



## BINARY LOGISTIC REGRESSION PROCEDURE WITH AN APPLICATION

Mustafa ŞAHİN<sup>1\*</sup>

<sup>1</sup>Kahramanmaraş Sütçü İmam University, Faculty of Agriculture, Department of Agricultural Biotechnology, 46100, Kahramanmaraş, Türkiye

**Abstract:** Binary logistic regression is widely used in health, social, and science sciences, particularly in studies involving extensive categorical independent variables. Considering that binary logistic regression calculates the probability of a specific event occurring or not occurring relative to the opposite event using one or more independent variables, its widespread use becomes even clearer. Another reason for its widespread use is that regression equations based on the least squares method cannot be used in data sets with numerous categorical variables affecting a two-level dependent variable. Another advantage is its ability to make future predictions and identify risk factors and impact magnitudes that influence the occurrence of events. However, in practice, many errors can occur, especially during model construction. Determining the individual effects of variables, joint effects, interaction effects, and how independent variables enter the model are highly sensitive processes. This study will examine the operations performed on independent variables in binary logistic regression until the final model is formed, using a numerical example. Thus, an important resource will be created for researchers to obtain the correct models with the correct process flow in binary logistic regression.

**Keywords:** Logistic regression, Procedure, Independent variable, Binary, Discrete data

\*Corresponding author: Kahramanmaraş Sütçü İmam University, Faculty of Agriculture, Department of Agricultural Biotechnology, 46100, Kahramanmaraş, Türkiye

E mail: ms66@ksu.edu.tr (M. ŞAHİN)

Mustafa ŞAHİN  <https://orcid.org/0000-0003-3622-4543>

Received: March 28, 2025

Accepted: May 06, 2025

Published: June 15, 2025

Cite as: Şahin, M. (2025). Binary logistic regression procedure with an application. *Black Sea Journal of Statistics*, 1(1), 22-26.

### 1. Introduction

Binary logistic regression (BNR) is a statistical modeling method used when the dependent variable consists of two categories. This model analyzes the relationship between one or more independent variables and the dependent variable to estimate the probabilities of observed outcomes. The aim of binary logistic regression is to estimate the probability that an observation belongs to a particular class. In this respect, the logistic regression model stands out as an effective tool for supporting decision-making processes and making probability-based predictions in a wide variety of disciplines. For example, in the field of healthcare (Bircan, 2004; Dasgupta and De, 2007; Şahin and Efe, 2018) (presence-absence of disease), education (Erath Şirin and Şahin, 2020) (success-failure), marketing (Maharani, 2021) (whether the customer will purchase the product), and animal husbandry (Adekunle, 2021; Gök and Tolun, 2023). Production being below or above a certain limit) has found a wide range of use. On the other hand, it is also used in many fields such as credit risk analysis, evaluation of insurance applications (Yin et al., 2020, Bakrie et al., 2023), estimation of crime rates, meteorology (Ruiz and Villa, 2008; Bertolin and De Santis, 2022), attitude and behavior analysis in social sciences (Kariyam, 2020; Okeke et al., 2020).

Above all, the effectiveness of binary logistic regression

analysis, which has found such widespread application, in relevant fields depends on the analysis being conducted using the correct statistical steps. For example, how independent variables with more than two discrete levels are entered into the model, or how these variables are transformed into design variables, is crucial. Similarly, whether continuous variables are entered into the model as continuous variables or transformed into discrete variables, or at what stage are interacting variables introduced into the model and how are their significance verified, is crucial. Consequently, the selection of variables and their evaluation are crucial. The process flow in binary logistic regression consists of various steps, from data preparation to model interpretation. First, the dependent variable is checked for binary nature. Dependent variables that are not binary are converted to binary, taking into account the researcher's experience or limitations outlined in the literature. Furthermore, the model is constructed after the appropriateness of the independent variables is assessed. Following model construction, regression coefficients are interpreted, significance tests are conducted, and the overall validity of the model is assessed using various statistical criteria.

In this study, after explaining the basic concepts of the binary logistic regression model, the step-by-step process flow will be detailed on a sample data set.



**2. Materials and Methods**

**2.1. Materials**

This study examined the stages of evaluating dependent variables in binary logistic regression, creating the model, and interpreting it using a numerical example. For this purpose, calf birth weight, live weight, birth type, calf sex, calf season, and feeding method were considered for 87 individuals of Jersey and Holstein dairy cattle. Calf birth weight (BW) was the dependent variable, while cattle breed (CB), live weight (LW), lactation order (LO), birth type (BT), calf gender (CG), birth season (BS), and feeding method (FM) were considered as independent variables. The coding of the variables is provided in Table 1.

**2.2. Methods**

Simple logistic regression analysis of candidate variables: In this initial stage of logistic regression, simple logistic regression analysis is performed on each variable individually. Three different tests can be used to determine whether variables are included in the model, i.e., to test the significance of the coefficients: the likelihood ratio test (G-test), the Wald test, and the score test. The G test will be used in this study. If the contribution of a candidate variable to the model is found to be insignificant by the G test, that variable is excluded from the model. The G statistic is in the form of given in Equation 1.

**Table 1.** Variables and coding types

		Variables						
Coding type	BW	LW	CB	LO	BT	CG	BS	FM
		0 if 30< 1 if 30≥	kg	Holstein=0 Jersey=1	1,2,3,4	Normal=0 Other=1	Female=0 Male=1	Spring=0 Summer=1 Autumn=2 Winter=3

$$G = 2 \left\{ \sum_{i=1}^n [y_i \log(\hat{P}_i) + (1 - y_i) \log(1 - \hat{P}_i)] - [n_1 \log(n_1) + n_0 \log(n_0) - n \log(n)] \right\} S \tag{1}$$

$$(\hat{P}_i = \hat{P}(x_i), n_1 = \sum y_i \text{ and } n_0 = \sum (1 - y_i))$$

$$\ell(\mathbf{x}_i) = \beta_0 + \sum_{k=1}^p \beta_k x_{ik} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} \tag{2}$$

Continuous variables are directly entered into the model as continuous variables when testing their individual contributions. On the other hand, variables with a level number of r>2 should be defined as design variables when checking their individual contributions. In simple logistic regression, all candidate variables whose contribution is found to be insignificant are excluded from the model, while all variables whose contribution is found to be significant are included in the multiple analysis.

**2.2. Multiple Analysis**

Discrete and continuous variables whose individual contributions are found to be significant are subjected to multiple analysis. In multiple analysis, the contribution of variables is again checked with the G test. A variable whose contribution is significant in simple logistic regression may be found to be insignificant in multiple analysis. In this case, the variable will be excluded from the model. A logit model for p independent variables,  $x=(x_1, x_2, \dots, x_p)$ , can be written as given in Equation 2:

**2.3. Deciding on How to Enter Continuous Variables into the Model**

At this stage, it is necessary to decide how to enter the continuous variable whose contribution is found to be significant in simple logistic regression and multiple regression into the model. For this purpose, the continuous variable is divided into quartile groups. The

first quartile group is taken as the reference group, and the design variables are defined. The continuous variable is then removed from the model, and multiple logistic regression analysis is performed using the defined design variable instead. The odds ratios for the estimated coefficients for the design variables are examined. If the odds ratios show a linear increase or decrease, the relevant independent variable is concluded to be linear with the logit method, and the variable is entered into the model as a continuous variable. If there is no linear increase or decrease, the variable is determined to be non-linear with the logit method, and the continuous variable remains in the model as a design variable. At this point, depending on the researcher's preference, the continuous variable can be made discrete and included in the model instead of the design variable. For this purpose, the continuous variable can be made discrete using a biologically significant or critical threshold. The contribution of the discrete variable to the model must be checked. Continuity checks are performed for all continuous variables in the model.

**2.4. Creating the Main Effects Model**

A main effects model is created using discrete variables whose contribution is deemed significant and continuous variables whose continuity is checked and their inclusion in the model is determined. Multiple logistic regression analysis is then performed by individually adding

interaction terms deemed likely or significant to the main effects model. The contribution of the added interaction terms to the main effects model is again tested using the G test. The interaction terms deemed significant are also added to the model to create the final model.

**2.5. Interpreting Coefficients in Logistic Regression**

In logistic regression, "odds" and "odds ratio" are used to interpret coefficients. The probability that  $y = 1$  for any group of  $X$  ( $X = 0$ ) is called "odds." "odds" is the logit without the natural logarithm, and for example, the odds values for  $X = 1$  and  $X = 0$  can be expressed as given in Equations 3 and 4, respectively.

$$\frac{P(x=1)}{1-P(x=1)} \tag{3}$$

$$\frac{P(x=0)}{1-P(x=0)} \tag{4}$$

The odds ratio will be denoted by  $\Omega$ . The odds ratio is the ratio of the odds value calculated for  $x=1$  to the odds value calculated for  $x=0$ . Therefore, the odds ratio can be written as given in Equation 5.

$$\Omega(1,0) = \frac{P(1)/[1-P(1)]}{P(0)/[1-P(0)]} \tag{5}$$

Accordingly, if the independent variable in logistic regression is binary and coded as 0,1, the odds ratio can be expressed as  $\Omega = e^{\beta_1}$ .

**2.6. Determining the Goodness of Fit of the Model**

The C statistic can be used to test the goodness of fit of the main effects model. From the main effects model, the logit estimate ( $\hat{\ell}(x_i)$ ) is obtained for each subject, respectively. Then, using the  $P(\mathbf{x}_i) = \frac{e^{\ell(\mathbf{x}_i)}}{1+e^{\ell(\mathbf{x}_i)}}$  equation,

the  $P(\mathbf{x}_i)$  and  $1-P(\mathbf{x}_i)$  values are obtained for each subject, respectively. They are sorted from smallest to largest according to  $P(\mathbf{x}_i)$ . Ten-fold risk groups T are then created, and their observed and expected frequencies are obtained. After obtaining the observed and expected frequencies for each ten-fold risk group, crosstabs are obtained, with the independent variable in the row and the ten-fold risk groups in the column.

The  $\hat{C}$  test statistic is then used to test the goodness of fit using expected and observed frequencies. The  $\hat{C}$  test statistic, which shows the  $\chi^2$  distribution with  $t-2$  degrees of freedom, is as given in Equation 6.

$$\hat{C} = \sum_{m=1}^t \left[ \frac{(g_{1m} - b_{1m})^2}{b_{1m}} + \frac{(g_{0m} - b_{0m})^2}{b_{0m}} \right] \tag{6}$$

**3. Results and Discussion**

The results of simple logistic regression analysis of variables thought to be related to birth weight such as cattle breed (CB), live weight (LW), lactation order (LO), birth type (BT), calf gender (CG), birth season (BS) and feeding method (FM) are given in Table 2.

The lactation order variable with more than two levels was defined as a design variable (DLO) in the model (1st and 2nd lactation=0, 2nd and 3rd lactation=1) and included in the main effects model together with breed, live weight and sex, which were found to be significant.

As shown in Table 3, the live weight variable, which is included as a continuous variable in the main effects model, should be checked to see if it can be included as a continuous variable. For this purpose, the live weight variable was converted into a design variable using quartile partitioning (LW1, LW2, LW3), and the results are presented in Table 4.

**Table 2.** Simple logistic regression analysis of variables thought to be associated with birth weight

Parameter	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\Omega}$	95% Confidence Limits	Chi-Square	-2 Log L	G	P
Intercept	1.209	0.254	-	-	22.511	93.810	-	-
CB	3.337	1.055	28.148	355-222.90	9.993	71.774	22.036	0.0016*
LW	0.0185	0.00493	1.019	1.009-1.029	14.017	64.50	29.31	0.0002*
LO	1534	0.390	4.639	2.158-9.972	15.441	68.196	25.614	0.0001*
BT	-0.518	0.513	0.595	0.218-1.519	1.205	92.792	1.018	0.312
CG	2.609	0.682	14.316	3.758-54.53	21.195	72.615	21.195	0.0001*
BS	-0.087	0.224	0.916	0.59-1.424	0.150	93.659	0.151	0.6977
FM	0.889	0.529	2.433	0.862-6.872	2.937	90.873	2.937	0.1093

\* Statistically significant at  $\alpha=0.10$  error level.

**Table 3.** Multiple logistic regression analysis results of the main effects model

Parameter	$\hat{\beta}$	$SE(\hat{\beta})$	Chi-Square	P
Intercept	-7.2223	2.6487	7.4351	0.0064*
CB	2.0715	0.8335	6.1771	0.0129*
LW	0.0131	0.00561	5.4769	0.0193*
DLO	-2.8991	1.4795	3.8395	0.0501*
CG	4.8539	1.6057	9.1382	0.0025*

\* Statistically significant at  $\alpha=0.10$  error level.

**Table 4.** Multiple logistic regression analysis for the live weight variable divided into quartiles and transformed into a design variable

Parameter	$\hat{\beta}$	$SE(\hat{\beta})$	Chi-Square	P
Intercept	-2.2379	0.9951	5.06	0.0245
CB	1.7861	0.8751	4.17	0.0413
LW1	1.5166	1.1436	1.76	0.1848
LW2	1.5615	1.1069	1.99	0.1583
LW3	3.6140	1.4186	6.49	0.0108*
DLO	-2.7252	1.5571	3.06	0.0801*
CG	4.8828	1.6510	8.75	0.0031*

\* Statistically significant at  $\alpha=0.10$  error level.

An examination of Table 4 reveals that the odds ratios for live weight, as a result of the quartile analysis, are lowest in Group 2, higher in Group 3, and highest in Group 4. Since the estimated coefficients do not follow an increasing trend (4.55, 4.76, 37.11), it is concluded that the variable is linear with the logit. Therefore, the live weight variable will be entered into the model as a continuous variable.

After obtaining the main effects model, interactions between the variables in the model are examined. In this step, the contribution of each interaction term to the main effects model is examined. Therefore, the -2 log likelihood value, likelihood ratio test statistic (G), degrees of freedom, and P-value for each possible interaction term are presented in Table 5.

**Table 5.** Values obtained by adding interaction terms one by one to the main effects model

Parameter	-2 Log L	DF	G	P
Main Effects Model	43.80			
CB*LW	42.02	1	1.78	0.256
CB*DLO	41.13	1	2.67	0.951
CB*CG	41.31	1	2.49	0.953
LW*DLO	41.61	1	2.19	0.174
LW*CG	43.41	1	0.39	0.551
DLO*CG	43.73	1	0.07	0.966

\* Statistically significant at  $\alpha=0.10$  error level.

**Table 6.** Observed and expected frequencies for ten risk groups

	1	2	3	...	9	10	$\Sigma$
$y=1$	0	1	1	...	2	1	18
	0.114	0.783	2.615	...	2.921	1.048	18
$y=0$	6	5	7	...	7	5	69
	5.915	4.745	6.135	...	5.987	4.589	69

An examination of Table 5 reveals that all potential interaction effects added to the main effects model are statistically insignificant ( $P>0.10$ ). At this stage, the final model can be tested for goodness of fit. Some of the observed and expected frequencies of the resulting ten-risk groups are presented in Table 6. Using these values, the C statistic can be calculated for the goodness of fit test of the main effects model.

From Table 6, the Y value can be calculated as given in Equation 7.

$$\hat{C} = \sum_{m=1}^t \left[ \frac{(g_{1m} - b_{1m})^2}{b_{1m}} + \frac{(g_{0m} - b_{0m})^2}{b_{0m}} \right] = 5.253 \quad (7)$$

Here, since  $\hat{C} = 5.253 < X_{8,0.05}^2 = 15.507$ , it can be said that the final model fits the data well.

As a result, the logistic regression model estimate for any subject from the main effects model can be written as

given in Equation 8.

$$\hat{\varrho}(CB, LW, DLO, CG) = \hat{\beta}_0 + \hat{\beta}_1(CB) + \hat{\beta}_2(LW) + \hat{\beta}_3(DLO) + \hat{\beta}_4(CG) = -7.2223 + 2.0715(CB) + 0.0131(LW) - 2.8991(DLO) + 4.8539(CG) \quad (8)$$

The main effects model does not include birth type, calving season, and feeding method variables, and therefore, these variables have no statistically significant effect on calf birth weight. The main effects model shows that cattle breed (CB), live weight (LW), lactation order (DLO), and calf gender (CG) have an effect on calf birth weight. The odds ratios for these effects are as follows; for cattle breed (CB);  $\hat{\Omega} = \exp(2.0715) = 7.936$ , live weight (LW);  $\hat{\Omega} = \exp(0.0131) = 1.031$ , for lactation order (DLO);  $\hat{\Omega} = \exp(-2.8991) = 0.055$ , and for calf sex (CG);  $\hat{\Omega} = \exp(4.8539) = 128.22$ . In this case, it can be said that calf

gender is the most effective variable, while lactation order has the least effect.

#### 4. Conclusion

In binary logistic regression, variable selection and how they are included in the model play a critical role in the model's accuracy, interpretability, and generalizability. Including unnecessary or irrelevant variables can lead to problems such as multicollinearity and overfitting, thereby reducing model performance.

In this study, the linearity of the live weight variable, as a continuous independent variable, was tested using the logit method. It was concluded that it was linear using the logit method, and it was decided to include it as a continuous variable in the model. Furthermore, the lactation order variable, which has more than two levels, was also defined as a design variable and included in the model. The study demonstrated that without checking continuity for the continuous variable and defining a design variable for the independent variable with more than two levels, very different coefficient estimates, and therefore different odds ratios, would have been obtained.

In conclusion, it can be said that in binary logistic regression analysis, determining the factors affecting the dependent variable with a statistically sound process and the method of evaluating the variables are extremely important.

#### Author Contributions

The percentages of the author' contributions are presented below. The authors reviewed and approved the final version of the manuscript.

	M.Ş.
C	100
D	100
S	100
DCP	100
DAI	100
L	100
W	100
CR	100
SR	100
PM	100
FA	100

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

#### Conflict of Interest

The author declare that there is no conflict of interest.

#### Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

#### References

- Adekunle, F. O. (2021). A binary logistic regression model for prediction of feed conversion ratio of *Clarias gariepinus* from feed composition data. *Marine Science and Technology Bulletin*, 10(2), 134–141.
- Bakrie, P. D., Aulia, R. N., & Taufiqillah, R. (2023). Customer churn prediction for life insurance using binary logistic regression. *Economics Review and Journal*, 3(3), 45–58.
- Bertolin, C., & De Santis, F. (2022). Prediction of snowmelt days using binary logistic regression in the Umbria-Marche Apennines (Central Italy). *Water*, 14(9), 1495.
- Bircan, H. (2004). Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama. *Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 2(1), 185–208.
- Dasgupta, S., & De, U. K. (2007). Binary logistic regression models for short term prediction of premonsoon convective developments over Kolkata (India). *International Journal of Climatology*, 27(6), 831–836.
- Eratlı Şirin, Y., & Şahin, M. (2020). Investigation of factors affecting the achievement of university students with logistic regression analysis: School of Physical Education and Sport example. *SAGE Open*, 10(3), 215–225.
- Gök, İ., & Tolun, T. (2023). A study on body weight and carcass characteristics and sex in broilers. *Türk Doğa ve Fen Dergisi*, 12(4), 129–133.
- Kariyam, F. H. (2020). Application of binary logistic regression in modelling women's participation in improving the welfare of fishermen families. İçinde *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)* (ss. 179–185).
- Maharani, O. C. (2021). A binary logistic regression analysis of factors influencing repeat purchase for snack product X. *Jurnal Teknik Industri Terintegrasi*, 7(4), 359–366.
- Okeke, J. U., Okeke, E. N., & Dakhin, Y. V. (2020). Binary logistic models of home ownership in Wukari Nigeria. *Open Journal of Statistics*, 10(1), 50–59.
- Ruiz, A., & Villa, N. (2008). *Storms prediction: Logistic regression vs random forest for unbalanced data* (arXiv Preprint, arXiv: 0804.0650). Cornell University.
- Şahin, M., & Efe, E. (2018). Determining the factors affecting birth weight by using logistic regression method. *Black Sea Journal of Health Science*, 1(2), 22–27.
- Yin, S., Dey, D. K., Valdez, E. A., Gan, G., & Vadiveloo, J. (2020). *Skewed link regression models for imbalanced binary response with applications to life insurance* (arXiv Preprint, arXiv: 2007.15172). Cornell University.