THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

USE OF MULTIPLE REGRESSION TECHNIQUES FOR ESTIMATING MUNICIPAL SEWAGE TREATMENT COSTS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

BY

JOSEPH ROBERT ASSENZO

Oklahoma City, Oklahoma

1963

USE OF MULTIPLE REGRESSION TECHNIQUES FOR ESTIMATING

MUNICIPAL SEWAGE TREATMENT COSTS

APPROVED BY

John C. Griffes

Robert D Morrison

Geo. W. Reid

Clinton M Miller III

Pearl D. Fisher

DISSERTATION COMMITTEE

ACKNOWLEDGMENT

I am indebted to Doctor John C. Brixey, my senior advisor for his guidance and help in the presentation of the material in this dissertation, and am also indebted to Professor George W. Reid for his advice concerning the sanitary engineering aspects of the dissertation. I am grateful to the National Science Foundation and the Department of Preventive Medicine, University of Oklahoma, for the fellowships which allowed me to pursue my studies and research on a full time basis. The financial support of the mail survey given by the Bureau of Water Resources Research, University of Oklahoma is greatly appreciated. I am very grateful to the Biostatistical Unit and Medical Research Computer Center for the use of their computer facilities and for the advice given to me by members of the unit, in specific, Doctors Brandt, Miller and Fisher. Finally, my deep appreciation goes to Mrs. Rose Titsworth for her valuable help in typing this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

USE OF MULTIPLE REGRESSION TECHNIQUES FOR ESTIMATING MUNICIPAL SEWAGE TREATMENT COSTS

CHAPTER I

INTRODUCTION

One of the major goals in water resources planning in the United States is to derive a mathematical model that will maximize the net benefit from the operation of a water resources region. To optimize the net benefit it is essential that all the variables, and their interactions, be well defined and that their effect on the operation of a water resources region be estimable.

One of these variables is the number of people within a water resources region. It is necessary, then, that means be available for estimating future population, the amount of water they will require, and the amount of waste they will generate.

Two more variables are the cost of treating municipal water and the cost of treating municipal wastes. There are direct relationships among these cost variables and the cost variables for other water uses, such as, industrial water and waste treatment. Hence, a knowledge of municipal water and waste costs will provide a means for estimating costs of other water uses and waste treatment.

Accompanying the estimated increases in population and in

their activities will be a great need for adequate and proper treatment of wastes before discharge into receiving streams. This is important in order to prevent pollution and contamination of the stream water to such an extent that it becomes unusable. It is essential that there is available in the planning stages of the development of water resources activities within any water basin a reliable estimate of the cost of the treatment of wastes of varying strength to any specified degree.

The degree of treatment needed, and cost thereof, is controlled by the nature of the receiving water, consequently, two more variables which must be considered with a discussion of waste disposal are dilution requirements and the cost of dilution water. That is, it is necessary, using current treatment methods, to have available high volumes of relatively clean water to dilute the treated wastes so as to maintain, in the stream, dissolved oxygen concentrations which will support fish and other biological life. This dilution requirement is greater than the sum of the water requirements for all other purposes, such as, navigation, agriculture, municipal water supply, etc. It follows from this that new treatment processes, requiring less dilution water, must be developed. Any study of the treatment cost variable should provide guidelines for extrapolating from present knowledge costs of treating wastes by new processes.

A few mathematical models for maximizing net benefit have been proposed, but unreliable input data for use in the models make them unworkable. This dissertation is concerned with one of the fore-mentioned variables, knowledge of which is needed to make the mathematical models

operable, the cost of waste treatment variable. Methods will be presented for estimating the cost of waste treatment as a function of several variables, such as, population equivalency, degree of treatment, and type of treatment plant.

Other variables, such as, cost of water treatment, and their interactions with cost of waste treatment will not be discussed herein because of the enormity of such a study. A study of these will be left for a later date.

The Need for Waste Treatment Facilities

The Bureau of the Census (1, 2) has estimated that by year 1980 the population of the United States may be as high as 274 million and that by year 2000 the population may exceed 420 million. Table 1 is a summary of the Bureau of the Census population projections.

Accompanying this increase in population, it is anticipated that there will be an increase in water consumption. It has been estimated by the Public Health Service (3) that by 1980 total withdrawals of water for all purposes might almost equal the 650 billion gallons per day of total developable supplies. By year 2000, such withdrawals could amount to almost twice the total supply. For municipal purposes alone, demands in year 2000 could equal five times present domestic requirements, or about 85 to 90 billion gallons per day.

What causes concern here is not only the total amount of water required, but the fact that water as it is used becomes polluted and may become unfit for further use for some purposes unless it is adequately treated.

TABLE 1

ESTIMATES OF THE TOTAL POPULATION OF THE UNITED STATES INCLUDING ARMED FORCES ABROAD (2)

Projections* Population in Millions				
Year	Series I	Series II	Series III	Series IV
1965	200	197	195	192
1970	221	215	209	204
1975	245	236	227	217
1980	274	261	247	232
1985	306	288	267	248
1990	340	317	288	263
1995	378	349	310	279
2000	420	384	333	295

*The population estimates for the four series were based on different fertility assumptions.

Since water can be, and is, used over and over again as it flows to the sea one is tempted to say that such huge demands are no cause for alarm. However, one of the principal requirements for water in the future is for the dilution of effluent resulting from the treatment and disposal of municipal and industrial wastes into the nation's streams. Reid (4) has estimated the dilution requirements, in relation to degree of treatment, for maintenance of four parts per million of dissolved oxygen in all portions of a stream for the projected waste discharges (5). The predictions by Reid (4) and Wollman (6) indicate that in years 1980 and 2000 approximately 64 per cent of the required stream flow for all purposes will be necessary for waste dilution. Table 2 shows the distribution of predicted required streamflow for various uses.

Of the twenty-two major water resources regions in the United States all will require water storage facilities to provide the required stream flow to meet the needs of all uses. One important problem to be solved, therefore, is the selection of the most economical and efficient combination of water storage facilities and new, improved, and more highly efficient waste treatment facilities. As a first step toward this selection it will be necessary to be able to estimate the cost of treating wastes. This dissertation is concerned with this portion of the problem.

Actually, the waste treatment facility problem is not only one of the future, but a current problem. There is a large backlog of needed sewage and industrial waste treatment construction. The Public Health Service (7) has estimated that nearly 2,900 new sewage treatment

TABLE 2

DISTRIBUTION OF REQUIRED STREAM FLOW BY USES,

UNITED STATES, 1980 and 2000 (4, 6)

	Per cent of Total Flow	
Use:	1980	2000
Agriculture	20.0	18.1
Mining	0.1	0.1
Manufacturing	1.7	3.0
Thermal Power	0.3	0.4
Municipal	0.7	0.8
Land Treatment	0.8	1.0
Fish and Wildlife Habitat	12.8	12.8
Sub-Total	36.4	36.2
Waste Dilution Flow	63.6	63.8
Total	100.0	100.0

works are needed to serve 19.5 million people living in communities that have never provided treatment for their wastes. Another 1,100 new plants are needed for 3.4 million people in communities where treatment works built in the past have become overloaded or obsolete. In addition to these 4,000 communities needing new plants, another 1,630 communities have sewage treatment facilities requiring enlargement or the addition of new units or processes to adequately serve populations totaling more than 25 million.

Growth in population and urbanization create new sewage treatment needs continuously, and existing treatment works become obsolete. The Public Health Service (7) estimated that if municipalities are to catch up with treatment needs by 1965, they will have to spend $1.9 billion to eliminate the backlog, $1.8 billion to provide for new population growth, and $0.9 billion to replace plants that will become obsolete. This is a total of $4.6 billion.

Thus, reliable estimates of sewage treatment costs are an immediate need.

CHAPTER II

PREVIOUS WORK BY OTHERS

Previous work concerning the estimation of the cost of waste treatment has involved the use of only one independent variable in the estimation equation. The dependent, or cost, variable has been expressed as either dollars per capita or dollars per million gallons per day (mgd) of waste flow. The independent variable used has been either design population or capacity of treatment plant in mgd. Regional differences in cost of sewage treatment plants have been taken into account through the use of the Engineering News-Record (ENR) Construction Cost Index.

Velz (8) related the unit construction cost of waste treatment works per mgd to size of plant in mgd. All plant costs were referred to 1926 as the base year of construction, adjusted by means of the United States average ENR Construction Cost Index. The costs were also referred to 100 per cent efficiency in the removal of the Biochemical Oxygen Demand (BOD) by considering primary, chemical coagulation, trickling filter, and activated sludge systems as effecting 35, 65, 85, and 90 per cent BOD removal, respectively. No correction for regional price differentials was made. The estimation equation used was of the form:

$$y = ax^b,$$

where y is unit cost per mgd, and x is size of plant in mgd. The unit-cost curves developed by Velz were based on a representative sample of waste treatment plants in northeastern and central United States, and as such are valid only in these regions.

In an effort to update the work of Velz, Diachishin in 1957 (9) analyzed data gathered by the Engineering News-Record. Diachishin, as did Velz, related the unit construction cost of waste treatment works per mgd to size of plant in mgd. All plant costs were referred to 1913 as the base year of construction, adjusted by means of the ENR Construction Cost Index which has a value of 100 for the year 1913. Two estimation equations were derived; the first to be used for primary treatment plants, plants of approximately 35% BOD removal, and the second for trickling filter and activated sludge plants. The form of both equations was:

$$y = ax^b,$$

where y and x are as defined in the equation by Velz. The cost curves were based on a random sample from all sections of the United States. No attempt was made to account for regional differences in the estimation equations.

In 1958, Thoman and Jenkins (10) reported on the results of a cost study conducted by the United States Public Health Service. Construction costs were adjusted to the ENR Construction Cost Index base year 1913. Three equations, one for primary treatment plants, one for secondary treatment plants, and one for oxidation ponds, were computed for estimating cost per capita as a function of design population. In an effort to account for regional differences in construction costs the

United States was partitioned into twenty regions on a county line basis. Each region corresponded to one of the twenty cities used in obtaining the United States Average ENR Construction Cost Index. A treatment plant selected within any region was assigned the ENR Cost Index for the ENR city. For example, a sewage treatment plant constructed in Oklahoma City in, say, 1950 was assigned the ENR construction cost index for Dallas for 1950. The form of the equations was,

$$y = ax^b,$$

where y is cost per capita, and x is design population.

In 1960, Rowan, Jenkins, and Butler (11) updated the Thoman and Jenkins study. Again all cost data were converted into 1913 dollars using the ENR cost index. The country was also divided, as before, into twenty areas. In this study six estimation equations were derived relating cost per capita to design population. The equations were for Imhoff tank treatment, primary settling and separate sludge digestion, activated sludge treatment, trickling filter treatment with separate sludge digestion and final settling, trickling filter treatment with contained digestion system, and oxidation pond treatment. The equations had the form:

$$y = ax^b,$$

where y and x are as defined in the Thoman and Jenkins study.

Rowan, Jenkins and Howells (12), realizing that insufficient attention is often given to the cost of operating and maintaining sewage treatment plants prior to their construction, reported on a cost study conducted by the Public Health Service. The inverse function,

$$\log y = \frac{1}{a+b \log x},$$

was the form of the estimation equation used because its use resulted in the smallest standard error of estimate. Both cost per mgd studies and cost per capita studies were conducted for the following types of waste treatment plants: primary, activated sludge, standard-rate trickling filter, and high-rate trickling filter. Thus, y, in the equation, is annual operation and maintenance cost either per mgd or per capita, and x is either average daily flow in mgd or population served.

Logan, Hatfield, Russell, and Lynn (13) reported on their investigation of the application of systems-analysis techniques to the preliminary design of waste-water treatment plants. Equations for estimating cost per mgd as a function of design capacity in mgd were derived for each of the unit processes in primary, high-rate trickling filter, standard-rate trickling filter, and activated sludge treatment plants. All equations were of the form;

$$y = ax^b,$$

where y is cost per mgd and x is design capacity in mgd.

Wollman (14), in developing a relationship for the estimation of operation and maintenance costs, was the first to use the multiple regression model. The estimation equation used by him was a linear one of the form:

$$Y = b_o + b_1x_1 + b_2x_2 + b_3x_3$$

where Y = the annual operation and maintenance cost per daily population equivalency (PE),

x_1 = treatment level in per cent of BOD removed,

x_2 = per cent of total waste that is industrial, and

x_3 = population served by sewage system.

The regression equation was computed for 38 non-Southwest cities.

For comparison of the estimation equations derived by the afore-mentioned authors, Table 3 gives values of expected construction cost for selected design conditions. All estimates were referred to 1913 dollars. The equations estimating cost per mgd are not directly comparable to those estimating cost per capita, hence they are presented separately. As can be seen in the table, in some instances the estimates are fairly close to each other, while in others they are quite different.

TABLE 3

WASTE TREATMENT PLANT CONSTRUCTION COST ESTIMATES

FOR SELECTED CONDITIONS USING EQUATIONS

DERIVED BY OTHER AUTHORS

Capacity mgd	Type of Plant	Estimated cost per mgd in 1913 dollars, by Velz	Diachishin	Logan
0.1	Prim	$43,700	$64,000	$36,900
0.1	TF	106,200	79,000	43,100
0.1	AS	112,500	79,000	73,900
1.0	Prim	28,600	31,000	23,400
1.0	TF	69,500	42,000	34,500
1.0	AS	73,500	42,000	49,300
3.0	Prim	21,900	22,000	20,900
3.0	TF	53,100	32,000	30,800
3.0	AS	56,200	32,000	39,400
10.0	Prim	16,800	15,000	14,700
10.0	TF	40,900	27,000	25,800
10.0	AS	43,300	27,000	34,500

TABLE 3-Continued

Design Population	Type of Plant	Estimated cost per capita in 1913 dollars, by Thoman	Rowan
1,000	Prim	$7.00	$7.42
1,000	TF	7.80	9.02
1,000	AS	7.80	8.73
10,000	Prim	3.80	3.53
10,000	TF	5.00	4.73
10,000	AS	5.00	4.87
30,000	Prim	1.50	2.50
30,000	TF	4.00	3.20
30,000	AS	4.00	3.30
100,000	Prim	1.80	1.68
100,000	TF	3.00	2.48
100,000	AS	3.00	2.72

Prim = Primary, TF = Trickling Filter, AS = Activated Sludge

CHAPTER III

VARIABLES USED IN THE STUDY

From theoretical considerations it would be more reasonable to relate unit cost to organic loadings rather than design population or capacity in mgd. This is so because the biological treatment unit processes of secondary treatment plants are designed on anticipated organic loadings rather than volumetric loadings. A commonly used measure of organic loading is population equivalency (PE) of the waste. The population equivalency of the wastes from a municipality may be computed in the following manner,

$$PE = \frac{8.33\ Qx}{c},$$

where PE is population equivalency, Q is average flow of waste into the treatment plant in mgd, x is average 5-day BOD of the waste in parts per million (ppm), c is generally assumed to be, as in this study, 1/6 of a pound of BOD per capita per day, and 8.33 is a conversion constant. The PE reflects the contribution to the organic loading from all sources within the community, e.g., domestic and industrial.

The PE was not used in the development of estimation equations in past studies because it was felt that this information was not available in sufficient volume (11) to obtain a high degree of precision. (The word precision is used herein in its statistical sense, i.e., high

degree of precision or high precision indicates a low value of the estimate of the residual mean square, or the variance, in the analysis of variance test for significance of regression (15)).

One obvious way to increase the precision of an estimator is to gather more data for use in the estimation procedure. Another method which may result in an increase in precision is the use of a more complex statistical model, such as the multiple regression model. The use of a multiple regression model may be indicated because of purely theoretical considerations, or because of the fact that a regression equation using only one of several independent variables does not give a high enough precision to be of great value, or because the use of additional variables significantly increases the precision. Williams (16) states that before collecting data for the derivation of a multiple regression equation some thought should be given to the selection of independent variables. Only those independent variables which are thought to add a significant amount to the sum of squares due to regression are worthwhile including in the relationship (15). Also, independent variables that are readily measurable or observable should be selected, both so that they can be used in deriving the estimated relationship and also, since the relationship may be required for later use in estimation, so that values can be determined for this purpose (16).

In addition to PE, another variable which may be important is the BOD of the effluent of the treatment plant. This variable is to be considered because it is a measure of the overall efficiency of the treatment plant. In almost every state a regulatory agency has, or will have, regulations concerning the required BOD of either the effluent or

of the stream water. Therefore, a fairly good estimate of the BOD of the effluent will be available for use in cost estimation equations.

In the study by Rowan, Jenkins and Butler (11) six different equations, one for each of six types of treatment plant, were computed. If it were possible to quantify the type of treatment plant by some rational method then one would have more observations to use in the development of a regression equation. The result would be an estimation equation with a higher degree of precision.

A search of past and current literature failed to find any references concerning the quantification of the type of treatment plant. However, in a paper by Reid (17) concerning biological treatment processes there is a table wnich presents the effectiveness of biological treatment processes relative to the sand filter process. The relative effectiveness for each process was derived by computing, from the accepted design criteria, the number of people per acre that the process was capable of handling; multiplying this value by the expected efficiency of the process and coding by dividing by the effectiveness of the sand filter.

It was decided by Reid and the author that with few modifications this method could be used to develop a continuous variable which would describe the type of treatment plant. For each of the unit processes within a treatment plant, (for example, the unit processes in a standard rate trickling filter plant with separate sludge digestion are primary settling, secondary settling, sludge digestion, and the standard rate trickling filter), the number of persons per acre, that the unit was capable of handling, was computed using accepted design criteria. Each

of these values was then multiplied by a weight factor, the expected efficiency of the unit operation, then summed to obtain the number of people per effective area. The people per effective area values for each of the types of treatment plants were coded by dividing all of them by the value for the primary type treatment plant. Table 4 shows the values of the type of treatment plant variable derived for use in this study.

TABLE 4

TYPE OF TREATMENT PLANT VARIABLE

(Based on people per effective area)

Type of Plant	Value
Imhoff Tank with Sand Filter	0.73
Primary*	1.00
Imhoff Tank and Standard Rate Trickling Filter	1.11
Standard Rate Trickling Filter*	1.62
High Rate Trickling Filter*	2.36
Contact Aerator*	2.46
Activated Sludge*	2.61

*Includes separate sludge digestion

Obviously one way to improve the values given in the table would be to study in detail the unit processes of a great number of treatment plants. From the results of the study one could derive more exact relationships for determing the people per effective area for each type of plant. The values in the table are expected values and,

therefore, should be valuable for the purposes to which the estimation equations are to be applied.

In previous studies (8, 9, 10, 11, 12, 13) the dimension of the dependent variable was either dollars per capita or dollars per mgd. In order to compute the expected total cost one computed the expected value of the dependent variable from either the design population or design flow and multiplied this value by the design population or flow, depending on which equation was used. In this study it was decided to derive equations which would predict costs in dollars per capita, dollars per PE of influent (PE produced), and dollars per PE treated. The total cost could be obtained then from a knowledge of design population or design PE.

In any economic study of optimum operation of a water resources region one must have some estimate of what it costs to treat waste. The cost per PE treated will provide the economist with this information. Of course, in order to obtain total cost, he must add to this the cost of diluting the effluent with relatively clean water in order to maintain the required in-stream standards. As stated previously some regulatory agency will in all probability set some limit on the strength of the effluent and this value can be used in the computation of the second part of the total cost. Being able to predict total treatment cost, the economist can then apply a cost minimizing analysis in his attempt to predict operation of a water resource region to obtain maximum net benefit.

If one has an estimate of the strength of the influent and the effluent then he has a measure of the overall efficiency of the

treatment plant. Consequently, he must choose a type of treatment plant which is consistent with the required efficiency. For example, a primary treatment plant, which is capable of removing perhaps as much as 35 per cent of the BOD, would not be chosen for the case where, say, 80 per cent removal was required.

CHAPTER IV

STATISTICAL MODELS AND TEST CRITERIA USED IN THE STUDY

The major aim of this study is to estimate the cost of municipal waste treatment. This is to be accomplished by finding an equation which relates certain quantities. In this study as in many engineering and scientific investigations the concept of cause and effect is obscure. The causes may be, perhaps, unidentifiable because of lack of knowledge, or if some are identifiable they may be unquantified. If the causes of the effect are not known then it is not possible to predict an effect exactly, but there may be variables, X_i, which can be observed and which are valuable for predicting the effect, Y. The use of these X_i may result in a difference between an estimated Y and an observed Y for the same values of the X_i. This disagreement in Y values may be due either to the fact that the derived relationship between Y and the X_i is not correct, or to the fact that all or some of the X_i cannot be measured exactly. Therefore an error is introduced; the first type of error mentioned above is called an equation error and the second, measurement error. Often both types of error occur simultaneously and since it is usually difficult, if not impossible, to separate, or identify, the components of these types of errors there is generally only one error term in any given model.

Since a great deal of mathematical theory has been developed

using linear equations it was decided to use the linear statistical model in this study. A linear model is an equation that involves random variables, mathematical variables and parameters. It is an equation that is linear in the parameters and in the random variables (18). The form of the model that was used is

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e \qquad (1)$$

where, Y is an observable random variable, X_1, X_2, ..., X_k are known mathematical variables, β_o, β_1, β_2, ..., β_k are unknown parameters and e is an unobservable random variable, the error term, with mean 0. For a review of the derivation of the distribution of pertinent statistics needed for estimation of the parameters in this model and for testing hypotheses about them and for the necessary assumptions see Graybill (19).

Previous studies (8, 9, 10, 11) indicated that in all likelihood a transformation of variables would be necessary in order that the assumptions of the statistical procedures be met. After the sample was collected a frequency polygon of the Y values was constructed. This resulted in a right skewed frequency distribution with a lower limit of zero. A transformation commonly used with this type of distribution is the logarithmic transformation. Although this procedure indicated the necessity for a transformation in the dependent variable it was not possible to determine whether or not transformation of the independent variables was required. Consequently, it was decided that, using the same sample data, the partial regression coefficients for the following linear equations would be computed. The form which gave the "best" fit was to be used as the estimation equation:

$$Y = b_o + b_1X_1 + b_2X_2 + \ldots + b_kX_k \quad (2)$$

$$\ln Y = b_o + b_1X_1 + b_2X_2 + \ldots + b_kX_k \quad (3)$$

$$\ln Y = b_o + b_1\ln X_1 + b_2\ln X_2 + \ldots + b_k\ln X_k \quad (4)$$

$$1/\ln Y = b_o + b_1\ln X_1 + b_2\ln X_2 + \ldots + b_k\ln X_k \quad (5)$$

$$1/Y = b_o + b_1X_1 + b_2X_2 + \ldots + b_kX_k \quad (6)$$

where Y is the random, or dependent, variable, ln is the base e logarithm, the X_i are the known mathematical, or independent variables, and the b_i are estimates of the unknown parameters β_i in equation (1). The b_i are called partial regression coefficients. They may be read, say in equation (2), as the partial regression of Y on X_i, or as the regression of Y on X_i for fixed values of the other variables.

The technique selected for computing the regression coefficients was the abbreviated Doolittle method (20, 21). A Fortran II program for the IBM 1620 was written to fit the same set of data to equations (2) through (6) and to test hypotheses concerning the estimation of the parameters. The program is on file at the Medical Research Computer Center of the University of Oklahoma Medical Center.

When one is using the same data to derive several forms of a linear equation it is believed that the criterion to use for selection of the form that fits "best" is to choose the form which gives the highest coefficient of determination, R^2, or the highest R, the coefficient of multiple correlation. This test criterion appears to be contrary to the general feeling. Many investigators would use the form which gives the smallest standard error of estimate of the population of Y values.

As a matter of fact, in the previous studies reported (10, 11, 12) the smallest standard error of estimate was the criterion used for selecting the form of the regression equation.

The standard error of estimate of the population of Y values measures the closeness with which the estimated values agree with the original values used to determine the regression coefficients. There are various methods for computing the estimate of the standard error of estimate; see, for example, Steel and Torrie (21) and Ezekiel and Fox (22). The standard error of estimate, however, is not a measure of the proportion of the variation in the dependent factor which can be explained by, or is associated with, variation in the independent factor or factors.

In multiple regression the relative importance of all the variables combined is measured by dividing the standard deviation of the estimated values by that of the original values (21). This ratio is called the coefficient of multiple correlation. It measures the combined importance of the several independent variables as a means of explaining the differences in the dependent factor. The square of the coefficient of multiple correlation, R^2 is called the coefficient of multiple determination. The coefficient of multiple determination is the proportion of the variance in the dependent variable which has been mathematically accounted for by regression. In other words, R^2 is the proportion of the total sum of squares attributable to regression (23). Computational methods for R and R^2 are given in references (21) and (22).

If one is faced with the problem of determining which set of independent variables gives the "best" fit for a given form of a

regression equation, say that given by equation (2), then the test criterion for "best" fit could be the standard error of estimate or the multiple correlation coefficient. This is because the residual variation for each set of independent variables is being compared with the same original standard deviation. In this case the correlation would increase as the standard error decreased.

At this point of the study it is not a question of which set of independent variables gives the "best" fit, but which transformation of the same set of variables gives the "best" fit. After transformation, the standard deviation of the dependent variable is not necessarily going to remain unchanged. Therefore, it is to be expected that the transformed and untransformed data will have different original deviations in the dependent variable. Because of this, the standard error of estimate would not necessarily decrease as the correlation increased. The former is an absolute measure whereas the latter is a relative measure. Thus, the multiple correlation coefficient was selected as the criterion for determining best fit. An approximate test for homogeneity of R values was derived. It will be discussed in more detail in a later chapter.

With only one independent variable it may be less tedious to determine the form to be used empirically by plotting the observed data on various scales and determining which scale gives a straight line fit of the data points. However, when more than one independent variable is involved it is quite difficult to plot the data. Hence, the necessity for a test criterion for choosing the form of the prediction equation.

One danger in the use of transformations is that the

transformed data may no longer meet the necessary assumptions of the statistical procedures used for deriving estimates of the regression coefficients and for testing hypotheses concerning these estimates (19). The more important assumptions which are usually affected by transformation of the data are the assumptions of normality and homoscedasticity. Often, however, one transforms data in order that the assumptions of the test criteria will be met.

For any given transformation, of equation (2) say, the assumption of normality can be tested relatively easily, if enough data is available, by the use of the chi square test for goodness of fit for continuous distributions (24). In order to test for homoscedasticity, or equality of variance, one needs several observations on the dependent variable for each of several given sets of values of the independent variables. It is not too often that one has enough data to test this hypothesis. Especially when a random and independent sampling procedure has been followed. Consequently, one does not often test this assumption.

There are certain ideal rules which should be considered when making transformations (25), they are:

(a) The variance of the transformed variate should be unaffected by changes in the mean.

(b) The transformed variate should be normally distributed.

(c) The transformed scale should be one for which an arithmetic average from the sample is an efficient estimate of the true mean.

(d) The transformed scale should be one for which real effects are linear and additive.

These were kept in mind, and as will be seen subsequently the transformation selected met the above requirements.

In an attempt to account for apparent regional differences in the cost of waste treatment the United States was divided into regions and an equation of the same form was derived for each region. The method of selection of the regions will be discussed in the next chapter.

The question arose as to whether or not the regional differences were real. This is equivalent to asking whether the same regression equation will apply for all regions of the United States. To answer this question it was decided to use the method described by Williams (26). The method tests for differences among regression coefficients, or parallelism of regression planes. If these are not significant the method then tests for differences of position or coincidence of regression planes.

A common set of regression coefficients is estimated from the combined sums of squares and cross products within the regions. The regression sum of squares is determined from the combined data, on the assumption, of course, that the regression coefficients in the populations are the same. This regression sum of squares would be the same as the sum of the regression sum of squares from each region, if the regression coefficients were in fact the same for each region. Williams (26) states that the difference between the sum of the regression sums of squares for each region and the combined regression sum of squares gives a criterion appropriate for an over-all test of differences among the coefficients.

CHAPTER V

SELECTION OF THE SAMPLE

In order to estimate the cost of waste treatment a study of presently operating waste treatment plants was conducted. Generally, whenever the task of gathering data on a large number of sampling units arises a sample study, or sample survey, will provide results of some desired precision at comparatively low cost. When one decides to proceed with the planning of the survey the following must be known at some stage of the planning (27):

(a) The population for which information is desired.

(b) The information wanted concerning this population.

(c) The required precision of the results.

If the sample is selected, and the estimate obtained, by methods that permit the use of the theory of probability, the precision of the sample estimate can be computed. As a matter of fact, methods of selecting samples based on the theory of probability are the only general methods known which can provide a measure of precision. It is necessary to be sure that the conditions imposed by the use of probability methods are satisfied (26).

It was decided that the sample should be representative of all types and sizes of treatment plants in all of the United States, and that only municipal waste treatment facilities would be studied. That

is, no treatment plant for only industrial wastes would be included in the sample.

There are in existence rough estimates of the ratio of the cost of municipal wastes to industrial waste, hence from an estimate of municipal waste cost one can obtain an estimate of industrial waste cost.

Since the scope of inference was to be all plants in the United States a mail survey was deemed more practicable, in terms of time and money, than an interview survey. The Bureau of Water Resources Research of the University of Oklahoma, under the directorship of Professor George W. Reid, agreed to pay for the cost of the survey, i.e., cost of preparation of the questionnaire, cost of mailing, and cost of self addressed and stamped return envelopes.

In probability sampling one must have a list, or frame, in order to assign a probability of selection to each sampling unit. In this study a list would be the names, types, and locations of all presently operating municipal waste treatment plants. Such a list is available in the 1957 Inventory of Municipal and Industrial Waste Facilities (28). The Inventory consists of nine volumes, one for each of the nine Public Health Service regions.

In order to account for regional differences in the cost of waste treatment, which have been noted by the author and by others (8, 9, 10, 11, 12, 13, 14), the United States was stratified into several regions, and a stratified simple random sampling plan was adopted. A stratified simple random sampling plan is one in which the sampling units of the population are divided into groups, called strata, such

that each element is contained in one and only one stratum. The sample is then chosen by selecting a simple random sample of elements from each stratum (29). A simple random sample is a sample so drawn that every combination of, say, n elements has the same chance of being selected. Consequently, a list of sampling units is needed for each stratum. After the data was collected by this method regression equations for each region could be computed, and the cost of treatment for one region could be compared with the cost in any other region.

The question arose as to how many primary strata (regions) should be used. It was first thought that the United States ought to be divided into twenty-two regions, one for each of the major water resources regions (4, 5, 6). However, it was concluded that it would be difficult to construct the necessary twenty-two lists. The water resources regions know no political boundaries. The assignment of several plants to any one of several regions might be debatable. Also, it was felt that twenty-two prediction equations would be too cumbersome and that many of the equations would be estimating the same cost. That is, the data from many of the twenty-two regions could be combined into a single prediction equation. This conclusion was quite subjective and based on personal experience of the author and fellow consulting engineers. It was finally decided that meaningful results could be obtained through the use of regions based upon political boundaries. Since the Public Health Service Inventory of Municipal Waste Facilities (28) was to be used as the population list and since the list was subdivided into nine volumes, one for each Public Health Service region, it was decided, for the sake of simplicity more than anything else, to use nine strate in the sampling

design. Table 5 gives the names of the states in each of the nine Public Health Service Regions used as primary strata in this study.

Stratification can be used to increase the reliability of sample results. The amount of increase in precision of sample estimates accomplished by stratification will depend on the degree of homogeneity that is achieved within strata. In other words, on how much of the variability in the characteristic being estimated is reflected in the differences among the strata. This in turn depends on how effectively the strata have been defined (29).

In order to obtain a representative sample of all types of plants in each state within each region proportionate stratified sampling was used. The United States was stratified by Public Health Service regions, each region was stratified by states within the region, and each state was stratified by the type of plant. The types of plant used in the stratification were Imhoff Tank, Primary, and Secondary. The list provided information for this type of stratification. Simple random sampling was performed at the last stage of stratification.

It was decided at the outset to eliminate Hawaii and Alaska and the Territories from the sampling plan, but to include the District of Columbia along with the remaining 48 states. It was also decided to eliminate from the list facilities whose only treatment process was the oxidation pond. Any treatment process such as a primary, followed by an oxidation pond was included in the list. The list provided enough information to make this decision. Septic tank treatment facilities were also eliminated from the list.

Septic tank treatment facilities and oxidation pond treatment

TABLE 5

STATES WITHIN PUBLIC HEALTH SERVICE REGIONS

Region I

Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont

Region II

Delaware, New Jersey, New York, Pennsylvania

Region III

District of Columbia, Kentucky, Maryland, North Carolina, Virginia, West Virginia

Region IV

Alabama, Florida, Georgia, Mississippi, South Carolina, Tennessee

Region V

Illinois, Indiana, Michigan, Ohio, Wisconsin

Region VI

Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota

Region VII

Arkansas, Louisiana, New Mexico, Oklahoma, Texas

Region VIII

Colorado, Idaho, Montana, Utah, Wyoming

Region IX

Arizona, California, Nevada, Oregon, Washington, Alaska, Hawaii

facilities were eliminated from the study because in any mathematical model for operation of a water basin they would not appear in any term, except perhaps as water loss or as groundwater additions. They contribute no effluent to receiving waters. The effluent from septic tanks normally flows to underground seepage beds and not directly to a stream. Most oxidation ponds, now in operation, have never overflowed their effluent weir. They act as evaporation and seepage ponds for the most part.

The population defined, the next requirement was the determination of information desired concerning this population. The variables that are to be studied were discussed in CHAPTER III. The mail survey questionnaire was constructed such that the answers to the questions either gave the variables directly or gave data from which the variables could be computed. The questions in a mail survey questionnaire must be relatively simple to answer, take very little time to answer, and be framed so that as much of the required information as possible is obtained. Questions which require calculations should not be asked. For example, rather than ask the population equivalency of the influent one needs to ask questions which will enable the sender to calculate the value. Table 6 is a copy of the questionnaire used in this study. The questionnaire along with a cover letter signed by the Director of the Bureau of Water Resources Research, University of Oklahoma, and a stamped return envelope was mailed to the Superintendent of Public Works of the community whose treatment plant was selected by the sampling scheme.

The third necessary item, of the three listed on the first page of this chapter, the required precision, and hence the sample size, was

TABLE 6

QUESTIONNAIRE USED IN MAIL SURVEY

Questionnaire for

SEWAGE TREATMENT COST STUDIES

Bureau of Water Resources Research, University of Oklahoma

Norman, Oklahoma

February, 1963

Please supply data for year 1960 (if figures are for any other year state the year).

1. Estimated Population Served _____
   (alternate: No. Water Connections_____)

2. Total Annual Flow into Treatment Plant (Million Gallons) _____
   (alternate: Average Daily Flow (mgd)_____)

3. (a) Year Construction of Plant Completed _____

   (b) Capital Cost of Plant _____

   (c) This cost included (check one of following):

      1. Plant _____, % of cost _____

      2. Pump Station _____, % of cost _____

      3. Sewers _____, % of cost _____

   (d) Annual Interest and Principal Costs _____

4. Average Annual Operation and Maintenance Costs _____
   (last 5 years) _____

_____

_____

_____

_____

TABLE 6-Continued

5. Type of Treatment Facility (check appropriate items):

Primary	Secondary	Final
Screens_____	Intermittent Sand Filter_____	Chlorination_____
Bar Racks_____	Trickling Filter:	Lagoon_____
Comminutor_____	Standard Rate____	Other(specify)___
Grit Chamber_____	High Rate_____	
Septic Tank_____	Aeration:	
Imhoff Tank_____	Activated Sludge___	
Settling Basin____	Contact Aerators___	
Lagoon (raw)_____	Biosorption_____	
Chemical_____	Percolation beds_____	
Effluent Outfall greater than 1/2 mile___	Sub Surface Application_____	
Other (specify)_____	Lagoon_____	
	Digestor_____	
	Complete Oxidation_____	
	Other (specify)_____	

6. (a) Biochemical Oxygen Demand (BOD) of Influent, ppm (parts per million) _____

(b) BOD of Effluent, ppm _____

7. (a) Solids Concentration of Influent, ppm _____

(b) Solids Concentration of Effluent, ppm _____

8. *(a) Nitrogen (as N) concentration of Influent, ppm _____

*(b) Nitrogen (as N) concentration of Effluent, ppm _____

9. *(a) Phosphorus (as PO_4) concentration of Influent, ppm _____

*(b) Phosphorus (as PO_4) concentration of Effluent, ppm _____

*Desirable if available but not necessary as it is realized that not many plants keep this information.

TABLE 6-Continued

10. Type of Sewer System Connected to Plant (check one):

(a) Combined (storm and sanitary) _____

(b) Separate (sanitary) _____

11. We desire a copy of the results of your study (check one):

(a) Yes _____

(b) No _____

the most difficult to specify. In order to intelligently specify a required precision one needs to know something about the distribution of the variable under study. In particular, he should have an estimate of the population variance. The precision is usually specified by defining the expected width of a specified confidence interval. One may specify, for example, that the expected width of a 95 per cent confidence interval for the population mean be of size w. For two populations with the same means and dimensions, if the variance of the first population is quite large and the variance of the second population is small then a larger confidence interval would be specified for the first population than would be specified for the second population. Thus, if one assumed too small a variance, when it actually was large, specified a narrow width, and calculated the sample size it is doubtful that his results would be meaningful. If on the other hand he did just the opposite, when the variance was in reality small, money would have been wasted

since a smaller sample would have yielded quite good results. Estimates of population variances can often be found from a perusal of the literature in the field of study.

In order to estimate the optimum sample size within each stratum of a stratified simple random sampling scheme one must know or have an estimate of the variance within each stratum. Optimum sample size is the selection of the number of observations from each stratum such that the overall variance is a minimum under the assumption of equal costs between strata. In this study costs between strata were equal. However, no estimate of the variance within strata, or regions, was available. There was available an overall estimate, that is an estimate for the entire United States, of the variance for each type of treatment plant considered in this study. Use was made of these data to determine the total sample size. In the article by Rowan, Jenkins, and Butler (11) there are graphs for estimating cost of sewage treatment construction for several types of treatment plants. These graphs include, in addition to the expected cost curve, a 68 per cent confidence belt. The sample size used to derive each graph is also given. For each type of treatment plant the standard error of estimate was computed. Using the standard errors of estimate 95 per cent confidence intervals were computed. The values obtained for the intervals by this procedure varied from six dollars per capita for trickling filter plants to fourteen dollars per capita for Imhoff tank plants.

Considering the sampling procedure and the proposed statistical analyses to be used in this study a 95 per cent confidence interval width

of six to eight dollars per capita was chosen. Using the required precision and the estimates of the variances, the number of observations required for each type of treatment plant was computed using the relationship (30),

$$n = \frac{t_1^2 s^2}{d^2}, \tag{7}$$

where n is the number of observations required, s^2 is an estimate of the population variance obtained from a previous study, t_1 is the tabulated t value for the desired confidence level and the degrees of freedom of the initial sample which were used in the computation of s^2, and d is the half-width of the desired confidence interval. The number of observations computed for each type of plant were added to give an estimate of total sample size. For width of interval of six dollars per capita this sum was 408, and for a width of interval of eight dollars per capita it was 230. It should be pointed out that the figures of six and eight dollars per capita are the interval sizes for the mean of the dependent variable when the independent variables take on their mean values.

Since mail surveys usually result in a low response rate it was hoped that more questionnaires than required could be sent out. However, only enough financial support for about 545 copies and stamped envelopes was available.

Since no estimates of the variance within strata were available it was not possible to estimate the number of observations required from each region. Consequently, a proportionate sampling plan was adopted. That is, the number of plants selected in a given region was in proportion

to the total number of plants in the region. For example, 7 per cent of the total number of treatment plants in the United States, on the list, were in Region I, therefore 7 per cent of the total number of questionnaires, 40, were to be sent to plants in Region I. The number of questionnaires to be sent to each state within each region was also determined on a proportion to size basis. For example, 30 per cent of the treatment plants on the list in Region I were in Connecticut, thus 30 per cent of the 40, or 12, questionnaires were to be sent to Connecticut. The number of questionnaires to be sent to each type of treatment facility was again based on a proportion to size basis. Thus, in Connecticut where 42 per cent of the treatment plants were of the secondary type 0.42 times 12 or 5 questionnaires were sent to communities with this type plant.

Each plant on the list was assigned a number and a table of random numbers was used to insure selection of a random sample. The questionnaires, 545 of them, were mailed at the beginning of March, 1963. It was decided that the processing of the data would begin on May 23, 1963, and that any data received after that day would be discarded. Only one questionnaire was returned after that date.

As of May 23, 1963, 219 of the questionnaires had been returned, for a response proportion of 0.401. Of these 219, twenty-two had to be discarded because of various reasons, such as, no cost data being given, data given was illegible, etc. Thus, the total number of questionnaires returned was not large enough even for the eight dollar per capita width.

There were not too many options which could be followed at this

point. First of all, one could have proceeded with the analysis using just the 219 observations. However, it was possible that the population of inference was not what it was originally thought to be. It turns out often that the non-responders are quite different from the responders. If this is the case, then the population of inference is only the responders. If it is not the case, then the population of inference is as it was originally described. One way to determine this would be by sampling the non-responders by an interview survey. This of course could take much time and money.

Secondly, one could send out reminder letters to the non-responders. This method would result in several more returns, but it would be, again, time consuming and a financial burden. Even after this one would still have a problem with the remaining non-responders.

It seemed apparent that data from an interview survey of a number of non-responders was needed. Fortunately, it was discovered that the United States Public Health Service had interview data available from an operation and maintenance cost study (12). They graciously offered the use of their data to the author. Interview data on 252 of the 326 non-responders was available. This was copied and a study comparing the responders to non-responders was made.

It was assumed that the logarithm of the cost per population equivalency was normally distributed. This was tested later by the goodness of fit test (24) and the conclusion was that there was no evidence to reject the hypothesis of normality at the 0.05 level. Using the assumption of normality it was hypothesized that the responders and non-responders came from the same normal population. For each region,

the hypothesis of equality of population variances was tested using the fact that the ratio of two independent variables each distributed as chi square and each divided by its degrees of freedom is distributed as Snedecor-Fisher's F (30). That is,

$$F = \frac{s_1^2}{s_2^2},$$

where s_1^2, is larger of the two sample variances for responders and non-responders, and s_2^2 is the smaller sample variance. The numerator degrees of freedom are those for s_1^2 and the denominator degrees of freedom those for s_2^2. For all nine regions, tested separately, there was no evidence to reject the hypothesis of equality of variance at the 0.05 level.

Having no evidence to reject the hypothesis of equality of variance the hypothesis of equality of population means for each region was tested using the unpaired t test (30). For each of the nine regions there was no evidence to reject the null hypothesis of no difference in the means for responders and non-responders. Thus, it was assumed that the responders and non-responders came from the same normal population. Consequently, the Public Health Service data was added to that collected from the mail survey.

Table 7 shows the number and per cent of questionnaires sent to each region, the number and per cent returned and the number of non-responders for which the Public Health Service data was used.

Adding the 252 from the Public Health Service to the 219 returned in the survey gave a total of 471 questionnaires. Twenty-two of

TABLE 7

RESULTS OF MAIL SURVEY

Region	Questionnaires				
	No. Sent	Per Cent of Total Sent	No. Returned	Per Cent Returned	No. Obtained From PHS
I	40	7	22	55	16
II	55	10	23	42	25
III	65	12	19	29	43
IV	65	12	32	49	28
V	75	14	31	41	38
VI	60	11	22	37	20
VII	50	9	19	38	11
VIII	40	7	16	40	17
IX	95	17	35	37	54
Total	545	100	219		252

the mail survey questionnaires and six of the Public Health Service questionnaires were discarded, leaving a total of 443 entirely or partially usable questionnaires. This number exceeded the number required for a six-dollar per capita confidence interval. Hence, a valid analysis could be performed with the population of inference the same as previously stated.

CHAPTER VI

RESULTS OF THE ANALYSES OF THE DATA

Having combined the data from the survey with the Public Health Service data analyses were conducted to determine which of the five equation forms shown in CHAPTER IV gave the best fit. As previously stated in CHAPTER IV the criterion for best fit would be use of the multiple correlation coefficient, or the square of it, the coefficient of determination. There was some indication from past studies (11, 12) that the form of the equation for estimation of construction costs which gave the best fit might differ from the form which gave best fit for annual operation and maintenance costs. Studies were conducted using one of these dependent variables at a time, in each of the five equations given in CHAPTER IV.

Best Fit Study for Prediction of Construction Costs

The following variables were used in this analysis:

Y_1 = Construction cost per PE produced,

X_1 = Design Population in thousands,

X_2 = Design Flow in mgd,

X_3 = BOD of Influent in ppm,

X_4 = BOD of Effluent in ppm,

X_5 = Type of Treatment Plant (see CHAPTER III).

The analysis was performed by pooling all of the data for the United States. This indicates that the assumption was made that all nine regions were homogeneous regarding construction costs. In other words, no real regional differences in construction costs existed. It was not believed that there were no regional differences. The procedure was followed for the sake of expediency and simplicity. It was believed that the form which resulted in the best fit would have the highest R value for the total United States as well as for the nine regions individually with the R value for the United States being smaller than the R value for most of the nine regions. This belief was tested and shown correct for two forms during the operation and maintenance cost studies. The form which gave the best fit using R as the test criterion was:

$$\ln Y_1 = b_o + b_1 \ln X_1 + b_2 \ln X_2 + b_3 \ln X_3 + b_4 \ln X_4 + b_5 \ln X_5.$$

Table 8 is a summary of the results of this study.

As can be seen from this table the R value for the form of equation chosen is about 1.71 times that of the next largest value. The table also gives the values of the residual mean squares, or estimates of the error variances. Because of the transformations performed these residual mean squares have different dimensions and hence, are not directly comparable. This is an illustration of the argument presented in CHAPTER IV for using R as the test criterion for selecting the form which gives the best fit.

Best Fit Study for Prediction of Operation and Maintenance Costs

The variables used in this analysis were: Y_6, Annual Operation

TABLE 8

SELECTION OF FORM OF EQUATION TO BE USED

FOR PREDICTION OF CONSTRUCTION COSTS

Form of Equation	R	Residual Mean Square
$Y_1 = b_o + \sum_{i=1}^{5} b_i X_i$	0.26	103.244
$\ln Y_1 = b_o + \sum_{i=1}^{5} b_i X_i$	0.41	0.798
$\ln Y_1 = b_o + \sum_{i=1}^{5} b_i \ln X_i$	0.70	0.488
$1/\ln Y_1 = b_o + \sum_{i=1}^{5} b_i \ln X_i$	0.15	75.429
$1/Y_1 = b_o + \sum_{i=1}^{5} b_i X_i$	0.38	0.122

R = Coefficient of Multiple Correlation = $\left[\frac{\text{Regression sum of squares}}{\text{Total sum of squares}}\right]^{\frac{1}{2}}$

and Maintenance Cost per Capita, and X_1, X_2, X_3, X_4 and X_5 as defined above. Again, the data were pooled to arrive at five regression equations for the total United States. The analysis indicated that the form:

$$\ln Y_6 = b_o + b_1 \ln X_1 + b_2 \ln X_2 + b_3 \ln X_3 + b_4 \ln X_4 + b_5 \ln X_5,$$

resulted in the highest R value, and hence would be used in further studies. Table 9 gives the values of R and the residual mean squares for each of the five forms investigated.

TABLE 9

SELECTION OF FORM OF EQUATION TO BE USED FOR PREDICTION OF OPERATION AND MAINTENANCE COSTS

Form of Equation	R	Residual Mean Square
$Y_6 = b_o + \sum_{i=1}^{5} b_i X_i$	0.18	4.763
$\ln Y_6 = b_o + \sum_{i=1}^{5} b_i X_i$	0.30	0.552
$\ln Y_6 = b_o + \sum_{i=1}^{5} b_i \ln X_i$	0.52	0.439
$1/\ln Y_6 = b_o + \sum_{i=1}^{5} b_i \ln X_i$	0.13	20.492
$1/Y_6 = b_o + \sum_{i=1}^{5} b_i X_i$	0.25	0.186

The R value for the form of equation selected is about 1.73 times that

the next largest R value.

The study by Rowan, Jenkins, and Howells (12) resulted in the use of an equation of the form $1/\ln Y = a + b \ln X$, where Y is annual operation and maintenance costs per capita and X is population. Because of this, it was decided to fit the form selected in this study and the form selected in their study to determine which gave the best fit for each region. The test criterion for best fit was, once again, the highest R value. The variables used were Y_6, the annual operation and maintenance cost per capita, X_1, the population, and X_5, the type of treatment plant. For each of the nine regions the R value was higher for the form $\ln Y_6 = b_o + b_1 \ln X_1 + b_5 \ln X_5$. Table 10 gives a summary of the results of the analyses.

Having selected the "best" form of the equation, the next step was to develop regression equations for construction costs and operation and maintenance costs for each of the nine regions. The studies for each of these types of cost will be discussed separately.

Construction Cost Studies

As stated in CHAPTER III, under given circumstances it may be of benefit to be able to estimate construction costs as cost per PE produced or cost per PE treated or cost per capita. Regression equations for all three dimensions of this dependent variable were derived.

It might be well at this point to define the variables, and their dimensions, that were investigated in these studies:

Y_1 = Construction cost per PE produced, in 1913 dollars

Y_2 = Construction cost per PE treated, in 1913 dollars,

TABLE 10

COMPARISON OF TWO FORMS OF EQUATIONS

FOR EACH OF NINE REGIONS

Region		Form: $\ln Y_6 = b_0 + b_1 \ln X_1 + b_5 \ln X_5$	$1/\ln Y_6 = b_0 + b_1 \ln X_1 + b_5 \ln X_5$
I	R	0.41	0.14
	Res MS	0.591	147.952
II	R	0.50	0.46
	Res MS	0.654	453.537
III	R	0.46	0.25
	Res MS	0.409	74.022
IV	R	0.56	0.27
	Res MS	0.518	9.492
V	R	0.53	0.18
	Res MS	0.386	9.687
VI	R	0.44	0.25
	Res MS	0.355	60.229
VII	R	0.37	0.18
	Res MS	0.558	9.996
VIII	R	0.79	0.40
	Res MS	0.423	7.371
IX	R	0.66	0.42
	Res MS	0.223	48.776

R = Coefficient of Multiple Correlation

Res MS = Residual Mean Square

Y_3 = Construction cost per capita, in 1913 dollars,

X_1 = Design population x 0.001,

X_2 = Design flow, in mgd,

X_3 = BOD of influent, in ppm,

X_4 = BOD of effluent, in ppm,

X_5 = Type of Plant (see CHAPTER III),

X_6 = PE of influent x 0.001.

The cost data used in the study was for construction of the plant. It did not include cost of land, pumping station, interceptors or influent sewers. Prior to any studies, all of the construction costs were adjusted by means of the United States average Engineering News-Record Construction Cost Index base year 1913, taken as 100. Consequently, the regression equations will give cost estimates in 1913 dollars. To convert 1913 dollars to, say, 1962 dollars multiply the cost estimates given by the regression equations by the ratio, index for year of interest to 100, in this example, 871.84/100 or 8.7184.

Studies of Construction Cost per PE Produced

A relationship among Y_1 and X_1, X_2, X_3, X_4, and X_5 and a relationship among Y_1 and X_4, X_5, and X_6 was derived. The independent variable X_6 can be computed from a knowledge of X_2 and X_3 (see CHAPTER III). The independent variable X_1 was eliminated in the second relationship because it was felt that there might be a direct relationship between it and X_6. Consequently, two equations for estimating construction cost per PE produced were derived for each of the nine regions. The analysis of variance for each equation for each region indicated

that the sum of squares attributable to regression was significant either at the .05 or .01 level or both.

In five regions, II, IV, V, VI, IX, there was an increase in the sample variance, or residual mean square, when fewer independent variables were used in the estimation equation. These increases ranged from one per cent for Region VI to twenty per cent for Region IX. On the other hand there was a decrease in the residual mean square for four regions, I, III, VII, VIII, with the use of fewer independent variables. The decreases ranged from two per cent for Region III to twenty-five per cent for Region I.

The decision to use fewer independent variables in the estimation equation is not difficult in the case of the four regions for which there was a decrease in the residual mean square, however, for the regions for which there was an increase in residual mean square the only rule for determining whether or not to use fewer independent variables is to decide how much of an increase in variance one is willing to accept. Use of the equation with fewer variables was the equation of preference since it was felt that the decrease in precision was not great enough to warrant a more involved estimation equation. Table 11 presents the results of the study using three independent variables to estimate construction cost per PE produced. In the table df is degrees of freedom corresponding to the residual mean square, n is the number of observations and Res MS is residual mean square. The 95 per cent confidence limits for an estimated cost value is given by,

$$CL = \ln Y_1 \pm t\, s_{\ln Y_1},$$

where

CL = 95% confidence limits,

$\ln Y_1$ = estimated expected value for a given set of X's (X_6, X_4, X_5), (ln is the base e logarithm),

t = "student's t" value for confidence coefficient of 0.95 and degrees of freedom corresponding to the residual mean square,

and

$s_{\ln Y_1}$ = standard error of estimate for $\ln Y_1$ for which the set of X's is (X_6, X_5, X_5). The dimensions of $s_{\ln Y_1}$ are logarithmic (base e) units. The antilog (base e) of $\ln Y_1$ gives the expected construction cost in dollars per, PE produced, and the antilog (base e) of the upper and lower values of the computed confidence limits give the 95 per cent confidence limits in dollars per PE produced.

The next step was to test for differences among the regression coefficients, or parallelism of regression planes, for the nine regions using the method described by Williams (26). The F value, so obtained, in the analysis of variance was highly significant, (p less than .01), leading one to reject the hypothesis of parallelism. In other words, the same regression equation is not applicable to all regions. This does not exclude the combination of regions. It implies only that there must be at least two estimation equations. There is no known test analogous to the Duncan's multiple range test for all possible comparisons (32), and the only recourse would be to test every possible pair of equations using Williams' method. This would result in non-independent tests for which the Type I error would be unknown (32). Rather than

TABLE 11

EQUATIONS FOR ESTIMATING CONSTRUCTION COST PER PE PRODUCED

Region	Regression Coefficients				C_{ij}								
	b_o	b_4	b_5	b_6	C_{44}	C_{45}	C_{46}	C_{55}	C_{56}	C_{66}	df	n	ResMS
I	+3.926	-0.202	-0.751	-0.390	0.031	0.067	-0.004	0.373	-0.019	0.010	34	38	0.442
II	+3.107	-0.113	-0.003	-0.264	0.046	0.048	-0.004	0.178	-0.008	0.008	38	42	0.387
III	+3.187	-0.125	-0.195	-0.379	0.040	0.058	-0.003	0.186	-0.009	0.006	58	62	0.453
IV	+3.603	-0.336	+0.340	-0.321	0.025	0.044	-0.005	0.174	-0.007	0.007	55	59	0.565
V	+4.161	-0.420	-0.102	-0.320	0.030	0.050	-0.004	0.186	-0.014	0.006	58	62	0.419
VI	+3.472	-0.331	-0.056	-0.248	0.119	0.138	-0.016	0.384	-0.017	0.009	30	34	0.472
VII	-0.836	+0.517	+1.620	-0.261	0.142	0.148	-0.003	0.461	-0.005	0.027	20	24	0.349
VIII	+1.961	-0.054	+1.016	-0.434	0.065	0.089	+0.001	0.690	+0.050	0.021	19	23	0.334
IX	+2.821	-0.111	+0.139	-0.388	0.019	0.028	-0.002	0.113	0.000	0.006	82	86	0.393

Region	Deviations		
	x_4	x_5	x_6
I	$\ln X_4 - 3.938$	$\ln X_5 - 0.183$	$\ln X_6 - 2.899$
II	$\ln X_4 - 3.814$	$\ln X_5 - 0.463$	$\ln X_6 - 3.178$
III	$\ln X_4 - 3.871$	$\ln X_5 - 0.448$	$\ln X_6 - 2.427$
IV	$\ln X_4 - 3.510$	$\ln X_5 - 0.499$	$\ln X_6 - 2.706$
V	$\ln X_4 - 3.726$	$\ln X_5 - 0.547$	$\ln X_6 - 2.386$
VI	$\ln X_4 - 3.531$	$\ln X_5 - 0.616$	$\ln X_6 - 2.682$
VII	$\ln X_4 - 3.307$	$\ln X_5 - 0.663$	$\ln X_6 - 2.877$
VIII	$\ln X_4 - 3.586$	$\ln X_5 - 0.645$	$\ln X_6 - 2.895$
IX	$\ln X_4 - 3.723$	$\ln X_5 - 0.505$	$\ln X_6 - 3.348$

Regression equation:

$$\ln Y_1 = b_o + b_4 \ln X_4 + b_5 \ln X_5 + b_6 \ln X_6$$

Standard error of an estimated expected value:

$$s_{\ln Y_1} = (\text{ResMS}\,(1/n + C_{44}x_4^2 + C_{55}x_5^2 + C_{66}x_6^2 + 2C_{45}x_4x_5 + 2C_{46}x_4x_6 + 2C_{56}x_5x_6))^{\frac{1}{2}}$$

make non-independent comparisons it was decided to use nine regression equations.

Table 12 gives a summary of the test for parallelism of regression planes.

TABLE 12

TEST OF DIFFERENCE AMONG REGRESSION COEFFICIENTS DERIVED IN STUDIES OF CONSTRUCTION COST PER PE PRODUCED

Source	df	SS	MS	F
Combined Regression	3	174.0		
Difference of Regressions	24	31.5	1.31	3.05*
Combined Residual	394	171.3	0.43	
Total Within Groups	421	376.8		

*Significant at p less than 0.01

Studies of Construction Cost Per PE Treated

In these studies a relationship among Y_2 and X_1, X_2, X_3, X_4, and X_5, and a relationship among Y_2 and X_6, X_5, and X_4 were derived. The analysis of variance for regression for each region was significant at the 0.05 or 0.01 level. The use of three independent variables, in place of five, in the derivation of the regression equations resulted in an increase in the residual mean square for every region except Region I, where a twenty per cent decrease occurred. The per cent increase in residual mean square ranged from 10 to 14, with a modal value of 11. Experience indicates that this is not too great a loss in

precision. Thus the use of the regression equations with three independent variables might be preferred to the use of the equations with five independent variables.

Table 13 contains the results of the study for estimation of cost per PE treated. The test for differences among regression coefficients, or parallelism of regression planes, for the equations in Table 13 resulted in a highly significant, (p less than .01), F value. The interpretation of this was that the same regression plane would not apply for all regions.

Studies of Construction Cost Per Capita

Two equations, one relating Y_3 and X_1, X_2, X_3, X_4 and X_5, and one relating Y_3 and X_1, X_4, and X_5, were derived for each region in the construction cost per capita studies. Since it was a per capita study the variable population equivalency was not used in either equation. In order to get total cost one must have an estimate of design population, that is, total cost is the product of cost per capita and design population.

The analysis of variance for each equation for each region showed that the sum of squares attributable to regression was significant at the .05 or .01 level. The equations with three independent variables instead of five resulted in a higher residual mean square in four regions, in a lower residual mean square in four regions, and in the same residual mean square for one region. Experience indicates that the increases were not too great. Thus, one ought to be able to estimate cost per capita by an equation involving three independent variables

TABLE 13

EQUATIONS FOR ESTIMATING CONSTRUCTION COST PER PE TREATED

Region	Regression Coefficients b_o	b_4	b_5	b_6	C_{ij} C_{44}	C_{45}	C_{46}	C_{55}	C_{56}	C_{66}	df	n	ResMS
I	+3.664	+0.077	-1.326	-0.401	0.031	0.067	-0.004	0.373	-0.019	0.010	34	38	0.491
II	+2.747	+0.155	-0.450	-0.253	0.046	0.048	-0.004	0.178	-0.008	0.008	38	42	0.621
III	+2.263	+0.250	-0.651	-0.342	0.040	0.058	-0.003	0.186	-0.009	0.006	58	62	0.640
IV	+4.335	-0.317	-0.519	-0.297	0.026	0.044	-0.005	0.174	-0.007	0.008	53	57	0.651
V	+4.296	-0.261	-0.729	-0.320	0.030	0.051	-0.004	0.190	-0.015	0.007	57	61	0.509
VI	+3.074	-0.077	-0.478	-0.241	0.119	0.001	-0.016	0.384	-0.017	0.009	30	34	0.432
VII	-1.737	+0.868	+1.826	-0.315	0.142	0.148	-0.003	0.461	-0.005	0.027	20	24	0.461
VIII	+2.478	-0.026	+0.141	-0.391	0.065	0.089	+0.001	0.690	+0.050	0.021	19	23	0.545
IX	+2.179	+0.156	-0.152	-0.336	0.019	0.029	-0.002	0.113	0.000	0.006	82	86	0.529

Region	Deviations x_4	x_5	x_6
I	$\ln X_4$ - 3.938	$\ln X_5$ - 0.183	$\ln X_6$ - 2.899
II	$\ln X_4$ - 3.814	$\ln X_5$ - 0.463	$\ln X_6$ - 3.178
III	$\ln X_4$ - 3.871	$\ln X_5$ - 0.448	$\ln X_6$ - 2.427
IV	$\ln X_4$ - 3.459	$\ln X_5$ - 0.516	$\ln X_6$ - 2.740
V	$\ln X_4$ - 3.736	$\ln X_5$ - 0.541	$\ln X_6$ - 2.411
VI	$\ln X_4$ - 3.531	$\ln X_5$ - 0.616	$\ln X_6$ - 2.682
VII	$\ln X_4$ - 3.307	$\ln X_5$ - 0.663	$\ln X_6$ - 2.877
VIII	$\ln X_4$ - 3.586	$\ln X_5$ - 0.645	$\ln X_6$ - 2.895
IX	$\ln X_4$ - 3.723	$\ln X_5$ - 0.505	$\ln X_6$ - 3.348

Regression equation:

$$\ln Y_2 = b_o + b_4 \ln X_4 + b_5 \ln X_5 + b_6 \ln X_6$$

Standard error of an estimated expected value:

$$s_{\ln Y_2} = (\text{ResMS}\,(1/n + C_{44}x_4^2 + C_{55}x_5^2 + C_{66}x_6^2 + 2C_{45}x_4x_5 + 2C_{46}x_4x_6 + 2C_{56}x_5x_6))^{\frac{1}{2}}$$

with about as much precision as one with five independent variables. The estimation equations will not be given here since the use of cost per PE produced or per PE treated is preferred. The equations are on file if needed.

The question arose as to whether fewer than three independent variables might be used in the regression equations. It was found that variable X_6, population equivalency, in the cost per PE produced and cost per PE treated studies contributed a highly significant amount to the regression sum of squares. The variable X_1, design population, used in the cost per capita studies was also found to contribute a highly significant amount to the regression sum of squares. The conclusion was that these two variables could not be deleted from their respective regression equations without a great loss in precision.

The variable X_5, type of treatment plant, contributed a significant amount to the regression sum of squares for Regions I, II, III, IV, V, VII, and VIII. Thus, variable X_5 could not be deleted from the equations for these regions without great loss in precision. However, it could be deleted from Regions VI, and IX without much loss in precision. The variable X_4, BOD of effluent, was found to contribute a significant amount to the regression sum of squares for Regions III, IV, V, VI, VII and IX. Thus, variable X_4 could be deleted from the equations for Regions I, II and VIII. It seemed reasonable that both variables, X_4 and X_5, should contribute a significant amount to the regression sum of squares for each region. It is possible that the sample size was too small for those regions where they did not contribute significantly.

One thought that occurred was that the BOD of effluent, X_4, and the type of treatment plant variable, X_5, were measuring the same characteristic, efficiency of the plant. However, if this actually were the case it is hardly likely that both variables would contribute a significant amount to the regression sum of squares in any one region. They did so in Regions III, IV, V, and VII. The variable X_5, type of treatment plant, is a measure of the expected effectiveness of a treatment plant, whereas the BOD of the effluent is a direct measure of the actual efficiency. What a treatment plant is capable of doing and what it actually does are rarely the same. The actual performance of a plant is a measure of the operation and maintenance of the plant and of variation in actual loading conditions, which are not the same as design conditions. Thus, it is believed that the two variables are not measuring the same phenomenon.

It was decided to use the equations given in Tables 11 and 13 for estimating construction costs because the use of the three variables contributed a significant amount to the regression sum of squares for four regions and because it seemed simpler to use the same form of equation with the same number of variables for each region. The use of a variable which does not contribute significantly to the regression sum of squares cannot decrease the precision of an estimate. It does, however, result in a loss of one degree of freedom for the residual mean square. If the loss of a degree of freedom is relatively unimportant as is the case for each of the Regions, I, II, VI, VIII, and IX, then there is no serious objection to including the third variable.

Comparison of Construction Cost Equations

The three equations derived for estimating construction cost differ in the dimensions of the cost. The dimensions were 1913 dollars per PE produced, 1913 dollars per PE treated, and 1913 dollars per capita. A comparison among the three equations was made to determine the differences in precision of the construction cost estimation equation. Table 14 gives the coefficients of multiple correlation and the residual mean squares for each equation for each region.

The multiple correlation coefficient of the cost per PE produced equation for Regions II, III, IV, V, VI, VIII and IX is higher than those for the other two. The inference would be that the cost per PE produced equation estimated cost with greater precision than did the other two equations for seven out of the nine regions and thus is the one of choice. However, if there were a statistical test for homogeneity of multiple correlation coefficients the apparent differences might be non-significant.

There is a test for homogeneity of simple or partial correlation coefficients using Fisher's transformation (33). This test cannot be extended for testing homogeneity of multiple correlation coefficients as they range in value from zero to plus one (34), rather than from minus one to plus one as do simple or partial correlation coefficients. The fact that the multiple correlation coefficient (R) cannot be negative means that Fisher's z value cannot be negative. Hence, it cannot be assumed to be approximately normally distributed under Fisher's transformation.

An approximate test for homogeneity of multiple correlation

TABLE 14

COMPARISON OF CONSTRUCTION COST EQUATIONS

Region		Equation		
		Cost/PE Produced	Cost/PE Treated	Cost/Capita
I	R	0.78	0.82	0.73
	ResMS	0.42	0.491	0.363
II	R	0.65	0.59	0.65
	ResMS	0.387	0.621	0.234
III	R	0.71	0.70	0.68
	ResMS	0.453	0.640	0.495
IV	R	0.73	0.63	0.70
	ResMS	0.565	0.659	0.563
V	R	0.75	0.71	0.56
	ResMS	0.419	0.509	0.396
VI	R	0.71	0.66	0.59
	ResMS	0.472	0.432	0.532
VII	R	0.73	0.74	0.66
	ResMS	0.349	0.461	0.394
VIII	R	0.86	0.70	0.79
	ResMS	0.334	0.545	0.401
IX	R	0.73	0.57	0.65
	ResMS	0.393	0.529	0.367

coefficients was developed using a procedure similar to Bartlett's test for homogeneity of variance and using the fact that $-(n-1-\frac{2p+5}{6})\ln R$ is distributed approximately as chi square with $(\frac{1}{2}\, p(p-1))$ degrees of freedom (35), where, n is the number of observations, p is the number of variables, dependent plus independent, R is the multiple correlation coefficient, and ln is the base e logarithm. The method is illustrated below for Region V.

Test for Homogeneity of R Values

	n	p	R	ln R	$-(n-1-\frac{2p+5}{6})=k$	k ln R
Cost/PE Prod.	62	4	0.75	-.28768	-48.833	16.92508
Cost/PE Treat.	61	4	0.71	-.34249	-57.833	19.80722
Cost/Capita	62	4	0.56	-.57982	-58.833	34.11255
					-175.499	70.84485

Pooled R^2 = Sum of Regression Sum of Squares divided by sum of Total Sum of Squares for each equation

$$= \frac{70.425}{146.820} = 0.48$$

Pooled R = 0.69

-175.499 ln (0.69) = 65.12059

chi square = 70.845 - 65.121 = 5.724 with 2 df

The tests indicated that there was no evidence to reject the hypothesis of homogeneity of R values at the 0.05 level for every region. However, the chi square values for Regions IX and V were borderline. If the smallest R value in each region had been 0.01 smaller the chi square would have been significant. Since the test is approximate, in all likelihood a significant difference exists among the R values in

Regions V and IX.

Thus it was shown, by an approximate method, that the three equations for Regions I, II, III, IV, VI, VII, and VIII would estimate construction cost with about equal precision. For Regions V and IX, it appeared that the cost per PE produced equation might estimate construction cost with greater precision than either of the other equations.

Study of Operation and Maintenance Costs

The operation and maintenance costs, unlike the construction costs, cannot be referred to a base year. Thus, it would be desirable to get information from treatment plants for several years, determine if there is a change in cost with time, and derive a relationship based on the results of this determination. Almost all of the respondents to the mail survey questionnaires reported operation and maintenance costs for year 1960 only. The Public Health Service operation and maintenance costs data were for 1955 through 1958. In most cases the annual costs for these years was of approximately the same magnitude. It was assumed in these studies that the operation and maintenance costs were the same for both 1958 and 1960. There was insufficient information to test this assumption statistically for there were only relatively few plants in the Public Health Service study for which there was information collected by the mail survey. A visual inspection of the few plants common to both studies indicated that the assumption might be valid. Consequently, the operation and maintenance costs apply to the year 1960. These costs do not include annual interest and payments to principal for the plant, nor do they include costs of billing and

collection of sewer charges. They include only those costs directly related to the actual operation and maintenance of the sewage treatment plant.

The following is a list of variables investigated in the operation and maintenance cost studies:

Y_4 = Annual operation and maintenance cost per PE produced per year, in dollars x 1000,

Y_5 = Annual operation and maintenance cost per PE treated per year, in dollars x 1000,

Y_6 = Annual operation and maintenance cost per capita, in dollars,

X_1 = Population x .001,

X_2 = Flow, in mgd,

X_3 = BOD of influent, in ppm,

X_4 = BOD of effluent, in ppm,

X_5 = Type of plant (see CHAPTER III),

X_6 = PE of influent x 0.001.

Studies of Operation and Maintenance Cost per PE Produced per Year

In these studies relationships among Y_4 and X_1, X_2, X_3, X_4, and X_5 and among Y_4 and X_6, X_4 and X_5 were derived. The residual mean square for seven of the nine regions was smaller for model with a greater number of variables. However, experience indicated that not much precision would be lost in the use of the equation with fewer variables. The test for differences among the regression coefficients, or

parallelism of regression planes, for the nine regions, using the method described by Williams (26) resulted in a highly significant (p less than .01) F value. The inference is that there is a highly significant difference among regions as far as estimation of operation and maintenance costs are concerned. The equations are not presented herein because it was found that one could estimate operation and maintenance cost with relatively high precision using only population and type of treatment plant. They are on file if needed.

Studies of Annual Operation and Maintenance Cost per PE Treated per Year

Two equations for each region were derived in these studies. The analysis of variance for each equation indicated that the regression sum of squares was significant at the 0.01 or 0.05 level. The residual mean square increased for all nine equations in which fewer independent variables were used. However, the increase was relatively small for all regions. The test for parallelism of regression planes was highly significant, p less than 0.01. That is, it is quite unlikely that a single regression plane would apply to all nine regions. The equations are not given herein as it was found that use of population and type of treatment plant would estimate the cost with relatively high precision. They are on file if required.

Studies of Annual Operation and Maintenance Costs per Capita

In these studies three equations for each region were derived. The first equation expressed a relationship among Y_6 and X_1, X_2, X_3, X_4,

X_5, the second a relationship among Y_6 and X_1, X_4 and X_5, and the third a relationship among Y_6 and X_1 and X_5. All of these variables have been defined at the beginning of the section entitled Operation and Maintenance Cost Studies. Table 15 gives the values of the multiple correlation coefficients and residual mean squares obtained in the studies. Not much precision is lost when fewer variables are used. There is a gain in precision using fewer variables for three regions.

Table 16 gives the results of the study using two independent variables to estimate operation and maintenance cost per capita.

The F value obtained in the test for differences among regression coefficients was highly significant, p less than 0.01. Thus the hypothesis that the same regression plane would apply for all regions was rejected.

It appeared from the results of the operation and maintenance cost per capita studies that perhaps two or fewer independent variables would estimate the cost with only a minor loss in precision. Consequently, tests were made to determine whether or not the use of a particular independent variable contributed a significant amount to the regression sum of squares.

In the per capita studies, all three variables, X_1, X_4 and X_5, contributed a significant amount (at the 0.05 or 0.01 level) to the regression sum of squares for Region II. For Regions I, III, and IX, variables X_1 and X_5 contributed a significant amount to the regression sum of squares, X_4 did not. For Regions IV, V, VI, VII and VIII, variable X_1 was the only variable that contributed significantly to the regression sum of squares.

TABLE 15

VALUES OF MULTIPLE CORRELATION COEFFICIENTS AND RESIDUAL MEAN SQUARES OBTAINED IN THE STUDIES OF OPERATION AND MAINTENANCE COST PER CAPITA

Region		Equation $\ln Y_6 = b_o + \sum_{i=1}^{5} b_i \ln X_i$	Equation $\ln Y_6 = b_o + b_1 \ln X_1 + b_4 \ln X_4 + b_5 \ln X_5$	Equation $\ln Y_6 = b_o + b_1 \ln X_1 + b_5 \ln X_5$
I	R	0.59	0.46	0.41
	ResMS	0.506	0.577	0.591
II	R	0.60	0.57	0.50
	ResMS	0.602	0.599	0.654
III	R	0.50	0.46	0.46
	ResMS	0.406	0.414	0.409
IV	R	0.62	0.58	0.56
	ResMS	0.491	0.507	0.518
V	R	0.67	0.53	0.53
	ResMS	0.312	0.393	0.386
VI	R	0.69	0.46	0.44
	ResMS	0.254	0.357	0.355
VII	R	0.57	0.40	0.37
	ResMS	0.508	0.570	0.558
VIII	R	0.80	0.79	0.79
	ResMS	0.465	0.444	0.423
IX	R	0.66	0.66	0.66
	ResMS	0.229	0.225	0.223

TABLE 16

EQUATIONS FOR ESTIMATING ANNUAL OPERATION

AND MAINTENANCE COST PER CAPITA

Region	Regression Coefficients b_o	b_1	b_5	C_{ij} C_{11}	C_{15}	C_{55}	df	n	ResMS	Deviations x_1	x_5
I	+0.939	-0.213	+0.514	0.011	-0.013	0.253	36	39	0.591	$\ln X_1$ - 2.906	$\ln X_5$ - 0.156
II	+1.459	-0.263	+0.534	0.010	-0.006	0.139	37	40	0.654	$\ln X_1$ - 3.106	$\ln X_5$ - 0.440
III	+0.397	-0.127	+0.631	0.005	-0.003	0.091	62	65	0.409	$\ln X_1$ - 2.350	$\ln X_5$ - 0.466
IV	+1.080	-0.277	+0.517	0.008	-0.001	0.104	52	55	0.518	$\ln X_1$ - 2.716	$\ln X_5$ - 0.486
V	+1.153	-0.265	+0.277	0.009	-0.007	0.102	56	59	0.386	$\ln X_1$ - 2.399	$\ln X_5$ - 0.561
VI	+0.279	-0.066	+0.667	0.008	+0.002	0.224	31	34	0.355	$\ln X_1$ - 2.490	$\ln X_5$ - 0.616
VII	+0.766	-0.194	+0.063	0.204	-0.012	0.307	21	24	0.558	$\ln X_1$ - 3.143	$\ln X_5$ - 0.667
VIII	+1.867	-0.534	+0.081	0.023	+0.033	0.497	21	24	0.423	$\ln X_1$ - 2.568	$\ln X_5$ - 0.654
IX	+1.158	-0.288	+0.270	0.007	+0.001	0.072	78	81	0.223	$\ln X_1$ - 3.099	$\ln X_5$ - 0.524

Regression equation:

$$\ln Y_6 = b_o + b_1 \ln X_1 + b_5 \ln X_5$$

Standard error of an estimated expected value:

$$s_{\ln Y6} = (\text{ResMS } (1/n + C_{11}x_1^2 + C_{55}x_5^2 + 2C_{15}x_1x_5))^{\frac{1}{2}}$$

In the per PE treated per year studies all three variables, X_6, X_4, X_5, contributed a significant amount to the regression sum of squares for Regions IV and VII. For Regions I, II, and IX, variables X_6 and X_4 contributed significantly to the regression sum of squares. For Region V, variables X_6 and X_5 contributed significantly to the regression sum of squares. The variable X_6 was the only one that contributed significantly to the sum of squares for Regions III, VI, and VIII.

In the cost per PE produced studies all three variables, X_6, X_4, and X_5, contributed a significant amount to the regression sum of squares for Region VII. Variables X_6 and X_4 contributed a significant amount to the regression sum of squares for Regions IV, V, and VI. Variables X_6 and X_5 contributed a significant amount to the regression sum of squares for Regions II and IX. For Regions I, III and VIII, only variable X_6 contributed significantly to the regression sum of squares. The tests for significance of independent variables are summarized in Table 17.

Using one independent variable, X_1 or X_6, seemed just as helpful as referring to three independent variables. It is possible that either X_4 or X_5 or both do contribute a significant amount to the regression, but this was not detected since the sample size was too small for each region. There is also the possibility that they do not contribute a significant amount to regression and that there are other independent variables which do contribute significantly. This will be discussed further in the next chapter.

There were no equations derived which utilized only one independent variable, since the use of the per capita equations with two

TABLE 17

SIGNIFICANT VARIABLES FOR DERIVED ESTIMATION EQUATIONS FOR OPERATION AND MAINTENANCE COSTS

Region	Regression Equation for cost per: Capita — Significant Variables	Regression Equation for cost per: PE Treated per year — Significant Variables	Regression Equation for cost per: PE Produced per year — Significant Variables
I	X_1, X_5	X_6, X_4	X_6
II	X_1, X_4, X_5	X_6, X_4	X_6, X_5
III	X_1, X_5	X_6	X_6
IV	X_1	X_6, X_4, X_5	X_6, X_4
V	X_1	X_6, X_5	X_6, X_4
VI	X_1	X_6	X_6, X_4
VII	X_1	X_6, X_4, X_5	X_6, X_4, X_5
VIII	X_1	X_6	X_6
IX	X_1, X_5	X_6, X_4	X_6, X_5

variables, X_1 and X_5 gave quite reasonable results. In the per capita study involving the two variables X_1 and X_5, X_5 contributed a significant amount to the regression for Regions III, IV, VI, and IX. The variable X_1, contributed significantly for all regions. If, in fact, X_5 does not contribute a significant amount to regression using it in the regression equations will not decrease the precision of the estimate. However, a degree of freedom will be lost in the calculation of the confidence limits.

Population Equivalency Versus Population Study

It was believed that from a knowledge, or estimate, of design population one could determine the population equivalency produced within narrow limits. Regression studies were conducted for each region. The form,

$$\ln Y = a + b \ln X,$$

where Y is the population equivalency produced, and X is the design population, was selected because its use resulted in the highest correlation coefficient.

After the regression equations were derived for the nine regions the hypothesis of homogeneity of regression coefficients was tested. The result of this test indicated that there was no evidence to reject the hypothesis of homogeneity. It was assumed that the regression lines for the nine regions were parallel. The hypothesis that the lines were coincident was tested. The results indicated that there was no evidence to reject this hypothesis. Thus, it was assumed that one equation would apply for all nine regions.

The regression equation and necessary data for computing 95% confidence limits of an estimated expected value are given in Table 18. Use of the regression equation will allow a planner, who has no information except design population, to compute an estimate of the population equivalency for use in the estimation of construction costs. There is some danger in using an estimated value in the regression equations for construction cost because it is assumed that the independent variables are measured without error. However, it must be remembered that design population is also an estimate, yet it is used as the basis of design for a treatment plant. Thus, it is not a variable in the mathematical sense because it is used as an exactly known value. This will also be true of population equivalency. Once a planner arrives at a value he feels is reliable he will use it as an exactly known value. The above regression equation provides a rational method for selecting a reliable value for population equivalency.

Table 19 gives the results of the application of the estimation equation for selected values of the design population. The width of interval is quite narrow at low values of design population, and although it is wider at higher values of design population it is still within reason. For example, the width of interval for a design population of 10,000 is 1200 PE and for 100,000 population is 18,000 PE.

TABLE 18

EQUATION FOR ESTIMATING POPULATION EQUIVALENCY

Regression equation:

$$\ln Y = 0.052 + 1.022 \ln X$$

where Y = Population equivalency x 0.001,

X = Design population x 0.001,

and ln is the base e logarithm.

Standard error of an estimated expected value for which $X = X_o$:

$$s_{\ln Y} = \left(0.332 \left(1/443 + \frac{(\ln X_o - 2.734)^2}{1144.311}\right)\right)^{\frac{1}{2}}$$

95% confidence limits: $\ln Y \pm t\, s_{\ln Y}$

where t is student's t value: $t = 1.97$

and the degrees of freedom are 441.

TABLE 19

ESTIMATES OF POPULATION EQUIVALENCY

FOR SELECTED DESIGN POPULATION

Design Population x.001	Expected Population Equivalency x.001	95% Confidence Limits x 0.001	
		Lower Limit	Upper Limit
1.00	1.05	0.95	1.17
5.00	5.46	5.12	5.83
10.00	11.10	10.50	11.70
25.00	28.30	26.70	29.90
50.00	57.40	53.70	61.40
75.00	86.90	69.50	80.90
100.00	110.70	102.00	120.00

CHAPTER VII

SUMMARY AND DISCUSSION

As an aid to water resources planning, equations were derived for estimating construction and operation and maintenance cost of municipal sewage treatment facilities. The use of the equations is not meant to replace detailed engineering studies for any given project. The equations are meant to provide an answer to the question: Of all possible values for cost, what is the most likely construction and operation and maintenance cost for a given set of conditions in a given region?

The collection of the data and derivation of the equations were based on modern statistical techniques. It is believed that the data, in the form used, met all of the assumptions required for use of these statistical techniques. Whenever possible these assumptions were tested using accepted statistical methods.

The stratification of the United States into nine regions, and the subsequent derivation of equations for each region and statistical tests on these equations provided evidence to support the hypothesis that there is a wide regional variation in cost. The use of the multiple linear regression technique made it possible to derive equations such that the confidence intervals on estimates of construction costs are rather narrow for each region.

For estimation of construction cost, regression equations with three independent variables gave relatively precise figures when compared to those with five variables. In a few of the regions it was found that the use of two independent variables was just as good as the use of three, as far as the contribution to regression sum of squares was concerned. It is possible that the sample was too small to indicate that all three independent variables contributed significantly to the regression. The sample size within each region was not determined by a statistical method because no estimate of the variance within the regions was available.

Relative to operation and maintenance cost, only in very few regions did the use of all three independent variables, PE or population, BOD of effluent and type of treatment plant, contribute significantly to the regression. In a few regions only two of these variables contributed significantly to the regression, and in a majority of regions only one of the variables, population, contributed significantly to the regression.

It was thought that the type of treatment plant variable would contribute a significant amount to the regression for all regions, since one would expect that, for example, an activated sludge plant in a given region designed for a given population equivalency would have a higher annual total operation and maintenance cost than a standard rate trickling filter plant, in the same region, designed for the same population equivalency. It is possible that the sample size was not large enough to indicate this expected significance. It is also possible that there

are other variables, not used in the study, which would contribute significantly to the regression and that the use of the type of treatment plant variable is not really significant in the statistical sense. The number of plant employees, number of work shifts, power costs, and use of sludge gas within the plant for power and heat are examples of variables which are very important in determining operation and maintenance costs. They were not used since information concerning them is very scarce. It was hoped that the BOD of effluent and the type of plant variable would be adequate substitutes.

It could be argued that for a given loading condition, on a given treatment plant, the BOD of the effluent is fixed. However, the operation of the plant determines to a great extent, within quite wide limits, the BOD of the effluent. That is, if the operators are inexperienced, or if there is a great turnover in operators, or if there are not enough operators a plant cannot be expected to operate as efficiently as a similar plant with an adequate number of highly trained and experienced operators. Thus, BOD of the effluent ought to be an important variable in operation and maintenance costs. It is possible that most of the plants were operating close to their maximum efficiency and if this were the case then the BOD of the effluent would not contribute much to the regression. An inspection of the data revealed that approximately 80 per cent of the plants in the study were operating at or very close to their expected efficiency. This is evidence that the BOD of the effluent would not contribute much to the regression. This was quite unexpected in view of past experience.

The type of data that is required for more reliable operation

and maintenance cost estimation can only be obtained from an interview survey. The survey might have to last a few days at each chosen municipality.

A few examples are cited to illustrate application of the equations. Assume that a planner desires to estimate the construction cost per PE produced for a municipality in Region I. The design PE of the influent is 4,000 and the stream conditions are such that the BOD of the effluent must be about 330 ppm. A primary type treatment plant would suffice for this BOD requirement and PE of influent. The equation for Region I given in Table 11 is appropriate for estimating the expected cost:

$$\ln Y_1 = 3.926 - 0.390 \ln 4. - 0.202 \ln 330. - 0.751 \ln 1.00 = 2.216$$

the antilog of 2.216 (an estimate of the expected cost per PE produced in 1913 dollars) = \$9.17 for the 95% confidence interval the following are computed:

$$x_4 = (\ln 330. - 3.938) = 1.661$$

$$x_5 = (\ln 1.00 - 0.183) = -0.183$$

$$x_6 = (\ln 4. - 2.899) = -1.513$$

$$s_{\ln Y_1} = (0.442\,(1/38 + 0.010\,(-1.513)^2 + 0.031\,(1.661)^2 + 0.373\,(-0.183)^2$$
$$+2\,(0.067)(1.661)(-0.183) + 2\,(-0.004)(1.661)(-1.513)$$
$$+2\,(-0.019)(-0.183)(-1.513))^{\frac{1}{2}} = 0.234$$

$$t\, s_{\ln Y_1} = (2.03)(0.234) = 0.475$$

95% confidence limits $= \ln Y_1 \pm t\, s_{\ln Y_1}$

$$= 2.216 \pm 0.475$$

$$= 1.741 \text{ to } 2.691$$

antilog of these (95% confidence interval in 1913 dollars per PE produced) = \$5.70 to 14.70. The probability is 0.95 that the true mean construction cost per PE produced in 1913 dollars for a primary treatment plant in Region I with a PE of 4,000 and BOD of effluent of 330 lies in the interval \$5.70 to \$14.70 and that the best estimate of this expected value is \$9.17. To obtain the total cost for say 1962 (ENR index = 871.84 for 1962):

$$9.17(4000)\ \frac{871.84}{100} = \$320{,}000 \text{ (approximate).}$$

An estimate of the annual operation and maintenance cost for this plant is given by:

$$\ln Y_4 = 1.716 - 0.320 \ln 4. + 0.135 \ln 330 + 0.622 \ln 1.00 = 2.055$$

antilog of this = 7.81

decoding gives the estimated cost per PE produced per year is \$0.00781

for 95% confidence limits compute:

$$x_4 = (\ln 330 - 4.060) = 1.739$$

$$x_5 = (\ln 1.00 - 0.156) = -0.156$$

$$x_6 = (\ln 4. - 2.959) = -1.573$$

$$s_{\ln Y_4} = (0.405\ (1/39 + 0.009\ (-1573)^2 + 0.034\ (1.739)^2 + 0.375\ (-0.156)^2$$
$$+2(-0.004)(-1.573)(1.739) +2\ (-0.018)(-0.156)(-1.573)$$
$$+2(0.064)(-0.156)(1.739))^{\frac{1}{2}} = 0.237$$

$$t\ s_{\ln Y_4} = (2.03)(.237) = 0.480$$

$$\ln Y_4 \pm t\ s_{\ln Y_4} = 2.055 \pm 0.480 = 1.575 \text{ to } 2.535$$

antilog of these = 4.83 to 12.60

decoding gives cost per PE produced per year = \$0.00483 to \$0.0126.

To obtain total annual cost:

expected value = (4000)(365)(.00781) = $11,000 (approximate)

Lower Limit of 95% confidence interval = (4000)(365)(.00483) = $7,000 (approximate)

Upper Limit of 95% confidence interval = (4000)(365)(.0126) = $18,000 (approximate).

Table 20 presents the results of a few applications of the estimation equations for construction cost per PE produced in 1913 dollars.

TABLE 20

ESTIMATED CONSTRUCTION COSTS PER PE

PRODUCED FOR SELECTED CONDITIONS

Region	PE Influent	BOD of Effluent	Type of Plant	Estimate of Mean Cost	95% Confidence Limits Lower	Upper
I	4,000	330	1.00	$9.17	$5.70	$14.70
II	60,000	60	1.62	4.78	4.47	5.10
III	10,000	30	1.62	6.02	5.08	7.14
III	100,000	20	2.61	2.42	1.72	3.38
IX	30,000	30	2.36	3.47	2.88	4.17

Table 21 gives the results of the use of the construction cost per capita equations for selected conditions. The cost is in 1913 dollars per capita.

The confidence intervals are narrow except for a few cases. It was expected that these would be wide for they are at the lower limits of the valid population or population equivalency. The valid range for

TABLE 21

ESTIMATED CONSTRUCTION COSTS PER CAPITA FOR SELECTED CONDITIONS

Region	Design Population	BOD of Effluent	Type of Plant	Estimate of Mean Cost	95% Confidence Limits Lower	Upper
II	1,000	30	1.62	$16.50	$11.70	$23.30
V	1,000	20	2.61	15.10	9.88	23.10
V	10,000	30	2.61	8.47	6.67	10.80
V	100,000	30	2.61	4.94	3.61	6.68

population and population equivalency is from about 500 to 1,300,000. That is, there were only few municipalities in the study with a PE or design population of lower than 500 or greater than 1,300,000. An inspection of the equation for computing the standard error of estimate reveals that the greater the deviation between a design population and the mean design population used in the study (measured in logarithmic units) the greater will be the value of the standard error, hence the wider the confidence interval. The range in BOD of effluent values used in the study was 3 to 380.

There is some risk in extrapolating beyond the range of values for the independent variables observed in the study. It is possible that the same regression function will not apply to values outside the range. If this is the case the estimates will be either much too large or much too small.

It is believed that use of the regression equations will give water resources planners reliable estimates of the cost of sewage treatment. Use of the estimates in present, or new, mathematical models for water basin operation will greatly aid in deciding the proper balance between dilution of wastes and degree of treatment of wastes as well as in aiding in the decision of overall operation. That is, in any given basin it may be possible that provisions for storage of dilution water will be more economical than requiring, say 90 per cent BOD removals by all of the communities in a given portion of the basin. On the other hand, just the opposite may prove to be true. In either event this study does provide information, in the form of cost estimation, to enable the planner to make a more rational decision concerning operation of a water basin.

BIBLIOGRAPHY

1. Bureau of the Cenus. Current Population Reports. Series P-25, No. 187.

2. Bureau of the Census. "Population Projections and Economic Assumptions", Committee Print No. 5, Water Resources Activities in the United States, United States Senate Select Committee on National Water Resources, March, 1960.

3. Public Health Service. "Future Water Requirements for Municipal Use", Committee Print No. 7, Water Resources Activities in the United States, United States Senate Select Committee on National Water Resources, January, 1960.

4. Reid, George W., "Water Requirements for Pollution Abatement", Committee Print No. 29, Water Resources Activities in the United States, United States Senate Select Committee on National Water Resources, July, 1960.

5. Public Health Service. "Pollution Abatement", Committee Print No. 9, Water Resources Activities in the United States, United States Senate Select Committee on National Water Resources, January, 1960.

6. Wollman, Nathaniel, "Water Supply and Demand", Committee Print No. 32, Water Resources Activities in the United States, United States Senate Select Committee on National Water Resources, August, 1960.

7. Public Health Service. "Water Quality Management", Committee Print No. 24, Water Resources Activities in the United States, United States Senate Select Committee on National Water Resources, February, 1960.

8. Velz, C. J., "How Much Should Sewage Treatment Cost?", Engineering News-Record, October 14, 1948, pp. 84-86.

9. Diachishin, A. N., "New Guide to Sewage Plant Costs", Engineering News-Record, October 17, 1957, pp. 316-318.

10. Thoman, J. R. and Jenkins, K. H., "How to Estimate Sewage Plants Quickly", Engineering News-Record, December 25, 1958, pp. 64-65.

11. Rowan, P. P., Jenkins, K. H. and Butler, D. W., "Sewage Treatment Construction Costs", Journal Water Pollution Control Federation, June, 1960, pp. 594-604.

12. Rowan, P. P., Jenkins, K. L. and Howells, O. H., "Estimating Sewage Treatment Plant Operation and Maintenance Costs", Journal Water Pollution Control Federation, February, 1961, pp. 111-121.

13. Logan, J. A., Hatfield, W. D., Russell, G. S. and Lynn, W. R., "An Analysis of the Economics of Waste-Water Treatment", Journal Water Pollution Control Federation, September, 1962, pp. 860-882.

14. Wollman, N., "Cost of Treatment", unpublished paper, University of New Mexico, New Mexico, 1963.

15. Williams, E. J., Regression Analysis. John Wiley and Sons, Inc., 1959, page 24.

16. Williams, E. J., Regression Analysis. John Wiley and Sons, Inc., 1959, page 23.

17. Reid, G. W., "Biological Treatment of Domestic Wastes", Public Works, July, 1955, pp. 78-89.

18. Graybill, F. A., An Introduction to Linear Statistical Models, Volume I. McGraw-Hill Book Co., Inc., 1961, page 97.

19. Graybill, F. A., An Introduction to Linear Statistical Models, Volume I. McGraw-Hill Book Co., Inc., 1961, Chapter 6.

20. Graybill, F. A., An Introduction to Linear Statistical Models, Volume I. McGraw-Hill Book Co., Inc., 1961 Chapter 7.

21. Steel, R. G. D. and Torrie, J. H., Principles and Procedures of Statistics. McGraw-Hill Book Co., Inc., 1960, Chapter 14.

22. Ezekial, M. and Fox, F. A., Methods of Correlation and Regression Analysis, John Wiley and Sons, Inc., 1959, Chapter 12.

23. Ezekial, M. and Fox, F. A., Methods of Correlation and Regression Analysis, John Wiley and Sons, Inc., 1959, Appendix 3, Note 2.

24. Steel, R. G. D. and Torrie, J. H., Principles and Procedures of Statistics. McGraw-Hill Book Co, Inc., 1960, Chapter 17.

25. Federer, W. T., Experimental Design, The Macmillan Company, 1955, Chapter 2.

26. Williams, E. J., Regression Analysis, John Wiley and Sons, Inc., 1959, Chapter 8.

27. Hansen, M. H., Hurwitz, W. N. and Madow, W. G., Sample Survey Methods and Theory. John Wiley and Sons, Inc., 1953, Vol. I, Chapter 1, Vol. II, Chapter 1.

28. Public Health Service. Inventory of Municipal and Industrial Waste Facilities in the United States - 1957. Publication No. 622, Government Printing Office, 1958.

29. Hansen, M. H., Hurwitz, M. N., and Madow, W. G., Sample Survey Methods and Theory. John Wiley and Sons, Inc., 1953, Vol. I, Chapter 5, Vol. II, Chapter 5.

30. Steel, R. G. D. and Torrie, J. H., Principles and Procedures of Statistics. McGraw-Hill Book Company, Inc., 1960, Chapter 5.

31. Ostle, B. Statistics in Research. The Iowa State University Press, 1963, Appendix 5.

32. Steel, R. G. D. and Torrie, J. H. Principles and Procedures of Statistics. McGraw-Hill Book Company, Inc., 1960, Chapter 7.

33. Steel, R. G. D. and Torrie, J. H. Principles and Procedures of Statistics. McGraw-Hill Book Company, Inc., 1960, Chapter 10.

34. Steel, R. G. D. and Torrie, J. H. Principles and Procedures of Statistics. McGraw-Hill Book Company, Inc., 1960, Chapter 14.

35. Keeping, E. S. Unpublished class notes. PM&PH 359, Advanced Principles of Statistics, 1963.