*ORiON, Vol. 1, No. 1, pp. 2-15*

# CONSTRAINED REGRESSION MODELS FOR OPTIMIZATION AND FORECASTING *

P.J.S. BRUWER ** and
J.M. HATTINGH ***
Potchefstroom University for CHE
Potchefstroom

## ABSTRACT

Linear regression models and the interpretation of such models are investigated. In practice problems often arise with the interpretation and use of a given regression model in spite of the fact that researchers may be quite "satisfied" with the model.

In this article methods are proposed which overcome these problems. This is achieved by constructing a model where the "area of experience" of the researcher is taken into account. This area of experience is represented as a convex hull of available data points. With the aid of a linear programming model it is shown how conclusions can be formed in a practical way regarding aspects such as optimal levels of decision variables and forecasting.

* This article is based on a lecture given at a simposium on "CONTROL AND DECISION-MAKING IN THE INDUSTRIAL ENVIRONMENT" at the VAAL TRIANGLE CAMPUS of the PU for CHE by J M Hattingh. - April 23, 1979.

** Department of Computer Science

*** Department of Statistics and Operations Research

## 1. INTRODUCTION

In paragraph 2 the linear regression model is briefly discussed and some of the problems which may arise with the interpretation and use of the model are indicated.

In paragraph 3 the restricted linear regression model is developed and discussed and the application of the model illustrated by means of an example. In the last paragraph an indication of the applicability of the method in general is given with possible extensions. It is possible to extend these results to the nonlinear regression model.

## 2. THE LINEAR REGRESSION MODEL AND ITS USE

Consider the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i$$

where $\varepsilon_i$ are independent random variables with

$E(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all i.

For this model:

a. Values $X_{i1}, X_{i2}, \dots X_{ik}$ are specified by the researcher,

b. $X_{ij}$ is the $i^{th}$ value of the $j^{th}$ independent variable, and

c. $\beta_0, \beta_1, \dots \beta_k$ are unknown parameters to be estimated.

Several methods to estimate the parameters exist. The estimates of the parameters $\beta_j (j=0,1, \dots, k)$ will be denoted by $b_0, b_1, b_2, \dots, b_k$. The coefficients $b_j (j=0,1, \dots, k)$ are to be determined from N observations $(Y_i, X_{i1}, \dots, X_{ik})$, $i=1,2,\dots,N$.

One of the best known methods is the method of least squares where $b_0, b_1, \dots, b_k$ are determined in such a way that

$\sum_{i=1}^{N} \Delta_i^2$ is minimized where

$$\Delta_i = Y_i - b_0 - b_1 X_{i1} \dots - b_k X_{ik}.$$

Some other well known methods to obtain the estimates are given below:

a. Find $b_0, b_1 \ldots b_k$ to minimize

$$\sum_{i=1}^{N} |\Delta_i| .$$

b. Make use of the Mallows $C_p$ criterion [3].

Some of the stepwise methods try to eliminate variables from the regression model to obtain a less complicated model (see for example [3] and [5] ).

Some disadvantages of the stepwise approach are given below:

a. The data is often not obtained from a designed experiment and quite often interdependence exists between decision variables.

b. Non-measureable variables which could play a role are not considered by the model.

c. It is difficult to eliminate inaccuracy and outliers from the data.

d. It is not always clear which variables should be used in the model.

e. A priori information and constraints on parameters are often not be taken into account by the model.

f. Even in situations where we are quite satisfied with the model some questions like the following may still exist:

1) How do we interpret the regression function?

2) What is the influence of a specific "independent" variable?

3) How do we use the estimated function when we are interested in forecasting or optimization of the dependent variable?

In this article the discussion is restricted to the questions raised in f.

Some embarassing situations often arise when a researcher tries to use an estimated regression equation.

For example:

1) A totally "different" estimating equation may be available from another regression study. (Although both fits might be of significance).

2) The sign and numerical values of the regression coefficients are often in contrast with physical reality.

We claim that in order to optimize the dependent variable or to make useful forecasts it is often essential to know the area of experience that gave rise to the regression funcion obtained. Below we give a hypothetical example to illustrate this point.

EXAMPLE

Consider a chemical reactor where raw gas mixtures are transformed into final products. Suppose a catalyst is used to further the reaction and suppose the variables that could be measured are

$Y \equiv$ efficiency of the reaction

$X_1 \equiv$ temperature

$X_2 \equiv$ activity of the catalyst (often measured by the inverse of age)

It is customary to raise the temperature to improve the reaction when the transformation process deteriorates (as result of the declining of the catalyst). Therefore a simple regression of Y on $X_1$ alone will often show a negative regression coefficient (as result of the customary production procedure). Regression of Y on both $X_1$ and $X_2$ will often give more meaningful results.

Suppose we obtain the regression funcion

$\hat{Y} = b_0 + b_1X_1 + b_2X_2$, with $b_1$, $b_2 > 0$ when this is done.

Consider a situation where we want to propose levels of the decision variables $X_1$ and $X_2$ that will maximize $\hat{Y}$. In the first place we notice that $X_2$ is a so called state variable in the sense that this variable is not under control of the experimenter. $X_1$ on the other hand is a control or decision variable but can not be varied independently of $X_2$ in practical situations. Note that in practice a high level of $X_2$ usually corresponds to a low level of $X_1$ and vica versa.

This is an example of a situation where inderdependence between the "independent" variables exists. It is therefore clear that this situation should be analyzed further to make any realistic proposal to control the production process. In the next paragraph attention is given to a method which could be used in this (and other) situations.

## 3. THE CONSTRAINED REGRESSION MODEL METHOD

### 3.1 THE AREA OF EXPERIENCE OF THE MODEL

In this paragraph we take a look at the area of experience measured by the existence of data points in this area. For this purpose we consider the convex set obtained by looking at all possible convex combinations of the observed data points. This is enclosed by the convex hull of the data points.

Suppose we have data points.

$$V_i = (X_{i1}, X_{i2}, \ldots X_{ik}) \text{ for } i = 1, 2, 3, \ldots N$$

The convex hull could now be represented as the following set:

$$C = \{Z/Z \in E^k \text{ and } Z = \sum_{i=1}^{N} \lambda_i V_i \text{ with } \lambda_i \geq 0$$

$$\text{and } \sum_{i=1}^{N} \lambda_i = 1\} \text{, and where } E^k \equiv \text{ k-dimensional Euclidian space.}$$

We are now interested in the behaviour of the regression function in this area C.

### 3.2 THE INFLUENCE OF AN "INDEPENDENT" VARIABLE

A researcher is usually interested in the influence of a specific independent variable upon the dependent variable. When no inderdependence exists between decision variables, we need only take notice of the regression coefficient of the specific variable. In such a situation it could be possible theoretically to obtain for example the maximum of the dependent variable $\hat{Y}$ by setting the values of the decision variables with positive regression coefficients as high as possible and those with negative regression coefficients as low as possible. The problem that arises with this approach is the fact that it could be impossible physically to implement the proposed combination of levels of the variables. No experience may exist where a spesific variable $X_j$ has a high level and another variable $X_m$ also has a high level. If the proposal is that both these variables should be at a high level, it could be an unrealistic proposal which could not be implemented physically. In this discussion we assume that we only want to propose levels of variables which fall within the area of experience of the experimenter. That is, levels of decision variables that fall in the convex hull C. Suppose we want to determine the influence of a specific variable $X_p$ on the dependent variable $\hat{Y}$. In the first place we could determine the range of onservations on $X_p$. Suppose the minimum value observed is K' and the maximum K".

If we let $X_p = q$ where $q \in (K',K")$ we now want to determine the values of the remaining decision variables so that $\hat{Y}$ is maximum or minimum.

For this purpose we have to solve the following linear program:

$$\begin{matrix}\text{Max}\\ \text{Min}\end{matrix} \quad \hat{Y} = b_0 + b_1 X_1 + \ldots + b_k X_k$$

subject to the following constraints:

$$\sum_{i=1}^{N} \lambda_i X_{ij} = X_j \geq 0 \text{ for } j = 1 \ldots k$$

$$\sum_{i=1}^{N} \lambda_i = 1, \lambda_i \geq 0 \text{ for } i = 1, \ldots N$$

$$X_p = q$$

The solution of the linear programming problem then yields the maximum (minimum) of $\hat{Y}$ as well as the levels of the decision variables $X_1, X_2, \ldots, X_k$, where this optimal level is reached. The above solutions may be obtained by using parametric linear programming techniques (see [4] Chapter 8).

Note that the difference between the maximum and minimum values of $\hat{Y}$ at any level q of $X_p$ is an indication of the influence of the remaining decision variables on $\hat{Y}$. This influence is relatively small (large) when the difference is small (large). In the discussion above only one decision variable $X_p$ was restricted at a particular level and the linear program could then be solved. It is of course possible to restrict more than one decision variable and then solve the linear programming problem. Indeed in practice more than one state variable can often be present and incorporated in the model.

## 3.3 SUMMARY OF THE CONSTRAINED REGRESSION MODEL METHOD

3.3.1 Obtain a regression model that is "satisfactory"(Make use of ridge regression when high degree of interdependence exists).

3.3.2 Determine the area of experience of the regression model by identifying the convex hull of the available data points.

3.3.3 Identify the state variables whose influence on the dependent variable have to be investigated.

3.3.4 Select a specific level for this variable.

3.3.5 Optimize the regression function over the area within the convex hull where this variable is at a specific level. Obtain maximum and minimum values. Select another level and repeat the procedure.

3.3.6 Graph the optimum values (maximum and minimum) of the regression function against different levels of the chosen variable.

### EXAMPLE

Suppose we want to evaluate the benefits of a proposed irrigation scheme in a certain area. This evaluation must be done by using data available of production (Y) and rainfall ($X_1$) over a number of years. Table 1 contains these data.

TABLE 1[1)]

| YEAR | PRODUCTION(Y) | RAINFALL($X_1$) | AVERAGE TEMPERATURE ($X_2$) |
|---|---|---|---|
| 1963 | 60 | 8 | 56 |
| 1964 | 50 | 10 | 47 |
| 1965 | 70 | 11 | 53 |
| 1966 | 70 | 10 | 53 |
| 1967 | 80 | 9 | 56 |
| 1968 | 50 | 9 | 47 |
| 1969 | 60 | 12 | 44 |
| 1970 | 40 | 11 | 44 |

In this case we can think of Y as the dependent variable and $X_2$ as an independent variable (state variable). If $X_1$ can be viewed as the application of moisture in the process, it could be considered as an independent (decision) variable.

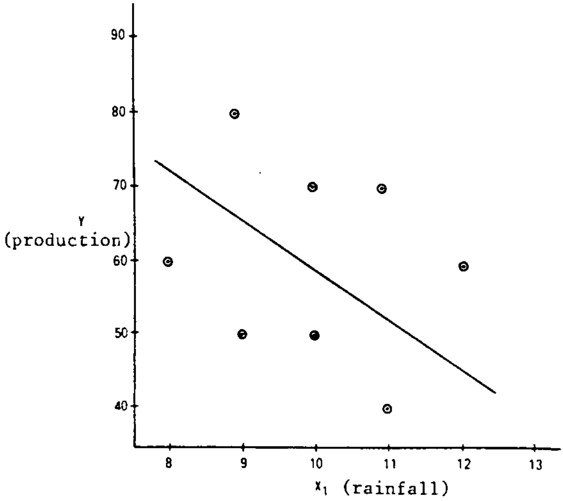
[1)] From: Wonnacott & Wonnacott: Introductory Statistics.
New York, Wiley, 1972, p.304

Using ordinary least squares we find the following regression equation (of Y on $X_1$ alone):

$$\hat{Y} = 76,7 - 1,67X_1$$

The standard error of the regression coefficient of $X_1$ is $S_{b_1} = 4,0$. See the graph in figure 1. Notice the poor fit of the equation - in fact $R^2 = 0,028$! In contrast with our intuition we further notice a negative regression coefficient.

FIGURE 1



Let us now look at the multiple regression of Y on $X_1$ and $X_2$.

The equation derived is:

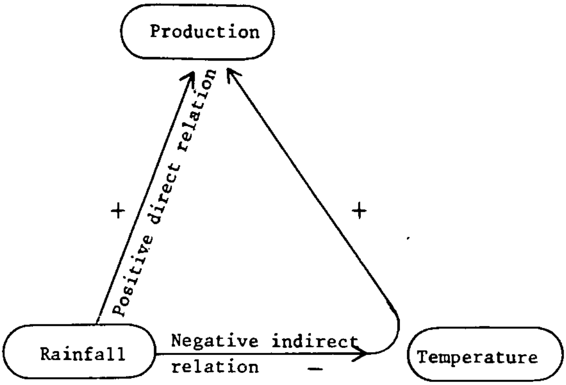
$$\hat{Y} = -144,6 + 5,71\ X_1 + 2,95\ X_2$$

The standard errors of the regression coefficients of $X_1$ and $X_2$ are respectively:

$S_{b_1} = 2{,}68$ and $S_{b_2} = 0{,}69$

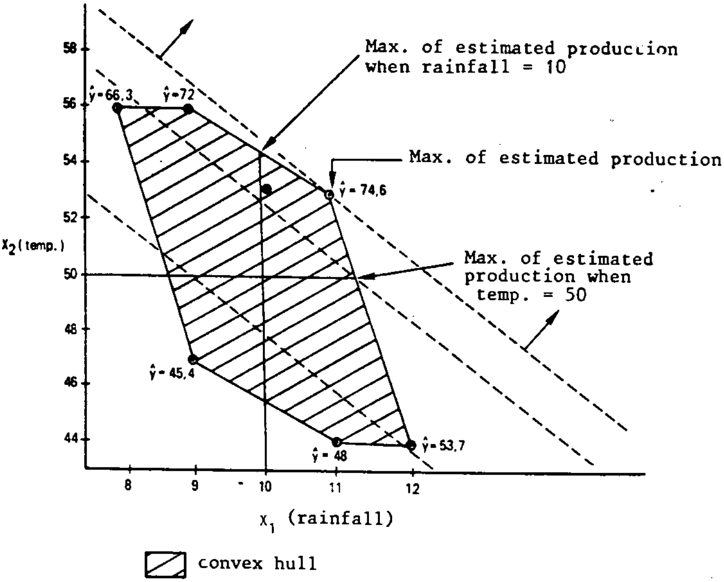
The fit of this equation is somewhat better and both regression coefficients are positive. This indicates the posibility of an indirect effect of rainfall via temperature as shown logically for this process in figure 2.

FIGURE 2



Now apply the constrained regression algorithm to this example. The main components are shown in figure 3.

FIGURE 3



From this figure we notice a probable interdependence between $X_1$ and $X_2$. The convex hull is also shown in this figure. Suppose we look at the effect of temperature ($X_2$) on Y. If we select a level of $X_2$ say $X_2 = 50$ we see that the feasible region of the linear programming problem is the intersection of the sets represented by $X_2 = 50$ and the convex hull.

If we maximinze the objective function

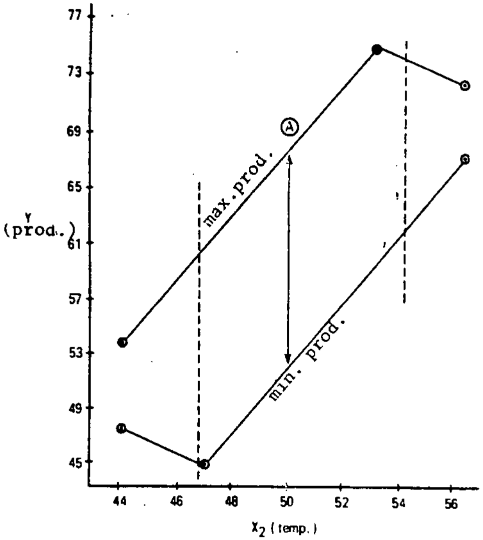
$$\hat{Y} = -144,6 + 5,71\ X_1 + 2,95\ X_2$$

subject to the feasible region and the constraint $X_2 = 50$, we find a solution where

$$X_1 = 11,33 \text{ with } \hat{Y} = 67,6$$

In the same way other levels of $X_2$ could be selected and the procedure repeated. This maximum and minimum values can be drawn and the results are contained in figure 4.

FIGURE 4

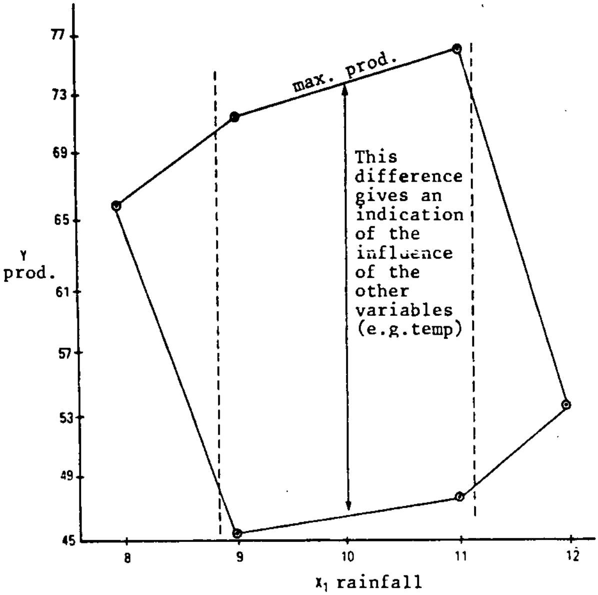


Notice the strong positive influence of temperature on production. The point A is found when $X_2$ = 50 and $X_1$ is chosen as 11,33. In this particular case $\hat{Y}_{max}$ = 67,6.

The vertical distance between the maximum and minimum values gives an indication of the relative importance of the variable $X_2$ on the dependent variable $\hat{Y}$.

In the same way the method could be applied to determine the influence of $X_1$ (rainfall) on $\hat{Y}$. This procedure is analogous to the one above and the resulting graph is given in figure 5. Notice in this case that a relatively big difference between the maximum and minimum values exists, which is an indication of the other variable(s) relative strong influence.

FIGURE 5



## 4. CONCLUSIONS AND EXTENSIONS

The method discussed above could be applied to any linear regression model in general. In this discussion only one decision or state variable was restricted at a certain level but it is of course possible to restrict more than one variable at a certain level at a time and it might also be necessary to do so in some applications.

When the model is not a linear regression model but for instance a quadratic regression function, the possibility exists to apply other optimizing techniques such as quadratic programming to do the optimization.

This method and model was used in a research project to evaluate the performance of the computer-based information systems of a large organization in South Africa (see [2] ). Several publications resulted from this project. The method seemed to be a very useful one when realistic proposals which can physically be implemented has to be make from a linear regression model.

REFERENCES

[1] BEALE E.M.L., KENDALL M.G. and MANN D.W., *"The discarding of variables in multivariate Analysis"*, Biometrika, 54, 3 and 4, 2967, p.357.

[2] BRUWER P.J.S., *"Evaluating the Performance of Computer-Based Information Systems using a Restricted Linear Regression Model"*. Quaestiones Informaticae, Vol. 2, No.3, September 1983, pp.1-6.

[3] DANIEL C. and WOOD F.S., *Fitting equations to Data*, Wiley Interscience, New York, 1971.

[4] GASS S.I., *Linear Programming*, McGraw-Hill Book Company, Inc., New York, 1958.

[5] GRAYBILL F.A., *Theory and Application of the linear model*, Duxbury Press, Massachusets, 1976.