Baseline-Category Logits

A Paper Presented to the

Department of Mathematics and Statistics

College of Science

De La Salle University

In Partial Fulfillment

of the Requirements for the Course

STT245A - N01

Submitted By:

ALCALA, Chester Paulo B.

BOCATO, Lyka Janine A.

ESTACIO, Jenerose A.

GO, Albert Thimothy M.

MEDALLO, Kiah Kyhle S.

July, 2022

INTRODUCTION

Ordinary Logistic Regression

Ordinary Logistic Regression is used when the dependent variable has two possible outcomes. Usually, the response variable is coded as 1 or 0, with 1 being the event of a "success" and 0 being the event of a "failure" (Agresti, 2007). Using this model, it is possible to estimate the probability of occurrence of the characteristic value of the dependent variable.

Recall that Logit p(x) is

$$log\frac{p(x)}{1-p(x)} = \log\frac{P[Y=1]}{P[Y=0]} = \log \text{ odds}$$

From the ordinary logistic model, the probability of success, Y=1 is derived as follows:

$$log\frac{p(x)}{1-p(x)} = \alpha + \beta \text{x}$$

$$\frac{p(x)}{1-p(x)} = \exp(\alpha + \beta \text{x})$$

$$P[Y=1] = \frac{e^{\alpha+\beta_x}}{1+e^{\alpha+\beta_x}}$$

Polychotomous Responses

In ordinary logistic regression, the response variable Y is dichotomous or binary. When response variable Y is polychotomous, an extension is to make dummy variables and form many ordinary logistic regression models equal to the number of categories in Y (Fienberg, 2004). Each category of the response in Y corresponds to 1 dummy variable, thus in the data preparation phase, one must make *J* dummy variables if *Y* has *J* categories.

The disadvantage of forming many ordinary logistic models arises in comparing the odds of two specific categories (Fienberg, 2004). For instance, if the response variable Y has three response categories *A*, *B*, and *C*, it will be challenging to compare the odds of the response being *A* over *B* since the ordinary logistic model only models the odds of Y being *A* over *B* or *C*, and B over *A* or *C*. It will be difficult to get an estimation of the ratio of the probability of getting *A* over *B*. In order to estimate these odds more conveniently, one may use the Baseline-Category Logits.

Baseline Category Logits (Adopted from Agresti (2007)).

In order to model a nominal response variable with more than 2 possible responses, a better model than using many logistic models is the use of Baseline Category Logit Models (Agresti, 2007). Suppose the response variable has J total number of possible categories. When the last category (J) is the baseline, the baseline category logits are of the form

$$\log\left(\frac{\pi_k}{\pi_J}\right) \text{ for } k = 1, 2, ..., J - 1$$

For example, in a data set where the response has 5 possible categories, there are 4 baseline category logits, given by:

$$\log\left(\frac{\pi_1}{\pi_5}\right), \log\left(\frac{\pi_2}{\pi_5}\right), \log\left(\frac{\pi_3}{\pi_5}\right), \text{and } \log\left(\frac{\pi_4}{\pi_5}\right)$$

The choice of the baseline category does not always have to be the last category in the list – any of the possible categories in the response variable Y can be used as the baseline (Agresti, 2007). For the purposes of the discussion of the model in this paper, the last category is used as the baseline.

The Baseline Logit Model

The Logistic Model for Baseline Category Logits has $J - 1$ equations, each corresponding to the $J - 1$ categories that are compared to the baseline. In general, if there are multiple predictors, $X_1, X_2, ... X_p$ then each of the equations in the model has the form

$$\log\left(\frac{\pi_k}{\pi_J}\right) = \beta_{0k} + \beta_{1k}X_1 + \beta_{2k}X_2 + ... + \beta_{pk}X_p \text{ for } k = 1, 2, ..., J - 1$$

Where β_{pk} corresponds to the coefficient of predictor p for the kth category logit (Agresti, 2007). Note that if $J = 2$, the model simplifies to the ordinary Logistic Regression Model. Furthermore, each predictor corresponds to $J - 1$ parameters, since each equation of the $J - 1$ equations are linear combinations of all of the predictors in the analysis. Thus, each predictor in the model corresponds to $J - 1$ degrees of freedom in the model.

The Baseline Logit Model can also be used to compare the odds of two specific categories in the response. This is possible since the equations in the model can be used to determine the equations to compare the log of the odds for the other pairs of categories (Agresti, 2008). For instance, consider two categories, K_1 and K_2, where $K_1, K_2 \in \{0, 1, 2,..., J - 1\}$, $K_1 \neq K_2$, then the logarithm of their odd ratio can be computed as follows:

$$\begin{aligned}\log\left(\frac{\pi_{K_1}}{\pi_{K_2}}\right) &= \log\left(\frac{\pi_{K_1}}{\pi_{K_J}} \times \frac{\pi_{K_J}}{\pi_{K_2}}\right) \\ &= \log\left(\frac{\pi_{K_1}}{\pi_{K_J}}\right) + \log\left(\frac{\pi_{K_J}}{\pi_{K_2}}\right) \\ &= \log\left(\frac{\pi_{K_1}}{\pi_{K_J}}\right) - \log\left(\frac{\pi_{K_2}}{\pi_{K_J}}\right)\end{aligned}$$

Then, using the form of the equations in the baseline category logits model, it follows that:

$$\begin{aligned}\log\left(\frac{\pi_{K_1}}{\pi_{K_2}}\right) &= (\beta_{0K_1} + \beta_{1K_1}X_1 + \beta_{2K_1}X_2 + ... + \beta_{pK_1}X_p) - (\beta_{0K_2} + \beta_{1K_2}X_1 + \beta_{2K_2}X_2 + ... + \beta_{pK_2}X_p) \\ &= (\beta_{0K_1} - \beta_{0K_2}) + (\beta_{1K_1} - \beta_{1K_2})X_1 + ... + (\beta_{pK_1} - \beta_{pK_2})X_p\end{aligned}$$

The coefficients of each parameter for the resulting equation is simply the difference of the parameters of each predictor, computed using category *J* as the baseline. This also holds true if another category is used as the baseline as long as the coefficients of both equations use the same baseline category.

Estimating Response Probabilities

The model for the baseline logits model can be used to estimate the response probability of each category (Agresti, 2007). The equation used to estimate the response probability for the jth category is:

$$\pi_j = \frac{\exp[\beta_{0j} + \beta_{1j}X_1 + ...\beta_{pj}X_p]}{\sum_{l=1}^{J}\exp[\beta_{0l} + \beta_{1l}X_1 + ...\beta_{pl}X_p]}$$

Notice that the coefficients for the predictors of category J are 0, since the computation of the coefficients uses category J as the baseline. Thus, another equation that can be used to estimate response probabilities is:

$$\pi_j = \frac{\exp[\beta_{0j} + \beta_{1j}X_1 + ...\beta_{pj}X_p]}{1 + \sum_{l=1}^{J-1}\exp[\beta_{0l} + \beta_{1l}X_1 + ...\beta_{pl}X_p]}$$

METHODOLOGY

Fitting a Baseline-Category Logits Model in SAS 9.4

In order to fit a baseline-category logits model, the researchers used SAS/STAT 9.4. Similar to ordinary logistic regression in SAS 9.4, the `LOGISTIC` procedure was also used to fit a baseline-category logit model (SAS, 2022). However instead of a logit link, the link specified in the `MODEL` statement is *glogit,* or generalized logits. Backward selection was also implemented using the selection option in the `MODEL` statement. In backward selection, the variable that is eliminated is the variable that is not significant according to the overall analysis of effects, also called the type III analysis of effects in the output.

The baseline category selected by the `LOGISTIC` procedure of SAS is the category of the last variable in the dataset by default. However, this can be changed by using the `ref=` option in the `MODEL` statement (SAS, 2022). The procedure also calculates odds ratios assuming an increment of 1 for each predictor by default. The increments can also be changed by specifying the increments of each predictor using the `UNITS` statement of the procedure (SAS, 2022).

Type of House using some Parts of the Food, Income, and Expenditure Survey 2016.

The researchers demonstrate the use of the baseline-category logits model in SAS by predicting the type of residential building a person has based on their spending habits using parts of the Food, Income, and Expenditure Survey 2016. The dataset is posted by Francis Paul Flores in Kaggle in 2017. The dataset has 60 variables, however, for the sake of the discussion of the paper, a part of the dataset has been selected by the researchers.

In this paper, the response variable used in the model is the type of household of a respondent. This has three categories: Single House, Duplex, and Multi-unit Residential. These categories are assigned as categories 1, 2, and 3 respectively. Table 1 below shows the frequency distribution of the categories.In the model, the researchers used Multi-unit Residential Category, House = 3, as the baseline category.

Table 1: Frequency Distribution of the Types of House

House	Frequency	Percentage
1 - Single House	39069	94.18%
2 - Duplex	1084	2.61%
3 - Multi Unit Residential	1329	3.20%

In this baseline-category logits model, the researchers used 8 predictors, 7 of which are quantitative and 1 is qualitative. The 7 quantitative variables are the total monthly expenditures on food, clothing, utilities, medical care, education, transportation, and communication. The odds ratio estimate computation of SAS used increments of 2000 in all of the quantitative variables. The qualitative variable used in the study is the type of household. In the data set, there are 3 types of households: Single Family, Extended Family, and Two or more Non-related Persons. In order to properly incorporate the type of household in the analysis, 2 dummy variables are generated: Extended Family, and Non-related Persons. In this case, the Single Family category for type of household was used as the baseline.

The full model to predict the type of house based on the predictors of the study has the form

$$\begin{aligned}\log\left(\frac{\hat{\pi}_k}{\hat{\pi}_3}\right) =& \hat{\beta}_{0k} + \hat{\beta}_{NonRelated,k}X_{NonRelated} + \hat{\beta}_{ExtendedFamily,k}X_{ExtendedFamily} + \hat{\beta}_{FoodExp,k}X_{FoodExp} \\ &+ \hat{\beta}_{ClothingExp,k}X_{ClothingExp} + \hat{\beta}_{HousingExp,k}X_{HousingExp} + \hat{\beta}_{MedicalExp,k}X_{MedicalExp} \\ &+ \hat{\beta}_{TransportExp,k}X_{TransportExp} + \hat{\beta}_{CommExp,k}X_{CommExp} + \hat{\beta}_{EducationExp,k}X_{EducationExp}\end{aligned}$$

for $k = 1, 2$. The model contains two equations – one for each of the two baseline category logits. Thus, each predictor corresponds to two degrees of freedom in the model, since each predictor corresponds to two parameters.

RESULTS

Full Model

Tables 2 and 3 below show the Model Fit Statistics and the hypothesis tests for significance of the full model. The results reported that the full model is significant for predicting the type of house building ($\Delta G^2 = 541.0808$, p < 0.0001). Therefore, this model has a good fit, and can be used to predict the type of house.

Table 2: Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	21734.063	21228.982
SC	21751.329	21401.642
-2 Log L	21730.063	21188.982

Table 3: Testing Global Null Hypothesis

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	541.0808	18	<.0001*
Score	789.8513	18	<.0001*
Wald	616.7940	18	<.0001*

Table 4 shows the analysis of the overall significance of each effect in the model. The degrees of freedom of the model is 18 since there are 9 predictors – 7 quantitative and 2 dummy variables. Each predictor corresponds to two parameters, one parameter for each equation in the model. The results show that the effects that are not significant in the overall model are the dummy variable Non-Related ($\chi^2 = 4.9808$, $p = 0.0829$) and quantitative variable Education Expenditure ($\chi^2 = 4.3065$, $p = 0.1161$) at the 5% significance level. The other effects are found to be significant for predicting the type of house building.

Table 4: Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Non-Related	2	4.9808	0.0829
Extended Family	2	27.2518	<.0001*
Total Food Expenditure	2	179.6910	<.0001*
Clothing Expenditure	2	14.1777	0.0008*
Housing Expenditure	2	22.4787	<.0001*
Medical Expenditure	2	7.2293	0.0269*

Transportation Expenditure	2	7.5598	0.0228*
Communication Expenditure	2	47.7895	<.0001*
Education Expenditure	2	4.3065	0.1161

Table 5.1 and 5.2 below shows the coefficients of the effects of the equations for the two baseline-category logits comparing the odds of owning house 1 compared to house 3, and the odds of owning house 2 compared to house 3 respectively. Since the type of house building has three possible responses, there are two baseline category logits, thus the model has 2 equations. The equations are:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_3}\right) =& 4.014 - 0.647X_{NonRelated} + 0.337X_{ExtendedFamily} - 0.00000729X_{FoodExp} \\ &+ 0.000015X_{ClothingExp} - 0.00000147X_{HousingExp} + 0.0000373X_{MedicalExp} \\ &+ 0.00000416X_{TransportExp} - 0.00003X_{CommExp} + 0.00000294X_{EducationExp}\end{aligned}$$

and

$$\begin{aligned}\log\left(\frac{\hat{\pi}_2}{\hat{\pi}_3}\right) =& 0.0674 - 0.360X_{NonRelated} + 0.338X_{ExtendedFamily} - 0.000004X_{FoodExp} \\ &+ 0.000011X_{ClothingExp} + 0.000000435X_{HousingExp} + 0.0000343X_{MedicalExp} \\ &+ 0.00000236X_{TransportExp} - 0.00002X_{CommExp} + 0.000000953X_{EducationExp}\end{aligned}$$

From the two equations of the model, the equation for $\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right)$ can also be found through subtraction of the coefficients of the two models. The equation for the said logit is given by:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) =& 3.947 - 0.287X_{NonRelated} - 0.001X_{ExtendedFamily} - 0.00000329X_{FoodExp} \\ &+ 0.000004X_{ClothingExp} - 0.000001905X_{HousingExp} + 0.00003X_{MedicalExp} \\ &+ 0.00000180X_{TransportExp} - 0.00001X_{CommExp} + 0.00000199X_{EducationExp}\end{aligned}$$

Table 5.1: Analysis of Maximum Likelihood Estimates

Effect	House	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	4.0139	0.0524	5857.4533	<.0001*
Non-Related	1	1	-0.6469	0.2986	4.6923	0.0303*
Extended Family	1	1	0.3369	0.0646	27.2076	<.0001*
Total Food Expenditure	1	1	-7.29E-6	5.587E-7	170.3936	<.0001*
Clothing Expenditure	1	1	0.000015	4.039E-6	13.7673	0.0002*
Housing Expenditure	1	1	-1.47E-6	4.05E-7	13.2137	0.0003*
Medical Expenditure	1	1	3.727E-6	1.389E-6	7.2029	0.0073*
Transportation Expenditure	1	1	4.155E-6	1.568E-6	7.0202	0.0081*
Communication Expenditure	1	1	-0.00003	4.025E-6	46.7713	<.0001*
Education Expenditure	1	1	2.294E-6	1.194E-6	3.6941	0.0546

Table 5.2: Analysis of Maximum Likelihood Estimates

Effect	House	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	2	1	0.0674	0.0776	0.7546	0.3850
Non-Related	2	1	-0.3597	0.4987	0.5201	0.4708
Extended Family	2	1	0.3377	0.0919	13.5140	0.0002*
Total Food Expenditure	2	1	-4E-6	8.779E-7	20.7729	<.0001*
Clothing Expenditure	2	1	0.000011	5.75E-6	3.5281	0.0603
Housing Expenditure	2	1	4.345E-7	5.052E-7	0.7397	0.3897
Medical Expenditure	2	1	3.425E-6	1.687E-6	4.1296	0.0421*
Transportation Expenditure	2	1	2.356E-6	2.33E-6	1.0224	0.3119
Communication Expenditure	2	1	-0.00002	6.405E-6	8.3346	0.0039*
Education Expenditure	2	1	9.532E-7	1.827E-6	0.2722	0.6019

In the model, since the dummy variable *NonRelated* is not significant, these effects are not interpreted. The resulting odds ratios for the qualitative predictors are shown on table 6. The confidence intervals of the odds ratio of the qualitative is also visualized in figure 1.

Table 6: Odds Ratio Estimates and Wald Confidence Intervals

Odds Ratio	Estimate	95% Confidence Limits	
House 1: Extended Family 0 vs 1	0.714	0.629	0.810
House 2: Extended Family 0 vs 1	0.713	0.596	0.854
House 1: Non Related 0 vs 1	1.910	1.064	3.429
House 2: Non Related 0 vs 1	1.433	0.539	3.808

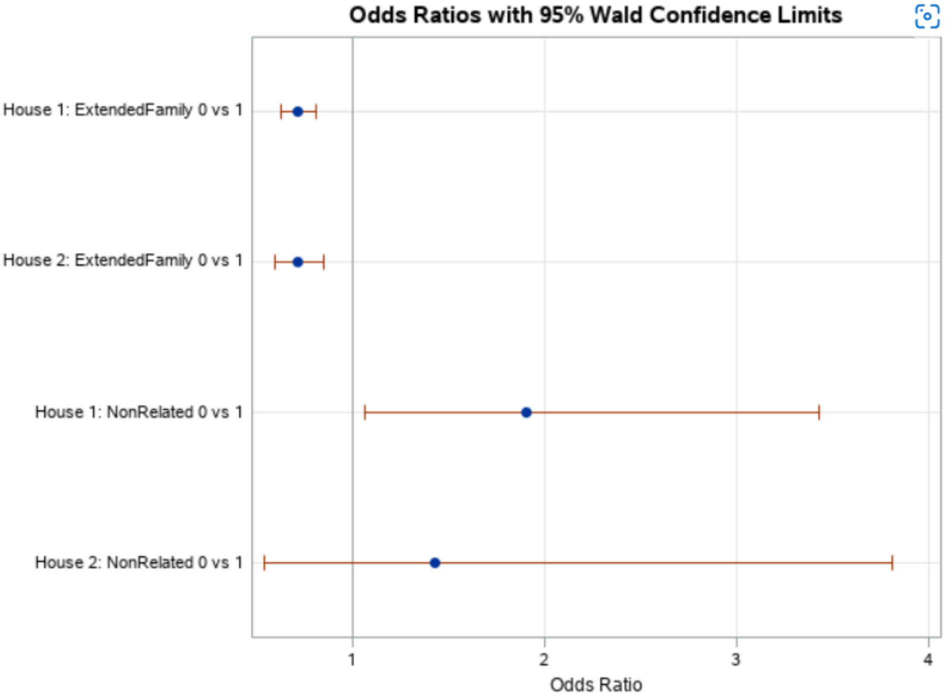

Figure 1. Odds Ratio for the Dummy Variables in the Full model

From the estimates given in the model, households with housing extended family members are more likely to be either single or duplex houses than multi-unit residential areas holding all other variables constant when compared to households with single families. To be exact, the model estimates that the odds of single family households owning single houses against multi residential units are 0.714 times that of the odds for households with extended families. Meanwhile, the model also estimates that the odds of households housing single families owning a duplex against owning a multi residential area is 0.713 times as of households with extended families.

After interpreting the qualitative predictors, the quantitative predictors are also interpreted. The odds ratio estimates are shown in table 7.1 and 7.2, while the confidence intervals of these odds ratios are visualized in figure 2. Since education expenditure is not a significant predictor in the model, the effect of education expenditure is not interpreted.

Table 7.1: Odds Ratio Estimates and Wald Confidence Intervals

Effect	House	Unit	Estimate	95% Confidence Limits	
Total Food Expenditure	1	2000.0	0.986	0.983	0.988
Clothing Expenditure	1	2000.0	1.030	1.014	1.047
Housing Expenditure	1	2000.0	0.997	0.995	0.999
Medical Expenditure	1	2000.0	1.007	1.002	1.013
Transportation Expenditure	1	2000.0	1.008	1.002	1.015
Communication Expenditure	1	2000.0	0.946	0.932	0.961
Education Expenditure	1	2000.0	1.005	1.000	1.009

Table 7.2: Odds Ratio Estimates and Wald Confidence Intervals

Effect	House	Unit	Estimate	95% Confidence Limits	
Total Food Expenditure	2	2000.0	0.992	0.989	0.995
Clothing Expenditure	2	2000.0	1.022	0.999	1.045
Housing Expenditure	2	2000.0	1.001	0.999	1.003
Medical Expenditure	2	2000.0	1.007	1.000	1.014
Transportation Expenditure	2	2000.0	1.005	0.996	1.014
Communication Expenditure	2	2000.0	0.964	0.940	0.988
Education Expenditure	2	2000.0	1.002	0.995	1.009

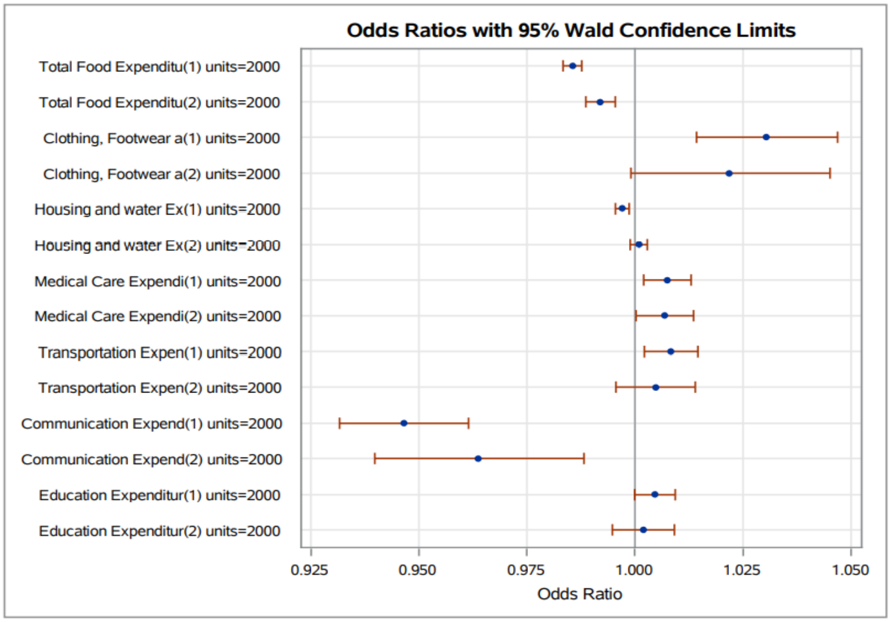

Figure 2. Odds Ratio for the Quantitative Predictors of the full model

The model estimates that increases in total food expenditure and communication expenditure decreases the odds of owning either a single house or duplex against owning a multi-residential unit. Specifically, an increase of 2000 pesos in total food expenditures is associated with a significant multiplicative decrease in the odds of owning a single house and a duplex with factors of 0.986 and 0.992 respectively when compared to owning a multi residential area. Meanwhile, an increase of 2000 pesos in communication expenses is associated with a decreasing factor of 0.946 and 0.964 in the odds of owning a single house and duplex respectively against owning a multi residential unit.

Results also show that increases in expenditure in medical care increases the odds of owning either a single house or duplex against owning a multi residential unit. Specifically, increments of 2000 pesos in medical expenses is associated with a multiplicative increase in both the odds of owning a single house and duplex against owning a multi residential unit by a factor of 1.007.

Some of the factors also show no significant changes in the odds ratio for owning one type of house, but significant changes in the odds ratio in owning another type of house. First, an increase of 2000 pesos to clothing expenditures increases the odds of owning a single house to owning a multi residential area by a factor of 1.03. However it was found that increments of 2000 pesos in clothing expenditures does not significantly increase the odds of owning a duplex against owning a multi residential area. It is also estimated that an increase of 2000 pesos in housing and water expenditures is associated with a decrease in the odds of owning a single house against owning a multi residential unit by a factor of 0.997. However, increments of 2000 pesos do not significantly affect the odds of owning a duplex against owning a multi-residential area. Lastly, an increase of 2000 pesos in the transportation expenses increase the odds of owning a single house against owning a multi residential unit by a factor of 1.008. However, increases in the same expenses does not significantly change the odds ratio of owning a duplex against owning a multi residential unit.

Backward Selection

In the full model, there were effects that were not significant in the model. Since a simpler model to predict the type of house is desired, backward selection was implemented to remove non-significant effects. Table 8 below summarizes the model fit statistics of them candidate models formed using backward selection.

Table 8: Backward Selection

Model	G^2	ΔG^2 with full model	df	Δdf	Predictors Removed	Decision
Model 1: Full Model	21188.982	0	18	0	-	-
Model 2	21193.634	4.652	16	2	Education Expenditure	Not significant
Model 3	21197.873	8.891	14	4	Education Expenditure, NonRelated	Not significant

The results show that the final model chosen by backward selection is the model without the predictors education expenditure and non-related. Note that in contrast to ordinary logistic regression, the decrease in the degrees of freedom of the model with each predictor removed corresponds to a decrease of 2 degrees of freedom (df) in the model instead of 1. As explained previously, this is because each predictor in the model corresponds to 2 parameters: one for each of the 2 equations in the model.

In order to ensure that the fit of the models do not significantly differ with the full model in terms of their fit, the difference in the G^2 statistic is used to compare the models. The results show that Model 2, the model with Education Expenditure removed, is not significantly different when compared to the full model ($\Delta G^2 = 4.652 < \chi^2_{0.05, df=2} = 5.991$). Also, Model 3, the model with both Education Expenditure and NonRelated removed, is not significantly different when compared to the full model ($\Delta G^2 = 8.891 < \chi^2_{0.05, df = 4}$= 9.488). Since the two models are not significantly different from the full model in terms of fit, and the third model has the least number of parameters, the third model is chosen as the best model, and used for interpretation.

Tables 9 and 10 below provide the Model Fit Statistics and the hypothesis tests for Model 3 significance. According to the findings, Model 3 is significant for predicting the type of house building ($\Delta G^2 = 532.1900$, p < 0.0001). As a result, this model fits well and may be used to determine the type of house.

Table 9: Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	21734.063	21229.873
SC	21751.329	21368.001
-2 Log L	21730.063	21197.873

Table 10: Testing Global Null Hypothesis

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	532.1900	14	<.0001*
Score	773.3515	14	<.0001*
Wald	608.4338	14	<.0001*

Table 11 below summarizes the overall significance of each effect in the model. The model has 14 degrees of freedom since there are 7 predictors – 6 quantitative and 1 dummy variable. Each predictor corresponds to two parameters in lhe model: one parameter for each equation. Since all predictors are significant at the 0.05 significance level, no further variable deletions can be made.

Table 11: Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Extended Family	2	27.8034	<.0001*
Total Food Expenditure	2	179.2448	<.0001*
Clothing Expenditure	2	14.9669	0.0006*
Housing Expenditure	2	22.1722	<.0001*
Medical Expenditure	2	6.7942	0.0335*
Transportation Expenditure	2	9.6711	0.0079*
Communication Expenditure	2	44.4050	<.0001*

The coefficients of the equations' effects for Model 3's two baseline-category logits are shown in table 12.1 and 12.2. Because the type of house building has three possible responses, there are two baseline category logits, resulting in two equations in the model shown below.

$$
\begin{aligned}
\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_3}\right) =& 4.0089 + 0.3387X_{ExtendedFamily} - 0.00000723X_{FoodExp} \\
&+ 0.000015X_{ClothingExp} - 0.00000146X_{HousingExp} + 0.000003582X_{MedicalExp} \\
&+ 0.000004612X_{TransportExp} - 0.00003X_{CommExp}
\end{aligned}
$$

and

$$
\begin{aligned}
\log\left(\frac{\hat{\pi}_2}{\hat{\pi}_3}\right) =& 0.0657 + 0.3399X_{ExtendedFamily} - 0.00000399X_{FoodExp} \\
&+ 0.000011X_{ClothingExp} + 0.0000004245X_{HousingExp} + 0.000003311X_{MedicalExp} \\
&+ 0.000002548X_{TransportExp} - 0.00002X_{CommExp}
\end{aligned}
$$

Table 12.1: Parameter Estimates and Wald Confidence Intervals

Effect	House	Estimate	95% Confidence Limits	
Intercept	1	4.0089	3.9062	4.1117
Extended Family	1	0.3387	0.2127	0.4647
Total Food Expenditure	1	-7.23E-6	-8.32E-6	-6.15E-6
Clothing Expenditure	1	0.000015	7.496E-6	0.000023
Housing Expenditure	1	-1.46E-6	-2.25E-6	-6.67E-7
Medical Expenditure	1	3.582E-6	8.849E-7	6.28E-6
Transportation Expenditure	1	4.612E-6	1.567E-6	7.637E-6
Communication Expenditure	1	-0.00003	-0.00003	-0.00002

Table 12.2: Parameter Estimates and Wald Confidence Intervals

Effect	House	Estimate	95% Confidence Limits	
Intercept	2	0.0657	-0.0865	0.2178
Extended Family	2	0.3399	0.1605	0.5192
Total Food Expenditure	2	-3.99E-6	-5.7E-6	-2.28E-6
Clothing Expenditure	2	0.000011	-3.5E-7	0.000022
Housing Expenditure	2	4.245E-7	-5.69E-7	1.418E-6
Medical Expenditure	2	3.311E-6	2.551E-8	6.596E-6
Transportation Expenditure	2	2.548E-6	-1.95E-6	7.041E-6
Communication Expenditure	2	-0.00002	-0.00003	-5.46E-6

The equation for $\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right)$ may also be derived by subtracting the coefficients of the two models from the two equations of the previous model. The equation for the aforementioned logit is as follows:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) =&4.0233 - 0.0012X_{ExtendedFamily} - 0.00000324X_{FoodExp}\\ &+ 0.000004X_{ClothingExp} - 0.0000018845X_{HousingExp} + 0.000000217X_{MedicalExp}\\ &+ 0.000002064X_{TransportExp} - 0.00001X_{CommExp}\end{aligned}$$

According to the model, households with extended family members are more likely to be single or duplex houses than multi-unit residential areas when all other factors are held constant as compared to households with single families. Table 13 shows the odds ratio estimates for the effect of Extended Family with the corresponding 95% confidence interval, while Figure 3 visualizes these odds and its 95% confidence interval. To be more precise, the model estimates that the probabilities of extended family members having single houses against multi residential units are $e^{0.3387}$= 1.403 times those of single families. Meanwhile, the model forecasts that the odds of households with extended family members buying a duplex against having a multi residential unit are $e^{0.3399}$= 1.405 times higher than the odds of households with single families.

Table 13. Odds ratios and Confidence intervals for Extended Family

Odds Ratio	Estimate	95% Confidence Limits	
House 1: Extended Family	1.403	1.27	1.592
Hosue 2: Extended Family	1.405	1.174	1.681

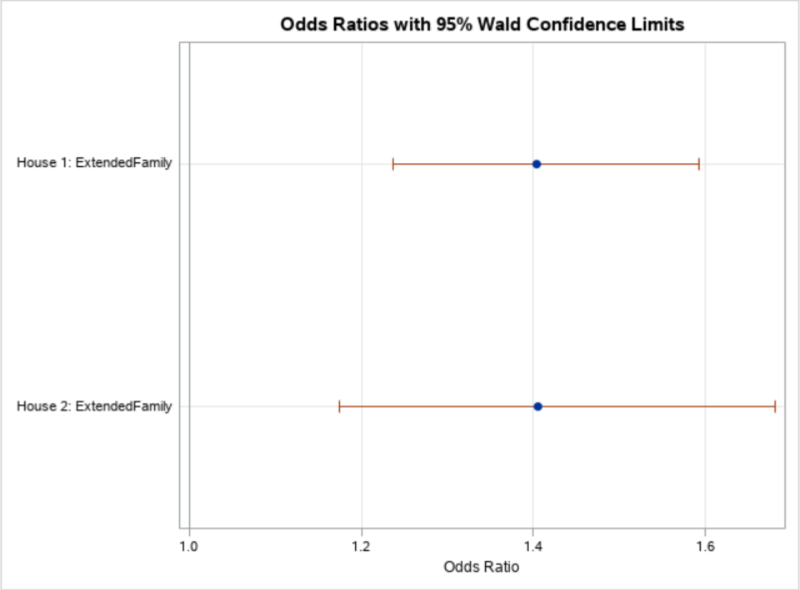

Figure 3. Odds ratio for the dummy variable extended family in the final model

The quantitative predictors are interpreted after the qualitative predictors. The odds ratio estimates are presented in table 14, and the confidence intervals for these odds ratios are depicted in figure 4.

Table 14.1: Odds Ratio Estimates and Wald Confidence Intervals

Effect	House	Unit	Estimate	95% Confidence Limits	
Total Food Expenditure	1	2000.0	0.986	0.983	0.988
Clothing Expenditure	1	2000.0	1.031	1.015	1.048
Housing Expenditure	1	2000.0	0.997	0.996	0.999
Medical Expenditure	1	2000.0	1.007	1.002	1.013
Transportation Expenditure	1	2000.0	1.009	1.003	1.015
Communication Expenditure	1	2000.0	0.949	0.934	0.964

Table 14.2: Odds Ratio Estimates and Wald Confidence Intervals

Effect	House	Unit	Estimate	95% Confidence Limits	
Total Food Expenditure	2	2000.0	0.992	0.989	0.995
Clothing Expenditure	2	2000.0	1.022	0.999	1.046
Housing Expenditure	2	2000.0	1.001	0.999	1.003
Medical Expenditure	2	2000.0	1.007	1.000	1.013
Transportation Expenditure	2	2000.0	1.005	0.996	1.014
Communication Expenditure	2	2000.0	0.965	0.941	0.989

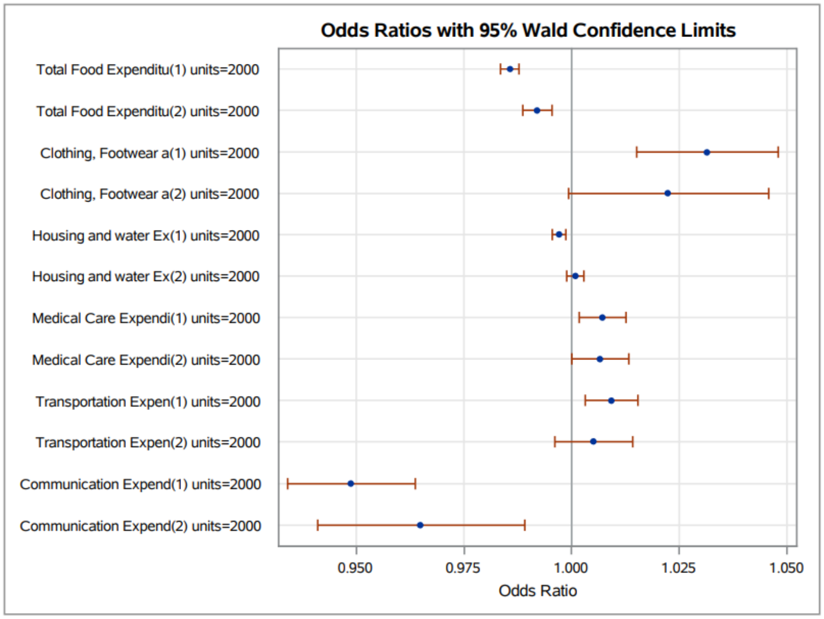

Figure 4. Odds Ratio for the Quantitative Predictors of the final model

According to the model, increases in total food and communication expenses reduce the likelihood of having a single house or duplex over owning a multi-residential unit. When compared to owning a multi-residential area, an increase in total food expenditures of 2000 pesos is related with a substantial multiplicative decline in the probabilities of owning a single house and a duplex with factors of 0.986 and 0.992, respectively. Meanwhile, a 2000 peso increase in communication expenses is connected with a decreasing factor of 0.949 and 0.965 in the probability of having a single house or duplex over owning a multi-residential unit.

Additionally, increases in medical care expenditure boost the likelihood of having a single house or duplex over owning a multi-residential unit. In particular, a 2000 peso increase in medical expenditures is connected with a 1.007 multiplicative increase in the probability of having a single house or duplex over owning a multi-residential unit.

It is also projected that a 2000 peso rise in housing and water expenses is linked with a 0.997 percent drop in the likelihood of having a single house over owning a multi residential unit. However, an increase of 2000 pesos has little effect on the probability of owning a duplex as against a multi-residential area, only by a ratio of 1.001. Finally, a 2000 peso increase in transportation costs increases the probability of having a single house over owning a multi residential unit by a ratio of 1.009. However, increases in the same expenditures have no significant effect on the chances ratio of owning a duplex over owning a multi residential unit, which is only by a factor of 1.005.

Estimating Response Probabilities

The estimates in Table 12 contrast “Single House” and “Duplex” to “Multi-unit Residential” as the base- line category. The estimated probabilities (12) of the outcomes (Single House, Duplex, and Multi-unit Residential) equal

$$\hat{\pi}_1 = \frac{\exp[4.0089 + 0.3387X_{ExtendedFamily} + \ldots -0.00003X_{CommExp}]}{1 + \exp[4.0089 + 0.3387X_{ExtendedFamily} + \ldots -0.00003X_{CommExp}] + exp[0.0657 + 0.3399X_{ExtendedFamily} + \ldots -0.00002X_{CommExp}]}$$

$$\hat{\pi}_2 = \frac{\exp[0.0657 + 0.3399X_{ExtendedFamily} + \ldots -0.00002X_{CommExp}]}{1 + \exp[4.0089 + 0.3387X_{ExtendedFamily} + \ldots -0.00003X_{CommExp}] + exp[0.0657 + 0.3399X_{ExtendedFamily} + \ldots -0.00002X_{CommExp}]}$$

$$\hat{\pi}_3 = \frac{1}{1 + \exp[4.0089 + 0.3387X_{ExtendedFamily} + \ldots -0.00003X_{CommExp}] + exp[0.0657 + 0.3399X_{ExtendedFamily} + \ldots -0.00002X_{CommExp}]}$$

The “1” in each denominator and in the numerator of $\hat{\pi}3$ represents the e$^{\hat{\alpha}_3 + \hat{\beta}_3 x}$ for $\hat{\alpha}_3 = \hat{\beta}_3 = 0$ with the baseline category

For example, a person that does not have an extended family, total food expenditure is ₱117,848, clothing expenditure is ₱4,607, housing expenditure is ₱63,636, medical expenditure is ₱3,457, transportation expenditure is ₱4,776, and communication expenditure is ₱2,880, the estimated probability that type of residential building of a respondent is a Multi-unit Residential equal to

$$\hat{\pi}_3 = \frac{1}{1 + \exp[4.0089 + 0.3387(0) + \ldots - 0.00003(2,880)] + exp[0.0657 + 0.3399(0) + \ldots - 0.00002(2,880)]} = 0.04$$

Likewise, it can also be shown that that $\hat{\pi}_1 = 0.93$ and $\hat{\pi}_2 = 0.03$. Presumably, the type of residential building the person has is a single house.

SUMMARY AND CONCLUSIONS

The Baseline Category Logit is a better model to use when the response variable is nominal with more than 2 possible responses compared to using many ordinary logistic models. An application of the Baseline-Category Logit model was shown by predicting the type of residential building a person has based on their spending habits using parts of the Food, Income, and Expenditure Survey 2016. Since the response variable, the type of household of a respondent, has 3 possible categories, it is appropriate to use the Baseline Category Logit Model.

The final model chosen by the backward selection was Model 3. The model where extended family, total food expenditure, clothing expenditure, housing expenditure, medical expenditure, transportation expenditure, and communication expenditure were found to be significant. In terms of fit, it was also found that the best model to be used for interpretation was also Model 3 since it was not significantly different from the full model and had the least number of parameters. Response probabilities for each category in the response were also estimated using the formed equations in the model.

REFERENCES

Agresti, A. (2007). *An Introduction to Categorical Data Analysis (2nd edition)*. Wiley.

Fienberg, S. (2004). *The Analysis of Cross-Classified Data (2nd edition).* Springer.

Flores, F. P. (2017) Filipino family income and expenditure. https://www.kaggle.com/datasets/grosvenpaul/family-income-and-expenditure

SAS (2022). *PROC LOGISTIC Statement*. SAS/STAT 9.22 User's Guide. https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect004.htm

SAS CODES

```
data catdata.fiessort;
set catdata.fiesproj;
if 'Type of Building/House'n = 'Single house' then House =1;
if 'Type of Building/House'n = 'Duplex' then House =2;
if 'Type of Building/House'n = 'Multi-unit residential' then
House =3;
if 'Type of Household'n = 'Extended Family' then ExtendedFamily
= 1;
else ExtendedFamily = 0;
if 'Type of Household'n = 'Two or More Nonrelated
Persons/Members' then NonRelated=1;
else NonRelated = 0;
run;

proc sort data = catdata.fiessort;
by House;
run;

proc freq data = catdata.fiessort;
table House;
run;

title "Full Model";
proc logistic data = catdata.fiessort order = data;
model House = ExtendedFamily NonRelated 'Total Food
Expenditure'n 'Clothing, Footwear and Other Wea'n
'Housing and water Expenditure'n 'Medical Care Expenditure'n
'Transportation Expenditure'n
'Communication Expenditure'n 'Education Expenditure'n / ctable
corrb link = glogit
alpha = 0.05 clparm = wald clodds = wald influence ctable;
units 'Total Food Expenditure'n = 2000
'Clothing, Footwear and Other Wea'n = 2000
'Housing and water Expenditure'n = 2000
'Medical Care Expenditure'n = 2000
'Transportation Expenditure'n = 2000
'Communication Expenditure'n = 2000
'Education Expenditure'n = 2000;
output out = results1 p = predicted;
```

```
oddsratio ExtendedFamily;
oddsratio NonRelated;
run;

title "Backward Selection";
proc logistic data = catdata.fiessort order = data;
model House = ExtendedFamily NonRelated 'Total Food
Expenditure'n 'Clothing, Footwear and Other Wea'n
'Housing and water Expenditure'n 'Medical Care Expenditure'n
'Transportation Expenditure'n
'Communication Expenditure'n 'Education Expenditure'n / ctable
corrb link = glogit selection = backward slentry = 0.05 slstay =
0.05
alpha = 0.05 clparm = wald clodds = wald influence;
units 'Total Food Expenditure'n = 2000
'Clothing, Footwear and Other Wea'n = 2000
'Housing and water Expenditure'n = 2000
'Medical Care Expenditure'n = 2000
'Transportation Expenditure'n = 2000
'Communication Expenditure'n = 2000
'Education Expenditure'n = 2000;
output out = results1 p = predicted;
oddsratio ExtendedFamily;
run;
```