AAEC/E486

AAEC/E486

(2)

AUSTRALIAN ATOMIC ENERGY COMMISSION

RESEARCH ESTABLISHMENT

LUCAS HEIGHTS

FACTORS AFFECTING LUNG CAPACITY IN MALE EMPLOYEES AT THE AAEC RESEARCH ESTABLISHMENT

by

PAMELA MACE*

* Vacation student

April 1980

ISBN 0 642 59685 9

AUSTRALIAN ATOMIC ENERGY COMMISSION
RESEARCH ESTABLISHMENT
LUCAS HEIGHTS

FACTORS AFFECTING LUNG CAPACITY IN MALE EMPLOYEES AT THE AAEC RESEARCH ESTABLISHMENT

by

PAMELA MACE*

*Now completing postgraduate studies at the Institute of Animal Resource Ecology, University of British Columbia, Canada. This work was undertaken in 1976 while the author was employed at the AAEC Research Establishment as a vacation student.

ABSTRACT

In 1976 a detailed analysis was made of the influence of age, height, weight, smoking habits, occupation and birthplace of male employees at the AAEC Research Establishment, Lucas Heights, on two measures of human lung capacity - forced expiratory volume after one second (FEV_1) and forced vital capacity (FVC). The analysis was performed on medical data accumulated over ten years. The most important determinants of lung capacity are age and height; age explains more of the variation in FEV_1 and height explains more of the variation in FVC. Regression equations have been based on these variables. The analyses also supported the hypothesis that smoking habits and occupation have comparatively small, but nevertheless distinct, effects on lung capacity. Results indicate that smoking has the greater effect on FEV_1 and occupation has the greater effect on FVC, but it is difficult to quantify

(Continued)

these effects and incorporate them into the prediction equations. Making allowance for birthplace and weight did not increase the reliability of the predictions, the latter because of its high correlation with height.

National Library of Australia card number and ISBN 0 642 59685 9

The following descriptors have been selected from the INIS Thesaurus to describe the subject content of this report for information retrieval purposes. For further details please refer to IAEA-INIS-12 (INIS: Manual for Indexing) and IAEA-INIS-13 (INIS: Thesaurus) published in Vienna by the International Atomic Energy Agency.

LUNGS; VOLUME; PERSONNEL; AAEC

CONTENTS

1. INTRODUCTION

Many studies have been carried out on the vital capacity of human lungs but most have been concerned only with elucidating the important factors that determine the capacity, by using regression techniques. Invariably, the only factors found to explain a significant amount of the variance in lung capacity are the height and age of the subject and, occasionally, the country of origin. Intuitively, it would seem likely that other variables such as working conditions and smoking habits could have an effect if only by accelerating the deterioration of vital capacity with age. Such variables are rarely found to be important when regression methods are employed, owing to the high variability in vital capacity between subjects. As this high variability also results in wide confidence limits for the regression equation, it is virtually impossible to determine whether or not, for example, a person's respiratory system is affected adversely by working environment. This is obviously of importance to an employer who is not concerned with a natural age-related deterioration in lung capacity but is, or should be, concerned about the possible effects of dust, chemicals, smoke, radiation, etc., in the working area. Thus, height and age may not be the most important variables from some standpoints, even though they are the major determinants of vital capacity.

The main problem in determining the effects of the other variables is that extremely large sample sizes are required to detect statistically significant differences between samples of subjects of the same height and age over a wide range of heights and ages. Nevertheless, subtle effects of minor variables can often be elucidated by a thorough analysis of the available data. Irrefutable evidence for the effects of the minor variables may not emerge, but the evidence may be sufficient to warrant further experimentation and analysis.

In the present study, a detailed analysis of the relative importance of the factors contributing to vital capacity and the form of the distribution of the data was made. The main objectives were to obtain as much information as possible from the data by use of standard statistical techniques and then to examine the possibility that variables other than age and height are important in the deterioration of vital capacity with age.

2. METHODS

Data were collected for a period of ten years from annual medical examinations of male employees at the AAEC Research Establishment, Lucas Heights, and analysed in 1976. Two measurements were made for each subject - forced expiratory volume after one second (FEV_1) and forced vital capacity (FVC). Expired gas volumes were measured against the 'air temperature and pressure' scale of a recording Vitalograph spirometer. The room is kept at approximately 18-22°C and is 150 metres above sea level. The age, weight, and height of the subject, the type of work he does, his country of birth and his smoking habits were also recorded. A further variable, body surface area (BSA) was calculated from the formula [Guyton 1971]:

$$BSA = (0.007184)(\text{weight})^{0.425}(\text{height})^{0.725}$$

Because of the small sample size, females were excluded from the analysis, as were incomplete data sets. Sample sizes, means and ranges for each of the variables were determined from the remaining data. Forced vital capacity measured over the period 1966-1975 yielded 7452 data points and FEV_1, recorded only during the period 1972-1975 (except for one measurement in 1971), provided 2935 data points.

The ranges of the continuous variables and the categories used for the nominal variables are listed in Table 1. The continuous variables were divided into classes for some of the analyses. The number of classes employed were: age (9 or 5), height (8 or 4), weight (6 or 4), surface area (6 or 4), FEV_1 (6 or 4) and FVC (6 or 4). The classes were all of equal size except the first and last classes which were sometimes slightly different.

Data were analysed in several stages, each of which resulted in hypotheses that could subsequently be tested. These stages were:

- Basic statistics and tests for normality.

- Analysis of variance and *a posteriori* tests.

- Correlation analysis.

TABLE 1

RANGES OF CONTINUOUS VARIABLES AND CATEGORIES OF NOMINAL VARIABLES

Age:	range = 15-66 years
Height:	range = 152-198 cm
Weight:	range = 44-135 kg
Surface area:	range = 1.40-2.53 m^2
FEV_1:	range = 0.6-6.5 l; mean = 3.84 l
FVC:	range = 1.5-7.5 l; mean = 4.42 l
Occupation categories:	PR - professional and administrative
	TC - technicians
	TR - tradesmen
Birthplace categories:	AUS - Australia
	A - Asia
	E - Europe
	NZ - New Zealand
	UK - United Kingdom
	USA - United States of America
Smoking categories:	N - non-smokers
	X - ex-smokers
	P - predominantly pipe smokers
	L - light smokers (<15 cigarettes per day)
	M - medium smokers (15-25 cigarettes per day)
	H - heavy smokers (>25 cigarettes per day)

- Regression analysis.

- Median tests.

Most of these tests required that the data comprise a set of independent observations; thus all of the data from the entire period of ten years could not be analysed together. Another complicating factor was that in 1972 a new type of machine was used to record vital capacity. The first machine, a simple water-filled spirometer which was not able to record FEV_1, was replaced by the more versatile Vitalograph spirometer. The calibration of these machines varied by about 0.5 litres.

As the 1973 population had the greatest number of data points (923 for both FEV_1 and FVC), it was selected for the more detailed analysis.

3. RESULTS

3.1 Basic Statistics and Normality Tests

The distribution characteristics of the data and selected sub-samples were examined to determine the validity of subsequent tests (e.g. tests based on the assumption of a normal distribution). The means, medians, variances, standard deviations, coefficient of variation, skewness and kurtosis were calculated.

In a normal frequency distribution, both skewness (g_1) and kurtosis (g_2) are zero. A negative g_1 indicates skewness to the left (left tail drawn out), a positive g_1 indicates skewness to the right. A negative g_2 indicates platykurtosis, whereas a positive g_2 shows leptokurtosis. These statistical parameters were tested for significance by the procedure outlined by Sokal and Rohlf [1969].

Two other normality tests were made. The first was the Kolmogorov-Smirnov test which compares the sample statistic D_{max}, the maximum distance between the observed population curve and a normal frequency curve with the same mean and variance, with the test statistic D_α. If

$$D_{max} > D_{\alpha} = \sqrt{\frac{-\ln(\frac{1}{2}\alpha)}{2n}}$$

where α is the level of significance, the hypothesis of normality is rejected. The second was the Lilliefors test. This test is basically the same as the first except that the sample statistic, D_{max}, is computed from the standardised sample data instead of the raw values. The test statistic, D_{α}, is the same as that defined above. (Refer to Conover [1971] for more information of these normality tests.)

The results from the tests for skewness, kurtosis and the two normality tests are summarised in Table 2. The first row in the table shows that the FEV_1 values for all data taken over the entire 10 year period are highly abnormal. This is to be expected because of the dependence of each subject's FEV_1 on his previous measurements. However, the other subsets of the data form groups of independent observations; it is evident for these groups, on the basis of the four methods of determining normality, that the hypothesis that the distribution of the data is normal at the $\alpha = 0.05$ level of confidence is usually accepted. The hypothesis is rejected less often for FVC than it is for FEV_1.

In the cases where the hypothesis is rejected, as shown in Table 2 (rejection occurs when the sample statistic is greater than the test statistic), the sample statistic is usually only slightly greater than the test statistic. Hence, the normal distribution does not appear to be an unreasonable approximation to the true unknown distribution for all subsamples of the data except when all the data are considered together. Further evidence that the assumption of normality is reasonable for these subsamples is that in all cases the means and medians differ by a maximum of 0.05 litres.

Hence, it was considered that in subsequent analyses, statistical tests which assume a normal distribution could be used, although it was realised that significance levels might not always be exact and equivalent non-parametric tests were used where possible.

Variances and standard deviations have not been recorded but it should be noted that all subsamples of the data had extremely high variability. Values for the coefficient of variation are listed in Table 3 for the last five subsamples of Table 2. This number is useful for comparison of the variation

TABLE 2

RESULTS OF NORMALITY TESTS FOR SELECTED SUBSAMPLES OF THE DATA

A = accept hypothesis of normality; R = reject hypothesis of normality (at the 5 per cent level). The number above the symbol R is the sample statistic, the number below is the test statistic.

SUBSAMPLE		Skewness	Kurtosis	Kolmogorov-Smirnov Test	Lilliefors Test
All data	FEV_1	3.24 R 1.96	6.67 R 1.96	0.047 R 0.025	
	FVC				
All data for 1973	FEV_1	2.01 R 1.96	4.49 R 1.96	0.046 R 0.045	0.046 R 0.029
	FVC	A	A	A	0.036 R 0.029
Non-smokers for 1973 (all nationalities)	FEV_1	A	A	A	0.044 R 0.041
	FVC	A	A	A	0.049 R 0.041
Non-smoking Australians for 1973	FEV_1	A	A	A	0.051 R 0.048
	FVC	A	A	A	A
Heavy smokers for 1974	FEV_1	A	A	A	
	FVC	A	A	A	
All Australians for 1974	FEV_1	A	2.06 R 1.96	A	
	FVC	A	A	A	

of populations, independent of the magnitude of their means. It can also be used to determine whether a given biological sample is more variable for one character than for another.

Two main points are evident from Table 3: (i) FVC values are less variable between subjects than FEV_1 values; and (ii) the most variable group of those tested is the heavy smokers for 1974 - but this may be partly a reflection of the small sample size for this group.

3.2 Analysis of Variance and a posteriori Tests

The analyses in this and subsequent sections were performed on the FEV_1 and FVC values recorded for 923 subjects in 1973. Within this group, there were 478 non-smokers, 102 ex-smokers, 46 pipe smokers, 113 light smokers, 158 medium smokers and 26 heavy smokers. Three hundred and sixteen people were classified as professional workers, 245 as technicians and 362 as tradesmen.

Single classification analysis of variance (ANOVA) was used to determine whether there were any differences between mean FEV_1 and mean FVC for each category or class of the variables: age, birthplace, height, smoking habits, occupation and weight. Single classification ANOVA is a statistical technique that assesses the effect of a categorically independent variable (factor), measured at any level, upon a continuously dependent variable (FEV_1 or FVC in this case). It is used to determine whether or not the variances of the means of the dependent variable among categories are greater than expected on the basis of variances of the dependent variable within categories.

The validity of ANOVA as a method of separating the total variation in a set of observations into components from different sources does not depend upon any assumption of normality. It requires only that the observations are independent and arise from the usual type of additive model. "Normality of the distribution of the random errors is required only for strict validity of the usual tests of significance and of calculations of fiducial limits to estimates..." [Finney 1963].

ANOVA was carried out on the entire 1973 sample, excluding subjects found to be more than 20 per cent heavier than their predicted weights, all non-smokers, and non-smokers excluding those who were overweight. The formula used to determine the 'correct' weight for a given age and height was

TABLE 3

THE COEFFICIENT OF VARIATION FOR THE

LAST FIVE SUBSAMPLES IN TABLE 2

SUBSAMPLE	FEV_1	FVC
All data for 1973	21.48	17.98
Non-smokers for 1973 (all nationalities)	17.81	15.84
Non-smokers Australians for 1973	18.11	15.83
Heavy smokers for 1974 (all nationalities)	30.83	22.44
All Australians for 1974	20.53	17.77

$$y = a \cdot \exp(bx)$$

where y = weight in kg, x = age in years, a = -37.83 + 0.58 height, b = 0.0048 -0.0000049 height. This formula was derived from an analysis of weights,ages and heights adopted as standards for acceptance of superannuation candidates in the Commonwealth service [Commonwealth Medical Officers' Handbook, 1965]. The predictions were found to be reasonable in the range 160 to 180 cm (5' 3" to 6' 1"). Overweight subjects were excluded to see if there was a resultant increase in the significance of weight as a factor contributing to lung capacity.

Results of these analyses are summarised in Table 4, together with η^2 values [Nie et al. 1975] for each of the independent variables. The η^2 value is a measure of the proportion of the total (linear and non-linear) variance explained by the independent variable. It should not be taken literally as its computation depends on the normal distribution; however the values obtained suggest that, at least in the case of non-smokers, age has the greater effect on FEV_1 and height has the greater effect on FVC.

Table 4 also shows that the F values for age and height were highly significant (α = 0.001), whereas birthplace was not found to be a significant factor for any of the subsamples, even at the α = 0.05 level. Occupation was a highly significant factor when all of the data were considered together but, when the smokers were removed from the analysis, the F-values, although still significant at α = 0.001, decreased considerably. This suggests that there is some association between occupation and smoking habits. The F-values for weight were often significant but as those for height were so much higher, the significance of weight as a factor may be partly a result of the positive correlation between weight and height and that between height and FEV_1 or FVC. Exclusion of overweight subjects increased the significance of weight as a factor in determining vital capacity.

Although sample sizes for countries other than Australia and the United Kingdom were small, birthplace did not appear as a significant factor of vital capacity.

The main result from these single classification ANOVAs is an indication that many factors may have an effect on vital capacity and that some may have their greatest effect on FEV_1 whereas others may have more of an effect on

TABLE 4

F-VALUES FROM SINGLE CLASSIFICATION ANALYSIS OF VARIANCE OF FEV_1 AND FVC FOR FOUR SUBSAMPLES

NS = result not significant; * = significance at 5 per cent level; ** = significance at 1 per cent level; *** = significance at 0.1 per cent level.

	Entire Sample		Entire Sample excluding Overweight Subjects		Non-smokers		Non-smokers excluding Overweight Subjects	
	FEV_1	FVC	FEV_1	FVC	FEV_1	FVC	FEV_1	FVC
Age	74.69***	52.58***	75.95***	52.54***	27.28***	17.10***	26.57***	17.31***
Eta-square	0.40	0.31	0.41	0.33	0.32	0.23	0.32	0.24
Height	36.17***	59.38***	35.42***	58.71***	26.42***	38.45***	29.28***	41.53***
Eta-square	0.22	0.31	0.22	0.32	0.28	0.36	0.32	0.40
Weight	3.62**	7.26***	5.16**	12.09***	1.96NS	3.93**	2.23NS	5.98***
Eta-square	0.02	0.04	0.02	0.04	0.02	0.04	0.01	0.04
Birthplace	1.00NS	0.61NS	0.65NS	0.43NS	0.79NS	0.60NS	0.71NS	0.58NS
Eta-square	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.01
Smoking habits	19.79***	10.63***	19.46***	10.41***				
Eta-square	0.10	0.05	0.10	0.06				
Occupation	27.89***	29.82***	27.73***	29.84***	4.96**	7.11***	2.24NS	5.98***
Eta-square	0.06	0.06	0.06	0.06	0.02	0.03	0.01	0.04

FVC. In particular, age seems to be a more important determinant of FEV_1 than height, and height seems to be the most important factor determining FVC. However, it should be noted that in a multivariate system such as the present one, the simple design of a single classification ANOVA is not sufficient to represent the total complexity of the system. If the null hypothesis of equal mean capacities for all categories of each independent variable (factor) is rejected, it may not mean that this factor has a significant effect on lung capacity, but rather that this factor is highly correlated with some other factor which itself has a direct effect on lung capacity. Therefore the results of these analyses should be interpreted with caution.

To verify the results of the single classification ANOVAs, insofar as they can be accepted, an equivalent non-parametric test, the Kruskal-Wallis test [Sokal and Rohlf 1969], was also carried out. Such non-parametric tests usually operate for a wide range of distributions. The null hypothesis is not concerned with specific parameters (such as the mean in the analysis of variance) but only with the distribution of the variates. (But in the case where normality holds even approximately, ANOVA is generally the more efficient statistical test for detecting departures from the null hypothesis according to Sokal and Rohlf [1969].)

The Kruskal-Wallis test is based on the idea of ranking the variates in a sample after pooling all groups and considering them as a single sample for the purpose of ranking. The hypothesis tested is the same as that in a single classification ANOVA, viz.,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

where μ_i is the mean of the i^{th} level of the factor being considered. The test statistic is $\chi^2_\nu(\alpha)$, where ν is the number of degrees of freedom.

The values of the sample statistics are displayed in Table 5 for the entire sample and for the non-smokers in the sample. A comparison of these results with those in Table 4 for ANOVA shows that the two types of tests give qualitatively similar results.

At this stage, we cannot say whether each treatment is different from every other treatment, or whether it is simply true that there is one group which is different from the rest but all groups are not different from each

TABLE 5

χ^2 VALUES FROM THE KRUSKAL-WALLIS ANALYSIS OF FEV_1 AND FVC FOR TWO SUBSAMPLES

NS = result not signficant; * = signficance at 5 per cent level; ** = significance at 1 per cent level; *** = significance at 0.1 per cent level.

	Entire Sample		Non-smokers	
	FEV_1	FVC	FEV_1	FVC
Age	387.19***	296.75***	157.05***	111.02***
Height	204.15***	281.38***	120.73***	161.96***
Weight	16.45**		8.65NS	15.50**
Birthplace	4.5INS	2.79NS	5.80NS	2.62NS
Smoking habits	80.05***	88.44***		
Occupation	52.25***	55.36***	8.14*	13.34**

TABLE 6

SNK TESTS FOR THE EFFECT OF AGE ON FEV_1 AND FVC

Lines connect the age classes not found to be significantly different from each other. Class 1 = youngest, class 9 = oldest.

	CLASS	
	(increasing mean FEV_1 →)	(increasing FVC →)
All data	9 8 7 6 5 4 3 1 2	9 8 7 6 5 4 1 3 2
Non-smokers	8 9 7 6 5 4 3 1 2	8 9 7 6 5 4 1 2 3

other. Hence, the results from the single classification ANOVA were examined in greater detail, testing which means are different from which other individual or group means. The procedure used was the a posteriori Student-Newman-Keuls (SNK) test [Sokal and Rohlf 1969]. This is a stepwise method that uses the range as a statistic to measure differences among means.

The results of this analysis have been summarised by drawing lines connecting those categories of the independent variable that are not significantly different from each other at the 5 per cent level of significance. These categories have been arranged in order of increasing mean FEV_1 and FVC for age, height, weight, smoking habits and occupation in Tables 6-10. Birthplace was not included as it was never significant in the ANOVA tests.

The similarity between the end classes of the continuous variables in Tables 6-8 may be a reflection of small sample sizes in these classes. However, the fact that age group 2 (20-25 years) always had a higher mean FEV_1 than age group 1 (15-19) is more likely to be due to the peak in FEV_1 observed at about 27 years of age [A.D. Tucker, personal communication]. The similarity, especially for FVC, of the last five age classes suggests that vital capacity may converge for older age groups. It also seems that FEV_1 may converge for taller height classes, although this convergence is not so marked for FVC. The only weight class found to be different from the rest is that consisting of the lightest subjects. Weight seems to have a slightly greater effect on FVC than on FEV_1 (possibly because of its positive correlation with height). The ranking of the categories in terms of mean FEV_1 and mean FVC when overweight subjects are removed suggests that the increase in vital capacity with increasing weight falls off slightly after a certain weight is reached.

Table 9 shows that, from the SNK test, the category of heavy smokers is always considered to be different from the other categories. The categories are divided into more groups for FEV_1 than for FVC, suggesting that smoking has its greatest effect on FEV_1. It is interesting to note the ordering of the smoking categories in terms of mean FEV_1 and FVC. The categories are heterogeneous in the sense that pipe smokers may smoke some cigarettes and that no record was made of the length of time since the ex-smokers gave up smoking nor of whether they took it up again.

TABLE 7

SNK TESTS FOR THE EFFECT OF HEIGHT ON FEV_1 AND FVC

Lines connect the height classes not found to be significantly different from each other. Class 1 = shortest, class 8 = tallest.

	CLASS	
	(increasing mean FEV_1 →)	(increasing mean FVC →)
All data	1 2 3 4 5 6 8 7	1 2 3 4 5 6 7 8
Non-smokers	1 2 3 4 5 6 7 8	1 2 3 4 5 6 7 8

TABLE 8

SNK TESTS FOR THE EFFECT OF WEIGHT ON FEV_1 AND FVC

Lines connect the weight classes not found to be significantly different from each other. Class 1 = lightest, class 6 = heaviest.

	CLASS	
	(increasing mean FEV_1)	(increasing mean FVC)
All data	1 5 2 3 4 6	1 2 5 3 4 6
All data excluding overweight subjects	1 2 4 3	1 2 4 3
Non-smokers	Result nonsignificant	1 5 2 4 3 6
Non-smokers excluding overweight subjects	Result nonsignificant	1 2 4 3

TABLE 9

SNK TESTS FOR THE EFFECTS OF SMOKING ON FEV_1 AND FVC

Lines connect the smoking categories not found to be significantly different from each other. N = non-smokers, X = ex-smokers, P = pipe smokers, L = light smokers, M = medium smokers, H = heavy smokers.

	CATEGORY	
	(increasing mean FEV_1 →)	(increasing mean FVC →)
All data	H M X P L N	H M X P L N

TABLE 10

SNK TESTS FOR THE EFFECT OF OCCUPATION ON FEV_1 AND FVC

Lines connect the occupation categories not found to be significantly different from each other. PR = professional workers, TC = technicians, TR = tradesmen.

	CATEGORY	
	(increasing mean FEV_1 →)	(increasing mean FVC →)
All data	TR TC PR (all different)	TR TC PR (all different)
Non-smokers	TR TC PR	TR TC PR

From Table 10, it is evident that smoking accounts for most of the difference observed between technicians and tradesmen, but not for that observed between these groups and professional workers. Professional workers were always found to have higher vital capacities than the other occupation groups, even when smokers were excluded from the analysis.

Although the information obtained from a single classification analysis of variance on a multivariate system is limited, higher order classifications were not considered since further subdivisions of the data resulted in extremely small sample sizes for many of the combinations of factors and levels of factors. In any case, the usual higher order ANOVA design was not valid as the subclass numbers were unequal and hence the analyses were not orthogonal.

3.3 Correlation Analysis

A significant result from a single classification analysis of variance may be due to a complex interrelation of factors rather than to direct dependence of vital capacity on each factor. Some idea of the interdependence between factors can be gained by analysing the correlation patterns in the data. Correlation analysis is a technique for measuring the linear relationship between two variables and produces a single summary statistic which describes the strength of the association. One such measure is the Pearson correlation coefficient which is best suited for normally distributed data. The Kendall rank correlation coefficient is a related non-parametric measure. Both coefficients were used in the initial correlation analysis. Each ranges from +1 for perfect positive correlation to -1 for perfect negative correlation. Sample correlation coefficients were tested for significance by comparing their values with those obtained from standard tables. The hypothesis to be tested was

$$H_0 : p = 0 \quad \text{vs.} \quad H_1 : p \neq 0 ,$$

where p is the value of the correlation coefficient.

Correlation coefficients were calculated for all pairs of the metric variables: age, height, weight, body surface area, FEV_1 and FVC. Body surface area was introduced as a new independent variable to determine whether or not this combination of height and weight would be more highly correlated

with vital capacity than simple height or weight. When all data were considered, all pairs tested were significant at the 5 per cent level except for the combination age x surface area (Table 11). The same result held when smokers were excluded from the analysis. Table 11 shows that most of the correlations that were significant at the 5 per cent level were also significant at the 1 per cent level.

Another important measure associated with a correlation analysis is the coefficient of determination, r^2, the square of the correlation coefficient. The coefficient of determination is a measure of the proportion of the variation in one variable explained linearly by the variation in the other variable. The values of r^2 for age and height as independent variables and FEV_1 and FVC as the dependent variables are given in Table 12. These results suggest, once again, that age explains the greater proportion of the linear variation in FEV_1 and height explains more of the linear variation in FVC.

The values of the Pearson correlation coefficients should be compared with those obtained using the Kendall rank correlation coefficient (Table 13), which does not assume that the two variables have a bivariate normal distribution. The coefficient values are similar in both cases, except that the Kendall coefficient tends to give slightly lower values.

Simple correlation is not really what is needed to analyse these data since vital capacity is influenced by several independent variables. From the simple correlation coefficient we cannot conclude that the relationship between, say, weight and FEV_1 will be the same if a different range of heights is considered. In fact, when height is controlled (in a statistical sense), the relationship between weight and FEV_1 may no longer be apparent. This can be tested by calculating partial correlation coefficients which determine the relationship between two variables while adjusting or controlling the effects of one or more additional variables. The effect of the control variable(s) is assumed to be linear throughout its range; hence it is necessary to test for linearity before the analysis. Scattergrams of all pairs of variables were plotted but are not reproduced here. For each pair, the scatter of data points was very large but no distinct trends indicating non-linearity were evident in any case.

TABLE 11

PEARSON PRODUCT MOMENT CORRELATION COEFFICIENTS BETWEEN THE METRIC VARIABLES

NS = not significant, * = significance at 5 per cent level, ** = significance at 1 per cent level. A = entire sample considered, B = non-smokers only.

		Height	Weight	Surface Area	FEV_1	FVC
Age	A.	-0.201**	0.079**	-0.005NS	-0.621**	-0.541**
	B.	-0.240**	0.108**	0.010NS	-0.541**	-0.4444**
Height	A.		0.475**	0.727**	0.480**	0.575**
	B.		0.426**	0.696**	0.533**	0.612**
Weight	A.			0.948**	0.119**	0.175**
	B.			0.945**	0.102**	0.165**
Surface Area	A.				0.264**	0.344**
	B.				0.271**	0.351**
FEV_1	A.					0.919**
	B.					0.903**

TABLE 12

COEFFICIENT OF DETERMINATION FOR AGE AND HEIGHT, AND FEV_1 AND FVC (AS A PERCENTAGE)

		FEV_1	FVC
Age	Entire sample	38.56	29.27
	Non-smokers	29.27	19.17
Height	Entire sample	23.04	33.06
	Non-smokers	28.41	37.45

TABLE 13

KENDALL RANK CORRELATION COEFFICIENT BETWEEN THE METRIC VARIABLES

	Height	Weight	Surface Area	FEVI	FVC
Age	-0.144	0.059	-0.003	-0.467	-0.396
Height		0.344	0.551	0.356	0.430
Weight			0.810	0.072	0.113
Surface Area				0.174	0.230
FEV_1					0.769

Partial correlation coefficients were computed by controlling first for age, then for height, and finally (in selected cases) for both age and height. The equation used was

$$r_{ij.k} = \frac{r_{ij} - (r_{ik})\ (r_{jk})}{\sqrt{1-r_{ik}^2}\ \sqrt{1-r_{jk}^2}}$$

where k is the control variable, i and j are the independent and dependent variables, and r_{ij} is the ordinary bivariate correlation coefficient between i and j. When control of more than one variable is desired, this formula can be used to define and compute recursively each higher order partial from the previous one.

Partial correlation analysis was carried out on the data to determine whether the relationship between weight and FEV_1 and that between weight and FVC would disappear if height, or height and age, were controlled. This proved to be so (Table 14), but only when both age and height were controlled. The relationships between surface area x FEV_1 and surface area x FVC also disappeared when these controls were applied. This suggests that the only reason weight and surface area are positively correlated with FEV_1 and FVC is that height is positively correlated with all of these variables, i.e. relationships such as that between weight and FEV_1 are spurious.

3.4 Association Tables

As a correlation analysis can only be carried out on metric variables, association tables were used to examine patterns of association between the nominal variables - occupation, smoking habits, and birthplace - and between these and the metric variables. Associations between metric variables were also calculated, for comparison with the results of the correlation analysis.

Initially, to test the hypothesis of a strong relationship between pairs of variables, a two-way test of association was made. The sample statistic used was the 'G' statistic defined by

$$G = 2\left[\sum_i \sum_j f_{ij} \ln f_{ij} - \sum_i f_{ij} \ln f_{ij} - \sum_j f_{ij} \ln f_{ij} + n \ln n\right],$$

where f_{ij} is the number of entries in row i, column j, and n is the total

TABLE 14

PARTIAL CORRELATION COEFFICIENTS BETWEEN THE METRIC VARIABLES

NS = not significant, * = signficance at 5 per cent level, ** = significance at 1 per cent level. A = control for age, B = control for height, C = control for both age and height.

		Weight	Surface Area	FEV_1	FVC
Age	A.				
	B.	0.203**	0.209**	-0.610**	-0.531**
	C.				
Height	A.	0.503**	0.741**	0.463**	0.566**
	B.				
	C.				
Weight	A.		0.952**	0.215**	0.260**
	B.		0.998**	-0.142**	-0.136**
	C.			-0.023NS	-0.034NS
Surface Area	A.			0.333**	0.406**
	B.			-0.141**	-0.132**
	C.			-0.017NS	-0.025NS
FEV_1	A.				0.884**
	B.				0.895**
	C.				0.851**

sample size. The test statistic is χ^2 (a-1) (b-1) (α), where a is the number of rows in the table, b is the number of columns, and (a-1) (b-1) is the number of degrees of freedom for a χ^2 statistic at level α. The G statistic was used as it is usually a more exact test than the usual χ^2 test for large tables [Sokal and Rohlf 1969].

The results of the two-way test of association between all pairs of variables using all the data points are summarised in Table 15. With few exceptions, almost every variable of the nine considered was associated with every other variable at the 5 per cent level of significance. The main exceptions were the general lack of association between birthplace and the other variables, and between smoking and the variables height, weight, and body surface area. The next step in the analysis was to determine whether or not these relationships persisted when controlling for levels of height and age. In contrast to a partial correlation analysis the control, in this case, is literal rather than statistical. For example, the method of 'controlling' for height is to test each of the other pairwise associations over a number of classes of height. This exerts a tremendous drain on average cell frequencies and hence, for subsequent analyses, the class intervals for the metric variables were made larger so that the frequencies in each class were higher. Five age classes and four classes each of height, weight, body surface area, FEV_1 and FVC were used. This is the classical approach to 'elaboration analysis'. Initially, the bivariate table is examined and then test variables are entered successively to determine their role in the basic relationship.

The original bivariate analysis was carried out for the smaller number of size classes to see if the results were similar to those from the original larger number of classes. A comparison of the two analyses (Table 15) showed that the associations were generally similar except that those between surface area and the variables age, occupation and birthplace were no longer significant for the smaller number of classes. When the analysis of the smaller number of classes was carried out for non-smokers, weight was no longer significantly associated with FEV_1 or FVC and the associations between occupation and the other variables, particularly FEV_1 and FVC, generally decreased (Table 15).

At this stage, the main aim was to determine if the associations of occupation with FEV_1 and FVC persisted in the non-smoking group when age and height were controlled. Other pairwise combinations were also tested so that

TABLE 15

ASSOCIATION TABLE FOR ALL PAIRS OF VARIABLES WITH NO CONTROLS

NS = not significant, * = significance at 5 per cent level, ** = significance at 1 per cent level, *** = significance at 0.1 per cent level. A = original classes of the metric variables, B = smaller number of classes (see text), C = smaller number of classes for non-smokers.

		Height	Weight	Surface Area	Smoking Habits	Occupation	Birthplace	FEV_1	FVC
Age	A.	91.95***	65.67**	59.97*	117.20***	181.49***	60.28*	442.96***	332.80***
	B.	44.22***	35.68***	11.94NS	96.65	146.55***	47.13***	338.65***	263.36***
	C.	35.43***	31.86**	10.09NS		68.91***	33.04*	126.78***	80.11***
Height	A.		199.72***	568.79***	36.35NS	32.06**	35.46NS	223.86***	354.80***
	B.		112.68***	373.41***	19.00NS	20.80**	17.76NS	152.35***	237.01***
	C.		56.98***	169.80***		12.52NS	13.92NS	114.20***	127.22***
Weight	A.			911.38***	28.12NS	65.67**	27.39NS	38.81*	48.22**
	B.			496.80***	10.67NS	35.68***	24.07NS	18.42*	17.09*
	C.			240.79***		31.86**	19.72NS	15.52NS	13.17NS
Surface Area	A.				18.42NS	22.48*	38.22*	93.76***	138.47***
	B.				20.89NS	10.54NS	11.97NS	69.06***	109.39***
	C.					4.70NS	13.28NS	41.73***	63.56***
Smoking Habits	A.					80.88***	25.09NS	101.72***	69.44***
	B.					80.90***	25.13NS	75.43***	51.43***
	C.								
Occupation	A.						18.23NS	69.08***	60.76***
	B.						18.25NS	57.43***	54.04***
	C.						17.60NS	16.72*	15.91*
Birthplace	A.							14.27NS	15.05NS
	B.							13.26NS	11.48NS
	C.							13.46NS	12.36NS
FEV_1	A.								1019.84***
	B.								732.22***
	C.								327.06***

any previously 'hidden' relationships could be located. When age was controlled, and only non-smokers were considered (Table 16), relationships emerged between birthplace and height, and between birthplace and FVC. On the other hand, the association between occupation and FEV_1 was lost. When height was controlled and only non-smokers were considered (Table 17), there was no longer any association between birthplace and FVC but associations between birthplace and age and weight became evident. Again, association between occupation and FEV_1 was not evident. When both age and height were controlled, each pairwise combination had to be considered over five age classes x four height classes, or 20 combinations of age and height. Consequently, the entire table has not been reproduced here; however Table 18 summarises those combinations for which significance (at the 5 per cent level or better) was observed for at least one combination of age class and height class. When all data were considered, the combinations that exhibited no association at all were weight x FVC, weight x occupation, surface area x birthplace, and birthplace x smoking habits. Of those that did exhibit association, the association was lost for the combinations weight x birthplace and occupation x birthplace when the non-smoking group was analysed (Table 18). The associations between occupation and both FEV_1 and FVC were retained.

In summary, it seems that although age and height are obviously the most important factors associated with FEV_1 and FVC, others such as occupation and smoking habits may also be associated with these measures of lung capacity. The association table analysis is not conclusive, however, as the sizes of the samples between which comparisons were made were very small when both age and height were controlled. Hence, the above analyses serve as a first descriptive technique for estimating the degree of association among the variables. The next technique to be employed, multiple regression analysis, leads to further insights into the relationships between the variables.

3.5 Multiple Regression Analysis

Multiple regression analysis allows the researcher to study the linear relationship between a set of independent variables and a dependent variable while taking into account the interrelationships among the independent variables [Nie et al. 1975]. The basic goal of multiple regression is to produce a linear combination of independent variables which will correlate as highly as possible with the dependent variable. This can then be used to predict values of the dependent variable, and the importance of each of the

TABLE 16

ASSOCIATION TABLE FOR PAIRS OF VARIABLES OVER FIVE AGE CLASSES

NS = not significant, * = significance at 5 per cent level, ** = significance at 1 per cent level, *** = significance at 0.1 per cent level.

	Number	Weight	Surface Area	Occupation	Birthplace	FEV_1	FVC
Height	1.	17.94**	18.54**	6.41NS	3.36NS	15.69**	8.76NS
	2.	29.30***	57.97***	6.58NS	14.96NS	45.53***	60.78***
	3.	21.79**	66.04***	2.69NS	10.65NS	14.89*	30.33***
	4.	17.89**	41.76***	6.05NS	15.77*	29.82***	16.21**
	5.	7.24NS	8.98NS	3.86NS	6.84NS	3.57NS	4.58NS
Weight	1.		31.49***	11.02NS	12.65NS	9.54NS	8.29NS
	2.		80.37***	7.11NS	10.08NS	20.30**	9.23NS
	3.		63.74***	3.87NS	11.95NS	9.08NS	9.95*
	4.		55.09***	7.35NS	6.85NS	2.23NS	4.28NS
	5.		11.44*	2.22NS	1.95NS	0.57NS	3.31NS
Surface Area	1.			15.76*	12.77NS	15.22*	14.18*
	2.			13.67*	12.29NS	26.93***	33.77***
	3.			9.99NS	8.12NS	7.81NS	20.81**
	4.			2.86NS	7.06NS	9.64NS	10.72NS
	5.			1.95NS	7.47NS	1.13NS	3.41NS
Occupation	1.				7.98NS	5.55NS	10.82*
	2.				6.85NS	6.67NS	5.71NS
	3.				12.85NS	6.49NS	11.82*
	4.				6.18NS	8.17NS	4.26NS
	5.				11.52NS	3.40NS	3.69NS
Birthplace	1.					2.95NS	2.03NS
	2.					14.34NS	16.64*
	3.					12.59NS	6.99NS
	4.					6.52NS	7.28NS
	5.					3.17NS	7.76NS
FEV_1	1.						36.99***
	2.						102.29***
	3.						42.95***
	4.						70.52***
	5.						14.87**

TABLE 17

ASSOCIATION TABLE FOR PAIRS OF VARIABLES OVER FOUR HEIGHT CLASSES

NS = not significant, * = significance at 5 per cent level, ** = significance at 1 per cent level, *** = significance at 0.1 per cent level.

	Height Group Number	Weight	Surface Area	Occupation	Birthplace	FEV_1	FVC
Age	1.	6.86NS	8.11*	7.27NS	10.94NS	8.79NS	4.59NS
	2.	39.82***	15.98*	55.06***	36.13*	83.95***	48.01***
	3.	19.53NS	6.24NS	10.65NS	22.60NS	20.34**	17.69*
	4.	2.81NS	3.26NS	8.99NS	1.02NS	8.99*	3.26NS
Weight	1.		23.83***	0.02NS	3.24NS	2.63NS	3.48NS
	2.		115.32***	4.31NS	23.86**	2.45NS	1.99NS
	3.		54.66***	7.85NS	6.62NS	12.76*	12.53*
	4.		2.21	2.95NS	2.21NS	0.17NS	2.21NS
Surface Area	1.			0.35NS	4.06NS	1.71NS	2.24NS
	2.			3.61NS	12.34NS	4.91NS	4.33NS
	3.			3.49NS	10.15NS	6.97NS	9.67*
	4.			0.62NS	0.28NS	3.26NS	6.03*
Occupation	1.				7.20NS	7.55NS	6.88NS
	2.				10.82NS	8.25NS	6.87NS
	3.				14.75NS	3.82NS	9.46*
	4.				0.62NS	3.59NS	0.62NS
Birthplace	1.					2.43NS	5.62NS
	2.					11.70NS	6.87NS
	3.					9.26NS	13.44NS
	4.					0.62NS	10.75NS
FEV_1	1.						4.79NS
	2.						162.73***
	3.						66.60***
	4.						3.26NS

TABLE 18

ASSOCIATIONS BETWEEN PAIRS OF VARIABLES WHEN BOTH AGE AND HEIGHT ARE CONTROLLED

NS = association never significant for any combination of age class and height class, * = significance (at 5 per cent level or better) observed for at least one combination of age class and height class. A = entire sample, B = non-smokers only.

		Surface Area	Smoking Habits	Occupation	Birthplace	FEV_1	FVC
Weight	A.	*	*	NS	*	*	NS
	B.	*		NS	NS	*	NS
Surface Area	A.		*	*	NS	*	*
	B.			*	NS	*	*
Smoking Habits	A.			*	NS	*	*
	B.						
Occupation	A.				*	*	*
	B.				NS	*	*
Birthplace	A.					*	*
	B.					*	*
FEV_1	A.						*
	B.						*

independent variables in that prediction can be assessed. When there are many independent variables, multiple regression techniques also enable the subset of variables (that gives the best linear prediction equation) to be found.

The method employed for this analysis was a stepwise multiple regression which is a statistical technique for selecting the independent variables in the order of their importance. The criterion of importance is based on the reduction of sums of squares of deviations from the regression, and the independent variable most important in this reduction in a given step is entered into the regression. The program STEPR from the IBM scientific subroutine package (SSP) was used. Regressions were carried out separately for FEV_1 and FVC as dependent variables, and a number of combinations of independent variables including interactions between them.

The independent metric variables used were age, height, weight, and surface area. (Pairwise interactions between the metric variables were also considered, together with polynomial terms; however, these additions made no significant improvement to any of the regression equations.) Next, the categorically independent variables, smoking habits and occupation, were introduced to determine their effect, if any, when age, height, and the other independent variables were not controlled. Other independent variables were formed from the six categories of smoking, three categories of occupation, and all possible interactions of the smoking categories with each of the occupation categories. Finally, the categorical variables were combined with the metric variables to assess the importance of the former when the metric variables were controlled.

Regressions were performed on the categorical variables by defining them as dummy variables. A set of dummy variables is 'created' by treating each category of a nominal variable as a separate variable and assigning arbitrary scores for all cases, depending upon their presence or absence in each of the categories. All cases (individuals) in a sample can be assigned arbitrary scores of, say, 1 or 0 on each variable. Then each person in the sample would be scored a 1 on the appropriate category and a zero on all others [Nie et al. 1975].

The results of the regression analyses are summarised in Table 19. All regressions were significant at the $\alpha = 0.01$ level. The variables that explained at least 1 per cent of the variation in the system are given on the

TABLE 19

SUMMARY OF REGRESSION RESULTS FOR VARIOUS SUBSAMPLES OF THE DATA AND VARIOUS COMBINATIONS OF THE INDEPENDENT VARIABLES

Variables were entered into the regression unless the addition accounted for at least another 1 per cent of the variation in vital capacity.

Independent Variables Used		Sample Size	Proportion SS Explained	Multiple Correlation Coefficient	Variables Entered in Order and Additional SS Explained
Entire sample					
All metric variables	FEV_1	923	51.7	0.719	Age 0.385, Ht 0.132
	FVC	923	51.9	0.720	Ht 0.330, Age 0.189
All metric and categorical variables	FEV_1	923	54.6	0.738	Age 0.385, Ht 0.132, TR 0.019, H 0.010
	FVC	923	53.7	0.733	Ht 0.330, Age 0.189, TR 0.018
Entire sample excluding 'overweight' and 'asthmatic' subjects					
All metric variables	FEV_1	797	55.5	0.741	Age 0.385, Ht 0.179
Smoking categories	FEV_1	797	5.3	0.225	M 0.021, H 0.019 X 0.013
	FVC	797	1.1	0.104	H 0.011
Occupation categories	FEV_1	797	4.1	0.199	TR 0.030, TC 0.011
	FVC	797	4.8	0.215	TR 0.037, TC 0.010
Smoking and Occupation categories and pairwise combinations of these	FEV_1	797	7.4	0.265	TR 0.030, M 0.016, H 0.017, X 0.011
	FVC	797	4.8	0.215	TR 0.037, TC 0.010
Same as above plus metric variables	FEV_1	797	56.7	0.752	Age 0.376, Ht 0.179, TR 0.012
	FVC	797	55.7	0.745	Ht 0.374, Age 0.094
Non-smokers					
Metric variables	FEV_1	478	46.5	0.681	Age 0.292, Ht 0.173
	FVC	478	46.8	0.683	Ht 0.374, Age 0.094
Non-smokers excluding 'overweight' and 'asthmatic' subjects					
Metric variables	FEV_1	478	50.7	0.711	Ht 0.328, Age 0.179
	FVC	478	49.7	0.704	Ht 0.406, Age 0.091

right hand side of the table in the order that they were entered into the regression using the stepwise method. The proportion of the total sums of squares explained by the addition is also included.

The regression models analysed for different subsets of the sample are not directly comparable because even though there would seem to be less variation in a system from which extreme data points have been excluded, the significance of a regression decreases (in general) with decreasing sample size. Nevertheless, it is evident from Table 19 that when overweight and 'asthmatic' individuals (those with FEV_1/FVC ratios less than 0.7) were excluded, a larger proportion of the total variation was explained than was the case for the entire sample. The same was true for the third and fourth subsamples.

It is also true that when the independent variables are measured on different units (such as age in years and height in centimetres), the regression coefficients within a given model are not directly comparable. The only sensible way to compare the relative effects of each independent variable on the dependent variable is to standardise each variate by subtracting its mean and dividing by its standard deviation. The regression equation also becomes somewhat simpler to interpret in this case, as the constant term is zero. When the entire sample was considered, the resulting equations were

$$FEV_1 = -0.546\ \text{Age} + 0.370\ \text{Height}$$
$$FVC = -0.443\ \text{Age} + 0.486\ \text{Height}$$

The regression coefficients of these equations are directly comparable. Thus, as was found in previous analyses, age is the most important variable determining FEV_1 and height is most important in determining FVC.

Even if the regression coefficients are clearly statistically significant, it is not uncommon to find that the fraction of the variance in the dependent variable attributable to the regression is less than half. For the above equations, the percentages of the variances explained were 51.7 and 51.9 respectively. These numbers are an expression of the overall accuracy of the equation as given by the square of the multiple correlation coefficient (R^2). They indicate that much of the variation in FEV_1 and FVC must be due to other variables not included in the regression. Weight and body surface area are not good candidates for these other variables because when they were

entered into the regression as independent variables along with age and height, R^2 was not increased to any significant degree. The nominal variables had a slightly greater effect. Table 19 shows that, when smoking habits and occupation were added to the equation, the tradesman category of occupation resulted in an increase in R^2 of more than 1 per cent, both for FEV_1 and FVC. The heavy smoking group was also added for FEV_1. These results are difficult to interpret, however, as the method used to set up the dummy regressions requires omission of one of the categories of each of the nominal variables. The omitted category is then a reference category and its effects will contribute in some way to the constant term of the regression equation.

The main regression models to be noted from the second subsample in Table 19 (the entire sample excluding overweight and asthmatic subjects) are those in which the nominal variables were considered without the metric variables. When only those categories of the nominal variables that explain at least 1 per cent of the variation were included, it was found that 5.3 per cent of the variation in FEV_1 and 1.1 per cent of the variation in FVC was explained by smoking. Occupation accounted for 4.1 per cent of the variation in FEV_1 and 4.8 per cent of that in FVC. Medium, heavy, and ex-smokers contributed a significant amount to the regression of FEV_1, but only heavy smokers to FVC. Both technicians and tradesmen contributed to FEV_1 and FVC. (Note that this does not mean that no contribution is made by professionals since this group was used as the reference category in the dummy regression. The same applies for smoking categories; in this case, non-smokers were used as the reference category.) When smoking habits, occupation and interactions between the categories of these two variables were considered, the variation explained was 7.4 per cent for FEV_1 and 4.8 per cent for FVC.

From these results, it can be surmised that smoking has a greater effect on FEV_1 than it does on FVC and occupation may have a slightly greater effect on FVC than on FEV_1. However, these equations are the results of regressions from which age and height have been excluded even though they are known to be the two most important variables. When age and height are controlled, the nominal variables may not have such a great effect. This was found to be the case (Table 19). For both FEV_1 and FVC, the only variable other than age and height which explained more than 1 per cent of the variation was the tradesman category. This category must be interpreted in relation to the reference category of the dummy regression for occupation (professionals) and also in relation to the high degree of association between smoking habits and

occupation. Thus, the addition of this variable represents the addition of a combination of effects of smoking and occupation. Nevertheless, it only accounts for an additional 1.2 per cent of the variation in FEV_1 and 1.3 per cent of the variation in FVC, after height and age have been considered.

When the non-smoking group was considered (subsample 3 in Table 19), results were similar to those obtained for the first set of equations in the table except that the multiple correlation coefficient was lower. Once again, the proportion of the variation explained was greater when 'asthmatic' and overweight individuals were excluded.

The residuals from several of the regressions in Table 19 were examined to detect any non-linearity in the system. The pattern of residuals, observed on a scatterplot, may indicate a need for adding terms to the equations, such as multiplicative terms to handle interaction or polynomial terms to handle curvilinearity. Outliers are also clearly visible in a scatterplot. Residual analysis is also useful for determining whether the assumptions about the error terms in a regression are met. In regression analysis, it is assumed that the error components (i.e. the deviations off the data points from the regression line) (i) are independent, (ii) have a mean of zero, and (iii) have the same variance throughout the range of the dependent variable. An examination of the scatter of the residuals showed that although the scatter was very large, none of the above assumptions were violated for any of the regression models. Hence, there was no reason to assume that the systems were nonlinear. The scatterplot for Equation (1) shows the typically wide scatter of the residuals about the regression line (Figure 1). The magnitude of the variance for the other regression models was similar.

If a regression model with the narrowest possible confidence limits is desired, then we can do no better than to consider the two regression equations for which the entire sample was used as, in general, the confidence limits become smaller as the sample size becomes larger. These equations were

$$FEV_1 = 0.04796 \text{ Height} - 0.03859 \text{ Age} - 2.85609 \qquad (1)$$

and

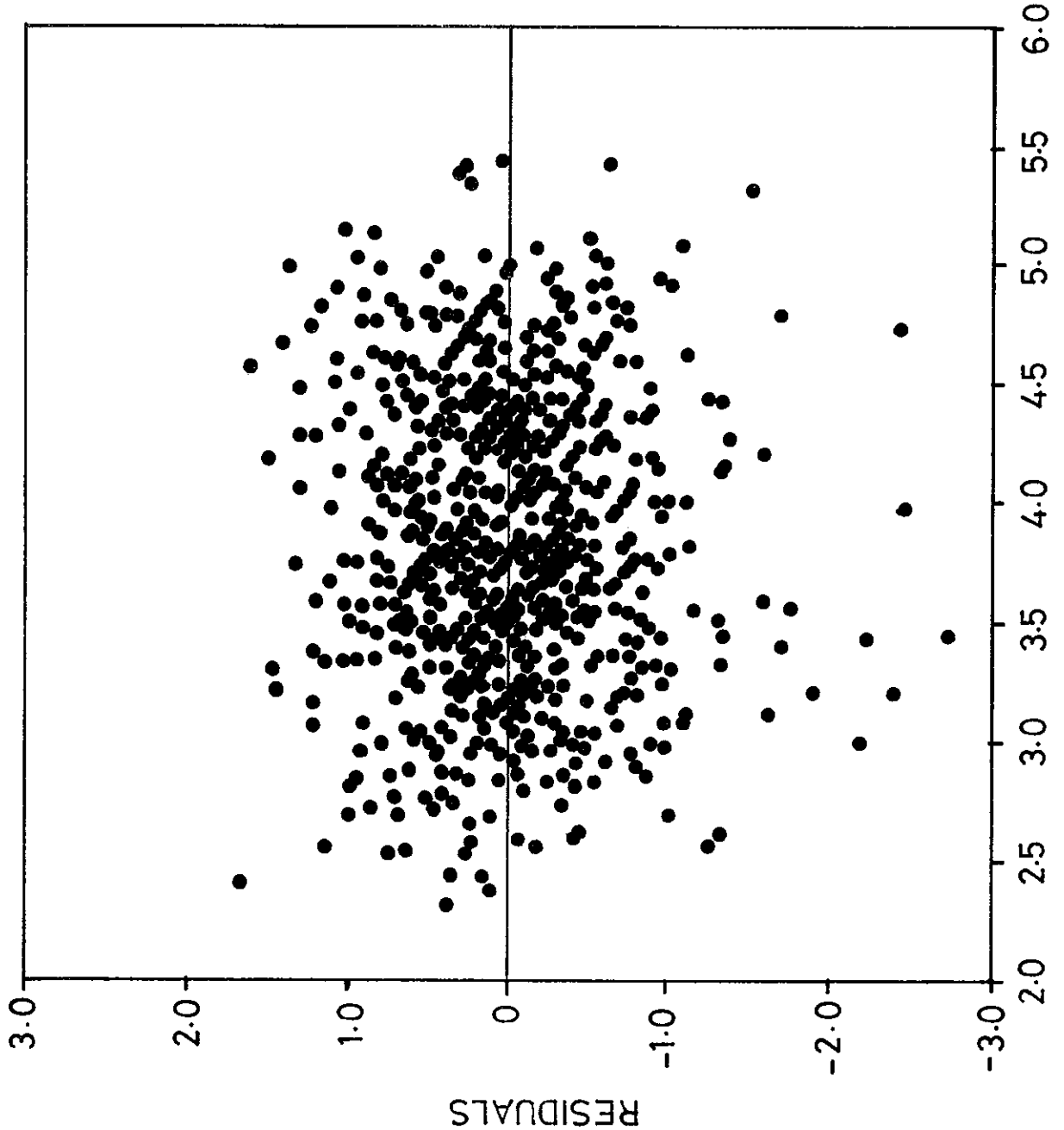

Figure 1. Residual differences between observed and predicted FEV_1 for the model FEV_1 - 0.04796 Height - 0.03859 Age - 2.85609

$$FVC = 0.06609\ Height - 0.03288\ Age - 5.24753 \qquad (2)$$

They may be used to predict the vital capacity for a new member of the population. The 95 per cent confidence interval for this prediction is given by

$$Y^1 - t_{920}(0.05)S\hat{y} < \alpha + \beta_1 x_1 + \beta_2 x_2 < Y^1 + t_{920}(0.05)S\hat{y} \quad ,$$

where Y^1 is the value predicted by the equations for some specific values of age and height, $\alpha + \beta_1 x_1 + \beta_2 x_2$ is the prediction equation, $t_{920}(0.05)$ is Student's t with n-3 degrees of freedom where n is the sample size, and

$$S\hat{y} = S * \sqrt{1 + \frac{1}{n} + c_{11}x_1^2 + c_{22}x_2^2 + 2c_{12}x_1x_2}$$

with $x_i = \hat{x} - x_i$ for variable i

$$c_{11} = \varepsilon x_2^2/D$$

$$c_{12} = \varepsilon x_1 x_2/D$$

$$c_{22} = \varepsilon x_1^2/D$$

$$D = (\varepsilon x_1^2)(\varepsilon x_2^2) - (\varepsilon x_1 x_2)^2$$

$S = \sqrt{MS_{dev}}$ - the mean - squared deviation from the analysis of variance [Snedecor and Cochran 1971].

The 95 per cent confidence limits are then given by

$$Y^1 \pm (1.96)S\hat{y} \text{ litres,}$$

or, since only the 5 per cent of the population which has an unusually low FVC or FEV_1 is of interest, we would use the one-sided confidence interval given by

$$Y^1 - (1.65)S\hat{y} \quad .$$

Some examples of the predictions obtained with Equation (1) for FEV_1, and their confidence limits, are given in Table 20. The predictions for a similar

TABLE 20

PREDICTED FEV_1 VALUES FROM EQUATION (1) AND THEIR CONFIDENCE LIMITS, FOR SOME VALUES OF AGE AND HEIGHT

The values in brackets are predicted FEV_1 values from Equation (3).

Age (years)	Height (cm)	Predicted FEV_1	Confidence Limits Two-sided	One-sided
20	160	4.05(3.61)	±1.132	-0.953
20	190	5.48(5.03)	±1.134	-0.955
30	140	2.70(2.35)	±1.145	-0.964
30	170	4.14(3.77)	±1.128	-0.950
40	140	2.31(1.14)	±1.144	-0.963
40	180	4.23(3.92)	±1.128	-0.950
40	200	5.19(4.87)	±1.139	-0.959
50	160	2.89(2.65)	±1.131	-0.952
60	140	1.54(1.39)	±1.146	-0.965
60	190	3.94(3.76)	±1.134	-0.954

model derived by McCullagh and Balaam [1975] are also included. Their prediction equation was

$$FEV_1 = 0.04731 \text{ Height} - 0.03188 \text{ Age} - 3.3208 \quad (3)$$

which tends to give lower values than Equation (1). The confidence intervals for the predictions from Equation (3) were calculated by considering the average standard error of the regression rather that the standard error for each prediction. This gave uniform confidence intervals over the whole range of ages and heights: ± 0.8761 for the two-sided limit and -0.7331 for the one-sided limit. The method used in the present study gives a more accurate indication of the confidence with which predictions can be made for a particular height and age. In most cases, these confidence limits were larger than those obtained by McCullagh and Balaam [1975].

3.6 Median Tests

The regression analyses described in the previous section resulted in prediction equations that give the best possible estimates for vital capacity when the height and age of the subject are known. Most studies on vital capacity stop at this point. However, the regression equation can only be used to predict the 'average' vital capacity for the population at a specific time point; it does not indicate how vital capacity falls off with age for each individual. Trends in vital capacity that are evident for individuals over time may not be evident when the sample is considered as a whole.

An hypothesis suggested by A.D. Tucker (personal communication) is that vital capacity falls off at a different rate and a different age, between the ages of 32 and 47, for the six categories of smokers. On the basis of clinical tests of AAEC employees during the decade 1965-1975 there is reason to suspect that trends such as those illustrated in Figure 2 may occur. Vital capacity seems to fall more rapidly and at an earlier age for smokers than it does for non-smokers. The fall-off in vital capacity may also be different for men engaged in different occupations. For example line 1, in Figure 2 may represent the decrease in vital capacity (FEV_1 or FVC) for non-smokers, line 2 that for medium smokers, and line 3 that for heavy smokers; or line 1 may represent the decrease in vital capacity for individuals classified as PR, line 2 that for individuals classified as TC, and 3 for the category TR. Thus we could hypothesise that non-smoking professionals will have a smaller

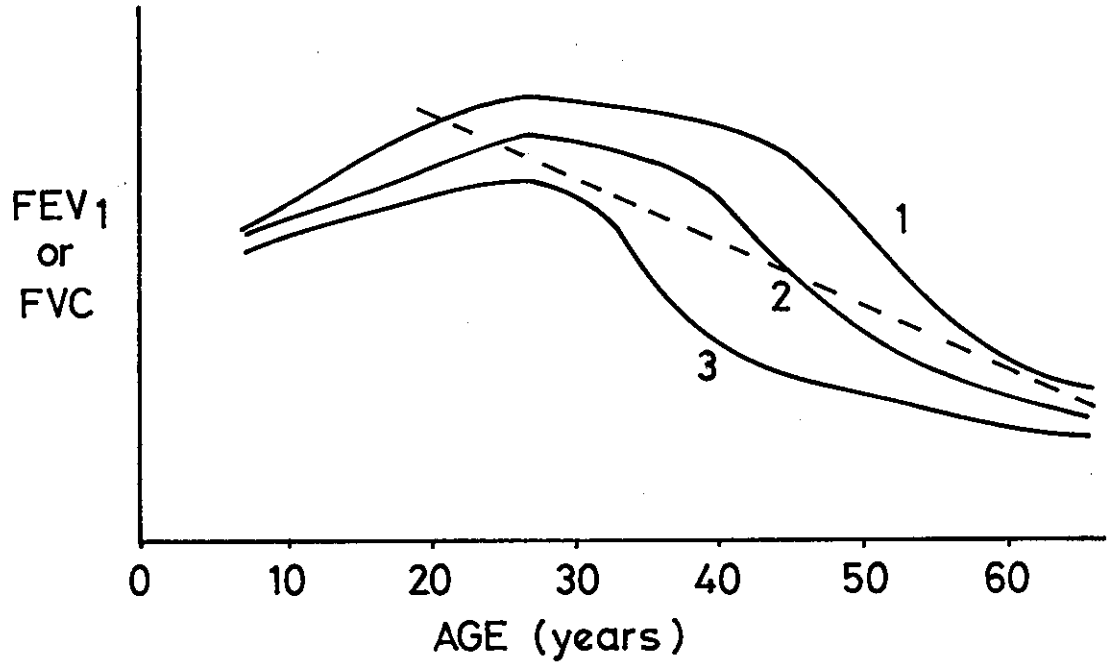

Figure 2. Hypothesised decreases in vital capacity over time. Lines 1, 2, and 3 could represent nonsmokers, light-medium smokers, and medium-heavy smokers respectively. The dashed line is the regression line that would be constructed through the points on the graph.

average decrease in vital capacity, beginning at a later age than heavily smoking tradesmen.

It is difficult to test this hypothesis from the time series data, as there are very few subjects who have been monitored for a long enough period to show a significant decrease in vital capacity. It may be possible to do so at some later date. In the meantime, some support for the hypothesis represented by Figure 2 can be gained from the previous analyses and from the following median tests. If the hypothesis were true, then the median vital capacity would decrease along some sort of continuum from non-smoking professionals to heavily smoking tradesmen. Testing for differences between medians may actually give better support for this hypothesis than testing for differences between means, because the median is less sensitive to large deviations in the vital capacity.

The hypothesis of unequal medians was tested for all combinations of each of the three categories of occupation with each of the six categories of smoking using the median test [Conover 1971]. These categories combined to provide a total of 18 subsamples. The two hypotheses to be tested were

H_0: all 18 subsamples have the same median vs.
H_1: at least two of the sample medians are unequal.

The median for the entire array of sample values was calculated and a 2 x 18 contingency table was constructed consisting of the number of observations which were greater than the grand median and the number less than the grand median for each of the 18 subsamples. The sample statistic was then defined as

$$T = \frac{N^2}{ab} \sum_{i=1}^{18} \left[\frac{O_{1i} - \frac{n_i a}{N}}{n_i} \right]^2$$

where N is the total sample size, a is the total number of observations above the grand median, b is the total number of observations less than or equal to the grand median, O_{1i} is the number of observations greater than the grand median for each subsample, and n_i is the sample size for the i^{th} subsample.

The test statistic was $\chi_{(1-\alpha)}$, the $(1-\alpha)$ quantile of a χ^2 random variable with c-1 degrees of freedom, where c is the number of groups being compared. The decision rule was to reject H_0: all 18 medians were equal, if T exceeded $\chi_{(1-\alpha)}$ - otherwise H_0 was accepted. The test is only approximate if more than 20 per cent of the sample sizes are less than 10, or if any are less than 2.

A sample statistic of 72.43 was obtained for FVC when the test was carried out for all subsamples. This was significant at 0.001 ($\chi_{(1-0.001)}$ = 40.79). A value of 87.63, which was also highly significant, was obtained from the same analysis for FEV_1. When overweight individuals and those with FEV_1/FVC ratios less than 0.7 were excluded from the analysis, the values of the sample statistics decreased for both FEV_1 and FVC but they were still significant at the α = 0.001 level. These analyses are summarised in Table 21 by ranking the combined categories on the basis of the proportion of the subsample greater than the grand median. If there was no difference in the way the occupation categories were ranked in Table 21, we would expect there to be two occurrences of each of the three occupation categories for every set of six ranking. However, Table 22 shows that the professional category (PR) tends to have its greatest representation in the six highest ranked subsamples, the technician category (TC) in the middle six, and the tradesmen category (TR) in the lowest six. When the same analysis was carried out for the smoking categories (Table 23), non-smokers (N) and light smokers (L) had their greatest densities in the first 12 ranked subsamples; ex-smokers (X), pipe-smokers (P) and medium smokers (M) showed no marked concentration in any of the three sets of six subsamples; and heavy smokers (H) occurred predominantly in the last six ranked subsamples.

In all cases, the hypothesis, H_0, was rejected and the alternative hypothesis, H_1: at least two of the sample medians are unequal, was accepted. The next problem was to determine which of the medians were different from one another. Hence each of the 18 subsamples were tested against the other 17 subsamples. Of the 153 pairs considered during examination of the entire sample, 48 had significantly different medians for FVC and 59 had significantly different medians for FEV_1 at the α = 0.05 level of significance. For both FEV_1 and FVC the subsamples that were the most different from the other categories most often were TRX and TRM - tradesmen who were ex-smokers and tradesmen who were medium smokers. The test was not valid for TRH because of its very small sample size.

TABLE 21

COMBINED SMOKING AND OCCUPATION CATEGORIES RANKED ON THE BASIS OF PERCENTAGE OF THE SUBSAMPLE GREATER THAN THE GRAND MEDIAN

This percentage is indicated in brackets and followed by the sample size. The subsamples are denoted by three letters, the first two corresponding to the occupation category and the last one to the smoking category, as outlined in Table 1.

Entire Sample		Entire Sample Excluding Overweight and Asthmatic Subjects	
FEV_1	FVC	FEV_1	FVC
PRP(71.43)14	PRP(78.56)14	PRL(60.87)23	PRL(78.26)23
PRL(70.83)24	PRL(75.00)24	TCN(54.82)166	PRP(76.92)13
TCN(63.48)178	PRM(63.89)36	PRP(53.85)13	PRM(66.67)27
PRN(61.17)206	TCN(57.30)178	PRN(50.27)187	TCN(57.83)166
PRX(55.56)27	PRN(56.80)206	PRM(48.15)27	PRN(57.22)187
PRM(52.78)36	TCL(48.78)41	TRN(44.30)79	PRX(50.00)24
TCL(48.78)7	PRX(48.15)27	TRL(42.11)38	TRH(50.00)6
TRN(47.87)94	TCX(45.83)48	PRX(41.67)24	TRL(47.37)38
TCX(45.83)48	TRL(45.83)48	TCX(39.53)43	TRN(44.30)79
TRL(41.67)48	TEN(41.49)94	TCP(35.29)23	TCX(44.19)43
TCM(35.00)60	TRM(38.33)60	TCL(34.29)35	TCL(42.86)35
TCP(33.33)30	TCP(36.67)30	TCM(26.92)52	TCM(40.38)52
TRX(25.93)27	TRH(25.00)12	TRX(16.67)24	TC)(39.13)23
TCH(20.00)5	TRM(20.97)62	TRH(16.67)6	TRM(23.91)46
TRM(17.74)62	TCH(20.00)5	TRM(15.22)46	TRX(20.83)24
TRH(16.67)12	TRX(18.52)27	TCH(0.0)4	TCH(0.0)4
PRH(11.11)9	PRH(11.11)9	PRH(0.0)5	PRH(0.0)5
TRP(0.0)2	TRP(0.0)2	TRP(0.0)2	TRP(0.0)2

TABLE 22

NUMBER OF OCCURRENCES OF EACH OF THE OCCUPATION CATEGORIES IN EACH SET OF SIX RANKINGS IN TABLE 21

	Entire Sample		Entire Sample Excluding Overweight and Asthmatic Subjects	
	FEV_1	FVC	FEV_1	FVC
	5	4	4	5
PR	0	1	1	0
	1	1	1	1
	1	2	1	1
TC	4	3	4	3
	1	1	1	2
	0	0	1	0
TR	2	2	1	3
	4	4	4	2

TABLE 23

NUMBER OF OCCURRENCES OF EACH OF THE SMOKING CATEGORIES IN EACH SET OF SIX RANKINGS IN TABLE 21

	Entire Sample		Entire Sample Excluding Overweight and Asthmatic Subjects	
	FEV_1	FVC	FEV_1	FVC
	2	2	3	2
N	1	1	0	1
	0	0	0	0
	1	0	0	1
X	1	2	2	1
	1	1	1	1
	1	1	1	1
P	1	1	1	0
	1	1	1	2
	1	2	1	1
L	2	1	2	2
	0	0	0	0
	1	1	1	1
M	1	1	1	1
	1	1	1	1
	0	0	0	0
H	0	0	0	1
	3	3	3	2

Obviously, the next step in this analysis would be to amalgamate those groups whose medians were not significantly different from one another at the $\alpha = 0.05$ level of significance and then to subdivide the total sample into a number of categories that can be considered to differ from one another on the basis of their median vital capacities. This step was not carried out because it was felt that the median tests had already served their purpose: they have confirmed that FEV_1 and FVC analyses should be considered separately within a number of different smoking and occupation categories.

4. GENERAL DISCUSSION AND CONCLUSIONS

This study has resulted in prediction equations for FEV_1 and FVC specific to the male employees of the AAEC Research Establishment. Although these equations, viz:

$$FEV_1 = 0.04796 \text{ Height} - 0.03859 \text{ Age} - 2.85609$$

and

$$FVC = 0.06609 \text{ Height} - 0.03288 \text{ Age} - 5.24753$$

depend only on the variables age and height, it has been shown that more subtle, but nevertheless important, effects on vital capacity may be brought about by smoking habits and occupation. The problem now is one of quantifying the contribution of the latter variables to vital capacity. We need to be able to separate the effects of smoking habits and occupation both from each other and from the effects of the other variables. This task will be more readily accomplished by analysis of trends in vital capacity over time for individual subjects belonging to different smoking and occupation categories, rather than analysing the entire sample from a single year.

However, the strength of the conclusions that can be derived from data of this type is limited. Little support for a hypothesis can be gained unless experiments are set up to test that hypothesis explicitly. Analysis of the type of data on which the present study is based can only give an indication of the factors that might be important to the structure of the data; it cannot result in conclusive evidence to verify the relevance of these factors.

Another problem which precludes the possibility of conclusive results is the way in which individuals have been categorised. Many of the categories are extremely heterogeneous. For example, the occupation categories are assumed to reflect the working environment, yet obviously there is a great diversity of working conditions within each of the categories of professional workers, technicians and tradesmen. It would probably be more profitable for future studies to classify individuals in terms of the areas in which they work. The smoking categories could also be defined in a different way. The ex-smokers, in particular, form a very heterogeneous group. The length of time since they gave up smoking and whether or not they took it up again after their first medical examination has not been recorded. Nor was the age at which each subject began smoking. Perhaps it would be worthwhile to define smoking as a quantitative variable by obtaining a rough estimate of the number of cigarettes each subject has smoked over his life, and combining this in some way with the length of time since he stopped smoking.

The collection of measurements of vital capacities on employees of the Research Establishment is a long-term project and it is hoped that eventually it will be possible to determine in a more precise way how smoking habits and occupation affect vital capacities. The ultimate aim is to find separate regression equations for each of the categories of smokers and to determine their one-sided confidence intervals. Then, for any individuals falling outside these confidence limits, further examinations could be carried out, and the possibility of their working environments having some effects on their respiratory systems could be investigated.

Commonwealth Medical Officers' Handbook [1965] - Australian Government Publishing Service, Canberra.

Conover, W.J. [1971] - Practical Non-parametric Statistics. John Wiley and Sons, Inc., New York.

Finney, D.J. [1963] - An Introduction to the Theory of Experimental Design. University of Chicago Press.

Guyton, A.C. [1971] - Textbook of Medical Physiology. Saunders Philadelphia, p.829.

McCullagh, S.F. and Balaam, L.N. [1971] - Vital Capacity and One-second Forced Expiratory Volume in Australian Male Factory Workers. Med. J. Aust., 2:173-175.

Nie, N.H., Hull, C.H. Jenkins, J.G., Steinbrenner, K. and Bent, D.H. [1975] - Statistical Package for the Social Sciences. McGraw-Hill, New York.

Snedecor, G.W. and Cochran, W.G. [1971] - Statistical Methods. Iowa State University Press.

Sokal, R.R. and Rohlf, F.J. [1969] - Biometry. Freeman and Co., San Francisco.