Modeling Mortality Rates using GEE Models

Liberato Camilleri[1] and Kathleen England[2]

[1] Department of Statistics and Operations Research, University of Malta, Malta
(E-mail: liberato.camilleri@um.edu.mt)
[2] Directorate of Health Information and Research, Malta
(E-mail: kathleen.england@gov.mt)

Abstract. Generalised estimating equation (GEE) models are extensions of generalised linear models by relaxing the assumption of independence. These models are appropriate to analyze correlated longitudinal responses which follow any distribution that is a member of the exponential family. This model is used to relate daily mortality rate of Maltese adults aged 65 years and over with a number of predictors, including apparent temperature, season and year. To accommodate the right skewed mortality rate distribution a Gamma distribution is assumed. An identity link function is used for ease of interpretating the parameter estimates. An autoregressive correlation structure of order 1 is used since correlations decrease as distance between observations increases. The study shows that mortality rate and temperature are related by a quadratic function. Moreover, the GEE model identifies a number of significant main and interaction effects which shed light on the effect of weather predictors on daily mortality rates.

Keywords: Generalised estimating equation, Daily mortality rates, Apparent temperature

1 Introduction

The efficient heat regulation system of the human body enables healthy adults to cope effectively with heat and cold stress; however, this is not the case with more vulnerable older adults. The vulnerability of adults suffering from cardiovascular, respiratory and other health problems increases exponentially when temperatures exceed certain threshold limit in cold winter spells and hot summer heat waves. The cause of death during extreme temperature spells is very often attributed to the medical condition of the individual but rarely attributed to the hot or cold temperatures. There is sufficient evidence in literature of the relationship between mortality rates and temperature.

Numerous studies report a quadratic relationship between mortality rates and temperature; however the trough of the function varies by location. In warmer climatic regions, the minimum mortality rates occur at higher temperatures than colder regions. This minimum mortality temperature range varies from 14.3-17.3°C in Finland, 19.3-22.3°C in London and 22.7-25.7°C in Athens. Optimal

4rd SMTDA Conference Proceedings, 1-4 June 2016, Valletta, Malta
C. H. Skiadas (Ed)
© 2016 ISAST

temperature ranges yielding minimum mortality rates vary between regions due to the physiological adaptation of the people living in a particular region to its climate, where individuals living in hot climatic regions use air conditioners and other cooling facilities, while persons living in cold regions wear appropriate clothing and dwell in well-insulated houses.

The aim of the study is to identify the effect of temperature on mortality in the Maltese Islands and identify the optimal temperature band that yields minimum daily mortality rates. Moreover, Generalized Estimation Equation (GEE) is used to identify the significant predictors of daily mortality rates in Malta given that these outcomes are not independent.

2 Methodology

The number of daily deaths among Maltese adults aged 65 years and over was recorded over a fourteen year period. Since Malta had one of the largest population rises in the EU throughout the last decade and life expectancy is on the increase it was appropriate to rescale the number of daily deaths to yield daily mortality rates per 100,000 adults. Given the daily air temperature and relative humidity it was possible to calculate the saturated vapour pressure, vapour pressure, dew point temperature and apparent temperature.

The saturated vapour pressure (*SVP*) is related to the air temperature (*T*) by:

$$SVP = 6.11 \text{ x} 10^{\left(\frac{7.5T}{237.7+T}\right)}$$

The vapour pressure (*VP*) is related to relative humidity (*RH*) and saturated vapour pressure (*SVP*) by:

$$VP = \frac{RH \text{ x } SVP}{100}$$

The dew point temperature (*DPT*) is related to the vapour pressure (*VP*) by:

$$DPT = 237.7\left(\frac{\ln(VP) - \ln 6.11}{19.08 - \ln(VP)}\right)$$

The apparent temperature (*AT*) is related to dew point temperature (*DPT*) and air temperature (*T*) by:

$$AT = -2.653 + 0.994T + 0.0153DPT^2$$

Daily mortality rates (per 100,000) were analyzed using generalized estimating equation (GEE). To allow for skewness in the daily mortality rate distribution the model assumed a Gamma distribution and an identity link function. An auto regressive correlation structure with lag one was used since it was evident that correlations decrease as distance between observations increase. Season, year of death and apparent temperature were included both as main and interaction effects in the model fit to explain optimally thevariations in the daily mortality rates.

3 Theoretical Framework

One of the most far-reaching contributions in statistical modelling is the concept of generalized linear models introduced by Nelder and Wedderburn[8]. These models overcome the limitations of regression models, which rely heavily on the normality assumption. Generalized linear models relate the outcome variable to the linear predictor (non-random component) through an invertible link function and accommodate any error distribution within the exponential family. These models provide a unified theoretical and conceptual framework for categorical modelling procedures – Logistic and Probit regression models for Binomial data and Loglinear models for Poisson data – with the traditional regression and ANOVA methods for Normal response data. Although Generalized Linear models accommodate most of the assumptions of Regression models they still rely on the assumption that the responses are independent. These models are not well suited for the analysis of highly correlated responses when the assumption of independence is violated. On the other hand, generalized estimating equation procedures extend generalized linear models because they accommodate correlated longitudinal and clustered data. Given the correlated structure of the responses, these procedures are well suited to analyze daily mortality rates (per 100000) in the Maltese Islands and relate these rates to a number of predictors.

The seminal work authored by Liang and Zeger[6] on Generalized estimating equations (GEE) introduced an extension to the standard array of Generalized linear models (GLM) for the analysis of longitudinal data. The major limitation of GLM is that the observations are assumed to be independent; however, GEE overcome this restrictive assumption of independence. This type of estimating equations has become increasingly popular in biomedical and health science in handling existing correlated data. This category of estimating equations can also be identified as an extension of repeated measures models for non-Normal data. Parameters from GEE are estimated with a possible unknown correlation between outcomes, which are also consistent when the covariance structure is misspecified, under mild regularity conditions. The focus of the GEE is on estimating the average response over the population rather than the regression parameters that would enable prediction of the effect of changing one or more covariates on a given individual. These models are appropriate to model panel data which includes all forms of correlated data ranging from repeated measures, clustered or multilevel data.

Primarily, in a panel data set we assume that we have $i = 1, \ldots, k$ panel (clusters) for which each panel have $t = 1, \ldots, n_i$ correlated observations. In balanced panels $n_i = n_j$ for all $i \neq j$ and in unbalanced panels $n_i \neq n_j$ for at least one $i \neq j$. The probability density function of y_{it} is assumed to follow the form of the exponential family of distributions

$$f(y_{it}) = \exp\left\{\frac{y_{it}\theta_{it} - b(\theta_{it})}{a_i(\phi)} + c(y_{it}, \phi)\right\} \tag{1}$$

where the repeated observations within a given panel i are assumed to be correlated. In GLM, maximum likelihood estimation is used to estimate the parameters and hence the linear predictors and fitted values. These parameter estimates and their standard errors are consistent if the observations are independent; however their efficiency will deteriorate when the correlation between observations increases. On the other, the quasi-likelihood estimation method used in GEE models takes correlation between observations into account, thus increasing the efficiency of the estimators. This leads to consistent estimates of the parameters and their standard errors, even when the covariance structure is misspecification.

Wedderburn[9] show that the mean μ_{it} and variance function $V(\mu_{it})$ are part of the estimating equation even when the distribution being used is not a member of the exponential family. The log-likelihood implied by the estimating equation is called quasilikelihood and the rresulting parameter estimates are called maximum quasilikelihood estimates. The quasilikelihood estimates is a generalization of the likelihood. In fact, all estimates obtained from a GLM can be referred to maximum quasilikelihood estimates irrespective of the source distribution of the applied mean and variance functions. Hence the quasilikelihood estimating equation for GLMs with no restriction on the choice of the mean and variance functions is given by:

$$\Psi(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{t=1}^{n_i} \frac{(y_{it} - \mu_{it})}{a(\phi)} \frac{1}{V(\mu_{it})} \frac{\partial \mu_{it}}{\partial \eta_{it}} x_{jit} = \mathbf{0} \text{ for } \quad j = 1, 2, ..., p \tag{2}$$

Rewriting it in matrix terms of the panels

$$\Psi(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \mathbf{x}_{ji}^{'} \mathbf{D} \left[V(\boldsymbol{\mu}_i) \right]^{-1} \frac{(\mathbf{y}_i - \boldsymbol{\mu}_i)}{a(\phi)} \right\} = \mathbf{0} \text{ for } j = 1, 2, ..., p \tag{3}$$

where **D** is a diagonal matrix of derivatives $\partial \mu_i / \partial \eta_i$ and $V(\boldsymbol{\mu}_i)$ is an $(n_i \text{ x } n_i)$ diagonal matrix which can be decomposed into:

$$V(\boldsymbol{\mu}_i) = \mathbf{D}\left[V(\mu_{it})\right]^{1/2} \mathbf{I}_{(n_i \times n_i)} \mathbf{D}\left[V(\mu_{it})\right]^{1/2} \tag{4}$$

The estimating equation is treating each observation within a panel as independent. If we focus on the marginal distribution of the outcome, for which the expected value and variance functions are averaged over the panels, then the identity matrix in (4) is the within-panel correlation matrix. The GEE proposed by Liang and Zeger[5] is a modification of quasilikelihood estimating equations for GLMs that simply replaces the identity matrix with a more general correlation matrix, since the variance matrix for correlated data does not have a diagonal form.

$$V(\boldsymbol{\mu}_i) = \mathbf{D}\left[V(\mu_{it})\right]^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{D}\left[V(\mu_{it})\right]^{1/2} \tag{5}$$

The correlation matrix $\mathbf{R}_i(\boldsymbol{\alpha})$ is estimated through the parameter vector $\boldsymbol{\alpha}$. Liang and Zeger[6] stated that if the correlation matrices $\mathbf{R}_i(\boldsymbol{\alpha})$ are correctly specified, the estimator $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically Normal. Moreover, $\hat{\boldsymbol{\beta}}$ is fairly robust against mis-specification of $\mathbf{R}_i(\boldsymbol{\alpha})$. Moreover, GEE models yields both robust and model-based standard errors implying that correct specification of $\mathbf{R}_i(\boldsymbol{\alpha})$ is not essential. Liang and Zeger[6] used the term working correlation matrix for $\mathbf{R}_i(\boldsymbol{\alpha})$ and suggested that knowledge of the study design and results from explanatory analysis should be used to select a plausible form. Preferably, $\mathbf{R}_i(\boldsymbol{\alpha})$ should depend on a small number of parameters, using assumptions such as equicorrelation or autoregressive correlation.

Efficiency in the estimation of regression parameters is gained by choosing to formally include a hypothesized structure to the within-panel correlation. If the observations within a panel follow no specific order and that they are equally correlated then only one additional scalar parameter needs to be estimated. If the observations within a panel follow a more complicated structure having a specific order then a vector of additional parameters needs to be estimated. The simplest form of the working correlation matrix is the identity matrix assumed by the independence model, which imposes no additional ancillary parameters. A simple extension to the independence model correlation structure is when the observations within a panel are equally correlated, implying an additional ancillary parameter.

$$R_{ij}(\alpha) = \begin{cases} 1 & \text{if } i = j \\ \alpha & \text{if } i \neq j \end{cases} \tag{6}$$

This structure is suitable for clustered data and repeated measurements that have no time dependence. The term $\boldsymbol{\alpha}$ is called the intra-class correlation coefficient. This type of correlation goes under several names including the equicorrelation,

exchangeable, spherical and compound symmetry. If the repeated measurements are time dependent then it would be more appropriate that the observations within the panels have a natural order. If panels include observations with repeated measures recorded over time the first-order autoregressive correlation structure AR(1) assumes that:

$$R_{ij}(\alpha) = \begin{cases} 1 & \text{if } i = j \\ \alpha^{|i-j|} & \text{if } i \neq j \end{cases} \tag{7}$$

The Toeplitz working correlation matrix is similar to autoregressive correlation and assumes that any pair of observations that are equally separated in time have the same correlation, implying that the correlation structure has *n* parameters. Technically, the first-order autoregressive model is a special case of the Toeplitz.

$$R_{ij}(\alpha) = \begin{cases} 1 & \text{if } i = j \\ \alpha_{|i-j|} & \text{if } i \neq j \end{cases} \tag{8}$$

This banded correlation matrix is an alternative to autoregressive correlation where the correlations exist for small number of time units. A maximum time difference *k* is specified for which observations are correlated.

$$R_{ij}(\alpha) = \begin{cases} 1 & \text{if } i = j \\ \alpha_{|i-j|} & \text{if } |i-j| \leq k \\ 0 & \text{if } |i-j| > k \end{cases} \tag{9}$$

The unstructured correlation matrix is the most general correlation structure and it imposes no structure to $\mathbf{R}_{ij}(\alpha)$. It is only practical to use this form when $\mathbf{R}_{ij}(\alpha)$ is not large since the number of estimated parameters $n(n-1)/2$ depends on *n*. The working correlation matrix is given by:

$$R_{ij}(\alpha) = \begin{cases} 1 & \text{if } i = j \\ \alpha_{ij} & \text{if } i \neq j \end{cases} \tag{10}$$

The procedure used to estimate the vector of parameters $\boldsymbol{\beta}$ for the GEE models is equivalent to the iteratively weighted least square method used for the GLMs. This algorithm is a modification of the Newton-Raphson algorithm in which the expected Hessian matrix is substituted for the observed Hessian. The modification is known as the method of Fisher scoring. The iterative procedure is initiated by setting the correlation matrix $\mathbf{R}_i(\boldsymbol{\alpha})$ as the identity matrix and setting $a_i(\phi) = \phi = 1$. The parameters $\boldsymbol{\beta}$ are estimated by solving equations (3).

The estimates are then used to calculate the fitted values $\hat{\mu}_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$ and hence the residuals $y_i - \hat{\mu}_i$. Consequently these are used to estimate $V(\boldsymbol{\mu}_i)$, $\mathbf{R}_i(\boldsymbol{\alpha})$ and ϕ. By solving (3) and using (4) iteratively the estimates $\boldsymbol{\beta}$ can be updated.

$$\hat{\boldsymbol{\beta}}^{(r)} = \hat{\boldsymbol{\beta}}^{(r-1)} - \left\{\sum_{i=1}^{n} \mathbf{D}_i' \left[V(\boldsymbol{\mu}_i)\right]^{-1} \mathbf{D}_i\right\} \left\{\sum_{i=1}^{n} \mathbf{D}_i' \left[V(\boldsymbol{\mu}_i)\right]^{-1} \mathbf{S}_i\right\} \qquad (11)$$

where $\mathbf{D}_i = D\left[V(\mu_{it})\right] D\left(\partial\mu_i / \partial\eta\right) \mathbf{X}_i$ and $\mathbf{S}_i = \mathbf{y}_i - g^{-1}(\boldsymbol{\eta}_i)$

The solution $\hat{\boldsymbol{\beta}}$ entails the alternating estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ until the iterative procedure converges when a predetermined criterion is reached.

4 Data Results

The sample comprises 5102 daily mortality rates, rescaled per 100,000 individuals, recorded over a 14-year period. The daily mortality distribution is right skewed and follows closely the Gamma distribution. It peaks at around 13 deaths daily.

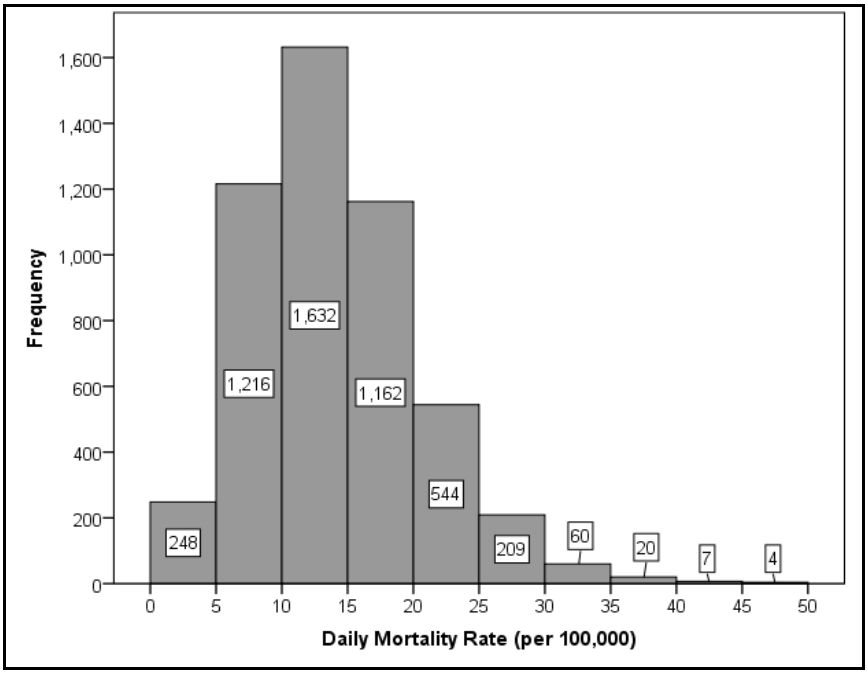

Fig.1. Daily mortality rate distribution per 100,000 adults aged 65 years and over

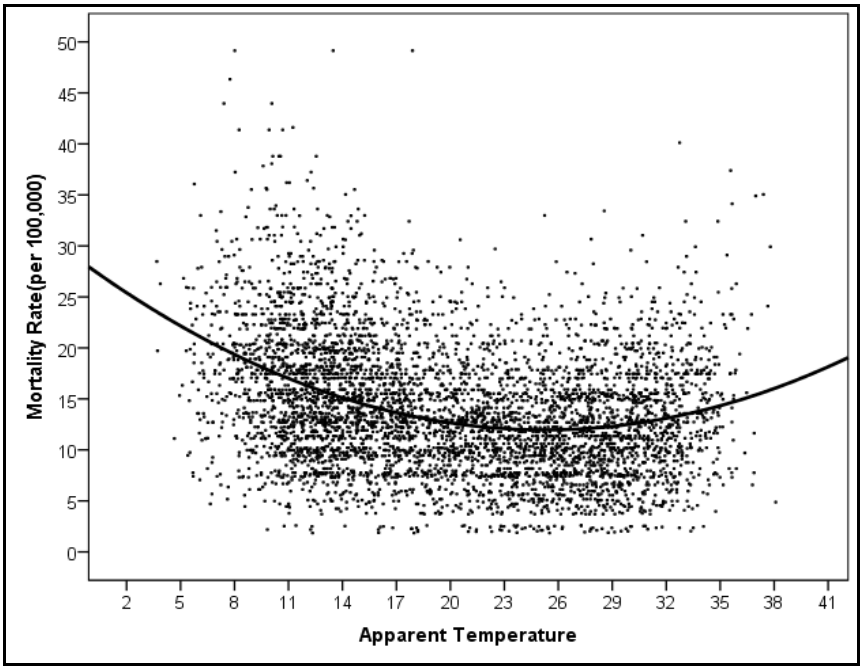

Fig.2. Relationship between daily mortality rate and apparent temperature

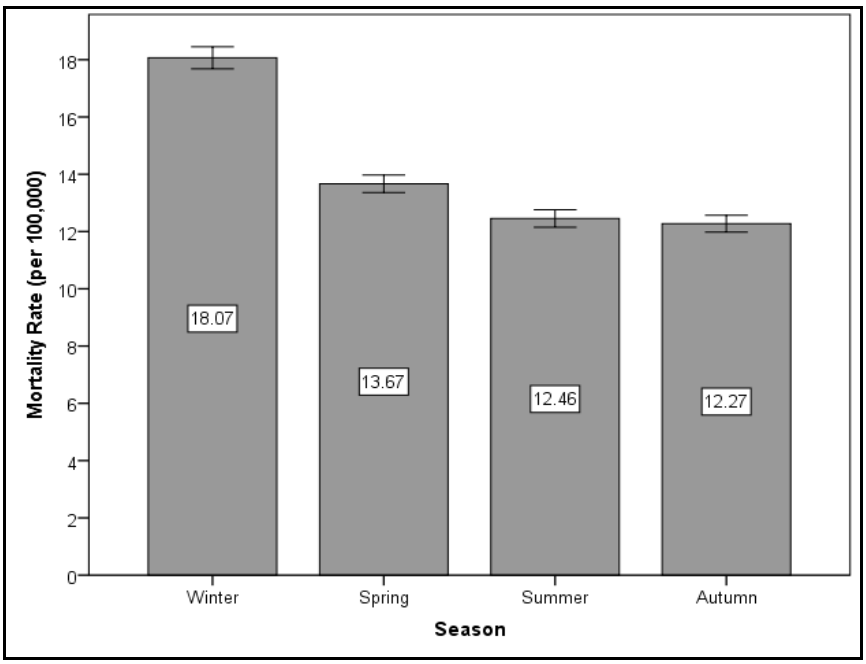

Fig.3. 95% confidence intervals for the actual mean daily mortality rates by season

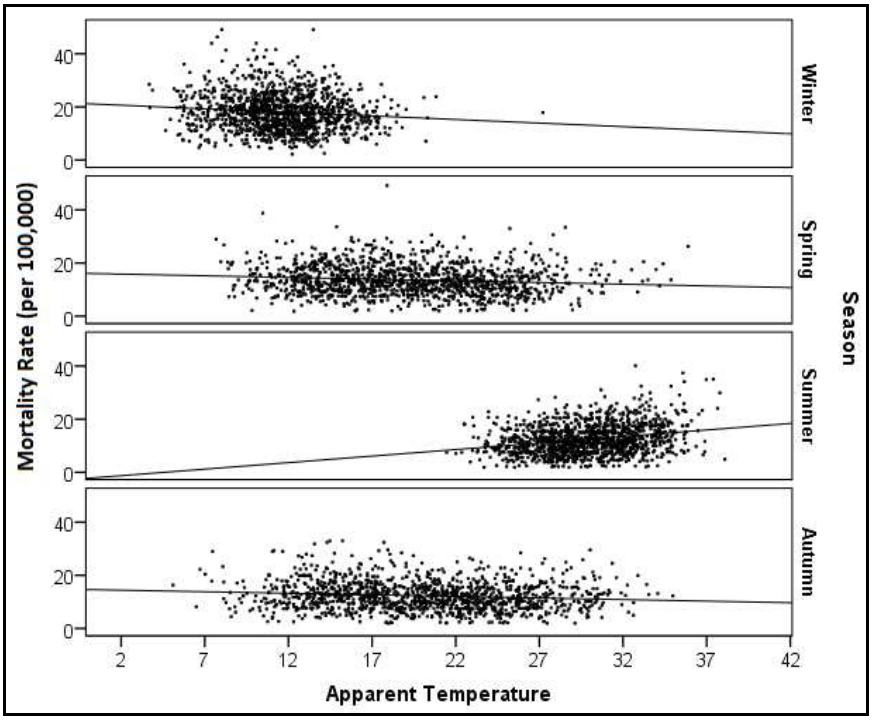

Fig.4. Daily mortality rate trends by apparent temperature and season

Figure 2 evidently shows that daily mortality rates increases at both low and high temperatures. The relationship between mortality and apparent temperature is a quadratic function, reaching a minimum daily mortality rate at around 26^{o}C. Figure 3 clearly shows that daily mortality rate in winter is significantly larger than other seasons. Figure 4 shows an interaction effect for daily mortality rate between apparent temperature and season. Daily mortality rate tend to decrease with an increase in temperature in winter and tend to increase with an increase in temperature in summer; however, daily mortality rate is not affected by a change in temperature in spring and autumn.

In the fitted GEE model, daily mortality rate is the dependent variable, whereas apparent temperature, season, and year are the predictors. The model includes season and year as main effects, a quadratic function of apparent temperature and an interaction effect of season and apparent temperature to encompass the strong results derived from the descriptive statistical analysis. Other interaction terms were excluded from the GEE model fit since their contribution in explaining variations in the daily mortality rates were not found to be significant. An AR(1) correlation structure was selected on the merit that mortality rates recorded from close days were more correlated than mortality rates recorded from distant days. This is partly explained by the high mortality rates during seasonal influenza spells or pandemic episodes and the low mortality rates during more favourable

climatic conditions. Moreover, the quasi-likelihood under the independence model criterion yielded the lowest QIC (981.92) indicating the AR(1) as the best correlation structure. Table 1 displays the results of the tests of model effects. All models effects contributed significantly in explaining total variance of the daily mortality rates and their p-values are considerably lower than the 0.05 level of significance.

Table 1. P-values of model effects

Model Effects	Wald Chi-Square	df	P-value
Season	1436.64	4	0.000
Year	68.446	1	0.000
Temperature	25.256	1	0.000
Temperature^2	17.255	1	0.000
Season * Temperature	26.157	3	0.000

Table 2 displays the estimates and standard errors of the model parameters. The regression coefficients for the season categories indicate that daily mortality rates per 100,000 in winter and spring are approximately 3.5 and 1.1 deaths higher than autumn, while the mean daily mortality in summer is comparable to autumn. The regression coefficient for Year indicates that daily mortality rate in Malta is decreasing by 0.167 yearly, given that other effects are kept fixed. The regression coefficients of the quadratic function of apparent temperature indicate that minimum daily mortality is attained at an apparent temperature of 26.4^{o}C. Hence the 3-degree temperature band of minimum mortality for Malta ranges from 24.9^{o}C to 27.9^{o}C, which is similar to other Mediterranean countries.

$$\text{Minmum Apparent Temperature} = -\frac{b}{2a} = -\left(\frac{-0.633}{2 \text{ x } 0.012}\right) = 26.37$$

Table 2. Estimates and standard errors of model parameter

Model Terms	Parameter Estimate	Standard Error
Intercept	20.837	1.255
Season = Winter	3.516	1.209
Season = Spring	1.093	0.513
Season = Summer	0.213	2.181
Season = Autumn	0	.
Year	-0.167	0.020
Temperature	-0.633	0.124
Temperature^2	0.012	0.003
Season = Winter * Temperature	-0.112	0.083
Season = Spring * Temperature	0.011	0.038
Season = Summer * Temperature	0.339	0.079
Season = Autumn * Temperature	0	.
Scale	0.167	

The regression coefficients of the interaction effect between apparent temperature and season indicate that in very hot summer days, daily mortality rates increase by 0.34 deaths for every 1^{o}C rise in temperature compared to autumn. Conversely, in very cold winter days, daily mortality rates increase by 0.11 deaths for every 1^{o}C drop in temperature compared to autumn. This implies that while more deaths occur in the winter months, daily mortality rates can soar up more rapidly with a abrupt increase in temperature during hot summer days than a sudden decrease in temperature during cold winter days.

5 Conclusion

This study shows that an optimal apparent temperature around 25 ºC to 27ºC results in minimum daily mortality rates. This indicates that minimum mortality rates in warmer regions occurs at higher temperatures than colder regions. This is mainly attributed to physiological adaptation of the people living in a particular region to its climate. People living in warm regions are better adapted to the hot weather through the use of air conditioners and cooling facilities. On the other hand, people living in cool regions are better adapted to the cold weather through the use of central heating, insulated houses and warm clothing.

Extreme cold and hot temperatures increase the number of deaths, particularly adults aged 65 years and over. Thermoregulation of body temperature of older adults is less effective compared to their younger counterparts. This fact together with other health-related problems increases mortality risks in elderly persons. Basu et. al.[1] remark that individuals with pre-existing cardiovascular and respiratory problems have higher risks of death associated with ambient heat exposure. Since most influenza spells occur in winter, health precautions are more likely to be taken in winters than summers; however, this study reveals that abrupt increment in temperatures during hot summer periods are more fatal than sudden drop in temperatures during cold winter periods.

In very humid conditions, a hot day feels hotter and a cold day feels colder. In hot summer days, sweat evaporates more rapidly in a low humidity environment. So perspiration, which is the body's cooling mechanism, is less effective in humid conditions, resulting in a slower sweat evaporation rate and slower cooling process. Conversely, on cold rainy days our clothing absorbs moisture from the humid air causing a drop in body temperature. Since Malta is an island, humidity tends to be high with very little seasonal variation. The logic of using apparent temperature rather than actual air temperature in this study is that it combines humidity and air temperature. This is important because humidity actually accentuates the body discomfort in very low and very high temperatures. It is highly recommended that health warnings are issued on different media by public health departments when temperatures fall below 10^{o}C or rise above 35^{o}C to caution vulnerable individuals of the mortality risks.

References

1. Basu R and Samet J. Relation between elevated ambient temperature and mortality: a review of the epidemiologic evidence. Epidemiol Rev, 24:190-202, 2002.
2. Donaldson G, Ermakov S, Komarov Y, McDonald C and Keatinge W. Cold related mortalities and protection against cold in Yakutsk, eastern Siberia: observation and interview study. BMJ, 317: 978-982, 1998.
3. Keatinge W, Donaldson G, Cordioli E, et al. Heat related mortality in warm and cold regions of Europe: observational study. BMJ, 321: 670-673, 2000.
4. Hardin J and Hilbe J. Generalized Estimating Equations. Chapman & Hall/ CRC, 2012.
5. Kunst A, Looman C and Mackenbach J. Outdoor air temperature and mortality in the Netherlands: a time series analysis. Am J of Epidemiol, 137: 331-341, 1993.
6. Liang K and Zeger S. Longitudinal data analysis using generalized linear models. Biometrika, 73, 12-22, 1986.
7. McCullagh P and Nelder J. Generalized Linear models (2nd Edition) London, Chapman & Hall/CRC Press (1989).
8. Nelder J and Wedderburn R. Generalized Linear Models. Journal of the Royal Statistical Society, 135, 370-384, (1972).
9. Wedderburn R. Quasi-Likelihood Functions, Generalized Linear Models and the Gauss-Newton Method, Biometrika, 61, 3, 439-447, 1974
10. Zeger S, Liang K and Albert P. Models for longitudinal data. A generalized estimating equation approach. Biometrics, 44, 1049-1060, 1988.
11. Ziegler A, Blettner M and Kastner C. The Generalized Estimating Equations, Biometrical Journal, 40, 115-139, 1998.
12. Ziegler A. Generalized Estimating Equations. New York, Springer (2011).