ISSN 2090-3359 (Print)
ISSN 2090-3367 (Online)

ΑΔΣ

Advances in Decision Sciences

Volume 25
Issue 2
June 2021

Michael McAleer

Editor-in-Chief
University Chair Professor
Asia University, Taiwan

Published by Asia University, Taiwan

ADS@ASIAUNIVERSITY

Specification and Estimation of a Logistic Function, with Applications in the Sciences and Social Sciences*

Kim-Hung Pho

Fractional Calculus, Optimization and Algebra Research Group
Faculty of Mathematics and Statistics
Ton Duc Thang University
Ho Chi Minh City, Vietnam

Michael McAleer **

Department of Finance, College of Management
Department of Bioinformatics and Medical Engineering
College of Information and Electrical Engineering
Asia University, Taiwan
Discipline of Business Analytics, University of Sydney Business School, Australia
Econometric Institute, Erasmus School of Economics
Erasmus University Rotterdam, The Netherlands
Department of Economic Analysis and ICAE
Complutense University of Madrid, Spain

Revised: May 2021

* For financial support, the second author wishes to thank the Australian Research Council and the Ministry of Science and Technology (MOST), Taiwan.

** Corresponding author: michael.mcaleer@gmail.com

Abstract

This research makes a theoretical contribution by providing straightforward and coherent derivation of a logistic model, and then estimating the parameters of the model with a fishing data set. The logistic model is frequently considered as a convenient regression model to find the associations between a binary outcome variable and several covariates. This is also a model that has numerous practical applications, as in banking, engineering, social sciences, medical research and biostatistics. In the paper, we briefly summarize the function and estimating equation of the logistic model. We next investigate the large sample properties of this model under some regularity conditions. We then provide a simulation study of the work. A factual application of the logistic model is illustrated using a fishing data set. The results have consilience with practice. It also shows that this is a reliable model to maximize the number of fish while fishing. Finally, some applications in decision sciences, some concluding remarks, and future research directions are discussed.

Keywords: Estimation, Logistic, Regression models, Fishing data, Decision Sciences.

JEL: J16, K38, M14.

1. Introduction

In statistics and in several other sciences, a linear regression model is one of the most essential and widespread statistical models to delineate the associations between a continuous dependent variable and numerous covariates. However, in several practical situations, the dependent variable is not a continuous variable, but a binary variable such as yes/no, success/failure, and several other choie sets. Linear models are no longer pertinent to analyze these kinds of data. In order to overcome this issue, Cox (1958) pioneered a logistic regression model that is now used widely in the literature.

Up to the present time, the logistic model has become a primary tool for studying the associations between binary outcome variables and covariates. This model is generally considered the most commonly used regression model, and has many real and practical applications in regression analysis. Starting from the pioneering work of Cox (1958), hitherto there have been many scientists studying this model, as in Hosmer and Lemesbow (1980), who presented goodness of fit tests for the multiple logistic model.

Albert and Anderson (1984) supplied the existence of maximum likelihood estimates for the logistic model. DeMaris (1995) offered a tutorial on the logistic model. Hosmer et al. (1997) provided a comparison of goodness of fit tests for the logistic model. A monograph on applied logistic regression analysis is available in Menard (2002).

Recently, this model continues to be studied strongly in both theory and application, as in Allison (2012), who introduced SAS software for the logistic model. Hosmer et al. (2013) presented a book about the applied logistic model. Austin and Merlo (2017) offered several important topics in multi-level logistic regression analysis.

Wang et al. (2020) predicted epidemic trends COVID-19 by using the logistic model and machine learning technique. Furthermore, Chang et al. (2021) used the logistic model to estimate the parameters in a two-stage randomized response technique.

The logistic model is a model that has numerous actual applications in banking, finance, engineering, social science, medical, biostatistics, among many others. Some preeminent references to these can be found in Enea and Lovison (2019) and Ugwuanyim et al. (2020). Regression analysis aims to understand the nature of things and phenomena, from which it is possible to forecast and estimate future problems. To the best of our knowledge, the logistic model is the most crucial and meaningful model in regression analysis, which is which is a reason it has been investigated extensively in the sciences and social sciences.

As an example, the fishing data set is a ubiquitous data set in practice, and it is also a very frequent problem with great interested in the current era. Albeit there are several research papers about regression analysis, the use of logistic model to analyze the specific fishing data set has not yet been studied. An important goal of this paper is to present a detailed and complete recipes and estimate parameters of a logistic model with the fishing data set. We then discuss a number of important applications of the logistic model to some important areas in the sciences and social sciences.

The reminder of the paper is described as follows. Section 2 introduces the formula, log-likelihood and score function of the logistic model. The theory of the large sample properties of the logistic model under some regularity conditions is investigated in Section 3. A simulation study is investigated in Section 4, with the fishing data set studied in Section 5. Section 6 presents applications of the logistic model in some important disciplines in the sciences and social sciences. Some concluding remarks and future areas of research are oconsidered in the last section.

2. Logistic Regression Model

Let Y be a binary outcome variable, and X and Z be the covariates. The logistic model can be described as follows:

$$P\left(Y_i=1 \mid X_i, Z_i\right)=H\left(\eta_0+\eta_1^T X_i+n_2^T Z_i\right)=\frac{e^{\eta^T \mathrm{X}}}{1+e^{\eta^T \mathrm{X}_i}}=H\left(\boldsymbol{\eta}^T \mathrm{X}_i\right), \tag{1}$$

where $H(u)=(1+e^{-u})^{-1}$ is the logistic distribution function, $\mathrm{X}_i=\left(1, X_i^T, Z_i^T\right)^T$, and $\boldsymbol{\eta}=\left(\eta_0, \eta_1^T, \eta_2^T\right)^T$ is a vector of parameters to be estimated.

The log-likelihood of $\boldsymbol{\eta}$ is $\ell(\boldsymbol{\eta})=\ln[L(\boldsymbol{\eta})]=\sum_{i=1}^{n} \ell_i(\boldsymbol{\eta})$, which is given as follows:

$$\begin{aligned}
\ell(\boldsymbol{\eta}) &= \ln[L(\boldsymbol{\eta})] \\
&= \ln \prod_{i=1}^{n}\left[\left(\frac{e^{\eta^T \mathrm{X}_i}}{1+e^{\eta^T \mathrm{X}_i}}\right)^{Y_i}\left(1-\frac{e^{\eta^T \mathrm{X}_i}}{1+e^{\eta^T \mathrm{X}_i}}\right)^{1-Y_i}\right] \\
&= \sum_{i=1}^{n} Y_i \ln\left(\frac{e^{\eta^T \mathrm{X}_i}}{1+e^{\eta^T \mathrm{X}_i}}\right)+\sum_{i=1}^{n}(1-Y_i) \ln\left(\frac{1}{1+e^{\eta^T \mathrm{X}_i}}\right) \\
&= \sum_{i=1}^{n} Y_i\left[\ln(e^{\eta^T \mathrm{X}_i})-\ln(1+e^{\eta^T \mathrm{X}_i})\right]-\sum_{i=1}^{n}(1-Y_i)\left[\ln(1+e^{\eta^T \mathrm{X}_i})\right] \\
&= \sum_{i=1}^{n} Y_i\left[(\boldsymbol{\eta}^T \mathrm{X}_i)-\ln(1+e^{\eta^T \mathrm{X}_i})\right]-\sum_{i=1}^{n}(1-Y_i)\left[\ln(1+e^{\eta^T \mathrm{X}_i})\right] \\
&= \sum_{i=1}^{n}\left[Y_i(\boldsymbol{\eta}^T \mathrm{X}_i)-\ln(1+e^{\eta^T \mathrm{X}_i})\right].
\end{aligned}$$

The estimating score function for the complete data is stated below:

$$\boldsymbol{U}_{F,n}(\boldsymbol{\eta}) = \frac{1}{\sqrt{n}} \frac{\partial \ell(\boldsymbol{\eta})}{\partial \, \boldsymbol{\eta}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{X}_i \left(Y_i - \frac{e^{\eta^T \mathrm{X}_i}}{1+e^{\eta^T \mathrm{X}_i}} \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{S}_i(\boldsymbol{\eta}), \tag{2}$$

where

$$\boldsymbol{S}_i(\boldsymbol{\eta}) = \mathrm{X}_i \left(Y_i - \frac{e^{\eta^T \mathrm{X}_i}}{1+e^{\eta^T \mathrm{X}_i}} \right). \tag{3}$$

As $E[\boldsymbol{U}_{F,n}(\boldsymbol{\eta})] = \mathbf{0}$, $\boldsymbol{U}_{F,n}(\boldsymbol{\eta})$ is an unbiased score function, we can obtain the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\eta}}_F$ of $\boldsymbol{\eta}$ by addressing $\boldsymbol{U}_{F,n}(\boldsymbol{\eta}) = \mathbf{0}$.

3. Large-Sample Properties

We review the asymptotic properties of the MLE of the logistic model in this section. It is known that the MLE has an asymptotic normal distribution, with a mean and asymptotic standard deviation. Nevertheless, we have approached a new, detailed and straightforward proof, as given below. The proof is provided in detail in the Appendix, which gives a new theoretical contribution to the paper.

The asymptotic property of $\hat{\boldsymbol{\eta}}_F$ is developed under some regularity conditions, which are much the same as those given in Lee et al. (2021). Recall that:

$$U_{F,n}(\boldsymbol{\eta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\boldsymbol{\eta}),$$

where $S_i(\boldsymbol{\eta})$ is mentioned in (3).

The regularity conditions are described as follows:

(C1) $E\left[(S_1(\boldsymbol{\eta}))^{\otimes 2}\right]$ **is positive definite in a neighbourhood of the true** η **. For any column vector a, we define** $a^{\otimes 2} = aa^T$ **.**

(C2) The first derivatives of $U_{F,n}(\boldsymbol{\eta})$ **with respect to** η **exist almost surely in a neighbourhood of the true** η **. In addition, in such a neighbourhood, the first derivatives are uniformly bounded above by a function of** (Y, X, Z) **, whose expectations exist.**

Before the asymptotic property of $\hat{\boldsymbol{\eta}}_F$ is presented, consider:

$$G_{F,n}(\boldsymbol{\eta}) = -\frac{1}{\sqrt{n}} \frac{\partial U_{F,n}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial S_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T},$$

As every constituent of $G_{F,n}(\boldsymbol{\eta})$ is a mean of independent and identically distributed random variables, we have:

$$E[G_{F,n}(\boldsymbol{\eta})] = \frac{1}{n} \sum_{i=1}^{n} E\left\{ E\left[\left(\frac{-\partial S_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T} \right) \middle| Y_i, X_i, Z_i \right] \right\}$$

$$= E\left\{ E\left[\left(\frac{-\partial S_1(\boldsymbol{\eta})}{\partial\, \boldsymbol{\eta}^T} \right) \Big| Y_1, X_1, Z_1 \right] \right\} = E\left[\frac{-\partial S_1(\boldsymbol{\eta})}{\partial\, \boldsymbol{\eta}^T} \right] = G_F(\boldsymbol{\eta}).$$

Performing the weak law of large numbers (WLLN), it can be proved that $G_{F,n}(\boldsymbol{\eta}) \xrightarrow{p} G_F(\boldsymbol{\eta})$.

Let $\hat{\boldsymbol{\eta}}_F$ be the root of $U_{F,n}(\boldsymbol{\eta}) = 0$. The asymptotic property of $\hat{\boldsymbol{\eta}}_F$ is introduced in Theorem 1. The detailed proof of Theorem 1 is displayed in **Appendix**.

Theorem 1

It is presumed that conditions (C1) and (C2) satisfied. As $n \to \infty$, then $\hat{\boldsymbol{\eta}}_F$ is a consistent estimator of $\boldsymbol{\eta}$ and $\sqrt{n}(\hat{\boldsymbol{\eta}}_F - \boldsymbol{\eta})$ is asymptotically normally distributed with mean 0 and covariance matrix Δ_F, where $\Delta_F = G_F^{-1}(\boldsymbol{\eta}) Q_F(\boldsymbol{\eta}) [G_F^{-1}(\boldsymbol{\eta})]^T$ and $Q_F(\boldsymbol{\eta}) = E\left[(S_1(\boldsymbol{\eta}))^{\otimes 2} \right]$.

The detailed calculations of $G_F(\boldsymbol{\eta})$ and $Q_F(\boldsymbol{\eta})$ are illustrated in the **Appendix**.

Aa the variance of the score function is also the inverse of the Fisher information matrix, then $G_F(\boldsymbol{\eta}) = Q_F(\boldsymbol{\eta})$ is a recipe of Δ_F, which is described as $\Delta_F = G_F^{-1}(\boldsymbol{\eta})$.

4. Simulation Study

In this section, Monte Carlo experiments are used to investigate the finite sample performance of the MLE $\hat{\boldsymbol{\eta}}_F$, that is, the root of the equations in (2). We consider two cases such that X can be both univariate and bivariate, whereas Z is only univariate. In the current framework, the sample size considered was n = 500; 1,000; 1,500; 2,000 and 2,500. The number of replications is 2,000. We then calculated the bias, asymptotic standard error (ASE), standard deviation (SD), and coverage probability (CP).

The examples are presented sequentially.

4.1 Case 1: Both X and Z are univariate

We used $N(0,1)$ and $Ber(1, 0.4)$ to generate the data of X and Z, respectively. The data of Y were created by relying on the formula given in (1). The simulation results in this case are given in Table 1.

As seen Table 1, the bias is very small. Specifically, when the sample size was increased, the SDs and ASEs were decreased. The empirical CPs overall were close to the nominal probability of 0.95. Based on these findings, it can be strongly concluded that the results obtained in the simulation are very reliable.

4.2 Case 2: X is bivariate and Z is univariate

We considered $\boldsymbol{X} = (X_1, X_2)^T$ is bivariate and Z is univariate, and used $N(\mu, \boldsymbol{\Sigma})$ to generate the data of $\boldsymbol{X}$, where:

$$\mu = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.4\sqrt{2} \\ 0.4\sqrt{2} & 2 \end{bmatrix}.$$

The data of Z were generated from the distribution that had the three different values of $-0.5, 0.5$ and 1, with probabilities 0.3, 0.3 and 0.4, respectively. The data of Y were generated as the first case. The simulation results in this situation are given in Table 2.

Similar to the first case, as seen Table 2 the bias is very small. When the sample size was increased, the SDs and ASEs were decreased. Furthermore, the empirical CPs overall were close to the nominal probability of 0.95. From the simulation results in both cases, these results demonstrate that the MLE is a reliable and accurate estimator in finite samples.

We present this simulation study for the main purpose of checking the working efficiency of the MLE method for the logistic model. Based on the findings above, it can be concluded that this method works very well for the logistic model. In the next sections, we use the MLE method to investigate practical application with the fishing data set. In order to emphasize the enormous and widespread application of the logistic model in practice, we presented a number of illustrative examples from some disciplines in the sciences and social sciences.

5. Practical Applications

The fishing data set that is performed in this section can be found at the following website: https://github.com/rmcelreath/rethinking/blob/master/data/Fish.csv.

This data set includes 250 groups that were surveyed. Each group was queried how many fish they had caught (***fish_caught***), whether or not they had used live bait (***livebait***), whether or not they had brought a camper to the park (***camper***), how many people were in the group (***persons***), how many children were in the group (***child***), and how long they stayed (***hours***).

We should also note that the sets of values of the variable Y are a count data set, and we would like to perform a logistic regression model to study the data set. It is assumed that if the angler caught the fish, the value of Y was denoted as 1, otherwise 0. Of the 250 groups that was surveyed, 108 groups (43.2%) had caught fish. The diagram of the Y variable is displayed in Figure 1.

As shown in Figure 1, the values of the outcome variable Y are only 0 and 1. Here 0 indicates there were no fish caught when fishing, or did not go fishing, while 1 represents whether any fish was caught, but we do not care about the number of fish actually caught. As the data set only has the values of 0 and 1, the logistic regression model is used to study the relationship between the binary outcome variable and covariates.

Details on the number of variables are given in Figure 2. As the fishing times of the 250 surveyed groups were very different, we cannot describe these in detail. Interested readers can see the link mentioned above. For the example presented here, we conducted a logistic regression model to study the association of ***fish_caught*** as a binary outcome variable (Y), where the independent covariates that were investigated for incorporation into the final model included ***livebait***, ***camper***, ***persons***, ***child*** and ***hours***. The results for this analysis are provided in Figure 3.

As observed Figure 3, the estimated results of all the variables examined in the analysis are statistically significant, except for the estimated results of ***hours***. The estimated results of the three variables of ***camper***, ***livebait*** and ***persons*** are 0.83, 1.07 and 1.16, respectively. These values are all positive, indicating that, in increasing a unit of these variables, the ability to catch fish also increased. That means the number of fish collected in fishing will increase accordingly.

It should be remarked that, of the three positive values, the estimated result of the ***persons*** variable is the greatest, and the estimated result of the ***camper*** variable is the smallest. It means that, if the same increases were one unit in the three variables of ***camper***, ***livebait*** and ***persons***, then the probability of catching a fish increases one unit in the ***persons*** variable is the highest, followed by the ***livebait*** variable, then the ***camper*** variable. In short, increasing a unit in the ***persons*** variable is such that the number of fish collected accounts for the highest proportion.

Furthermore, it has been seen that the estimated result of the ***child*** variable is -2.08. This value is negative, indicating that if adding a unit in the ***child*** variable, the amount of fish collected is not affected, and may even cause a reduction in the results. Hence, it is better not to increase the number of this variable if the purpose is to catch fish.

6. Applications in Decision Sciences

We next turn to discuss applications of the logistic model in the banking sector, companies and business, and medical research. We present numerous different applications of the logistic model in these and related fields. It is discussed with the aim to provide an overview of the practicality of the logistic model. This is also a meaningful and complete reference for the applications of the logistics model in a wide range of disciplines in the sciences and social sciences.

6.1. Banking Sector

The first application is in the banking sector, where the logistic model is often used to assess credit risk. It is assumed that there is a bank employee in charge of the loan segment, who wants to identify the characteristics of customers with potential (inability to repay) debts after borrowing. These signals will then be used to determine whether credit risk is a good or bad decision for each customer.

More specifically, assume that the bank has data for 900 customers, of which 650 have received the loan. It is then possible to use a random sample of these 650 customers to generate a logistic regression model and to classify the remaining 250 customers at (possible) risk of bad debt, based on the database of 650 customers, some of whom have paid off the entire loan, and some have not yet paid off the loan.

Moreover, in the banking sector, in order to prevent the risks, the logistic model can also help increase profits by helping the bank to reach the right customers with the appropriate products and services. Specifically, on this issue are the following considerations: a bank wants to build a regression model for the purpose of estimating or forecasting whether or not to issue a credit card to any particular customer. The target variable is a binary variable consisting of only two values, namely 0 and 1.

If the customer has not been granted a credit card, then the target variable value is 0, otherwise it will be1. Meanwhile, a set of independent variables can be considered, such as age, income, assets, occupation, employment status, and so on. In using a logistic model, it is possible to predict the ability or probability to issue a credit card to any customer. Some interesting studies on this topic can be found in Zaghdoudi (2013), Serener (2016), and Pho et al. (2019), among others.

6.2. Companies and Businesses

The second most important application can be mentioned is in companies, businesses, and business establishments. Logistic models are often used in predicting the likelihood of an event or situation occurring in the future. For example, a company wants to know the ability of customers to access the website, with the target variable being to choose or not to choose the company's offers/promotions. The following characteristics or attributes are used to consider them as independent variables: other websites that they visited before visiting the company's website, how often they visited on the website of company, or selected actions on the company's website.

Hence, the logistic model can be used to determine the likelihood that the type of customer who visited the website of company is likely to choose an offer/promotion, or not. From that it can be seen that the company will be able to make better decisions about the company's incentive advertising strategies, or create better policies for these incentives.

A logistic model could also be executed in building predictive models for companies. Therefore, it is seen as an effective method to make a difference in its competitive advantage. Predictive models can help companies to fully exploit relationships, which will have an impact on sales and profits in the future, through a better understanding o customer behaviour. From there, the company will be able to construct more effective decisions and strategies. A typical example of how a logistic model could be used to predict the potential failure of a machine's components or parts based on how long they are stored in stock.

Thus, we can see that with the obtained results from the logistic model analysis process, a company can make its maintenance and installation plans more reasonable and efficient. Furthermore, we can also use the logistic model to predict the likelihood of customers leaving the service, and customer segmentation by target production group based on purchasing characteristics and personal information. Some excellent research on this can be found in Sandberg et al. (2011) and Shrivastava et al. (2018), among others.

6.3. Medical Sector

Another important area where logistic models can be used lies in the medical sector. Logistic models can be used to predict the possibility of a certain disease of a population group in a particular country, region, state, province, or city. If there is any disease, we denote the outcome variable as 1, and otherwise 0. As the outcome variable only takes the values 0 and 1, the logistic model is perfectly suitable for analysis and research for this data set.

Needless to say, an exceedingly concrete example in today's world is the COVID-19 pandemic. This issue is a tremendously painful and complicated problem in the current period as every country and territory in the world is facing this pandemic. Effective vaccines are still being considered and studied, so the way to prevent and predict the disease is very important and meaningful at this stage.

A noteworthy research for this problem is shown in Wang et al. (2020), who used the logistic model to predict the epidemic trends in COVID-19. From there, the president or head of that county or district, will make the appropriate policies or decisions to take preventative measures. It can be seen that the application of the logistic model in reality is extremely rich and diverse. Furthermore, for extensive literature reviews in this regard see, for example, Royston and Altman (2010), Ugwuanyim et al. (2020), and McAleer (2021a, 2021b), among others.

We should also pay attention that, albeit the logistic model's result variable only accepts binary values of 0 and 1, in practice many events may take these two values. In addition, as in the first paragraph of this section, we also mentioned that several count data sets can be expanded to become binary data sets if we do not care how many times the event happened, but only care whether it happened or not.

6.4. Missing data

It is often seen in fact that data sets, which we derive from some survey or from some data source, often contain missing values. This issue is frequently encountered in the areas of, for instance, medical research, economics, finance, accounting, education, and transportation. Hence, it is necessary to have a suitable parameter estimation method for regression models, specifically for the logistic model.

Generally speaking, the problems related to missing data can normally be classified into two different categories, namely missing outcomes and missing covariates. To the best of our knowledge, up to the present time, there have been substantial research examining both the theory and application to this problem. We are mainly interested in the parameter estimation methods for the logistic model in the case of missing data.

Some prominent studies may be noted as in Wang et al. (2002), who considered the joint conditional likelihood (JCL) method to estimate the parameters in the logistic model in the case of missing covariates. This method was later adopted by Lee et al. (2012), who expanded it to develop the semiparametric estimation of the logistic model with missing covariates and outcomes. It was later adopted by Hsieh et al. (2013), who examined logistics models with outcomes and covariates that were missing separately or simultaneously.

Based on these studies, we can see that the JCL method is known to be a powerful method for estimating the parameters for the logistic model in the case of survey data sets containing missing values. Furthermore, in these studies, the authors also argued that the JCL method worked more effectively than other common methods, such as the inverse probability weighting (IPW) or the complete cases (CC) methods.

When applying alternative methods to solve problems containing missing data, it should be noted if the missing ratio in the data set is missing low, medium or high as the methods for solving this problem may be more suitable for different data sets according to the missing ratio. Therefore, it can be seen that there are several alternative methods in order to choose the most suitable and appropriate method to deal with missing data.

6.5. Other Applications

Furthermore, we should also pay attention that the logistic model is only applied to regression analysis for data sets where the outcome variable only takes binary values. In this work, we considered the fishing data set as applied in regression analysis. Actually, this data set is a count data set but, in the current paper, the number of fish collected is not important, but only that the fish are caught or not.

In practice, we do not care how many times the event has occurred, but only if the events occurred or not. If the event occurs, the outcome variable Y is set to 1, otherwise it is set to 0. When several count data sets can be expanded to become data sets, the outcome variable Y is only given binary values. From there, we can use the logistic model for the subsequent analysis.

7. Concluding Remarks and Future Research

This paper is meaningful in that there is a theoretical contribution by providing an explicit and straightforward derivation of the logistic model. We summarized the formula, concerned functions, and estimating equation of the logistic model. Theory of large sample properties for this model under some regularity conditions was considered. The detailed proof is provided in the **Appendix**.

To the best of our knowledge, so far there have been no detailed and complete presentations regarding the related formulas of this model. Therefore, some of its inferences may not be straightforward for those who have not majored in mathematics. Thus, this paper may make a useful contribution to the literature.

A simulation study and a factual application are also investigated in the paper. We estimated the parameters of a logistic model with a practical fishing data set. The results that we have obtained in the current paper are significant in practice. It helps us to better understand how to increase the number of fish collected while fishing.

We also discussed numerous applications of the logistic model in important disciplines in the sciences and social sciences, with the aims of providing an overview of the issue. It can be observed that the practical applications of the logistic model are extremely diverse and abundant. From these detailed and complete discussions, we can apply the logistic model in a variety of future studies. This can assist banks, companies, and businesses to predict future situations, help to obtain higher profits, and a range of related decisions in business and finance. These are also potentially useful and meaningful research directions that can be researched and developed in the near future.

Furthermore, it should be noted that the studies on parameter estimation for the logistic model with missing data mentioned in the previous section are only research in missing in the random (MAR) case. This can be extended to consider other types of missing data, such as missing completely at random (MCAR), and also missing not at random (MNAR). We could also examine the problem by incorporating the randomized response technique, or goodness of fit tests. Other variations of the logistic model with missing data are given in Hsieh et al. (2009, 2010).

References

Albert, A., and Anderson, J.A. (1984), On the existence of maximum likelihood estimates in logistic regression models, *Biometrika*, 71(1), 1-10.

Allison, P.D. (2012), Logistic regression using SAS: Theory and application, *SAS institute*.

Austin, P.., and Merlo, J. (2017), Intermediate and advanced topics in multilevel logistic regression analysis, *Statistics in Medicine*, 36(20), 3257-3277.

Chang, P.C., Pho, K.H., Lee, S.M., and Li, C.S. (2021), Estimation of parameters of logistic regression for two-stage randomized response technique, *Computational Statistics*, 1-23.

Cox, D.R. (1958), The regression analysis of binary sequences, *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215-232.

DeMaris, A. (1995), A tutorial in logistic regression, *Journal of Marriage and the Family*, 956-968.

Enea, M., and Lovison, G. (2019), A penalized approach for the bivariate ordered logistic model with applications to social and medical data, *Statistical Modelling*, 19(5), 467-500.

Foutz, R.V. (1977), On the unique consistent solution to the likelihood equations, *Journal of the American Statistical Association*, 72(357), 147-148.

Hosmer, D.W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997), A comparison of goodness-of-fit tests for the logistic regression model, *Statistics in Medicine*, 16(9), 965-980.

Hosmer, D.W., and Lemesbow, S. (1980), Goodness of fit tests for the multiple logistic regression model, *Communications in Statistics-Theory and Methods*, 9(10), 1043-1069.

Hosmer, D.W., Lemeshow, S., and Sturdivant, R.X. (2013), Applied logistic regression (Vol. 398), *Wiley*.

Hsieh, S., Lee, S.M., and Shen, P.S. (2010), Logistic regression analysis of randomized response data with missing covariates, *Journal of Statistical Planning and Inference*, 140(4), 927-940.

Hsieh, S.H., Lee, S.M., and Shen, P.S. (2009), Semiparametric analysis of randomized response data with missing covariates in logistic regression, *Computational Statistics and Data Analysis*, 53(7), 2673-2692.

Hsieh, S.H., Li, C.S., and Lee, S.M. (2013), Logistic regression with outcome and covariates missing separately or simultaneously, *Computational Statistics and Data Analysis*, 66, 32-54.

Lee, S.M., Li, C.S., Hsieh, S.H., and Huang, L.H. (2012), Semiparametric estimation of logistic regression model with missing covariates and outcome, *Metrika*, 75(5), 621-653.

Lee, S.M., Pho, K.H., and Li, C.S. (2021), Validation likelihood estimation method for a zero-inflated Bernoulli regression model with missing covariates, J*ournal of Statistical Planning and Inference*.

McAleer, M. (2021a), A critique of recent medical research in JAMA on COVID-19, *Advances in Decision Sciences*, 25(1), 40-142.

McAleer, M. (2021b), A critical analysis of some recent medical research in Science on COVID-19, *Advances in Decision Sciences*, 25(1), 216-332.

Menard, S. (2002), Applied logistic regression analysis (Vol. 106), *Sage*.

Pho, K.H., Ly, S., Ly, S., and Lukusa, T.M. (2019), Comparison among Akaike information criterion, Bayesian information criterion and Vuong's test in model selection: A case study of violated speed regulation in Taiwan, *Journal of Advanced Engineering and Computation*, 3(1), 293-303.

Royston, P., and Altman, D.G. (2010), Visualizing and assessing discrimination in the logistic regression model, *Statistics in Medicine*, 29(24), 2508-2520.

Sandberg, E., Kihlén, T., and Abrahamsson, M. (2011), Characteristics of a logistics-based business model, *Journal of Marketing Channels*, 18(2), 123-145.

Serener, B. (2016), Statistical analysis of internet banking usage with logistic regression, *Procedia Computer Science*, 102, 648-653.

Shrivastava, A., Kumar, K., and Kumar, N. (2018), Business distress prediction using bayesian logistic model for indian firms, *Risks*, 6(4), 113.

Ugwuanyim, G.U., Osuchukwu, C.O., Bartholomew, D.C., and Obite, C.P. (2020), Medical choices for a wealthy nation - A multinomial logistic model, *Asian Journal of Probability and Statistics*, 1-12.

Wang, C.Y., Chen, J.C., Lee, S.M., and Ou, S.T. (2002), Joint conditional likelihood estimator in logistic regression with missing covariate data, *Statistica Sinica*, 555-574.

Wang, P., Zheng, X., Li, J., and Zhu, B. (2020), Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics, *Chaos, Solitons and Fractals*, 139, 110058.

Zaghdoudi, T. (2013), Bank failure prediction with logistic regression, I*nternational Journal of Economics and Financial Issues*, 3(2), 537.

Appendices

Table 1

Simulation results of X (univariate) with $\boldsymbol{\eta}_1 = (\eta_0, \eta_1, \eta_2)^T = (0.7, -1.7, 0.5)^T$ and $\boldsymbol{\eta}_2 = (\eta_0, \eta_1, \eta_2)^T = (-0.7, -1.3, -0.5)^T$

Sample size		$P(Y=0)=0.36$					$P(Y=0)=0.6$				
		500	1,000	1,500	2,000	2,500	500	1,000	1,500	2,000	2,500
η_0	Bias	0.0052	0.0050	0.0037	0.0010	0.0007	-0.0014	-0.0035	-0.0044	-0.0006	-0.0017
	ASE	0.1506	0.1060	0.0865	0.0748	0.0669	0.1350	0.0952	0.0777	0.0672	0.0601
	SD	0.1500	0.1057	0.0865	0.0746	0.0679	0.1342	0.0951	0.0758	0.0655	0.0618
	CP	0.9560	0.9530	0.9540	0.9540	0.9460	0.9520	0.9500	0.9550	0.9600	0.9440
η_1	Bias	-0.0213	-0.0101	-0.0098	-0.0069	-0.0053	0.0118	0.0064	0.0061	0.0031	0.0035
	ASE	0.1672	0.1175	0.0958	0.0828	0.0740	0.1212	0.0853	0.0696	0.0601	0.0538
	SD	0.1702	0.1156	0.0966	0.0809	0.0740	0.1227	0.0823	0.0700	0.0597	0.0534
	CP	0.9510	0.9560	0.9540	0.9670	0.9500	0.9480	0.9580	0.9480	0.9520	0.9510
η_2	Bias	0.0058	0.0001	0.0052	0.0053	0.0013	-0.0009	0.0053	0.0017	0.0003	-0.0001
	ASE	0.2387	0.1680	0.1371	0.1187	0.1060	0.2064	0.1454	0.1187	0.1027	0.0918
	SD	0.2345	0.1657	0.1365	0.1176	0.1048	0.2034	0.1410	0.1191	0.1023	0.0929
	CP	0.9600	0.9510	0.9540	0.9470	0.9480	0.9540	0.9540	0.9540	0.9480	0.9450

The average observed percentages of $Y = 0$ was approximately 36% and 60% with respect to $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$.

Table 2

Simulation results of $\boldsymbol{X}=(X_1,X_2)^T$ (bivariate) with

$\boldsymbol{\eta}_3=(\eta_0,\eta_1,\eta_2,\eta_3)^T=(0.7,0.8,0.7,0.5)^T$ **, and** $\boldsymbol{\eta}_4=(\eta_0,\eta_1,\eta_2,\eta_3)^T=(-0.7,0.8,0.7,-0.5)^T$

Sample size		$P(Y=0)=0.36$					$P(Y=0)=0.6$				
		500	1,000	1,500	2,000	2,500	500	1,000	1,500	2,000	2,500
η_0	Bias	0.0084	0.0051	0.0005	0.0012	0.0019	-0.0075	-0.0037	0.0006	-0.0060	-0.0026
	ASE	0.1420	0.0999	0.0813	0.0704	0.0629	0.1442	0.1014	0.0824	0.0714	0.0638
	SD	0.1435	0.1012	0.0828	0.0705	0.0633	0.1487	0.1019	0.0835	0.0701	0.0627
	CP	0.9480	0.9480	0.9440	0.9540	0.9490	0.9470	0.9520	0.9460	0.9520	0.9560
η_1	Bias	0.0209	0.0108	0.0012	0.0045	0.0051	0.0260	0.0157	0.0052	0.0078	0.0035
	ASE	0.1556	0.1091	0.0886	0.0768	0.0687	0.1723	0.1209	0.0981	0.0849	0.0759
	SD	0.1566	0.1109	0.0872	0.0762	0.0673	0.1728	0.1215	0.0992	0.0844	0.0752
	CP	0.9500	0.9460	0.9500	0.9510	0.9560	0.9530	0.9480	0.9500	0.9560	0.9480
η_2	Bias	0.0111	0.0062	0.0052	0.0049	0.0019	0.0137	0.0064	0.0026	0.0015	0.0033
	ASE	0.1043	0.0732	0.0596	0.0516	0.0460	0.1068	0.0748	0.0608	0.0526	0.0470
	SD	0.1062	0.0730	0.0593	0.0531	0.0466	0.1071	0.0758	0.0606	0.0532	0.0464
	CP	0.9460	0.9480	0.9540	0.9430	0.9480	0.9510	0.9500	0.9510	0.9520	0.9520
η_3	Bias	0.0062	0.0062	0.0045	0.0061	0.0009	-0.0018	-0.0043	-0.0041	0.0017	-0.0005
	ASE	0.1888	0.1328	0.1081	0.0936	0.0836	0.1941	0.1366	0.1111	0.0961	0.0859
	SD	0.1938	0.1342	0.1087	0.0925	0.0832	0.1971	0.1384	0.1122	0.0952	0.0859
	CP	0.9440	0.9480	0.9440	0.9500	0.9520	0.9470	0.9480	0.9460	0.9520	0.9470

The average observed percentages of $Y=0$ was approximately 36% and 60% with respect to $\boldsymbol{\eta}_3$ and $\boldsymbol{\eta}_4$.

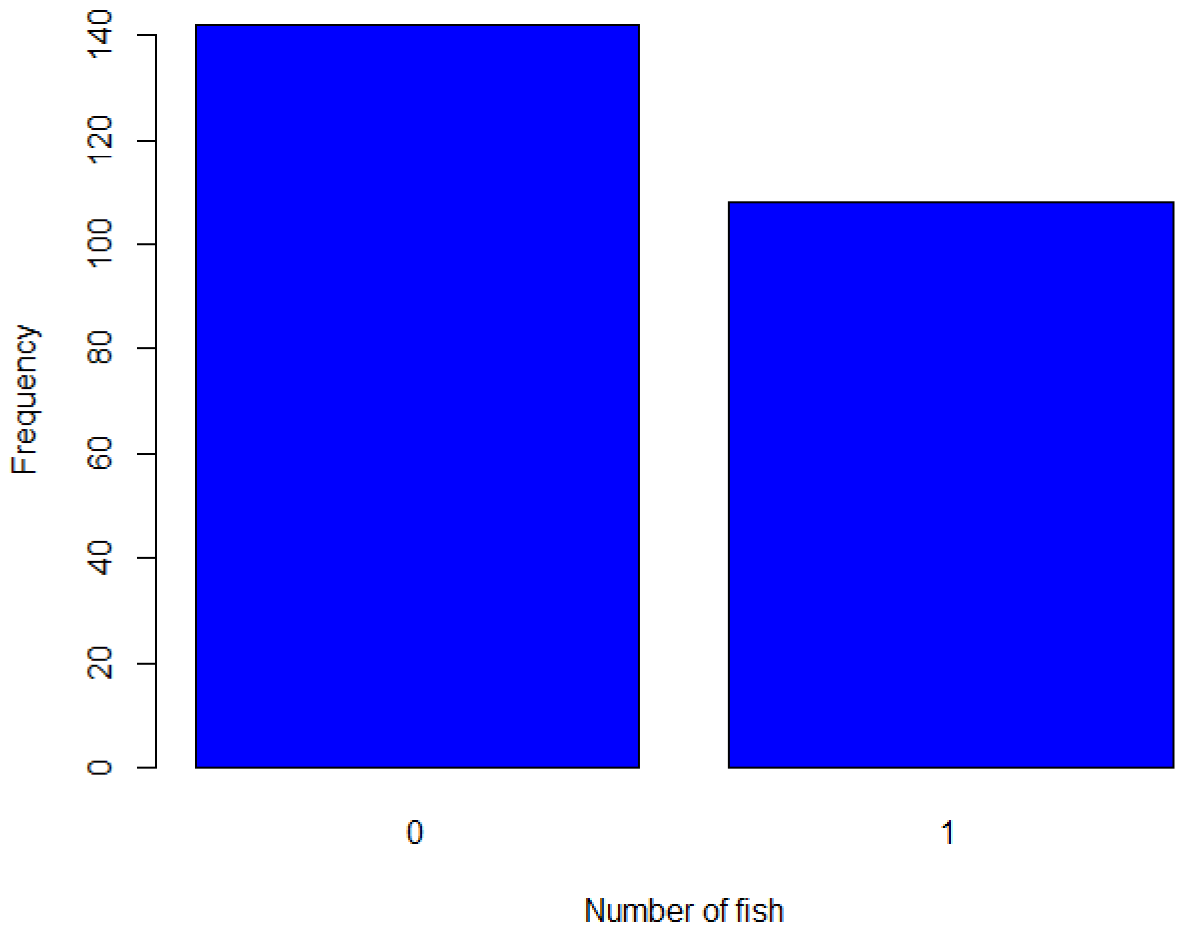

Figure 1

Diagram of Y

```
> table(fish_caught)      ### The number of fish_caught
fish_caught
  0   1
142 108
> table(livebait)         ### The number of livebait
livebait
  0   1
 34 216
> table(camper)           ### The number of camper
camper
  0   1
103 147
> table(persons)          ### The number of persons
persons
 1  2  3  4
57 70 57 66
> table(child)            ### The number of child
child
  0   1   2   3
132  75  33  10
```

Figure 2

Variables of the fishing data set

```
> summary(glm(fish_caught~ 1 + livebait + camper + persons +
+ child + hours,family="binomial"))

Call:
glm(formula = fish_caught ~ 1 + livebait + camper + persons +
    child + hours, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1942  -0.7594  -0.3062   0.7723   2.1456

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.44496    0.69277  -4.973 6.60e-07 ***
livebait     1.07863    0.50022   2.156   0.0311 *
camper       0.83171    0.33910   2.453   0.0142 *
persons      1.16270    0.19998   5.814 6.10e-09 ***
child       -2.08065    0.33606  -6.191 5.97e-10 ***
hours        0.02962    0.02198   1.348   0.1778
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 341.94  on 249  degrees of freedom
Residual deviance: 245.78  on 244  degrees of freedom
AIC: 257.78

Number of Fisher Scoring iterations: 5
```

Figure 3

Logistic regression analysis with the fishing data set

Proof that $G_F(\boldsymbol{\eta}) = Q_F(\boldsymbol{\eta})$ **.**

We have:

$$\boldsymbol{S}_i(\boldsymbol{\eta}) = \mathrm{X}_i\left(Y_i - \frac{e^{\eta^T \mathrm{X}_i}}{1+e^{\eta^T \mathrm{X}_i}}\right),$$

so that:

$$\frac{\boldsymbol{S}_i(\boldsymbol{\eta})}{\partial\ \boldsymbol{\eta}} = -\frac{\mathrm{X}_i^{\otimes 2} e^{\eta^T \mathrm{X}_i}(1+e^{\eta^T \mathrm{X}_i}) - \mathrm{X}_i^{\otimes 2} e^{\eta^T \mathrm{X}_i} e^{\eta^T \mathrm{X}_i}}{(1+e^{\eta^T \mathrm{X}_i})^2} = -\frac{\mathrm{X}_i^{\otimes 2} e^{\eta^T \mathrm{X}_i}}{(1+e^{\eta^T \mathrm{X}_i})^2},$$

and

$$\begin{aligned}
[S_1(\boldsymbol{\eta})]^{\otimes 2} &= \mathrm{X}_1^{\otimes 2}\left(Y_1 - \frac{e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\right)^2 \\
&= \mathrm{X}_1^{\otimes 2}\left[Y_1^2 - 2Y_1\left(\frac{e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\right) + \left(\frac{e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\right)^2\right] \\
&= \mathrm{X}_1^{\otimes 2}\left[Y_1 - 2Y_1\left(\frac{e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\right) + \left(\frac{e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\right)^2\right] \\
&= \mathrm{X}_1^{\otimes 2}\left[Y_1\left(1 - \frac{2e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\right) + \left(\frac{e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\right)^2\right] \\
&= \mathrm{X}_1^{\otimes 2}\left[Y_1\left(\frac{1-e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\right) + \left(\frac{e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\right)^2\right] \\
&= \mathrm{X}_1^{\otimes 2}\left[\left(Y_1 - \frac{e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\right)\frac{1-e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}} + \frac{e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\,\frac{1-e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}} + \left(\frac{e^{\eta^T \mathrm{X}_1}}{1+e^{\eta^T \mathrm{X}_1}}\right)^2\right]
\end{aligned}$$

$$= X_1^{\otimes 2}\left[\left(Y_1 - \frac{e^{\eta^T X_1}}{1+e^{\eta^T X_1}}\right)\frac{1-e^{\eta^T X_1}}{1+e^{\eta^T X_1}}\right] + X_1^{\otimes 2}\left[\frac{e^{\eta^T X_1}}{1+e^{\eta^T X_1}}\frac{1-e^{\eta^T X_1}}{1+e^{\eta^T X_1}} + \left(\frac{e^{\eta^T X_1}}{1+e^{\eta^T X_1}}\right)^2\right]$$

$$= X_1^{\otimes 2}\left[\left(Y_1 - \frac{e^{\eta^T X_1}}{1+e^{\eta^T X_1}}\right)\frac{1-e^{\eta^T X_1}}{1+e^{\eta^T X_1}}\right] + X_1^{\otimes 2}\frac{e^{\eta^T X_1}}{1+e^{\eta^T X_1}}.$$

Thereafter, we obtain:

$$G_F(\boldsymbol{\eta}) = E\left[-\frac{\partial S_1(\boldsymbol{\eta})}{\partial\ \boldsymbol{\eta}^T}\right] = E\left[\frac{X_i^{\otimes 2} e^{\eta^T X_i}}{(1+e^{\eta^T X_i})^2}\right] = E\left[[S_1(\boldsymbol{\eta})]^{\otimes 2}\right] = Q_F(\boldsymbol{\eta}).$$

Proof of Theorem 1

In order to illustrate that $\hat{\boldsymbol{\eta}}_F$ is a consistent estimator of $\boldsymbol{\eta}$, it is repeated as follows:

$$U_{F,n}(\boldsymbol{\eta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\boldsymbol{\eta})$$

$$G_{F,n}(\boldsymbol{\eta}) = -\frac{1}{\sqrt{n}}\frac{\partial U_{F,n}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial S_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T}.$$

Performing the WLLN, then $G_{F,n}(\boldsymbol{\eta}) \xrightarrow{p} G_F(\boldsymbol{\eta})$,and by using Condition $(C2)$, the convergence of $G_{F,n}(\boldsymbol{\eta})$ to $G_F(\boldsymbol{\eta})$ is uniform in a neighborhood of the true $\boldsymbol{\eta}$.

Executing the inverse function theorem of Foutz (1977), it can be demonstrated that a unique consistent root of $U_{F,n}(\boldsymbol{\eta}) = 0$ subsists in a neighborhood of the true.

Thereafter, we obtain that $\hat{\boldsymbol{\eta}}_F$ **is a consistent estimator of** $\boldsymbol{\eta}$ **.**

We next consider the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\eta}}_F - \boldsymbol{\eta})$.

As $\hat{\boldsymbol{\eta}}_F$ is a unique root of $U_{F,n}(\boldsymbol{\eta}) = 0$, using a Taylor's expansion of $U_{F,n}(\hat{\boldsymbol{\eta}}_F)$ at $\boldsymbol{\eta}$ gives:

$$0 = U_{F,n}(\hat{\boldsymbol{\eta}}_F) = U_{F,n}(\boldsymbol{\eta}) + \left(\frac{1}{\sqrt{n}}\frac{\partial U_{F,n}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T}\right)\sqrt{n}(\boldsymbol{\eta}_F - \boldsymbol{\eta}) + o_p(1),$$

so that:

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\eta}}_F - \boldsymbol{\eta}) &= G_{F,n}^{-1}(\boldsymbol{\eta})U_{F,n}(\boldsymbol{\eta}) + o_p(1)\\ &= G_F^{-1}(\boldsymbol{\eta})U_{F,n}(\boldsymbol{\eta}) + [G_{F,n}^{-1}(\boldsymbol{\eta}) - G_F^{-1}(\boldsymbol{\eta})]U_{F,n}(\boldsymbol{\eta}) + o_p(1)\end{aligned}$$

where $i = 1, 2, ..., n, E[S_i(\boldsymbol{\eta})] = 0$ and $Var[S_i(\boldsymbol{\eta})] = E\left[[S_i(\boldsymbol{\eta})]^{\otimes 2}\right]$.

Executing the central limit theorem and Condition $(C2)$ gives:

$$U_{F,n}(\boldsymbol{\eta}) \xrightarrow{d} N(0, Q_F(\boldsymbol{\eta})),$$

where $Q_F(\boldsymbol{\eta}) = Var[U_{F,n}(\boldsymbol{\eta})]$. As $G_{F,n}(\boldsymbol{\eta}) - G_F(\boldsymbol{\eta}) \xrightarrow{p} 0$ by the WLLN, and $G_F(\boldsymbol{\eta})$ is nonsingular by Condition $(C2)$, we apply Slutsky's theorem to obtain:

$$\left[G_{F,n}^{-1}(\boldsymbol{\eta}) - G_F^{-1}(\boldsymbol{\eta})\right] U_{F,n}(\boldsymbol{\eta}) \xrightarrow{d} 0$$

so that $\left[G_{F,n}^{-1}(\boldsymbol{\eta}) - G_F^{-1}(\boldsymbol{\eta})\right] U_{F,n}(\boldsymbol{\eta}) \xrightarrow{p} 0$.

In addition, from Slutsky's theorem, it can be shown that:

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_F - \boldsymbol{\eta}) \xrightarrow{d} N(0, \Delta_F)$$

where $\Delta_F = G_F^{-1}(\boldsymbol{\eta}) Q_F(\boldsymbol{\eta}) [G_F^{-1}(\boldsymbol{\eta})]^T = G_F^{-1}(\boldsymbol{\eta})$ due to $G_F(\boldsymbol{\eta}) = Q_F(\boldsymbol{\eta})$.