GENERAL LINEAR REGRESSION

Kirti Gadekar and Sharad Gore
J.J.T. University,
Vidyanagri, Rajasthan, India

Abstract : Linear regression is the most basic prediction model in Statistics. It includes simple and multiple linear regressions. Literature on regression includes several variations of regression. The most common variation is brought about by introducing non-linearity in the regression model. For instance, polynomial regression introduces non-linearity in predictor variables, while logistic regression introduces non-linearity in the response variable. A recent variation in the form of regression tree discretizes the response variable through partitioning data so that every leaf node of the regression tree has a distinct predicted value for the response variable. Some researchers have introduced the concept of link functions in order to develop paths. This approach has resulted in the development of structural equations modeling and path analysis.

This paper proposes a general form of the regression model. The main feature of the proposed model is linearity in coefficients and non-linearity, if any, of predictors in the form of link functions. Link functions are identified by examining the nature of the relationship between predictor variables and response variable.

The purpose of this paper is to present a unified regression equation that will be useful in all applications. Some illustrative examples are given to illustrate the procedure and discuss how the results can be interpreted.

Keywords : Regression , Simple Regression , multiple regression , linear regression , polynomial regression , logistic regression , regression tree , link function , regression path , path analysis.

I. INTRODUCTION

Regression is the most popular statistical model for predicting the response or outcome variable. Statistical literature is rich with a large number of research articles on various regression models. One reason for the large number of publications is the great variety of regression models. Even a casual literature survey can provide information on the variety of different types of regression models. Never the less, all the models aim at the following benefits

a) Regression analysis indicates significant relationships between the predictor variables and the response variable
b) Regression analysis indicates the strength of the impact that each predictor variable has on the response variable.

Regression analysis also allows comparisons between effects of predictor variables even when they are measured on different scales. It is therefore possible to drop or eliminate variables that are not really useful while identifying the best set of variables for building a predictive model.

When it comes to the type of model, three important considerations become the determining factor These three considerations are i) the number of predictor or independent variables in the model , ii) the shape of the regression line or the functional form of regression and iii) the nature of the response or dependent variable in the model. Most of the classical regression models are based on one or more of the above considerations in varying proportions. It is still possible to develop a new type of regression by using a new combination of the above considerations. However, it is necessary that the following seven commonly used regression models are well understood before an attempt is made to develop a new model. The seven most commonly used regression models are mentioned and briefly described below.

1. Linear Regression

Linear regression is one of the most popular techniques of predictive modeling. In its simplest form, apply known as simple linear regression. The linear regression is represented by a straight line. The regression line is optimal in the sense that it minimizes the total squared error of prediction. Linear regression in its more general form is multiple linear regression and it accommodates two or more independent or predictor variables under the following assumptions-

a) The variation in the response variable caused by every proctor variable is linear in nature,

b) The effects of different predictor variables on the response variable are added to obtain their combined, joint or total effect on the response variable . This property of multiple linear regressions is called the additivity property and the corresponding model is described as an additive model.

c) The effect of any particular prediction variable on the response variable is independent of other prediction variables.

These assumption impose some restrictions on the multiple linear regression model and must therefore be verified for validating the model . Voilation of any of these assumptions makes the model inappropriate and the results are not as good as desired or excepted. For example , if the effects of certain predction variables on the response variable are not independent of other prection variables, the problem of multicollinearity arises . If the prediction variables themselves have dependent observations , the problem of autocorrelation arises . If the variability of the response variable changes over its range , there is the problem of heteroskedasticity. The literature on regression addresses these possible anomalies in the linear regression model. For example , the problem of multicollinarity is addressed through ridge regression , lasso regression , or , more generally regression under regularization or shrinkage methods. An alternative way of addressing the problem of multicollinearity is to use forward selection , backward elimination, or stepwise approach in order to avoide overlap of effect of one prediction variable on that of other.

2. Logistic Regression

This regression model is appropriate when the response variable is binary , that is, when there are only two possible responses. In such cases , the more desirable response is called success so that the other response is called failure. Instead of predicting success or failure , the logistic regression model predicts the probability of success. Moreover , since this probability is restricted to the unit interval from 0 to 1 , the response variable is converted into the odds ratio p/ (1-p) that occupies the positive half of real line. In order to avoid situations where predicted values may not belong to this permissible range the odds ratio is further transformed through the logarithm function ln[p/(1-p)], so that the response variable becomes a real number . Once the response variable is modified as described above, the rest of the procedure is the same as that of linear regression. The following points are important in the context of logistic regression

(a) Since the original response variable is binary , the original problem can be called a classification problem . It is converted to a linear regression problem through the transformation from p to ln[p/ (1-p)]

(b) Logistic regression is not limited to a linear relationship between the prediction and the response variables. As a matter of the fact , the transformation of the response variable is non-linear and hence the logistic regression investigates a non-linear relationship between the predictor and the response variable.

(c) It is necessary to include all important predictor variables in the model in order to avoide under fitting . At the same time , it is also necessary to exclude predictor variables that have no significant effect on the response variables in order to avoid over fitting. This may best be achieved through a stepwise approach

(d) A sample size required for logistic regression is larger than that required for linear regression because maximum likelihood estimates have low power for small samples.

(e) The problem of multicollinearity can aries in case of logistic regression and has to be handled in the same way as in case of linear regression .

(f) Logistic regression is called ordinal logistic regression if the response variable is measured on the ordinal scale.

(g) If the response variable is not binomial , but has multiple possible values , then logistic regression is called multinomial logistic regression.

3. Polynomial Regression

The regression is said to be polynomial regression if predictor variables are raised to powers higher than 1 in the regression formula. Polynomial regression was the first non-linear form of regression. What may be interesting is to note that polynomial regression is non-linear in predictor variables, but is still linear in regression parameters. .the following points are important in the context of polynomial regression.

(a) Once it is decided to include terms that involve powers of predictor variables in the regression formula , it may be tempting to include terms with higher degrees in order to reduce error. This can however, lead to over fitting. It is therefore advised to plot data in order to get an idea about the reasonable values for the degree of the polynomial to be used in the regression formula. The guiding principle here is the principle of parsimony. If two models have similar performances, then the preference should be given to the simpler model. It is then called the parsimonious model.

(b) Care must be taken near the two extremes of the range of the values of the predictor variable. This is so because polynomials of higher degrees may exhibit weird behaviour when extrapolated beyond data range .

4. Ridge regression

When predictor variables are highly correlated , the data set is said to be suffering from multi co linearity the condition of multi co linearity does not influence the unbiased nature of the least square estimates of regression coefficients , but sampling variances of these estimates get inflated , resulting in a great loss of precision . Mathematically speaking , the ordinary least squares(OLS) estimates of the regression coefficients are given by the formula

$$\hat{\underline{\beta}} = (X^{l} X) X^{l} \underline{Y} \qquad (1)$$

where X is the dada matrix of predictor variables and $\underline{Y}$ is the data vector of the response variable.

When predictor variables are highly correlated, the matrix $(X^{l}X)$ is near singular.

Ridge regression introduces a ridge parameter Δ and modifies the OLS estimator to

$$\hat{\underline{\beta}}_{\Delta} = (X^{l} X + \Delta I)^{-1} X^{l} \underline{Y} \qquad (2)$$

It is obvious that the two estimates of egression coefficients are no more unbiased. An alternative definition of ridge regression without a ridge parameter is linear regression model that minimizes the total squared prediction error subject to the condition that

for $\Sigma \beta_i^2 < C$ some positive constant C

The last condition is called a regularization condition and its effect is restricting regression coefficients within a hyper share with radius C

Following points are important for ridge regression

(a) Ridge regression makes the same assumption that linear regression makes , except for the assumption of normality.

(b) Ridge regression shrinks the values of regression coefficients but does not reduce them to zero. As a result , ridge regression does not lead to variablr selection that can avoide multycollinearity.

(c) Ridge regression is called a regularization method and it uses L_2 regularization because it controls the L_2 norm of the regression coefficients .

(d) Depending on the form of the ridge regression model , the parameter Δ or C is known as the shrinkage parameter . It is also called the biasing parameter due to the fact that it causes the estimates of regression coefficients to be biased.

5. Lasso Regression

Lasso Regression is similar to ridge regression except for the fact that lasso regression results in selection of variables as a result of regularization. The name is the acronym of the descriptive name "Least Absolute Shrinkage and Selection Operator" . It is obtained by subjecting the regression coefficients to the linear constraint

$$\Sigma| \beta_i |< C \text{ for some } C > 0$$

The constraint is also called the penalty .Lasso regression uses the L_1 norm as penalty while ridge regression uses the L_2 norm as penalty. As a result of L_1 norm, lasso regression reduces some regression coefficients to zero , leading to removal of the corresponding variables from the model. Predictor variables that have non-zero coefficients in the lasso regression are the variables that are selected for inclusion in the model. Larger the penalty applied , closer the estimates get to zero.

Following points are important in the con text of lasso regression

(a) Lasso regression has same assumptions as linear regression except for the assumption of normality.

(b) Lasso Regression shrinks some of the regression coefficients to zero, thus effecting selection of predictor variables for inclusion in the model.

(c) Lasso Regression is a regularization method that uses the L_1 norm for regularization

(d) If there is group of highly correlated predictor variables, lasso regression retains only one of these variables in the model and shrinks coefficients of others to zero.

6. Elastic Net Regression

Elastic Net Regression combines the regularization conditions of ridge regression and lasso regression. More precisely Elastic Net Regression minimizes the total squared error subject to the following two constraints.

Constraint 1 : for $\Sigma| \beta_i |< C_1$ for some constant $C_1 > 0$

Constraint 2 : for $\Sigma \beta_i^2 < C_2$ for some constant $C_2 > 0$

Elastic Net Regression is supposed to inherit benefits of both ridge regression and lasso regression , while avoiding disadvantages or limitations of any one of the two . However , Elastic Net Regression appears to be more of a theoretical interest than of much practical use. The only mentionable application of elastic

Following points are important when one consider Elastic Net Regression

(a) Elastic Net Regression allows group effect of predictor variables when there are highly correlated predictor variables.

(b) Variable selection does not require iterations as are required instepwise methods in linear regression

(c) Elastic Net Regression can be badly affected by double shrinkage through L_1 norm as well as L_2 norm

7. Quantile Regression

Quantile Regression is an extension of linear regression to be used in the presence of outliers , high degree of skewness and heteroskedasticity. The objective of the quantile regression is to predict the specified quantile of the response variable instead of predicting it's arithmetic mean. In particular, median regression is a quantile regression model . since quantiles are partition values of the distribution of the response variable, quantile regression is not sensitive to presence of outliers , deviation from normality or heteroskedasticity. Quantile regression is very useful for describing the distribution of the response variable when it is known to be non-normal and hence cannot be described by only the mean and the variance. Quantile regression can be useful in estimating the average income of low income group since it is known that the income distribution is not normal.

8. Principle Component Regression (PCR)

Principle component regression (PCR) is used in presence of multicollinearity or even when the number of predictor variables is too large . since principle components are independent of one another ,PCR has no problem of multicollinearity. Also, since principle components reduce the dimentionality of a data set ,PCR also achieves reduction in dimentionality of data. PCR requires computation of principle components before fitting the regression model, and hence is a two step procedure. It is important to note that even though PCR uses only a few principle components in the regression model, these principle components are obtained from all the predictor variables in the given data set. As a result PCR is a featured extraction method and not feature selection method. As a consequence, PCR cannot provide any information on which predictorvariables are more dominating in their effect on the response variable.PCR cannot even find the extent to which any particular predictor variable affects the response variable. Nevertheless , except for these two drawbacks,PCR provides multicollinearity without regularization.

9. Support Vector Regression (SVR)

Support Vector Regression (SVR) is a very recent development. SVR uses support vector machines as it's basis and modifies the problem of classification of observations when the response variable is categorical or discrete. SVR uses the same method except for the fact that the response variable is continuous. Nevertheless the main feature of maximum margin is maintained in SVR .

It is interesting to note that SVR takes care of multicollinearity, if it is present among the predictor variables, through SVM that uses only important variables and ignores redundant variables. It is interesting to note that SVR also caters for non-linearity in the response variable without involving any non linear function of a predictor variable.

It is however important to also note that SVR being a recent development not much is known about its limitations or drawbacks. It is getting more attention from the machine learning community than from statistics community.

The purpose of describing all these regression models in order to establish a need of unifying the variety of models, so that every model can emerge as the most appropriate case of the unified model. The major advantage of having unified model is uniformity of processing rather than uniformity of model components as can be found in all the models that are in use. The second point of consideration is the fact that linear regression over-emphasizes normality of the distribution of residuals. There is no need for any consideration of this distribution as long as model fitting and predicting values of the response variable are concerned. It is only when some test of significance are to be carried out that the distribution becomes relevant.

II. General Linear Regression

The purpose of this paper is to propose a general form of linear regression where regressors can be functions od predictor variables, that is need not be predictor variables themselves. Consider the sitation where Y is the response variable and X_1 , X_2 ,….. X_p are the P predictor variables for some positive integer P. let G_1 ,G_2 , …..G_k be function of P predictor variables. Denoting the P -dimentional vector (X_1 , X_2 ,….. X_p) by $\underline{X}$, we write $G_1(\underline{X})$,$G_2(\underline{X})$,…..$G_k(\underline{X})$ and use them as regressors in the general linear regression , written as follows

$$Y = \beta_0 + \beta_1 G_1(\underline{X}) + \beta_2 G_2(\underline{X}) + \ldots + \beta_k G_k(\underline{X}) \qquad (3)$$

The positive integer k has no relation with dimension P of predictor variable space. When k < P, we have a sparse regression model that can accomodate the case of variable selection in presence of multicollinearity, when k = P , the multiple linear regression comes out as the special case with

$$G_i(\underline{X}) = X_i \text{ for } i = 1, 2, \ldots P$$

This case is also called the case of canonical regression model, when k > P, it is possible to incorporate interaction terms as well as polynomial terms if linear regression is not adequate to describe the relationship of the response variable with the predictor variables. In any case, the general linear regression is linear in parameters as well as linear in regressors, eventhough the regressors themselves can be non-linear functions of the predictor variables.

The proposed model can accommodate discrete as well as continuous predictor variables. It can also cover regression trees by identifying corresponding $G_i(\underline{X})$ as indicator functions. The proposed model includes the polynomial regression by defining some of the $G_i(\underline{X})$ as powers of one of the variables. Interaction terms involve cross product terms in some of the $G_i(\underline{X})$.

III. Fitting General Linear Regression

The method of fitting the general linear regression model is stepwise method. The first step involves the canonical form and hence involves each predictor variable separately . a scatter plot of the two variables Xi (predictor) and Y (response) is drawn in order to identify the curve of best fit so that the corresponding $G_i(\underline{X})$ can be identified with help of this scatter plot. After all predictor variables are accommodated with appropriate powers two predictor variables are taken at a time for identifying the optimal regressor in the model. It is also possible to use other graphical methods of visualizing multivariate data for identifying the most appropriate regressors in the model. The general linear regression model can further be modified by regularization. The proposed model aims at unifying the large variety of regression models that have appeared in the literature. The research is going on and more results will be published soon. For the time being, the current paper is concluded with an illusrttive example given in the next section.

IV. Illustrative Example

The illustrative example is constructed using the following R script. The output is too large to include here. The script given here can be implemented to find use of the proposed general linear model.

```
x=matrix(0,nrow=100,ncol=5)
x[,2]=rnorm(100,2,4)
x[,3]=rnorm(100,3,4)
x[,4]=rnorm(100,1,9)
x[,5]=rnorm(100,0,4)
x[,1]=7+3*x[,2]-5*x[,3]+2.5*x[,4]-3.5*x[,5]+x[,2]*x[,3]+x[,4]*x[,5]+2*x[,4]^2
m1=lm(x[,1]~x[,2])
m2=lm(x[,1]~x[,3])
m3=lm(x[,1]~x[,4])
m4=lm(x[,1]~x[,5])
m5=lm(x[,1]~x[,2]+x[,3]+x[,4]+x[,5])
m6=lm(x[,1]~x[,2]+x[,3]+x[,4]+x[,5]+x[,2]*x[,3])
m7=lm(x[,1]~x[,2]+x[,3]+x[,4]+x[,5]+x[,4]^2)
m8=lm(x[,1]~x[,2]+x[,3]+x[,4]+x[,5]+x[,2]*x[,3]+x[,4]^2)
g1=x[,2]
g2=x[,3]
g3=x[,4]
g4=x[,5]
g5=x[,2]*x[,3]
g6=x[,4]*x[,5]
g7=x[,4]^2
m10=lm(x[,1]~g1+g2+g3+g4+g5+g6+g7)
m11=lm(x[,1]~x[,2]+x[,3]+x[,4]+x[,5]+x[,2]*x[,3]+x[,4]*x[,5]+x[,4]*x[,4])
summary(m1)
summary(m2)
summary(m3)
summary(m4)
summary(m5)
summary(m6)
```

```
summary(m7)
summary(m8)
summary(m10)
summary(m11)
plot(x[,2],x[,1])
plot(x[,3],x[,1])
plot(x[,4],x[,1])
plot(x[,5],x[,1])
plot(x[,2]*x[,3],x[,1])
plot(x[,4]*x[,5],x[,1])
```

REFERENCES

1. Chatterjee, S., Hadi, A. S. and Price B(2000). Regression Analysis by Example , 3rd edition, John Wiley and sons , New York.
2. Kutner,, M.H. ,Nachtcheim C.J. ,and Neter J.(2004) Applied Linear Regression Models , 4th edition Mc Graw- Hill. Irwin, Chicago.
3. Fan, J.(1996) Local Polynomial Modeling and its Applications: from linear regression to non-linesr regression .Monogrphs on statistics and Applied Probability. Chapman and Hall/CRC
4. Hoerl ,A. E. and Kennard, R.(1970). Ridge Regrssion : Biased estimation for non orthogonal problems. Technomrtrices, vol – 12,pp – 55-67.
5. Zou, H. and Hastie, T.(2005) Regularization and variable selection via the elastic net ,Journal of the Royal Statistical Society, Ser.B. Vol – 67, pp=301-320
6. Hosmer Jr, D.W., Lemeshow ,S. and Sturdivant, R.X., (2013) Applied Logistic Regression, Volume 398, John Willy and Sons
7. Osborne, M.R., Presnell, B.,and Turlach, B.A.,(2000). On the lasso and its dual. Journal of Computational and Graphical Statistics,Vol.-9,No-2, pp.319-337
8. Rao,C.R.,(1973) Linear Statistical inference and its Applications, John Willy and Sons
9. Tibshirani ,R.(1996) Regularised Shrinkage and selection via the lasso,Journal of the Royal Statistical Society, B., Vol-58, No-1,pp 267-288
10. Gunn,S.R.(1998) Support V ector Machines for Classificaton and Regression,I.S.I.S. Technical Report
11. Smola,A. J., and Scholkopf,B.,(2004)A tutorial on Support Vector regression, Journal of Statistical Computing, Vol-14, No-3,pp 199-222
12. Gelman ,A., and Hill ,J., (2007) Data Analysis using Regression and Multilevel/ Hierarchical Models. Cambridge University Press, New York ,pp 79-108