

TREATMENT OF NON NORMAL RESPONSES FROM DESIGNED EXPERIMENTS

G. M. Oyeyemi
Department Of Statistics
University Of Ilorin
Ilorin -NIGERIA

Abstract

Many experimental designs, most especially industrial designs produce non-normal response variables. The Least Squares method of modeling may therefore not produce efficient estimates. Models are built using β -technique and Box-Cox methods of data transformation or by using the GLM to overcome the non-normality nature of the data. In this paper, these techniques are compared using the histograms of the estimated mean responses and by examining the length of the confidence interval about the mean responses. It is observed that, β -technique is the best method of data transformation but GLM provides an excellent alternative if the experimental design points are not replicated.

Key Words: β -Technique, Robust Parameter Design, Generalized Linear Model

1. Introduction

Design of experiment has come to play an important role in the application of Robust Parameter Design (RPD) methodology to achieve a process that is robust or insensitive to uncontrollable variables. RPD was pioneered by Taguchi (1987) as a strategy for improving the performance of a system, a process or a product. This approach is based on the premise that the key to improve performance is by reducing variability. Thus experiments should be designed to search for ways of reducing variability (that is, selecting the appropriate setting that will reduce variability).

Taguchi identified two main types of RPD applications, which he referred to as static characteristics and dynamic characteristics. Static characteristics refer to situations where the signal factors are constant. A dynamic characteristic is the situation where the relationship between two quantities, the signal factors and response, are of interest (that is, signal factors are varied in the experiment).

The objective of Robust Parameter Design is to design a system or product in which its performance is insensitive to uncontrollable variables. This is done by systematically investigating the relationship between the appropriate control factors and the uncontrollable

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda + 1}{\lambda} & \lambda \neq 0 \\ \log(Y) & \lambda = 0 \end{cases}$$

which holds for $Y > 0$

and a two-parameter family indexed by λ_1 and λ_2

$$Y^{(\lambda)} = \begin{cases} \frac{(Y + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \lambda_1 \neq 0 \\ \log(Y + \lambda_2) & \lambda_1 = 0 \end{cases}$$

which holds for $Y > -\lambda_2$

β -Technique

In Robust Parameter Design methodology, one of the drawbacks is often the ambiguity in classification of variables (factors) that affect variability or target, which prevents a distinction between the factors affecting the variation and those affecting only the mean response. A large overlap in classification is as a result of violation of the assumption that the variance of the treatments within an experimental trial (cell) should not change with mean level of the measurement (constancy of error variance). In order to make trial variance independent of the mean we need a suitable data transformation. Therefore if we can establish the functional relationship between the trial variance and the trial mean, this will determine the type of transformation required for the data.

Suppose that $\sigma_y^{(i)}$ and $\mu_y^{(i)}$ are respectively standard deviation and mean in the original scale of the observations of Y in the i^{th} trial and are connected by the relation

$$\sigma_y^{(i)} = f(\mu_y^{(i)}) \quad \text{for all } i.$$

Let the transformation $T(Y)$ of Y transform Y such that the new cell variance of the transformed data is independent of the mean. Using Taylor series expansion near $[\mu, T(\mu)]$

$$T(Y) = T(\mu) + (Y - \mu)T'(\mu) + \frac{1}{2}(Y - \mu)^2 T''(\mu) + \dots$$

$$\text{Var}[T(Y)] = [T'(\mu)]^2 \text{var}(Y - \mu) + \text{higher order terms},$$

where T' is the first derivative of T

$$\sigma_T^2 \equiv [T'(\mu)]^2 \sigma_Y^2$$

$$\sigma_T = [T'(\mu)] \sigma_Y$$

$$[T'(\mu)] = \frac{\sigma_y}{\sigma_x}$$

$$T(\mu) = \int \frac{\sigma_y}{\sigma_x} d\mu$$

Since the transformation T is supposed to make the variance σ_y^2 constant, ($=c^2$ say) and $\sigma_y = f(\mu)$, we have:

$$T(\mu) = \int \frac{c}{f(\mu)} d\mu$$

If we assume the functional relationship

$$\sigma_y = f(\mu) = c \cdot \mu^\beta$$

and integrate the above equation, we have

$$T(\mu) = \begin{cases} c_1 \mu^{1-\beta} & \text{if } \beta \neq 1 \\ c_1 \log(\mu) & \text{if } \beta = 1 \end{cases}$$

where c_1 is a constant.

From the above equation, we need a transformation to the power of $(1-\beta)$ if $\beta \neq 1$ and log transformation if $\beta = 1$.

This is equivalent to the one-parameter Box-Cox family of transformation with $\lambda = 1-\beta$.

One way to establish whether $\sigma_y = c \cdot \mu^\beta$ is a reasonable approximation is to estimate β by the least square method from the relationship between sample mean $\bar{Y}_i$ and s_i given by

$$\log(s_i) = \alpha + \beta \log(\bar{Y}_i) + \epsilon_i, \quad i = 1, 2, 3, \dots, n$$

(n is the number of experimental trials).

The significance of β can be tested through the usual t -test. The limitation of β -technique is that it cannot deal with cases that do not have replicate in each cell.

3. Generalized Linear Model (GLM)

The Generalized Linear Model introduced by Nelder and Wedderburn (1972), is a unification of linear and non-linear regression models that allows the experimenter to select a response distribution that is a member of a very rich and flexible family; the Exponential family. The Normal, Binomial, Poisson, Exponential, Gamma, Inverse Normal and Negative Binomial distributions are all members of the exponential family. With the Generalized Linear Model, normality and constant variance are no longer required; in fact, in all

exponential family members except the normal distribution, the variance is a function of the mean.

A Generalized Linear Model consists of three components:

1. The response distribution, which is a member of the exponential family. This component identifies the response Y and its probability distribution (usually exponential family). It is assumed that x_i, y_i are (conditionally) independent, and the (conditional) distribution of y_i belongs to a simple exponential family with (conditional) expectation, $E(y_i / x_i) = \mu_i$.
2. A linear predictor, which is the linear combination of the design variables, including mean effects, interaction and possibly polynomial terms for quantitative factors. This component specifies the explanatory variables used in the model.

$$\eta = X' \beta$$

$$= \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$$

where x can be main effect, interaction or polynomial. β 's are the unknown parameters to be estimated through MLE.

3. A link function that relates the natural mean for the specified exponential family member to the linear predictor. This component describes the functional relationship between the systematic components of η and mean μ of the random component.

4. Confidence Intervals

The confidence intervals shall be used to measure the predictive performance of the fitted model. The confidence interval of the mean response for each of the experimental run and its associated length or precision, provides the experimenter with critical information about the uncertainty associated with the estimation of the mean response. This will indicate the most appropriate model and hence the best method of selecting the important variables or covariates.

The statistical inference for Generalized Linear Model is usually based on asymptotic theory. The overall model significance is evaluated with an analysis of deviance. This reduces to the familiar analysis of variance when the response distribution is normal. Inference on individual parameters in the model is often performed via Wald inferences.

which are t-like ratio of parameter estimates to standard errors that have an asymptotic normal distribution. Wald inference can be used to obtain confidence intervals on the model parameters (Fears *et al.*, 1996 and Hanck and Donner, 1997).

5. Examples

The first example is data on the printing study of Vining and Myers (1990) (adapted from Box and Draper (1987, p.247)). The experiment was conducted to investigate the effect of Speed (A), Pressure (B), and Distance (C) on a printing machine's ability to apply coloured ink to package labels (Y).

A three level factorial design on the control variables with three runs at each design point was used to fit the responses. The data is shown in Table 1.

Table 1. The Printing Study Data.

	A	B	C	y_1	y_2	y_3	$\bar{y}$	s
1.	A1	B1	C1	34	10	28	24.00	12.49
2.	A2	B1	C1	115	116	130	120.33	8.39
3.	A3	B1	C1	192	186	263	213.67	42.83
4.	A1	B2	C1	82	88	88	86.00	3.46
5.	A2	B2	C1	44	178	188	136.67	80.41
6.	A3	B2	C1	322	350	350	340.67	16.17
7.	A1	B3	C1	141	110	86	112.33	27.57
8.	A2	B3	C1	259	251	259	256.33	4.62
9.	A3	B3	C1	290	280	245	271.67	23.63
10.	A1	B1	C2	81	81	81	81.00	0.00
11.	A2	B1	C2	90	122	93	101.67	17.67
12.	A3	B1	C2	319	376	376	357.00	32.91
13.	A1	B2	C2	180	180	154	171.33	15.01
14.	A2	B2	C2	372	372	372	372.00	0.00
15.	A3	B2	C2	541	568	396	501.67	92.50
16.	A1	B3	C2	288	192	312	264.00	63.49
17.	A2	B3	C2	432	336	513	427.00	88.61
18.	A3	B3	C2	713	725	754	730.67	21.08
19.	A1	B1	C3	364	99	199	220.67	133.82
20.	A2	B1	C3	232	221	266	239.67	23.46
21.	A3	B1	C3	408	415	443	422.00	18.52
22.	A1	B2	C3	182	233	182	199.00	29.45
23.	A2	B2	C3	507	515	434	485.33	44.64
24.	A3	B2	C3	846	535	640	673.67	158.21
25.	A1	B3	C3	236	126	168	176.67	55.51
26.	A2	B3	C3	660	440	403	501.00	138.94
27.	A3	B3	C3	878	991	1161	1010.00	142.45

From the table, the estimates of the mean ($\bar{y}$) and standard deviation (s) are obtained. The regression line of $\ln(s)$ on $\ln(\bar{y})$ is obtained i.e.

The fitted model is

$$\ln(s) = -0.425 + 0.767 \ln(\bar{y})$$

The test statistic $t = 2.54$ shows that the β is significantly different from zero, hence there is need for transformation with $\beta = 0.767$

For β -technique

$$Y_{TP} = Y^{1-\beta} = Y^{0.23}$$

For Box-Cox

$$Y_{BC} = \frac{Y^\lambda - 1}{\lambda} = \frac{Y^{0.23} - 1}{0.23} \quad \text{where } \lambda = 1 - \beta$$

The third approach is to fit a model using Generalized Linear Model, with Gamma as the response distribution and log link function. The histogram of the original data in Fig. 1 shows an exponential distribution. This suggests the use of Gamma distribution in the Generalized Linear Model.

After the data transformation, using both β -technique and Box-Cox methods, the model is fitted for each case. The estimates of the mean responses are obtained from the fitted model using the original data, β -technique, Box-Cox and the GLM. The results are shown in Table 2.

Fig. 1 Histogram of the Original Data

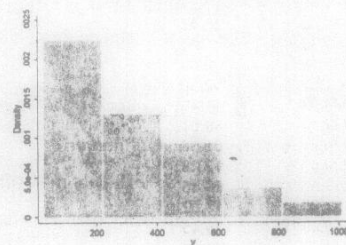


Fig. 2 Histogram of Transformed Data Using β -Technique

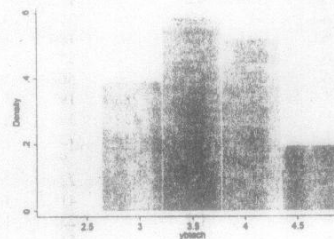


Fig.3 Histogram of Transformed Data Using Box-Cox Method

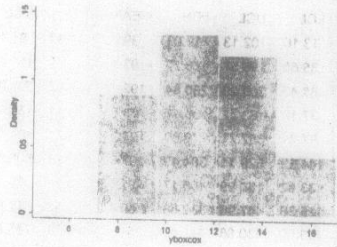
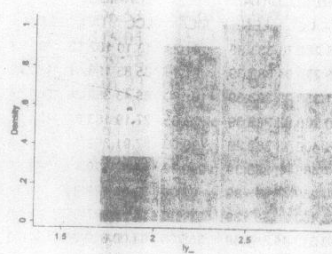


Fig. 4 Histogram of GLM Data



The graph in Fig. 2 shows the histogram of the mean responses of transformed data using β -technique, while Fig. 3 shows the histogram of the mean responses when Box-Cox method is used to transform the data. Fig. 4 is the histogram of the mean responses when GLM is used to fit the model.

Fig. 1 shows a total departure from normal distribution hence the reason for transformation. Fig. 2 and Fig. 3 (histograms of β -technique and Box-Cox) give an approximate normal distribution curve while Fig. 4 (histogram of GLM) is better than untransformed data but not as good as those of β -technique and Box-Cox. The 95% confidence intervals (Table 2.) of the estimated means also provide good explanation as to which of the methods of transformation is better. The estimated mean responses for all the three methods are approximately the same; particularly, β -technique and Box-Cox methods have the same mean responses.

Table 2 shows that the 95% confidence intervals of all the three methods contain the true means of the response variables. The length of the 95% confidence interval of each method is compared with that of the untransformed data. A shorter length is preferred to longer one. Both β -technique and Box-Cox methods behave in a similar manner. The confidence lengths of both methods are shorter than that of untransformed data on 15 different occasions while that of GLM method is shorter on 11 occasions. This shows that β -technique and Box-Cox methods have higher precision in estimating the mean responses.

Table 2. The mean responses and confidence intervals (length)

TRUE ^a UNTRANSFORMED DATA					B-TECHNIQUE				BOX-COX				GLM			
MEAN	MEAN	LCL	UCL	LEN.	MEAN	LCL	UCL	LEN.	MEAN	LCL	UCL	LEN.	MEAN	LCL	UCL	LE
24.00	97.03	-94.66	288.73	383.39	41.24	13.10	102.13	*89.03	41.24	13.10	102.13	*89.03	39.72	19.43	81.18	*6
120.33	94.35	-97.35	286.04	383.39	91.18	35.85	196.57	*160.72	91.18	35.85	196.57	*160.72	93.59	45.79	191.28	*14
213.67	166.60	-25.11	358.28	383.39	191.60	88.43	369.26	*280.84	191.58	88.43	369.26	*280.84	192.26	94.07	392.94	*29
86.00	77.22	-114.50	268.92	383.39	73.05	27.19	163.37	*136.17	73.05	27.19	163.37	*136.17	73.81	36.11	150.86	*11
136.87	208.40	16.74	400.13	383.39	190.60	87.91	367.71	*279.80	190.64	87.91	367.71	*279.80	197.71	96.73	404.08	*30
340.67	277.70	85.97	469.36	383.39	290.30	144.82	529.11	384.49	290.27	144.62	529.11	384.49	297.80	145.71	608.64	46
112.33	48.08	-143.60	239.77	383.39	87.00	33.82	188.99	*155.17	87.00	33.82	188.99	*155.17	91.18	44.61	186.36	*14
256.33	210.50	18.82	402.21	383.39	239.60	115.38	447.96	*332.58	239.64	115.38	447.96	*332.58	250.44	122.53	511.84	38
271.67	381.70	190.10	573.44	383.39	355.00	183.00	630.86	447.86	355.01	183.00	630.86	447.86	364.83	178.51	745.65	56
81.00	95.08	-96.62	286.77	383.39	83.08	31.94	181.84	*149.90	83.08	31.94	181.84	*149.90	81.21	39.73	165.98	*12
101.67	123.50	-68.18	315.21	383.39	121.30	50.93	250.01	*199.08	121.31	50.93	250.01	*199.08	122.36	59.87	250.09	*19
357.00	321.10	129.40	512.77	383.39	307.50	154.71	556.31	401.61	307.46	154.71	556.31	401.61	304.65	149.08	622.66	47
171.33	175.30	-16.44	366.95	383.39	182.30	83.35	353.91	*270.56	182.34	83.35	353.91	*270.56	182.99	89.53	374.00	*28
372.00	337.60	145.90	529.29	383.39	314.70	158.95	567.66	408.71	314.65	158.95	567.66	408.71	313.43	153.36	640.60	48
501.67	532.20	340.50	723.84	383.39	560.50	310.08	943.59	633.51	560.46	310.08	943.59	633.51	572.20	279.97	1169.46	88
264.00	246.00	54.30	437.69	383.39	246.20	119.11	458.53	*339.42	246.18	119.11	458.53	*339.42	247.58	121.14	506.01	38
427.00	439.60	247.90	631.25	383.39	443.10	236.62	766.49	529.87	443.06	236.62	766.49	529.87	434.82	212.75	888.68	67
730.67	736.10	544.40	927.81	383.39	753.10	434.26	1227.5	793.22	753.05	434.26	1227.47	793.22	767.72	375.63	1569.08	119
220.67	133.60	-58.14	325.25	383.39	151.40	66.64	301.81	*235.16	151.38	66.64	301.81	*235.16	157.89	77.25	322.69	*24
239.67	243.80	52.08	435.47	383.39	261.40	127.87	483.08	*355.21	261.44	127.87	483.08	*355.21	271.26	132.72	554.41	42
422.00	505.00	313.30	696.70	383.39	533.90	293.28	903.82	610.53	533.88	293.28	903.82	610.53	588.37	287.88	1202.52	91
199.00	203.90	12.16	395.55	383.39	215.10	101.50	407.99	*306.48	215.08	101.50	407.99	*306.48	221.45	108.35	452.60	*34
485.33	448.00	256.30	639.66	383.39	440.30	234.92	762.30	527.38	440.32	234.92	762.30	527.38	432.52	211.62	883.98	67
673.67	706.20	514.50	897.88	383.39	699.70	399.50	1149.6	750.09	699.74	399.50	1149.59	750.09	687.87	338.56	1405.87	106
176.67	258.90	67.23	450.62	383.39	237.80	114.32	444.95	*330.62	237.78	114.32	444.95	*330.62	251.57	123.09	514.17	39
501.00	534.30	342.60	725.96	383.39	516.20	282.14	877.21	595.07	516.17	282.14	877.21	595.07	503.81	246.50	1029.68	78
1010.0	894.50	702.80	1086.2	383.39	802.90	466.97	1299.9	832.89	802.86	466.97	1299.86	832.89	774.93	379.16	1583.81	120

* Shows the length values that are less than 383.39 (the confidence length of the untransformed data).

The second example is from a 2^4 factorial experiment on four control variables with three runs at each experimental point. The designs as well as the responses are shown in Table 3.

The estimates of the mean ($\bar{y}$) and standard deviation(s) at each run are obtained and these are used in finding the regression of $\ln(s)$ on $\ln(\bar{y})$ i.e.

$$\ln(s) = \alpha + \beta \ln(\bar{y})$$

The fitted model is

$$\ln(s) = -0.7028 + 1.001 \ln(\bar{y})$$

The test statistic $t = 553.36$ shows that β is significant, hence there is need for log-transformation since $\beta \neq 1$.

Table 3. Data from a 2^4 Factorial Experiment.

	A	B	C	D	y_1	y_2	y_3	$\bar{y}$	s
1.	A1	B1	C1	D1	121.55	242.22	362.65	242.74	120.56
2.	A2	B1	C1	D1	159.98	318.47	476.89	318.45	158.46
3.	A1	B2	C1	D1	132.59	264.70	396.53	264.61	131.97
4.	A2	B2	C1	D1	172.25	343.93	515.05	343.73	171.41
5.	A1	B1	C2	D1	146.70	293.21	439.28	293.06	146.29
6.	A2	B1	C2	D1	164.40	327.52	491.23	327.72	163.42
7.	A1	B2	C2	D1	146.26	291.85	437.05	291.67	145.40
8.	A2	B2	C2	D1	155.97	310.82	465.40	310.73	154.71
9.	A1	B1	C1	D2	170.07	339.67	509.06	339.60	169.50
10.	A2	B1	C1	D2	192.75	384.58	576.52	384.61	191.88
11.	A1	B2	C1	D2	145.56	290.94	435.34	290.61	144.89
12.	A2	B2	C1	D2	190.22	378.82	567.70	378.91	188.74
13.	A1	B1	C2	D2	153.13	306.40	458.47	305.99	152.67
14.	A2	B1	C2	D2	185.84	370.89	555.79	370.84	184.98
15.	A1	B2	C2	D2	159.26	318.22	476.30	317.92	158.52
16.	A2	B2	C2	D2	189.68	378.05	566.58	378.10	188.45

Though, all the three approaches (original data, log-transformation and GLM methods) produced almost the same mean responses, log-transformation and GLM methods are very close (Table 4). The 95% confidence intervals of all the methods contain the true means of the response variable. The confidence length shows that both the log-transformation and GLM performed better than the untransformed data in terms of level of precision in estimating the mean responses because of their shorter length of confidence intervals.

Fig. 5 shows the histogram of mean responses of the original data, while Fig. 6 shows the histogram of mean responses of the transformed data. The histogram of the transformed data gives an approximate normal distribution curve than the untransformed data. GLM and Log-transformed produced the same histogram.

Fig. 5. Histogram of the Original Data

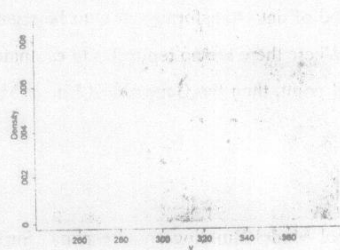


Fig. 6. Histogram of Log Transformed Data

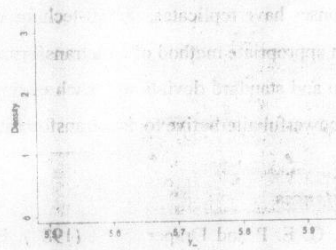


Table 4. The mean responses and confidence intervals (length)

TRIAL	UNTRANSFORMED					LOG-TRANSFORMED					GLM METHOD			
	TRUE	DATA				DATA								
	MEAN	MEAN	LCL	UCL	LENGTH	MEAN	LCL	UCL	LENGTH		MEAN	LCL	UCL	LENGTH
1	242.14	243.99	153.14	334.84	181.70	285.18	233.69	300.92	67.23		244.00	215.18	276.75	61.56
2	318.45	316.61	225.76	407.46	181.70	317.75	280.01	360.57	80.56		316.00	278.65	358.38	66.24
3	264.61	262.76	171.91	353.61	181.70	264.03	232.67	299.61	66.94		262.60	231.54	297.78	87.39
4	343.73	345.58	254.73	436.43	181.70	344.48	303.57	390.91	87.34		346.40	305.46	392.81	73.37
5	293.06	291.22	200.37	382.07	181.70	292.42	257.69	331.83	74.14		290.80	256.44	329.81	83.32
6	327.72	329.56	238.71	420.41	181.70	328.43	289.42	372.69	83.27		330.30	291.23	374.55	74.15
7	291.67	293.51	202.66	384.36	181.70	292.31	257.59	331.70	74.11		293.90	259.18	333.33	77.79
8	310.73	308.88	218.03	399.73	181.70	310.05	273.23	351.84	78.61		308.40	271.90	349.69	77.79
9	339.61	337.75	246.90	428.60	181.70	338.86	298.62	384.53	85.91		337.00	297.15	382.17	95.01
10	384.61	386.46	295.61	477.31	181.70	385.45	339.67	437.40	97.72		387.60	341.79	439.57	97.78
11	290.61	292.46	201.61	383.31	181.70	291.24	256.65	330.49	73.84		292.90	258.25	332.13	73.88
12	378.91	377.07	286.22	467.92	181.70	378.09	333.19	429.04	95.86		376.00	331.56	426.42	94.86
13	305.99	307.84	216.99	398.69	181.70	306.66	270.24	347.99	77.75		308.40	271.93	349.72	77.80
14	370.83	368.99	278.14	459.84	181.70	370.03	326.08	419.90	93.81		368.00	324.49	417.32	92.84
15	317.92	316.08	225.23	406.93	181.70	317.23	279.55	359.98	80.43		315.50	278.19	357.79	79.59
16	378.11	379.95	289.10	470.80	181.70	378.93	333.93	430.00	96.07		381.00	336.00	432.12	96.13

6. Conclusion

Illustrations have been made with examples that data transformation and Generalized Linear Model will often produce a better model in terms of potential performance in response estimation and prediction than what one obtained by using data that are not normally distributed.

Though β -technique method of data transformation produced a better model than Box-Cox method as well as Generalized Linear Model, the drawback it has is that the experimental points have to be replicated in order to have estimates of the mean and standard deviation.

It is therefore recommended that if experimental design points with non-normal responses have replicates, then β -technique method of data transformation is to be regarded as an appropriate method of data transformation. Where there are no replicates to estimate the mean and standard deviation at each experimental point, then the Generalized Linear Model is a powerful alternative to data transformation.

References

- Box, G. E. P and Draper, N. R (1987). Empirical Model-Building and Response Surface. John Wiley & Sons, New York, N.Y.

- Fear, T. R; Benichou, J; and Gail, M. H (1996). A Reminder of the Fallibility of the Wald statistics. *The American Statistician* 50, pp. 226-227.
- Hanck, W. W. Jr. and Donner, A. (1997). Wald's Test as applied to Hypotheses in Logit Analysis. *Journal of the American Statistician Association* 72, pp. 851-853.
- Koksoy, O. and Doganaksoy, N. (2003). Joint Optimization of Mean and Standard Deviation Using Response Surface Methods. *Journal of Quality Technology* 35, No. 3, pp. 243-244.
- Lewis, S. L; Montgomery, D. C and Myers, R. H (2003). Examples of Designed Experiments with Non-normal Responses. *Journal of Quality Technology* 33, No. 3.
- Miller, A. (2003). Analysis of Parameter Design Experiment for Signal-Responses Systems. *Journal of Quality Technology* 34, No. 2.
- Montgomery *et al* (2001). *Design and Analysis of Experiment* 5th Edition. John Wiley & Sons, New York, N.Y.
- Myers, R. H and Montgomery, D. C (1997). *Response Surface Methodology*. John Wiley & Sons Inc., New York, N.Y.
- Myers, R. H And Montgomery, D. C (1997). A Tutorial on Generalized Linear Models. *Journal of Quality Technology* vol. 29.
- Nelder, J. A and Wedderburn, R. W. M (1972). Generalized Linear Model. *Journal of the Royal Statistical Society A* 135, pp. 370-384.
- Taguchi, G. (1987). *System of Experimental Design*. Unipub/Kraus International Publications, White Plains, New York, N.Y.
- Vining, G. G and Myers, R. H (1990). Combining Taguchi and Response Surface Philosophies: A Dual Response Approach. *Journal Quality Technology*, 22.