IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Statistical Modeling (Linear/Non-Linear) and Model Diagnostics

Nidhi Arora Dhingra,
Assistant Professor,
Department of Mathematics, Ramjas College,
University of Delhi, Delhi.

Abstract: Through this article, we intend to describe the process of modelling real world problems into known statistical models and understand how the solutions are obtained with the help of different statistical techniques.

1) Introduction: The statistical procedures have a well-defined path to follow:

(i) **Identification of the problem**: A process is observed, and a problem (objective) is identified. The process may or may not be generated by the observer. For example, a survey is conducted into a city, to study the effect of exposure of increased screen timings on health of school going children during Covid Pandemic.

(ii) **Collection of Data**: The data is gathered for the process.

(iii) **Analysis:**

(a) We usually start with plotting the data to find if any known patterns are being exhibited by the data. For this, either scatter plot or a histogram is used. Whereas scatter plot gives plot of the data on a continuous time scale, a histogram will plot it according to time intervals. A cumulative frequency curve can also be drawn, after arranging the data in ascending order.

(b) After this, descriptive statistics, like mean, variance, standard deviation, skewness etc. are computed. These are the estimates of the corresponding population parameters.

(c) If the data is a good quality data, then the conditions of **normality**, **randomness** and **independence** would be met, although to different extent, i.e., the error term will be distributed normally, a condition that can be tested by statistical tests, e.g., a *t*-test.

Randomness means that the selection of each unit is equally likely and there is no reason to expect preference for selection of one unit over another. A simple scatter plot would indicate at the (non)randomness by exhibiting (absence) presence of a trend.

Independence means selection of one unit will not affect selection of another and independence can be established by plotting the residuals. Generally, independence is not present, nor it is desired, particularly in case of time-series data.

Normality of Error term: Most of the statistical theory is built on the assumption of normality of error terms i.e. $\underset{\sim}{\mu} \sim N(\mu, \sigma^2)$. A normal distribution is signified by its bell shaped, symmetric curve. For $\mu = 0$ and $\sigma^2 = 1$, the distribution is a standard normal distribution. A distribution can be standardized by a transformation $Z = \dfrac{X-\mu}{\sigma}$.

In practice, the population parameters μ and σ^2 are not known. If σ is not known, it is replaced by its sample equivalent $s \left(= \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \right)$; however the resultant statistic is not normal but a *t*-statistic $\left(= \dfrac{\bar{X}-\mu}{s} \right)$ which is very similar to a normal variable and can be used to test if sample mean tends to population mean or not (in case of normal parent population).

2) Methods:

2.1 Sampling distributions of sampling statistics

All the sample statistics are random variables and hence have a probability distribution attached with them. Such a distribution is called a **sampling distribution**. So, for example, the distribution of sample mean $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ (or tends to normality even if the parent population is not normal). **Standard deviation** of sampling statistic is called the **standard error** of the statistic. A standardized *t*-statistic is defined as $t = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$, '*n*-1' is the number of "degrees of freedom" and signifies the number of free observations to make inference.

Similarly, for a parent normal population, the sampling distribution of the sample variance is a χ^2 distribution, i.e., $\chi^2 = \upsilon \frac{s^2}{\sigma^2} \sim \chi^2_{\upsilon}$; υ is the number of "degrees of freedom". The shape of a χ^2 curve depends on υ.

2.2 Testing of hypothesis, Significance levels and Confidence intervals

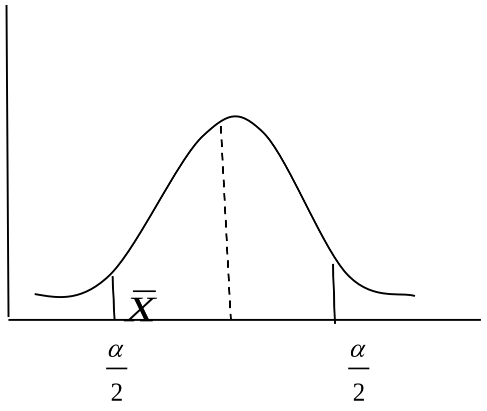
a) A **hypothesis** is a proposition that needs to be tested, e.g. we may want to test if the sample mean is a true representative of the population mean. We set the null hypothesis

$$H_0 : \bar{X} = \mu$$

The hypothesis is to be tested at a (specified) **level of significance** which is the largest probability of rejecting a true hypothesis. It is represented by α. So the test statistic is

$t = \frac{\bar{X}-\mu}{s} \sim t_{n-1,\alpha}$ if $H_1 : \bar{X} < \mu$ or $H_1 : \bar{X} > \mu$ i.e. a single tail test; and

$t = \frac{\bar{X}-\mu}{s} \sim t_{n-1,\frac{\alpha}{2}}$ if $H_1 : \bar{X} \neq \mu$, i.e. a two-tail test.



b) A **confidence interval** is a random interval which may contain the true value of the population parameter with a given confidence 1-α , so if $t = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1,\frac{\alpha}{2}}$, then $\mu = \bar{X} \mp t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ so a $(1-\alpha)\times 100\%$ confidence interval for population mean μ is $\left(\bar{X} - t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{X} + t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$.

Post the data collection, test statistics namely Sample Mean, Sample Variance are calculated and are treated as random variables, having a probability distribution. This paves the way of modelling the data. The data presented in the graphical form does not help much particularly in the scatter form. The data is adjusted by drawing a smooth curve known as **Smoothing**.

3) Model Diagnostics

Modelling is the science of using a common function to fit the data. It can be linear or non-linear. Also, it can be descriptive or mechanist. Model diagnostics seek to assess the validity of a model in several different ways. Multicollinearity and Autocorrelation are key issues in model diagnostics.

I) Multicollinearity

It means that there is a correlation among independent variables of our sample. This correlation can be invariably present in the data or may be inevitable theoretically.

Effect:

a) Regression Coefficients become Indeterminate.

b) Standard errors are not defined.

Methods to detect:

a) High values of R^2.

b) In case of two independent variables, correlation among the independent variables gives the idea of multi collinearity.

It needs to be taken care of statistically, if significant. It involves problem of less information, Addition of more data, Dropping of some variables (particularly collinear variable), Statistical techniques like Ridge regression and Lasso regression.

II) Autocorrelation

Autocorrelation is correlation among the terms of a time series data. It may be desirable theoretically. It invalidates the assumptions of normality or a t-test. It needs to be taken care of statistically. Smoothing of data is required. External reference distribution is carried out to identify the underlying distribution.

III) Heteroscedasticity

Heteroscedasticity is another key issue in assessing assumptions of the model. The problem of non-constant variance is mainly due to measurement errors or non-identical measurement scales for different variables. It is more prominent in non-normal data. It invalidates the set statistical tests. One of the methods to remove heteroscedasticity is power transformation model as it is suitable for any kind of statistical transformation and satisfies the conditions of normality and homoscedasticity simultaneously.

4) Regression Analysis

Generally, real life data is obtained in the form of numbers. While many analyses can be performed on numbers, many-a-times, it is required to find a common function to fit the data. This can be a simple linear function or an unsteady-state non-linear model. The process of finding such a function is called **model fitting**. A **model** can be a **descriptive** model or it can be a **mechanistic** model involving mathematical or statistical functions. In statistical modelling, regression analysis is a statistical technique of fitting a model. It can be linear or non-linear.

I) Linear Regression

The dependent variable is assumed to have a linear with the other independent variables. For example, in a two variable scenario: The variable X is assumed to be fixed (error-free) but the variable Y is a statistical variable, i.e. contains experimental errors. The errors are independent of levels of X and are independent among themselves also. The measurement errors are proportional to independent variable X.

Methods for solving Linear Regression (Two Variable)

a) Method of Least squares

b) Hypothesis testing

Multiple Linear Regression

In this scenario, there is one dependent variable Y(regressand) which is dependent on two or more explanatory variables(regressors). The two variable phenomena can be extended to three variables with new concepts of Partial correlation, Multiple correlation coefficient etc. Stepwise regression is a technique which retains only significant variables in the result and the insignificant variables are eliminated in each step. Structural equation Modelling is another technique of solving the multiple linear regressions.

II)Non-Linear Regression

In this form of regression, there is a functional relationship between the dependent and independent variables. The function can be exponential, logarithmic. A non-linear model involves non-linear parameters. Successive approximations are used to fit the data.

Common technique to solve the nonlinear regression models is the conversion into linear models wherever it is possible. The nonlinear least square method is also commonly used to solve the non-linear models.

For linear models, the joint intervals are typically symmetric. But the same may not be true for non-linear models, in which case the shape of confidence intervals may be irregular. If the size of the confidence intervals is small, it indicates that the true values of the parameters are likely to fall in a narrow range and hence the precision of the estimates is high (or the variance is low). The orientation and the shape of confidence intervals (region) may indicate about the precision of the estimates.

In general, the sizes of the confidence intervals decrease as the number of observations increase, but it also depends on the actual levels at which measurements are made. This is especially true for non-linear models.

5) Conclusion

There are various techniques available to model the data, depending on the size and parameters involved. The presence of software's like Excel, R have made it easy to analyze the data and summarize the results.

Regression types can be summarized as follows:

1. Linear parametric
2. Linear non-parametric
3. Non-linear
4. Semi parametric
5. Non-parametric
6. Multivariate regression
7. Generalized linear Models

Tools to solve the regression models:

1. Fitting distributions
2. Multivariate Analysis
3. Semi parametric methods
4. Non-parametric methods
5. Bayesian inference.

References:

1) Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons.
2) Damodar, N. G. (2004). *Basic Econometrics-Damodar N. Gujarati*. McGraw− Hill.
3) Eregno, F. E. (2014). *Multiple linear regression models for estimating microbial load in a drinking water source case from the Glomma river, Norway* (Master's thesis, Norwegian University of Life Sciences, Ås).
4) Berthouex, P. M., & Brown, L. C. (1994). Statistics for environmental engineers.