tweedie distribution gives better alternative goodness-of-fit measures than the old best model.

Table 6: Alternative goodness-of-fit measures to compare fitted *GLMs* validations for the *US* diabetes mortality data from the Tweedie distribution

Models	MAE	RMSE	NRMSE	RSR	mNSE	md	VE
Theold best *GLM*	0.16	0.22	6.8	0.07	0.94	0.97	0.99
The new *GLM*	0.11	0.15	4.5	0.05	0.96	0.98	1.00

From Table 6, it is easily observed that *MAE*, *RMSE*, *NRMSE*, and *RSR* values of the new *GLM* are smaller than the old fitted best *GLM* for the *US* diabetes mortality data. On the other hand, *mNSE*, *md*, and *VE* values of the new *GLM* are greater than the old fitted best *GLM* for the *US* diabetes mortality data. These values of the alternative goodness-of-fit measures and the *IRLS* estimates of the dispersion parameter as 0.0002366898 and 9.568814e-05 for the old fitted best *GLM* and the new *GLM*, respectively

indicate that the new *GLM* for the *US* diabetes mortality data from the Tweedie distribution better fit to the data than the old best *GLM* after omitting extreme observations.

In Table 7, according to the variance power parameter=1.9 and the link function power=1, the old best *GLM*, and the new *GLM* for the *US* diabetes mortality data from the Tweedie distribution are compared by using the values of *AIC* goodness-of-fit test statistic and also Pearson chi-square and deviance statistics for the residuals and the dispersion parameter.

Table 7: Comparison of the old fitted best *GLM* and the new *GLM* for the*US* diabetes mortality data from the Tweedie distribution in the case of the variance power parameter=1.9 and the link function power=1

Model	AIC	Pearson chi-square residual statistic	Deviance residual statistic	Pearson chi-square dispersion statistic	Deviance dispersion statistic
The old best *GLM*	62.28802	0.0050373702	0.005043739	0.0003598122	0.0003602671
The new *GLM*	22.89971	0.0019396973	0.0019416	0.0001939697	0.0001941602

The smallest values of *AIC*, and also Pearson chi-square and deviance statistics for the residuals and the dispersion parameter indicate that the new *GLM* for the *US* diabetes mortality data from the Tweedie distribution better fits to the data than the old best *GLM*.

The values of the alternative goodness-of-fit measures in table 6 and the statistics in Table 7 indicate that the new *GLM* for the *US* diabetes mortality data from the Tweedie distribution better fits to the data than the old best *GLM* after omitting extreme observations.

The values of the Pearson residuals and the deviance residuals for the *US* diabetes mortality rates between 1960 and 2010 by the new *GLM* with the Tweedie distribution are given in Table 8.

The new *GLM* for the*US* diabetes mortality data from the Tweedie distribution with "identity" link function by using the parameter estimates given in Table 5 is as follows;

$$\begin{aligned} diabetes\,mortality\,rate = & -1.449627(mal.neo.) - 2.217560(mal.neo.lung) \\ & + 1.102750(mal.neo.skin) - 3.871801(endocrine\,diseases) \\ & + 2.260208(dis.musculoskeletal) + 38.944957(obesity) \\ & + 4.914137(sugar\,int\,ake) \times (endocrine\,diseases) \\ & + 1.040975(sugar\,int\,ake) \times (mal.neo.) \\ & - 3.939097(alcohol) \times (obesity) \end{aligned} \quad (7)$$

As seen from Table 8, the deviance residuals and the Pearson residuals lie within the range (-0.02102, 0.01502) and (-0.02091, 0.01508) belonging to the *US* diabetes mortality rates between 1960 and 2010 by the new *GLM* with the Tweedie distribution, respectively.

On the other hand, the deviance residuals and the Pearson residuals lie within the range (-0.02295, 0.02025) and (-0.02281, 0.02036) belonging to the *US* diabetes mortality rates between 1960 and 2010 by the old best *GLM* with the Tweedie distribution, respectively.

Line graph of the actual vs. predicted values of the *US* diabetes mortality rates between 1960 and 2010 by the new *GLM* with the Tweedie distribution is given in Figure 5.

Table 8: Pearson and deviance residuals values for the *US* diabetes mortality rates between 1960 and 2010 by the new *GLM* with the Tweedie distribution

Years	**1960**	**1962**	**1963**	**1964**	**1965**	**1966**	**1967**	**1968**
Pearson residuals	-0.00223	0.00309	0.00036	-0.00452	0.00542	0.00217	-0.00432	-0.00903
Deviance residuals	-0.00223	0.00309	0.00036	-0.00453	0.00541	0.00216	-0.00433	-0.00905
Years	**1969**	**1970**	**1971**	**1972**	**1973**	**1974**	**1976**	**1977**
Pearson residuals	0.00364	0.01262	-0.00912	0.00352	0.00163	-0.00340	0.00000	-0.00331
Deviance residuals	0.00364	0.01258	-0.00915	0.00352	0.00163	-0.00340	0.00000	-0.00332
Years	**1978**	**1979**	**1980**	**1981**	**1982**	**1983**	**1984**	**1985**
Pearson residuals	0.00333	-0.00012	0.00335	-0.00668	0.00338	0.00610	-0.00594	0.00000
Deviance residuals	0.00332	-0.00012	0.00335	-0.00669	0.00337	0.00609	-0.00595	0.00000
Years	**1986**	**1987**	**1988**	**1990**	**1991**	**1992**	**1993**	**1994**
Pearson residuals	0.00141	0.00163	-0.00316	-0.00328	0.00478	0.00151	-0.00320	-0.00520
Deviance residuals	0.00140	0.00163	-0.00316	-0.00328	0.00478	0.00151	-0.00320	-0.00521
Years	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**
Pearson residuals	0.00000	0.01180	-0.00569	-0.00073	0.00199	-0.02091	0.01013	0.00764
Deviance residuals	0.00000	0.01177	-0.00570	-0.00073	0.00199	-0.02102	0.01010	0.00763
Years	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Pearson residuals	0.00473	-0.00388	0.01508	-0.00484	-0.00947	-0.00565	0.00702	-0.00167
Deviance residuals	0.00472	-0.00388	0.01502	-0.00485	-0.00949	-0.00565	0.00701	-0.00167

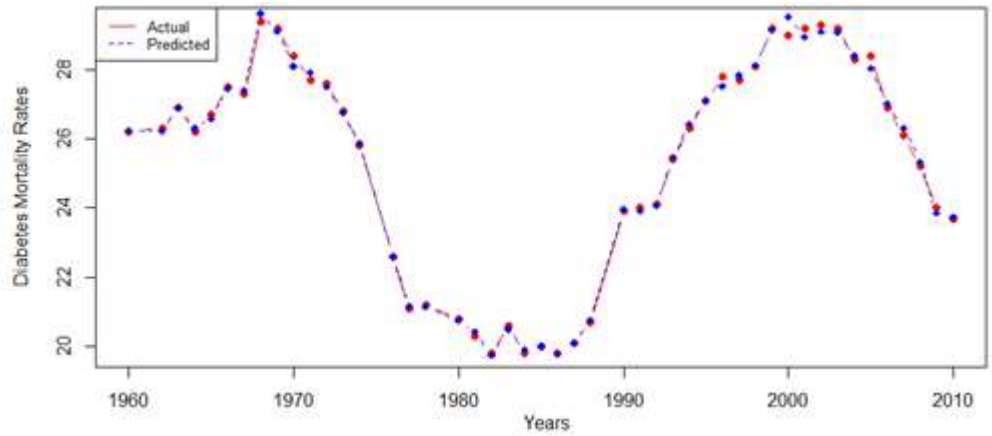

Figure 5: Line graph of the actual vs. predicted values of the *US* diabetes mortality rates between 1960 and 2010 by the new *GLM* with the Tweedie distribution

As seen from Figure 5, the new *GLM* with the Tweedie distribution best fits to the *US* diabetes mortality rates between 1960 and 2010.

5. Conclusions

In this study, especially the main interest is on the effects of changing the variance power parameter (p)($1<p<2$) in the variance of the Tweedie distributed response variable and the index of the power link function from -1 to +1 on the *AIC* goodness-of-fit test statistic and also Pearson chi-square and deviance statistics for the residuals and the dispersion parameter in the constituted *GLMs* for the *US* diabetes mortality data.

The smallest values for *AIC*, Pearson chi-square and the deviance statistics for the residuals and the dispersion parameter indicate that the best link function is "identity" with the variance power parameter "1.9" and the link function power "1" belonging to the Tweedie distribution in the *GLM* for the *US* diabetes mortality data.

Although the values for the Pearson chi-square and the deviance statistics for the residuals and the dispersion parameter seem very close to each other, the deviance statistics give larger values than the Pearson chi-square statistics in the *GLMs* for the *US* diabetes mortality data from the Tweedie distribution. In this case, for examining the residuals and the dispersion parameter, the deviance statistics tend to be preferable than the Pearson chi-square statistics.

Also in this study, it is emphasized that the model diagnostic plots based on the residuals, Cook's distance and leverage need to be investigated to determine the extreme observations that may cause some problems for parameter estimations, hypothesis tests, and statistical inferences in the *GLM* for the *US* diabetes mortality data from the Tweedie distribution. So determining and then omitting these extreme observations from the data set gives better results for the statistically insignificant independent variables to become statistically significant in the *GLM* for the *US* diabetes mortality data. Also alternative goodness-of-fit measures give better results and the ranges for the deviance residuals and the Pearson residuals become narrower in the *GLM* for the *US* diabetes mortality data from the Tweedie distribution after omitting these extreme observations.

It can also be concluded that over-dispersion is not determined in the data structure because the *IRLS* estimates of the dispersion parameter are not greater than 1 in the *GLMs* constituted for the *US* diabetes mortality data. Checking over-dispersion in the data structure is so important because it may cause a statistically significant independent variable to become statistically insignificant in the model.

6. Acknowledgements

NeslihanIyit is the major authorof this study and also Oznur Ozaltin's M.Sc. supervisor. This study is based on and an extended version of Oznur Ozaltin's M.Sc. Thesis titled "An Application of Generalized Estimating Equations (GEE) Approach on Organisation for Economic Cooperation and Development (OECD) Countries Mortality Data" supervised by Assist. Prof. Dr. NeslihanIyit submitted by Statistics Department, Graduate School of Natural Sciences, Selcuk University. An earlier version of this study is presented by the same authors at 3rd International Researchers, Statisticians and Young Statisticians Congress (IRSYSC-2017), 24-26May, 2017, Selcuk University, Konya, Turkey.

References

[1] A., Agresti, "Foundations of linear and generalized linear models", John Wiley & Sons, Cambridge, 2015.
[2] H., Akaike, "A new look at the statistical model identification", IEEE Transactions on Automatic Control, 19(6), 716-723, 1974.
[3] D. K., Blough, C. W., Madden, M. C., Hornbrook, "Modeling risk using generalized linear models", Journal of Health Economics, 18(2), 153-171, 1999.
[4] J. E., Brown, P. K., Dunn, "Comparisons of Tobit, linear, and Poisson-gamma regression models: An application of time use data", Sociological Methods &Research, 40(3), 511-535, 2011.
[5] A. C., Cameron, P. K., Trivedi, "Econometric models based on count data comparisons and applications of some estimators and tests", Journal of Applied Econometrics, 1(1), 29-53, 1986.

[6] S., Candy, "Modelling catch and effort data using generalized linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects", Ccamlr Science, 11, 59-80, 2004.
[7] L. A., Cifuentes, J., Vega, K., Kopfer, L. B., Lave, "Effect of the fine fraction of particulate matter versus the coarse mass and other pollutants on daily mortality in Santiago Chile", Journal of the Air & Waste Management Association, 50(8), 1287-1298, 2000.
[8] R. J., Cook, "Generalized linear model, " in Encyclopedia of Biostatistics, eds. P. Armitageand T. Colton, John Wiley&Sons, Chichester, UK, 1998.
[9] P., Diggle, "Analysis of longitudinal data", Oxford University Press, Oxford, UK, 2002.
[10] A. J., Dobson, A., Barnett, "An introduction to generalized linear models", CRC Press, Boca Raton, FL, 2008.
[11] P. K., Dunn, "Occurrence and quantity of precipitation can be modelled simultaneously", International Journal of Climatology, 24(10), 1231-1239, 2004.
[12] P. K., Dunn, G. K., Smyth, "Tweedie family densities: methods of evaluation", Paper presented at the Proceedings of the 16th International Workshop on Statistical Modelling, Odense, Denmark, 2001.
[13] P. K., Dunn, G. K., Smyth, "Series evaluation of Tweedie exponential dispersion model densities", Statistics and Computing, 15(4), 267-280, 2005.
[14] P. K., Dunn, G. K., Smyth, "Evaluation of Tweedie exponential dispersion model densities by Fourier inversion", Statistics and Computing, 18(1), 73-86, 2008.
[15] D., Firth, "Generalized linear models in statistical theory and modelling", Chapman&Hall, London, 1991.
[16] G. M., Fitzmaurice, N. M., Laird, J. H., Ware, "Applied longitudinal analysis", John Wiley & Sons, Hoboken, New Jersey, 2012.
[17] M., Franco, P., Ordunez, B., Caballero, J. A., Tapia Granados, M., Lazo, J. L., Bernal, E., Guallar, R. S., Cooper, "Impact of energy intake, physical activity, and population-wide weight loss on cardiovascular disease and diabetes mortality in Cuba, 1980–2005", American Journal of Epidemiology, 166(12), 1374-1380, 2007.
[18] J., Fuller, J., Elford, P., Goldblatt, A., Adelstein, "Diabetes mortality: new light on an underestimated public health problem", Diabetologia, 24(5), 336-341, 1983.
[19] G., Goodkin, "Mortality factors in diabetes. A 20 year mortality study", Journal of Occupational Medicine.: official publication of the Industrial Medical Association, 17(11), 716-721, 1975.
[20] G., Grover, A. S. A., Sabharwal, J., Mittal, "An application of gamma generalized linear model for estimation of survival function of diabetic nephropathy patients", International Journal of Statistics in Medical Research, 2(3), 209-219, 2013.
[21] J. M., Hilbe, A. P., Robinson, "Methods of statistical model estimation", CRC Press, Boca Raton, FL, 2013.
[22] T., Hothorn, B. S., Everitt, "A handbook of statistical analyses using *R*", CRC Press, Boca Raton, FL, 2014.
[23] N., Iyit, H., Yonar, A., Genc, "Generalized linear models for European Union countries energy data", ActaPhysicaPolonica A, 130(1), 397-400, 2016.
[24] B., Jorgensen, M. C., Paes De Souza, "Fitting Tweedie's compound Poisson model to insurance claims data", Scandinavian Actuarial Journal, 1994(1), 69-93, 1994.
[25] R., Kaas, "Compound Poisson distribution and GLMs—Tweedie's distribution", MATHEMATICS DAY, 3, 2005.
[26] G., Kaati, L. O., Bygren, S., Edvinsson, "Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period", European Journal of Human Genetics, 10(11), 682, 2002.
[27] P., Krause, D. P., Boyle, F., Base, "Comparison of different efficiency criteria for hydrological model assessment", Advances in Geosciences, 5, 89-97, 2005.
[28] D. R., Legates, G. J., McCabe, "Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation", Water Resources Research, 35(1), 233-]1, 1999.
[29] T. F., Liao, "Interpreting probability models: Logit, probit, and other generalized linear models", Sage Pub., London, UK, 1994.
[30] J. K., Lindsey, "Applying generalized linear models", Springer Science & Business Media, New York, 2000.
[31] D. N., Moriasi, J. G., Arnold, M. W., Van Liew, R. L., Bingner, R. D., Harmel, T. L., Veith, "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations", Transactions of the ASABE, 50(3), 885-900, 2007.
[32] P., McCullagh, J. A., Nelder, "Generalized Linear Models (2nd ed.)", Monograph on Statistics and Applied Probability, Chapman & Hall, Boca Raton, FL, 1989.
[33] J. A., Nelder, R. W. M., Wedderburn, "Generalized linear models", Journal of the Royal Statistical Society, Series A 135, 370-384, 1972.
[34] OECD, "Database Health Status: Mortality", Retrieved from http://stats.oecd.org/, [Accessed: Jan.15, 2016].
[35] A. E., Renshaw, S., Haberman, "Lee–Carter mortality forecasting with age-specific enhancement", Insurance: Mathematics and Economics, 33(2), 255-272, 2003.
[36] SAS, "Fitting Tweedie's Compound Poisson-Gamma Mixture Model by Using PROC HPGENSELECT", Retrieved from https://support.sas.com/rnd/app/stat/examples/tweedie/tweedie.pdf, [Accessed: March 03, 2017].
[37] A.M., Secrest, R.E., Washington, T.J., Orchard, "Mortality In Type 1 Diabetes", Diabetes in America, Chapter 35, 3rd edition, 35-1, 35-16, 2014.
[38] H., Shono, "Application of the Tweedie distribution to zero-catch data in CPUE analysis", Fisheries Research, 93(1-2), 154-162, 2008.
[39] G. K., Smyth, B., Jorgensen, "Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling", Astin Bulletin, 32(01), 143-157, 2002.
[40] U., Simsekli, A. T., Cemgil, B., Ermis, "Learning mixed divergences in coupled matrix and tensor factorization models", Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, 2015.
[41] M., Tweedie, "An index which distinguishes between some important exponential families", Paper presented at the Statistics, Applications and new directions, Proc.

Indian statistical institute golden Jubilee International conference, 1984.
[42] C. P., Wen, T. Y. D., Cheng, S. P., Tsai, H. L., Hsu, S. L., Wang, "Increased mortality risks of pre-diabetes (impaired fasting glucose) in Taiwan", Diabetes Care, 28(11), 2756-2761, 2005.
[43] D. F., Williamson, T. J., Thompson, M., Thun, D., Flanders, E., Pamuk, T., Byers, "Intentional weight loss and mortality among overweight individuals with diabetes", Diabetes Care, 23(10), 1499-1504, 2000.
[44] World HealthOrganization, Diabetes, 2017. Retrievedfromhttp://www.who.int/mediacentre/factsheets/fs312/en/ [Accessed: December 18, 2017].
[45] M. V., Wuthrich, "Claims reserving using Tweedie's compound Poisson model", ASTIN Bulletin: The Journal of the IAA, 33(2), 331-346, 2003.
[46] Y., Zhang, "Likelihood-based and bayesian methods for tweedie compound poisson linear mixed models", Statistics and Computing, 23(6), 743-757, 2013.
[47] M. Zambrano-Bigiarini, "Package 'hydroGOF' ", 2017. Retrievedfromhttps://cran.R version 3.4.0 version 3.4.0.org/web/packages/hydroGOF/hydroGOF.pdf, [Accessed: September 03, 2017].

Author Profile

Oznur Ozaltin received the B.S. degree in Actuarial Science Department from Hacettepe University in 2013 and M.Sc. degree in Statistics Department from Selcuk University in 2016, respectively. She has been a Ph.D. student at Hacettepe University since 2017. She has been working as a research assistant at Ataturk University, Turkey since 2016. Her research interests are R programming, linear models, generalized linear models, generalized estimating equations, heuristic methods and operations research.

NeslihanIyit received the B.S. degree in Statistics Department from Dokuz Eylul University in 2000, M.Sc. degree in Statistics Department from Selcuk University in 2003 and Ph.D. degree in Mathematics Department from Selcuk University 2008, respectively. She has been working as an assistant professor doctor at Selcuk University, Turkeysince 2011. Her research interests are applied statistics, statistical modelling techniques, categorical data analysis, linear mixed models, generalized linear models and generalized estimating equations.