**Multiple Logistic Regression for Dichotomous Variables in R**

**Statistics in R Series**



Photo by [Kimberly Farmer](https://unsplash.com/@kimberlyfarmer?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText) on [Unsplash](https://unsplash.com/s/photos/education?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText)

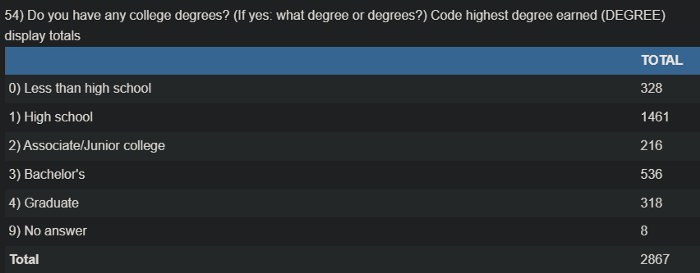
**Introduction**

Simple logistic regression incorporates only one predictor variable and we have implemented it using R previously. We have also discussed on the statistics for goodness-of-fit. Oftentimes, the real world data has several predictor variables. Sometimes we simply don’t know if including the extra variable as a predictor will make the model more robust or not. So we need to always compare the full model with the nested model and draw conclusion from the data. Here, we will increase the number of independent parameters and expand our study on multiple logistic regression.

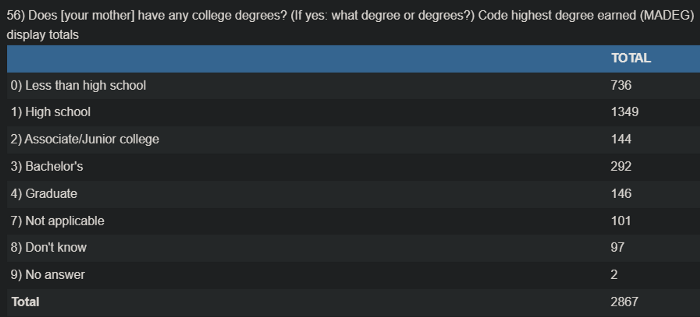
**Dataset**

The data from the 2016 General Social Survey (GSS) will be used to demonstrate logistic regression. Data was downloaded from the [Association of Religion Data Archives](https://www.thearda.com/data-archive?fid=GSS2016&tab=3) and collected by Tom W. Smith. This dataset includes responses collected from nearly 3,000 respondents and contains data related to several socioeconomic factors. It contains various types of data, such as information pertaining to a person’s marital status, the status of their education, their working hours, their employment status, etc. Let’s take a closer look at this dataset to gain a better understanding of it.

For each individual, the DEGREE column provides the level of education, while the MADEG column provides the level of education for each individual mother. In this study, we aim to determine whether the mother’s bachelor’s degree level is a good predictor of the children’s bachelor’s degree level. This dataset contains categorical data that is encoded ordinally.



DEGREE data [Image by Author]



MADEG data [Image by Author]

**Answer we are trying to find**

In the previous multiple logistic regression article (link below),

**[Simple Logistic Regression for Dichotomous Variables in R](https://towardsdatascience.com/simple-logistic-regression-for-dichotomous-variables-in-r-8befbe3e95b4" \t "_blank)**

**[Statistics in R Series](https://towardsdatascience.com/simple-logistic-regression-for-dichotomous-variables-in-r-8befbe3e95b4" \t "_blank)**

[towardsdatascience.com](https://towardsdatascience.com/simple-logistic-regression-for-dichotomous-variables-in-r-8befbe3e95b4" \t "_blank)

we tried to answer the following question.

Does the education level of mother impact the education level of the children?

Our analysis of the data revealed a positive coefficient for MADEG, which represents the mother’s educational level. As can be seen in the interpretation of result section of the following figure, the intercept estimate is 0.257 and the MADEG coefficient estimate is 0.316. Accordingly, for every one unit increase in the predictor variable, which is mother’s education level, the logit probability that a child’s education level will have a value of 1 will increase by 0.31598. However, this still represents a positive slope, indicating an increase in the response variable with an increase in the predictor variable. Thus, the probability that the child will receive a bachelor degree is increased if his or her mother has a bachelor degree.

Now, we will pose a new question.

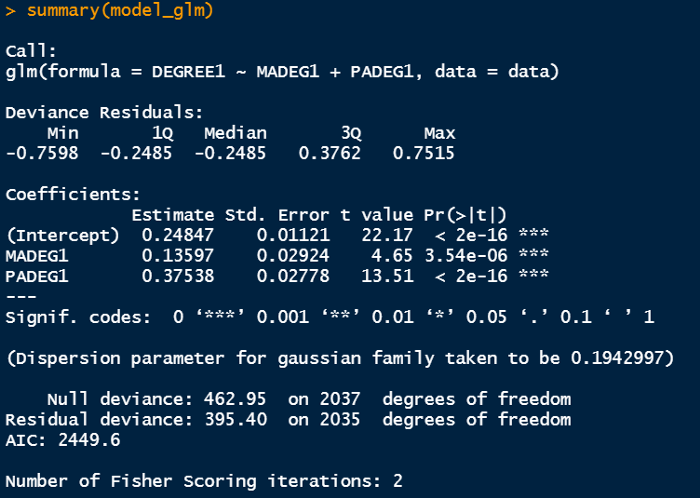
Does the education level of father have anything to do with the children’s education or is it uncorrelated?

When we incorporate another independent variable, the simple logistic regression become multiple logistic regression.

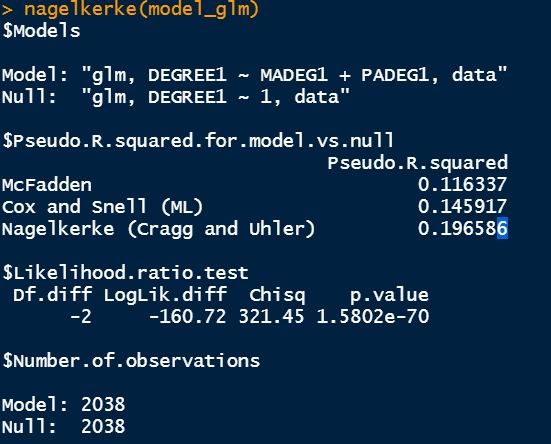
**Implementation in R**

To execute this regression study in R, we will need the following libraries to be installed. The data is stored in an excel file and we will use the glm() function. The difference now is the addition of PADEG which represents the education level of father.

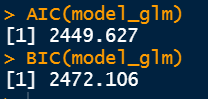
R code [by Author]



Output window in R [Image by Author]



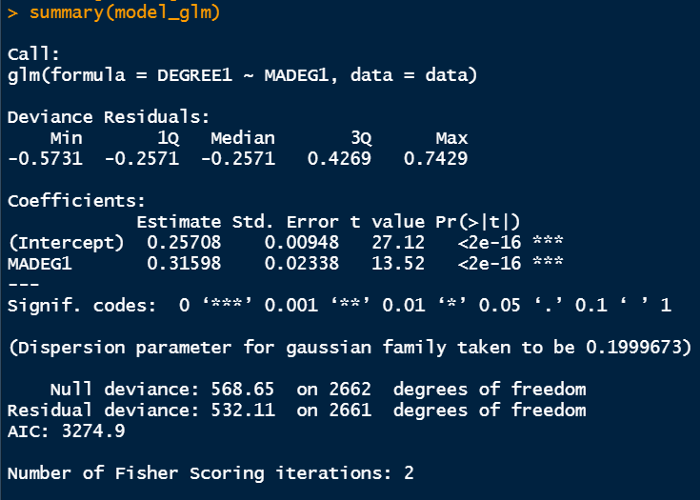
Output window in R [Image by Author]



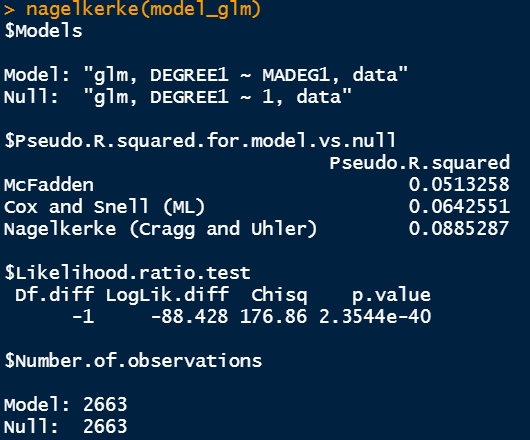
Output window in R [Image by Author]

***Interpretation of result***

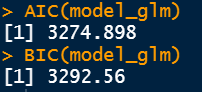
As a first step in the process of implementing logistic regression, we need to convert the probability of output success into logarithmic measures, in order to determine the coefficient and intercept of the predictor variable. I have provided a brief interpretation of the data below. In addition, I have placed the output windows of a simple logistic regression in the same place as well for the sake of comparison.



Coefficients from simple logistic regression when only MADEG is considered



pseudo R² from simple logistic regression when only MADEG is considered



AIC/BIC from simple logistic regression when only MADEG is considered

1. The MADEG coefficient is 0.136 and PADEG coefficient is 0.375 and the intercept’s coefficient remains similar. We can conclude that for every one unit increase in the mother’s education level, the logit probability of child’s education level to have value of 1 increases by 0.135 which is still positive and for every one unit increase in the father’s education level, the logit probability of child’s education level to have value of 1 increases by 0.375 which is also positive. In other words, when both mother’s and father’s bachelor degree are considered, there is increase in the probability of the child’s bachelor degree.
2. The associated p-value is less than 0.05 which also tells us to reject the null hypothesis. The null hypothesis here is “the predictor variables has coefficient of 0 and essentially do not impact the response variable”. Therefore, we can conclude that mother’s and father’s bachelor education significantly impact the child’s bachelor degree.
3. The pseudo R² value can also compared with the simple logictic regression counterpart. It is obvious the pseudo R² values in this case have increased when father’s education level is included. It means the the full model fits better than the simple logistic model.
4. The AIC/BIC statistics can also be compared. It is also evident that AIC/BIC values are smaller in the full model. Smaller AIC/BIC values indicate a better fit and this also supports the statement from pseudo R² where we also concluded that the full model is better.
5. The deviance of the simple logistic regression model is 532.11 where as the deviance of the new model is 395.40 which implies that the new model has smaller deviance from the saturated model. In a saturated model, the number of parameters equals the sample size since it contains one parameter for each observation. The difference between the null deviance and the residual deviance is used to determine the significance of the current model.
6. We can also calculate the log-likelihood ratio by subtracting the deviances between these two models. This is for comparison between several nested models with respect to the full model where all the possible predictor variables are considered.

Log-likelihood ratio = Deviance of the reduced model — Deviance of the full model

***Conclusion***

We have discussed about multiple logistic regression and its implementation in R. We have also walked though the R outputs and interpret the results from General Society Survey. The positive coefficient for the predictor variables indicate that with the increase of mother’s and father’s bachelor degree’s value from 0 to 1, the probability of the child’s bachelor degree becoming 1 increases by 0.135 and 0.375 respectively or in other words it can be concluded that mother’s and father’s education significantly impact the child’s education in our dataset.

Thanks for reading.