International Journal of Statistical Distributions and Applications
2021; 7(4): 95-101
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20210704.13
ISSN: 2472-3487 (Print); ISSN: 2472-3509 (Online)

An Analysis of Corona Virus Disease (COVID-19) Predictors: Logistic Regression Model Approach

Usman Aliyu[1, *], Abubakar Umar Bashar[1], Umar Usman[2]

[1]Department of Statistics, Waziri Umar Federal Polytechnic, Birnin Kebbi, Nigeria

[2]Department of Statistics, Usmanu Danfodiyo University Sokoto, Nigeria

Email address:

usmanaliyusta@gmail.com (U. Aliyu), uusman07@gmail.com (U. Usman), abumar24@yahoo.com (A. U. Bashar)

[*]Corresponding author

To cite this article:

Usman Aliyu, Abubakar Umar Bashar, Umar Usman. An Analysis of Corona Virus Disease (COVID-19) Predictors: Logistic Regression Model Approach. *International Journal of Statistical Distributions and Applications*. Vol. 7, No. 4, 2021, pp. 95-101.
doi: 10.11648/j.ijsd.20210704.13

Received: August 15, 2021; **Accepted**: September 17, 2021; **Published**: November 19, 2021

Abstract: Although Corona Virus disease (COVID-19) is a contagious disease cause by severe acute respiratory syndrome which affects mostly people whose immune system are weak or not resistance to the disease, there exists no vaccine that is 100% effective for its cure though efforts are being intensify by researchers in discovering the vaccine as well as model for prediction of Corona Virus Disease. In this era of advanced information and communication technology, as well as evidence-based medicine, statistical modeling has become as necessary the medical practitioners who are interested in lasting solution to diagnosed problems. In this work a logistic regressions model has been proposed to serve the purpose. The data was obtained from Nigeria Centre for Disease Control (NCDC) and was analyzed using binary logistic regression model in which Corona Virus disease was considered as categorical dependant variable (COVID-19 status: chance of being positive or negative) and the predictors considered are; Age, any of either Headache or Vomiting, Fever, Sore throat/runny nose, Any of Cold, cough or sweating, Loss of Smell or taste, and Breathing Difficulties. The results shows the significant predictors for predicting Corona Virus Diseases are; Loss of Smell or taste, Breathing Difficulties, Fever, Sore throat or runny nose, Age, any of either Headache or Vomiting, and Any of Cold, cough or sweating. The logit model obtained was: Logit (P(y=1)) = -3.748 + 0.356 Age +2.938 any of either Headache or Vomiting + 0.752 Fever + 2.792 Sore throat or runny nose – 0.028 Any of Cold, cough or sweating + 1.872 Loss of Smell or taste + 0.844 Breathing Difficulties. So also from the same results, it was found among predictors that; Sex/Gender, Temperature >37.5 degree and Fatigue or Muscle Pain were not good predictors of Corona Virus disease.

Keywords: Logit Function, COVID-19, Logistic Regression, Maximum Likelihood Estimation

1. Introduction

Logistic regression deals with the binary case, where the response variable consists of just two categorical values [23]. Logit model is mainly used to identify the relationship between two or more explanatory variables. X_i and the dependent variable Y. Logistic regression model has been used for prediction and determining the most influential explanatory variables on the dependent variable [11, 12]. The Logistic regression model for the dependence of p_i (response probability) on the values of k explanatory variables x_1, x_2,……..x_k is given below [10].

$$Logit(P_i) = Log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \quad (1)$$

$$\text{Or } P_i = \frac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)} \quad (2)$$

Which is linear and similar to the expression of multiple linear regression.

Where $\left(\frac{P_i}{1-P_i}\right)$ is the ratio of the probability of a failure and called odds, β_0, β_i are parameters to be estimated and pi is the response probability.

Logistic regression predict the outcome of a dichotomous dependent variable based on one or more predictor variables (features) that is, in estimating empirical values of the parameters in a qualitative response model [4]. The probabilities describing the possible outcomes of a single trial are modeled, as a function of the explanatory (predictor) variables, using a logistic function [1]. Logistic regression measures the relationship between a categorical dependent variable and one or more independent variables, which are usually (but not necessarily) continuous, by using probability scores as the predicted values of the dependent variable [2].

Logistic regression is binomial or binary logistic regression when the observed outcome for a response variable can have only two possible types (for example, "the patient is Corona Virus positive" vs. "the patient is Corona Virus negative") and Logistic regression is Multinomial logistic when the outcome of the response variable can have three or more possible types (e.g., " the patient A is Corona Virus positive" vs. "the patient B is Corona Virus positive" vs. "the patient C is Corona Virus positive") [6]. Logistic regression is used to predict the odds of being a case based on the values of the independent variables (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a non-case [18].

Logistic regression models are adequate for those situations where the dependent variable of the regression problem is binary. That is, the dependent variable has only two possible outcomes, e.g., "success/failure" or "normal/abnormal". We assume that these binary outcomes are coded as 1 and 0. [19]. The application of linear regression models to such problems would not be satisfactory since the fitted predicted response would ignore the restriction of binary taking on values for the observed data. When studying linear regression, we attempted to estimate a population regression equation;

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \tag{3}$$

By fitting the model of the form;

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k \tag{4}$$

The response Y was continuous, and was assumed to follow a normal distribution. We were concerned with predicting or estimating the mean value of the response corresponding to a given set of values for the explanatory variable [20]. In general, the value 1 is used to represent a "success" or the outcome we are not interested in, and 0 represents a "failure" [5]. The mean of the dichotomous random variable Y, designated p, is the proportion of times that it takes the value 1. Equivalently;

$$p = P(Y = 1) = P(success)$$

Just as we estimated the mean value of the response when Y was continuous, we would like to be able to estimate the probability p associated with a dichotomous response (which, of course, is also its mean) for various values of an explanatory variable. To do this, we use the technique of logistic regression. A simple regression model for this situation is:

$$Y_i = g(x_i) + \varepsilon_i \tag{5}$$

With $y_i \in \{0,1\}$

According to [24], logistic regression is identified as the most popular method used in analyzing epidemiological data when the outcome variable is binary. The response variable is coded with the value 0 or 1 and it is used in categorical data.

Logistic regression provides a method for modeling a binary response variable. For example, we may wish to investigate how death (1) or survival (0) of patients can be predicted by a level of one or more metabolic markers. According to [25], a logistic regression is considered as a parametric model and is a form of generalized linear model. This is because the probability distribution for the response variable is specified as well as the error terms. Logistic regression makes use of several predictor variables which may be categorical or numerical. The odds ratio is usually of interest in a logistic regression due to its ease of interpretation. Odds ratio is a statistic that measures the odds of an events compared to the odds of another event [for 2 x 2 contingency table, the odds ratio is a measure of association [3]. Combination of the odds and the logistic regression leads to the interpretation of any logistic regression result [19].

A large sample size is needed for testing of hypothesis in logistic regression since it does not require much assumption for the hypothesis to be accurate. This is because of the nature of probabilities which logistic regression principles are based. A logit transformation is used [15, 16, 21].

Hauck, W. W, and Donner, A. [17] examined the performance of the Wald test and likelihood ratio test. They found that Wald test behaved in an aberrant manner, often failing to reject the null hypothesis when the coefficient was significant. Therefore, they recommended the likelihood ratio.

Hosmer, David W. et al [20] highlighted that it is possible to construct a model that fits the data (good estimation of the relationship between response and explanatory variables) but is a poor predictive model.

According to Michael, H. K et al [26], logistic regression is an important nonlinear regression model and could be considered for use when the response variable is qualitative with two possible outcomes, such as financial status of firm (sound status, headed towards insolvency) or blood pressure status (high blood pressure, low blood pressure).

2. The Logistic Function

The term "Logit" as a contraction of the phrase "logarithmic unit" was introduced by [7]. This is in analogy to the term "Probit" as a contraction of the phrase "Probability unit" as introduced by [8]. Recall that if an event occurs with probability p, odds in favour of the event are;

$$\frac{p}{1-p} \quad to \quad 1$$

Thus, if a success occurs with probability

$$p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \tag{6}$$

the odds in favour of success are

$$\frac{p}{1-p} = \frac{e^{\alpha+\beta x}/(1+e^{\alpha+\beta x})}{1/(1+e^{\alpha+\beta x})} = e^{\alpha+\beta x} \tag{7}$$

So that we have: $p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$

This expression on the right, called a logistic function (logit model), cannot yield a value that is either negative or greater than 1; consequently, it restricts the estimated value of p to the required range [9].

3. The Method of Maximum Likelihood

Logistic regression uses the Maximum Likelihood Estimation method to estimate the model coefficients and the method of maximum likelihood uses the information in a sample to find the parameter estimates that are most likely to have produced the observed data. This method yields values of α and β which maximize the probability of obtaining the observed set of data [14]. Conceptually, it works like this:

First construct a likelihood function which expresses the probability of the observed data as a function of the unknown parameters α and β. In the univariate case, the contribution to the likelihood function for a given value of the predictor X, is

Let Y_i represent response variable, X_i represent covariates, we get:

$$P(Y_i = 1) = \pi_i = \frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)} \tag{8}$$

We can extend the simple logistic regression model easily to more than one predictor variable.

Let us define,

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{p-1} \end{bmatrix}_{p\times 1} \quad X = \begin{bmatrix} 1 \\ X_1 \\ \dots \\ X_{P-1} \end{bmatrix}_{P\times 1} \quad X_i = \begin{bmatrix} 1 \\ x_{i1} \\ \dots \\ x_{i,p-1} \end{bmatrix}_{p\times 1}$$

Then, we get,

$$X'\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} \tag{9}$$

$$X_i'\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} \tag{10}$$

$$\text{So } E\{Y_i\} = \pi_i = \frac{exp(X_i'\beta)}{1+exp(X_i'\beta)} \tag{11}$$

Recall that, the joint probability function for binary logistic regression is:

$$g(Y_1, \dots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \tag{12}$$

$$log_e g(Y_1, \dots, Y_n) = log_e \prod_{i=1}^{n} f_i(Y_i) = log_e \prod_{i=1}^{n} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \tag{13}$$

$$= \sum_{i=1}^{n} [Y_i log_e \pi_i + (1-Y_i) log_e (1-\pi_i)] \tag{14}$$

$$= \sum_{i=1}^{n} \left[Y_i log_e \left(\frac{\pi_i}{1-\pi_i} \right) \right] + \sum_{i=1}^{n} log_e (1-\pi_i) \tag{15}$$

$$\text{Since } 1-\pi_i = \frac{1}{1+exp(\beta_0+\beta_1 x_i)} \text{ and } log_e \left(\frac{\pi_i}{1-\pi_i} \right) = \beta_0 + \beta_1 x_i \tag{16}$$

Therefore,

$$log_e L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} log_e [1 + exp(\beta_0 + \beta_1 x_i)]. \tag{17}$$

We are trying to find β_0 and β_1 to maximize the log-likelihood function:

$$ln = log_e L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} log_e [1 + \exp(\beta_0 + \beta_1 X_i)]. \tag{18}$$

Define:

$$\underset{\sim}{y}U = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix} \quad X_U = \begin{bmatrix} X_1^T \\ X_2^T \\ \dots \\ X_N^T \end{bmatrix}$$

The model is $Y = X^T\beta$. The estimator of B is $\underline{\hat{\beta}} = (X_U^T \sum_U^{-1} X_U)^{-1} X_U^T \sum_U^{-1} y_U$, where $\sum_u is$ a diagonal matrix with ith diagonal element σ_i^2.

3.1. WALD Test

The Wald test will be familiar to those who use multiple regressions. In multiple regressions, the common t-test for testing the significance of a particular regression coefficient is a Wald test [25]. In logistic regression, the Wald test is calculated in the same manner. The formula for the Wald statistic is

$$Z_j = (b_j / sb_i) \tag{19}$$

where sb_i is an estimate of the standard error of b_j provided by the square root of the corresponding diagonal element of the covariance matrix, *v(β)*.

With large sample sizes, the distribution of Z_j is closely approximated by the normal distribution. With small and moderate sample sizes, the normal approximation is described as 'adequate' [27].

3.2. Likelihood Ratio Tests

The likelihood ratio test (LRT) is based on the difference in -2LL between the researcher's model and the null model, with a finding of significance indicating a good model. The likelihood ratio test statistic is -2 times the difference between the log likelihoods of two models, one of which is a subset of the other [17]. The distribution of the LR statistic is closely approximated by the chi-square distribution for large sample sizes. The degrees of freedom (DF) of the approximating chi-square distribution is equal to the difference in the number of regression coefficients in the two models. The test is named as a ratio rather than a difference since the difference between two log likelihoods is equal to the log of the ratio of the two likelihoods. That is, if L_{full} is the log likelihood of the full model and L_{subset} is the log likelihood of a subset of the full model, the likelihood ratio *js* defined as

$$LR = -2\,[L_{subset} - L_{full}] = -2\,[ln\,(l_{subset}/l_{full})] \quad (20)$$

Note that the – 2 adjusts LR so the chi-square distribution can be used to approximate its distribution [17].

3.3. Goodness of Fit of the Model

In logistic regression, instead of R^2 as the statistic for overall fit of the model, we have chi-square instead. when we studied chi-square analyses, chi-square was said to be a measure of "goodness of fit" of the observed and the expected values. We use chi-square as a measure of model fit here in a similar way. It is the fit of the observed values (Y) to the expected values (Y'). The bigger the difference (or "deviance") of the observed values from the expected values, the poorer the fit of the model [22, 13]. So, we want a small chi-square if possible. As we add more variables to the equation the deviance should get smaller, indicating an improvement in fit. The difference between these two deviance values is often referred to as G for goodness of fit.

$$G = \chi^2 = D(\textit{for the model without the variable}) - D(\textit{for the model with the variable}) \quad (21)$$

or, using the Pedhazur notation, $G = \chi^2 = -2LL_R - (-2LL_F)$ (22)

An equivalent formula is: $G = \chi^2 = -2\ln\left(\frac{likelihood_R}{liklihood_F}\right)$ (23)

where the ratio of the ML values is taken before taking the log and multiplying by –2. This gives rise to the term "likelihood ratio test" to describe G.

3.4. The Hosmer-Lemshow Test

H-S test was considered a model goodness of fit for logistic regression [20]. H-L test divides the sample into portions (usually 10 deciles) and compares observed and expected (predicted) values of the DV within each decile, then uses a type of averaging to get a whole-model result for which a finding of non-significance indicates a good model. The data are divided into approximately ten groups defined by increasing order of estimated risk. The observed and expected number of cases in each group is calculated and a Chi-squared statistic is calculated as follows:

$$x^2_{HL} = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - E_g/n_g)} \quad (24)$$

with O_g, E_g and n_g the observed events, expected events and number of observations for the g^{th} risk decile group, and G the number of groups. The test statistic follows a Chi-squared distribution with G-2 degrees of freedom.

A large value of Chi-squared (with small p-value < 0.05) indicates poor fit and small Chi-squared values (with larger p-value closer to 1) indicate a good logistic regression model fit.

4. Material and Methods

The data was obtained from Nigeria Centre for Disease Control (NCDC) consist 500 people diagnosed for Corona Virus disease (COVID-19) out of which some happens to be positive while others are negative. For easy analysis of data, the following coding were made;

Corona Virus disease status (1= Patient is COVID-19 Positive, and 0 = Patient is COVID-19 Negative).

Any of either Headache or Vomiting (1 = Yes 0 = No).

Ages (between 1-25 years=1, 26-50 years=2, 50 years and above=3).

Sex/Gender (1 = male, 0 = female).

Fever (1 = Yes, 0= No).

Temperature ≥ 37.5°C (1 = Yes, 0 = No)

Sore throat or runny nose (1 = Yes, 0 = No)

Loss of Smell or taste (1 = Yes, 0 = No)

Breathing Difficulties (1 = Yes, 0 = No)

Fatigue or Muscle Pain (1 = Yes, 0 = No)

Any of cough, cold or sweating, etc. (1 = Yes, 0 = No).

5. Results and Discussions

Analysis was performed using Statistical Package for Social Sciences (SPSS) version 20 and the output of the analyzed data are as follows;

Table 1. *Logistic Regression Coefficients for Likelihood of Corona Virus Diseases (COVID-19).*

	B	S.E.	Wald	Df	P-value	Odd Ratio	95% C.I. for Odd Ratio	
							Lower	Upper
Sex/Gender	-0.237	0.340	0.487	1	0.485	0.789	0.405	1.535
Age	0.356	0.325	1.199	1	0.027	5.701	1.370	9.325
Headache or Vomiting	2.938	0.303	94.27	1	0.000	18.878	10.432	34.16
Fever < 7 Days	0.752	0.263	8.144	1	0.004	2.121	1.265	3.553
Temperature ≥ 37.5°C	0.600	0.519	1.336	1	0.248	1.822	0.659	5.041
Sore throat or Runny nose	2.792	0.460	52.615	1	0.012	9.371	7.536	24.263
Cold, Cough or Sweating	-0.028	0.010	8.544	1	0.003	2.973	0.955	6.991
Loss of smell or taste	1.872	0.290	41.765	1	0.016	6.503	3.686	11.474
Breathing Difficulties	0.844	0.324	6.792	1	0.009	2.325	1.233	4.386
Fatigue or Muscle Pain	0.868	0.477	3.32	1	0.068	2.383	0.936	6.063
Constant	-3.748	0.543	47.699	1	0.000	0.024		

a. Variable(s) entered on step 1: Any of either Headache or Vomiting, Ages, Sex/Gender, Fever, Temperature ≥ 37.5°C, Sore throat or runny nose, Loss of Smell or taste, Breathing Difficulties, Fatigue or Muscle Pain, Any of Cold, cough or sweating, etc.

It can be noted from Table 1 that the predictors such as Age, any of either Headache or Vomiting, Fever, Sore throat or runny nose, Any of Cold, cough or sweating, Loss of Smell or taste, and Breathing Difficulties, with the significance values 0.027, 0.000, 0.004, 0.012, 0.003, 0.016 and 0.009 respectively are each less than $\alpha = 0.05$. This means that these predictors are each important to be included in the final model. Therefore, there are enough bases to conclude that these predictors are relevant predictors in predicting Corona Virus in kebbi state. From the same table 1, it is revealing to note that, the predictors such as Sex/Gender, Temperature >37.5 degree and Fatigue or Muscle Pain were dropped from the model. Since the p – values 0.485, 0.248 and 0.068 were each greater than $\alpha = 0.05$, that means there is sufficient evidence to indicate that these predictors were not important to be included in the model. Hence the predictor's Sex/Gender, Temperature ≥ 37.5°C and Fatigue or Muscle Pain were not relevant in predicting Corona Virus.

5.1. Interpretation of THE Odds Ratios and Wald Statistic

As in Table 1 the strongest predictor of the outcome of Corona Virus patient was Severe Headache or Vomiting, recording an odds ratio of 18.878 (95% C.I. = 10.432 to 34.16). This indicated that patients who had been checked and referred as having Severe Headache or Vomiting is likely to estimate the success of Corona Virus as to those who were not referred, controlling for all other factors in the model. The odds ratio 9.371 with (95% C.I. = 7.536 - 24.263) for Sore throat or runny nose indicating that for every treatment per patient, there were more Corona Virus due to a pain in throat lasting for some time caused by changes in pressure in the blood vessels leading to and from the brain, controlling for other factors in the model. Again, the odds ratio with respect to Loss of Smell or taste was 6.503 (95% C.I. = 3.686 to 11.474) meaning that more of the Corona Virus positive was estimated by Loss of Smell or taste, holding other factors constant.

The Wald Chi-Square statistic, which tests the unique contribution of each predictor, in the context of the other predictors -- that is, holding constant the other predictors -- that is, eliminating any overlap between predictors. from the analysis it was observed that the Headache or Vomiting contribute more in predicting Corona Virus Diseases as it records 94.27 followed by Sore throat or Runny nose recording 52.615, then Loss of smell or taste as it has 41.765, followed by Cold, Cough or Sweating 8.544 then Fever less than 7 days i.e 2 to 3 days as it records 8.144 and Breathing Difficulties recording 6.792.

5.2. Model Fit Assessment Test

H_0: The hypothesized model fits the data.
H_1: The hypothesized model does not fit the data.

Table 2. *Assessing Model Fit by Hosmer and Lemeshow Test.*

Step	Chi-square	Df	Sig.
1	15.898	8	0.044

We can observe that from Table 2, for model assessment using Hosmer Lemeshow test, since the P-value which is 0.044, is less than significance value i.e alpha = 0.05, we accept the null hypothesis (H_0) and conclude that there is enough evidence to show that the hypothesized model fits the data set used in predicting Corona Virus. Hence, this indicates that the model adequately fits the data.

Table 3. *Regression Model Summary.*

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	455.436[a]	0.369	0.494

It is observed from above table that, between 36.9% and 49.4% of the variance in predicting whether or not Corona Virus patient would be positive or negative was explained by the predictors; Age, any of either Headache or Vomiting, Fever, Sore throat or runny nose, Any of Cold, cough or sweating, Loss of Smell or taste, and Breathing Difficulties.

Meanwhile, Nagelkerke or Pseudo R-square was 12.8% more than the Cox & Snell R-Square value.

Table 4. *Classification Table.*

Observed		Predicted		
		Corona Virus Status		Percentage Correct
		The Patient is Corona Virus Negative	The Patient is Corona Virus Positive	
Corona Virus Status	The Patient is Corona Virus Negative	85	9	90.4
	The Patient is Corona Virus Positive	7	399	98.3
Overall Percentage				96.8

However, it was shown from the classification table (table 4) that, about 98.3% could be predicted as the patient has Corona Virus whilst about 90.4% can be predicted as the patient has no Corona Virus but any other related issue. It is worth noting that, overall, about 96.8% of the cases were correctly classified.

In this research, it was discovered that the chance of being Corona Virus disease positive in dependant on predictors; Age, any of either Headache or Vomiting, Fever, Sore throat or runny nose, Any of Cold, cough or sweating, Loss of Smell or taste, and Breathing Difficulties. The logistic regression model obtained was:

$$\text{Logit}(P(y=1)) = -3.748 + 0.356\ \text{Age} + 2.938\ \text{any of either Headache or Vomiting} + 0.752\ \text{Fever} + 2.792\ \text{Sore throat or runny nose} - 0.028\ \text{Any of Cold, cough or sweating} + 1.872\ \text{Loss of Smell or taste} + 0.844\ \text{Breathing Difficulties}$$

Again, from the same results, it was found among predictors that; Sex/Gender, Temperature >37.5 degree and Fatigue or Muscle Pain were not good predictors of Corona Virus disease.

6. Conclusion

This study provides evidence for the predictors of Corona Virus diseases (COVID-19) in Nigeria. The model indicates that Headache, Sore throat/runny nose and difficult breathing contributes more in terms of predicting Corona Virus disease. So also it was observed that loss of taste or smell, feeling cold/cough or sweating were at higher chance of Corona Virus disease.

Therefore, based on the data collected and analysis made on the data we conclude that the predictors which actually influence Corona Virus disease are Age, any of either Headache or Vomiting, Fever, Sore throat or runny nose, Any of Cold, cough or sweating, Loss of Smell or taste, and Breathing Difficulties.

7. Recommendations

The following are suggested recommendation:

1) It is recommended that the model built by this research should be should be used as a means of justification concerning prediction of Corona Virus diseases.
2) It is recommended that physical or social distance should be maintain as well as quarantining and covering coughs and sneezes, maintaining regular hand washing and keep hands away from the face or mouth.
3) Use of face mask or coverings is recommended o as to minimize the risk of transmissions.

References

[1] Afifi, W. A. Dillow, M. R. and Morse, C. (2004). Examining Predictors and Consequences of Information seeking in Close Relationships, Personal relationships, 11, 429-449.

[2] Agresti, A. (1996). An Introduction to Categorical Data Analysis. John Wiley and Sons, Inc. http://lib.stat.cmu.edu/datasets/agresti

[3] Agresti, A. (2007), An Introduction to Categorical Data Analysis, Second Edition, Wiley, Inc., New York.

[4] Agresti, Alan. (2002). Categorical Data Analysis. New York: Wiley-Interscience. ISBN 0-471-36093-7.

[5] Amemiya, T. (1985). Advanced Econometrics. Harvard University Press. ISBN 0-674-00560-0.

[6] Balakrishnan, N. (1991). Handbook of the Logistic Distribution. Marcel Dekker, Inc. ISBN 978-0-8247-8587-1.

[7] Berkson, J. (1944). Application of Logistic function to bio-assay. *Journal of the American Statistical Association* 39, 357-365.

[8] Bliss, C. J. (1935). The calculation of the dosage mortality curve. Annals of Applied Biology 22, 134-167.

[9] Christensen, R. (1997), Log-linear Models and Logistic Regression, Second edition. Springer-Verlag, New York.

[10] Collett, D. R. (2003). *Modeling Binary data*, Chapman & Hall, London.

[11] Cox, D. R. and Snell, E. J (1994). *Analysis of binary data*. Chapman & Hall, London.

[12] Cox, D. R. and Snell, E. J. (1989), Analysis of Binary Data (second edition), Chapman & Hall/CRC.

[13] Devita, V. T., Hellman, S. L. and Rosenberg's, S. A. (2008), Principles and Practice of Oncology, Volume 2.

[14] Greene, William H. (2003). Econometric Analysis, fifth edition. Prentice Hall. ISBN 0-13-066189-9.

[15] Grimms, L. G. and Yarnold, P. R. (1995), Reading and Understanding Multivariate Statistics, American Psychological Association, Washington, D. C.

[16] Grizzle, J. E., Starmer, C. F. and Koch, G. G. (1969), Analysis of Categorical data by Linear Models, Biometrics, 25, 489-504.

[17] Hauck, W. W, and Donner, A., (1997), Wald's test as Applied to Hypothesis in Logit Analysis. Journal of the American Statistical Association, 72, 851 – 853.

[18] Hilbe, Joseph M. (2009). Logistic Regression Models. Chapman & Hall/CRC Press. ISBN 978-1-4200-7575-5.

[19] Hosmer D. W and Lemeshow S. (2000). Applied Logistic Regression. 2nd ed. New York, USA: John Wiley and Sons.

[20] Hosmer, David W.; Lemeshow, Stanley (2000). Applied Logistic Regression (2nd ed.). Wiley. ISBN 0-471-35632-8.

[21] Howell, David C. (2010). Statistical Methods for Psychology, 7th ed. Belmont, CA; Thomson Wadsworth. ISBN 978-0-495-59786-5.

[22] Ingeles, C. J.; Garcia-Fernandez, J. M. Castejon, J. L.; Valle Antonio, B. D. and Marzo, J. C (2009), Reliability and Validity Evidence of Score on the Achievement Goal Tendencies: Questionnaire in a sample of Spanish students of compulsory secondary education, Psychology in the school, Vol. 46. 1048 – 1060, Wiley Periodicals, Inc; A Wiley company.

[23] Jennings, D. E. (1986), Judging Inference Adequacy in Logistic Regression, Journal of the American statistical Association, 81, 471 – 476.

[24] Kleinbaum, D. G. (1994), Logistic Regression, A self-learning text. Springer - Verlag, New York, 104 – 119.

[25] McCaullagh, H. and Nelder, J. N., (1992), Generalized linear Models (2nd Edition). Chapman and Hall, Madras.

[26] Michael, H. K.; Christopher, J. N., John N., and William. L (2005). "Applied Linear Statistical Model", fifth Ed.; 555-623., McGraw Hill International, New York.

[27] Peduzzi, P.; J. Concato, E. Kemper, T. R. Holford, A. R. Feinstein (1996). "A simulation study of the number of events per variable in logistic regression analysis". Journal of Clinical Epidemiology. 49 (12): 1373–1379. PMID 8970487.