Logistic Regression

Logistic regression is a class of regression where the independent variable is used to predict the dependent variable. When the dependent variable has two categories, then it is a binary logistic regression. When the dependent variable has more than two categories, then it is a multinomial logistic regression. When the dependent variable category is to be ranked, then it is an ordinal logistic regression (OLS). To obtain the maximum likelihood estimation, transform the dependent variable in the logit function. Logit is basically a natural log of the dependent variable and tells whether or not the event will occur. Ordinal logistic regression does not assume a linear relationship between the dependent and independent variable. It does not assume homoscedasticity. Wald statistics tests the significance of the individual independent variable.

Assumptions: This test is popular because it can overcome many restrictive assumptions of OLS regression.

1. In OLS regression, a linear relationship between the dependent and independent variable is a must, but in logistic regression, one does not assume such things. The relationship between the dependent and independent variable may be linear or non-linear.
2. OLS assumes that the distribution should be normally distributed, but in logistic regression, the distribution may be normal, poisson, or binominal.
3. OLS assumes that there is an equal variance between all independent variables, but ordinal logistic does not assume that there is an equal variance between independent variables.
4. Does not assume normally distributed error term variance.

Still, violation of these OLS assumptions in logistic regression assumes the following:

- Data level: The dependent variable should be dichotomous in nature for binary regression.
- Error Term: The error term is assumed independently.
- Linearity: Does not assume a linear relationship, but between the odd ratio and the independent variable, there should be a linear relationship.
- No outliers: Assumes that there should be no outliers in data.
- Large sample: Uses the maximum likelihood method, so a large sample size is required for logistic regression.

Key terms and concepts:

- **Dependent variable:** Dichotomous in nature, for the binary logistic regression dependent variables are in two categories. Usually we predict the higher category (assumed as 1) by

taking the lower reference category (assumed as 0). In multinomial logistic regression, the dependent variable has more than two categories. We can predict the other category by the reference category. In ordinal logistic regression, we predict the cumulative probability of the dependent variable order.

- **Factor:** The independent variable is dichotomous in nature and is called the factor. Usually we convert them into a dummy variable.

- **Covariate:** The independent variable that is metric in nature is called the covariate.

- **Interaction term:** The covariate shows the individual effect on the dependent variable. The interaction effect is the combination of two variable effects on the dependent variable. For example, when we predict the dependent variable based upon age and education category, there will be two impacts: one is individual impact on the dependent variable and the other is the interaction impact.

- **Maximum likelihood estimation:** This method is used to predict the odd ratio for the dependent variable. In OLS estimation, we minimize the error sum of the square distance, but in maximum likelihood estimation, we maximize the log likelihood.

- **SPSS and SAS:** In SPSS, this test is available in the regression option and in SAS, we can use this method by using "command proc logistic" or "proc catmod."

- **Significance test:** Hosmer and Lemeshow chi-square test is used to test the overall model of goodness-of-fit test. It is the modified chi-square test, which is better than the traditional chi-square test. Significant p value shows the goodness-of- fit model. Omnibus tests table in SPSS output shows the traditional chi-square and Hosmer and Lemeshow chi-square test value. Pearson chi-square test and likelihood ratio test are used in multinomial logistic regression to estimate the model goodness-of-fit.

- **Stepwise:** The three methods available are enter, backward, and forward. In the enter method, all variables will be included, whether it is significant or insignificant. In the backward method, it will start dropping non-significant variables from the list. In forward method, it will move forward while dropping non-significant variables.

- **Parameter estimate and logit:** In SPSS statistical output, the "parameter estimate" is the b coefficient used to predict the log odds (logit) of the dependent variable. Let z be the logit for a dependent variable, then the logistic prediction equation is:

$$z = \ln(\text{odds(event)}) = \ln(\text{prob(event)}/\text{prob(nonevent)}) = \ln(\text{prob(event)}/[1 - \text{prob(event)}])$$
$$= b0 + b1X1 + b2X2 + ... + bkXk$$

Where b0 is constant and k is independent (X) variables. In ordinal logistic regression, the threshold coefficient will be different for every order of dependent variables. The coefficient will give the cumulative probability of every order of dependent variables.

- **Odd ratio:** Exponential beta gives the odd ratio of the dependent variable. We can find the probability of the dependent variable from this odd ratio. When the exponential beta value is greater than one, than the probability of higher category increases, and if the probability of exponential beta is less than one, then the probability of higher category decreases. Exponential beta value is interpreted with the reference category, where the probability of the dependent variable will increase or decrease. In continuous variables, it is interpreted with one unit increase in the independent variable, corresponding to the increase or decrease of the units of the dependent variable.

- **Measures of Effect Size:** R^2 is no more accepted because R^2 tells us the variance extraction by the independent variable, but here, variance is split into two categories. Cox and Snell's R^2, Nagelkerke's R^2, McFadden's R^2, and Pseudo-R^2 are now more realizable then simple R^2.

- **Classification Table:** The classification table shows how these two categories are correctly predicted. For example, from two categories, only 85% were predicted correctly, this is shown in the classification table.

Logistic Regression Resources

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods and Research, 28*(2), 186-208.

DeMaris, A. (1992). *Logit modeling: Practical applications*. Newbury Park, CA: Sage Publications.

Greenland, S., Schwartzbaum, J. A., & Finkle, W. D. (2000). Problems due to small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology, 151*(5), 531-539.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: John Wiley & Sons.

Jaccard, J. (2001). *Interaction effects in logistic regression*. Thousand Oaks, CA: Sage Publications.

Jennings, D. E. (1986). Outliers and residual distributions in logistic regression. *Journal of the*

American Statistical Association, 81(396), 987-990.

Kleinbaum, D. G., Klein, M., & Pryor, E. R. (2004). *Logistic regression: A self-learning text* (2nd ed.). New York: Springer.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105-142). New York: Academic Press.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications.

O'Connell, A. A. (2005). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: Sage Publications.

Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage Publications.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research*. New York: Holt, Rinehart & Winston.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12), 1373-1379.

Peng, C. -Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research, 96*(1), 3-14.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association, 73*(364), 699-705.

Rice, J. C. (1994). Logistic regression: An introduction. In B. Thompson (Ed.), *Advances in social science methodology* (pp. 191-245). Greenwich, CT: JAI Press.

Wright, R. E. (1994). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 217-244). Washington, DC: American Psychological Association.

- Logistic Regression: A Self-Learning Text (Statistics for Biology and Health)

- Logistic Regression: A Primer (Quantitative Applications in the Social Sciences)

- Logistic Regression: A Self-Learning Text (Statistics for Biology and Health)

Related Pages:

- Conduct and Interpret a Logistic Regression
- What is Logistic Regression?
- The Logistic Regression Analysis in SPSS