ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1280

Remedy for Inadequacy of Assumptions in Statistical Models

Chris Uchechi Onyemachi[1]

[1]Faculty of Physical Sciences, Department of Statistics, Abia State University

Publication Date: 2026/11/02

Abstract: Models that specify a constant variance for the error terms perhaps form the "backbone" of statistical modeling as applied to much of scientific investigation. Error variance estimation is one of the most important issues in regression models. The estimation of the error variance is essential to assess the variability of the estimated mean of Y_i given X_i. Therefore, the error variance plays an important role in regression analysis but uncorrelated in sense that errors in successive measurements cannot be forecast from a knowledge of errors in previous measurements. One of the central messages of this study is that (a) unequal variances may be more prevalent than typically recognized in statistical analysis, and (b) the usefulness of "Levene's test when considering tests of equal variances, as test is actually a family of techniques. Depending on which of the Levene tests that are being implemented, and particularly the Levene's test based on means which is found in widely used software like SPSS, one may be using a statistical technique that is as bad (if not worse) than the F test which the Levene test was intended to replace. The primary goals of this study are to (a) demonstrate that the current statistical practice of testing for equality of variances in hypothesis testing (as prescribed by textbooks and statistical software programs) is insufficient, hence data transformation is suggested.

Keywords: *Constant Variance, Error Variance, Estimation, Levene's Test, F-Test, Transformation.*

How to Cite: Chris Uchechi Onyemachi (2026) Remedy for Inadequacy of Assumptions in Statistical Models. *International Journal of Innovative Science and Research Technology*, 11(10), 2484-2491. https://doi.org/10.38124/ijisrt/26oct1280

I. INTRODUCTION

Regression analysis is one of the most widely used methodological tools in applied statistics. The main aim of regression analysis is to find a general relationship between a response variable and one or more predictor variables. For example, in the regression analysis with one predictor variable, the aim is to estimate the unknown mean of the response variable for a given value of the independent variable. For given data (X_i, Y_i), $i = 1,2,\ldots,n$, which represent n values observed on the response Y, corresponding to the n values of the independent variable X. Relationship between X and Y can be modeled as.

$$Y_i = m(x_i) + \varepsilon_i \quad \text{for } i = 1,2,\ldots,n \tag{1}$$

Where ε_i denotes errors which are assumed to be independent and identically distributed random variables, while $m(x_i)$ represents the mean function, $E[Y_i \mid x_i]$.

The two most commonly used approaches to estimate the mean function in (1) are parametric and nonparametric methods. In the parametric approach, a specific functional form is assumed for the mean function. For instance, if we assume $m(x_i)$ to be a linear function, the model becomes

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \; i = 1, 2, \ldots, n \tag{2}$$

Where β_0 and β_1 are unknown parameters.

Among others [4] and [8] have presented in several books and literature developed theory for this type of model and for linear models in general. Efficient inferential procedures are available when the functional form of the relationship between the response and predictor variables complies with parametric assumptions.

➢ *Rationale*

The purpose of this study is on the inference concerning the regression function and the prediction of Y the response variable which requires an estimation of the variance σ^2 of the error terms ε_i in the model to obtain an indication of the variability of the probability distributions of Y.

Before a statistical analysis is performed on data, assumptions are made about the data driven by a model for the data which is simply a statement of those assumptions. An assumption of the distribution of the error terms is necessary in order to carry out statistical analysis. The basic assumptions entail:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1280

- The relationship between Y and X is linear or can at least be approximated by a straight line.
- The mean of the error terms $\varepsilon_i = 0$ i.e., $E(\varepsilon_i) = 0$
- The error terms are normally and identically distributed with mean equal to zero and variance σ^2
- The error terms have equal and constant variance
- The errors are independently distributed i.e., the errors are uncorrelated with each other. $E(\varepsilon_i \varepsilon_j) = 0$ for $i \neq j$.

This study is to provide an understanding of the importance of the role played by measurement uncertainty analysis in any test or experimental measurement process with the following objectives:

- To verify fundamental assumptions related to statistical tests
- To evaluate error variance estimation
- To evaluate the significance of a test's measurement uncertainty in order to better plan experiments and tests in order to save cost and proffer remedial measure where required assumptions are violated.

➢ *Conceptual Framework*

- *Error Variance Estimation*

Error variance estimation is one of the most important issues in regression models. The estimation of the error variance is essential to assess the variability of the estimated mean of Y_i given X_i. Therefore, the error variance plays an important role in regression analysis but uncorrelated in sense that errors in successive measurements cannot be forecast from a knowledge of errors in previous measurements.

- *Research Importance*

There are at least three possible occasions where testing for equality of variances are a concern:

The first is when one wants to make inferences about population variances because they are of scientific interest on their own. For example, a health researcher may be interested in studying the effects of a new drug that helps prevent mood swings on some members of a mood management program.

The researcher hypothesizes that the drug will decrease the severity of mood swings in patients. In this case, the researcher is interested in the overall increase or decrease in the severity of mood swings (operationalized as the change in the range of mood scores from high extremes to low extremes to more moderate shifts in mood) in which case a test for equal variances would be conducted to test for differences because those in the group that received the program would be hypothesizes to have less severity in the range of their scores.)

The second is when there is suspected heterogeneity of variances in a t- test or analysis of variance (ANOVA) in which not all of the factors have fixed effects and thirdly, when one might suspect heterogeneity of variances in a fixed-effect analysis of variance in which the numbers of observations in the groups are widely disparate. When there is a reasonable evidence suggesting that the variances of two or more groups are unequal, a preliminary test of equal variances is conducted prior to conducting the t-test or ANOVA.

II. MATERIAL AND METHODS

➢ *Transformation of Variables*

Transformations of statistical variables are used in applied sciences mainly for two purposes among others. The first one is for variance stabilization while the other is for normalization. Thus, a transformation is used to facilitate the tail sums of the distribution by the aid of the normal probability integral [1] posits. It is very useful to deriving a linear relationship between two variables, so that linear regression analysis can be applied. A review of the literature shows that a considerable number of transformations of binomial, negative binomial, Poisson and χ^2 variables have been proposed. For more references see [3], [2]. Further, examples of stochastic modeling using the truncated normal random variable include the modeling of demand functions in the determination of safety stocks by [7]. [6] studied the relationship between the periodic mean and variance and showed that a plot of the periodic means and standard deviations could determine the appropriate model to adopt in the analysis of time series model. They averred that if the periodic means appear to increase with the standard deviation or the variance, then a multiplicative time series model would be most appropriate. On the other hand, if the periodic means does not appear to increase or decrease with the standard deviation or the variance, then the additive model would be suggested. [12] noted that the analysis of seasonal time series with periodicity, s (length of the period interval) requires the arrangement of the series in a two-dimensional table called the Buys-Ballot table. With this type of arrangement, the within and between periods of the relationships were displayed. Similarly, the relationship within the periods shows the non-seasonal part of the series. The between periods represent the seasonal aspect. The intrinsic information contained in the Buys-Ballot table has attracted a lot of attention in Time Series analysis [2].

- *The Variance of the Transformed*

In data analysis transformation is the replacement of a variable by a function of that variable: for example, replacing a variable x by the square root of x or the logarithm of x. In a stronger sense, a transformation is a replacement that changes the shape of a distribution or relationship.

Reasons for using transformation

- ✓ Convenience
- ✓ Reducing skewness
- ✓ Equal spreads
- ✓ Linear relationships
- ✓ Additive relationship

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1280

- *Error Variance Estimation*

Error variance estimation is one of the most important issues in regression models. The estimation of the error variance is essential to assess the variability of the estimated mean of Y_i given X_i. Therefore, the error variance plays an important role in regression analysis but uncorrelated in sense that errors in successive measurements cannot be forecast from a knowledge of errors in previous measurements.

- *Brown-Forsythe Test*

The test which is a modification of the Levene test, does not depend on normality of the error terms. It is robust against serious departures from normality in the sense that the norminal significance level remains approximately correct when the error terms have equal variance whether the error terms distribution is far from normal. The BF consists of sample t test based on one-sided and two-sided tests concerning μ_1-μ_2 given by

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (3)$$

Where $\bar{d}_1$ and $\bar{d}_2$ are the sample means of the $\bar{d}_{i1}$ and $\bar{d}_{i2}$ respectively, and the pooled variance s^2 becomes

$$s^2 = \frac{\sum[(d_{i1} - \bar{d}_1) + (d_{i2} - \bar{d}_2)^2]}{n-2} \quad (4)$$

We denote the test for Brown-Forsythe test by t^*_{BF}.

If the error terms have constant variance and n_1 and n_2 are not extremely small, t^*_{BF}. follows approximately the t distribution with n-2 degrees of freedom. Large values of t^*_{BF}. indicate that the error terms do not have constant variance.

- *Batch Means*

The batch means methodology is frequently used to calculate Confidence Intervals for the mean μ of a stationary process $Y_1, \ldots, Y_n$ Wadsworth (1998).

Suppose $Y_1, \ldots, Y_n$ is partitioned into b non overlapping contiguous batches of m Y_k's each (assume that $n = bm$). So the i^{th} batch consists of the random variables $Y_{(i-1)m+1}, Y_{(i-1)m+2}, \ldots, Y_{im}$, $i = 1, \ldots, b$. The i^{th} batch mean is defined as the average of the m Y_k's from the i^{th} batch.

$$\bar{Y}_i = \frac{1}{m}\sum_{k=1}^{m} Y_{(i-1)m+k} \quad i = 1, \ldots, b.$$

The batches and associated batch means are given in the table below:

Batch 1	Batch 2	, . . . ,	Batch b
$Y_1, Y_2, \ldots, Y_m$	$Y_{m+1}, Y_{m+2}, \ldots, Y_{2m}$	, . . . ,	$Y_{(b-1)m+1}, Y_{(b-1)m+2}, \ldots, Y_{bm}$
$\bar{Y}_1$	$\bar{Y}_2$	, . . . ,	$\bar{Y}_3$

It is implicitly assumed that the batch means are approximately $i.i.d$ normal random variables with unknown mean μ and variance. We further define

$$\bar{Y} \equiv \frac{1}{n}\sum_{k=1}^{n} Y_k = \frac{1}{b}\sum_{i=1}^{b} \bar{Y}_i$$

And

$$V \equiv \frac{m}{b-1}\sum_{i=1}^{b} \left(\bar{Y}_i - \bar{Y}\right)^2$$

As point estimators for the mean μ and the variance parameter $\sigma^2 \equiv \lim_{n\to\infty} nVar(\bar{Y})$, respectively. Then standard statistical manipulations yield the following approximate $100(1-\alpha)\%$ CI estimator for mean μ:

➢ *Data*

Data was obtained from Innoson Nigeria Limited, located at Umudim Nnewi in Anambra State. Innoson manufactures vehicles and their product line includes heavy duty vehicles, middle and high level buses, special environmental friendly vehicles. The Company carries out optimization design and assembly according to African road condition so as to produce suitable products at affordable

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1280

prices (https:www.innosongroup.com). One of the major replacement parts is produced periodically in lots of varying sizes. The study focuses on one key input for the model to ascertain the optimum lot size by considering the relationship between lot size and man hour required to produce the lot in determining the optimum lot size for producing this part. To obtain this relationship, data on lot size and work hours for 25 recent production were used as presented in Table 1 below.

Table 1 Distribution of Lotsize/Workhour Innoson Group of Companies Nnewi January - June 2022

S/N	Lot-size	Work-hour	S/N	Lot-size	Work-hour
1	80	399	14	20	113
2	30	121	15	110	435
3	50	221	16	100	420
4	90	376	17	30	212
5	70	361	18	50	268
6	60	224	19	90	377
7	120	546	20	110	421
8	80	352	21	30	273
9	100	353	22	90	468
10	50	157	23	40	244
11	40	160	24	80	342
12	70	252	25	70	323
13	90	389			
TOTAL				1,750	7,807
MEAN				70.0	312.28

III. RESULTS AND DISCUSSIONS

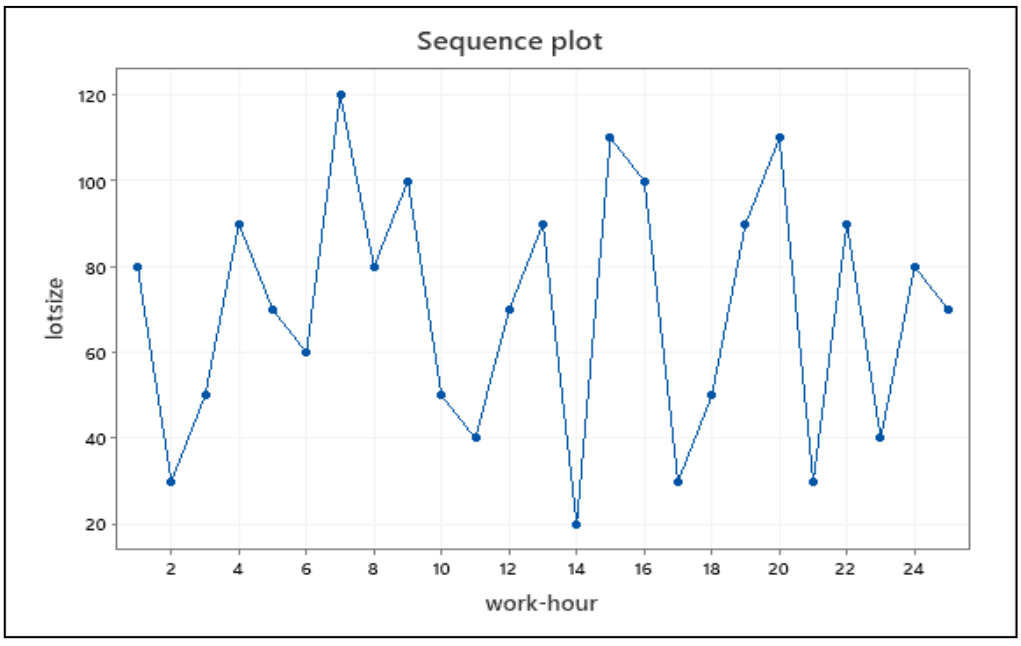


Fig 1 Time Plot of Lot Size and Man-Hour

➢ *Regression Analysis:*
The regression equation is Y = 62.37 + 3.570 X

- *Analysis of Variance*

Source	DF	SS	MS	F	P
Regression	1	252378	252378	105.88	0.000
Error	23	54825	2384		
Total	24	307203			

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1280

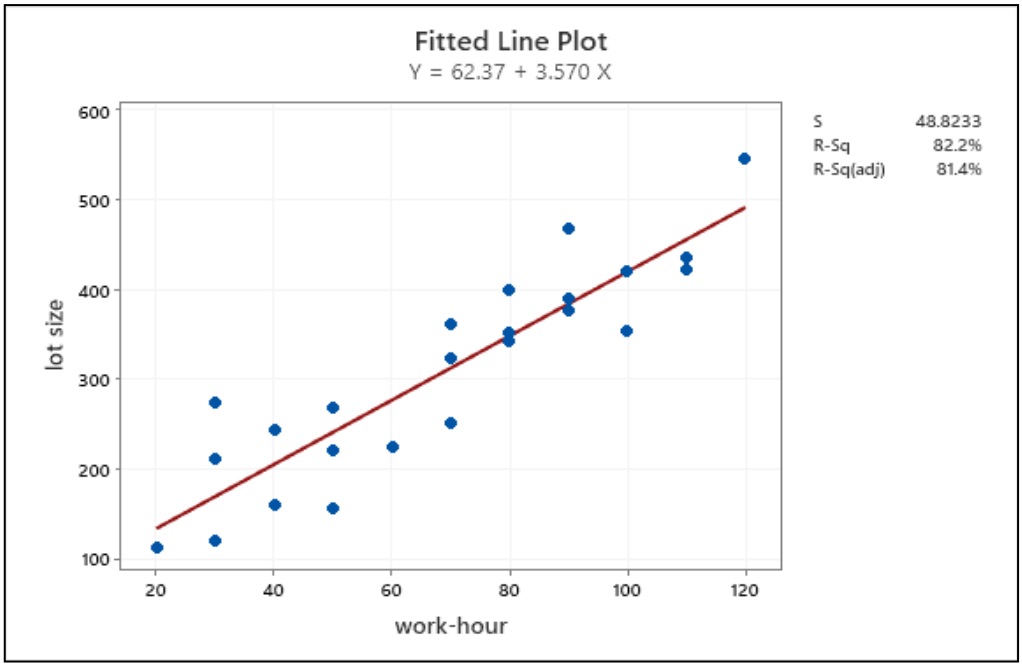


Fig 2 Fitted Line Plot of Lot Size Versus Work Hour

Table 2 Analysis of Lot Size/Workhour Innoson Group of Companies Nnewi January - June 2022

Lotsize x	**Work-hour** y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$\hat{y}$	$(y-\hat{y})=e_i$	$(y-\hat{y})^2=e_i^2$
80	399	10	86.72	867.2	100	347.97	51.03	2604.1
30	121	-40	-191.28	7651.2	1600	169.47	-48.47	2349.3
50	221	-20	-91.28	1825.6	400	240.87	-19.87	394.8
90	376	20	63.72	1274.4	400	383.67	-7.67	58.8
70	361	0	48.72	0.0	0	312.27	48.73	2374.6
60	224	-10	-88.28	882.8	100	276.57	-52.57	2763.6
120	546	50	233.72	11686.0	2500	490.77	55.23	3050.4
80	352	10	39.72	397.2	100	347.97	4.03	16.2
100	353	30	40.72	1221.6	900	419.37	-66.37	4405.0
50	157	-20	-155.28	3105.6	400	240.87	-83.87	7034.2
40	160	-30	-152.28	4568.4	900	205.17	-45.17	2040.3
70	252	0	-60.28	0.0	0	312.27	-60.27	3632.5
90	389	20	76.72	1534.4	400	383.67	5.33	28.4
20	113	-50	-199.28	9964.0	2500	133.77	-20.77	431.4
110	435	40	122.72	4908.8	1600	455.07	-20.07	402.8
100	420	30	107.72	3231.6	900	419.37	0.63	0.4
30	212	-40	-100.28	4011.2	1600	169.47	42.53	1808.8
50	268	-20	-44.28	885.6	400	240.87	27.13	736.0
90	377	20	64.72	1294.4	400	383.67	-6.67	44.5
110	421	40	108.72	4348.8	1600	455.07	-34.07	1160.8
30	273	-40	-39.28	1571.2	1600	169.47	103.53	10718.5
90	468	20	155.72	3114.4	400	383.67	84.33	7111.5
40	244	-30	-68.28	2048.4	900	205.17	38.83	1507.8
80	342	10	29.72	297.2	100	347.97	-5.97	35.6
70	323	0	10.72	0.0	0	312.27	10.73	115.1

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1280

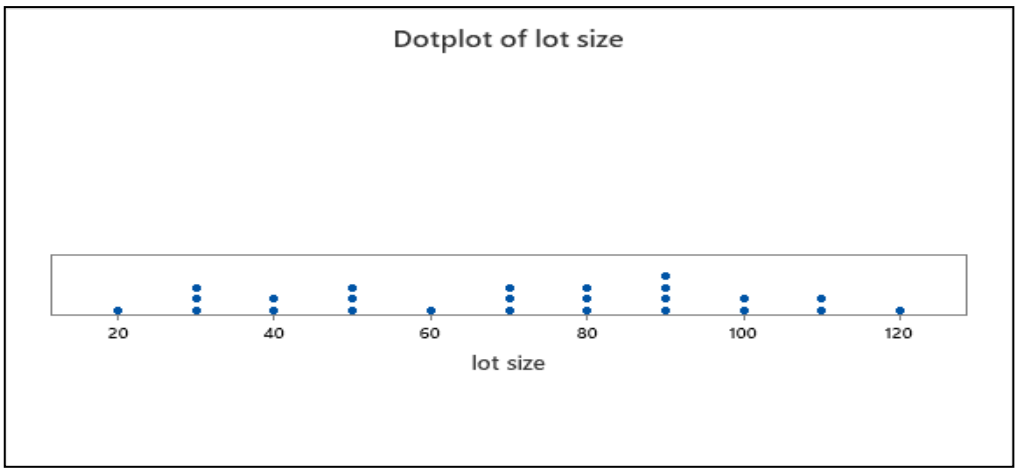


Fig 3 Dot Plot of the Distribution Lot-Size/Man Hour

➢ *Estimation of Error Terms Variance* σ^2

We calculate SSE for the Innoson Group of Companies based on the data. The residuals were obtained in Table 1. The table shows the squared residuals in column 5, from where we obtain these results:

Since 25 – 2 = 23 degrees of freedom are associated with SSE, thus

$$s^2 = MSE = \frac{54{,}825}{23} = 2{,}384$$

A point estimate of σ the standard deviation of the probability distribution of Y for X, is obtained $s = \sqrt{2{,}384} = 48.8$ hours.

➢ *Test for Constancy of Error Variance*

We shall employ the Brown-Forsythe test. The choice of this test is based on the fact that is does not depend on normality of the error terms. We wish to use the test to determine whether or not the error term variance varies with the level of X. since the X levels are spread fairly uniformly (see Figure 3 dot plot), we divided the 25 cases into two groups with approximately equal X ranges. The first group consists of the 13 runs with lot sizes from 20 to 70. The second group consists of the 12 runs with lot sizes from 80 to 120. Table 2 presents a portion of the data for each group. In columns 1 and 2 are repeated the lot sizes and residuals from Table 1. The median residual is $\bar{e}_1 = -19.88$ for group 1 and $\bar{e}_2 = -2.68$ for group 2 respectively. Obtained with MINITAB software.

Table 3 Calculations for Brown-Forsythe Test for Constancy of Error Variance

Group 1					
		(1)	(2)	(3)	(4)
i	Run	Lot Size	Residual e_{i1}	d_{i1}	$(d_{i1} - \bar{d}_1)^2$
1	14	20	-20.77	.89	1929.41
2	2	30	-48.47	28.59	263.25
…	…	…	…	…	…
12	12	70	-60.28	40.40	19.49
13	25	70	10.72	30.60	202.07
	Total			582.60	12566.60
		$\bar{e}_1 = 19.88$		$\bar{d}_1 = 44.815$	
Group 2					
		(1)	(2)	(3)	(4)
i	Run	Lot Size	Residual e_{i2}	d_{i2}	$(d_{i2} - \bar{d}_2)^2$
1	1	80	51.02	53.70	637.56
2	8	80	4.02	6.70	473.06
…	…	…	…	…	…
11	20	110	-34.09	31.41	8.76
12	7	120	55.21	57.89	866.71
	Total			341.40	9610.20
		$\bar{e}_2 = 2.68$		$\bar{d}_2 = 28.450$	

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1280

$$d_{11} = | e_{11} - \bar{e}_1 | = | -20.77 - (-19.88) | = .89$$

$$d_{12} = | e_{12} - \bar{e}_2 | = | 51.02 - (-2.68) | = 53.70$$

The means of the absolute deviations are obtained in the usual fashion:

$$\bar{d}_1 = \frac{582.60}{13} = 44.815 \qquad \bar{d}_2 = \frac{341.40}{12} = 28.450$$

Finally, column 4 contains the square of the deviations of the d_{i1} and d_{i2} around their respective group means. For instance, we have:

$$(d_{11} - \bar{d}_1)^2 = (.89 - 44.815)^2 = 1929.41$$

$$(d_{12} - \bar{d}_2)^2 = (53.70 - 28.450)^2 = 637.56$$

We now calculate the test statistic

$$s^2 = \frac{12566.60 + 9610.20}{25 - 2} = 964.21$$

$$s = \sqrt{964.21} = 31.05$$

$$t^*_{BF} = \frac{44.815 - 28.450}{31.05\sqrt{\frac{1}{13} + \frac{1}{12}}} = 1.32$$

To control the α risk at .05 significant level, we require $t(.975;23) = 2.069$. The decision rule therefore is:

If $| t^*_{BF} | \leq 2.069,$ conclude the error variance is constant

If $| t^*_{BF} | > 2.069,$ conclude the error variance is not constant

Since $| t^*_{BF} | = 1.32 \leq 2.069,$ we conclude that the error variance is constant and does not vary with the level of X. The two sided $P-value$ of this test is .20

WORKSHEET 1

Regression Analysis: Lot Size versus Workhour

Regression Equation

Lot = 64.14

Size + 5.55 Workhr

➢ *Coefficients*

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	64.14	7.88	8.14	0.000	
Frac Def	5.55	5.13	1.08	0.290	1.00

➢ *Model Summary*

S	R-sq	R-sq(adj)	R-sq(pred)
28.6191	4.86%	0.72%	0.00%

➢ *Analysis of Variance*

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	961.8	961.8	1.17	0.290
Frac Def	1	961.8	961.8	1.17	0.290
Error	23	18838.2	819.1		
Total	24	19800.0			

We fit a simple linear regression model with predictor variable Lot-size (the number of motor parts in a production run) and response variable Workhours (labour required to produce the lot of the motor spare parts).

➢ *Summary*

The study has shown the appropriateness of the simple linear regression model, the basic principles applied to statistical models using graphical methods. It also examined the aptness of the model before inferences are undertaken.

➢ *Recommendation*

We recommend thorough investigations of compliance of the basic assumptions in statistical modeling and recommend appropriate transformation to the dataset.

IV. CONCLUSION

The assumption of normally distributed population distributions is important for parametric statistics (e.g., t-test, test of equal variances based on the mean) and is based on a normal theory approach for conceptualizing variables that are

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1280

being studied. The normal (Gaussian) distribution approach is justified through the central limit theorem which basically states, in nonmathematical terms, that as the size of a sample of scores increases, the sampling distribution about the mean approaches the normal distribution. When distributions are non-normal (e.g., highly skewed, multimodal, or heavily-tailed), the ability to identify a viable probability distribution using the normal theory approach is reduced. This is because, under normal theory statistics, the probability distributions are symmetric and more or less bell-shaped and, when distributions are non-normal, especially in the ways listed above, the probability distribution is distorted which effects the estimation of means and variances leading to erroneous statistical results.

REFERENCES

[1.] Alessandro, B. (2014). Data Transformation for Confidence Interval Improvement: An application to the Estimation of Stress-Strength Model Reliability. Available at http://digitalcommons.wayne.edu/jmasm/vol14/iss2/15

[2.] Bartlett, M. S. (1947). The Use of Transformations. Biometrics. Vol. 3. Pp 39 - 52

[3.] Box, G. E. P. and Cox, D. R (1964) An Analysis of Transformation. Journal of Royal Statistical Society, Series B (Methodological), Vol. 26, No. 2, pp. 211-252.

[4.] Iwueze I.S (2007). Some Implications of Truncating the Distribution to the left at Zero. JASA, Vol 73, N0. 363, 536 - 544

[5.] Johnson, A. C. and Thomopoulos, N. T. (2002). Use of the Left Truncated Normal Distribution for Improving Achieved Service Levels, Decisions Sciences Institute 2002 Annual Meeting. Journal of Applied Sciences. 2007;7(2):189-195.

[6.] Wei, W. (1990) Time Series Analysis: Univariate and Multivariate Methods. Addison-Wesley Publishing Company, Inc., New York.