Lifecycle Analysis of Bicycle Effect

Jorge Loría *
Max Woodbury *
Gordon Yang *

1 Introduction

The Life Cycle program is a local nonprofit organization that donates bicycles to anyone 16 years or older in need. We are testing the effectiveness of this program by comparing the number of charity visits for participants versus non-participants. If we can show a significant difference in charity usage between these two groups, then we will consider the program effective. This analysis will be done by using a Poisson model with charity usage as the response and descriptive predictors such as gender, ethnicity, employment, and marital status. We will also analyze the significance of the association between these descriptive variables and charity usage. This will provide suggestions on what types of people use charities more often and seem to need the most help. Note that some of these results from our sample may not fully portray the population because of the high percentage of missing values for some variables, such as veteran and disabled status.

2 Data description

Using the data provided by Life Cycle, which corresponds to 4 different years: 2015, 2016, 2017, and 2018. We use the following variables to control for demographics and other charities: gender, ethnicity, education, employment, marital status, receives medicaid, receives social security, receives veterans benefits, veteran status, receives WIC (Women, Infants, and Children), receives medicare, at risk of being homeless, disabled, homeless, and poverty level. Finally, the variable we care about is if the person is in the Life Cycle program. It should be noted, that all these variables were self reported, and often the participants chose not to respond. In those cases, we first impute using the most recent observation for that participant in that variable where the participant answered. In the case the participant hasn't previously reported that variable we assign it "*missing*".

*StatCom, Department of Statistics Purdue University

A brief summary of the variables is presented next, for each participant, by looking at their last observation. We have a total of 2987 cases. Of those: 1141 report to be female, 493 male, and 1353 decided not to respond. Regarding ethnicity, 38 participants report to be African-American, 491 Caucasian, 6 Hispanic, and 2452 didn't respond this question. On education, we have 22 on the college level, 23 with incomplete high-school, 46 on High-School/GED, and 2896 didn't report their education level. For employment, 262 reported to be full-time, 260 on part-time, 850 unemployed, and 1615 didn't report this variable. For marital status, 47 reported to be divorced, 362 married, 613 reported to be single, and 1965 chose not to report their marital status. The poverty level that is reported takes the following values: below 150% of the poverty level, with 870 cases; below 200% and above 150%, 77 participants; below the poverty level, 183 participants; and finally 1877 participants marked as missing. The rest of the variables, shown in Table 1 correspond to binary variables, which means they take one value or the other. On the original data, we only had a few values that were marked as "Yes". Upon further inquiry we were informed that these came from a checkbox option, so the respondents would either mark or not. We assume that if they didn't mark that checkbox, that they don't belong to that category.

Variable	No	Yes
At risk of being homeless	2686	301
Disabled	2973	14
Homeless	2935	52
Receives food stamps	2276	711
Receives medicaid	2939	48
Receives medicare	2970	17
Receives social security	2376	611
Receives veterans benefits	2979	8
Receives wic	2976	11
Veteran	2974	13

Table 1: Variable counts for each of the participants' last observation

We have a record of each visit during that time lapse. We consider how many times each participant visits any charities (within Charity Tracker) every month, and summarize the observations for each participant and every month. When the participant stops attending the charities, we stop considering it, unless it eventually comes back (according to the current data), in which case we say that the participant visited 0 times until they come back again. This number of visits to charities is going to be our response variable. We also include the month and the case number for each of the participants recorded. It should be noted, that we use the case number as an identification, and don't have access to any other information describing the person.

3 Model description

The type of model we use is a generalized linear mixed effects model (see: Faraway, ch. 3 and ch. 10[1]), using a Poisson regression and accounting for the effects of each person and the time effect. The output we consider is the number of times each person has a recorded visit in the charity tracker. We are interested in the effect of receiving a bike from the Life Cycle project on the previously mentioned number of visits per month.

Note that the value we care about can only take non-negative values, since we are trying to work with the number of visits per month each person makes. Which is why we use a Poisson regression. The Poisson regression relies on the *Poisson* probability distribution, which has probability mass given by:

$$P(Y = k) = \frac{e^{-\mu}\mu^k}{k!};\ k = 0, 1, 2, \ldots \tag{1}$$

We denote this by $Y \sim Poi(\mu)$, and a well known property is that the mean value of Y is μ, this and other properties can be consulted in any introductory probability books, for example: Castañeda [2]. Given this, we model the mean of a person with characteristics given by a vector $X = (x_1, x_2, \ldots, x_p)$, by:

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{2}$$

Where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, are the regression coefficients which we want to find. Note that since we are modelling the logarithm of the mean, we have to be careful when interpreting the values of the coefficients β, as they won't have a linear effect on the number of visits, but instead will have a multiplicative effect.

Now, we must also include the information of *when* and *who* the person is. For this, we use *random effects*, which contemplate the effect of each person and each month in the observed data. Since we have 48 months, let's denote the time index using $t = 1, \ldots, 48$ for the 48 possible months, and for each of the 2987 participants observed in the data, we have a person index denoted by $j = 1, \ldots, 2987$. Then we model the mean $\mu_{j,t}$ of the $j-$th participant on the $t-$th month by:

$$\log(\mu_{j,t}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon_{j,t} \tag{3}$$

$$\varepsilon_{j,t} = \eta_j + \alpha_t \tag{4}$$

$$\alpha_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_a^2) \tag{5}$$

$$\eta_j \overset{iid}{\sim} \mathcal{N}(0, \sigma_c^2) \tag{6}$$

This is a way of specifying that the $t-$th month affects everyone in a similar way, and each person will behave consistently throughout time, but that we

don't think the effect of time and person will be too large. Also, note that this doesn't change the interpretation mentioned above for the multiplicative effect on the mean.

4 Model results and interpretation

All variables included in the model are listed in the data description section. That is, covariates were included in the model even if they were not found to be significant, to control for their effects. The model selection part was skipped since we care to interpret the average effect of having or not having a bike. Again, it can't be stressed enough that for some variables most of the data was missing, to reserve some doubt that the model and data reflects reality. If it were available, the data may lead to different results. For the missing data the only assumption we make is treating it as a group that all behave in the same way. The missing group masks the significance of other predictors. When we say that a predictor is significant, this means that we are 95% sure that it's related to the outcome.

4.1 Main result

On average, keeping other factors fixed, Life Cycle participants make 3.660 times more visits per month than non-participants. We are 95% confident that Life Cycle participants make 3.076 to 4.355 more visits to charity associations per month on average. Possible reasons for the effect may be:

1. Bicycle gives ability to travel more/further.
2. Bicycles are given to those who needed it the most.

The same model from the main results was used for the fixed effects and random effects. By fixed effects we mean it is calculated from the data and no population is assumed. The levels of a variable are how many values it can take on.

4.2 Fixed effects

Levels = 2

The effects that are significant are: receives social security, at risk of being homeless, disabled, homeless, in lc.

Interpretation: The numbers in the estimate column of Table 2 represent how many times more likely a person in that category (e.g. receives social security, is disabled, is homeless) is to make a monthly visit than someone who is not in that category. For example, on average, keeping other factors fixed, people who receive social security make 1.569 times more visits per month than those who don't. We are 95% confident that people who receive social security make 1.425 to 1.728 more visits to charity associations per month on average.

Variable	Estimate	95% CI
Receives Social Security	1.569	(1.425, 1.728)
At Risk of Being Homeless	1.533	(1.357, 1.732)
Disabled	1.903	(1.218, 2.973)
Homeless	1.424	(1.092, 1.857)
Participates in Life Cycle	3.660	(3.076, 4.355)

Table 2: Coefficients of fixed effects with 2 levels

For variables with more than two levels we only discuss variables that were found to be significant as a group. If the group was found to be significant, we consider all the levels to be significant regardless of whether they're significant individually.

Levels $>$ 2

Significant effects: gender, ethnicity, poverty level, employment.

Variable	Value	Estimate	CI
Gender	Male	0.891	(0.807, 0.985)
Gender	Missing	0.831	(0.74, 0.933)
Ethnicity	African-American	1.169	(0.866, 1.579)
Ethnicity	Hispanic	1.437	(0.695, 2.971)
Ethnicity	Missing	1.575	(1.407, 1.762)
Employment	Missing	0.585	(0.505, 0.679)
Employment	Part time	0.867	(0.739, 1.018)
Employment	Unemployed	0.826	(0.719, 0.949)
Poverty Level	Below 200%, but Above 150%	1.213	(1.029, 1.428)
Poverty Level	Below Poverty Level	1.081	(0.948, 1.231)
Poverty Level	Missing	0.688	(0.636, 0.744)

Table 3: Factors with more than two levels, with their confidence intervals and estimated effects

Where the baselines are the absent category for each factor. Namely:

- Gender: Female,
- Ethnicity: Caucasian,
- Employment: Employed,
- Poverty Level: Below 150% of the Poverty Level.

The interpretation of Table 3 is as follows, using the previously itemized baseline variables:

1. Gender: on average, keeping all other factors fixed, males make 0.891 times less visits to charities per month compared to females.

2. Poverty level: "...", those below 200% but above 150% poverty level make 1.213 times more visits to charities per month compared to those below 150% poverty level.

3. Employment: "...", the unemployed make 0.826 times less visits to charities per month than employed.

4. Ethnicity was only significant due to missing data.

Note that these values are interpreted as a general effect on the estimates, and these are **not** for how these populations behave within the Life Cycle program, but rather how they affect the Charity Tracker population in general.

4.3 Random effects

We verified if it's required to include the person and time effects, and our tests indicate they are a good addition (person: $\chi^2(1) = 6302.1$, $p < 2.2e-16$; time: $\chi^2(1) = 1817$, $p < 2.2e-16$). A usual tool when including random effects is to determine which of the random effects is explaining most of the variability, specifically we want to see what percentage of the variability in the observations corresponds to the time-effect over the total variation: time-effect *plus* person-effect. This percentage is 22.92%, and the person-effect explains 77.08%.

5 Conclusion

We found that being a participant in the Life Cycle program has a strong positive association on charity usage. That is, people that received a bicycle from the program are likely to use charities more often compared to those that did not. One possible reason for this relationship is that participants were able to travel to more charities using their donated bicycles compared to other people who attend these charities. Another possible reason is that the types of people who participate in this program need more help than the average non-participants. We cannot determine the true reason for this effect, but in any case we can see that this program has a significant positive impact on its participants.

In this analysis we focused on comparing the charity usage of Life Cycle participants versus non-participants. For next steps, we recommend obtaining more data from other charity providers to further improve the accuracy of the results. To answer the question of "*Are the people that need it the most getting a bike*?", a survey would probably be an appropriate tool; this would help identify how many people don't have access to bikes and don't know about the Life Cycle program. Additionally, it would help to reduce the percentage of missing observations in order to have stronger sample results that can more closely approximate the population. One way to do this would be to require certain key information from the participants before donating a bicycle to them.

The predictors that had the most significant impact on charity usage (and the ones we recommend requiring first) are ethnicity, employment, poverty level, receives social security, at risk of being homeless, disabled, and homeless.

We understand the main goal of the program is to give people bicycles, not to collect information. However, we suggest collecting more information from the participants, to improve the quality of the program. One way to do this is to offer an extra incentive for completely filling the form, like a coupon or offering a meal.

References

[1] Faraway, J. (2006) "Extending the linear model with R"

[2] Castañeda, L., Arunachalam, V., et al. (2012) - "Introduction to Probability and Stochastic Processes with Applications".