THE EGYPTIAN POPULATION AND FAMILY PLANNING REVIEW.
ISSR, CAIRO UNIV., VOL. 28, NO.1, 1994, PP. 1-18

# LOGISTIC REGRESSION FOR SOCIAL-ECONOMIC AND CULTURAL FACTORS AFFECTING DIARRHEA DISEASES IN CHILDREN UNDER TWO YEARS IN EGYPT

BY
SAMIR K. ASHOUR* AND MOHIY EL-DIN M. AHMED*.

## ABSTRACT

In this study, we will consider the problem of relating a qualitative dependent variable to one or more independent variables, which may or may not be qualitative. This problem has its multivariate analogous as well, when the dependent and independent variables are related by a logistic function, the model is often referred to as a logistic regression. A step-up (forward) selection procedure from the set of the independent variables, involving all the 63 variables in our data set, was carried out. The forward selection procedure begins by fitting the intercept. It then solves each of the models with only one independent variable. A Linear Logistic Regression (LLR) models was fitted for the response variable Y, is 0 for a child having diarrhea disease and 1 for a child having no diarrhea using the selected subset of variables obtained from the step-up procedure for Dakahlia and Sohag Govemates together based on 1020 observations, for Dakahlia Govemate based on 660 observations, for Sohag Govemate based on 330 observations.

## 1) INTRODUCTION

The problem presented here is concerned a large data set from Central Agency for Public Mobilization and Statistics (CAPMAS), Egypt, in which attempts are made to study which subset of a large set of socio-economic and cultural factors, as the independent variables, predicts diarrhea disease, as the dependent variable, in children under two years in Egypt. A child's health depends on several interacting factors, some of which are related to demographic characteristics, environmental conditions, income level and others are related to knowledge, attitude, and practices. These interacting factors might have a direct or indirect effect on the care of children having diarrhea disease, which is one of the major causes of morbidity and mortality among infants and children, particularly in developing countries. It is one of the big three killers of children in the tropics, along with protein energy malnutrition and respiratory tract infections (see for example, Sen et al. (1985), Giugliano et al. (1985), Saran and Gaur (1981)).

Diarrhea diseases in Egypt particularly afflict infants especially in rural areas. It is a common pediatric and community health problem. The predisposing factors for diarrheoa, may be artificial feeding, low income, poorly educated mothers, poor housing, big family size and lack of adequate medical care. Environmental factors, poor sanitation with lack of clean plentiful water supply. Unsanitary waste disposal and high prevalence of flies are important in precipitating infection.

The fieldwork of the (NCDDP) research was carried out in two govemates, Dakahlia govemate (representing lower Egypt) and Sohag govemate (representing Upper Egypt). The choice of the two govemates was based on the results of projects entitled "Establishment and Implementation of an Internal Evaluation system for the National Campaign of Diarrhea Diseases control". The primary results of the project showed that the Dakahlia govemate has a lower incidence rate of diarrhea (Rural 41.27% and Urban 36.66%) than the Sohag govemate (Rural 55.6% and Urban 47.22%), so both Dakahlia and Sohag govemates were chosen to be the study area to conduct the household survey.

A sub sample was chosen for the study from the sample of the original project previously mentioned. In Dakahlia out of twenty eight clusters, eleven clusters (3 Urban and 8 Rural) were chosen at random. In Sohag govemate, out of fourteen clusters, six clusters (2 Urban and 4 Rural) were selected at random. That makes a total of seventeen clusters, each of which consists of sixteen households. Thus the total sample consisted of 1020 households or mothers. Table (1) displays the number of clusters, as well as, the number of children by govemates.

* Institute of Statistical Studies and Research, Cairo University.

**Table (1): Number of Children Surveyed and Clusters by Governates**

| Governate | No. of Clusters | | Total | No. of Children | | Total |
|---|---|---|---|---|---|---|
| | Urban | Rural | | Urban | Rural | |
| Dakahlia | 3 | 8 | 8 | 180 | 480 | 660 |
| Sohag | 2 | 4 | 6 | 120 | 240 | 360 |
| Total | 5 | 12 | 17 | 300 | 720 | 1020 |

Population census data were obtained from central Agency For Public Mobilization and Statistics (CAPMAS) for the two governates under study for (1984) at the sheeakha and village level. The number of clusters to be chosen from each governate was determined based upon the area's population as a proportion of the population living in the two study governates, and the number of clusters chosen from rural against urban areas within the proportion of the population living in rural against urban districts. Population lists at the sheeakha and village level were constructed separately for urban and rural areas within each governate. The procedure used to select clusters, was simple random sampling without replacement.

## 2) Logistic Regression Analysis

Logistic regression analysis provide an approach of handling the binary data by modeling the association between the response variable and a set of explanatory variables. Relating qualitative variable to other variables through a logistic distribution functional form is logistic regression. Logistic regression have been widely used for the analysis of binary response data with continuous or categorical explanatory variables. Parameter estimates are usually obtained through direct maximum likelihood estimation.

Now let $Y_i$ is a Bernoulli random variable and its probability of success conditional on the levels of explanatory variables will be denoted as

$$p_i = pr\{y_i = 1 \mid X_1, X_2, \dots, X_n\}. \qquad (1)$$

where $\underset{\sim}{X}$ is the (n by p) data matrix of independent variables and, $y_i$ is the observed value of the outcome variable $Y$ which is either 1 or 0.

The logistic model equates the logarithm of the conditional odds of success to failure to a linear function of the explanatory variables. Thus it proposes that

$$\text{Logit}(p_i) = \log(\frac{p_i}{1-p_i}) = \lambda\text{, say} \qquad (2)$$

from which we get $\lambda = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$. This implies that

$$p_i = \frac{\exp\{\beta_0 + \beta' X_i\}}{1+\exp\{\beta_0 + \beta' X_i\}} = \frac{1}{1+\exp\{-(\beta_0 + \beta' X_i)\}}, \; i = 1, 2, \dots, n \qquad (3)$$

which depends on $\beta_0$ and $\underset{\sim}{\beta}$. Thus, we have to estimate $\beta_0$ and $\underset{\sim}{\beta}$, taking the model as

$$\lambda_i = \beta_0 + \beta' X_i \;, i = 1, \dots, n \qquad (4)$$

The logistic regression model is an extension of the general linear model to the case involving a dichotomous variable $Y$. (It should be noted that we could not take P itself as a linear function of the explanatory variables as it was restricted to the interval (0,1), however, the logistic transformation removed that restriction, since $\lambda$ can take values in $(-\infty, \infty)$). After assuming $logit(p) = \underset{\sim}{\beta_0} + \beta' \underset{\sim}{X_i}$, the next step would be estimate the unknown parameters β. Maximum likelihood estimates can be obtained by iterative procedures.

Hartz and Rosenberg (1975) describe an iterative computing technique for obtaining maximum likelihood estimates of multiple logistic regression coefficients when analyzing categorical data. The methodology is based upon the marginal iterative proportional fitting method of Deeming and Stephan (1940), Studied by Ireland and Kullback (1968), among others. This iterative procedure is commonly applied in the context of log-linear analysis and thus draws heavily on the work of Bishop (1969), Goodman (1970), and others. Kleinbaum et al.(1982) describe and illustrate techniques for both unconditional and conditional maximum likelihood estimation of the parameters in the logistic model.

Fortunately, there are computer packages such as BMDP (program LR), reported by Cox (1987) and its applications indicated by Greenland (1985), GLIM programs provided by Baker and Nelder (1979), and NCSS by Hintze (1992), which provide maximum likelihood estimates of the logistic regression model.

Once we have fitted a particular multiple logistic regression model, we begin the process of assessment of the model. The first step in this process is usually assessing the significance of the variables in the model using the likelihood ratio test for overall significance of the p coefficients for the independent variables in the model. Under the null hypothesis that p "slope" coefficients for the covariates in the model are equal to zero, the distribution of G will be chi-square with p degrees of freedom. Rejection of the null hypothesis in this case has an interpretation analogous to that in multiple linear regression; we may conclude that at least one, and perhaps all p coefficients are different from zero.

## 3) Results of Logistic Regression

A step-up (forward) selection procedure from the set of the independent variables, involving all the 63 variables in our data set, was carried out. The forward selection procedure begins by fitting the intercept. It then solves each of the models with only one independent variable. The variable having the maximum $\chi^2$ is added. Next, each of the remaining variables is considered and the variable having the largest $\chi^2$ value is added to the model. This procedure continues until none of the remaining variables have a $\chi^2$ value greater than the tabulated $\chi_0^2 = \chi^2(1,0.05) = 3.84$.

A $\chi^2$ value is used to test the hypothesis of all of the $\beta$'s (all coefficients except the intercept term) are equal to zero. The number of degrees of freedom of the corresponding chi-square statistic is equal to the number of variable coefficients estimated. The $\chi^2$ is calculated as two times the difference between the log likelihood of the current model and the log likelihood of the intercept term model. The calculated $\chi^2$ value will be termed the model chi-square.

$\chi^2 = 2(\mathit{log}$ likelihood of the current model - $\mathit{log}$ likelihood of the intercept only model)

which is the likelihood ratio $\Lambda$.

For the hypothesis that a single component $\beta_i$ of $\beta$ is zero, an alternative test compares $\hat{\beta}_i$ with its asymptotic standard error $V_{ii}^{1/2}$ and then treats $t_i = \frac{\hat{\beta}}{\hat{V}^{\frac{1}{ii}}}$ as approximately $N(0,1)$ under the hypothesis

For large samples this test will be closely equivalent to the likelihood ratio test, with $t_i^2 \approx \chi^2$.

The classification table displays the results of classifying the data based on the logistic regression equation compared to the actual allocation. Considering a child to be having diarrhea or having no diarrhea is based on the predicted probability of having no diarrhea ($prob(y_i = 1)$). A probability near to one suggests strongly that this child is more likely to have not diarrhea. In practice, predicted probabilities below certain classification cut-off point value (normally 0.5) are classified as having no diarrhea. The simple method (based on consideration of the correct classification rates p(0|0), probability of predicted a child have diarrhea while he is actually having diarrhea, and p(1|1), probability of predicted a child have no diarrhea while he is actually having no diarrhea) of comparing classification matrices will be used in making comparisons between different proposed models.

## 3.1) Logistic Regression for Dakahlia and Sohag Governates

A Linear Logistic Regression (LLR) model was fitted for the response variable Y, is 0 for a child having diarrhea disease and 1 for a child having no diarrhea using the selected subset of variables obtained from the step-up procedure for Dakahlia and Sohag Governates together based on 1020 observations. Twelve variables out of the sixty three independent variables (the $X_1$ "governate", $X_2$ "type of residence" (urban or rural), $X_7$ "age of mother", $X_{12}$ "number of children in a family", $X_{20}$ "garbage disposal", $X_{23}$ "own land", $X_{31}$ "watery stool", $X_{32}$ "change of stool texture", $X_{34}$ "increase number of defecation", $X_{35}$ "thirsty", $X_{41}$ "dentition" and $X_{54}$ "special fluid") were chosen by the step-up procedure in the LLR model. Four of these variables are demographic variables ($X_1$ "governate", $X_2$ "type of residence", $X_7$ "age of mother" and $X_{12}$ "number of children in a family"), one variable is environmental condition ($X_{20}$ "garbage disposal"), one variable is economic condition ($X_{23}$ "own land") and four variables are knowledge of mother the signs and symptoms of diarrhea disease ($X_{31}$ "watery stool", $X_{32}$ "change of stool texture", $X_{34}$ "increase number of defecation", $X_{35}$ "thirsty"), one variable is knowledge of mother of the causes of diarrhea ($X_{41}$ "dentition") and one variable is type of treatment ($X_{54}$ "special fluid").

The $\chi^2$ for testing the equality of all the β's to zero show that the model is significantly different from zero (The $\chi^2$ value and its significance are given in table (2)).

**Table (2): Variable Selection for Logistic Regression of Dakahlia and Sohag Governates**

| Variable | Beta Estimate | Standard Error | Chi-square | Prob. | R-Square |
|---|---|---|---|---|---|
| Intercept | -2.626583 | 0.937963 | 7.84 | 0.0051 | 0.0077 |
| $X_1$ | 0.3122587 | 0.149329 | 4.37 | 0.0365 | 0.0043 |
| $X_2$ | -.4805794 | 0.158395 | 9.21 | 0.0024 | 0.0091 |
| $X_7$ | -.4367442 | 0.170620 | 6.55 | 0.0105 | 0.0065 |
| $X_{12}$ | 0.2960627 | 0.169752 | 3.04 | 0.0811 | 0.0030 |
| $X_{20}$ | 0.3070792 | 0.081151 | 14.32 | 0.0002 | 0.0140 |
| $X_{23}$ | -.760263 | 0.157972 | 23.16 | 0.0000 | 0.0225 |
| $X_{31}$ | 0.7573459 | 0.146607 | 26.69 | 0.0000 | 0.0258 |
| $X_{32}$ | -.424761 | 0.221965 | 3.66 | 0.0557 | 0.0036 |
| $X_{34}$ | 0.2971732 | 0.148375 | 4.01 | 0.0452 | 0.0040 |
| $X_{35}$ | 1.423958 | 0.791995 | 3.23 | 0.0722 | 0.0032 |
| $X_{41}$ | 0.4136936 | 0.165141 | 6.28 | 0.0122 | 0.0062 |
| $X_{54}$ | 0.5353867 | 0.317986 | 2.83 | 0.0922 | 0.0028 |

In table (2) we present, for each variable listed in the first column, the following information: (1) the estimated slope coefficient for the logistic regression model, (2) the estimated standard error of the estimated slope coefficient, (3) the likelihood ratio test statistic, G, for the hypothesis that the slope coefficient is zero. This is obtained as minus twice the difference between the log likelihoods for the constant only model given in the first row and the model containing the respective variable. Under the null hypothesis this quantity will follow the chi-square distribution with 1 degree of freedom.

The positive and negative values of the estimated logistic regression coefficient beta values are important in deciding which variables are significant for predicting a low probability of no diarrhea (Y= 1) and which for predicting a higher probability of no diarrhea. The estimator linear logistic regression, $\hat{p}(Y=1|X)=\hat{\pi}(X)$, for both Dakahlia and Sohag governates based on 1020 observations is:

$$p(y=1|X)=\frac{1}{1+\exp(-(-2.6265+0.3122X_1-0.4805X_2-0.4367X_7+0.296X_{12}}$$

$$\frac{1}{+0.3071X_{20}-0.7602X_{23}+0.7573X_{31}-0.4247X_{32}+0.2972X_{34}+1.4239X_{35}}$$

$$\frac{1}{+0.4137X_{41}+0.5354X_{54}))}.$$

Then the estimated logit, $\hat{g}(x)$, of the multiple logistic regression model is

$$\hat{g}(X)=-2.6265+0.3122X_1-0.4805X_2-0.4367X_7+0.296X_{12}+0.3071X_{20}$$
$$-0.7602X_{23}+0.7573X_{31}-0.4247X_{32}+0.2972X_{34}+1.4239X_{35}$$

$$+0.4137X_{41}+0.5354X_{54}.$$

Corresponding to the logistic regression model given above, a positive beta estimate is an indication of a decrease in the probability of no diarrhea as the variable score moves from 0 to 1, therefore positive coefficients for governates $X_1$, number of children in a family $X_{12}$, garbage disposal $X_{20}$, mother's knowledge the signs and symptoms of diarrhea disease $X_{31}$, $X_{34}$ and $X_{35}$, mother's knowledge the causes of diarrhea disease $X_{41}$, and type of treatment $X_{54}$ are associated with poor outcomes (Y=1). So, from table (2), it can be concluded that the probability of no diarrhea is low in children who are live in Sohag governate, who live with number of children less than two in a family, whose refuse disposal by thrown near house or in surface water (unhealthy methods), mothers who know signs and symptoms of diarrhea disease like watery stool and increase number of defecation and thirsty their children, mothers who know causes of diarrhea disease like dentition and who use special fluid as treatment for diarrhea disease.

Negative beta estimates for type of residence $X_2$, age of mother $X_7$, own land $X_{23}$ and knowledge of mother the signs of diarrhea disease as change of stool texture $X_{32}$ variables indicate that as the score for these variables move from 0 to 1, the probability of no diarrhea increases. So, mothers who live in rural, mothers whose age over 25 years, who owned land and who know signs and symptoms of diarrhea disease like change of stool texture their children are more likely to have children with no diarrhea.

This does not mean, however, that all of these variables are efficient and accurate in their prediction of no diarrhea under these conditions. The signs of the beta estimates are only an indication of which response of the independent variable is more favorable for prediction. By looking at the relative values or importance of the variables, it can be seen that some variables may be more significant than others. From table (2), results show that mother's knowledge, attitude and practices as regards diarrhea disease "watery stool" $X_{31}$, "dentition" $X_{41}$, "increase number of defecation" $X_{34}$, "change in stool texture" $X_{32}$, thirsty $X_{35}$ and "special fluid" $X_{54}$, economic condition "own land" $X_{23}$, environmental condition "refuse disposal" $X_{20}$ and demographic information "type of residence" $X_2$, "governates" $X_1$, "age of mother" $X_7$ and "number of children in a family" $X_{12}$ are useful in predicting no diarrhea, respectively. This observation is supported by the corresponding chi-square values and probability values. Probability levels given by the chi-squared test indicate the relative significance of the variables.

Although the most significant variables or the strongest predictors can be identified from the results of the logistic regression analysis and the chi-squared significance tests, no exact preference of ordering can be deduced. Identifying the most significant variable and ranking the others in order of importance can be done by using the 'Stepwise' method which discards the non significant variables and orders the remaining variables.

The reliability of this model for predicting levels is examined using "sensitivity" and "specificity" analysis. Clearly, we would like this model to be highly sensitive and specific, that is, as close to 1 as possible. To define the sensitivity and specificity in simple notation, let p(1|1) be the probability that predicted a child has no diarrhea when he actually has no diarrhea and let p(0|0) be the probability that predicted a child has diarrhea when he actually has diarrhea. Given non having diarrhea p(Y=1) or having diarrhea p(Y=0), the probability that the model will predict not having diarrhea or having diarrhea is characterized by the following formula:

Test sensitivity = p(1|1) and Test specificity = p(0|0).

**Table (3): Classification Matrix for Dakahlia and Sohag Governates**

| | Predicted | | |
|---|---|---|---|
| **Actual** | 0 | 1 | **Total** |
| 0 | 318 | 192 | 510 |
| 1 | 171 | 339 | 510 |
| **Total** | 489 | 531 | 1020 |

Probabilities of no diarrhea defined by $\text{Prob}(Y_i=1 \mid X\text{'s})$ were calculated and cross tabulated with the observed frequencies of non diarrhea and having diarrhea in table (3) at cut-off point=0.5, where $\text{Prob}(y=1 \mid X\text{'s})=0.6647$, $\text{Prob}(y=0 \mid X\text{'s})=0.6235$ and the percentage of all correctly classified predictions =64.41 %.

The comparison of observed to predicted values using the likelihood function is based on the following expression:

$$D = -2\ln\left[\frac{(\text{likelihood of the current model})}{(\text{likelihood of the saturated model})}\right].$$

The quantity inside the large brackets in the expression above is called the likelihood ratio. The statistic, D, is called the deviance by some authors [see for example, McCullagh and Nelder (1983)], and plays a central role in some approaches to assessment of goodness-of-fit. The deviance for logistic regression plays the same role as the residual sum of squares plays in linear regression. in fact, the deviance as shown above, when computed for linear regression, is identically equal to SSE.

For purposes of assessing the significance of an independent variable we compare the value of D with and without the independent variable in the equation. The change in D due to including the independent variable in the model is obtained as follows:

$$G = D(\text{for the model without the variable}) - D(\text{for the model with the variable})$$

This statistic plays the same role in logistic regression as does the numerator of the partial $F$ test in linear regression. Because the likelihood of the saturated model is common to both values of D being differences to compute G, it can be expressed as

$$G = -2\ln[\frac{(\text{likelihood without the variable})}{(\text{likelihood with the variable})}]$$

Under the hypothesis that $\beta_1$ is equal to zero, the statistic G will follow a chi-square distribution with 1 degree of freedom. Additional mathematical assumptions are also needed; but for the above case they are rather non-restrictive and involve having a sufficiently large sample size, n.

The calculation of the log likelihood and likelihood ratio test are standard features of any good logistic regression package. This makes it possible to check for the significance of the addition of new terms to the model as a matter of routine. In the simple case of a single independent variable, we can first fit a model containing only the constant term. We can then fit a model containing the independent variable along with the constant. This gives rise to a new log likelihood. The likelihood ratio test is obtained by multiplying the difference in these two values by -2.

Likelihood Ratio (G) statistic= Model chi-square with 12 degrees of freedom= 113.11 with p = 0.0000.

The classification table displays the results of classifying the data based on the logistic regression model compared to the actual allocation. Considering a child to be have no diarrhea or having diarrhea is based on the predicted probability of have no diarrhea (Prob($y_i$ =1)). A probability near to one suggests that this child is more likely to have no diarrhea. (In practice, predicted probabilities below certain classification cut-off value (normally 0.5) are classified as having diarrhea. However, the cut-off point, need not be the same for all the models, it may be equal to any value between 0 and 1. Hence, the cut-off point, for each model, is chosen in such a way that best classification is obtained). Full details of the cut-off point for predicted probability of have no diarrhea to yield high overall agreement with the actual condition, which are obtained from ROC analysis are given later.

From the above results, the sensitivity of the model is estimated as $\frac{339}{510}x100 = 66.47\%$ and the specificity is estimated to be $\frac{318}{510}x100 = 62.35\%$. Overall accuracy or percent correctly classified is $\frac{657}{1020}x100 = 64.41\%$. The model therefore appears to be moderately sensitive and specific in prediction.

## 3.2) Logistic Regression for Dakahlia Governate

A Linear Logistic Regression (LLR) model was fitted for the response variable Y, is 0 for a child having diarrhea disease and 1 for a child having no diarrhea using the selected subset of variables obtained from the step-up procedure. The logistic regression analysis gave the following details of results about the estimated (LLR) model, shown in table (4) for Dakahlia Governate based on 660 observations.

Six variables out of the sixty three independent variables (the $X_{20}$ garbage disposal, $X_{23}$ own land, $X_{31}$ watery stool, $X_{35}$ thirsty, $X_{41}$ dentition and $X_{62}$ previous use of ORT) were chosen by the step-up procedure in the LLR model. One of these variables is environmental condition ($X_{20}$), one variable is economic condition ($X_{23}$) and two variables are knowledge of mother of the signs and symptoms of diarrhea disease ($X_{31}$ and $X_{35}$), one variable is knowledge of mother the causes of diarrhea (X41) and one variable is type of treatment ($X_{62}$).

The $\chi^2$ for testing the equality of all the β's to zero show that the model is significantly different from zero (The $\chi^2$ value and its significance are given in table (4)). This together with the relatively high overall agreement between the actual and the predicted have not diarrhea suggests that this model is worth considering.

**Table (4): Variable Selection for Logistic Regression of Dakahlia Governate**

| Variable | Beta Estimate | Standard Error | Chi-square | Prob. | R-Square |
|---|---|---|---|---|---|
| Intercept | -3.86571 | 1.155332 | 11.2 | 0.0008 | 0.0169 |
| $X_{20}$ | 0.408070 | 0.091187 | 20.03 | 0.0000 | 0.0298 |
| $X_{23}$ | -0.44777 | 0.183719 | 5.94 | 0.0148 | 0.0090 |
| $X_{31}$ | 0.566325 | 0.177377 | 10.19 | 0.0014 | 0.0154 |
| $X_{35}$ | 1.979543 | 1.075532 | 3.39 | 0.0657 | 0.0052 |
| $X_{41}$ | 0.477081 | 0.196664 | 5.88 | 0.0153 | 0.0089 |
| $X_{62}$ | 0.498392 | 0.188380 | 7.00 | 0.0082 | 0.0106 |

The estimated (LLR) for Dakahlia governate based on 660 observations is:

$$p(y=1|X)=\frac{1}{1+\exp(-(-3.8657+0.4081X_{20}-0.4477X_{23}+0.5663X_{31}+1.9795X_{35}}$$

$$\frac{1}{+0.4771X_{41}+0.4984X_{62}))}.$$

Then the estimated logit, $\hat{g}(x)$, of the multiple logistic regression model is

$$\hat{g}(X)=-3.8657+0.4081X_{20}-0.4477X_{23}+0.5663X_{31}+1.9795X_{35}$$
$$+0.4771X_{41}+0.4984X_{62}.$$

Corresponding to the logistic regression model given above, a positive beta estimate is an indicatior of a decrease in the probability of having diarrhea as the variable score moves from 0 to 1, therefore positive coefficients are associated with poor outcomes (Y=0). So, from table (4), it can be concluded that the probability of no diarrhea is low in children whose family carry out refuse disposal by throwing it near to the house or in surface water (unhealthy methods), mothers who know signs and symptoms of diarrhea disease like watery stool, mothers who know signs and symptoms of diarrhea disease like thirst, mothers who know causes of diarrhea disease like dentition, mothers who previous used of ORT as treatment of diarrhea disease.

Negative beta estimates for children's family with their own land as an economic condition variable indicate that as the score for this variable move from 0 to 1, the probability of no diarrhea increases. So owning land as an economic condition is give a higher probability of no diarrhea.

This does not mean, however, that all of these variables are efficient and accurate in their prediction of no diarrhea under this conditions. The signs of the beta estimates are only an indication of which response of the independent variable is more favorable for prediction. By looking at the relative values or importance of the variables, it can be seen that some variables may be more significant than others. From table (4) results show that mother's knowledge, attitude and practices as regards diarrhea disease, economic conditio (land ownership), environmental condition (refuse disposal), and demographic information (age of mother are useful in predicting no diarrhea, respectively. This observation is supported by the corresponding chi square values and probability values. Probability levels given by the chi-squared test indicate the relativ significance of the variables.

**Table (5): Classification Matrix for Dakahlia Governate**

| | **Predicted** | | |
|---|---|---|---|
| **Actual** | 0 | 1 | **Total** |
| 0 | 254 | 94 | 348 |
| 1 | 136 | 176 | 312 |
| **Total** | 390 | 270 | 660 |

Likelihood Ratio (G) statistic= Model chi-square with 6 degrees of freedom= 72.49 with p=0.0000. From the above results, the sensitivity of the model is estimated as $\frac{176}{312}x100 = 56.41\%$ and the specificity is estimated to be $\frac{254}{348}x100 = 72.99\%$. Overall accuracy or percent correctly classified is $\frac{430}{660}x100 = 65.15\%$. The model therefore appears to be moderately sensitive and highly specific in prediction.

## 3.3) Logistic Regression for Sohag Governate

A Linear Logistic Regression (LLR) model was fitted for the response variable Y, is 0 for a child having diarrhea disease and 1 for a child having no diarrhea using the selected subset of variables obtained from the step-up procedure. The logistic regression analysis gave the following details of results about the estimated (LLR) model, shown in table (6) for Sohag Governate based on 360 observations.

Ten variables out of the sixty three independent variables (the $X_2$ type of residence, $X_{23}$ own land, $X_{27}$ own radio, $X_{31}$ watery stool, $X_{34}$ increase number of defecation, $X_{42}$ other diseases, $X_{49}$ causes of choice place of treatment, $X_{57}$ breast feeding during diarrhea, $X_{58}$ eating during diarrhea and $X_{63}$ who takes the treatment decision) were chosen by the step-up procedure in the LLR model. One of these variables is demographic information ($X_2$), two of these variables are economic condition ($X_{23}$ and $X_{27}$) and three variables are knowledge of mother the signs and symptoms of diarrhea disease ($X_{31}$, $X_{34}$ and $X_{42}$), one variable is causes of choice place of treatment (X49) and three variables are the attitude of mother during diarrhea ($X_{57}$, $X_{58}$ and $X_{63}$).

The $\chi^2$ for testing the equality of all the β's to zero show that the model is significantly different from zero (The $\chi^2$ value and its significance are given in table (6)). This together with the relatively high overall agreement between the actual and the predicted have not diarrhea suggests that this model is worth considering.

Table (6): Variable Selection for Logistic Regression of Sohag Governate

| Variable | Beta Estimate | Standard Error | Chi-square | Prob. | R-Square |
|---|---|---|---|---|---|
| Intercept | -1.19117 | 1.703283 | 0.49 | 0.4843 | 0.0014 |
| $X_2$ | -0.6510303 | 0.2731406 | 5.68 | 0.0171 | 0.0160 |
| $X_{23}$ | -0.9404006 | 0.280479 | 11.24 | 0.0008 | 0.0312 |
| $X_{27}$ | -1.006365 | 0.4822272 | 4.36 | 0.0369 | 0.0123 |
| $X_{31}$ | 0.7660261 | 0.2492349 | 9.45 | 0.0021 | 0.0264 |
| $X_{34}$ | 0.3070792 | 0.2565061 | 8.16 | 0.0043 | 0.0229 |
| $X_{42}$ | -1.219059 | 0.6085835 | 4.01 | 0.0452 | 0.0114 |
| $X_{49}$ | -0.1562579 | 0.0621164 | 6.33 | 0.0119 | 0.0178 |
| $X_{57}$ | 0.2912931 | 0.1494755 | 3.80 | 0.0513 | 0.0108 |
| $X_{58}$ | 0.3881866 | 0.1841258 | 4.44 | 0.0350 | 0.0126 |
| $X_{63}$ | 0.2584016 | 0.1141531 | 5.12 | 0.0236 | 0.0145 |

The estimator linear logistic regression for Sohag governate based on 330 observations is:

$$p(y=1|X)=\frac{1}{1+\exp(-(-1.19117-0.65103X_2-0.9404X_{23}-1.006X_{27}}$$

$$\frac{1}{+0.766X_{31}+0.73281X_{34}-1.219X_{42}-0.1562X_{49}+0.2913X_{57}}$$

$$\frac{1}{+0.3882X_{58}+0.2584X_{63}))}.$$

Then the estimated logit, $\hat{g}(x)$, of the multiple logistic regression model is

$$\hat{g}(X)=-1.19117-0.65103X_2-0.9404X_{23}-1.006X_{27}+0.766X_{31}+0.73281X_{34}$$
$$-1.219X_{42}-0.1562X_{49}+0.2913X_{57}+0.3882X_{58}+0.2584X_{63}.$$

Corresponding to the logistic regression model given above, a positive beta estimate is an indication of a decrease in the probability of no diarrhea as the variable score moves from 0 to 1, therefore positive coefficients are associated with poor outcomes (Y=1). So, from table (6), it can be concluded that the probability of no diarrhea is low in children whose mothers know signs and symptoms of diarrhea as "watery stool" $X_{31}$ and "increase number of defecation" $X_{34}$, children whose mothers continue "breast feeding during diarrhea" $X_{57}$, children whose mothers continue eating during diarrhea $X_{58}$, mothers who takes the decision of treatment $X_{63}$.

Negative beta estimates for type of residence $X_2$, economic condition "own land" $X_{23}$, "own radio" $X_{27}$, mother's knowledge the causes of diarrhea "other diseases" $X_{42}$, and causes of choice place of treatment $X_{49}$ variables indicate that as the score for these variables move from 0 to 1, the probability of no diarrhea increases So, mothers who live in rural, mothers whose own land and radio, mothers who know the causes of diarrhea as other diseases and who choose a good place of treatment, their children are more likely to curing from diarrhea

By looking at the relative values or importance of the variables, it can be seen that some variables may be more significant than others. From table (6), results show that own land, knowledge of mothers the signs and symptoms of diarrhea disease as watery stool, increase number of defecation, causes of choice place of treatment, type of residence (urban or rural), eating during diarrhea and who takes the treatment decision, are useful in predicting no diarrhea. This observation is supported by the corresponding chi-square values and probability values. Probability levels given by the chi-squared test indicate the relative significance of the variables.

**Table (7): Classification Matrix for Sohag Governate**

| | Predicted | | |
|---|---|---|---|
| **Actual** | 0 | 1 | **Total** |
| 0 | 97 | 65 | 162 |
| 1 | 55 | 143 | 198 |
| **Total** | 152 | 208 | 360 |

Likelihood Ratio (G) statistic= Model chi-square with 10 degrees of freedom= 61.55 with p=0.0000. From the above results, the sensitivity of the model is estimated as $\frac{143}{198}x100 = 72.22\%$ and the specificity is estimated to be $\frac{97}{162}x100 = 59.88\%$. Overall accuracy or percent correctly classified is $\frac{240}{360}x100 = 66.67\%$. The model therefore appears to be highly sensitive and moderately specific in prediction.

## 4) Evaluating The Models

The Logistic Regression Models (LLR) enable us to define three models of each technique for predicting a child having no diarrhea for the Dakahlia and Sohag governates, Dakahlia governate and Sohag governate. We also identified the key variables which are good predictors of no diarrhea.

$X_{23}$ and $X_{31}$ without any doubt are the most important variables. They appeared in the six models for predicting a child have no diarrhea. $X_{20}$, $X_{34}$, $X_{35}$, appeared in the four models. $X_{41}$, $X_{42}$, $X_{49}$, $X_{54}$ appeared in the three models.

One of the most interesting results was the ability of predicting a child have no diarrhea using one governate data (Sohag governate) in logistic regression models. The overall agreement between the predicted and observed was 66.67%.

Comparing the different three models for the two techniques, we find differences in the variable selected, the signs and in relative importance of the independent variables. A summary of our findings from the logistic regression models is presented in table (8).

**Table (8): Comparison Between Logistic Regression Models**

| Dakahlia and Sohag | Dakahlia | Sohag |
|---|---|---|
| X1 | X20 | X2 |
| X2 | X23 | X23 |
| X7 | X31 | X27 |
| X12 | X35 | X31 |
| X20 | X41 | X34 |
| X23 | X62 | X42 |
| X31 | | X49 |
| X32 | | X57 |
| X34 | | X58 |
| X35 | | X63 |
| X41 | | |
| 54 | | |
| (1) 62.35% | (1) 72.99% | (1) 59.88% |
| (2) 66.47% | (2) 56.41% | (2) 72.22% |
| (3) 64.41% | (3) 65.15% | (3) 66.67% |
| (4) 113.11 | (4) 72.49 | (4) 61.55 |

where: (1), (2) and (3) are the correct classification rates p(0|0), p(1|1) and the percentage of all correct classifications, respectively. (4) is the likelihood ratio.

The linear logistic regression for Dakahlia and Sohag governates together is of little practical use as does not attain the required correct classification rates with specificity p(0|0)= 62.35%, sensitivity p(1|1)= 66.47%, and the percentage of correct classifications= 64.41%, with likelihood ratio= 113.11 and p= 0.0000.

The linear logistic regression for Dakahlia governate does not satisfy the correct classifications criterion with specificity p(0|0)= 72.99%, sensitivity p(1|1)= 56.41%, and the percentage of correct classifications= 65.15%, with likelihood ratio= 72.49 and p= 0.0000.

The linear logistic regression for Sohag governate satisfy the correct classifications criterion with specificity p(0|0)= 59.88%, sensitivity p(1|1)= 72.22%, and the percentage of correct classifications= 66.67%, with likelihood ratio= 61.55 and p= 0.0000.

Linear logistic regression model for Sohag governate predict the response variable with much greater accuracy than other linear logistic regression models.

From the above, the logistic regression model for Sohag governate performed very well in prediction. But the number of independent variables involved in the Sohag model (10) is much greater than the number of independent variables (6) involved in Dakahlia model and less than the number of independent variables (12) involved in Dakahlia + Sohag model.

Finally we tested the performance of LLR models against chance allocation in the last section using ROC curve analysis. The area under the ROC curve is used to test if LLR allocation is significantly different from chance allocation. All our models were significantly different from the chance allocation assumption.

## 5) ROC Curve Analysis

ROC (Receiver Operating Characterestic, Hanley & McNeil (1982)) curve plots sensitivity (p(0|0)) against the false positive rate (p(1|0)) for a selection of cut-off points for any score. These scores in this case would be the predicted probabilities under the three models.

The area of the entire graph that lies beneath the ROC curve is the preferred measure of accuracy. This area is vary from 0.5 to 1. Thus, the area =0.5 when no classification exists, that is, when the curve lies along the major diagonal, where the sensitivity and false positive proporations are are equal. A system can achieve that performance by chance alone. The area =1 for perfect classification, that is, when the curve follows the left and upper axes, such that the sensitivity proportion is 1 for all values of the false positive proportion. (However, the measure of area sufficient quit well as a single-valued measure of the locus of curves of the sort widely observed.) Calculation of the area and its standard error can be accomplished graphically but is usually performed by a computer program.

If the model is good we will expect that an approximate 95% confidence interval for the area (area ± 2SE) would not contains 0.5, ROC curve also help with choosing an optimal cut-off point where a balance between sensitivity and specificity may be obtained.

## 5.1) ROC Curve Analysis for Logistic Regression Model of Dakahlia and Sohag Governates

Figure (1) shows the ROC curve that resulted from Dakahlia and Sohag logistic regression model and the estimate of the area under the ROC curve is found to be 0.68129945 with a 95% confidence interval between 0.60583421 and 0.756764879 which is above 0.5. (Note that ROC curve analysis was performed on each model in order to choose the cut-off point p). Therefore this model is significantly differcnt from the chance allocation. The heavy dots on the curve are the points resulted from the different cut-off points considered by the program. The points on the curve which is furthest from the diagonal can be taken as the best compromise between SEN and SPEC. We therefore took the cut-off point p= 0.4801561.

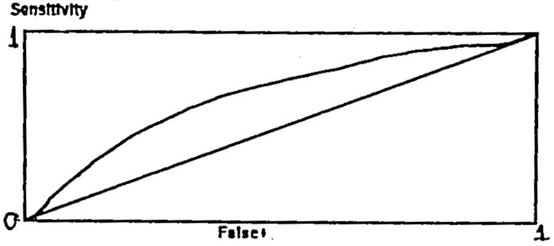

Figure (1): ROC Curve of Dakahlia and Sohag Governates Logistic Regression Model

The ROC curve plots the proportion of those actually had no diarrhea who were correctly diagnosed by the logistic regression model on the vertical axis against the proportion of those who had the diarrhea disease who were falsely diagnosed as having it on the horizontal axis. Hence, a test procedure whose ROC curve proceeds from the lower-left corner vertically until it reaches the top and then horizontally across the top to the right side is the optimum. The 45 degree line represents what you would expect from a chance (flip of the coin) classification procedure.

At the cut-off point p= 0.4801561 the classification matrix will be as follows:

| Predicted Group by | Actual Group | | |
|---|---|---|---|
| Logistic Regression Model | Diarrhea A(0) | No Diarrhea A(1) | A(All) |
| Diarrhea P(0) | 308 | 164 | 472 |
| No Diarrhea P(1) | 202 | 346 | 548 |
| P(All) | 510 | 510 | 1020 |

The classification matrix for Dakahlia and Sohag governates using the selected cut-off point (p= 0.4801561 shows that the zeros (having diarrhea) correctly classified percentage is 60.39%, the ones (having no diarrhea) correctly classified percentage is 67.84% and the all correctly classified percentage is 64.12% with model sensitivity = 0.6039 and model specificity = 0.6412.

Table (9): Area Under ROC Curve

| Variable | Area Under Curve | Standard Error | Sample Size 1 | Sample Size 2 |
|---|---|---|---|---|
| Predicted of Y= 0 | 0.68129945 | 0.03850277 | 510 | 510 |

Table (9) lists the criterion variable followed by the area under the ROC curve and the associated standard error. Note that this area depends on the number of cut-off points that were selected.

All variables in Dakahlia and Sohag logistic regression model are binary variables, so the coefficients may be directly compared.

## 5.2) ROC Curve Analysis for Logistic Regression Model of Dakahlia Governate

Figure (2) shows the ROC curve that resulted from Dakahlia governate logistic regression model and the estimate of the area under the ROC curve is found to be 0.68894601 with a 95% confidence interval between 0.597593 and 0.7802988 which is above 0.5. Therefore this model is significantly different from the chance allocation. The heavy dots on the curve are the points resulted from the different cut-off points considered by the program. The points on the curve which is furthest from the diagonal can be taken as the best compromise between SEN and SPEC. We therefore took the cut-off point p= 0.4906204.

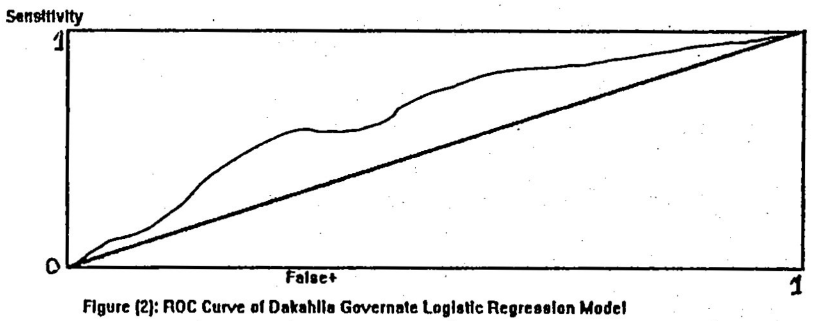

Figure (2): ROC Curve of Dakahlia Governate Logistic Regression Model

At the cut-off point p= 0.4906204 the classification matrix will be as follows:

| Predicted Group<br>by Logistic Regression Model | Actual Group<br>Diarrhea A(0) | <br>No Diarrhea A(1) | <br>A(All) |
|---|---|---|---|
| Diarrhea P(0) | 249 | 132 | 381 |
| No Diarrhea P(1) | 99 | 180 | 279 |
| P(All) | 348 | 312 | 660 |

The classification matrix for Dakahlia governates using the selected cut-off point (p= 0.4906204 shows that the zeros (having diarrhea) correctly classified percentage is 56.67%, the ones (having no diarrhea) correctly classified percentage is 71.14% and the all correctly classified percentage is 64.02% with model sensitivity = 0.5667 and model specificity = 0.7137.

Table (10): Area Under ROC Curve

| Variable | Area under curve | Standard Error | Sample Size 1 | Sample Size 2 |
|---|---|---|---|---|
| Predicted of Y= 0 | 0.68894601 | 0.04660856 | 348 | 312 |

Table (10) lists the criterion variable followed by the area under the ROC curve and the associated standard error. Note that this area depends on the number of cut-off points that were selected.

All variables in Dakahlia logistic regression model are binary variables, so the coefficients may be directly compared.

## 5.3) ROC Curve Analysis for Logistic Regression Model of Sohag Governate

Figure (3) shows the ROC curve that resulted from Sohag governate logistic regression model and the estimate of the area under the ROC curve is found to be 0.72646219 with a 95% confidence interval between 0.5885589 and 0.8643348 which is above 0.5. Therefore this model is significantly different from the chance allocation. The heavy dots on the curve are the points resulted from the different cut-off points considered by the program. The points on the curve which is furthest from the diagonal can be taken as the best compromise between SEN and SPEC. We therefore took the cut-off point p= 0.5137191.

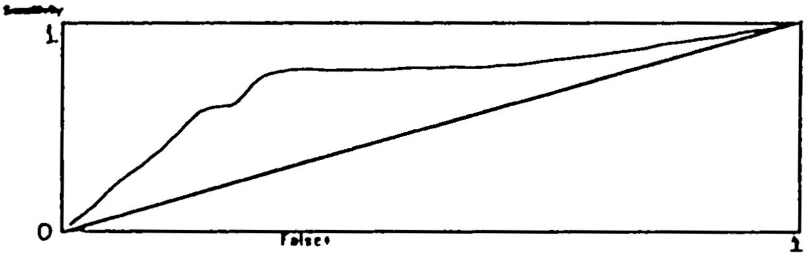

Figure (3): ROC Curve of Sohag Governate Logistic Regression model

At the cut-off point p= 0.5137191 the classification matrix will be as follows:

| Predicted Group by | Actual Group | | |
|---|---|---|---|
| Logistic Regression Model | Diarrhea A(0) | No Diarrhea A(1) | A(All) |
| Diarrhea P(0) | 108 | 61 | 169 |
| No Diarrhea P(1) | 54 | 137 | 191 |
| P(All) | 162 | 198 | 360 |

The classification matrix for Sohag governates using the selected cut-off point (p= 0.5137191 shows that the zeros (having diarrhea) correctly classified percentage is 66.67%, the ones (having no diarrhea) correctly classified percentage is 69.19% and the all correctly classified percentage is 68.06% with model sensitivity = 0.6667 and model specificity = 0.6919.

**Table (11): Area Under ROC Curve**

| Variable | Area under curve | Standard Error | Sample Size 1 | Sample Size 2 |
|---|---|---|---|---|
| Predicted of Y= 0 | 0.72646219 | 0.07034318 | 162 | 198 |

Table (11) lists the criterion variable followed by the area under the ROC curve and the associated standard error. Note that this area depends on the number of cut-off points that were selected.

All variables in Sohag logistic regression model are binary variables, so the coefficients may be directly compared.

## 6) Conclsion

Table (12) summarizes the results obtained with a computer by NCSS package for logistic regression models of Dakahlia + Sohag, Dakahlia and Sohag governates.

**Table (12): Summary of ROC Analysis Results**

| ROC Criteria | Dakahlia+Sohag Governates | Dakahlia Governate | Sohag Governate |
|---|---|---|---|
| Cut-off Point | 0.4801561 | 0.4906204 | 0.5137191 |
| Area Under ROC Curve | 0.68129945 | 0.68894601 | 0.72646219 |
| Standard Error | 0.03850277 | 0.04660856 | 0.07034318 |
| A 95% C.I. of the Area | 0.60583421 & 0.756764879 | 0.597593 &<br>0.7802988 | 0.5885589 &<br>0 8643348 |
| Sensitivity | 0.6039 | 0.5667 | 0.6667 |
| Specificity | 0 6412 | 0.7137 | 0.6919 |
| All Correctly Classified | 64.12% | 71.14% | 68 06% |

In order to compare the obtained models we presented summary of the four ROC analysis criterion (area under ROC curve, standard error, sensitivity and specificity) in table (12) Ideally a model is the "best model", if it has the maximum value of three criterion (area under ROC curve, sensitivity and specificity) and minimum value of standard error criteria. However, a compromise is needed if many models satisfy one of the criteria as the highest comparable to the others

Generally speaking the three different models are very close in terms of prediction with area under ROC curve between 0.68129945 and 0.72646219, standard error between 0.03850277 and 0.07034318, sensitivity between 0.5667 and 0.6667 and specificity between 0.6412 and 0.6919. Hence, we may be able to chose the linear logistic regression analysis for Sohag governate model as a "best model" among them.

## REFERENCES

Baker, R. J. & Nelder, J. A. (1978). *The GLIM System -Release 3.77.* Oxford, UK.: Numerical Algorithms Group. Oxford University, Institute of Statistics.

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Boston.

Cox, G. (1987).Threshold Dose - Response Models in Toxicology. *Biometrics*, 43, 511-523.

Deming, M. E. & Stephan, F. (1940). On Least Squares Adjustment of a sampled Frequency Table When Expected Marginal Totals are Known. *Ann. Math. Statist.*, 11,427-444.

Giugliano, L. G. et al. (1985). Study of Diarrhea Disease in a Peri - Urban Community in Manaus(Amazon - Brazil). *Annals of Tropical Medicine and Parasitology*, 30, 443-450.

Goodman, L. A. (1970). The Multivariate Analysis of Qualitative Data: Enteractions Among Multiple Classifications. *JASA*, 65, 226-256.

Greenland, S. (1985). An Application of Logistic Models to the Analysis of Ordinal Responses. *Biometrical J.*, 27, 189-197.

Hanley, J. A. & McNeil, B. J (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143 (1), 29-36.

Hartz, S. C. & Rosenberg, L. A. (1975). The Computation of Maximum Likelihood Estimates for Use with Categorical Data. *Journal of Chronic Diseases*, 28, 421- 429.

Hintze, J. L. (1992). Number Cruncher Statistical System (NCSS). Product 5.3. Kaysville, Utah 84037.

Ireland, C. T., & Hulback, S. (1968). Contingency Tables with Given Marginals. *Biometrika*, 55, 179-188.

Kleinbaum, D. G., Kupper, L. L. & Chambless, L. E. (1982). Logistic Regression Analysis of Epidemiologic Data: Theory and Practice. *Communication in Statist: Theory and Methods*, 11, 485-547.

McCullagh, P. & Nelder, J. A. (1983). Quasi-Likelihood Functions. *Annals of Statistics*, 11, 59-67.

Saran, M. & Gaure, S. D. (1981). Epidemiologic Correlates of Diarrhea in a Slum Community in Varanasi. *Indian. J. Pediat*, 48, 441-445.

Sen, D. et al. (1985). Etiological Spectrum of Acute Diarrhea in Hospitalized Patients in Calcutta. *Indian J. Med. Res.*, 82, 286-291.