Pakistan Journal of Applied Sciences 3 (4): 231-235, 2003
ISSN 1607-8926

Choice of Using Appropriate Statistical Technique in Data Analysis

Yousaf Hayat, [1]Qamaruz-Zaman and [2]S.M. Suhail
Department of Maths/Stat/Computer Science, NWFP Agricultural University Peshawar, Pakistan
[1]Department of Statistics University of Peshawar, Pakistan
[2]Department of Live Stock Management and Animal Breeding, NWFP,
Agricultural University Peshawar, Pakistan

ABSTRACT

Statistics is one of the most important discipline used in Basic and Applied research including Agricultural Sciences, Biological, Medical, Engineering and Social Sciences etc. Most of the researchers are using the statistical techniques in a wrong way because of the lack of knowledge about these techniques. In this paper an effort has been made to discuss an appropriate statistical techniques used in agricultural research. The most commonly used techniques are Regression analysis and Experimental Designs.

Key words: Expected mean square, regression and correlation, experimental designs

INTRODUCTION

Statistical techniques are greatly used in Agricultural, Medical, Engineering, Social, Biological and Physical Sciences etc. Since all the research decisions and findings based on these techniques. Therefore careful statement of the problem and use of proper methods in data analysis must be made to come up with meaningful results. Use of inappropriate statistical technique can lead to wrong interpretation of the data. In this paper an effort has been made to discuss an appropriate statistical techniques used in agricultural research. The most useful techniques used are:

1. Regression and Correlation Analysis.
2. Basic Experimental Design.

Regression analysis is used to establishing the actual relationship between two or more variables, for example the grain yield depends on rainfall or different doses of fertilizers etc, the milk production of animal depends on feed, the systolic blood pressure depends on age etc (Agarwal, 1991). All these problems are dealt with simple linear regression analysis. After fitting the regression line we can estimate (forecast or predict) the fitted values of dependent variable for the fixed value of independent variable. Some times it is relevant to check the goodness of fit of the regression model. This is dealt with the coefficient of determination (R^2). The high value of R^2 determines that the fitted model is appropriate. But it is not necessary that for high value of R^2 the model will be appropriate, because the value of R^2 inherently increases with the increase of independent variables in the regression model. So in such a case it is suggested to use the criteria of R^2-adjusted instead of R^2, because R^2-adjusted is the rescalling of "R^2"by degrees of freedom so that it involves a ratio of mean squares rather than sum-of-squares and is given by the relation: R^2-adjusted = 1- MSE/MS (total).

Before applying this technique (linear regression) the following assumptions must be checked:

I). The variance of Yi's (dependent variable) at each Xi is the same i.e. it has constant variance (σ^2).
ii). The Yi's (dependent variable), for each Xi's are normally distributed. Where Xi's are fixed and Yi's are random variables.
iii). The regression line Y = a + bX will pass through the mean (X, Y).

If the above three conditions are not fully satisfied then simple linear regression technique will not be an appropriate. In this case the conclusion drawn about the parameters must be inappropriate and will give misleading results.

The best way to check all these assumptions is, to plot a Scatter diagram of the observed data. If this diagram gives us a straight line then we can use simple linear regression line of Y on X but if the plotted line is not straight (linear) and the value of R^2 is too large, in such a case the interpretation will lead missing results by applying simple linear regression technique. So, in this case the simple linear regression line will not be used for prediction purpose, because it will yield suspicious results.

If the line is not linear then it is advised to use quadratic, cubic or exponential curve depending on the relationship between X and Y obtained from scatter diagram. In this case we cannot say that there is no relationship between the variables because in such a case there will exist some relation between the variables but not linear. On the other hand if we have a multiple regression ($Y = \alpha + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k + \epsilon_i$) problem, like the grain yield of a crop depends on amount of rainfall and different doses of fertilizers etc. In such a case the following assumptions must be satisfied before applying the ordinary least square method, to estimate various number of parameters.

1. The residual terms ϵ_i and ϵ_j are independent of each other i.e. Cov $(\epsilon_i, \epsilon_j) = 0$.
2. The residual term ϵ_i has zero mean for all I. This implies that for given X_i's, $E(Y_i) = \alpha + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k$.
3. Var $(\epsilon_i) = \sigma^2$ for all I, i.e. the variance of error terms is constant.
4. E $(X, \epsilon_i) = 0$ for all regressors, i.e. ϵ and X variable are independent.
5. ϵ_i's are normally distributed with a mean of zero and a constant variance σ^2.
6. We assume further in a multiple regression model that there exist no exact linear relationship between any two of the regressors i.e. $E(X_i, Xj) = 0$

If any of one of the assumption is violated, the least square principle will yield wrong results i.e. if Cov $(\epsilon_i, \epsilon_j) \neq 0$ the problem will be dealt through autocorrelation i.e. first it is necessary to detect autocorrelation and then use the method of least square for estimation of parameters. If for all I the variance of error terms is not constant, the problem will be dealt with hetroscadasticity, if for all regressors E $(X, \epsilon_i) \neq 0$ this means that there exist error in variables. If the error of measurements are found in the explanatory variables but not in response (dependent) variable, it is better to use inverse least square method for the estimation of parameters. Similarly, if E $(X_i, X_j) \neq 0$ then this indicates that there is a problem of multicollinearity. So in this case it is required to remove the problem of multicollinearity from the data.

The term coefficient of correlation is used in a bivariate situation, which is used to measure the strength of association between two random variables. To be kept in mind that in regression problem Y is a random variable while X is taken as fixed variable (non-stochastic variable) but in correlation problem both the variables are random. The Spearman's coefficient of correlation is used to measure the strength of linear association between two variables. In certain situation the relationship between two variables is not linear (quadratic or curvilinear) but we conclude on the basis of calculated value that there is no relationship between two variables. This interpretation of the term correlation coefficient is wrong because in such a case there will exist relationship between variables, which will either be quadratic or curvilinear (the relationship is not linear).

In certain situations the response (Y) has more than one values for one value of X, in this case the simple correlation coefficient is not correct to use because these observations are not independent. The coefficient of correlation is not appropriate for comparing alternative methods of measurement of the same variable because it assesses association not agreement. It can also give grossly misleading results if we are relating change overtime to the initial value. Great care should be taken in comparing variables which both vary with time because it is easy to get apparent association which is spurious.

There are certain situations in which the response is dummy (dichotomous), in such a case the problem will not be dealt with usual regression model but it should be dealt with Linear Probability Model (LPM), Logit Model (LM) or Probit Model (PM).

Use of experimental designs

Once the experimental results are obtained they have to be analyzed and interpreted. In experimental situations we may have large number of treatments. Our interest is to test whether all the treatments have the same effect. In other words, our intention is to test the null hypothesis, H_0: $\mu_1 = \mu_2 = ... \mu_t$, against the alternative hypothesis, H_1: $\mu_1 \neq \mu_2 \neq ... \neq \mu_t$. For this purpose we may use Student's t-test by taking all possible combinations. But it is a tedious

procedure and theoretically not valid. The appropriate method for such tests is the analysis of variance (experimental design).

The analysis of variance is the systematic algebraic procedure of decomposing the overall variation in the responses observed in an experiment into different components. Each component is attributed to an identifiable cause or source of variation. The structure of these component part is determined by the design of experiments.

There are many types of experimental designs. They can be broadly classified as single factor experiments and multi-factor (factorial) experiments. The single factor experiments are grouped as complete block designs and incomplete block design.

When the treatments consist different levels of a single variable factor and all other factors are kept at a single prescribed level, it is known as a single factor experiment. For example, in fertilizer trials several rates of single fertilizer element, say nitrogen, may be tested. All other factors such as agronomic practices, water management, insect control, etc., are kept at a uniform level.

The second misuse of statistical technique is that of experimental design in agricultural experiments. The simplest design used in agricultural experiments is Completely Randomized Design (CRD). Completely randomized design is the basic single-factor experiment. All other designs like randomized complete block design and Latin square design stem from it by imposing restrictions upon the allocation of the treatments within the experimental material.

The CRD is appropriate only when the experimental units are homogeneous. In field experiments there is generally large variation among experimental plots due to soil heterogeneity etc. Hence CRD is not preferred in field experiments. In laboratory experiments and greenhouse studies, it is easy to achieve homogeneity of experimental materials. Therefore, CRD is most useful in such experiments.

This design is misused in nested data (hierarchical classification) because of taking wrong value of denominator of an F-ratio. Most of the researchers are taking MSE as denominator but this is actually wrong. In such a situation it is advised to a researcher, to take care of a proper F-test. To take a proper F-test we consider the expected values of the mean squares under the different assumptions about the treatment, given in Table 1. In case of CRD (nested) F_T estimate $(\sigma^2 + r\sigma^2_\epsilon + nr\sigma^2_\alpha)/(\sigma^2 + r\sigma^2_\epsilon)$ under random effect model, while F_T estimate $\{\sigma^2 + r\sigma^2_\epsilon + nr\sum\alpha^2/(t\text{-}1)\}/\sigma^2 + r\sigma^2_\epsilon$ under fixed effect model. So, F_T = MST/MSE will be the proper F-test for treatment effect (H_0: $\alpha_j = 0$ or H_0: $\sigma^2_\alpha = 0$). In similar way, to test H_0: $\sigma^2_\epsilon = 0$ (there is no variation among the sampling units) we shall use F_E = MSE/MSS as a test statistic instead of MSE/MSE, because F_E estimate $(\sigma^2 + r\sigma^2_\epsilon)/\sigma^2$ under the null hypothesis.

The second basic experimental design used almost in agricultural experiments is that of RCBD (randomized complete block design). As discussed above that CRD requires homogeneous experimental units. But usually the experimental units are not so homogeneous as required. In such situations the principle of local control is adopted and the experimental material is grouped into homogeneous sub-groups. The sub-group is commonly termed as block. The treatments are assigned randomly with in blocks. Each treatment will occur once in a block. Separate randomization is used for each block. The blocks are formed with units having common characteristics that may influence the response under study. In agricultural field experiments the soil fertility is an important character that influence the crop responses. The uniformity trial is used to identify the soil fertility of a field. If the soil fertility is found to run in one direction, then the blocks are made orthogonal of that. In animal experiments animals of same age or little or weight may form blocks.

All the F-tests are taken in a wrong way in case of RCBD nested (hierarchical classification), because in all effects MSE is taken in the denominator of an F-test by the researchers. To take a proper F-test one should take the help of expected mean squares. Expected mean square provides proper F-test about various effects present in the experiment. Since RCBD is a mixed effect model, in which the treatment effect is taken fixed while the effect of blocks is random. For replications/blocks and treatment MSE while for H_0: $\sigma^2_\epsilon = 0$ MSS will be taken in the denominator of an F-test.

Table 3 presents the ANOVA table of Latin Square Design. Expected mean square reveals that for each effect Mean Square Error (MSE) will be used as denominator to take a proper F-test. Latin Square Design is used in a situation if there exists two sources of variations in the experimental units, this variation should be controlled by double way blocking (row wise and column wise). This design is not useful if the number of treatments are fever than four and greater than 10. Also the design is not valid for the comparison of two treatments because the degrees of freedom becomes zero in that case.

Great miss happen occurs in factorial experiments while taking F-ratios for various effects, because it has been observed in various agricultural experiments that always MSE is taken in the denominator while taking F-ratios. By

Table 1: The ANOVA Table for CRD nested (with "t" treatments, "n" experimental units and "r" sampling units in each experimental unit)

S.O.V	DF	SS	MS	E [MS]	F-ratio
Treatment	(t-1)	SST	MST	$\sigma^2 + r\sigma^2_\epsilon + nr\sigma^2_\alpha$ (Random) $\sigma^2 + r\sigma^2_\epsilon + (nr\sum\alpha^2/(t-1)$ (Fixed)	F_T
Experimental Error	t (n-1)	SSE	MSE	$\sigma^2 + r\sigma^2_\epsilon$	F_E
Sampling Error	nt (r-1)	SSS	MSS	σ^2	
Total	(ntr-1)	SSTot			

Table 2: The ANOVA Table for RCBD nested (with "t" treatments, "b" blocks and "n" sampling units in each experimental unit)

S.O.V	DF	SS	MS	E [MS]	F-ratio
Replication/Block	(b-1)	SSB	MSB	$\sigma^2 + r\sigma^2_\epsilon + nr\sigma^2_\rho$(Random)	F_B
Treatment	(t-1)	SST	MST	$\sigma^2 + r\sigma^2_\epsilon + nr\sigma^2_\alpha$ (Random) $\sigma^2 + r\sigma^2_\epsilon + (nr\sum\alpha^2/(t-1)$ (Fixed)	F_T
Experimental Error	(t-1)(b-1)	SSE	MSE	$\sigma^2 + r\sigma^2_\epsilon$	F_E
Sampling Error	bt (n-1)	SSS	MSS	σ^2	
Total	(btn-1)	SSTot			

Table 3: The ANOVA Table for PxP Latin Square Design (LSD)

S.O.V	DF	SS	MS	E [MS]	F-ratio
Rows	(P-1)	SSR	MSR	$\sigma^2 + P\sigma_\gamma^2$ (random)	F_R
Columns	(P-1)	SSC	MSC	$\sigma^2 + P\sigma_\rho^2$ (random)	F_C
Treatment	(P-1)	SST	MST	$\sigma^2 + P\sum_k\alpha_k^2/(P-1)$ (fixed)	F_T
Error	(P-1)(P-2)	SSE	MSE	σ^2	
Total	(P^2-1)	SSTot			

Table 4: ANOVA Table showing only the source of variation, degrees of freedom and E(MS), for three factors A, B and C (all the factors are random)

Factor	DF	E(MS)
A	(a-1)	$\sigma^2_e + n\sigma^2_{\alpha\beta\gamma} + nb\sigma^2_{\alpha\gamma} + nc\sigma^2_{\alpha\beta} + bcn\sigma^2_\alpha$
B	(b-1)	$\sigma^2_e + n\sigma^2_{\alpha\beta\gamma} + an\sigma^2_{\beta\gamma} + nc\sigma^2_{\alpha\beta} + acn\sigma^2_\beta$
C	(c-1)	$\sigma^2_e + n\sigma^2_{\alpha\beta\gamma} + nb\sigma^2_{\alpha\gamma} + an\sigma^2_{\gamma\beta} + abn\sigma^2_\gamma$
AB	(a-1)(b-1)	$\sigma^2_e + n\sigma^2_{\alpha\beta\gamma} + cn\sigma^2_{\beta\gamma}$
AC	(a-1)(c-1)	$\sigma^2_e + n\sigma^2_{\alpha\beta\gamma} + nb\sigma^2_{\alpha\gamma}$
BC	(b-1)(c-1)	$\sigma^2_e + n\sigma^2_{\alpha\beta\gamma} + an\sigma^2_{\beta\gamma}$
ABC	(a-1) (b-1) (c-1)	$\sigma^2_e + n\sigma^2_{\alpha\beta\gamma}$
Error	abc(n-1)	σ^2_e

examining the expected mean squares column of Table 4, we have noticed that, there is no proper test for testing each of the main effects, because F-ratio is the ratio of two mean squares's having the same expectation under the null hypotheses. If we could assume that all two-factor interactions are negligible, then we could put $\sigma^2_{\beta\gamma} = \sigma^2_{\alpha\beta} = \sigma^2_{\alpha\gamma} = 0$ and test for main effect could be performed.

While this seems to be an attractive possibility, we must point out that there must be something in the nature of the process or some strong prior knowledge, in order for us to assume that one or more interaction are negligible. In general, this assumption is not easily made, nor should be taken lightly. We should not eliminates certain interaction from the model without conclusive evidence that it is appropriate to do so. A procedure recommended by some experimenter is to test the interaction first, then to set at zero those interaction found insignificant and then to assume these interaction are zero, when testing other effects in the same experiment. While some times done in practice, this procedure can be dangerous, because any decision regarding the interaction is subject to both type of error. (The conclusion obtained from preliminary testing can be wrong. If we falsely conclude in preliminary test, that one or more components in testing main effect is zero, the test based on these conclusions will be biased. In order to decrease the probability of type-II error (assuming H_0 is true, when it is false) preliminary tests may use larger significance levels than usual. There is no standard level of significance α for preliminary testing, such as the conventional 0.05 for final testing, but α's from 0.20 to 0.30 frequently used and even larger levels have sometimes been recommended. The larger α level increase the probability of type-I errors, but decrease the probability of type-II error). In such a situation it is advised to use the pooling method, approximate tests (proposed by Satterth Waite) and conservative tests.

Similar procedure is advised to be used in the factorial experiments, depending whether the model is fixed, random or mixed effect model or exact test is possible or not. This is possible only by examining the expected mean squares.

REFERENCES

Agarwal, B.L., 1991. Basic Statistics, 2nd Edition, Willey Eastern Limited, New Delhi.
Cochran, W.C. and G.M. Cox, 1957. Experimental Design, 2nd Edition, John Wiley and Sons, Inc; New York.
Draper, N.R. and H. Smith, 1996. Applied Regression Analysis, John Wiley and Sons, Inc; New York.
Gujrati, D., 1984. Basic Econometrics, Mc Graw Hill Book Co. Singapur.
John, J.A. and M.H. Quonouile, 1977. The Design and Analysis of experiments, 2nd Edition, Charles Griffin and Co. Ltd. London.
Johnston, J., 1984. Econometric Methods, 3rd Edition, Mc Graw Hill Book Co. Japan.
Montgomery, D.C. and B.A. Peck, 1982. Introduction to Linear regression Analysis, John Willey and Sons, Inc; New York.
Searle, S.R., 1971. Linear Models, John Wiley and Sons, Inc; New York.