



# Video Tutorial on Logistic Regression: Background Information Sheet

*This document is part of a video tutorial series that outlines the use of two statistical methods in the evaluation of routine data. This background document is on the second method, logistic regression.*

## What is logistic regression?

Logistic regression, also referred to as binary logistic regression, is a specialized form of regression analysis used when the outcome variable is a nonmetric, binary/dichotomous variable. Often, researchers want to analyze whether an event has occurred or if a person falls within a target group. It has extensive use in healthcare and social sciences and is often used to develop prediction models that predict the probability of developing a disease outcome. Additionally, logistic regression is easy to use, and produces robust estimates.

In logistic regression, the dependent/outcome variable is a dummy variable that is coded 0 and 1, where 1 represents the event occurring and 0 represents the event not occurring. The model produced in the regression is a nonlinear transformation of the linear regression and employs binomial probability theory to predict the probability of an event happening. Thus, the probability is transformed through the log link function to create a new variable, logit. The distribution in logit is constrained so that the estimated probabilities lie between 0 and 1. The model formation is:

$$\text{logit} = \ln\left(\frac{p}{(1-p)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

*where,  $p$  = probability when outcome occurs ( $y=1$ ),  $X$  = independent variables,  $\beta$  = regression coefficients.*

The antilog of the regression coefficients/odds produces odds ratio. The model formation is:

$$\text{Odds ratio} = \left(\frac{p}{(1-p)}\right) = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}$$

*where,  $p$  = probability when outcome occurs ( $y=1$ ),  $X$  = independent variables,  $\beta$  = regression coefficients.*

## What questions can you answer using logistic regression?

This method has been used to answer a variety of questions in public health (see Table 1 for a list of studies that have employed logistic regression analysis), including:

- How do the characteristics of clients who initiate antiretroviral therapy (ART) on the same day as HIV diagnosis differ from clients who initiate treatment at later time points after diagnosis?
- What factors are associated with retesting in patients previously diagnosed with HIV?
- Is area-level deprivation associated with adverse obstetric and perinatal outcomes?
- Which clinic and patient factors are associated with enrolment, eligibility, and start of ART?

## When do you use logistic regression?

Logistic regression can be used if the assumptions are met, and the sample size is adequate.

*Assumptions are met.* The analysis can be used when the following assumptions are not violated (Hair, et al., 2019).

1. **Binary outcome.** The outcome variable is a dichotomous variable, and the categories are mutually exclusive and exhaustive.

2. **Independent variables.** The model should have two or more independent variables measured at the continuous or nominal level.
3. **Independence of observations.** There should not be a relationship between observations.
4. **Multicollinearity.** Two or more independent variables should not be highly correlated with each other.
5. **Linear relationship.** There should be a linear relationship between the continuous independent variables and the logit transformation of the outcome.
6. **Outliers.** No significant outliers or high leverage points should be present in the data.

*Adequate sample size.* Calculating the sample size is complex, but some researchers have suggested general guidelines.

- Peduzzi, et al. (1996) developed the 10:1 recommendation to calculate the sample size using the following equation:
  - $n = 10 \times \frac{k}{p}$ 
    - n = sample size for logistic regression
    - k = number of independent variables/ covariates/predictors
    - p = proportion of successes/falling in target group
  - If the number is less than 100, they recommended that the sample size should be increased to 100.
- Long (1997) suggests the minimum sample size should be 100 and samples less than 100 are risky and can bias the estimates. Samples greater or equal to 500 are adequate for logistic regression analysis and there should be at least 10 cases per predictor.

These guidelines do not apply in all circumstances—for instance, if the event is rare.



## Useful resources

Below is a compilation of useful resources on logistic regression.

- Hair, J., Black, W., Babin, B., & Anderson, R. (2019). *Multivariate Data Analysis* (8th ed.). Cengage Learning EMEA. Retrieved from <https://www.cengage.com/c/multivariate-data-analysis-8e-hair/9781473756540PF/>
- Hilbe, J.M. (2016). *Practical Guide to Logistic Regression* (1st ed.). Chapman and Hall/CRC. Retrieved from <https://doi.org/10.1201/b18678>
- Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression* (3rd ed.). John Wiley & Sons, Inc. Retrieved from <https://doi.org/10.1002/9781118548387>
- Long, J., & Freese, J. (2014). *Regression Models for Categorical Dependent Variables using Stata* (3rd ed.). Stata Press. Retrieved from <https://www.stata.com/bookstore/regression-models-categorical-dependent-variables/>
- Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. SAGE. Retrieved from <https://us.sagepub.com/en-us/nam/regression-models-for-categorical-and-limited-dependent-variables/book6071>
- 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. Retrieved from [https://doi.org/10.1016/S0895-4356\(96\)00236-3](https://doi.org/10.1016/S0895-4356(96)00236-3)
- Pituch, K., & Stevens, J. (2016). *Applied Multivariate Statistics for the Social Sciences* (6th ed.). Routledge Taylor and Francis Group. Retrieved from <https://doi.org/10.4324/9781315814919>

**Table 1: Summary of studies using logistic regression**

Study	Title	Year	Research question or objective	Data source
Agorinya, I., Kanmiki, E., Nonterah, E., Tediosi, F., Akazili, J., Welaga, P., Azongo, D., Oduro, A.	<a href="#">Socio-demographic determinants of low birth weight: Evidence from the Kassena-Nankana districts of the Upper East Region of Ghana</a>	2018	What are the social, economic, and demographic factors that determine low birth weight?	Navrongo Health and Demographic Surveillance System
Beyerlein, A., Lack, N., Maier, W.	<a href="#">Associations of area-level deprivation with adverse obstetric and perinatal outcomes in Bavaria, Germany: Results from a cross-sectional study</a>	2020	Is area-level deprivation associated with adverse obstetric and perinatal outcomes?	German Bavarian Quality Assurance Institute for Medical Care
Drewitz, K., Loss, M., Loss, J., Apfelbacher, C.	<a href="#">Predictors of non-transplantation of adult donor organs--an observational study using routine data from Eurotransplant</a>	2014	Determine the importance of sociodemographic and medical donor characteristics in determining non-translation of the pancreas.	Medical database on deceased potential organ donors in the Eurotransplant region
Grylka-Baeschlin, S., Iglesias, C., Erdin, R., Pehlke-Milde, J.	<a href="#">Evaluation of a midwifery network to guarantee outpatient postpartum care: a mixed methods study</a>	2020	(1) How did users of services from the network differ from women organizing their postpartum care themselves? (2) What are the predictors of the use of the midwifery network?	Routine data collected by independent midwives
Hofmann, S., Hess, S., Klein, C., Lindena, G., Radbruch, L., Ostgathe, C.	<a href="#">Patients in palliative care—Development of a predictive model for anxiety using routine data</a>	2017	(1) Explore the differences between data sets of patients with high anxiety scores and those with low anxiety scores. (2) Discover possible predictors of anxiety. (3) Find and test a predictive model based on routine data.	Routine data collected using the Hospice and Palliative Care Evaluation
Jacob, N., Rice, B., Kalk, E., Heekes, A., Morgan, J., Hargreaves, J., Boule, A.	<a href="#">Utility of digitising point of care HIV test results to accurately measure, and improve performance towards, the UNAIDS 90-90-90 targets</a>	2020	What factors are associated with the retesting in patients previously diagnosed with HIV?	Health records, patient master index, and existing HIV cascade of the PHDC using patient folder numbers
Kayigamba, F., Bakker, M., Fikse, H., Mugisha, V., Asiimwe, A., Schim van der Loeff, M.	<a href="#">Patient enrolment into HIV care and treatment within 90 days of HIV diagnosis in eight Rwandan health facilities: a review of facility-based registers</a>	2012	Which clinic and patient factors are associated with enrolment, eligibility, and start of ART?	Hospital registries

Study	Title	Year	Research question or objective	Data source
Lalla-Edward, S., Ncube, S., Matthew, P., Hankins, C., Venter, W., Gomez, G.	<a href="#">Uptake of health services among truck drivers in South Africa: analysis of routine data from nine roadside wellness centres</a>	2017	Determine the factors associated with service utilization.	Routine data from nine Roadside Wellness Centers
Lilian, R., Rees, K., McIntyre, J., Struthers, H., Peters, R.	<a href="#">Same-day antiretroviral therapy initiation for HIV-infected adults in South Africa: Analysis of routine data</a>	2020	How do the characteristics of clients who initiate ART on the same day as HIV diagnosis differ from clients who initiate treatment at later time points after diagnosis?	Three Interlinked Electronic Registers.Net (TIER.net)
Msukwa, M., Keiser, O., Jahn, A., van Oosterhout, J., Edmonds, A., Phiri, N., Manjomo, R., Davies, M., Estill, J.	<a href="#">Timing of combination antiretroviral therapy (cART) initiation is not associated with stillbirth among HIV-infected pregnant women in Malawi</a>	2019	Assess the association between timing of maternal combination ART initiation and stillbirth among HIV-infected pregnant women in Malawi's Option B+ program	(1) Maternity register (2) Government-issued personal health passport
Onwere, C., Gurol-Urganci, I., Cromwell, D., Mahmood, T., Templeton, A., van der Meulen, J.	<a href="#">Maternal morbidity associated with placenta praevia among women who had elective caesarean section</a>	2011	Determine the risks of complications associated with placenta praevia.	Hospital Episode Statistics database
Rishard, M., Fahmy, F., Senanayake, H., Ranaweera, A., Armocida, B., Mariani, I., Lazzerini, M.	<a href="#">Correlation among experience of person-centered maternity care, provision of care and women's satisfaction: Cross sectional study in Colombo, Sri Lanka</a>	2021	Determine the factors associated with each quality-of-care domain: (1) degree of implementation of PCMC, (2) key indicators of provision of care, and (3) women's satisfaction with maternal care.	(1) Individual patient-level database at De Soysa Hospital (2) Exit interview survey
Woldesenbet, S., Kufa-Chakezha, T., Lombard, C., Manda, S., Cheyip, M., Ayalew, K., Chirombo, B., Barron, P., Diallo, K., Parekh, B., Puren, A.	<a href="#">Recent HIV infection among pregnant women in the 2017 antenatal sentinel cross-sectional survey, South Africa: Assay-based incidence measurement</a>	2021	What characteristics are associated with being a recently infected pregnant woman with HIV?	2017 Antenatal HIV Sentinel Survey, and medical records
Zango, S., Lingani, M., Valea, I., Samadoulougou, O., Bihoun, B., Rouamba, T., Derra, K., Rouamba, E., Donnen, P., Dramaix, M., Tinto, H., Robert, A.	<a href="#">Malaria and curable sexually transmitted infections in pregnant women: A two-years observational study in rural Burkina Faso</a>	2020	What maternal characteristics are associated with malaria and curable STI?	Health facility registries used during ANC and delivery

DHIS – District Health Information Software, version 2; HMIS – Health Management Information System; MCH – maternal and child health; M&E – monitoring and evaluation; RHIS – routine health information system



## References

Hair, J., Black, W., Babin, B., & Anderson, R. (2019). *Multivariate Data Analysis* (8th ed.). Cengage Learning EMEA.

Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. SAGE.

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. Retrieved from [https://doi.org/10.1016/S0895-4356\(96\)00236-3](https://doi.org/10.1016/S0895-4356(96)00236-3)

### For more information

D4I supports countries to realize the power of data as actionable evidence that can improve programs, policies, and—ultimately—health outcomes. We strengthen the technical and organizational capacity of local partners to collect, analyze, and use data to support their move to self-reliance. For more information, visit <https://www.data4impactproject.org/>

