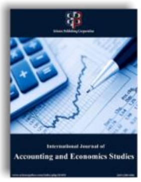
International Journal of Accounting and Economics Studies, 3 (2) (2015) 135-140

International Journal of Accounting and Economics Studies

Journal home page: www.sciencepubco.com/index.php/IJAES
doi: 10.14419/ijaes.v3i2.5317
Research Paper



Discrete choice models with response transformation : An application to beverage choice

Sunil K Sapra

Department of Economics and Statistics California State University 5151 State University Dr. Los Angeles, CA 90032 USA
E-mail: ssapra@calstatela.edu

Abstract

The paper studies various response transformation models for discrete choice and categorical data. These response transformation models are fitted to binary response data on beverage choice. Several models are compared, and the best model is selected using AICs and deviances. The transformations include extensions of the widely used Box-Cox transformation to Normality for continuous data to categorical data. The econometric techniques employed in the paper are widely applicable to the analysis of count, binary response, and duration types of data encountered in business and economics.

Keywords: *Box-Cox Transformation; Categorical Response Data; Logit Model; Generalized Linear Models; Link Functions*.

1. Introduction

Econometric models can often be improved by transforming response and predictor variables. A transformation replaces a variable by a function of that variable, which changes the shape of a distribution or relationship. Among the key reasons for transforming data are computational convenience, reducing skewness, achieving equal variance or homoscedasticity and linearity of the relationship between the response variable and the predictors as well as the ease of dealing with additive relationships than multiplicative relationships. The most common transformations in applied research are the reciprocal, logarithm, cube root, square root, and square.

Transformation of the response variable can lead to an improved fit as well as more reliable forecasts and inference. This has motivated empirical researchers to look for the best transformation models to fit their data. There is a vast literature on estimation of linear regression models with response transformations. Carrol and Ruppert (1988) provide a comprehensive account of transformations in regression models. Unfortunately, most empirical studies select the transformations in an ad hoc manner or use the Box-Cox transformation (Box & Cox (1964)), which is applicable to continuous response variables exclusively. For instance, the Box-Cox transformation cannot be used for several generalized linear models (GLMs), which allow the response variable to be categorical or count. There have been relatively few applications of response transformations in GLMs with categorical or count response variables. This paper focuses on discrete choice, categorical response GLMs with response transformation and studies their applications in business and economics. These transformations include various link functions suggested by Aranda-Ordaz (1981), Guerrero & Johnson (1982), Pregibon (1980), Koenker (2006), and Koenker & Yoon (2009) among others as alternatives to the popular Logit and Probit links.

This paper is organized as follows. Section 2 introduces GLMs with response transformations and an estimation algorithm for these models. Section 3 presents empirical applications of transformation models for categorical response GLMs. Various GLM response transformation models are compared to select the best model. Section 4 concludes and presents directions for future research.

2. Generalized linear models with response transformations

Generalized linear models consist of a random component, an additive component, and a link function relating these two components. The response y, the random component, is assumed to have a density in the exponential family

$$f_Y(y,\theta)=\exp\left\{\frac{y\theta-b(\theta)}{a(\varphi)}+c(y,\varphi)\right\}, \qquad (1)$$

where θ is called the natural parameter and ϕ is the scale parameter. Common distributions, such as normal, binomial, and poisson are all in this family. In generalized linear models (GLMs), the mean $\mu = E\,(y\,|\,x1,\ x2\ldots\ xp)$ is linked to the linear predictor $\sum xij\beta j$ through the link function

$$\eta = g(\mu)=\sum x_{ij}\beta_j . \qquad (2)$$

The regression transformation model takes the form

$$T(Y)=X\beta+U, \qquad (3)$$

where *T* is a strictly increasing function, Y is an observed dependent variable, X is an observed random vector, β is a vector of unknown parameters conformable with *X* and *U* is an unobserved random variable, which is independent of X.

The most popular choice for *T* if *T* is continuous is the Box-Cox transformation

$$T_{\lambda}(Y) = (Y^{\lambda} - 1)/\lambda \text{ for } \lambda \neq 0,$$

$$= \ln Y \quad \text{for } \lambda = 0. \tag{4}$$

The log-likelihood function is

$$l(\beta,\lambda,\sigma) = -n/2\ln\sigma - 1/2\sigma^2$$
$$\sum_{i=1}^{n}(T_{\lambda}(Y) - \sum_{j} x_{ij}\beta_j)^2 + \sum_{i=1}^{n}\ln\left|\partial T_{\lambda}(Y)/\partial Y\right| \tag{5}$$

We consider generalized linear models with categorical response, response transformation *T(Y)*, and parametric link function

$$\eta = g(E(T(Y)) = \alpha + \sum_{j=1}^{p}\beta_j x_j. \tag{6}$$

A Modified Iteratively Reweighted Least Squares Method for GLM with Response Transformation

The log-likelihood function is

$$\log L = \sum_{i=1}^{n} w_i\left[\frac{T(y_i)\theta_i - b(\theta_i)}{\varphi}\right] + c(y_i,\varphi)$$
$$+\sum_{i=1}^{n}\ln\left|\partial T(Y)/\partial Y\right|, \tag{7}$$

where

$$w_i = a_i(\varphi)/\varphi.$$

The log-likelihood function in (7) can be maximized via a modification of iteratively reweighted least squares (IRWLS) method of McCullagh and Nelder (1989) as follows.

Step 1: Set initial estimates $\hat{\eta} = \hat{\eta}_0$ and $\hat{\mu} = \hat{\mu}_0$.

Step 2: Form the adjusted dependent variable

$$z_0 = \hat{\eta}_0 + (T(y) - \hat{\mu}_0)d\eta/d\mu\big|_{\hat{\eta}_0}.$$

Step 3: Form the weights

$$w_0 = [(d\eta/d\mu)^2\big|_{\hat{\eta}_0} Var(\hat{\mu}_0)]^{-1}.$$

Step 4: Re-estimate β to get $\hat{\eta}_1$.

Step 5: Iterate steps 2 through 4 until convergence.

3. An empirical application

Parametric Transformations of Response with Binary Response Data

Following Koenker (2006), Koenker & Yoon (2009) and Zeileis et al. (2014), we employ several GLMs with symmetric and asymmetric parametric link functions, which are in effect various response transformations for the binary response data. These links include Guerrero-Johnson (1982) family, symmetric and asymmetric Aranda-Ordaz (1981) transformation, Pregibon (1980) two parameter family, Rocke (1993) family of links based on a linear transformation of the Beta distribution, the folded exponential family (Piepho (2003)), the t-alpha family (Doebler et al., 2011), the Gosset family of links, the Cauchy link, and the arcsine transformation. R-package glmx of Zeileis et al. (2014) was used for all of the computations. The package implements extended techniques for generalized linear models (GLMs), especially for binary responses, including parametric links and heteroskedastic latent variables. A key advantage of these non-standard links is that we are able to handle transformations to both symmetry and homoscedasticity in categorical response models.

The links employed for binary data are as follows. Each of these links involves a transformation of the response variable to achieve symmetry and homoscedasticity (equal variances).

Symmetric Aranda-Ordaz Transformation:

$$y = \frac{2x^{\varphi} - (1-x)^{\varphi}}{\varphi x^{\varphi} + (1-x)^{1-\varphi}}$$

Asymmetric Aranda-Ordaz Transformation:

$$y = \ln([(1-x)^{-\varphi} - 1]/\varphi)$$

Pregibon 2 parameter Transformation:

$$y = \frac{x^{a-b} - 1}{a-b} - \frac{(1-x)^{a+b} - 1}{a+b}$$

Guerrero-Johnson Transformation:

$$y = 1/\varphi([\frac{x}{1-x}]^{\varphi} - 1)$$

The t-alpha Transformation:

$$y = \alpha\log x - (2-\alpha)\log(1-x)$$

Angular (Arcsine) Transformation:

$$y = \arcsin(\sqrt{x})$$

An empirical application of generalized linear models (GLMs) to brand choice: GLMs Applied to Coke Data

The dataset is from ERIM public data base, James M. Kilts Center, University of Chicago Booth School of Business and consists of data on 1140 individuals who purchased Coke or Pepsi. It is also available in Hill et al. (2011).

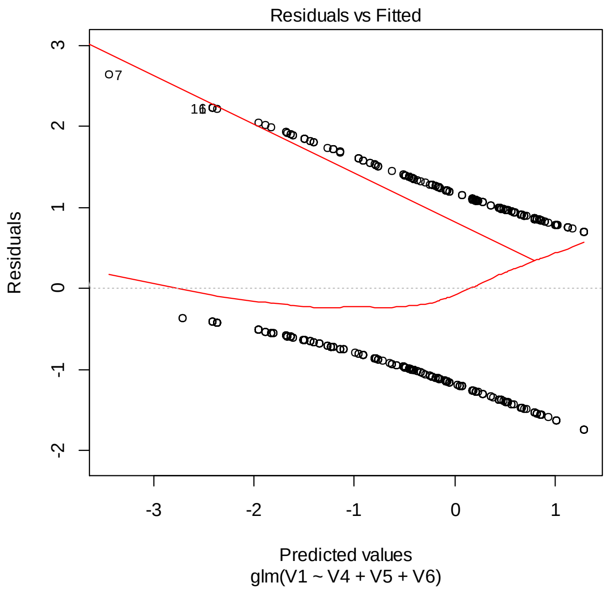
Data Description

The dependent variable COKE is defined as follows:
COKE =1 if Coke is chosen,
= 0 if Pepsi is chosen
PR_PEPSI = Price of 2 liter bottle of Pepsi
PR_COKE = Price of 2 liter bottle of Coke
DISP_PEPSI = 1 if Pepsi is displayed at time of purchase, otherwise = 0
DISP_COKE = 1 if Coke is displayed at time of purchase, otherwise = 0
PRATIO = Price of Coke relative to price of Pepsi

Table 1: Summary Statistics

Variable	Obs.	Mean	Std. Dev.	Min	Max
COKE	1140	0.4473684	0.4974404	0	1
PR_PEPSI	1140	1.202719	0.3007257	.68	1.79
PR_COKE	1140	1.190088	0.2999157	.68	1.79
DISP_PEPSI	1140	0.3640351	0.4813697	0	1
DISP_COKE	1140	0.3789474	0.4853379	0	1
PRATIO	1140	1.027249	0.286608	0.497207	2.324675

Source: ERIM public data base, James M. Kilts Center, University of Chicago Booth School of Business



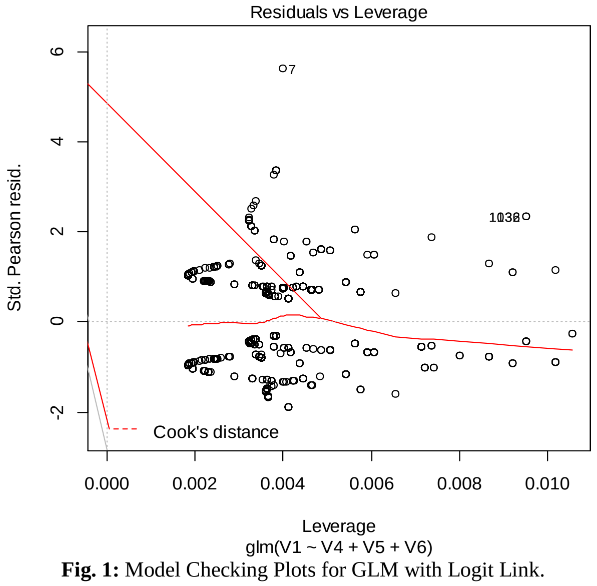
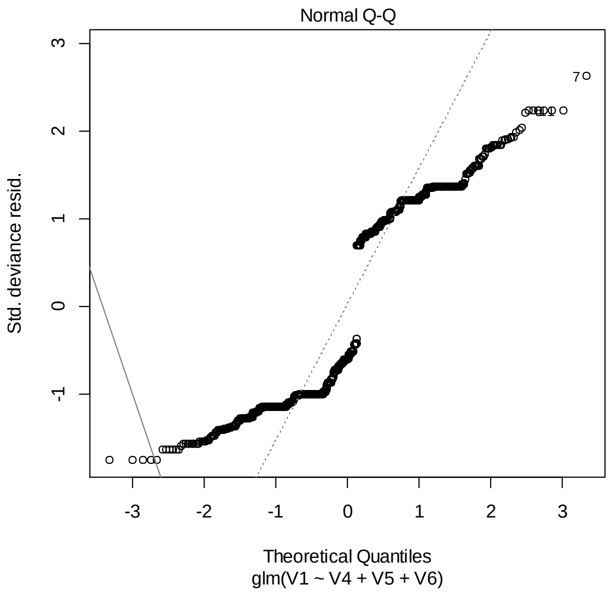


Fig. 1: Model Checking Plots for GLM with Logit Link.



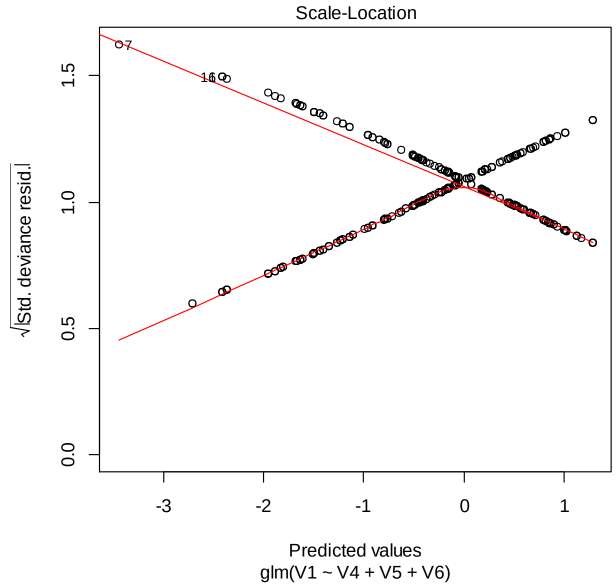


Table 2: GLM with Logit Link: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	1.9230	0.3258	5.902	3.59x10-09***
DISP_PEPSI	-0.7310	0.1678	-4.356	1.33x10-05***
DISP_COKE	0.3516	0.1585	2.218	0.0266*
PRATIO	-1.9957	0.3146	-6.344	2.23x10-10***

Signif Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1567.7 on 1139 degrees of freedom
Residual deviance: 1418.9 on 1136 degrees of freedom
AIC: 1426.9
Number of Fisher Scoring iterations: 3

Table 3: GLM with Logit Link and Variance Stabilizing Transform: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	3.6603	1.4036	2.608	0.00911**
DISP_PEPSI	-1.9654	1.2192	-1.612	0.10694
DISP_COKE	0.6018	0.3596	1.673	0.09429
PRATIO	-3.9500	1.5408	-2.564	0.01036*

Signif Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-likelihood: -701.5 on 7 degrees of freedom
LR test for homoskedasticity: 15.97 on 3 degrees of freedom, p-value: 0.001152
Dispersion: 1
Number of iterations: 10
AIC: 1416.926

Table 4: GLM with Guerrero-Johnson Link, Phi = 0.1: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	1.7406	0.3120	5.578	2.43x10-08***
DISP_PEPSI	-0.6966	0.1591	-4.378	1.20x10-05***
DISP_COKE	0.3352	0.1563	2.144	0.032*
PRATIO	-1.7941	0.2944	-6.094	1.10x10-09***

Signif Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1567.7 on 1139 degrees of freedom
Residual deviance: 1424.4 on 1136 degrees of freedom
AIC: 1432.4
Number of Fisher Scoring iterations: 5

Table 5: Model: GLM with Guerrero-Johnson Link, Phi = 0.1 and Variance Stabilizing Transformation: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	4.1171	1.5877	2.593	0.00951**
DISP_PEPSI	-1.9951	1.1598	-1.720	0.08539.
DISP_COKE	0.6623	0.3801	1.742	0.08147.
PRATIO	-4.4401	1.7335	-2.561	0.01043*

Signif Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -701.3 on 7 degrees of freedom
LR test for homoskedasticity: 21.78 on 3 degrees of freedom, p-value: 7.258x10-05
Dispersion: 1
Number of iterations: 11
AIC = 1416.62
Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6841	-1.0433	-0.6305	1.1062	2.8954

Table 6: GLM with Aranda-Ordaz 1 Link, Phi = 0.6: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	1.5099	0.2889	5.227	1.73x10-7***
DISP_PEPSI	-0.7185	0.1557	-4.614	3.95x10-6***
DISP_COKE	0.3629	0.1497	2.425	0.0153*
PRATIO	-1.5580	0.2724	-5.720	1.07x10-8***

Signif Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1567.7 on 1139 degrees of freedom
Residual deviance: 1426.1 on 1136 degrees of freedom
AIC: 1434.1
Number of Fisher Scoring iterations: 17

Table 7: GLM with Aranda-Ordaz 1 Link, Phi = 0.6 and Variance Stabilizing Transformation: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	3.4910	1.2170	2.868	0.00412**
DISP_PEPSI	-2.3074	1.4424	-1.600	0.10966
DISP_COKE	0.6238	0.3584	1.740	0.08178.
PRATIO	-3.7643	1.3357	-2.818	0.00483**

Signif codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -701.7 on 7 degrees of freedom
LR test for homoskedasticity: 22.7 on 3 degrees of freedom, p-value: 4.653x10-05
Dispersion: 1
Number of iterations: 10
AIC: 1417.394
Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8027	-0.9945	-0.5923	1.0916	2.5884

Table 8: GLM with Aranda-Ordaz 2 Link, Phi = 0.6: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	1.6070	0.2888	5.564	2.63x10-08***
DISP_PEPSI	-0.6620	0.1521	-4.354	1.34x10-05***
DISP_COKE	0.3177	0.1388	2.290	0.022*
PRATIO	-1.8353	0.2830	-6.484	8.92x10-11***

Signif codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1567.7 on 1139 degrees of freedom
Residual deviance: 1415.9 on 1136 degrees of freedom
AIC: 1423.9
Number of Fisher Scoring iterations: 4
Deviance residuals:

Min	1Q	Median	3Q	Max
-2.0473	-0.9311	-0.6784	1.1005	1.9398

Coefficients (binomial model with ao2 link):

Table 9: GLM with Aranda-Ordaz 2 Link, Phi = 0.6 and Variance Stabilization Transform: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	2.7920	1.0951	2.549	0.0108*
DISP_PEPSI	-1.9579	1.3123	-1.492	0.1357
DISP_COKE	0.5004	0.2673	1.872	0.0612.
PRATIO	-3.2311	1.2670	-2.550	0.0108*

Signif codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -701.6 on 7 degrees of freedom
LR test for homoskedasticity: 12.7 on 3 degrees of freedom, p-value: 0.005331
Dispersion: 1
Number of iterations: 9
AIC: 1417.183

Table 10: GLM With Gosset (2) Link: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	1.5764	0.2645	5.960	2.53x10-09***
DISP_PEPSI	-0.5540	0.1305	-4.244	2.20x10-05***
DISP_COKE	0.2671	0.1214	2.199	0.0278*
PRATIO	-1.6479	0.2604	-6.328	2.48x10-10***

Signif codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1567.7 on 1139 degrees of freedom
Residual deviance: 1413.6 on 1136 degrees of freedom
AIC: 1421.6
Number of Fisher Scoring iterations: 4

Table 11: Model: GLM with Gosset (2) Link and Variance Stabilization Transform: Coke Data

3	Estimate	Std. Error	t-statistic	p-value
Intercept	2.7241	1.1586	2.351	0.0187*
DISP_PEPSI	-1.1585	0.7268	-1.594	0.1110
DISP_COKE	0.4146	0.2592	1.600	0.1096
PRATIO	-2.9426	1.2705	-2.316	0.0206*

Signif codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -701.3 on 7 degrees of freedom
LR test for homoskedasticity: 11.04 on 3 degrees of freedom, p-value: 0.01152
Dispersion: 1
Number of iterations: 9
AIC: 1416.552

Table 12: GLM with Negative Binomial Logit Mixture Link, Parameter = 2: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	1.5272	0.2794	5.465	4.63x10-08***
DISP_PEPSI	-0.6449	0.1482	-4.352	1.35x10-05***
DISP_COKE	0.3089	0.1338	2.308	0.021*
PRATIO	-1.7933	0.2750	-6.522	6.95x10-11***

Signif codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1567.7 on 1139 degrees of freedom
Residual deviance: 1415.1 on 1136 degrees of freedom
AIC: 1423.1
Number of Fisher Scoring iterations: 4

Table 13: GLM with Negative Binomial Logit Mixture Link, Parameter = 2 and Variance Stablizing Transform: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	2.5864	1.0237	2.526	0.0115*
DISP_PEPSI	-1.9727	1.3570	-1.454	0.1460
DISP_COKE	0.4734	0.2465	1.921	0.0548.
PRATIO	-3.0556	1.2021	-2.542	0.0110*

Signif codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -701.6 on 7 degrees of freedom
LR test for homoskedasticity: 11.88 on 3 degrees of freedom, p-value: 0.007795
Dispersion: 1
Number of iterations: 9
AIC: 1417.258

Table 14: Model: GLM With Pregibon Link (0.5, 0.7): Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	1.6736	0.2452	6.824	8.83x10-12***
DISP_PEPSI	-0.5119	0.1147	-4.462	8.10x10-06***
DISP_COKE	0.2060	0.1293	1.594	0.111
PRATIO	-1.3319	0.2073	-6.424	1.32x10-10***

Signif codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1567.7 on 1139 degrees of freedom
Residual deviance: 1682.3 on 1136 degrees of freedom
AIC: 1690.3
Number of Fisher Scoring iterations: 9

Table 15: Model: GLM with Pregibon Link (0.5, 0.7) and Variance Stabilizing Transform: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	6.1428	2.2028	2.789	0.00529**
DISP_PEPSI	-1.9940	0.9587	-2.080	0.03753*
DISP_COKE	0.7880	0.5773	1.365	0.17223
PRATIO	-5.7022	2.1319	-2.675	0.00748**

Signif codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -701 on 7 degrees of freedom
LR test for homoskedasticity: 280.4 on 3 degrees of freedom, p-value: < 2.2e-16
Dispersion: 1
Number of iteration: 12
AIC: 1415.921

Table 16: GLM with Angular Link: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	1.13829	0.07008	16.244	<2x10-16***
DISP_PEPSI	-0.17896	0.03810	-4.698	2.63x10-06***
DISP_COKE	0.09162	0.03680	2.490	0.0128*
PRATIO	-0.36347	0.06565	-5.536	3.09x10-08***

Signif codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1567.7 on 1139 degrees of freedom
Residual deviance: 1428.3 on 1136 degrees of freedom
AIC: 1436.3
Number of Fisher Scoring iterations: 25

Table 17: GLM with Rocke (0.5, 1) Link: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	0.59767	0.09871	6.055	1.41x10-09***
DISP_PEPSI	-0.26593	0.05819	-4.570	4.87x10-06***
DISP_COKE	0.12672	0.04927	2.572	0.0101*
PRATIO	-0.62573	0.09745	-6.421	1.36x10-10***

Signif codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1567.7 on 1139 degrees of freedom
Residual deviance: 1417.3 on 1136 degrees of freedom
AIC: 1425.3
Number of Fisher Scoring iterations: 6

Table 18: GLM with Rocke (0.5, 1) Link and Variance Stabilizing Transform: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	0.8642	0.2666	3.242	0.00119**
DISP_PEPSI	-1.1515	1.0013	-1.150	0.25012
DISP_COKE	0.1937	0.1161	1.668	0.09533
PRATIO	-0.9311	0.2970	-3.135	0.00172**

Signif codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -702.5 on 7 degrees of freedom
LR test for homoskedasticity: 12.22 on 3 degrees of freedom, p-value: 0.006668
Dispersion: 1
Number of iterations: 10
AIC: 1419.065

Table 19: GLM with Complementary Log-Log Link: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	1.1192	0.2321	4.823	1.42x10-06***
DISP_PEPSI	-0.5607	0.1293	-4.335	1.46x10-05***
DISP_COKE	0.2624	0.1095	2.396	0.0166*
PRATIO	-1.5682	0.2335	-6.715	1.88x10-11***

Signif codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1567.7 on 1139 degrees of freedom
Residual deviance: 1411.7 on 1136 degrees of freedom
AIC: 1419.7
Number of Fisher Scoring iterations: 5

Table 20: GLM with Complementary Log-Log Link and Variance Stabilizing Transform: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	1.6343	0.7030	2.325	0.0201*
DISP_PEPSI	-2.2657	1.9181	-1.181	0.2375
DISP_COKE	0.3284	0.1577	2.083	0.0373*
PRATIO	-2.2069	0.9023	-2.446	0.0145*

Signif codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -701.8 on 7 degrees of freedom
LR test for homoskedasticity: 8.001 on 3 degrees of freedom, p-value: 0.046
Dispersion: 1
Number of iterations: 9
AIC: 1417.676

Table 21: GLM with Complementary Talpha (1.5) Link: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	1.3987	0.3256	4.295	1.74x10-05***
DISP_PEPSI	-0.7864	0.1825	-4.309	1.64x10-05***
DISP_COKE	0.3664	0.1525	2.403	0.0163*
PRATIO	-2.2139	0.3295	-6.719	1.83x10-11***

Signif codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1567.7 on 1139 degrees of freedom
Residual deviance: 1410.8 on 1136 degrees of freedom
AIC: 1418.8
Number of Fisher Scoring iterations: 5

Table 22: GLM with Complementary Talpha (1.5) Link with Variance Stabilizing Transform: Coke Data

VARIABLE	Estimate	Std. Error	t-statistic	p-value
Intercept	2.0707	0.9774	2.119	0.0341*
DISP_PEPSI	-3.4195	3.0697	-1.114	0.2653
DISP_COKE	0.4289	0.2094	2.048	0.0406*
PRATIO	-3.0647	1.3186	-2.324	0.0201*

Signif codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -701.8 on 7 degrees of freedom
LR test for homoskedasticity: 7.23 on 3 degrees of freedom, p-value: 0.06493
Dispersion: 1
Number of iterations: 10
AIC: 1417.533

Table 23: Models and the Aics

MODEL	AIC
Logit	1426.9
Logit with VST	1416.926
Guerrero-Johnson	1432.4
Guerrero-Johnson with VST	1416.62
Aranda-Ordaz 1	1434.1
Aranda-Ordaz 1 with VST	1417.394
Gossett-2	1421.6
Gosssett-2-VST	1416.552
Negative Binomial-Logit Mixture	1423.1
Negative Binomial-Logit Mixture with VST	1417.258
Pregibon (0.5,0.7)	1690.3
Pregibon (0.5,0.7) with VST	1415.921
Angular	1436.3
Rocke (0.5,1)	1425.3
Rocke (0.5,1) with VST	1419.065
Complementary Log Log	1419.7
Complementary Log Log with VST	1417.676
Complementary Talpha (1.5)	1418.8
Complementary Talpha (1.5) with VST	1417.533

4. Main results

Model checking plots for the Logit model in Fig. 1 suggest a nonlinear relationship between the predicted values and residuals, indicating a lack of fit. Furthermore, these plots also display nonnormality and unequal variances. Likelihood ratio test for heteroscedasticity confirmed heteroscedasticity for all the links necessitating the use of variance stabilizing to transform. This calls for a response transformation or a change in the link function. Accordingly, several discrete choice models with alternative links were estimated. Estimation and inference results for these links are presented in Tables 2 through 22. Table 23 presents the AICs for the various link functions used. Unsurprisingly, several links outperform the Logit, especially with the application of variance stabilizing transform as reflected in the lower AICs for these models relative to the Logit model. Without the variance-stabilizing transform, the Logit, Aranda-Ordaz-1, and Pregibon (0.5, 0.7) links provide the poorest models and Complementary Log, and Complementary Talpha (1.5) links produce the best models. However, with variance stabilizing to transform, Pregibon (0.5, 0.7) links have the lowest AIC and provide the best model among all links. Interestingly, Complementary Talpha (1.5) link provides a decent model with or without variance, stabilizing to transform and the difference between the AICs is very small.

All of the coefficients had expected signs across all of the models studied. The predictors DISP_PEPSI and PRATIO are highly statistically significant across all links without variance stabilizing to transform applied to the response variable. However, when the variance stabilizing transformation is applied to the response variable, these predictors become less significant. With respect to prediction performance, there was little difference between the logit link and other links. Following Hill et al. (2011), we employed the following prediction rule: predict that the consumer will choose coke if the predicted value $COK\hat{E}$ is greater than 0.5. For all of the links, of the 510 consumers who chose coke over Pepsi, 370 were correctly predicted without a variance stabilizing to transform, and 349 were correctly predicted if a variance stabilizing to transform was applied.

5. Conclusions

The paper has studied various response transformation models for categorical data. Several discrete choices GLMs with various response transformations and link functions were compared for selecting the best model. For asymmetric categorical data, several models with alternative link functions, notably Complementary Log Log and Complementary Talpha links perform better than the Logit model. The GLM response transformation models for categorical response data studied in the paper have wide applications in business and economics and can provide potential improvements in the quality of inference, and forecasts.

The GLMs studied in the paper can be extended to include transformations of predictors in the spirit of Box-Tidwell transformation. Other possible extensions include nonparametric smooth transformations for the response variable as well as predictors while assuming a known link function. Parametric models of response transformations can be restrictive and developing semiparametric techniques for GLMs and generalized additive models (GAMs) with response and predictor transformations with an unknown link function may be useful. Another idea is to extend a semi-parametric estimator for the Box-Cox model (Shin (2008)) to GLMs and GAMs to achieve root-n consistency of the estimators notwithstanding the large dimension of the data.

References

[1] Aranda-Ordaz F (1981) On Two Families of Transformations to Additivity for Binary Response Data. Biometrika. 68, 357–363. http://dx.doi.org/10.1093/biomet/68.2.357.

[2] Box, GEP. & Cox DR (1964) an Analysis of Transformations. Journal of the Royal Statistical Society, Series B, 26, 211-252

[3] Carrol, RJ & Ruppert D (1988) Transformations and Weighting in Regression, Chapman and Hall: New York http://dx.doi.org/10.1007/978-1-4899-2873-3.

[4] Doebler P, Holling H, Boehning D (2012) A Mixed Model Approach to Meta-Analysis of Diagnostic Studies with Binary Test Outcome. Psychological Methods, 17(3), 418–436. http://dx.doi.org/10.1037/a0028091.

[5] Guerrero V, Johnson R (1982) Use of the Box-Cox Transformation with Binary Response Models. Biometrika, 69, 309–314. http://dx.doi.org/10.1093/biomet/69.2.309.

[6] Hill, RC, WE Griffiths, and GG Lim (2011) Principles of Econometrics, New York: Wiley.

[7] Koenker R (2006) Parametric Links for Binary Response. R News, 6, 32–34.

[8] Koenker R & Yoon J (2009) Parametric Links for Binary Choice Models: A Fisherian-Bayesian Colloquy. Journal of Econometrics, 152, 120–130. http://dx.doi.org/10.1016/j.jeconom.2009.01.009.

[9] McCullagh, P & Nelder RA (1989) Generalized Linear Models, Chapman and Hall: New York. http://dx.doi.org/10.1007/978-1-4899-3242-6.

[10] Piepho H (2003) the Folded Exponential Transformation for Proportions. Journal of the Royal Statistical Society D, 52, 575–589. http://dx.doi.org/10.1046/j.0039-0526.2003.00509.x.

[11] Pregibon D (1980) Goodness of Link Tests for Generalized Linear Models. Journal of the Royal Statistical Society C, 29, 15–23. http://dx.doi.org/10.2307/2346405.

[12] Rocke DM (1993) On the Beta Transformation Family. Technometrics, 35, 73–81. http://dx.doi.org/10.1080/00401706.1993.10484995.

[13] Shin, Y (2008) Semiparametric estimation of the Box-Cox transformation model, The Econometrics Journal, 11, 517-537. http://dx.doi.org/10.1111/j.1368-423X.2008.00255.x.

[14] Zeileis, A, Koenker R & Doebler P (2014) R-package glmx.