Paper 1658-2014

Healthcare Services Data Distribution, Transformation and Model Fitting

Dawit Mulugeta, Jason Greenfield, Tison Bolen and Lisa Conley
Advanced Analytics, Pharmaceutical Distribution Pricing Strategy and Analytics, Cardinal Health, Dublin, Ohio 43017

ABSTRACT

Healthcare services data on products and services come in different shapes and forms. Data cleaning, characterization, massaging and transformation are essential precursors to any statistical model building efforts. In addition, data size, quality and distribution influence model selection, model life cycle and the ease with which business insights are extracted from data. Analysts need to examine data characteristics and determine the right data transformation and methods of analysis for valid interpretation of results. In this presentation we will demonstrate the common data distribution types for a typical healthcare services industry such as Cardinal Health and their salient features. In addition we will utilize Base and Stat SAS® for data transformation of both the response (Y) and the explanatory (X) variables in four combinations [RR (Y and X as row data), TR (only Y transformed), RT (only X transformed) and TT (Y and X transformed)] and the practical significance of interpreting linear, logistic, and completely randomized design model results using the original and the transformed data values for decision making processes. The reality of dealing with diverse forms of data, the ramification of data transformation, and the challenge of interpreting model results of transformed data are discussed. Our analysis showed that the magnitude of data variability is an overriding factor to the success of data transformation and subsequent tasks of model building and interpretation of model parameters. Although data transformation provided some benefits, it complicated analysis and subsequent interpretation of model results.

INTRODUCTION

Cardinal Health, a healthcare services supply chain company, collects diverse forms, large number and large volume of structured and unstructured data on a daily, weekly and monthly basis. These data may include information on product transactions, market data, and other metrics. Often analysts use these data to generate insights to make healthcare more affordable.

In order to extract actionable insights, analyst put considerable efforts to understand and characterize granular or aggregated data. In this crucial step of analysis, important considerations may include: 1) understanding data formatting and whether variables are numeric, character, or alpha-numeric; 2) determining if there are any egregious elements in the data including missing, truncated, outliers, undefined and ambiguous values; 3) assessing if the variables are discrete

(nominal or ordinal) or continuous (interval or ratio); 4) checking if the variables are symmetric or asymmetric, and if there is asymmetry, which directions are the outliers heavily concentrated; 5) ascertaining if there exist an upper or a lower boundary; 6) knowing the data distribution types; and 7) deciding if there is a need to have some form of data transformation prior to further analysis including model building tasks.

While most of the listed considerations are straight forward, the determination of data distribution types and the need for data transformation and subsequent steps of result interpretation require extra efforts. This is attributable to: 1) the presence of many distribution types including normal, Poisson, binomial, negative binomial, exponential, geometric, gamma, beta, multinomial and several others. Each of these distributions has its respective peculiar features and assumptions. There are also some overlapping and subtle similarities among some of the distribution types which make the quick identification and characterization a difficult task; 2) the presence of many data transformation types including logarithmic, square root, arcsine (angular), reciprocal, Box-Cox, power and others.

The intents of transformation are to stabilize the relationship between the variance and the mean, to convert data distribution to normal or to improve symmetry, to maximize correlations between response and explanatory variables, to reduce the influence of outliers, to improve linearity in regression, to minimize interaction effects and to reduce skewness and kurtosis [2]. Often the dilemma lies in the dependency of choices of transformation methods on the data distribution types, and also the interpretation of analysis results of the transformed data.

Although transaction and other forms of healthcare services data may assume many data distribution types, the ones that we encounter quite often are the binomial, the Poisson, and the normal distributions.

The binomial distribution is a discrete probability distribution [1, 2] of the number of successes in a sequence of n explanatory trials, and is described by the following function:

$$\mathbf{f(k,n,p) = Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}} \quad \textbf{(1)}$$

Where k = total success, n = total trials, p = event success, and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

The distribution of percentages with values from 0 to 100% follows the binomial distribution. Var7 is a good example (Table 3). Similarly many healthcare services data such as frequency of purchases and share of units among product substitutes lay in this category.

The Poisson distribution is a discrete distribution for count data [2] and takes on the values X = 0, 1, 2, 3, and so on. It is often used as a model for the number of events (such as the number units purchased or the number of items ordered, etc. per unit of time. It is determined by one parameter (λ) that serves as the mean and the variance of data distribution. The distribution function is described by the following formula:

$$f(k,\lambda) = Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (2)$$

Where λ = the mean and the variance, e = the base of the natural logarithm (e = 2.718...), and k! = the factorial of k (the number of occurrences of an event, the probability of which is given by the function).

Normal distributions are symmetric and have bell-shaped density curves with a single peak [2]. It has two parameters, the mean (also the median and the mode), where the peak of the density occurs, and the standard deviation, which indicates the spread of the bell curve. It has the following probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3)$$

Where μ = the mean, and σ = the standard deviation, π = 3.14159…, and e = 2.718….

Normal distributions have many convenient properties. Often random variables with unknown distributions are often assumed to be normal, and this is so possible due to the central limit theorem [2]. Many common attributes drawn from large samples such as average sales of a firm, test scores of students, etc., follow the normal distribution, with few values at the high and low ends and many in the middle.

The main objective of this study is to determine how some popular model building steps and model parameters of a diverse but typical form of healthcare services data are affected by diverse form of data transformations. In this paper we first provided a typical healthcare services data consisting of several variables that represent different distribution types. We then employed various data transformation methods and compared pre and post transformation values of various data metrics. Lastly we built linear and logistic models and conducted analysis of experimental design both on original and transformed data to measure the impact of data transformation on various model parameters and result interpretation.

METHODS

We created an artificial and hypothetical B2B dataset for a healthcare services firm. A typical example is Cardinal Health, a leading firm in the industry that serves thousands of customers (pharmacies) all over the country. The data consist of eight variables. These include: 1) Var1 is hypothetical identification of an entity; 2) Var2 with the following four distinct labels: N, Y, M and O; 3) Var3; 4) Var4; 5) Var5; 6) Var6; 7) Var7, and 8) Var8.

The variables displayed represent the typical data features encountered on a daily basis. We first described the distribution pattern of each variable in line with the salient features of the corresponding theoretical distribution reported in literature. We then used three kinds of transformations (logarithmic, square root and arc-sine) to transform each variable whenever

possible. Following these, we computed various statistical measures of central tendency, dispersion and measures of normality on both the original and the transformed values for all variables. We then developed linear [3] and logistic [4] models to examine model parameters, R-square, mean-square, significance tests, graphical residual analysis and other measures for the raw (original) and the transformed data of all variables. For the linear model we used the following functional form:

$$\mathbf{Y = \beta_o + \beta X + \varepsilon} \quad \mathbf{(4)}$$

Where Y is Var5 (raw or transformed); X is the raw or transformed values of Var3, Var4, Var6, and Var7 each regressed individually; β_o and β are parameter estimates of the intercept and the slope of the regression equation, respectively, and ε is the error term.

Four sets of simple linear regressions were run (Table 1). In the first, raw (R) data of the response (Y) and the explanatory (X) variables were used. This is designated as (Row-Row, RR). In the second, only Y was transformed (T) (Transformed-Row, TR). In the third, only X was transformed (RT), and in the fourth both Y and X were transformed (TT). These provided a total of 37 regression runs with their respective parameters.

Table 1 Combination of Transformations used to run Simple Linear Regression of the Response Variable, Var5 (Y), and Four Explanatory Variables (X's).

Transformation			Explanatory Variables (X)				
Status							
Response (Y)	Explanatory (X)	Type	Var3	Var4	Var6	Var7	Total
R	R	-	1	1	1	1	4
T	R	Log	1	1	1	1	4
		Sqrt	1	1	1	1	4
R	T	Log	1	1	1	1	4
		Sqrt	1	1	1	1	4
		Arcsine	-	-	-	1	1
T	T	Log	2	2	2	2	8
		Sqrt	2	2	2	2	8
Total							37

For the logistic regression we used the following functional form as described in Allison [4]

$$\mathbf{\log (P/(1-P)) = \beta_o + \beta X} \quad \mathbf{(5)}$$

Where P and P/(1-P)) are the probability and the odds of Var8, respectively; X is the raw or transformed values of Var3, Var4, Var5, Var6, and Var7 each taken individually; β_o and β are parameter estimates of the logistic equation. The coefficient of determination (R^2) for the logistic regression was computed following the method suggested by Allison [4]. Two sets of simple logistic regressions were run (Table 2). In the first, raw data of the response variable (Y) and the

explanatory variables (X) were used. In the second, only each of the explanatory variables (X) was transformed. A total of 16 logistic regressions and their respective parameters were generated.

Table 2 Combination of Transformations used to run Simple Logistic Regression of the Response Variable, Var8 (Y), and Five Explanatory Variables (X's).

Transformation			Explanatory Variables (X)					
Status								
Response (Y)	Explanatory (X)	Type	Var3	Var4	Var5	Var6	Var7	Total
R	R	-	1	1	1	1	1	5
R	T	Log	1	1	1	1	1	5
		Sqrt	1	1	1	1	1	5
		Arcsine	-	-	-	-	1	1
Total								16

Lastly we analyzed the same data using a completely randomized design to measure differences in Var3 among Var2 labels for the raw and transformed (logarithmic and square root) values using the following model as described in Steel et. al. [5]:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (6)$$

Where Y_{ij} is the j^{th} value of the response variable for the i^{th} treatment; μ the overall population mean of the response; α_i the difference between the population mean of the i^{th} treatment and the overall mean, μ; and ε_{ij} is the error term. Three completely randomized design analysis and their respective parameters were generated.

RESULTS AND DISCUSSION

Data Characteristics

The hypothetical data consisting of 52 rows and eight variables are shown in Table 3, and represent a typical data structure of a healthcare services firm such as Cardinal Health. The seven variables (aside from Var1) shown in Table 3 are different from each other, has its peculiar features, and each poses some unique challenges to model building efforts.

Var2 is a discrete variable with four nominal values. Var3 is a continuous variable with values ranging from zero to some unknown upper limit. There is great variability among data points of Var3, some values exceeded 8 Million and others are in few thousands. These types of variables where individual members show enormous variability pose a challenge to statistical analysts. Before further analysis, decisions have to be made on whether data transformation is needed or whether data segmentation followed by targeted analysis is essential. Alternatively some of the extreme observations can be designated as outliers.

Var4 and Var5 are continuous variables with modest and low variability, respectively. V6 is a variable that represents cases where values occur infrequently likely due to high temporal and / or spatial variation in data collection. This pattern of distribution is typical of events that occur infrequently.

Var7 is different from the rest in that it is a percentage data with upper and lower bounds, with values that range from 0% to 100%. These kinds of data are quite common and represent ratios or shares of entities. Var8 is a discrete variable with values of 1 or 0. Such classification of attributes into discrete entities, commonly referred as the creation of dummy variables, is an effective form of data differentiation and used to form unique groups based on differences in magnitude, availability, identity, location, etc. of the target variable. In general the data represented by the seven variables of Table 3 represent diverse forms in all possible ways.

Table 3 Artificial Data Consisting of Eight Variables (Var1 to Var8) used in the Analysis.

Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8
1	O	1,500,000	25,023	200	0	0.3	0	27	Y	49,896	68,576	400	0	0.1	1
2	O	2,000,000	53,863	400	0	0.4	0	28	O	4,000,000	39,471	300	0	0.1	0
3	Y	35,693	296,735	3,500	250	0.5	1	29	O	2,000,000	114,825	600	0	0.2	0
4	M	500,000	114,791	600	0	0.2	0	30	Y	45,362	493,047	4,800	1,000	0.2	0
5	M	520,000	145,073	900	50	0.7	0	31	Y	54,565	84,849	500	350	0.2	0
6	Y	45,635	210,050	2,000	580	0.1	0	32	O	3,750,000	169,322	900	0	0.2	0
7	M	550,000	125,616	700	0	0.4	0	33	O	4,275,550	305,473	3,500	0	0.3	1
8	O	8,000,000	34,348	300	0	0.2	0	34	Y	52,398	280,831	2,800	0	0.3	0
9	M	450,000	25,902	200	0	0.8	0	35	O	5,450,000	625,102	5,000	0	0.4	0
10	M	444,450	1,144,058	5,200	0	0.4	0	36	N	458	14,475	100	0	0.4	0
11	M	470,000	333,496	4,000	0	0.2	0	37	N	778	31,406	200	0	0.4	1
12	Y	52,350	66,429	400	100	0.1	0	38	N	852	90,668	500	0	0.4	0
13	O	9,250,000	28,372	200	0	0.05	0	39	Y	54,896	31,903	200	330	0.45	0
14	O	3,200,000	38,762	300	0	0.9	1	40	M	355,000	32,660	200	0	0.5	0
15	Y	56,667	431,827	4,300	0	0.6	0	41	N	356	45,885	300	0	0.5	0
16	Y	39,880	490,114	4,600	0	0.8	0	42	M	255,555	69,144	400	0	0.6	1
17	O	2,450,000	82,470	500	230	0.3	1	43	M	275,255	98,828	500	0	0.7	0
18	M	560,000	133,384	800	0	0.45	0	44	M	450,545	1,993,223	5,200	0	0.7	0
19	M	532,450	28,781	200	0	0.5	0	45	N	1,220	100,560	500	0	0.8	0
20	N	1,200	500,631	5,000	0	0.2	0	46	N	586	82,329	400	0	0.8	0
21	O	5,000,000	220,429	2,500	0	0.1	0	47	N	1,236	627,524	5,100	0	0.9	1
22	Y	58,978	167,258	900	0	0.7	1	48	N	1,158	82,467	400	0	0.9	1
23	Y	45,269	288,212	3,000	0	0.9	0	49	M	350,450	46,629	300	670	0.3	0
24	N	1,000	171,607	1,000	0	0.4	0	50	Y	42,087	120,974	600	0	0.1	0
25	O	3,000,000	130,189	700	0	0.05	0	51	N	1,255	113,830	500	40	0.3	0
26	N	1,856	132,601	700	0	0.1	0	52	N	1,300	121,561	600	0	0.5	0

Numerical Measures of Data Dispersion

Data characteristics including measures of central tendency (mean, median, minimum and maximum values), dispersion (standard deviation and coefficient of variation), data distribution (skewness, kurtosis, Shapiro-Wilk test and Kolmogorov-Smirnov test) are shown in Table 4 for

the raw data and following logarithmic, square root and arcsine transformations for five of the seven variables. None of the other three variables (Var1, the id; Var2, the discrete variable; and Var8, the dummy variable) were transformed. This approach is commonly practiced elsewhere. Various descriptive and theory driven graphical (stem-and-leaf plot, box plot, dot plot, histogram) and numerical methods (skewness, kurtosis, Shapiro-Wilk, Kolmogorov-Smirnov Anderson-Darling, and Jarque-Bera tests) are employed to determine the normality of data distribution [6]. In addition, before more rigorous statistical analysis begins, many statisticians perform basic inferential statistical tests such as chi-square and t-tests to assess unadjusted associations. These tests help to guide the direction of the more rigorous analysis [7].

The coefficient of variation (CV) which describes the relative magnitude of the standard deviation in relation to the overall mean is an important metrics when interest is in the size of variation relative to the average size of observations irrespective of units of measurements. CV values less than 20% are generally considered good indicator of low variability although this threshold is much lower for some studies. The reduction of variability as measured by CV in this study was affected both by the type of data and the method of transformation. While logarithmic transformation had reduced CV drastically for three variables (Var3, Var4 and Var5); it showed little impact on the other two variables (Var6 and Var7). Similarly the square root and the arcsine transformations impact on reducing variability as measured by CV values were inadequate for all variables.

Data transformation had reduced the Skewness (degree of departure from symmetry) and Kurtosis (degree of peaked-ness) of data distributions of almost all variables. The exception was Var7 for having about the same Skewness with and without transformation. The normality of data distribution as measured by the Shapiro-Wilk and Kolmogorov-Smirnov tests showed highly significant results at $p=0.01$ indicating that all data remain non-normal prior and after transformation. The exceptions were the logarithmic transformation of Var4 and the square root transformation of Var7 (for the KS test) where data distribution changed to normal after transformation. Transformation as measured by numerical methods showed little or no improvement in the normality of distribution for most of the data.

Table 4 Statistical Measures of Central Tendency, Dispersion, Symmetry and Normality of the Raw as well as Logarithmic, Square Root and Arc-Sine Transformed Data.

Variables	Data Transformation	Statistics									
		Min	Median	Max	Mean	STD	CV	Skewness*	Kurtosis*	Shapiro-Wilk**	KS**
Var3	Raw	356	157267	9250000	1158388	2081613	179.7	2.33	5.41	<0.0001	<0.0100
	Log	5.9	11.7	16.0	11.4	3.1	27.2	-0.38	-1.09	0.0008	<0.0100
	Sqrt	18.9	374.2	3041.4	719.5	808.3	112.3	1.3	0.76	<0.0001	<0.0100
Var4	Raw	14475	114808	1993223	217415	325611	149.8	3.83	17.96	<0.0001	<0.0100
	Log	9.6	11.7	14.5	11.7	1.1	9.2	0.39	-0.13	0.466	>0.1500
	Sqrt	120.3	338.8	1411.8	398.0	245.3	61.6	1.97	5.2	<0.0001	<0.0100
Var5	Raw	100	600	5200	1498	1737	116	1.22	-0.15	<0.0001	<0.0100
	Log	4.6	6.4	8.6	6.7	1.1	17.1	0.46	-1.02	0.0004	<0.0100
	Sqrt	10.0	24.5	72.1	33.1	20.3	61.3	0.95	-0.69	<0.0001	<0.0100
Var6	Raw	0	0	1000	69.23	192.86	278.58	3.39	12.16	<0.0001	<0.0100
	Log	0	0	6.9	1.1	2.2	211.4	1.76	1.36	<0.0001	<0.0100
	Sqrt	0	0	31.6	3.3	7.7	231.7	2.34	4.6	<0.0001	<0.0100
Var7	Raw	0.05	0.4	0.9	0.42	0.26	61.95	0.46	-0.87	0.0031	0.0399
	Log	-3	-0.92	-0.11	-1.12	0.78	-69.06	-0.69	-0.29	0.0022	<0.0100
	Sqrt	0.22	0.63	0.95	0.61	0.21	34.34	-0.05	-0.97	0.0322	0.1274
	ArcSine	0.05	0.41	1.12	0.46	0.31	69.09	0.75	-0.42	0.0005	<0.0100

* Skewness values of -3.0 to 3.0 and Kurtosis values of -0.8 to 0.8 are considered normal.

** The Shapiro-Wilk and the KS (Kolmogorov-Smirnov) tests used the W and the D statistics, respectively, at P=0.05 level to test the normality of data distribution.

Linear Regression

The two sets of simple linear regression results where both Y and X were analyzed as RR and TR are shown in Table 5. The remaining two (RT and TT) are shown in Table 6. Neither various data transformation types (logarithmic, square root or arcsine) nor the separate or simultaneous data transformations of Y and X had affected model or slope (β) significance. The P values are slightly different for each mode and combination of variable transformation but the overall data trend and relationship remain about the same. For a simple linear regression the t-test is equivalent to the F-test for each model as the F value equals t^2.

The model and slope values were significant only when Var5 (Y) was regressed against Var4 (X) for all combinations and methods of transformations (Tables 5 and 6). Graphical display of the relationship between the raw values of Y (Var5) and other four X variables shows only the Var5/Var4 graphs had significant positive slope (Figure 1).

The magnitude of the coefficient of determination (R^2) was affected by the combination of transformed variables (Y and/or X) both for the logarithmic and the square root transformations (Table 5 and 6). Logarithmic and square root data transformation of Y and X (TT) or only the X variable (RT) increased the R^2 values when the regression slope was significant (Var5 regressed against Var4). In contrast the R^2 value decreased slightly for the same set of variables when only the Y variable was transformed (TR). The intercept of the slope was drastically reduced only when Y (TR) or both Y and X (TT) were transformed for regression sets where the slopes were not significant (Var5 regressed individually against Var3, Var6 and Var7). Understandably as the

slope increased or become significant (Var5 regressed against Var4) the intercept values were correspondingly lower for most combinations of transformations especially for that of the RT.

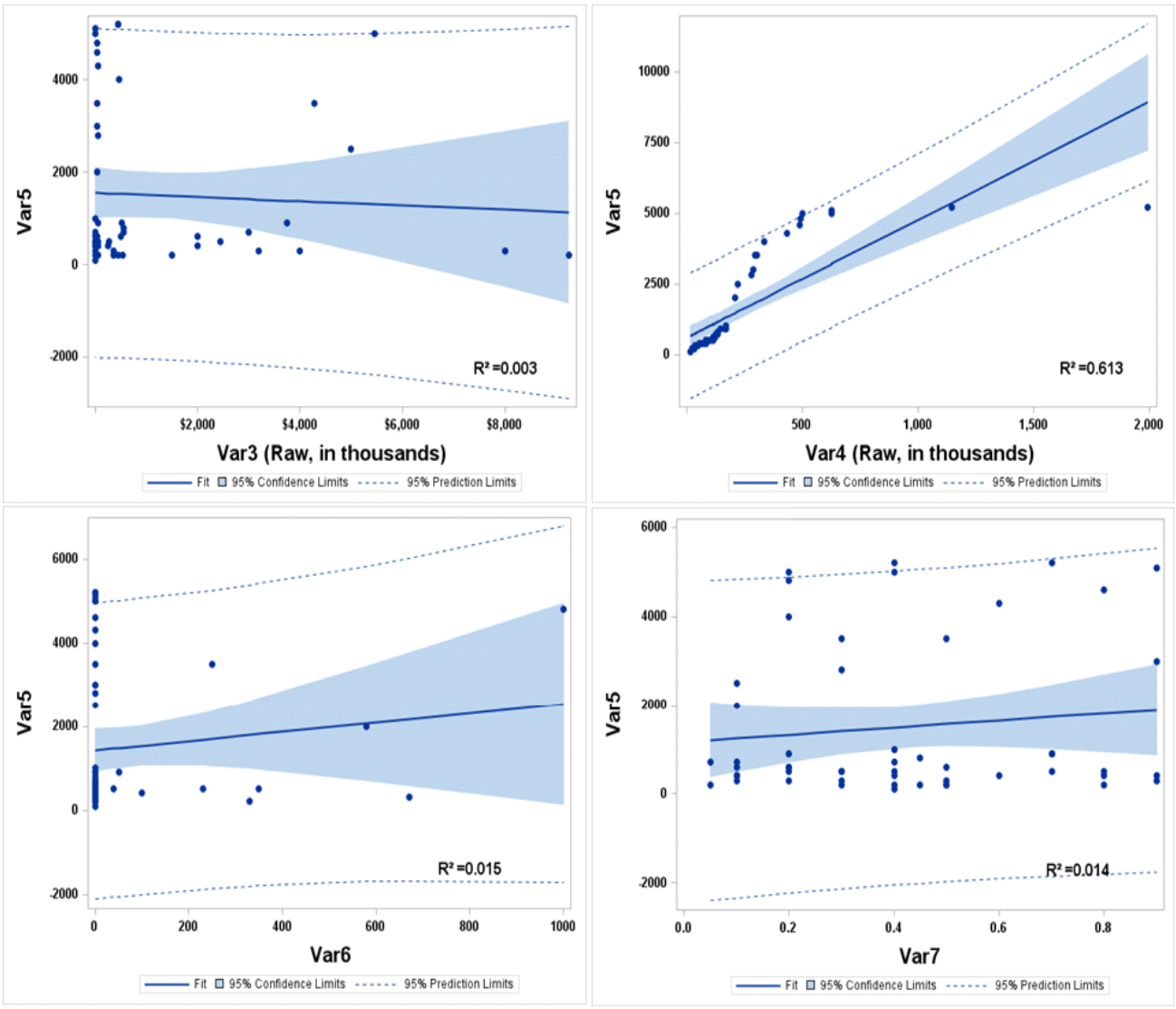
Table 5 Simple Linear Regression Model Parameters for Cases where both Y and X were not Transformed (RR) or only Y was Transformed (TR).

Case	Dependent Variable (Y)	Model Parameters	Independent Variables (X)			
			Var3	Var4	Var6	Var7
RR	Var5	β_0	1550.74	590.42	1421.29	1166.46
		β	0	0.004	1.109	798.34
		Pr > |t|	0.701	<.0001	0.384	0.404
		R-Square	0.003	0.612	0.015	0.014
TR	Log (Var5)	β_0	6.716	6.135	6.619	6.549
		β	0	0	0.0007	0.282
		Pr > |t|	0.571	<.0001	0.416	0.653
		R-Square	0.0065	0.489	0.013	0.004
	Sqrt (Var5)	β_0	33.819	22.905	32.196	29.99
		β	0	0	0.013	7.46
		Pr > |t|	0.649	<.0001	0.385	0.504
		R-Square	0.004	0.566	0.015	0.009

Table 6 Simple Linear Regression Model Parameters for Cases where only X was Transformed (RT) or both Y and X were Transformed (TT).

Case	Response Variable (Y)	Model Parameters	Explanatory Variables (X)								
			Var3		Var4		Var6		Var7		
			Log	Sqrt	Log	Sqrt	Log	Sqrt	Log	Sqrt	Arcsine
RT	Var5	β_0	1406.38	1586.3	-15355.5	-1053.15	1500.67	1456.07	1834.24	870.53	1192.91
		β	8.028	-0.123	1443.41	6.410	-2.468	12.637	299.505	1028.44	670.66
		Pr > |t|	0.919	0.688	<.0001	<.0001	0.982	0.693	0.345	0.381	0.391
		R-Square	0.0002	0.003	0.789	0.819	0	0.003	0.018	0.015	0.015
TT	Log (Var5)	β_0	6.535	6.735	-5.293	5.040	6.653	6.633	6.779	6.459	-
		β	0.012	-0.0001	1.024	0.004	0.013	0.010	0.101	0.339	-
		Pr > |t|	0.825	0.632	<.0001	<.0001	0.862	0.639	0.629	0.660	-
		R-Square	0.001	0.005	0.927	0.777	0.001	0.004	0.005	0.004	-
	Sqrt (Var5)	β_0	31.422	34.204	-172.424	3.373	33.00	32.542	36.214	27.310	-
		β	0.146	-0.002	17.601	0.075	0.09	0.165	2.783	9.468	-
		Pr > |t|	0.875	0.664	<.0001	<.0001	0.95	0.658	0.453	0.490	-
		R-Square	0.0005	0.004	0.861	0.816	0.0001	0.004	0.011	0.010	-

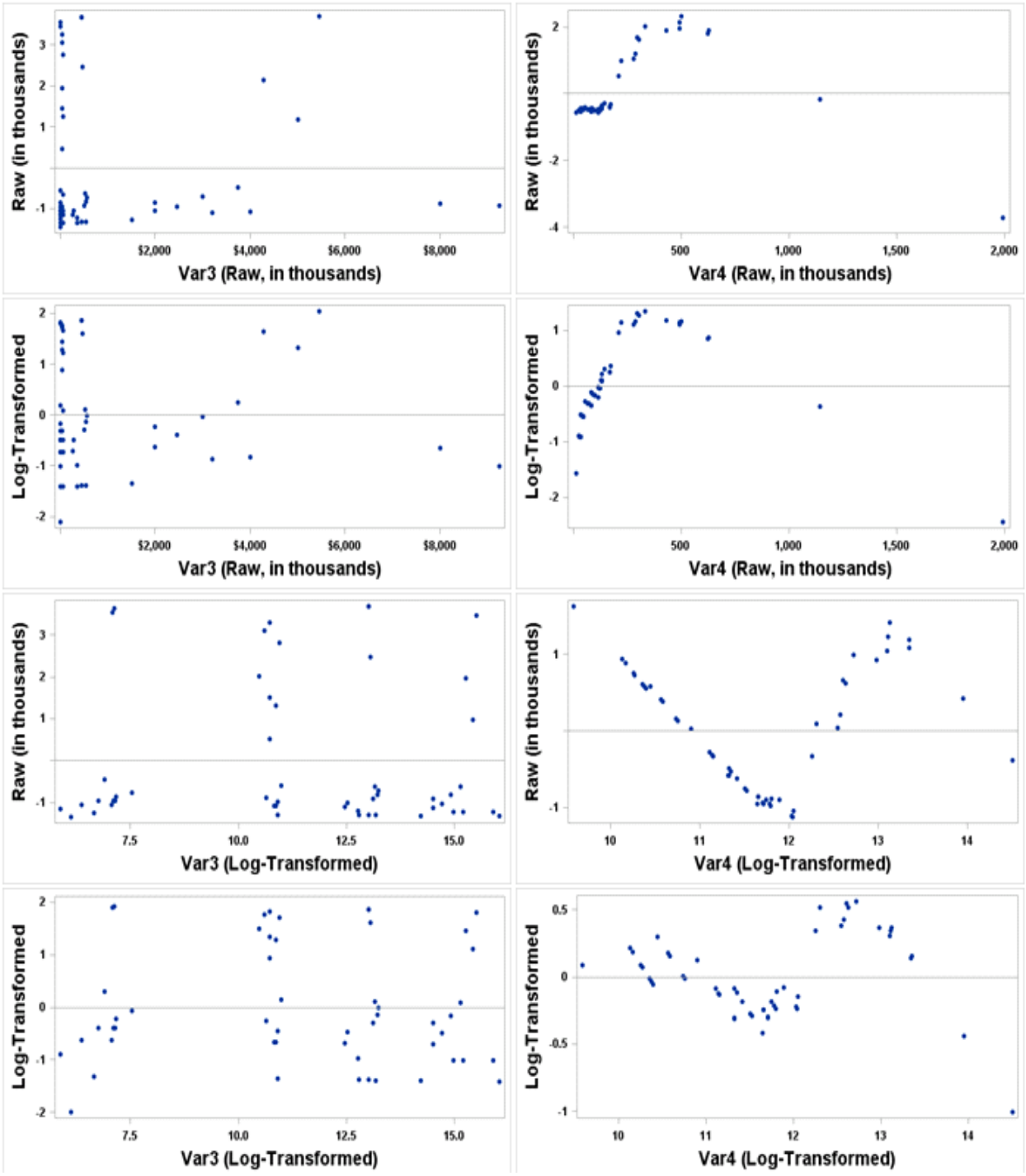
Figure 1 Graphical Display of the Relationship between Var5 (Y) and Four Response (X) Variables (Var3, Var4, Var6, and Var7) without Data Transformation (RR).



Residual plots can be used to assess the quality and to validate the results of an OLS regression. One can examine the underlying statistical assumptions about residuals such as constant variance, independence of variables and normality of the distribution. The Reg Procedure of SAS provides four distinct graphical residual analysis for a simple linear regression involving non-time-series data by plotting the residual given by the model against the explanatory variable (X) or the predicted variable ($\hat{Y}$); by sorting and quantile plotting of the residuals; and by plotting a histogram of residuals. Of the graphic residual plot analysis options provided by the SAS Reg Procedure, we examined several plots that depict the relationship between the residuals and the explanatory variable (X) for all the regression runs on the raw and the transformed variables.

Figure 2 shows eight representative residual analysis plots among the 37 simple linear regression runs. It shows the response variable (Var5) regressed separately against Var3 (where β was not significant) on the left side and Var4 (where β was significant) on the right side for all combinations of Y and X transformations (RR, TR, RT and TT).

Figure 2 Residual Analysis of the Linear Regression of Var5 (Y) and Raw or Transformed Values of Var3 and Var4 (X).



When Var5 regressed against Var3, the residual values plotted against X showed that the residual distribution was highly aggregated for the raw data of both Y and X (RR). Transformation of Y (TR) slightly improved the distribution, while transformation of either X (RT) or both Y and X (TT) showed better distribution of residuals (Figure 2). When Var5 regressed against Var4, the residuals for the RR showed both aggregation and positive correlation. Transformation of Y (TR), X (RT) and both (TT) has reduced the aggregation and yet the positive correlations among

residuals remain unchanged indicating that the constant variance assumption of the model was violated even after data transformation. A higher order model is likely a better fit for these data.

The distribution of residuals around zero is affected by the choice of the combination of Y and X transformations; and TT > RT > TR is the order of improvement of the distribution of residuals around zero. Data transformations have shown some improvement in the distributions of the residuals, however, upon examinations of various types of residual plot analysis described above we conclude that many of the data transformations done on the 37 linear regression runs still retained the non-normality of data distribution pattern inherent in the raw data.

Magnitude of data variability is critical to the success of data transformation and the ensuing data analysis. It was reported that logarithmic transformation of data is suitable when the variance is proportional to the square of the mean or the coefficient of variation is constant or where effects are multiplicative [9]. The reason why logarithmic transformation of some of our data failed to convert variables to a normal distribution is due to large variability in relation to the overall mean. With the exception of Var7, all of the other raw variables have CV values larger than constant (100%). In addition, even if it was suggested that counts can be transformed to near normality by taking their square roots [6], the magnitude of data variability appeared a critical factor whether transformation will convert a non-normal data to normal.

Logistic Regression

Results of the logistic regression where Y (Var8) was regressed against five individual X variables is summarized in Table 7. Logarithmic, square root and arcsine transformations of the X variables were used whenever possible. Parameters of the maximum likelihood estimates of the intercept and the X variables as well as the P value which measures the significance of the Wald Chi-Square statistics (obtained by dividing the respective coefficient by the standard error) and the Tjur R^2 are shown for each transformation.

Only Var7 showed significant coefficient at P=0.1 level for the raw and the three data transformations (Table 7). The coefficients tell the magnitude of change in the log-odds for every raw or transformed unit change in the explanatory variable. The odds ratio is often preferred to interpret logistic regression as it relates the odds (rather than the probabilities) with the explanatory variable (X), and is obtained by exponentiation of the coefficient. For the Var7 data the coefficients (odds ratio) were -2.731(0.07), 0.979(0.38), 0.071(0.03) and -2.254(0.11) for the raw, logarithmic, square root and arcsine transformations, respectively. Interpretation of the odds ratio for the transformed data is cumbersome. For the raw data an odds ratio of 1 indicate that the predicted odds of Var7 is the same whether Var8 was 1 or 0. For the transformed Var7 data, the odds ratios are difficult to interpret and do not convey the same message as that of the raw data.

Table 7 Logistic Regression Model Parameters for Cases where Var8 (Y) was Regressed Against Five Explanatory Variables (X) in a Row-Row (RR) and Row-Transformed (RT) Setups of the Y and the X Variables.

Case	Response Variable (Y)	Model Parameters	Explanatory Variables (X) Transformation Status	Var3	Var4	Var5	Var6	Var7
RT	Var8	Intercept	Raw	1.392	1.314	1.449	1.384	2.715
		Coefficent		0	0	0	0.001	-2.731
		Pr > Chi-Square		0.831	0.665	0.964	0.699	0.056
		Tjur R-Squared		0.001	0.003	0	0.002	0.082
		Intercept	Log	0.897	1.027	1.560	1.448	0.497
		Coefficent		0.048	0.035	-0.019	-0.012	-0.979
		Pr > Chi-Square		0.673	0.917	0.952	0.941	0.100
		Tjur R-Squared		0.004	0.000	0.0001	0.000	0.064
		Intercept	Sqrt	1.363	1.252	1.462	1.419	3.711
		Coefficent		0.0001	0.001	-0.001	0.005	-3.478
		Pr > Chi-Square		0.821	0.763	0.963	0.917	0.071
		Tjur R-Squared		0.001	0.002	0	0.0002	0.075
		Intercept	Arcsine	-	-	-	-	2.609
		Coefficent		-	-	-	-	-2.258
		Pr > Chi-Square		-	-	-	-	0.044
		Tjur R-Squared		-	-	-	-	0.092

Completely Randomized Experimental Design

Var2 has four discrete variables (Table 3). These four discrete variables were considered as treatments and the 13 distinct observations as replications to conduct a completely randomized design analysis to determine if there is any difference among the means of the four treatments. The F test indicated a highly significant difference among the means (P=0.01) for the raw, the logarithmic and the square root transformed data (Table 8). The size of the F statistics, computed as the mean square of the treatment (model) divided by the mean square of the error term increased substantially with transformation. This indicated that model significance (the rejection of the null hypothesis that there is no difference among treatments) is easily achieved with transformed than with the row data. This was possible due to data transformation effect on a greater reduction of error mean square in relation to that of the model.

Table 8 Analysis of Variance Table for a Completely Randomized Design Experiment where Treatments Constitute four treatments of Var2 for Raw, Logarithmic and Square Root Transformed Data.

Treatment	Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Var3	Model	3	1.5604E+14	5.20134E+13	38.44	<.0001
	Error	48	6.49485E+13	1.35309E+12	-	-
	Corrected Total	51	2.20989E+14	-	-	-
Log (Var3)	Model	3	485.474	161.825	1079.29	<.0001
	Error	48	7.197	0.150	-	-
	Corrected Total	51	492.671	-	-	-
Sqrt (Var3)	Model	3	29710175.52	9903392	131.78	<.0001
	Error	48	3607211.255	75150	-	-
	Corrected Total	51	33317386.77	-	-	-

Table 9 Mean Separation of Var2 Group Treatments using the Duncan Multiple's Range Test.

Treatment	Replications	Age	Mean	DMRT*
Var3	13	O	4,144,273	A
	13	M	439,516	B
	13	Y	48,744	B
	13	N	1,020	B
Log (Var3)	13	O	15	A
	13	M	13	B
	13	Y	11	C
	13	N	7	D
Sqrt (Var3)	13	O	1,968	A
	13	M	659	B
	13	Y	220	C
	13	N	31	C

* Duncan Multiple Range Test, Means follow by the same letter are not significantly different from each other (P = 0.05).

Among the handful of mean separation techniques, we have selected to use the Duncan Multiple Range Test (DMRT) due to its ease of identifying mean differences using alphabetical letters. Mean separation using DMRT provided different results for the raw, logarithmic and square root transformations (Table 9). For the raw data, only the mean of treatment O had significantly higher Var3 than each of the other treatments. Whereas for the logarithmic transformed values, each of the Var2 treatment mean was significantly different from each other. The square root transformed data showed that the treatments O and M were different from each other and the rest of the treatments, while the Y and the N treatment means were not significantly different from

each other. Mean separation in the analysis of experimental design is a critical component of decision making processes. This study shows that the same data can offer different results in declaring whether one group mean is different from the other. Thus experimentalists need to be aware of the implications of data transformation on mean separation and subsequent interpretation of results that affect pricing or product management decisions.

SUMMARY

Although many workers provided a detailed treatise of data transformation, empirical evidence on simultaneous comparisons of various transformation methods on diverse data forms of the healthcare services industry and their impact on diverse model building efforts, model parameters and model result interpretation are lacking. This study attempted to address this gap in our understandings of the healthcare services data, its transformations and model fitting.

The artificial healthcare services data provided in Table 3 is a typical representative of diverse data forms collected periodically in health care services industry. For such kinds of data, the practical importance of data transformation lies whether some improvements are obtained in normality of data distribution, level of model and parameter significance, coefficient of determination and other statistics. Even if improvements are obtained, the analyst has to grapple with the practicality of using model parameters from transformed data. In a simple linear regression, the β value indicates the rate of change in the response variable for one unit of change in the explanatory variable. Additional effort is required to explain this rate of change expressed in logarithmic scale, in square root values or other transformation metrics for decision making purposes. This can be challenging especially when the analyst is dealing with many data formats and sources, and when speed and accuracy of analysis is of paramount importance. In working with the Box-Cox power transformation, LaLonde [8] noted that transformations, in many eyes, complicate data analysis. The complication comes in the form of explaining to the non-statistician why one is modeling a transformed value of their favorite variable rather than the variable in its unadulterated form.

Although data transformations have improved the distribution of residuals around zero especially when both the response and the explanatory variables were transformed, it affected little of the significance of model parameters. The magnitude of data variability is an important factor in affecting the improvement of residual distributions through data transformations. While some improvement in residual distribution was possible with data transformation when regression was conducted, comparable improvement was not seen in the numeric measurements of normality. The two tests (Shapiro-Wilk and the Kolmogorov-Smirnov) improved little after logarithmic and square root transformations for most of the data.

Data transformation must offer clear advantage in improving the robustness of data analysis in terms of the significance of model parameters, improving the proportion of total variations explained by the model, mean separation and the ease with which analysis results of transformed data are explained. While the production, the management, and the analysis of diverse and complex forms of data that assume many distribution types will remain critical components of

business operations in the healthcare services industry, the deployment of data transformation must be weighed against its cost and potential benefits.

REFERENCES

[1] Collani E. and K. Drager. 2001. Binomial Distribution Handbook for Scientists and Engineers. Birkhauser. Boston. 357 pp.

[2] Forbes C., M. Evans, N. Hastings and M. Peacock. 2011. Statistical Distributions. Fourth Edition. John Wiley & Sons. Hoboken, New Jersey. 212 pp.

[3] Littell R. C, W. W. Stroup, and R. J. Freund. 2002.SAS® for Linear Models. Fourth Edition. SAS Institute Inc. Cary North Carolina. 466 pp.

[4] Allison P. D. 1999. Logistic Regression using SAS®. SAS Institute Inc. Cary North Carolina. 287 pp.

[5] Steel R. G, J. H. Torrie and D. A. Dickey. 1996. Principles and Procedures of Statistics: A Biometrical Approach. McGraw-Hill Higher Education. pp 450.

[6] Park, H. M.. 2008. Univariate Analysis and Normality Test Using SAS, Stata, and SPSS. Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University." 41 pp. http://www.indiana.edu/~statmath/stat/all/normality/index.html

[7] Waller J. L and M. H. Johnson. 2013. Chi-Square and T-Tests Using SAS®: Performance and Interpretation. SAS Global Forum Proceedings. Paper 430-2013. 12 pp.

[8] LaLonde S. M. 2012. Transforming Variables for Normality and Linearity-When, How, Why and Why Not's. SAS Global Forum 2012. Paper 430-2012. 8 pp.

[9] Statistical Concepts and Analytics Explained. Transformation of data. 2010. http://statisticalconcepts.blogspot.com/2010/02/transformation-of-data-validity-of.html

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please feel free to contact the authors at:

Dawit Mulugeta, Ph. D.
Director of Advanced Analytics, Pricing Analytics Team
Cardinal Health, Dublin, OH,
E-mail: dawit.mulugeta@cardinal.health.com

Jason Greenfield
Senior Consultant of Advanced Analytics, Pricing Analytics Team
Cardinal Health, Dublin, OH,
E-mail: jason.greenfield@cardinal.health.com

Tison Bolen
Manager, Operational Excellence,
Cardinal Health, Dublin, OH,
E-mail: tison.bolen@cardinal.health.com

Lisa Conley
Director of Advanced Analytics, Pricing Analytics Team
Cardinal Health, Dublin, OH,
E-mail: lisa.conley@cardinal.health.com

APPENDIX

SAS® CODE

```
**This code generates basic statistics including measures of central tendency both
for raw and transformed data. In addition it produces linear and logistic regression
parameter estimates and associated outputs. It also analyzes all data in a completely
randomized design setups.**;


%macro transform(var);
   if &var > 0 then &var.SQRT = sqrt(&var); else &var.SQRT = 0;
   if &var > 0 then &var.LOG = log(&var); else &var.LOG = 0;
   if abs(&var) < 1 then &var.ASIN = arsin(&var); else &var.ASIN = &var;
%mend;

data hs_service_data ;
input var1 var2 $ var3 var4 var5 var6 var7 var8 ;
   %transform(var3)
   %transform(var4)
   %transform(var5)
   %transform(var6)
   %transform(var7)

datalines;
```

1	O	1500000	25023	200	0	0.3	0
2	O	2000000	53863	400	0	0.4	0
3	Y	35693	296735	3500	250	0.5	1
4	M	500000	114791	600	0	0.2	0
5	M	520000	145073	900	50	0.7	0
6	Y	45635	210050	2000	580	0.1	0
7	M	550000	125616	700	0	0.4	0
8	O	8000000	34348	300	0	0.2	0
9	M	450000	25902	200	0	0.8	0
10	M	444450	1144058	5200	0	0.4	0
11	M	470000	333496	4000	0	0.2	0
12	Y	52350	66429	400	100	0.1	0
13	O	9250000	28372	200	0	0.05	0
14	O	3200000	38762	300	0	0.9	1
15	Y	56667	431827	4300	0	0.6	0
16	Y	39880	490114	4600	0	0.8	0
17	O	2450000	82470	500	230	0.3	1
18	M	560000	133384	800	0	0.45	0
19	M	532450	28781	200	0	0.5	0
20	N	1200	500631	5000	0	0.2	0
21	O	5000000	220429	2500	0	0.1	0
22	Y	58978	167258	900	0	0.7	1
23	Y	45269	288212	3000	0	0.9	0
24	N	1000	171607	1000	0	0.4	0

```
 25    O    3000000    130189    700     0    0.05    0
 26    N       1856    132601    700     0    0.1     0
 27    Y      49896     68576    400     0    0.1     1
 28    O    4000000     39471    300     0    0.1     0
 29    O    2000000    114825    600     0    0.2     0
 30    Y      45362    493047   4800  1000    0.2     0
 31    Y      54565     84849    500   350    0.2     0
 32    O    3750000    169322    900     0    0.2     0
 33    O    4275550    305473   3500     0    0.3     1
 34    Y      52398    280831   2800     0    0.3     0
 35    O    5450000    625102   5000     0    0.4     0
 36    N        458     14475    100     0    0.4     0
 37    N        778     31406    200     0    0.4     1
 38    N        852     90668    500     0    0.4     0
 39    Y      54896     31903    200   330    0.45    0
 40    M     355000     32660    200     0    0.5     0
 41    N        356     45885    300     0    0.5     0
 42    M     255555     69144    400     0    0.6     1
 43    M     275255     98828    500     0    0.7     0
 44    M     450545   1993223   5200     0    0.7     0
 45    N       1220    100560    500     0    0.8     0
 46    N        586     82329    400     0    0.8     0
 47    N       1236    627524   5100     0    0.9     1
 48    N       1158     82467    400     0    0.9     1
 49    M     350450     46629    300   670    0.3     0
 50    Y      42087    120974    600     0    0.1     0
 51    N       1255    113830    500    40    0.3     0
 52    N       1300    121561    600     0    0.5     0
;
run ;

/*************** UNIVARIATE SUMMARY ******************/
title 'Summary';
proc univariate data=hs_service_data normaltest;
   var var3 var3SQRT var3LOG
      var4 var4SQRT var4LOG
      var5 var5SQRT var5LOG
      var6 var6SQRT var6LOG
      var7 var7SQRT var7ASIN var7LOG;
   ods select Moments BasicMeasures Quantiles TestsForNormality;
run;

/************************************************/
/* UNIVARIATE LINEAR MODEL */
/* RAW */

%macro LINEARmod(depVAR=, indVAR=);
      title "Linear Model &depVar vs &indVar";
      proc reg data=hs_service_data ;
```

```
        model &depVAR = &indVAR;
        ods select FitStatistics ParameterEstimates;
    run;
    quit;
%mend;

/* Raw */
%LINEARmod(depVAR=var5, indVAR=var3)
%LINEARmod(depVAR=var5, indVAR=var4)
%LINEARmod(depVAR=var5, indVAR=var6)
%LINEARmod(depVAR=var5, indVAR=var7)

/* Dependent Only*/
%LINEARmod(depVAR=var5LOG, indVAR=var3)
%LINEARmod(depVAR=var5LOG, indVAR=var4)
%LINEARmod(depVAR=var5LOG, indVAR=var6)
%LINEARmod(depVAR=var5LOG, indVAR=var7)

%LINEARmod(depVAR=var5SQRT, indVAR=var3)
%LINEARmod(depVAR=var5SQRT, indVAR=var4)
%LINEARmod(depVAR=var5SQRT, indVAR=var6)
%LINEARmod(depVAR=var5SQRT, indVAR=var7)

/*Independent Only*/
%LINEARmod(depVAR=var5, indVAR=var3LOG)
%LINEARmod(depVAR=var5, indVAR=var4LOG)
%LINEARmod(depVAR=var5, indVAR=var6LOG)
%LINEARmod(depVAR=var5, indVAR=var7LOG)

%LINEARmod(depVAR=var5, indVAR=var3SQRT)
%LINEARmod(depVAR=var5, indVAR=var4SQRT)
%LINEARmod(depVAR=var5, indVAR=var6SQRT)
%LINEARmod(depVAR=var5, indVAR=var7SQRT)

%LINEARmod(depVAR=var5, indVAR=var7ASIN)

/* Both */
%LINEARmod(depVAR=var5LOG, indVAR=var3LOG)
%LINEARmod(depVAR=var5LOG, indVAR=var4LOG)
%LINEARmod(depVAR=var5LOG, indVAR=var6LOG)
%LINEARmod(depVAR=var5LOG, indVAR=var7LOG)

%LINEARmod(depVAR=var5LOG, indVAR=var3SQRT)
%LINEARmod(depVAR=var5LOG, indVAR=var4SQRT)
%LINEARmod(depVAR=var5LOG, indVAR=var6SQRT)
%LINEARmod(depVAR=var5LOG, indVAR=var7SQRT)

%LINEARmod(depVAR=var5SQRT, indVAR=var3SQRT)
%LINEARmod(depVAR=var5SQRT, indVAR=var4SQRT)
%LINEARmod(depVAR=var5SQRT, indVAR=var6SQRT)
%LINEARmod(depVAR=var5SQRT, indVAR=var7SQRT)

%LINEARmod(depVAR=var5SQRT, indVAR=var3LOG)
%LINEARmod(depVAR=var5SQRT, indVAR=var4LOG)
%LINEARmod(depVAR=var5SQRT, indVAR=var6LOG)
```

```
%LINEARmod(depVAR=var5SQRT, indVAR=var7LOG)

/* UNIVARIATE LOGISITIC MODEL */
/* RAW */
%macro LOGISTICmod(depVAR=, indVAR=);

      title "Logistic Model &depVar vs &indVar";
      proc logistic data=hs_service_data;
         model &depVAR = &indVAR;
         ods select ParameterEstimates;
         output out=p pred=yhat;
      run;

      /* Rsq */
      title 'Tjur R-Squared';
      proc ttest data=p ;
         class var8;
         var yhat;
         ods select Statistics;
      run;

%mend;

%LOGISTICmod(depVAR=var8, indVAR=var3)
%LOGISTICmod(depVAR=var8, indVAR=var4)
%LOGISTICmod(depVAR=var8, indVAR=var6)
%LOGISTICmod(depVAR=var8, indVAR=var7)
%LOGISTICmod(depVAR=var8, indVAR=var5)

/* LOG */
%LOGISTICmod(depVAR=var8, indVAR=var3LOG)
%LOGISTICmod(depVAR=var8, indVAR=var4LOG)
%LOGISTICmod(depVAR=var8, indVAR=var6LOG)
%LOGISTICmod(depVAR=var8, indVAR=var7LOG)
%LOGISTICmod(depVAR=var8, indVAR=var5LOG)

/* SQRT */
%LOGISTICmod(depVAR=var8, indVAR=var3SQRT)
%LOGISTICmod(depVAR=var8, indVAR=var4SQRT)
%LOGISTICmod(depVAR=var8, indVAR=var6SQRT)
%LOGISTICmod(depVAR=var8, indVAR=var7SQRT)
%LOGISTICmod(depVAR=var8, indVAR=var5SQRT)

/* ARCSINE */
%LOGISTICmod(depVAR=var8, indVAR=var7ASIN)

/*******************************************************/
/* COMPLETELY RANDOMIZED DESIGN                        */
title "Completely Randomized Design var3 vs var2";
proc glm data=hs_service_data;
     class var2 ;
     model var3 = var2  ;
       means var2 / duncan ;
       lsmeans var2 / stderr pdiff;
   ods select OverallANOVA MCLines ;
```

```
run ;
title "Completely Randomized Design var3LOG vs var2";
proc glm data=hs_service_data;
     class var2 ;
     model var3LOG = var2  ;
       means var2 / duncan ;
       lsmeans var2 / stderr pdiff;
   ods select OverallANOVA MCLines ;
run ;
title "Completely Randomized Design var3SQRT vs var2";
proc glm data=hs_service_data;
     class var2 ;
     model var3SQRT = var2  ;
       means var2 / duncan ;
       lsmeans var2 / stderr pdiff;
   ods select OverallANOVA MCLines ;
run ;
quit;
title;
```