Data Analysis Project

Yi Chen

Teachers College, Columbia University

Linear Model & Experimental Design

Data Analysis Project

# Introduction

The purpose of this project is to apply the linear regression and experimental design methodologies to explore the meaningful research questions with a real data set. In this project, we analysis Fiji Fertility Survey data. The primary outcome variable in this study is the number of children ever born to each woman. We will explore three independent factors: duration since their first marriage, type of place of residence, and educational level. We will begin by fitting one-way ANOVA, and two-way ANOVA models. Additionally, several linear contrasts will also be compared. Under the framework of log-linear model, we will also discuss the one-factor, two-factor, and three-factor models. Finally, we will summarize the implication and limitations.

# Data Set

The data set for this study is adapted from Little (1978) from Fiji Fertility Survey. Table 1 shows the data on the number of children ever born to married women of the Indian race classified by duration since their first marriage (noted as ` Marriage Duration `), type of place of residence (noted as `Residence`), and educational level (noted as `Education`). There are 70 grouped data in total. There are six categories of duration since their first marriage: 0-4 years, 5-9 years, 10-14 years, 15-19 years, 20-24 years, and 25-29 years. There are three categories for type of place of residence: Suva, other urban, and rural. There are four categories level for education: none, lower primary, upper primary, and secondary or higher.

All of these independent variables (i.e., duration since their first marriage, type of place of residence, and educational level) are discrete. In Table 1, each cell in the table shows the mean, variance, and the number of observations under difference situation. Consequently, if we times mean with number of observations, we can get the total number of children ever born. There are some missing group data since there is zero observation. We will ignore them since the proportion of missing data to the whole data set is small.

*Table 1*: Number of Children Ever Born to Women of Indian Race by Marital Duration, Type of Place of Residence and Educational Level

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Marriage Duration | Suva | | | | Urban | | | | Rural | | | |
| N | LP | UP | S+ | N | LP | UP | S+ | N | LP | UP | S+ |
| 0-4 | 0.5 | 1.14 | 0.9 | 0.73 | 1.17 | 0.85 | 1.05 | 0.69 | 0.97 | 0.96 | 0.97 | 0.74 |
| 1.14 | 0.73 | 0.67 | 0.48 | 1.06 | 1.59 | 0.73 | 0.54 | 0.88 | 0.81 | 0.8 | 0.59 |
| 8 | 21 | 42 | 51 | 12 | 27 | 39 | 51 | 62 | 102 | 107 | 47 |
| 5-9 | 3.1 | 2.67 | 2.04 | 1.73 | 4.54 | 2.65 | 2.68 | 2.29 | 2.44 | 2.71 | 2.47 | 2.24 |
| 1.66 | 0.99 | 1.87 | 0.68 | 3.44 | 1.51 | 0.97 | 0.81 | 1.93 | 1.36 | 1.3 | 1.19 |
| 10 | 30 | 24 | 22 | 13 | 37 | 44 | 21 | 70 | 117 | 81 | 21 |
| 10-14 | 4.08 | 3.67 | 2.9 | 2 | 4.17 | 3.33 | 3.62 | 3.33 | 4.14 | 4.14 | 3.94 | 3.33 |
| 1.72 | 2.31 | 1.57 | 1.82 | 2.97 | 2.99 | 1.96 | 1.52 | 3.52 | 3.31 | 3.28 | 2.5 |
| 12 | 27 | 20 | 12 | 18 | 43 | 29 | 15 | 88 | 132 | 50 | 9 |
| 15-19 | 4.21 | 4.94 | 3.15 | 2.75 | 4.7 | 5.36 | 4.6 | 3.8 | 5.06 | 5.59 | 4.5 | 2 |
| 2.03 | 1.46 | 0.81 | 0.92 | 7.4 | 2.97 | 3.83 | 0.7 | 4.91 | 3.23 | 3.29 | - |
| 14 | 31 | 13 | 4 | 23 | 42 | 20 | 5 | 114 | 86 | 30 | 1 |
| 20-24 | 5.62 | 5.06 | 3.92 | 2.6 | 5.36 | 5.88 | 5 | 5.33 | 6.46 | 6.34 | 5.74 | 2.5 |
| 4.15 | 4.64 | 4.08 | 4.3 | 7.19 | 4.44 | 4.33 | 0.33 | 8.2 | 5.72 | 5.2 | 0.5 |
| 21 | 18 | 12 | 5 | 22 | 25 | 13 | 3 | 117 | 68 | 23 | 2 |
| 25-29 | 6.6 | 6.74 | 5.38 | 2 | 6.52 | 7.51 | 7.54 | - | 7.48 | 7.81 | 5.8 | - |
| 12.4 | 11.66 | 4.27 | - | 11.45 | 10.53 | 12.6 | - | 11.34 | 7.57 | 7.07 | - |
| 47 | 27 | 8 | 1 | 46 | 45 | 13 | - | 195 | 59 | 10 | - |

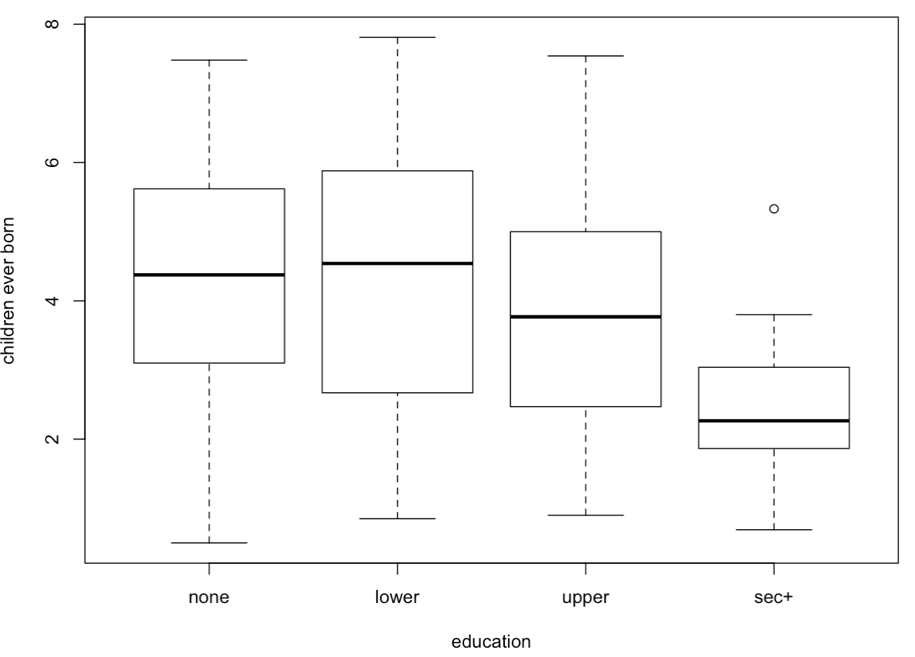
Note: Each cell shows the mean, variance and sample size

# Research Method & Analysis

This study is based on the grouped data. We do not analysis the children born count for each individual woman. Instead, women are categorized by three predictive factors. We focus on the average number of children born for the women under difference categories. One potential benefit of taking group data is: we can reduce the impact from outliers. In particular, we will have less zero count when we apply the generalized linear regression models. This will help to reduce the issue of overdispersion. In this project, we will start with the ANOVA analysis, in which we take the dependent variable (number of children ever born to married women) as a continuous variable. This assumption is not reliable. But it helps to simplify the analysis and still allows us to obtain useful insights. In the second part of this project, we will use the log-linear Poisson regression model, in which the dependent variable will be handled as count data.

* *One-way ANOVA*

We will begin by fitting the one-way ANOVA model using simple linear regression. The representative factor we use in this analysis is `Education`. *Figure 1* shows the relationship between marriage duration and average number of children born. There is a decreasing tendency of number of children born as the education level increase. The hypothesis of this model (H0) is: the main effect of `Education` on the average number of children born is significantly not equal to zero. In other words, `Education` has a significant effect on the number of children born. According to the result of F test, *F (3,66) = 3.87, p = 0.013*. The main effect is significant which reject the hypothesis. The women with education level at secondary or higher reported significantly fewer average children born (M=2.37, SD=1.21) than women have education at none level (M=4.28, SD=2.01), lower primary level (M=4.29, SD=2.19), and upper primary level (M=3.67, SD=1.85). Using dummy coding with the education at none level as reference level, the model result can be expressed as:

This result indicates that, the predicted average children born for the women with education at the none level is 4.28, for women with education at lower primary level is 4.29 (4.28+0.01), for women with education at upper primary level is 3.67 (4.28-0.61), and for the women with education at secondary or higher level is 2.37 (4.28-0.91). This finding is consistent with the finding of the empirical mean of each education level.

*Figure 1*. Relationship between marriage duration and children ever born

* *Two-way ANOVA*

Similarly, we can also add `Marriage Duration` as the secondary factor. The two-way ANOVA result is shown in the *Table 2*. Both `Education` and `Marriage Duration` have significant main effect on the prediction of average children born. The model results of two-way ANOVA can be expressed as:

In general, `Marriage Duration` have bigger and more significant effect than `Education`. According to this result, women with education at lower primary level and marriage duration between 25 to 29 years are predicted to have biggest average number of children born 6.69 (1.44+0.01+5.24). While, women with education at secondary or higher level and marriage duration within 5 years are predicted to have lowest average number of children born -0.17 (1.44-1.61). This prediction does not make real sense since the number of children born cannot be negative. However, it indicates the difference of number of children born among different marriage duration groups. The reason that prediction can be negative is because the two-way ANOVA have the limitation of handling the count data which is always positive.

*Table 2*: F table for Two-way ANOVA (Type 3)

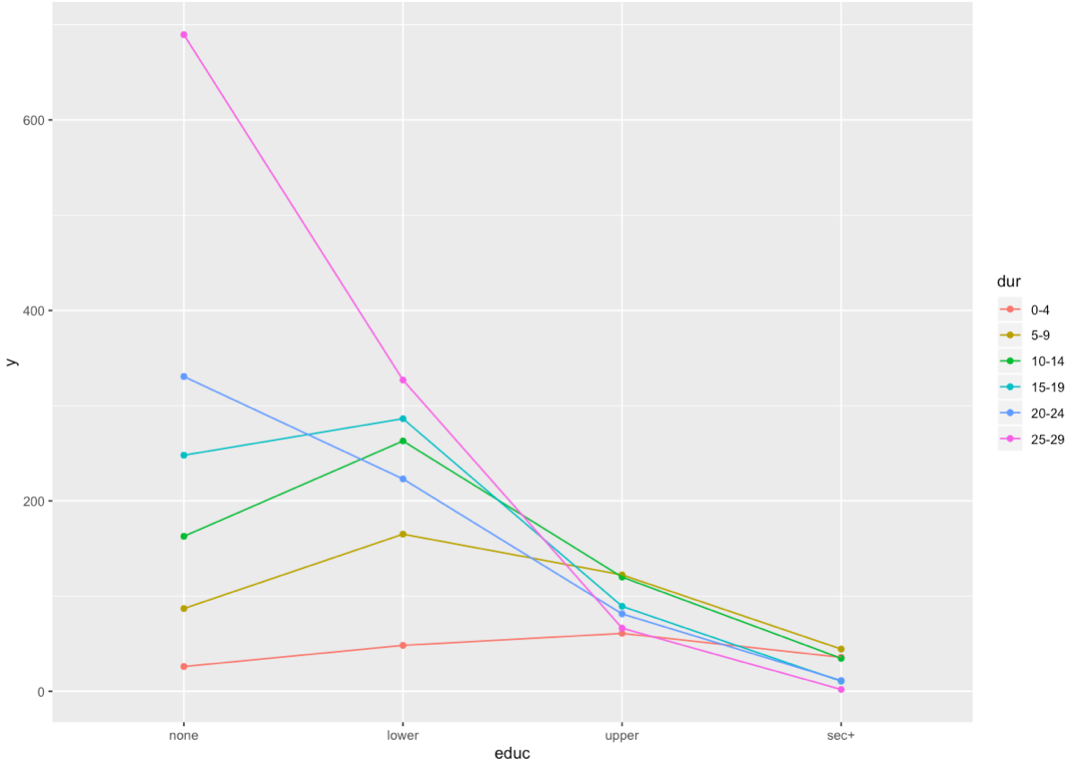
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Sum of Square | Degree of Freedom | F value | P value |
| Intercept | 16.51 | 1 | 24.82 | 0.00 |
| Education Level | 28.62 | 3 | 14.35 | 0.00 |
| Marriage Duration | 189.56 | 5 | 57.01 | 0.00 |
| Residuals | 40.57 | 61 |  |  |

When we take more than one factor into the model, there is always a question about whether there exists the interaction effect between two factors. Figure 2 shows the interaction plot between the `Education` and `Marriage Duration`. based on the evidence from the plot, we can conclude that there do exist the interaction effect. However, it is very common to see the interaction effect if the categorical predictors have more than two levels.

*Table 3* shows the results from F test. Most of variance in the dependent variable is explained by the variance in `Education` and interaction between `Education` and `Marriage Duration`. After adding the interaction effect into the model, the main effect from the `Education` become insignificant.

*Table 3*: F table for Two-way ANOVA with interaction (Type 3)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Sum of Square | Degree of Freedom | F value | P value |
| Intercept | 2.323 | 1 | 5.150 | 0.027 |
| Education | 0.134 | 3 | 0.099 | 0.960 |
| Duration | 64.838 | 5 | 28.747 | 0.000 |
| Education: Duration | 19.817 | 15 | 2.929 | 0.002 |
| Residuals | 20.750 | 46 |  |  |

*Figure 2*. Interaction Plot of Two-way ANOVA

* *Linear contrasts*

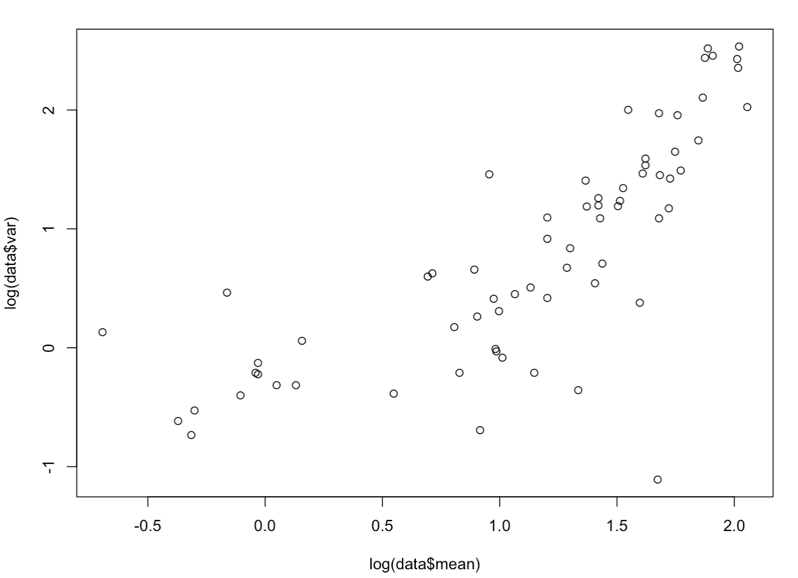
In this section, we will explore two linear contrast situations. Firstly, we want to check the whether the average number of children born for the women coming from Suva is equal to the women coming from other places. The first linear contrast (complex comparison) can be expressed as:

, where represent the average number of children born for women coming from Suva, other urban places, and rural places. According to the t test, *t = 3.04, p = 0.003.* Consequently, we reject the H0 hypothesis and conclude that there exists significant difference in the average number of children born for the women coming from Suva and the women coming from other places.

Similarly, the second linear contrast is applied for pairwise comparison of average number of children born between women with education at lower primary level and upper primary level. This linear contrast can be expressed as:

, where in this case represents the average number of children born for the four education levels. According to the t test, *t = -0.55, p = 0.58.* Consequently, we fill to reject the H0 hypothesis and conclude that there does not exist significant difference in the average number of children born between women with education at lower primary level and upper primary level.

* *Log-linear model*

Before we apply the Poisson regression in this data set. We need to firstly check the mean-variance relation in the data set. *In Figure 4*, we use a log-log scale. Clearly, the assumption of constant variance is not valid. Although the variance is not exactly equal to the mean, it is not far from being proportional to it since the mean of dependent variable is 3.92 and the variance is 3.69.

*Figure 3*. The mean-variance Relationship (log-log scale)

Let indicate the th individual observation in the th marriage duration, th place of residence, and th education level. indicates the total number of children born for the woman in each group. In the previous sections, we model the group average with different predictors. Our data set is grouped data which does not provide the individual data. However, with the number of observations in each group, we can transfer the model in the following way:

, where indicates the vector of covariate in our data set. For the purpose of comparation, we first test the additive two-factor log-linear model. is referred as offset. The estimated parameter is the same for individual data or group data.

However, under the framework of log-linear model, there are many other model design possibilities. We need to figure out which is the most efficient model. For simplicity, we only focus on the additive models in this section. Based on how many unique factors we used in the model, there are one-factor models, two-factor models, and three-factor models. When more than one factors are incorporated, there are model with and without the interaction effect. According to the result from *Table 4*, we can conclude that three-factor additive model is the best model. We can express the model result as:

Taking the estimated parameter value of `Marriage Duration from 15 to 19` as an example, represent for the women who have a marriage duration from 15 to 19 years are expected to have exp(1.54) = 4.66 more children ever born compared with who have a marriage within 5 years, if we ignore all the other covariate.

*Table 4*: Model Comparation for Poisson Log-linear models

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Degree of Freedom | AIC | BIC |
| Null | 69 | 149.88 | 154.37 |
| One-factor Models | | | |
| Duration | 64 | 33.17 | 48.85 |
| Residence | 67 | 152.65 | 161.64 |
| Education | 66 | 148.27 | 159.51 |
| Two-factor models | | | |
| D + R | 62 | 26.61 | 46.84 |
| D + E | 61 | -4.09 | 18.39 |
| R + E | 64 | 151.12 | 166.86 |
| Three-factor models | | | |
| D + R + E | 59 | -17.71 | 9.27 |

Note: D is `Marriage Duration`, R is ` Residence `, and E is `Education`.

# Conclusion

* Implications

In this study, we explore the conventional ANOVA analysis in the Fiji Fertility Survey data set. ANOVA analyses are based on the average number of children born for each group. Two different linear contrasts are discussed. Additionally, we also applied the log-linear models which handle the group data directly with offset. We also compare seven different additive models using AIC and BIC to find the best model design. In general, the finding of two parts of analyses are consistent with each other. And, the model result is reliable.

* Limitations

In this study, we do not incorporate the continuous variable. Consequently, we cannot discuss the ANCOVA. Additionally, we do not discuss the interactive models in the log-linear framework. This may be risky since there exists certain level of interaction effect according to the interaction plot. Other model comparison methods are not discussed in this study. For example, loglikelihood ratio test for the nest models.

References

Little, R. J. A. (1978). Generalized Linear Models for Cross-Classified Data from the WFS. *World Fertility Survey Technical Bulletins*, Number 5