**Logistic Regression Models Comparison**

**Statistics in R Series**

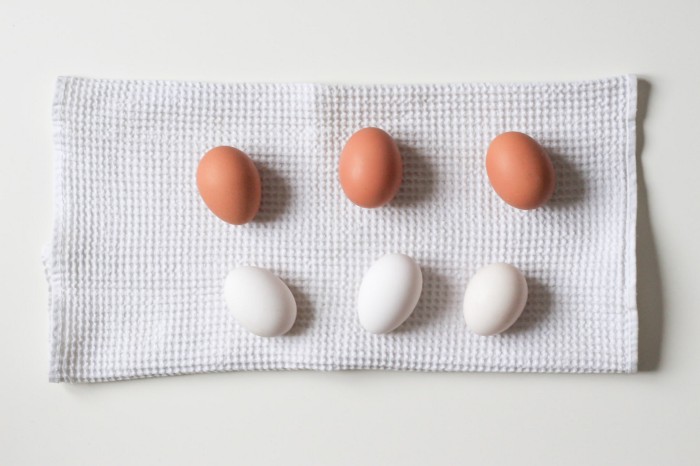


Photo by [NordWood Themes](https://unsplash.com/@nordwood?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText) on [Unsplash](https://unsplash.com/s/photos/comparison?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText)

**Introduction**

In simple logistic regression, we have only one predictor variable whereas in multiple logistic regression, there are more than one predictors. The response variable can be binary as well as ordered. For example, the response variable can be just a choice between two categories like city or village, healthy or sick…

We have discussed these binary and ordinal logistic regressions in previous articles with links below

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

[Multiple Logistic Regression for Dichotomous Variables in R](https://towardsdatascience.com/multiple-logistic-regression-in-r-aa65c2a91e22)

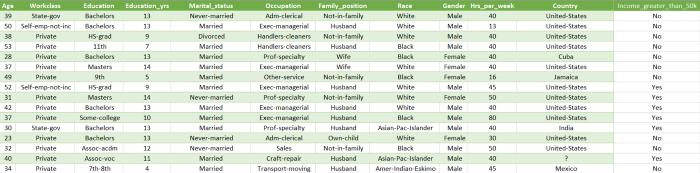
[Simple Logistic Regression for Ordinal Variables in R](https://towardsdatascience.com/simple-logistic-regression-for-ordinal-variables-in-r-1d95fa278c5e)

[Multiple Logistic Regression for Ordinal Variable and Predicted Probabilities in R](https://medium.com/towards-data-science/multiple-logistic-regression-for-ordinal-variable-and-predicted-probabilities-in-r-3e3ef3ba6ca2)

In this article, I am going to walk you through the comparison between different models and how to interpret R outputs.

**Dataset**

This case study will be based on a dataset from the [UCI Machine Learning Repository known as the Adult Data Set](https://archive.ics.uci.edu/ml/datasets/adult), which will be used as a data source. It has been estimated that more than 30000 individuals have been identified by their demographic information in this dataset, which includes, but is not limited to, their race, education, occupation, gender, salary, working hours per week, employment level, as well as their income level.



[Adult Data Set](https://archive.ics.uci.edu/ml/datasets/adult) from UCI Machine Learning Repository

For the purpose of performing an ordinal logistic regression analysis, it is necessary to make some modifications to the given data.

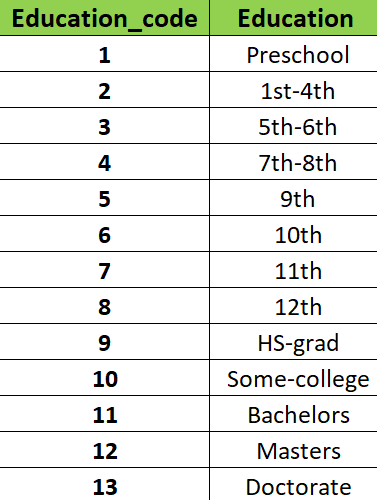
Here, I will have two datasets: one for the binary model where the response variable is binary and the other for the ordinal model where the response variable is ordered. The raw data was modified to perform simple and ordinal logistic regression which are saved in the github below.

Link to excel file simple logistic regression: [GSS — glm2.xlsx](https://github.com/mdsohelmahmood/Statistics-in-R-Series/tree/main/Simple%20Logistic%20Regression)

Link to excel file multiple logistic regression: [adult-v3.xlsx](https://github.com/mdsohelmahmood/Statistics-in-R-Series/tree/main/Multiple%20Logistic%20Regression)

Mother and father’s bachelor’s education education is binary in the first file denoted as MADEG and PADEG and output is also binary which is denoted by DEGREE1 and represents the bachelor’s education of each individual.

In the second file, the education level is transformed to make it ordered as below.



Education levels in the dataset

The other variables of gender and race are binary here.

**Implementation in R**

The github gist that I used to compare different models is below. I have used anova function for this purpose. The necessary library is loaded first.

<https://gist.github.com/mdsohelmahmood/96b275ec66ac976c987d64a8ed6e3e61#file-gistfile1-txt>

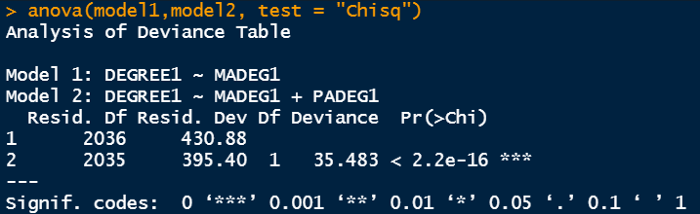
**Model Comparison and Result Interpretation**

In model1, the study question is:

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

In model2, I have added father’s education and the question becomes:

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



model1 and model2 comparison

Key observations are below:

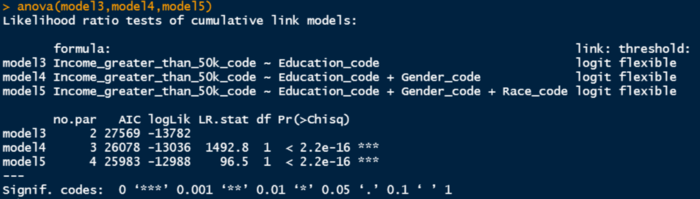
* The residual deviance for model1 is 430.88 and for model2, the value is 395.40 which is smaller. The deviance difference is 35.48 and it indicates that model2 is more robust than model1. Model2 brings us more information on each individual’s bachelor degree compared to model1. This residual deviance is the deviance of that particular model from the deviance of a saturated model where each observation brings one additional predictor parameter so that the model becomes perfect for that dataset. We can see the model2 is deviating less from perfection.
* The liklihood ratio chi-square test is significant here. Therefore we reject the null hypothesis and can come a conclusion that the model with two predictor variable is fits the data better.

In the subsequent models, I have developed ordinal logistics regressions using education, gender and race data as predictors and income level as response variable.

Model3 includes education data as predictor.

Model4 includes education and gender data as predictors.

Model5 includes education, gender and race data as predictors.



model3, model4 and model5 comparison

Since this is ordinal regression, the output window is different. Sometimes null model is referred to compare full or nested models. If we consider the null model as model0, we can define it as below:

model0 <- clm(Income\_greater\_than\_50k\_code ~ 1, data = data)

Key observations are below:

* The AIC statistics can be compared among all the models. Model5 has the smaller AIC value indicating more robustness for that model. Smaller the AIC value, better fit the model.
* The log-likelihood values for each model are shown in the next column. The likelihood ratio chi-square test statistics for moel4 is 1492.8 which is twice the difference between the log-likelihood values of these two models. Also p<0.001 which implies that model5 with two predictor variable fits better than model3 with one predictor variable. Therefore, when we incorporate gender into the model along with education, it has better predictability of each individual’s income level. Similarly when model4 and model5 are compared, the chi-square test statistics becomes 96.5 and p <0.001. Therefore model5, which includes race data, has better capability to describe income than model4.

**Conclusion**

Comparison between binary logistic regression models as well as comparison among ordinal logistic regression models has been demonstrated in this article with code implementation in R. There may be several models developed to address the same problem but comparison among those, gives us the option to check model robustness. We can check whether including one predictor is actually contributing to the robustness or not. It is also possible to obtain the extent of contribution. This is helpful is determining to deploy the final model.

**Acknowledgement for Dataset**