## Husbands and Happiness in Rural Ethiopia

STA-310 Final Project – Aleph Aseffa, Haruko Okada, Huandong Chang

## Introduction

This project aimed to visualize and analyze the factors that influence the happiness of women in rural villages in Amhara, Ethiopia. The data we used was from Professor McGavock’s research paper: “Her Time: A Time Use Study of Women Participating in Livelihoods Programs in Ethiopia”. One of the main topics of her research was to measure the recall bias of different types of surveys: phone calls and in-person interviews. However, our research focused on the reported “happiness level” of each woman. In particular, we wanted to identify if the happiness was impacted by the presence of their husband.

Happiness was evaluated on a five-point scale. However, 94.6% of the responses were either “very happy” or “fairly happy”. In this initial analysis, we chose to restrict our data to just the “very happy” or “fairly happy” responses and use a logistic regression to predict the probability of someone being “very happy. Our analysis shows that the presence of a husband had a significant effect the probability of a woman being “very happy” in the CC and PT surveys but not in the PD survey.

In this report, we detail the data cleaning, model construction, and provide short summary of our findings. The appendix provides instructions on a data visualization tool we created which is available at: <https://huandongchang.shinyapps.io/WomenHapiness>.

## Data

Appendix Part B contains the R code we used for the data cleaning. After this cleaning, the variables are now consistent across the treatments, and we only include variables that are meaningful to our analysis. In order to match the time slot between PD, CC(30 minutes slot survey) and PT(phone call asking what the respondent did in the last three hours), we created a new column called “Period” which are three hour slots and corresponds to all survey types. For the companion coding, we manipulated the data to figure out who they were with, and created a new column called “WithHusband” which is a 0,1-indicator variable. Our analysis focused on the following variables within the data:

* **Vesa**: Code that identifies the VESA (Village Economic and Social Associations; local community groups) that the respondent belongs to.
* **CompanionDuringActivity**: Was there anyone with the respondent (1 or 0)
* **PrimaryActivityMainCategory**: What category their primary activity during this time period falls into.
* **SecondaryActivityMainCategory**: What category their secondary activity during this time period falls into.
* **MainLocationDuringTimePeriod**: Where the respondent was for the majority of this time period.
* **CaringForChild**: If the respondent is caring for a child (1) or not (0).
* **WithHusband**: If the respondent is with her husband (1) or not (0).

More information on these variables can be found in the codebook (Appendix Part C).

### Methodology

This dataset consisted of three distinct survey types:

* **CC (Control):** Respondents are asked questions in in-person interviews.
* **PT (Phone Time Use Treatment):** Respondents receive 5 phone calls per day, and are asked about their primary and secondary time use activities
* **PD (Phone Diets Treatment):** Respondents are asked questions in in-person interviews.

In our dataset, each row of data represented an observation. The CC and PT surveys contained data reported every 30 minutes; while the PD survey had data based on three-hour time frames. To account for the bias that could be introduced by the varying time frames, we constructed one model for each household treatment status (i.e. CC, PD, PT), and one for the entire dataset (ALL), where PT is based on phone calls and CC and PD are based on recalls.

In each of these four models, we analyzed whether the presence of the husband affected self-reported happiness levels. For each model, we first create a reduced model including all main effect variables and their interaction terms, except for the WithHusband variable. We then conducted a Likelihood Ratio tests (LRT) on the WithHusband variable, allowing us to determine whether or not it had a significant impact on Happiness, *after accounting for all other variables in the model*. It is important to recognize that since all other variables were tested multiple times, their p-values can be very unreliable.

Create a reduced model:

* Variables that were linear combinations of (i.e. perfectly correlated with) any other variable were removed.
* A model with all possible main effects and two-way interactions was created.
* We used the Wald statistics, Likelihood Ratio tests (LRT), and each variables impact on happiness to determine which variables could be removed without hindering the accuracy of the model. In the LRT, the G-statistic and degrees of freedom corresponding to the smaller model are calculated and compared to those of the larger model. If the p-value associated with this G-statistic is greater than 0.05, we know that the removed variables are unimportant to the model and can be safely removed. This is a safety check since the p-values we used to select which variables to remove can be unreliable.

Compare the Reduced Model to the Full Model

* After the reduced model was created, we created a “Full Model” by adding the WithHusband variable to the model.
* We used the LRT to determine whether WithHusband is significant to the model. If the p-value we obtain is less than 0.05, we conclude that WithHusband should be included, and if the p-value is greater than or equal to 0.05, we conclude that WithHusband should be omitted from the model.

These steps are repeated for each model (CC, PD, PT, and ALL). The code used for each of these steps is presented in the appendix (Part D). Our evaluation of the effects of each individual variable is presented in Part E of the Appendix. Although not used here, it might be a helpful approach when dealing with even more variables than we have here.

## Findings

The following table lists the variable contained in each model. While we indicate significance for all terms, it is important to recognize that all variables, except WithHusband, are somewhat unreliable due to multiple comparisons and potential multicollinearity.

Legend:

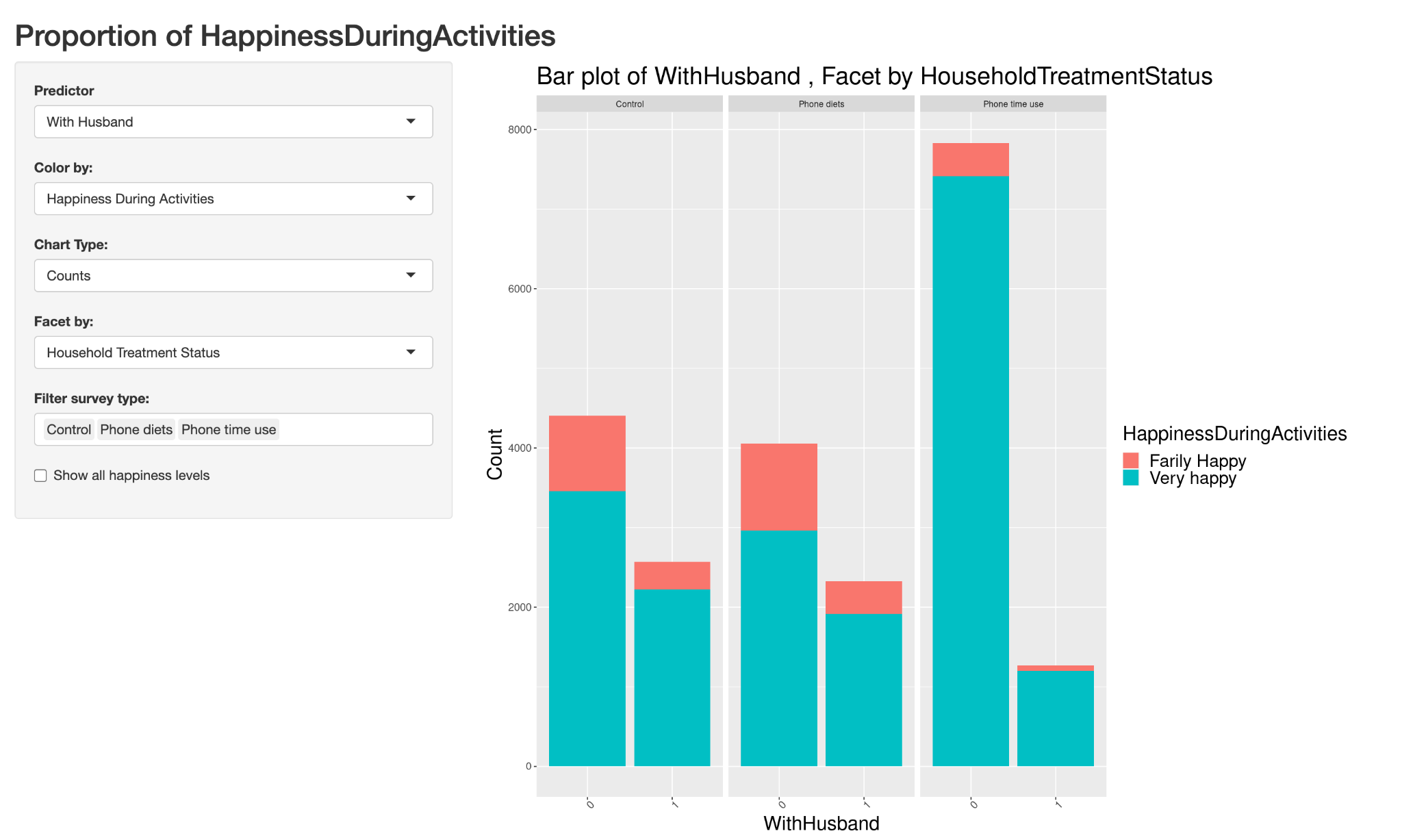
* Blank = Not included in the model
* x = Included, but p-value is greater than 0.05
* xx = p-value <= 0.05
* xxx = p-value <= 0.01

The p-value reported is the minimum of each group.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Model | | | |
| CC | PD | PT | ALL |
| Vesa | xxx | xx | x | x |
| CompanionDuringActivity | x | x | x | x |
| SecondaryActivityMainCategory | x |  | xxx | xxx |
| MainLocationDuringTimePeriod | xxx | xx | x | x |
| CaringForChild | xxx | x | x | x |
| PrimaryActivityMainCategory | x | x | x | x |
| Vesa\*CompanionDuringActivity | x | x | x | x |
| Vesa\*SecondaryActivityMainCategory | xxx | xx | xx | xxx |
| Vesa\*MainLocationDuringTimePeriod | xxx | xx | x | xxx |
| Vesa\*CaringForChild | xxx | x | xx | x |
| Vesa\*PrimaryActivityMainCategory | xxx | xx |  | xxx |
| CompanionDuringActivity\*PrimaryActivityMainCategory |  | x | x | xx |
| CompanionDuringActivity\*SecondaryActivityMainCategory |  |  |  |  |
| CompanionDuringActivity\*MainLocationDuringTimePeriod | x | xx | x | x |
| CompanionDuringActivity\*CaringForChild | x | x | x | x |
| PrimaryActivityMainCategory\*SecondaryActivityMainCategory |  |  |  |  |
| PrimaryActivityMainCategory\*MainLocationDuringTimePeriod |  |  |  |  |
| PrimaryActivityMainCategory\*CaringForChild |  |  |  |  |
| SecondaryActivityMainCategory\*MainLocationDuringTimePeriod | xxx | x |  |  |
| SecondaryActivityMainCategory\*CaringForChild | xx |  |  | xx |
| SecondaryActivityMainCategory\*PrimaryActivityMainCategory |  | x |  |  |
| CaringForChid\*PrimaryActivityMainCategory |  | x | x | xxx |
| MainLocationDuringTimePeriod\*PrimaryActivityMainCategory |  |  |  | x |
| MainLocationDuringTimePeriod\*CaringForChild |  |  |  |  |
| **WithHusband** | **xx** | **xxx** |  | **xxx** |

To explore how the WithHusband differs among the different household treatment statuses, we generated the following graphs. These visualizations can be easily created and modified with the shiny app.





These two graphs have the same variable settings, but one is based on counts and the other percentages. We can see that there are less respondents who report being with their husbands in the PT group, as compared to the other household treatment statuses. The PT model also has a consistent happiness level that does not change based on the husband’s presence when we examine the first graph. Perhaps this could be why the WithHusband variable was found to be insignificant for the PT model.

### Conclusion

We found that the WithHusband variable was significant for the CC (p-value < 0.0001), PT(p-value < 0.0001), and ALL (p-value < 0.0001), models, while it was insignificant for the PT (p-value = 0.65) model. While our analysis stops at uncovering the differences between these household treatment statuses, we believe that this result provides further evidence that the survey method influences the answers of respondents. We hope that our work here provides valuable insights into the collected data and simplifies further explorations.

Suggestions for future work:

* Use this app, (<https://huandongchang.shinyapps.io/WomenHapiness>) to find interesting patterns in the data.
* **Analyze all data based upon the three hour time frame variable (Period).** We could create a happiness score for each time period. For example, very happy could be rated as a 5, somewhat happy could be a 4, etc… Then we could calculate the average happiness score for every 3-hour time. We could then do a similar analysis using multiple regression instead of logistic regression, where the response would be average happiness score. Since the PD surveys would have little variation, we would not expect the residuals to be normally distributed.
* Create a model that looks at the interaction effect of WithHusband verses the other variables.
* Combine the primary and secondary activities to understand the impact of each activity.
* Control for the ID for each person within a model and/or control for the time within each model. Do some people just tend to always give a “very happy” response?
* Since the WithHusband variable is only insignificant for PT, different experimental methods might influence women’s responses. We can do further research to investigate the response bias associated with different experimental methods.

### Appendix

## Part A: App Instructions

<https://www.dropbox.com/s/r5mbzjl99gg2ft7/App%20Instructions.pdf?dl=0>

## Part B: Data Cleaning Code

<https://www.dropbox.com/s/uaphx9mp98n664k/Data_Cleaning.Rmd?dl=0>

## Part C: Codebook

<https://www.dropbox.com/scl/fi/scsxjahruu7tmsyup60di/Codebook-Final.xlsx?dl=0&rlkey=deg0n3ghehw2ymzob8cyv2hlo>

## Part D: Model Code

<https://www.dropbox.com/s/ek6rn0dcwbpeu6r/Model.Rmd?dl=0>

## Part E: Variable Selection

Code: <https://www.dropbox.com/s/p3ug39rytzcuorj/VariableSelection.Rmd?dl=0>

Explanation and results: <https://www.dropbox.com/scl/fi/1hiuq9pl8qh6h1l2ivtjf/Variable_Selection.docx?dl=0&rlkey=3dst18g5c58goflk1tgndy9ji>