Modeling and Analysis of Daily Temperature using Logistic Regression and Markov Chain

Mohammad Mahmood Faqe Hussein[1], Samira Muhamad Salh[1], Akhterkhan Saber Hamad[1]

[1]College of Administration & Economics,
University of Sulaimani, KRG, Iraq
Email: mohammad.faqe@univsul.edu.iq, samira.muhamad@univsul.edu.iq,
Email ID: *akhterkhan.hamad@univsul.edu.iq*

DOI: 10.23918/ICABEP2023p40

Abstract:

Weather and climate have a great influence on all aspects of life; any changes in weather and climate pose a challenge to all sectors, especially in the field of future planning. Temperature is considered the most important climatic element that has a direct or indirect effect on other climatic elements. The goal of the study is to control the effect of yesterday's daily temperature (x_1) and the day before it (x_2) on the current daily temperature (Y: current value equal to y_i and the yesterday value y_{i-1} = and the day before = y_{i-2})

Using a hybrid Markov-Logistic regression. The data on daily temperature were collected from the https://mesonet.agron.iastate.edu/request/daily.phtml?network=iq_asos# Network for about 4 years and 3 months on the daily temperature of Sulaymaniyah Governorate during the period January 2019 to March 2023. The first outcome illustrates that the daily temperature has a Markov chain with second order, and the logistic regression also expresses that a nice and sunny daily temperature followed by a nice and sunny daily temperature and a high daily temperature followed by a high daily temperature are more likely for the temperature of Sulaymaniyah. The second result shows that the model might achieve sufficient accuracy for many applications of temperature data reasonably, and the last result shows that yesterday's daily temperature and one day before yesterday's daily temperature have an impact on the daily temperature.

Keywords: Logistic regression, Markov chain, transition probability matrix, Maximum likelihood estimation (MLE), temperatures

1 Methodology

1-1 Introduction

The technical linear statistical methods widely used as part of this learning process [2]. Linear regression and many other models are special cases of generalized linear models [5]. In general, regression is an analysis that is regarded as the study of a single variable, known as the response variable, defined by one or more variables, known as the explanatory's variables. The regression analysis used to find the mathematical model. [2] There are many situations and conditions in the world,

such as education, social science, and others, where the response variable is dichotomous (binary) rather than continuous that can expand the linear regression to include explanatory variables that are binary or have more than two levels, but when the dependent variable is binary, the interpretation of the linear regression equation will not be direct, which means that when using a linear regression model, the output is uncorrected. [7]

1-2 Objective of this study

The study's objective is to use a hybrid Markov-Logistic regression to investigate the impact of the daily temperatures from yesterday and the day before on the current daily temperature.

1- 3 Definition of Logistic Regression (LR)

Logistic regression is a regression model where the response variable (DV) is binary; whenever the response variable is categorical, we apply logistic regression with respect to linear regression.

Using various types of data, LR may be used to categorize the observations, and it can also easily identify the best variables to apply for classification. The logistic function is displayed in the figure below[10]:

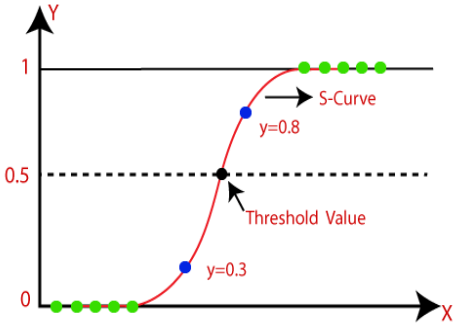

Fig (1): Represent the logistic function.

The sigmoid function used to chart the forecast values to probabilities, any real value into another value within a range of zero and one. The value of the LR must be between zero and one. The S-curve is called the sigmoid function. We use the concept of the threshold value, which defines the probability of either zero or one. Such values above the threshold value close to be one, and values below the threshold value close to be zero.[9]

1-4 The Model

The *logit model* is often used for classification and forecast. Logistic regression estimates the probability of an event occurring. Since the result is a probability, the response variable is bounded between 0 and 1. In LR, a logit transformation is applied to the odds. This is also generally known as the log odds, or the normal logarithm of odds, and this logistic function is represented by the following formulas:. [2],[5]

$$\text{logit}(P_i) = \frac{1}{(1+\text{Exp}(-Pi))} \qquad (1)$$

logit(pi) is the response variable and x is the explanatory variable

$$\ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta_1 X_1 + \cdots . + \beta_k X_k \qquad (2)$$

1-5 Characteristics of the logistic regression

The properties of the logistic regression equation include [7], [9]

- The response variable is distributed as a Bernoulli distribution
- y_i have different variances: $\text{Var}(y_i) = P_i(1 - P_i)$, i = 1,2, 3, …
- Each group $X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ik}$ has its own set of explanatory values.
- The coefficient of determination (R^2) in logistic regression is not calculated as it is in linear regression. Instead, a harmony is used to assess the model's fitness.
- Estimation/prediction is based on 'maximum likelihood method.

Also, odds are a way to express the probability that something happens compared to the probability of something else happening, and it is often expressed as a ratio between two numbers. The mathematical formula for the odds gives greater opportunity for a better understanding of the relationship between odds and probability. The description of odds (O_i)) is a proportion between the likelihood of an event happening (P_i) and the probability of the event not happening ($1 - P_i$),

$$\text{OddsRatio}(Oi) = \frac{P_i}{1-P_i} \qquad 0 \le P_i \le 1 \qquad (3)$$

Note that, the odds, which has the symbol $\mathbf{O_i}$, may solve the problem of the upper limits of the probability P_i so that the odds take any value from $(0 \to \infty)$, so the transformation odds have contributed to solving half the problem that is the upper limit of probability. The need for a transformation odds value, is to solve the minimum boundary of the allowable values for the odds.[9]

The general formula of the logistic regression model is:

$$\ln \text{odds} = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

The relationship between the logit (log odds) and explanatory variables is a linear relationship.

$$\log \text{odds} = \text{logit}(P_i) = \beta_0 + \sum_{i=1}^{j} \beta_i x_i \qquad (4)$$

1-6 The Likelihood Ratio Test, χ^2:

To fit a model, we compare the deviance with just the intercept (-2LLR: Likelihood of reduced model) to the deviance when the new predictor has been added (-2LLF: Likelihood of full model).

The difference between these two deviation values is called chi squared for goodness of fit.

H_0: The simpler model, 1 parameter vs .H_1 : The more general model, 2 parameters

$$\chi^2 = -2\ln\left(\frac{\text{LLR}}{\text{LLF}}\right) = -2(\text{LLR} - 2\text{LLF}) \qquad (5)$$

One can look up the significance of this test in a chi-square table using D.F. equal to the number of predictors added to the model.[2],[7],[9]

1-7 Stochastic Processes (SP)

SP is an indexed collection of arbitrary variables defined on a state space,$\{X(t); t \in T\}$, where T is some index set.[4],[3]

1-8 Markov Chains (M.C.) Model

A Markov chain is a discrete time and confidence net with stochastic variables in a sequence, each variable in the chain only having a direct relationship to the current value and not to any previous values. Markov chains are used to depict values in sequences, such the states of a dynamic system. A "stage" is used to describe each position in the sequence.[3],[4]

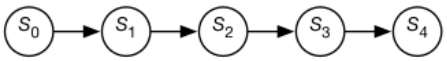

Fig (2): A confidence net using a Markov chain

Although the net has five stages, it need not conclude at stage 4; it can continue on indefinitely. The freedom premise is communicated by the belief net. :[6]

$$P(S_{t+1} = s_{t+1}|S_t = s_t) = P(S_{t+1} = s_{t+1}|S_o = s_o, \dots, S_t = s_t) \qquad (6)$$

Which is called a Markov property. The Markov property can be seen as "the future state dependent on the current state, not on the past state".

If all of the variables have the same domain and the transition probabilities are the same at each step, the Markov chain is stationary, i.e.,

$$P(S_{t+1} = s_{t+1}|S_t = s_t) = P(S_1 = s_1|S_o = s_o) \quad \text{for} \quad \text{all } t \geq 0 \qquad (7)$$

A transition probability matrix (T.P.M.) is defined as

$$P = [p_{ij}] = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

Where the elements satisfy

$$P_{ij} \geq 0(\text{positive number}),\ 0 \leq P_{ij} \leq 1 \quad \forall i, j \in S \ \text{ and } \ \sum_{j=0}^{\infty} P_{ij} = 1\ , i = 0,1,2, \dots \qquad (8)$$

1-9 The Probability Distribution of $\{S_t, t \geq 0\}$:

Let

$P_i^n = P[S_t = i] = [P_0^n \quad P_1^n \quad P_2^n \quad \dots]$

Where

$$\sum_{i=1}^{\infty} P_i^n = 1$$

P_i^n is called the state probability vector after *n* transitions.

$P_i^0 = P(X_0 = i)$ are the initial-state probabilities,

$$\text{initial} - \text{state probability vector} = P_i^{(0)} = \left[P_0^{(0)} \quad P_1^{(0)} \quad P_2^{(0)} \quad \dots\right] \quad , \quad i = 0,1,2,..$$

$$P_i^n = P^0 \times P^n \qquad (9)$$

Which indicates that the probability distribution of a homogeneous Markov chain is completely determined by the 1-step transition probability matrix P and the initial-state probability vector $p(0)$.[1],[10]

1-10 Maximum Likelihood (ML) for Parameters Estimation

Let $P_{j\to k}(t)$, $(j = 1,2,3 \text{ and } t = 1,2,3 \dots T, \ k = 1,2,3)$ be the probability of state k at time t, given that the state j at time $(t-1)$. If $P_{j\to k}(t)$ is stationary then $P_{j\to k}(t) = P_{j\to k}$, for all t.

Let a Markov chain with stationary transition probabilities $P_{j\to k}$ and finite number of states $1,2,3$. Let $f_{j\to k}$ be the no. of observations in transition from state j to the state k. The overall number of observations is [1],[6]

$$\sum_{j=1}^{2}\sum_{k=1}^{2} f_{j\to k} \qquad (10)$$

$$\sum_{k=1}^{2} f_{j\to k} = f_{j.} \quad \text{and} \quad \sum_{k=1}^{2} f_{j\to k} = f_{.k}$$

The Markov chain model is based on a set of explanatory trails. It reached the exact probability density of the observed $f_{j\to k}$

$$g(f_{i\to j}) = T(f_{i\to j}) \frac{\prod_{j=1}^{2}(f_{j.})!}{\prod_{j=1}^{2}\prod_{k=1}^{2}(f_{i\to j})!} \prod_{j=1}^{2}\prod_{k=1}^{2} P_{j\to k}^{f_{j\to k}} \qquad (11)$$

The factor $T(f_{j\to k})$ is the joint probability density of f_j's and is independent of $P_{j\to k}$'s. If we consider a Markov chain model with non-stationary transition probabilities $P_{j\to k}(t)$ then the probability density of $f_j(t)$ is given by

$$g\{f_{j\to k}(t)\} = T\{f_{j\to k}(t)\}\frac{\prod_{j=1}^{2}\{f_{j.}(t)\}!}{\prod_{j=1}^{2}\prod_{k=1}^{2}\{f_{j\to k}(t)\}!}\prod_{j=1}^{2}\prod_{k=1}^{2}P_{j\to k}^{f_{j\to k}}(t) \qquad (12)$$

In the case of non-stationary transition probabilities, the set $\sum_{t=1}^{T} f_{j\to k} = f_{j\to k}(t)$

$P_{i\to j\to k}(t)$. The set $f_{i\to j\to k}(t)$ is a multinomial set of sufficient statistics. The logarithm of the likelihood function for stationary transition probabilities $P_{i\to j\to k}$ as,

$$LP_{j\to k}(t) = C + \sum_{j=1}^{2}\sum_{k=1}^{2} f_{j\to k}\log P_{j\to k}$$

C : is the one that contains all terms that are independent of $P_{i\to j\to k}$. The MLE of $P_{i\to j\to k}$ are found to be

$$\hat{P}_{j\to k} = \frac{f_{j\to k}}{f_{j.}} \qquad (13)$$

For non-stationary transition probabilities, the MLE are

$$\widehat{P}_{j\to k}(t') = \frac{f_{j\to k}(t)}{f_{j.}(t)} = \frac{f_{j\to k}(t)}{f_{j.}(t-1)} \qquad (14)$$

Thus, we may estimate the T.P.M. of the Markov chain model. [4][8][10]

2 Analysis and Discussion

Analysis of the daily temperatures and their dependence by M.C. the explanatory variables are measured on various types of scales, but they are categorized in dichotomous form considering the extended past behavior of these meteorological factors in Sulaimaniyah. Notational, all response and explanatory variables are as follows: The response variable daily temperature (Y_i) is categorized: if the daily temperature is nice and sunny (< 30°C) it takes on the value 0, but if the daily temperature is high (>=30°C) it takes on the value 1. The data on daily temperatures was obtained from this site.

https://mesonet.agron.iastate.edu/request/daily.phtml?network=IQ__ASOS# during from (1/1/2019) to (31/3/2023)

x_1= Yesterday's daily temperature

$$x_1 = \begin{cases} 0 & x_1 < 30 \\ 1 & x_1 \geq 30 \end{cases}$$

x_2=Day before yesterday's daily temperatures is

$$x_2 = \begin{cases} 0 & x_2 < 30 \\ 1 & x_2 \geq 30 \end{cases}$$

2-11 First order of the transition probability matrix

Table 1, show the (↑ %0.973) goes to transition nice and sunny day to nice and sunny day and the (↓ %0.027) goes to transition nice and sunny day to the high. It is to be noted that transition to the nice and sunny day is higher than the percentage of transition to the high.

Table (1): First order of transition frequency distribution of daily temperature

Yesterday's state of daily temperature	Today's state of daily temperature		Sum
	Nice and Sunny (0)	High (1)	
Nice and Sunny (0)	1249	34	1283
High (1)	35	229	264
Sum	1284	263	1547

The table gives the MLE of T.P.M. for a 1st-order Markov chain found by using the formulation:

$$P_{i\to j} = P[X_n = j | X_{n-1} = i] = \frac{f_{i\to j}}{\sum_j f_{i\to j}}$$

Table (2): Maximum Likelihood Estimation (MLE) of T.P.M. of the 1st. order model

T.P.M.	P_{00}	P_{01}	P_{10}	P_{11}
M.L.E	0.973	0.027	0.133	0.867

2-12 Goodness-Of-Fit

A Chi-Square test has been used to test the T.P.M. In this study, we obtained the following expected and observed frequencies from the data:

H_0: The order of the Markov chain is zero. vs. H_1: The order of the Markov chain Model is one.

Table (3): Expected Frequency

	1	2
1	(1064.9)	(218.1)
2	(219.1)	(44.9)

From the chi-square analysis, we have

Table (4): Goodness of fit of first order of daily temperature.

	Test value	D.F.	P-value
Chi-Square Test	1097.270	1	0.000

From the table, we note that the P-value of the chi-square test (0.000) compares the value with the (α=0.05), we accept the alternative hypothesis H_1, which means that the Markov chain is of the order one.

2-13 Second order of the transition probabilities matrix

In Table 6, we noted that among 1448 days, 0.8483 percent remained in the nice and sunny state for three consecutive days, whereas 7.67% (0.1517) of the days remained in the wet state. The rest of the states have changed the temperature status at least once in the 3-consecutive days. The (↑ % 0.9799) belongs to the transition type nice and sunny at all consecutive days, and the (↓ %0.0200) is for the transition of the day before yesterday (nice and sunny) to yesterday (nice and sunny) to today (high). The MLE of the T.P.M. of a 2-order of the M.C. is found directly by using transition counts by the formula:

$$P_{i\to j\to k} = P[X_n = k | X_{n-1} = i, X_{n-2} = i] = \frac{f_{i\to j\to k}}{\sum_k f_{i\to j\to k}}$$

Table (5): Frequency distribution of the second order.

daily temperature states in the direct previous two days	daily temperature states in present day		Sum
	Zero	One	
Zero-Zero	1221	25	1246
Zero-One	12	23	35

One-Zero	25	8	33
One-One	23	111	134
Sum	1281	167	1448

Table (6): Maximum Likelihood Estimation (MLE) of T.P.M of the 2st. order model

T.P.M	M.L.E
P_{000}	0.9799
P_{001}	0.0200
P_{010}	0.3428
P_{011}	0.6571
P_{100}	0.7575
P_{101}	0.2424
P_{110}	0.1716
P_{111}	0.8283

2-14 Goodness-Of-Fit

A Chi-Square test has been used to test the T.P.M. In this study, we obtained the following expected and observed frequencies from the data:

Hypothesis Test:

H_0: The order of the Markov chain is one. vs. H_1: The order of the Markov chain is two.

Table (7): Expected Frequency

	1	2
1	(1102.3)	(143.7)
2	(31)	(4)
3	(29.2)	(3.8)
4	(118.5)	(15.5)

From the chi-square analysis, we have

Table (8): Goodness of fit of second order of daily temperature

	Test value	D.F.	P-value
Chi-Square Test	884.471	3	0.000

From the table, we note that the P-value of the chi-square test (0.000) compares the value with the (α=0.05), we accept the alternative hypothesis H_1, which means that the M.C. is of the order 2.

2-15 Logistic Regression Model:

The result of analysis is presented in the tables below. The overall number of cases is 1551. Observation 0 means that the Nice and Sunny ($<30^0$C) of daily temperature and observation 1 means the high ($\geq 30^0$C) of rainfall. Out of daily temperature cases, 12867(82.9%) cases are the nice and sunny of daily temperature and 265(17.1%) cases are high of daily temperature. The proportion of rainfall is 82.9%.

Table (9): Encoding of response variable

States	Coding value	Total Number of cases
Nice and Sunny	0	1551
High	1	

Table (10): Variables in the logistic Regression Model

Variables	β	S.E.	Wald	D.F.	Sig.
Yesterday's Daily Temperature	4.045	.313	167.277	1	.000

Day Before Yesterday's Daily Temperatures	2.004	.320	39.175	1	.000
Constant	-3.717	.178	435.083	1	.000

With the values of the test statistics (Wald) of 167.277 and 39.175 for yesterday and for the day before yesterday, with one degree of freedom, we get the significance level at (0.000 <α=0.05). Hence, we can say that both coefficients have a significant influence on the temperature. So now it can be interpreted that the probability that today is nice and sunny is approximately 57 times as likely as if yesterday was nice and sunny compared to that day.

The probability that today is nice and sunny is approximately 8 times as likely if the day before yesterday was nice and sunny compared to that day. Moreover, it is 7.69= [57.115/7.418] times more likely that the probability of today being high depends 7.69 times more on yesterday's high daily temperature as compared to the day before yesterday's high daily temperature.

Table (11) Test coefficient using Omnibus Tests

	Chi-Square	D.f.	Sig.
Step	922.145	2	0.000
Block	922.145	2	0.000
Model	922.145	2	0.000

Hypothesis Test:

H_0: The model is non-significance **vs**. H_1: The model is significance.

The results show that the value of chi-square (922.145) at a degree of freedom 2 and a significant level. (0.000<α=0.05), and this means that the statistical model that has been reconciled has a statistically significant, which indicates that the variables add in the model have an influence and a contribution to the classification.

Table (12): Hosmer and Lemeshaw test for quality of fit

Step	Chi-square	D.f.	Sig.	Step
1	2.351	1	0.125	1

Hypothesis Test:

H_0: The model is consistent with the data vs. H_1: The model is not consistent with the data

Accordingly, to the value of chi-square (2.351) at a degree of freedom of 1 and the level of significance (0.125> α =0.05), we accept the null hypothesis, and we conclude that the model is consistent with the study data, the model represents the data well, and we note that the differences between the observed and expected values are very simple.

Table (13): Coefficient of R square of Cox & Snell and Nagelkerke

Step	-2LL	Cox &Snell R Square	Nagelkerke R Square
1	495.474	0.449	0.748

The variables included in the model showed that they were explained about 0.748 using the coefficient Nagelkerke R Square and 0.449 using the coefficient Cox &Snell R Square of the changes that occur in the effect of the response variable, and this indicates that there is a percentage of changes in the response variable due to other variables not included in the model.

Table (14): Classification Table

			Predicted		
			Nice and Sunny	High	Percentage Correct
Step 1	Daily Temperature	Nice and Sunny	1249	35	97.3
		High	35	230	86.8
	Overall Statistics		1117.479	2	95.5

From the previous table, we note that the percentage of correct classification of daily temperatures (nice and sunny) amounted to %97.3, and the percentage of incorrect classification amounted to 2.7%. As for the daily temperatures (high), the percentage of correct classification amounted to 86.8% and the percentage of incorrect classification amounted to 13.2%.

3-16 Conclusion and Recommendations

3-16-1 Conclusions

1. Both the Markov chain model and the logistic regression model are the two main statistical methods that are the subject of the majority of our discussion in this work. We combine them to create a model, which we then employ to study the daily temperature.
2. A logistic regression model defines the dependence of one categorical explanatory variable on another dichotomous response variable.

3. A Markov chain of more than first order and logistic regression are showed to visualize the dependence of the present daily temperature upon the daily temperature of the earlier two-time period. It had been shown that the nice and sunny day of the previous two time periods compared to the high temperature day of the previous two time periods influenced completely the nice and sunny day of the current time period. In a logistic regression case, we consider that the current day's daily temperature is influenced by the daily temperature of the previous two days. But temperature is also influenced by other meteorological factors.

3-16-2 Recommendations

1. Generalizing the results of this study to other atmospheric variables.
2. Expansion of green spaces within the city to reduce the effects of high temperatures.
3. Orientation towards the construction of environmentally friendly buildings through the enactment of special laws.

References

Berchtold, A. (1999). The double chain Markov model. Communications in Statistics-Theory and Methods, 28(11), 2569-2589.

Forrest, T. L. (2007). Logistic regression models for predicting trip reporting accuracy in GPS-enhanced household travel surveys (Doctoral dissertation, Texas A&M University).

Hamdin, A. O., & Hussein, M. M. (2023). An application of Wavelet Markov Chains Model to Study Earthquake Occurrence. Tikrit Journal of Administrative and Economic Sciences, 19(61, 2), 334-353.

Hussein, M. M. F. (2022). Diagnosing the Causes of Car Accidents in the Garmian Region Using Multinomial Logistic Regression Model. Iraqi Journal of Humanitarian, Social and Scientific Research, 2(4).

Kozłowski, E., Borucka, A., & Świderski, A. (2020). Application of the logistic regression for determining transition probability matrix of operating states in the transport systems. Eksploatacja i Niezawodność, 22(2), 192-200.

Mohammed, M. F. (2018). Use the Markov chains to evaluate and predict the Probabilistic Waiting Time for Asia Cell Customers Company for the period (1/9/2014 to 1/6/2015). journal of kirkuk University For Administrative and Economic Sciences, 8(1).

Sinha, N. C., Islam, M. A., & Ahamed, K. S. (2011). Logistic regression models for higher order transition probabilities of Markov chain for analyzing the occurrences of daily rainfall data. Journal of Modern Applied Statistical Methods, 10(1), 31.

Tolver, A. (2016). An introduction to Markov chains. Department of Mathematical Sciences, University of Copenhagen.

Walker, J. (1996). Methodology Application Document: Logistic Regression Using the CODES Data (No. HS-808 460).

Website:https://www.javatpoint.com/logistic-regression-in-machine-learning