**Multiple Logistic Regression for Ordinal Variable and Predicted Probabilities in R**

**Statistics in R Series**

**Introduction**

We have covered simple logistic regression for ordinal variables in the last article which became quite interesting when looking at the result. In this article, I will discuss multiple logistic regression for similar ordinal data and also predict the probabilities using R packages. In ordinal logistic regression, the predictor variables can be either ordinal or binary or continuous and the response variable is ordinal.

For example, if we use the ordinal education levels to predict income which has only two levels of responses. We can have education levels from 1st grade to all the way doctorate degree and assign ordered numbers to carry out the regression. We can also use binary variables to predict income. For example, we can assign 1 to those people who have bachelors degree and 0 to those people who don’t have bachelors. In a sense, this can also be thought as ordinal variable with two kevels. Lastly, we can also use continuous variable like education years as predictor to predict income levels. he exact analysis is shown in the article below.

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

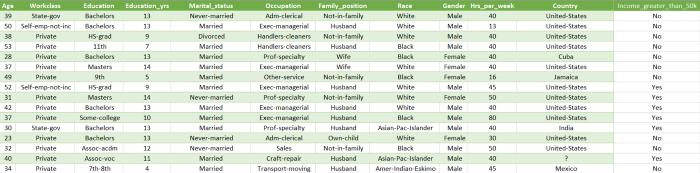
**[Statistics in R Series](https://towardsdatascience.com/simple-logistic-regression-for-ordinal-variables-in-r-1d95fa278c5e" \t "_blank)**

[towardsdatascience.com](https://towardsdatascience.com/simple-logistic-regression-for-ordinal-variables-in-r-1d95fa278c5e" \t "_blank)

Here, we are now interested to introduce multiple variables as predictors simultaneously.

**Dataset**

As a case study, we will use the [Adult Data Set](https://archive.ics.uci.edu/ml/datasets/adult) that is located in the UCI Machine Learning Repository. More than 30000 individual’s demographic data have been collected in this dataset, including their race, education, occupation, gender, salary, work 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

It is necessary to modify the given data in order to perform an ordinal logistic regression analysis. Nevertheless, let me begin by posing a question for the purpose of the study.

What is the impact of education level, gender and race on income?

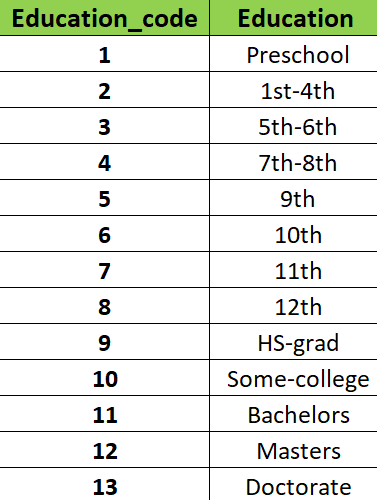
In order to answer this question, we will require data that is label encoded for both education and income level. As you can see in the dataset, there are different levels of education ranging from first grade through to doctorate level. As the income level is a binary one, it provides information on whether an individual has an income over $50000 or not. In this case, we are dealing with a binary response variable and an ordinal predictor variable (Education\_code column), two binary predictor variables (Gender\_code and Race\_code).

For gender, males are assigned 2 and females are assigned 1. Later we will swap this numbers and see the difference in the result. For race, we are interested to see if there is diference in income between white and non-white people. The non-white individuals are assigned 1 and white individuals are assigned 0. The analysis is going to be carried out in R, so let’s get started.

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

**Implementation in R**

We are going to use the same clm() function here. But at first, I would like to clarify the assigned values to different education level. It is shown below.



To add multiple variables, we can simply use the following format in clm() command.

model\_clm <- clm(Income\_greater\_than\_50k\_code ~ Education\_code + Gender\_code + Race\_code, data = data)

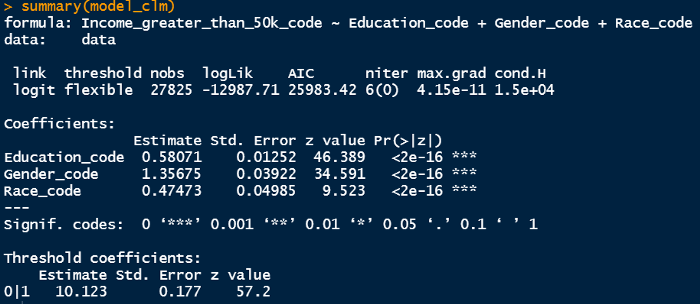
To show the effect of non-related variables, I have introduced a new column names “Random\_code” and we will see it’s effect on the result.

**Interpretation of Result**

In the code above, two models are shown. The first model evaluates the impact of education levels, gender and race on income. The second model introduces a new random variable.

1. Model 1: The predictor variables are ordinal education levels, binary gender and binary race variables. The response variable is binary income level.
2. Model 2: The predictor variables are ordinal education levels and a continuous random variable. The response variable is binary income level.

**Model 1 Result**

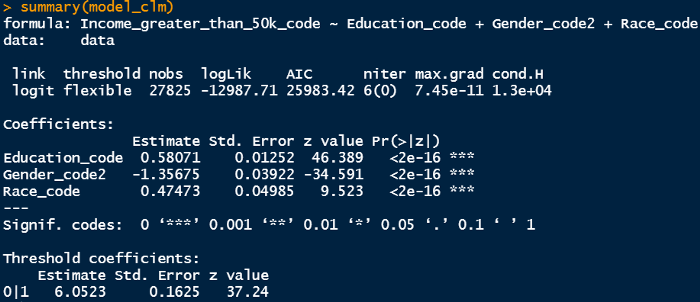


Model 1 result

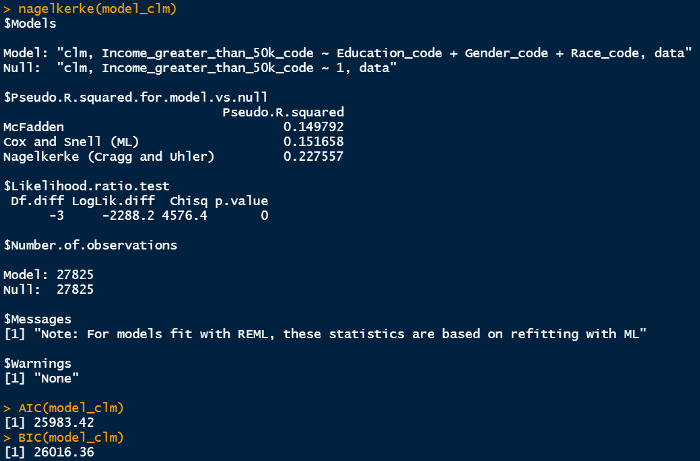
From the result summary, the key takeaways are:

* For every one level increase in education level, the logit or log odds of having income > $50000 increases by 0.581.
* For every one unit increase in Gender\_code, the logit or log odds of having income > $50000 increases by 1.357. This means if the individual under study is female (1) and we replace that individual by a male (2), the logit probability of higher incomes increases by 1.357.
* For every one unit increase in Race\_code, the logit or log odds of having income > $50000 increases by 0.475. This means if the individual under study is non-white (1) and we replace that individual by a white person (2), the logit probability of higher incomes increases by 0.475.
* All the predictor variables are significant (p<0.05).

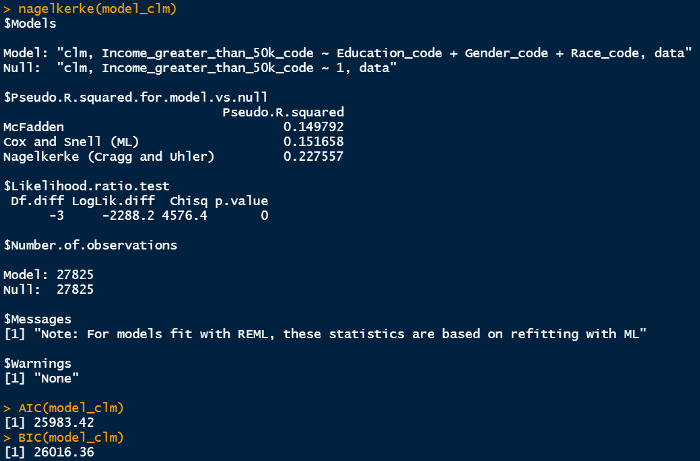
If we swap the gender value (assign 1 to male and 2 to female), the coefficient changes its sign as below



Model 1 result after swapping gender codes



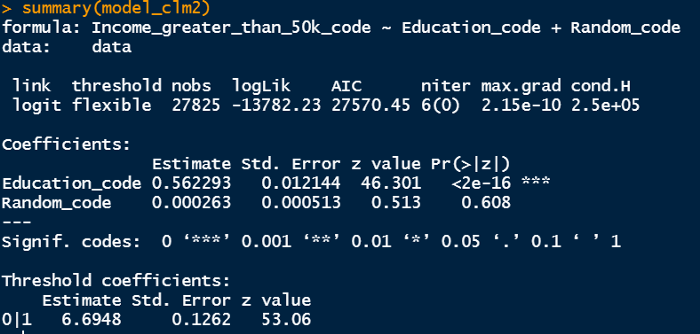
This shows us that for every one unit increase in Gender\_code, the logit or log odds of having income > $50000 decreases by 1.357. This means if the individual under study is male (1) and we replace that individual by a female (2), the logit probability of higher incomes decreases by 1.357. The same will happen if we swap the race codes.



Model 1 result

Regarding the pseudo R² value, McFadden gives us a value of 0.149, and we will compare this value with other models later on in the discussion. We will also compare the AIC and BIC statistics for the two different models since a single value for a single model does not have a great deal of significance when it comes to logistic regression.

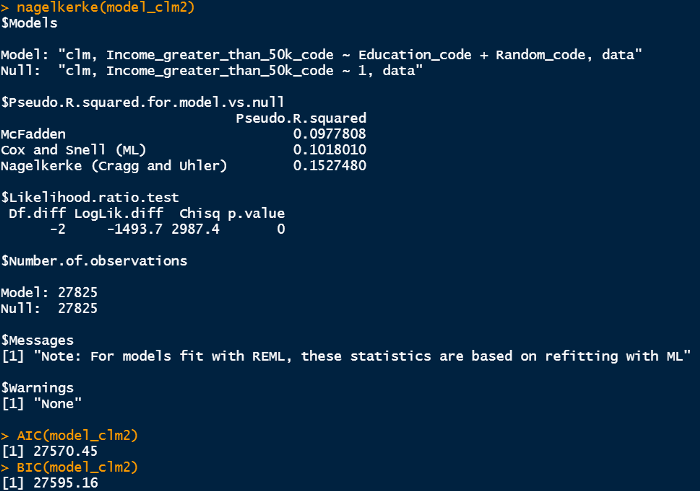
**Model 2 Result**



Model 2 result

From the result summary, the key takeaways are:

* For every one level increase in education level, the logit or log odds of having income > $50000 increases by 0.562.
* The random variable is not significant since p>0.05 in this case.

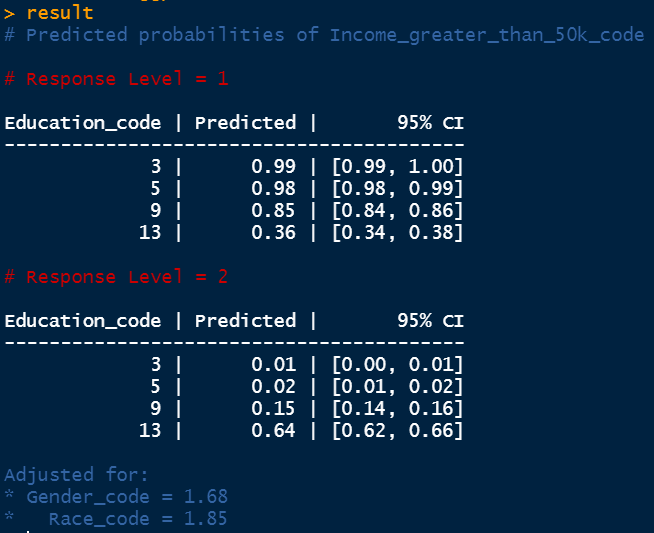


Model 2 result

The McFadden pseudo R² value ios smaller than model 1 and AIC/BIC statistics are higher than model 1. This means model 1 has better performance when we compare these goodness-of-fit statistics.

**Prediction**

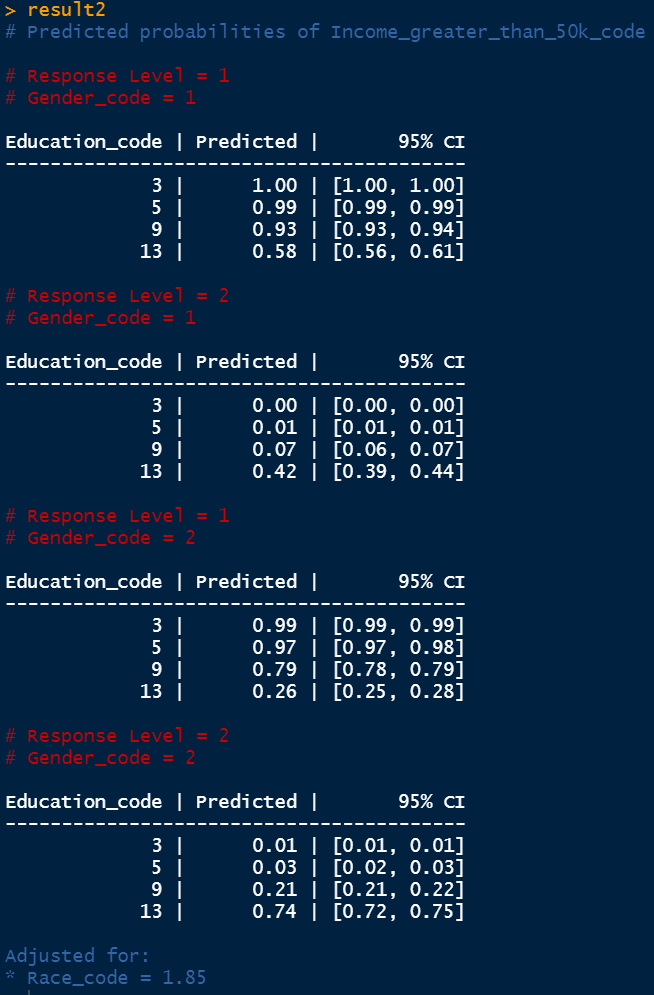
We are going to use ggpredict() command to effectively predict any given value for predictor variables. We need ggeffects libarary to implement that in R. Let’s predict a group of people’s income level given their education levels are 3,5,9 and 13.



Prediction result

Since we have two levels of income (income > $50000 and income ≤ $50000), the number of response levels are also 2 as shown in red above. The predicted probabilities for each education level are shown in the second column. When the education level is 3 (5th to 6th grade), the probability of income ≤ $50000 is 0.99 whereas if the education level is 13 (doctorate), the probability of income ≤ $50000 is 0.36. The same deduction can be made from the second response level prediction result.

We can introduce multiple predictor variable here. After including the gender, we obtain the following result.



Prediction result

Now we have 4 tables since gender has two levels of values similar to income level. For gender code 1 (female), if the individual has doctorate (education code 13), the predicted probability of income > $50000 is 0.42 whereas if the individual is male (gender code 2) and he has doctorate, the predicted probability of income > $50000 is 0.74. This indicates that females have some sort of unequal payment for the same level of education.

***Conclusion***

Two models have been discussed which incorporate logistic regression for an ordinal predictor variable and a binary response variable. The first model contains ordinal education variables, binary gender and race variables, as well as income variables. A second model consists of a continuous random variable, ordinal education variable, and a binary income variable. We compare each of these models on the basis of pseudo R2 and AIC/BIC statistics in order to determine their performance. Moreover, from a given set of data, predictions are made based on the probabilities and 95% confidence intervals of the predicted outcomes.

***Acknowledgement for Dataset***

[Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.](https://archive.ics.uci.edu/ml/datasets/adult)

Thanks for reading.