Monitoring infectious diseases using routine microbiology data II. An example of regression analysis used to study infectious gastroenteritis

BY HILARY E. TILLETT

Communicable Disease Surveillance Centre, Public Health Laboratory Service,
61 Colindale Avenue, London NW9 5EQ

(*Received* 2 *May* 1980)

SUMMARY

Routine data used to study infectious diseases may contain biases which obscure trends. A 16-year series (up to 1968) of routine laboratory data was used to study patterns of incidence of infective gastroenteritis for which no laboratory diagnosis could be made. An artificial pattern was detected. This arose because GPs tended to refer a greater proportion of their patients during dysentery epidemics. Multiple regression analysis was used to separate out this effect so that the underlying trends could be observed.

The seasonal pattern of undiagnosed cases showed an autumn peak. There were also early-winter epidemics of disease with little or no excretion of red blood or pus cells in the diagnostic faeces specimen. Some of the winter communicable disease among older children and adults appeared to be associated with signs of a temporary fat malabsorption in pre-school age cases. Undiagnosed cases in older children and adults were not related to the *E. coli* serotypes causing disease in infants during this period.

The statistical method applied increased the usefulness of these routine data. Although this series of laboratory records is now more than a decade old the results of the analysis can be compared with new observations as more is learned about the epidemiology of previously unrecognized pathogens, especially rotaviruses.

INTRODUCTION

A retrospective analysis of a long series of diagnostic results from a microbiology laboratory has been used to study causes of infectious gastroenteritis in patients referred by general practitioners (GPs) over a 16-year period up to 1968 (Thomas & Tillett, 1975). This analysis was made in collaboration with Dr Mair Thomas, director of the laboratory. An attempt has been made to assess the values and shortcomings of these routine data for monitoring disease experience in the area served by the laboratory (Tillett & Thomas, 1981). GPs were referring patients for laboratory diagnosis at their own discretion and were not taking part in any special survey, and thus these laboratory cases were a selected group. It was found

0022-1724/81/0042-1980 $01.00 © 1981 Cambridge University Press

https://doi.org/10.1017/S0022172400068741 Published online by Cambridge University Press

that the GPs had been referring greater proportions of their patients with relevant symptoms at those times when dysentery due to *Shigella sonnei* was causing outbreaks in local schools. Apart from this it appeared that GPs remained collectively consistent in the proportion of cases selected for laboratory investigation.

A total of 20273 index cases (first case in a household) were investigated by the laboratory during the 16 years 1953–68. There were 1868 cases of Sonne dysentery and 475 of salmonellosis. Among children under five years of age 232 cases yielded enteropathogenic *Escherichia coli*. Overall, 288 cases of *Giardia lamblia* were diagnosed by light microscopy and 122 cases yielded other pathogens. There were 1365 cases with signs of fat malabsorption typical of a disease we have called 'fatty diarrhoea', a condition which has been shown to be infectious (Thomas, 1952). There were 143 double and 4 triple infections leaving 16076 (79%) cases for which the laboratory could make no diagnosis. This long series of 16076 undiagnosed cases seen in 64 quarter years from 1953 to 1968, ranging from 146 to 455, gave a unique opportunity to study disease patterns of gastroenteritis of then unknown aetiology. The considerable fluctuation in cases suggests that most are of infective origin. It became obvious that any analysis of disease pattern must take into account the bias in the data due to the changing of referral habits by GPs during outbreaks of dysentery.

This paper describes how the statistical method of multiple regression analysis enables the underlying patterns of disease in routine data to be identified where there is known bias, and to derive a mathematical model describing the relationships.

STATISTICAL ANALYSIS

Preliminary analysis

Fig. 1 shows a plot of quarterly undiagnosed cases against quarterly cases diagnosed as Sonne dysentery. There is a strong positive correlation. It has already been shown that the laboratory was making very few false negative diagnoses of dysentery (Tillett & Thomas, 1974) and therefore this correlation is mainly due to GPs referring more of their diarrhoeal patients when dysentery was known to be in the area. The correlation between undiagnosed and Sonne dysentery cases was examined by performing linear and quadratic regression with quarterly undiagnosed cases as the dependent variable (Tillett, 1977). Linear correlation was highly significant, but the addition of a quadratic term improved the model by a small but significant amount, due mainly to one or two observations with exceptionally high numbers of dysentery cases. The residuals from this quadratic model were observed to have an upward trend when plotted against time (Fig. 2), indicating that average quarterly cases had increased by about 100 by the end of the study. Therefore a time trend was allowed for in the main analysis.

Because Sonne dysentery was seasonal, with most cases occurring in the first quarter, the same regression analysis was repeated four times for sets of 16 observations from the first, second, third and fourth quarters of the year. In each analysis the slope of the regression lines – and therefore the relationship between

https://doi.org/10.1017/S0022172400068741 Published online by Cambridge University Press

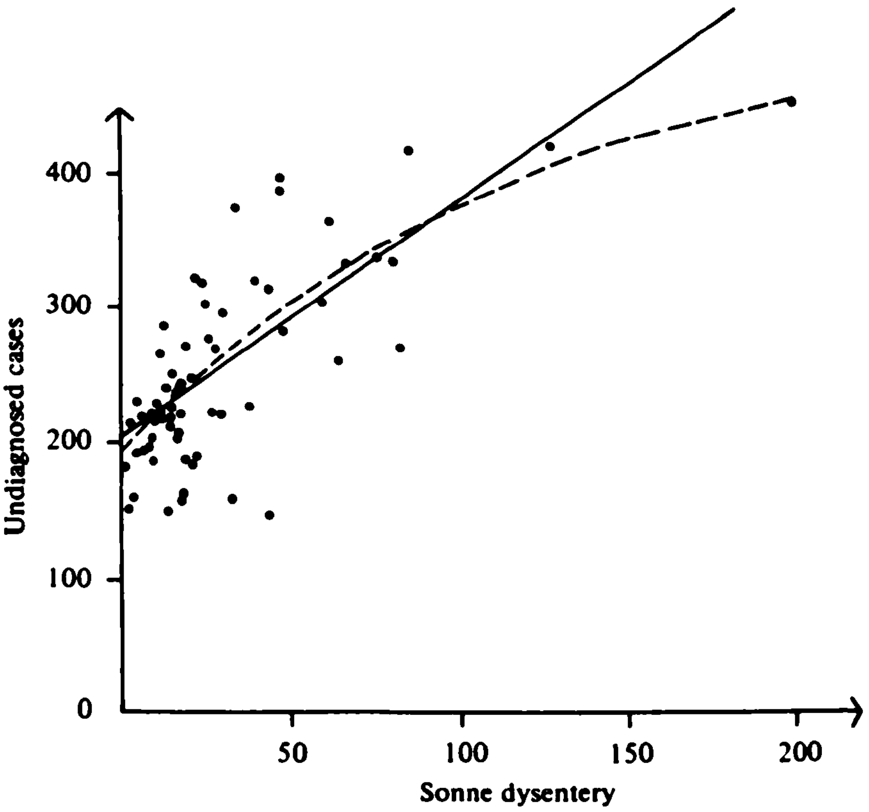

Fig. 1. Numbers of Sonne dysentery and undiagnosed index cases in 64 quarter years, with linear and quadratic regression lines.

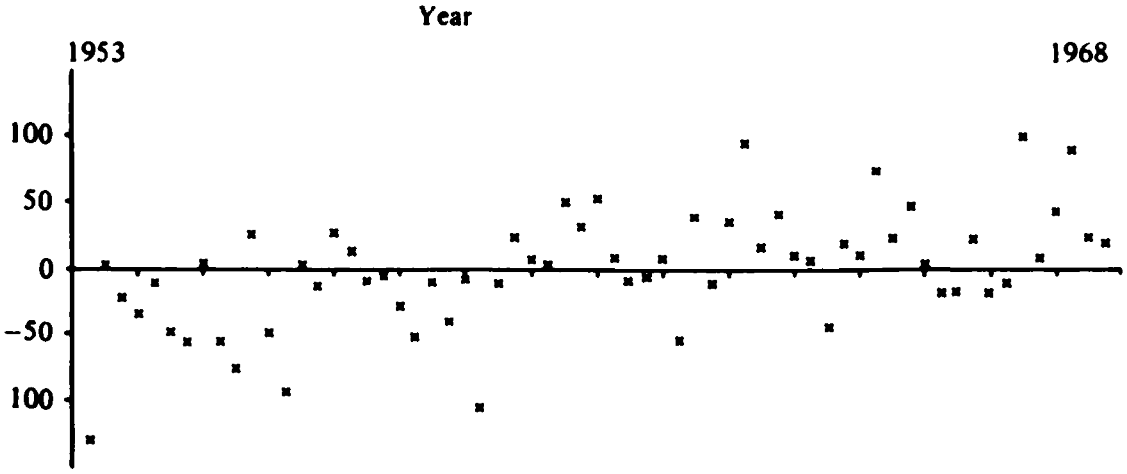

Fig. 2. Residual undiagnosed cases after fitting quadratic regression on quarterly Sonne dysentery cases, plotted against time.

undiagnosed and Sonne dysentery cases – was similar, but the intercepts were significantly different, suggesting an independent seasonal element in the fluctuation of undiagnosed cases. Therefore, variables to describe this seasonal pattern were introduced into the main analysis.

Multiple regression analysis

Stepwise regression analysis (Draper & Smith, 1966) was performed using the BMD 02R package program (Dixon, 1973). This analysis looks for linear relationships between y, the 'dependent' variable, here quarterly numbers of undiagnosed cases, and a set of 'regressor' variables x_1, x_2,... x_k. From these a subset of re-

https://doi.org/10.1017/S0022172400068741 Published online by Cambridge University Press

gressor variables x_1, $x_2, \ldots x_p$ is selected which are associated with systematic variation in y. In addition, there is assumed to be a random component in the fluctuation of y and thus the model can be described as:

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_p + \epsilon,$$

where ϵ is the random component, and x_0 is a dummy variable of value one. The βs are estimated by least squares. If the chosen regressor variables give the correct model (i.e. account for all the systematic variation in y) then the residual mean square will estimate the random component. The least squares method of fitting the model minimizes the error and provides unbiased estimates of the coefficients in the model. If the random component is normally distributed then the estimates of the coefficients also have maximum likelihood, and parametric tests can be used to evaluate the significance of partitioned sums of squares.

Forward selection is made of regressor variables, one at a time. At each step the one selected is that which, by its inclusion, would decrease the residual mean square by the greatest amount, provided that this decrease would be significant at the chosen level. After each step a check is made that every variable in the model is still maintaining its significance; if not it is removed from the equation. As a modification, the BMD program allows variables to be forced into the model regardless of their significance.

The regressor variables offered for selection fell into four groups. The first group describes time patterns:

x_1 = year (1953 set at $-7{\cdot}5$, 1968 at $+7{\cdot}5$);
x_2 = 1 for observations in the 1st quarter (Jan.–Mar.), 0 otherwise;
x_3 = 1 for observations in the 2nd quarter (Apr.–Jun.), 0 otherwise;
x_4 = 1 for observations in the 3rd quarter (Jul.–Sept.), 0 otherwise.

Observations from the 4th quarter (Oct.–Dec.) are uniquely defined by

$$x_1 = x_2 = x_3 = 0$$

The second group describes numbers of cases with positive diagnoses in that quarter:

x_5 = cases of Sonne dysentery,
x_6 = cases of salmonellosis,
x_7 = cases of *E. coli* infection,
x_8 = cases of giardiasis,
x_9 = cases of 'fatty diarrhoea',
x_{10} = square of x_5.

Other pathogens were recognized too infrequently to be included as regressor variables.

The third group describes severity of disease in terms of proportions of undiagnosed cases with or without blood or pus cells observed in the faeces specimen:

x_{11} = proportion excreting red blood cells (with or without pus),
x_{12} = proportion excreting pus cells (with or without red cells),
x_{13} = proportion not excreting red blood or pus cells.

The fourth group of variables was made up of meteorological variables, i.e. devia-

https://doi.org/10.1017/S0022172400068741 Published online by Cambridge University Press

Table 1. *Regression equations*

Age group analysed (total cases) (range of quarterly undiagnosed cases)	Coefficients of variables in every model						Coefficients of other selected variables			
	Constant	x_1 year	x_2 1st quarter	x_3 2nd quarter	x_4 3rd quarter	x_5 Sonne dysentery	x_9 Fatty diarrhoea	x_{10} Square of x_5	x_{11} or x_{12} Proportion blood/pus cells	R^2
Total (20273) (146–455)	198	+7·1	−31	+12	+22	+2·1	+1·3	−0·004	$-324x_{11}$	0·76
0–4 years (5653) (36–107)	65	+0·9	−0·55	+2·9	+5·0	+1·3	.	.	$-157x_{11}$	0·56
5–9 years (3406) (10–92)	30	+1·2	−2·6	−3·8	−11	+0·8	+2·0	.	.	0·66
15–39 years (4756) (32–122)	60	+1·3	−6·9	+8·9	+12	+0·59*	.	−0·002*	$-101x_{11}$*	0·59
40+ years (4713) (30–128)	52	+2·3	−9·8	+11	+15	+0·54*	+0·51	−0·002*	$-83x_{12}$*	0·59

* Variable relates to numbers of cases in the total group rather than in the age group being analysed.

Table 2. *Details of variables selected into regression equations, excluding those describing time, and their levels of significance expressed as probabilities of not being associated with fluctuations in undiagnosed cases*

	Age group analysed				
Variable	Total	0–4 years	5–9 years	15–39 years	40+ years
x_8 Number of Sonne dysentery cases	$\ll 0{\cdot}001$	$\ll 0{\cdot}001$	$\ll 0{\cdot}001$	$\ll 0{\cdot}001$	$< 0{\cdot}001$
x_9 Number of fatty diarrhoea cases	$< 0{\cdot}1$	—	$< 0{\cdot}05$	—	$< 0{\cdot}1$
x_{10} Square of Sonne dysentery cases	$< 0{\cdot}05$	—	—	$< 0{\cdot}1$	$< 0{\cdot}05$
x_{11} Proportion of undiagnosed cases excreting blood cells	$< 0{\cdot}05$	$< 0{\cdot}001$	—	$< 0{\cdot}05$	—
x_{12} Proportion of undiagnosed cases excreting pus cells	—	—	—	—	$< 0{\cdot}1$

tions from normal of temperature, rainfall, sunshine and air pollution. None of these was ever significant and so none ever appeared in the models.

Variables x_2, x_3 and x_4, dummy variables to identify the quarters, were forced into every equation because the preliminary analysis had shown that an independent seasonal element was present in undiagnosed cases. Allowing the models to have a seasonal structure prevented the incorrect selection of other regressor variables which were only indirectly correlated with undiagnosed cases because of their own seasonal pattern.

The selection of other regressor variables was made if their inclusion in the model accounted for a significant amount of variation in y independently of the influence of any other regressor variable. The level of significance was deliberately chosen to be low, only 10%, to allow the inclusion of variables with only moderate correlation and thus allow for broad discussion of possible associations.

Analysis was made using the total 20273 index cases and then for those four age groups considered to have large enough numbers for multivariate analysis. The age group not analysed separately was the 10–14-year-old (total of 1342 cases). The regression models achieved for total cases and for the four age groups analysed are shown in Table 1, together with the R^2 value, defined as the square of the multiple correlation coefficient, which measures the proportion of variation in quarterly undiagnosed cases explained by the regression equation. Details of the variables included in the models, apart from those describing time patterns, and their degree of significance are shown in Table 2. When the four age groups were analysed the regressor variables offered for selection included those relating to total cases as well as those relating to the age group being analysed. This was particularly useful for the adult age groups within which many fewer cases of Sonne dysentery were being diagnosed, and where numbers from total cases will have reflected changes in prevalence more accurately.

The quarterly seasonal structure was included in each model and described the

https://doi.org/10.1017/S0022172400068741 Published online by Cambridge University Press

numbers of referred undiagnosed cases which would have been observed quarterly independently of the influence of other regressor variables included in the model. In four models the highest number of undiagnosed cases was shown to occur in the third quarter, and in one model, that of the 5–9 year-olds, in the fourth quarter. For each model the most highly significant variable was x_5, cases of Sonne dysentery, which was first to be selected. The second variable to be selected in each model was x_1, which described time trend. The variable x_{10}, the square of the number of Sonne dysentery cases, was selected with a negative coefficient for the total-cases model and for both adult age groups.' This implied a levelling-off of undiagnosed cases once Sonne dysentery cases had reached a high level. The only other variable from the second group of regressor variables – frequencies of particular diagnoses – to appear in the models was the number of 'fatty diarrhoea' cases. This variable was selected for total cases, for 5–9 year-olds and for one adult age group. It only just failed to be selected into the other adult age group, but for 0–4 year-olds the partial correlation was close to zero.

The group of regressor variables describing proportions of undiagnosed cases excreting red blood or pus cells ($x_{11}x_{12}x_{13}$) was highly intercorrelated and once one had been selected into a model the partial correlations of the other two became very small. It made very little difference to the final R^2 value which one of the three was selected. In each of the four models where one of these variables was selected it had a negative coefficient, and the implication is that times of increased incidence of undiagnosed cases coincided with relatively fewer cases excreting blood or pus cells. Most patients excreting cells excreted both blood and pus. Out of a total of 16076 undiagnosed cases 1236 were observed with blood and pus, 247 blood only and 565 pus only.

Examination of residuals and extreme values

Residuals are defined as the difference between the number of undiagnosed cases observed in a quarter and the number predicted by the regression model. If the selected models fulfil the theoretical assumptions necessary for the significance tests that have been made, then the residuals should be normal random variables. Tests were applied to the residuals and no significant departures from normality and serial randomness were found (Tillett, 1977).

Fig. 1 shows that the 64 quarterly observations included one exceptionally high value of x_5, cases of Sonne dysentery, which was 199 whereas the median value was 20. Outlying points can have undue influence on regression analyses and therefore the analysis of total cases was repeated omitting this one observation. A similar model was achieved with small, insignificant changes in the regression parameters. The only difference was, as might be expected, that the quadratic variable, the square of Sonne dysentery cases, was no longer significant.

None of the five models explains more than 76% of the variation in quarterly undiagnosed cases (Table 1). This implies that there was considerable fluctuation in residuals, although no serial correlation was observed. Part of this must be due to sampling error, but part might be due to the constantly changing prevalence of numerous infectious agents, and it is of interest to see whether these

https://doi.org/10.1017/S0022172400068741 Published online by Cambridge University Press

Table 3. *Correlations between residuals from the four age group models*

Age group of regression model	5–9 years	15–39 years	40+ years
0–4 years	0·40	0·32	0·26
5–9 years	—	0·51	0·18
15–39 years	—	—	0·26

residual fluctuations were occurring at the same time in each age group. Simple correlation coefficients between the residuals from the four age groups analysed are shown in Table 3. The values are not high, especially for the older adults. The highest coefficient is between children within the primary school age range (5–9 years) and young adults (15–39 years).

DISCUSSION

More than 20000 index cases of diarrhoea were investigated at one Public Health Laboratory during the 16-year period ending in 1968, but in only one-fifth was an infective cause recognized. (A higher diagnostic rate might now be expected as the range of pathogens identified has been extended by technical advances.) There were 49 cases with non-infective diagnoses such as cancer or diverticulitis but the remaining 16076 cases remained undiagnosed. Statistical analysis has been used to examine the fluctuations with time of these undiagnosed cases, and to construct mathematical models to describe this fluctuation.

Multiple regression analysis was used to study quarterly undiagnosed cases and to look for linear associations with another set of variables, called regressor variables. If a strong association is observed with a particular regressor variable then that variable is selected into the model. But the association is calculated as the correlation between undiagnosed cases and this regressor variable, independently of the influence of other regressor variables already selected. This is particularly useful because it has already been shown that there is a bias in the way in which cases were selected for laboratory investigation: at times of Sonne dysentery epidemics the GPs referred greater proportions of their symptomatic patients, thus artificially increasing the numbers with non-Sonne diagnoses when there was no increase in incidence. These regression models have a variable – quarterly numbers of Sonne dysentery cases – to describe this bias. Other regressor variables selected into the models are correlated with fluctuation in undiagnosed cases independently of the bias introduced by dysentery epidemics. This is achieved because the method of analysis looks for correlations which would have existed under the imaginary conditions of having zero Sonne dysentery cases throughout the period. Thus it is hoped that the other regressor variables selected into the model describe factors directly associated with the changing incidence of undiagnosed cases.

Regression models were achieved for total cases and for four large age groups. In all models the correlation between undiagnosed cases and Sonne dysentery was large. A quadratic term in the models for total cases and the two adult age groups implied that the response of GPs to increases in Sonne dysentery incidence – that

https://doi.org/10.1017/S0022172400068741 Published online by Cambridge University Press

of referring more cases, including many in the undiagnosed group – tended to level off once dysentery incidence increased above a certain point. This indicates that, at the few times of exceptionally high incidence of Sonne dysentery, the GPs, perhaps because of workload, became more selective in referring patients for laboratory investigation. They investigated children with diarrhoea since children are more liable to Sonne infection than adults and were liable to spread the disease if allowed back to school. But they became more selective for the adult population, where the chance of Sonne infection and the public health implications were less.

The only other regressor variable to be selected from the group relating to particular diagnoses was the condition described as 'fatty diarrhoea', which was selected into the models for total cases, for the 5–9 age group and the 40+ age group. It only just failed to be selected into the 15–39 age group model. The diagnosis of 'fatty diarrhoea' was made when typical fat globules were seen by light microscopy (Thomas, 1952). The results of the regression analyses indicate that all 'fatty diarrhoea' cases in pre-school children (0–4 age group) were being correctly classified but that in older age groups some disease was arising from the same cause without displaying signs of fat malabsorption in the diagnostic specimen, and therefore these patients were put into the undiagnosed category. Overall, the model indicates that every 'fatty diarrhoea' case that was diagnosed was associated with an increase of 1·3 undiagnosed cases. There were 1365 'fatty diarrhoea' diagnoses and therefore the agent(s) of this disease probably accounted for about $1{\cdot}3 \times 1365 = 1775$ of the total 16076 undiagnosed cases or $1365 + 1775 = 15\%$ of the total 20273 index cases. Most cases were observed in the first quarter of the year, and sometimes there was a rise in the fourth quarter. It will be of interest to see whether this syndrome is related to one or more of the newly recognized, enteritis-associated viruses now being identified by electron microscopy and immunological methods. It is of interest that during the investigation of a recent outbreak of diarrhoea a comparison was made of observations by light and electron microscopy (EM). No bacterial or parasitic pathogens had been found. Rotavirus was seen by EM in 13 faeces specimens and fat globules or derivatives in 11 of these. A control group is being studied (Thomas, personal communication). The number of cases of diarrhoea attributed to rotavirus infection in reports to the Communicable Disease Surveillance Centre (CDSC) is now much larger than that for any other virus. Numbers have been highest in the first and second quarters of the year and lowest in the third. All ages have been affected, but most reports concern young children. Rotaviruses are now thought to cause a significant proportion of all acute gastroenteritis in young children in this country (Flewett, 1976).

Variables related to salmonellosis and *E. coli* diagnoses were not selected into any model, indicating that they were not associated with increases in undiagnosed cases nor did their prevalence affect the referral habits of GPs. Medical Officers of Health rather than GPs would have investigated incidents of salmonella food poisoning. *E. coli* were routinely sought only in pre-school age children, but had the serotypes circulating among the young also caused disease in older people then this variable would have correlated with undiagnosed cases and would have been selected into the models.

In four of the models increases in undiagnosed cases were associated with in-

https://doi.org/10.1017/S0022172400068741 Published online by Cambridge University Press

creases in the proportion of cases without blood or pus cells in the diagnostic faeces specimen. This suggests small epidemics associated with agents which caused little damage to the colon. This association was most strong in the preschool age group, and the rises in incidence were most frequent in the fourth quarter.

There was an upward trend in undiagnosed cases over the 16-year period, demonstrated by the inclusion of the year variable in every model. It is not possible to tell retrospectively whether this gradual increase was due to increased incidence or a change in the proportions of diarrhoeal patients having a laboratory investigation. The general seasonal pattern among the remaining undiagnosed cases showed fewest cases in the first quarter of the year and highest numbers in the third quarter.

Different age groups were compared to see whether there was correlation between the unexplained (residual) fluctuations. The residual fluctuations must have included purely random variation, but all comparisons between age groups showed some positive correlation indicating short-term outbreaks. Correlations were not particularly high, which might indicate that the different agents responsible were more prevalent among certain age groups; (we were analysing only index cases: the first case in each household). The strongest correlation was between school children and adults of 15–39, the adult group more likely to have contact with school children.

Since the period of this study several bacteria have been recognized as significant enteric pathogens (Hamilton, 1980). Of these campylobacter infections are being reported in the greatest numbers to CDSC. Reported cases have increased steadily over the past three years as more laboratories have sought them and are now reported as frequently as salmonellosis. As yet no seasonal pattern or epidemic period has emerged. Campylobacter enteritis is reported as being often associated with cellular stools (Karmali & Fleming, 1979). This agent seems an unlikely candidate for the autumn or non-cellular epidemics observed in Enfield.

These routine data were found suitable for multiple regression analysis after careful preliminary inspection of the data. The stepwise method was used, but with a seasonal structure imposed on the models to avoid selection of regressor variables with a seasonal pattern coincidental to that of undiagnosed cases. The effect of a quantifiable bias due to referral habits during dysentery epidemics in these data was removed. A check was made that an outlying observation had not distorted the analysis and the main analysis was repeated using a more recent method of fitting models to data, as described by Daniel & Wood (1971), but no advantage was found in this method over stepwise regression with these data (Tillett, 1977). Stepwise regression has also proved helpful in the study of influenza epidemics in England and Wales (Clifford *et al.* 1977; Tillett, Smith & Clifford 1980).

https://doi.org/10.1017/S0022172400068741 Published online by Cambridge University Press

REFERENCES

CLIFFORD, R. E., SMITH, J. W. G., TILLETT, H. E. & WHERRY, P. J. (1977). Excess mortality associated with influenza in England and Wales. *International Journal of Epidemiology* **6**, 115.

DANIEL, C. & WOOD, F. S. (1971). *Fitting Equations to Data*. New York: Wiley-Interscience.

DIXON, W. J. (1973). *Biomedical Computer Programs*. Berkeley: University of California Press.

DRAPER, N. R. & SMITH, H. (1966). *Applied Regression Analysis*. New York: Wiley-Interscience.

FLEWETT, T. H. (1976). Implications of recent virological researches. *CIBA Foundation Symposium* **42**, 237.

HAMILTON, J. R. (1980). Infectious diarrhoea: clinical implications of recent research. *Canadian Medical Association Journal* **122**, 29.

KARMALI, M. A. & FLEMING, P. C. (1979). Campylobacter enteritis. *Canadian Medical Association Journal* **120**, 1525.

THOMAS, M. E. M. (1952). Epidemic abdominal colic associated with steatorrhoea. *British Medical Journal* i, 691.

THOMAS, M. E. M. & TILLETT, H. E. (1975). Diarrhoea in general practice: a sixteen-year report of investigations in a microbiology laboratory, with epidemiological assessment *Journal of Hygiene* **74**, 183.

TILLETT, H. E. (1977). Ph.D. thesis. University of London.

TILLETT, H. E., SMITH, J. W. G. & CLIFFORD, R. E. (1980). Excess morbidity and mortality associated with influenza in England and Wales. *Lancet* i, 793.

TILLETT, H. E. & THOMAS, M. E. M. (1974). Culture of the faeces in the diagnosis of Sonne dysentery: a statistical method for estimating the true isolation rate. *International Journal of Epidemiology* **3**, 177.

TILLETT, H. E. & THOMAS, M. E. M. (1981). Monitoring infectious diseases using routine microbiology data. I. Study of gastroenteritis in an urban area. *Journal of Hygiene* **86**, 49.

https://doi.org/10.1017/S0022172400068741 Published online by Cambridge University Press