ISSN 1022-4734

JUJSS
Vol. 35, 2019, pp. 01-08

Construction of Tolerance Interval for Linear and Non-linear Regression Models

Md. Inzamamul Hoque
Lecturer, Safurannesa Mohila College, Satkhira, Bangladesh

Azizur Rahman
Assistant Professor, Department of Statistics, Jahangirnagar University, Bangladesh

Istiak Ahmed
Scientific Officer, BARI, Gazipur, Bangladesh
and
Mariam Akter
M. S. from Department of Statistics, Jahangirnagar University, Bangladesh

Abstract

This study is based on the linear and non-linear regression model to develop tolerance intervals. For the construction of tolerance interval first, we simulate the data series and then apply the package developed in R by Derek S. Young (1998) and then consider secondary data for empirical analysis. And the interval is developed for both linear and non-linear regression using R version 3.3.2. In all that case, we wish to be 95% confident that 95% of all observed responses are within certain limits for a given level of the predictor.

Keywords: Acceptance Limit, Coverage, Nonlinear Regression, Tolerance Interval.

1. Introduction

Tolerance intervals or enclosure intervals are similar to prediction intervals, but they cover a fixed proportion of the population. They are where we expect a certain population proportion to lie. For a particular confidence interval, it tells us lower and upper values which have a specific proportion (or percent) contained within them. They are lesser known relatives of confidence and prediction intervals. They can be very useful in many situations to make product or process quality assessments. Even for normally distributed data, their calculation is less trivial than confidence and prediction intervals, which makes them underutilized in practice. In addition, they are not always readily available in statistical software packages. As a result, there have been several approximate methods proposed in the literature to calculate them. The structure of the data and the assumptions made aspect the calculations of the desired tolerance intervals.

Based on this, the specific objective of the study is to construct tolerance interval i.e. we wish to be 95% confident that 95% of all observed responses are within certain

limits for a given level of the predictor both for linear and non-linear regression model. For the construction of tolerance interval first, we simulate the data series and then consider secondary data for empirical analysis.

2. Literature Review and Methodology

Linear regression models, that is, models that are linear in the parameters and/or models that can be transformed so that they are linear in the parameters. Basically with models that are linear in the parameters they may or may not be linear in the variables at a model that is linear in the parameters as well as the variables is a linear regression model and so is a model that is linear in the parameters but nonlinear in the variables. On the other hand, if a model is nonlinear in the parameters it is a nonlinear (in-the-parameter) regression model whether the variables of such a model are linear or not.

However, for some models may look nonlinear in the parameters but are inherently or intrinsically linear because with suitable transformation they can be made linear-in-the-parameter regression models. But if such models cannot be linearized in the parameters, they are called intrinsically nonlinear regression models (NLRM) (see Gujarati (2003)).

Burrows (1963) provides a general introduction to tolerance intervals and serves as a good starting point to understand the utility of tolerance intervals. Patel (1986) provides a review (which was fairly comprehensive at the time of publication) of tolerance intervals for many distributions as well as a discussion of their relationship with confidence intervals for percentiles and prediction intervals. One caveat with Patel (1986) is that there are some inconsistencies with the notation used, so it is best to refer back to the primary sources when studying the formulas. Many of the references cited within Patel (1986) were used in the development of the tolerance intervals discussed here. Finally, Krishnamurthy and Mathew (2009) provide one of the more detailed texts concerning the theory and application of statistical tolerance regions.

3. Tolerance interval for linear and non-linear regression

A tolerance interval is a statistical interval within which, with some confidence level, a specified proportion of a sampled population falls. More specifically, a [100×p%]/ [100×(1−α)] tolerance interval provides limits within which at least a certain proportion (p) of the population falls with a given level of confidence (1−α). A (p, 1−α) tolerance interval (TI) based on a sample is constructed so that it would include at least a proportion p of the sampled population with confidence 1−α; such a TI is usually

referred to as p-content (1−α) coverage TI. A (p, 1−α) upper tolerance limit (TL) is simply a (1−α) upper confidence limit for the (100×p) percentile of the population.

A tolerance interval can be seen as a statistical version of a probability interval. In the parameters-known case, a 95% tolerance interval and a 95% prediction interval are the same. If we knew a population's exact parameters, we would be able to compute a range within which a certain proportion of the population falls.

For example, if we know a population is normally distributed with mean μ and standard deviation σ, then the interval includes 95% of the population (1.96 is the z-score for 95% coverage of a normally distributed population).

Three types of questions can be addressed by tolerance intervals. Question (1) leads to a two-sided interval; questions (2) and (3) lead to a one-sided interval.

1. What interval will contain p percent of the population measurements?
2. What interval guarantees that p percent of population measurements will not fall below a lower limit?
3. What interval guarantees that p percent of population measurements will not exceed an upper limit?

As with tolerance regions for multivariate normal data, the calculated regression tolerance intervals can be overlayed on a scatterplot of the sample data using the **plottol** function in R.

All three regression settings we discuss can be plotted in a similar manner, provided that the regression model is a function of only one predictor.

4. Linear regression tolerance intervals

Suppose a quality engineer wishes to model the quality scores of small businesses as a function of the amount spent on program funding (in thousands of dollars). The engineer wishes to claim with confidence level $(1-\alpha)$, that a proportion P of all such businesses that spend a given amount are within certain limits on their quality scores. The calculation of two-sided linear regression tolerance intervals can provide such quantification for the engineer.

A multiple linear regression model is used to model the linear relationship between a response variable *Y* with a given set of predictor variables $X_1, X_2, \dots, X_p$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + l$$

where, $\beta_0, \beta_1, \beta_2, \dots, \beta_{p-1}$ are the p regression parameters and l is a normally distributed error term with mean 0 and variance σ^2. For a data set of size n, the estimated regression model (estimated using ordinary least squares) is given by

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_{p-1} x_{i,p-1} + l_i = \hat{y}_i + \hat{\alpha}_i \hat{l}_i k_{1,i}$$

where, the ℓ_i's are the residuals and the $\hat{y}_i$'s are the fitted values for this regression equation. Tolerance limits can then be constructed using these fitted values. $[100 \times (1-\alpha)\%]/[100 \times P\%]$ one-sided regression tolerance limits for each observation i are given by

$$L = \hat{y}_i - \hat{\alpha} k_{1,i}$$
$$U = \hat{y}_i + \hat{\alpha} k_{1,i}$$

respectively. $\hat{l}$ is estimated by the root mean square error and

$$k_{1,i} = \frac{t^*_{n-p;1-\alpha}\left(\sqrt{n_i^* z_p^*}\right)}{\sqrt{n_i^*}}$$

where, $t^*_{n-p;1-\alpha}(\gamma)$is the $(1-\alpha)^{\text{th}}$ quantile of a non-central t distribution with d degrees of freedom and non-centrality parameter γ, z^*_p is the p^{th} quantile of a standard normal distribution.

$$n_i^* = \frac{\hat{\sigma}^2}{s.e.(\hat{y}_i)^2}$$

Such that $s.e.(\hat{y}_i)$ is the standard error of $\hat{y}_i$, the value n_i^* is called effective numbers of observations, which means when n_i^* is divided into the variance of an observation,

then the result is the variance of the statistic. [100^$(1-\alpha)\%$] / [100^P%] two-sided regression tolerance limits for each observation i are given by

$$L = \hat{y}_i - \hat{\alpha} k_{2,i}$$
$$U = \hat{y}_i + \hat{\alpha} k_{2,i}$$

where, $k_{2,i}$ is estimated according to the formula in Krishnamoorthy and Mathew (2009).

Let f = n –p, then

$$k_{2,i} = \sqrt{\frac{f\chi^2_{1;p}\left(1/n_i^*\right)}{\chi^2_{f;p}}}$$

where, $\chi 2_{d;\alpha}(\delta)$ is the α^{th} quantile of a non-central $\chi 2$ distribution with d degrees of freedom and non-centrality parameter δ.

Returning to the example, suppose the quality engineer has data from n = 100 small businesses. The engineer wishes to be 95% confident that 95% of all such businesses that spend a given amount are within certain limits on their quality scores. The data were generated assuming $(\beta_0, \beta_1) = (20,5)$ and that the random error follows a normal distribution with mean 0 and standard deviation α = 3. Evaluation of 95%/95% two-sided regression tolerance limits is found by implementing the **regtol.int** function in R.

	α	P	y	$\hat{y}$	2-sided lower	2-sided upper
1	0.05	0.95	22.76428	21.33912	14.43466	28.24357
2	0.05	0.95	25.93468	21.72047	14.81959	28.62136
3	0.05	0.95	28.51155	22.64589	15.75344	29.53834
4	0.05	0.95	25.46921	24.39543	17.51807	31.27280
5	0.05	0.95	23.84511	25.98808	19.12346	32.85270

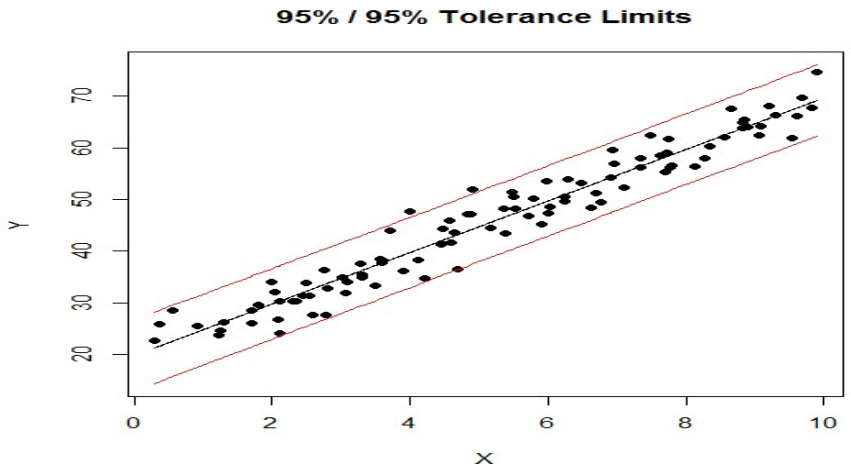

Figure 1. Tolerance intervals for linear regression

5. Nonlinear regression tolerance intervals

Suppose an engineer is dealing with a physical process which is known to have a specified nonlinear relationship. The engineer wishes to claim with confidence level $(1-\alpha)$ that a proportion P of all responses for a given level of the predictor is within certain limits. Calculation of two-sided nonlinear regression tolerance intervals can provide such quantification for the engineer.

A nonlinear regression model is used to model the nonlinear relationship between a response variable Y with a given set of predictor variables $X_1, X_2, \ldots, X_p$. A nonlinear regression model is defined as

$$y_i = f\left(\hat{\beta}, x_{i,1}x_{i,2} \ldots x_{i,p}\right) + \varepsilon$$

where, β is a vector of regression parameters and ε is an error term following a specified distribution which is not necessarily normal. For a data set of size n, the estimated nonlinear regression model is given by

$$y_i = f\left(\hat{\beta}, x_{i,1}x_{i,2} \ldots x_{i,p}\right) + l_i = \hat{y}_i + l_i$$

where, the ℓ_i's are the residuals and the $\hat{y}_i$'s are the fitted values for this regression equation. Tolerance limits are again constructed based on these fitted values. The

estimation done in the **nlregtol.int** function is through the nonlinear least squares routine. [100^$(1-\alpha)$%] / [100^P%] nonlinear regression tolerance limits are constructed in a similar manner as for the linear regression case. The only difference is how the effective sample size n_i^* is calculated.

For the nonlinear setting, n_i^* is a function of the partial derivatives of $f(\beta, x_{i,1}....x_{i,p})$ with respect to each of the regression parameters (i.e., the gradient of f(.)). The remaining formulas are the same. Further details can be found in Wallis (1946).

Returning to the example, suppose the physical process has the nonlinear relationship

$$Y = \beta_1 + (0.49 - \beta_2)l^{-\beta_2(X-8)}$$

where, Y is the response and X is some predictor. We wish to be 95% confident that 95% of all observed responses are within certain limits for a given level of the predictor.

The data were generated assuming $(\beta_1, \beta_2) = (0.39, 0.11)$ and that the random error follows a normal distribution with mean 0 and standard deviation 0.01. Evaluation of 95%/95% two-sided nonlinear regression tolerance limits is found by implementing the **nlregtol.int** function in R.

	α	P	y	$\hat{y}$	2-sided lower	2-sided upper
1	0.05	0.95	0.3950937	0.3763059	0.3875132	0.4227620
2	0.05	0.95	0.3942112	0.3763508	0.3833801	0.4210423
3	0.05	0.95	0.3946146	0.3812180	0.3678379	0.4213912
4	0.05	0.95	0.3959679	0.3838453	0.3693298	0.4226059
5	0.05	0.95	0.3976938	0.3881386	0.3711657	0.4242218

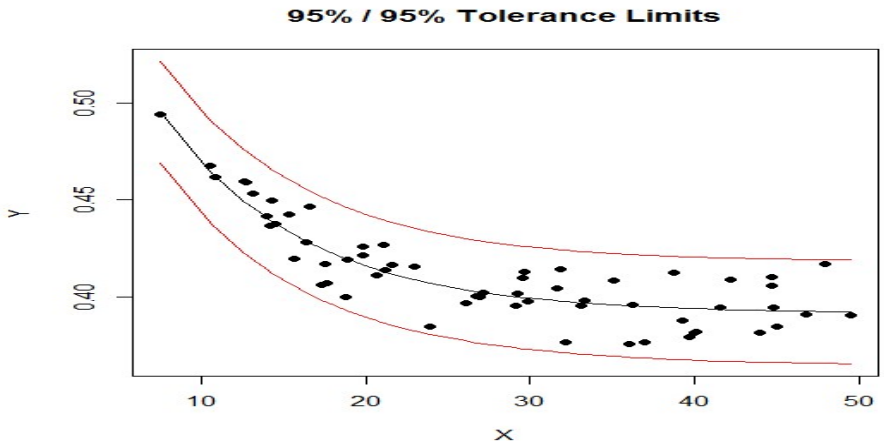

Figure 2. Tolerance intervals for non linear regression

Tolerance limits enjoy a fairly rich history in the literature and have a very important role in statistical applications. The tolerance package is applied to real data problem and provide a central collection of functions to estimate tolerance limits for some of the more frequent settings found in practice, including certain discrete and continuous univariate distributions, the multivariate normal distribution, and common regression settings.

References

Burrows G L (1963): Statistical Tolerance Limits – What Are They? *Applied Statistics*, 12, 133–144.

Derek S.,Young (1998): Tolerance: An R Package for Estimating Tolerance Intervals, *Journal of Statistical Software*, l 52, 34-44.

Gujarati, D N (2003): *Basic Econometrics*, 4th Edition. McGraw-Hill, New York.

Krishnamoorthy K, Mathew T (2009): *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley.

Patel J K (1986): Tolerance Limits – A Review. *Communications in Statistics: Theory and Methodology*, 15, 2719–2762.

Wallis W A (1946): Tolerance Intervals for Linear Regression. In J Neyman (ed.), *Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 43–51. University of California Press, Berkely, CA.