Statistical report on MINGAR

Investigation of market behavior and product issues

Report prepared for MINGAR by SPQS,LLC

2022-04-07

Contents

Executive summary

Existing literature illustrates that the main competition for the MINGAR fitness tracking wearable devices is Bitfit. Compared with MINGAR fitness tracking wearable devices, Bitfit devices are smaller, cheaper, and have more insights. In addition, the main customers for Bitfit devices are average consumers who are interested in living a healthy lifestyle. In order to grow with the market and gain market share, MINGAR introduces the "Active" and "Advance" lines. Comparing with original products (i.e. "Run", "Rush", etc.), new products are $100 to $200 cheaper. This report mainly focuses on comparing the differences between the new customers (who choose to buy "Active" or "Advance") and the traditional customers. In addition, this report also discussed whether MINGAR devices perform poorly for users with dark skin.

The results of the study are summarized below.

- The average age of new customers is 47.95 and it is slightly higher than the average age of traditional customers which is 46.51.
- Model illustrates when the income increases from lowest to highest, the odds (odds is the ratio of the probability of an event occurs to the probability that it does not occur) of being a new customer will decrease 92.7%.
- The odds of males becoming new customers are 3.8% higher than females when their average age and income are the same. But this result is not statistically significant.
- The average income of new customer is 68813.94 which is less than the average income of traditional customer which is 73168.02.
- The average regional population of new customers is 1519844 and it is higher than the mean regional population of traditional customers which is 1478529.
- when a customer's age increases from 18 to 92, the odds of being a new customer increased 45%.
- Number of times there was a quality flag during the sleep session for dark skin customers is 11 times larger than the light skin customers.

Additional Visualization of industry information and numerical summary are shown below:

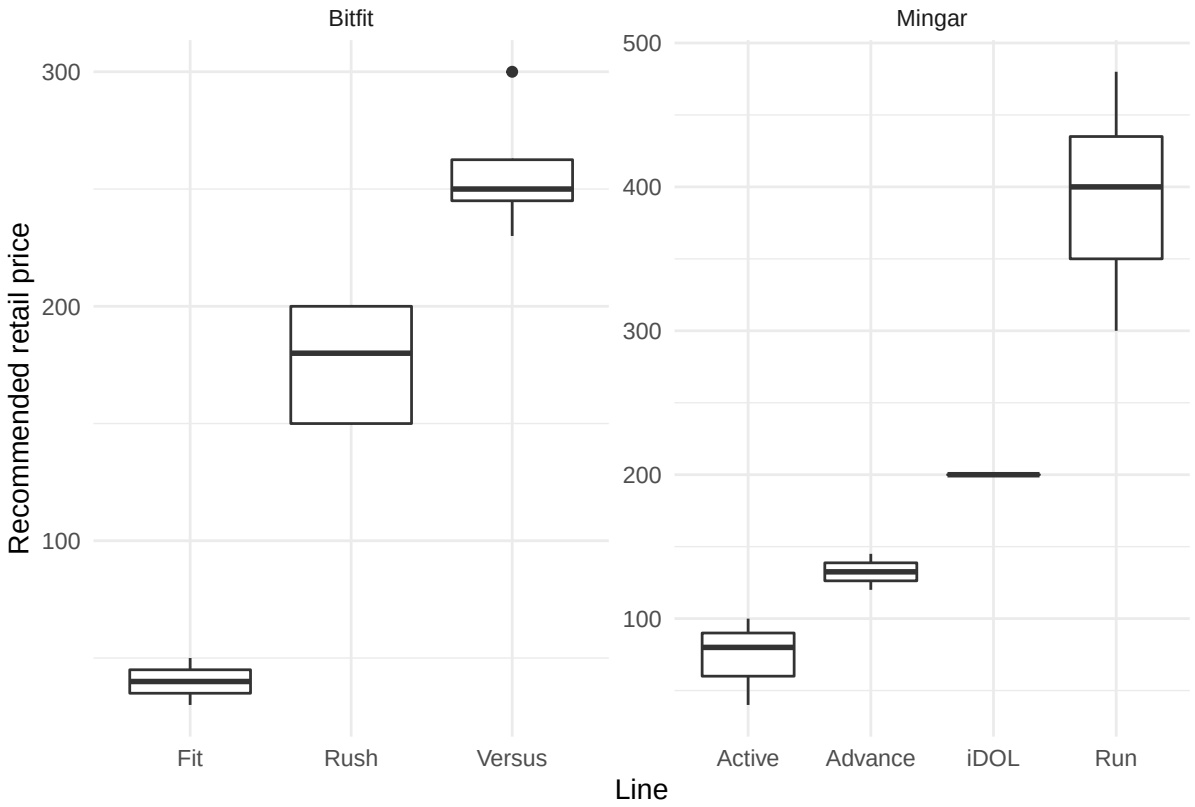

Figure 1: The retail prices for different lines of MINGAR and BITFIT

Table 1: Companring features of new and triditional customers

Customer type	Mean of regional median income	Mean of age	Mean of regional population
Triditional	73168.02	46.50614	1478529
New	68813.94	47.95307	1519844

Technical report

Introduction

According to an article by Ferreira, Rammal, and Veiga (2021), we can wearable technologies are transformed from fashioned and expensive products into daily products. This is the necessary key step for the development of wearable technologies. Thus, we can reasonably guess the customers are changing and the number of clients is increasing. For example, a larger percentage of the population applies products as part of daily life. younger people or elder people start to try the product. People with a lower average income start to use the product. This article provides thinking to help us investigate the following question:

- *How is the market changing? What kinds of people become the client of new products including "active" and "advence" types? How are they different from traditional customers?*
- *In the sleep score of this product, it trends to perform poorly like more flags(sleep quality reduction) data from darker skin. Thus, to avoid the company to to be called "racism". We are going to investigate on it.*

From the original data MINGAR provided. We apply further data manipulation. We use find median income data from Canada's census data to match with the original customer data. Because the skin color is not provided in the original dataset. We generate the potential skin color of customers through the emoji skin color. All data is pre-prepared before all the analyze. Also, all the missing values are removed from the orginal data. The detailed steps is provided in the appendix.

For each question in this report, we firstly did an exploratory data analysis(EDA) to investigate the data set with graphics and visualizations. For the first question, we use mixed logistic regression. We investigate variables including sex, age, skin, income, and population. Finally, we decide to keep age and income as the variable. For the second question, since the count of flags is a count data, we use mixed Poisson regression to investigate the impact of skin color on the count of flags after comparing this model with a larger model which contains the factor of income. The report follows the step:

- *Indicate problem*
- *Exploratory data analyse*
- *Medel*
- *Results*

Research questions

Marketing problem

- *Features of new customers*
- *Differences between new customers and traditional customers*

Exploratory data analyse (EDA)

Important variables Table 2: variable description

Variables	Description
Customertype	Type of the customer (new or traditional)
age	Age of the customer
sex	Sex of the customer
skin	Potential skin color of the customer
hhld_median_inc	median income of household's region
Population	population of household's region

From our data, these variables are likely to explain our research question. *Customertype* is generated from the prepared dataset. We categorize whether the customer is new or traditional by indicating the product line of which the customer purchased. The remaining variables are potential variables that can be used to explain the possibility of being a new customer. We decide whether to add one variable into our final model from the following analyze.

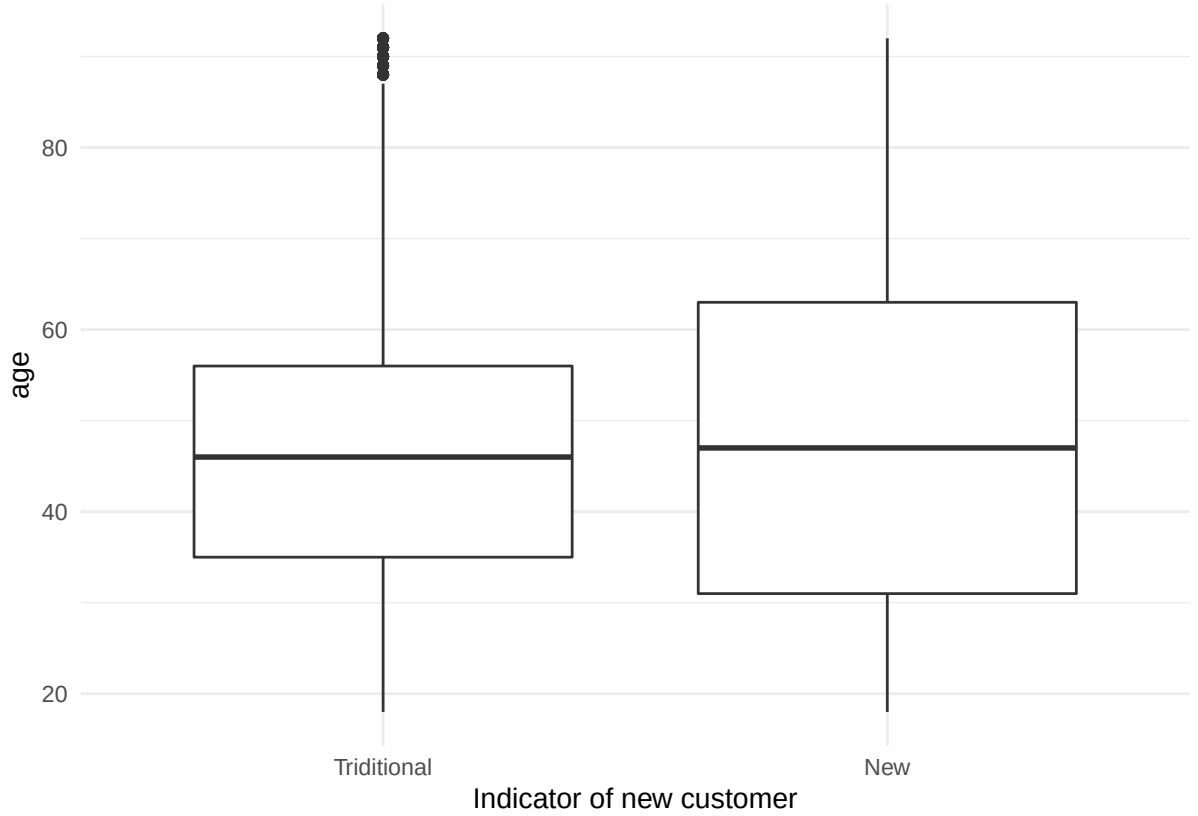

Figure 2: The variance of age is larger for new customer. Also, mean age for new customers is a bit higher than triditional customers. Therefore, we might see association between customer type and age

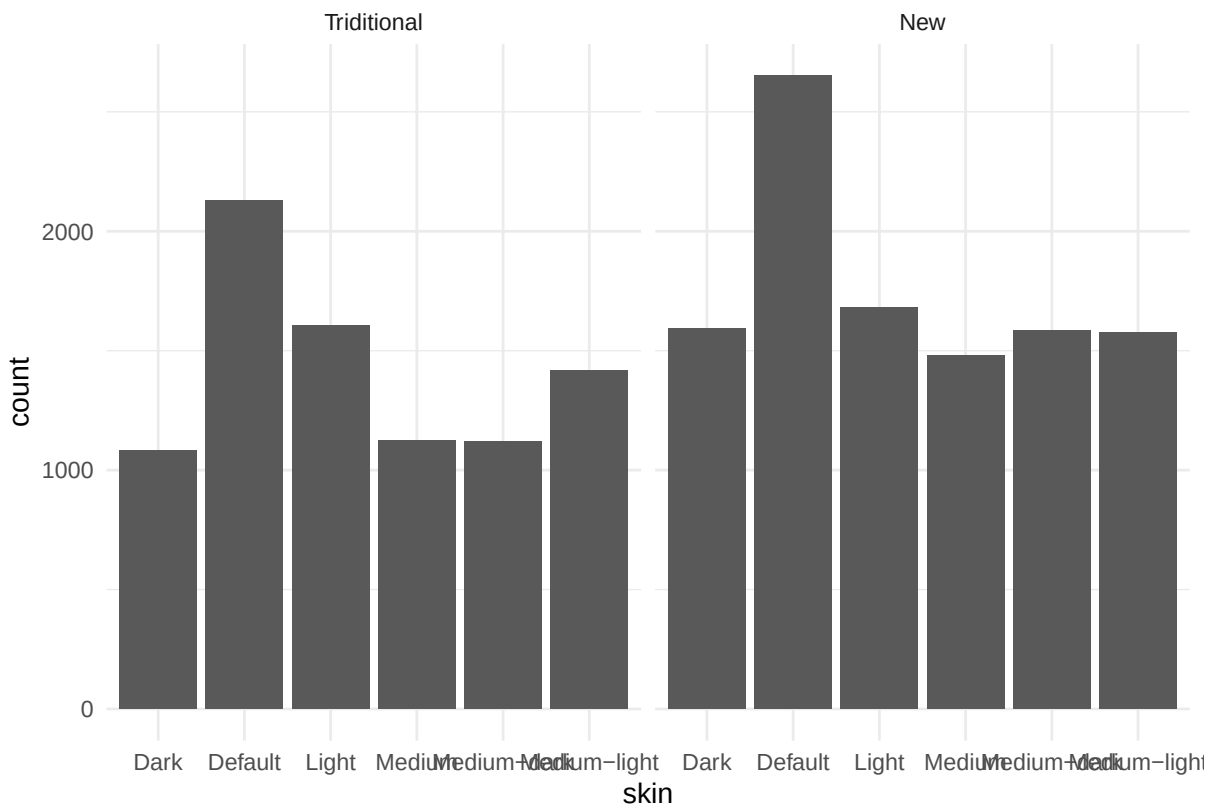

Figure 3: The variation in the count of two types of customers is very similar. Therefore, we believe the skin color is not helpful to explain the variance in customer typer. In other words, we will not include skin color in the model

Grapical summary

Numarical summary

- *number of unique* hhld_median_income *is 256*
- *number of unique* Population *is 255*

By observing the number of unique *hhld_median_income* and *Population*, the count of the group is similar for these two variables. Also, they both indicate the features of regions or neighborhoods of observations. Therefore, we may see dependency in observations with the same regional median income or population.

Table 3: Numarical summary of features of new customers

Customer type	Mean of regional median income	Mean of age	Mean of regional population
Triditional	73168.02	46.50614	1478529
New	68813.94	47.95307	1519844

According to the table, we have idea about who new customers are. The mean age of new customers is higher. They live in regions with lower median income and higher population.

Model

We worry about the violation of the independence of observations in the same region. By adding a random effect, we ensure that observations in the same group share the same random effect that indicates their differences from other groups. Further, we also assume that the impacts of age and sex are the same in different regions. Our model only contains random intercepts. In this model, our groups are dependent on *population* which means we tread the observations with the same population as one group. Since our variable of interest is a binary variable, a generalized linear model model is more appropriate. In this model, we use a link function $log(\frac{\pi}{1-\pi})$ to build a linear model with explanatory variables. Overall, our final model is a linear mixed model with a non-normal response which is a generalized linear mixed model:

$$log(\frac{\pi}{1-\pi})_{ij} = \beta_0 + \beta_1 Age_i + \beta_2 Intersex_i + \beta_3 Male_i + \beta_4 hhld\_median\_inc_i + b_j population$$

where

- $log(\frac{\pi}{1-\pi})_{ij}$ *is the log odd of* i^{th} *observation with* j^{th} *population group*
- β_0 *is observation i's log odd of being a new customer is a female with minimum age and reginal median income*
- $\beta_1, \beta_2, \beta_3, \beta_4$ *are the coefficient for explanatory variables (fixed effects)*
- $b_j \sim N(0, \sigma_b^2)$ *is the random intercept of* j^{th} *population group*

Note: In order to avoid issues of scaling, we rescale our explanatory variables by dividing each variable by the differences between maximum and minimum.

Results: Interpretation of Parameter Estimates

Table 4: Estimates of parameters

Parameters	Estimate	P-value
β_0	0.64295	3.11e-12
β_1	0.37404	8.89e-09
β_2	0.18568	0.186
β_3	0.03743	0.216
β_4	-2.61372	1.94e-12

Regarding the output of our estimation, we obtain estimates of our parameters among 255 population groups:

- $\hat{\beta_0} = 0.64295 =$ *the mean log odds of being a new customer who is female, with age of 18, and regional median income of 41880. With p-value of* 3.11×10^{-12}*, the estimate is statistically significant.*
- $\hat{\beta_1} = 0.37404 =$ *when a customer's age increases from 18 to 92, the log odds of being a new customer increase by 0.37404 with fixed sex and regional median income. With p-value of* 8.89×10^{-9}*, the estimate is statistically significant.*
- $\hat{\beta_2} = 0.18568 =$ *the mean log odds of an Intersex being a new customer is 0.18568 higher than Female. With p-value of 0.186, the estimate is not statistically significant.*
- $\hat{\beta_3} = 0.03743 =$ *similar interpretation as* $\hat{\beta_2}$
- $\hat{\beta_4} = -2.61372 =$ *when a customer's regional median income increases from 41880 to 195570, the log odds of being a new customer decrease by 2.61372 with fixed age and sex. With p-value of* 1.94×10^{-12}*, the estimate is statistically significant.*

Social media team problem

- *whether the devices are performing poorly for users with darker skin*

EDA

Important variables Table 5: Variable description

Variables	Description
flags	number of times there was a quality flag during the sleep session
duration	duration, in minutes, of sleep session

The additional variables is generated from merging customer sleep data with our previous customer data. *flags* is our variable of interests. It can represents the sleep scores of customers. In other words, more *flags* detected means lower sleep score of the customer. *Duration* is vary for different customs. For customer who sleep more, the *flags* detected tend to be more than those who sleep less. Therefore, we set *duration* as a offset term in the poisson regression. The description of other variables is the same as table 1.

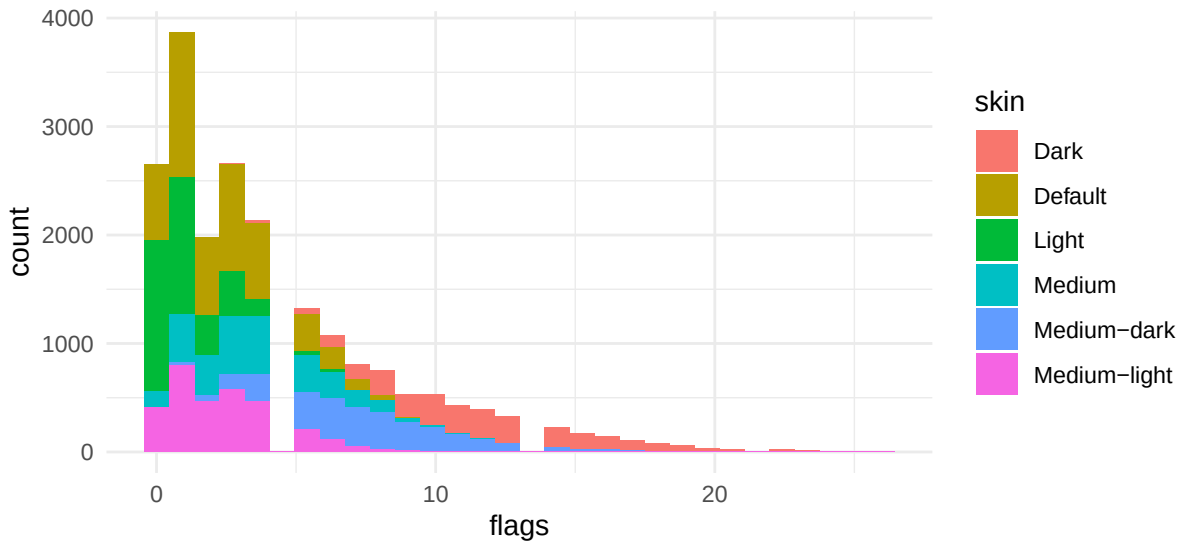

Figure 4: Most of Customers with darker skin is located on the right side. Therefore, we expect to see associeation between skin and count of flag

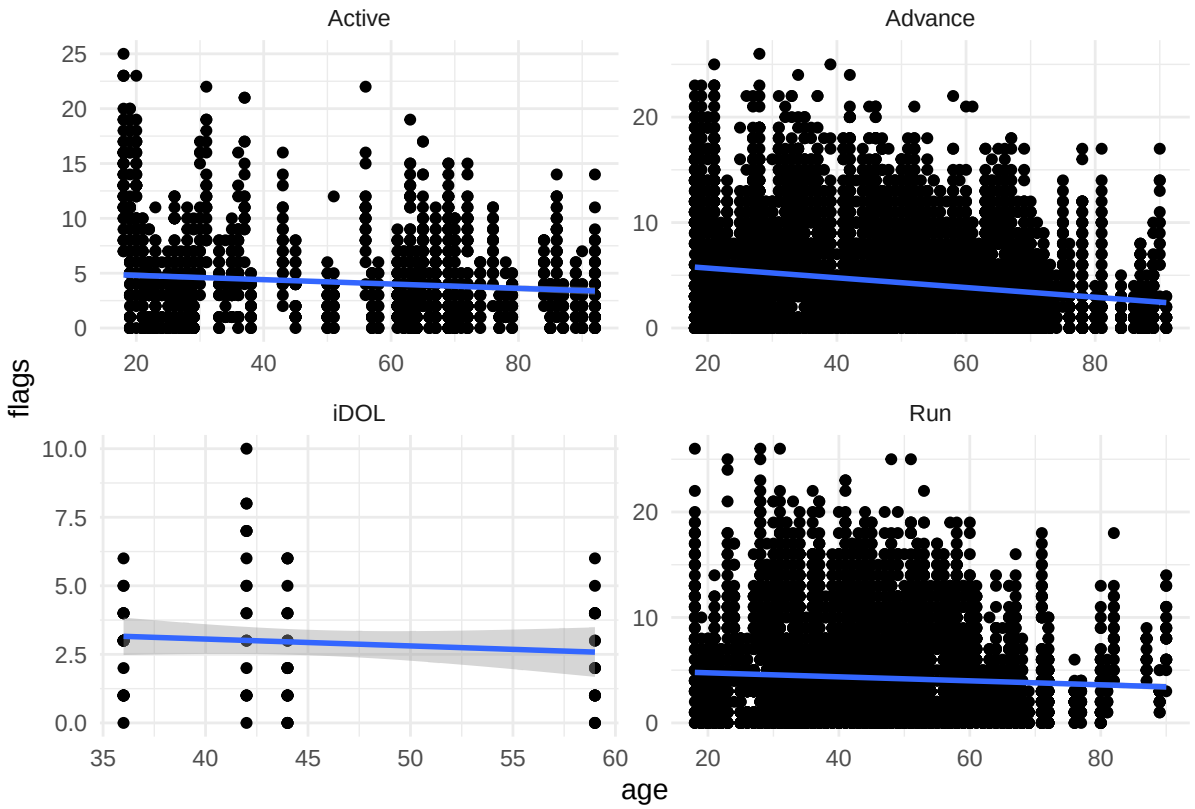

Figure 5: The impacts of age on flags tends to not be vary by different product lines. Therefore, we don't include random effect of product lines in our model.

Geographic analyse

Model

For the same reason as our model for the marketing problem, we are concerned about the dependence on observations. We also assign observations to different groups by population. Furthermore, in this data, the flags of one customer can be recorded more than one time. Therefore, we build our model with two grouping units (population group and customer group). Thus, the observational unit is count of flag for a particular customer in a particular population group. In our model, we assume the impacts of explanatory variables are not randomized by the random effects.

We notice that the count of flags is a count data. Therefore, Poisson regression is more appropriate than a normal linear regression. In Poisson regression, we assume a linear relationship between transformed *flags* (*log(flags)*) and explanatory variables. Therefore, we choose to use the mixed

Poisson model.

We first constructed a full mixed Poisson model with all features of the customer (sex, skin, age, regional median income). However, we found the estimates for sex and regional median income are not statistically significant. Therefore, we narrowed our model by reducing these two variables. We generated a p-value of 0.3853 which means we have no evidence against the simpler model explaining the data just as well as the full model. So, our final model is:

$$log(\mu_{ij}) = \beta_0 + \beta_1 skinDefault_i + \beta_2 skinLight_i + \beta_3 skinMedium_i + \beta_4 skinMedium\_dark_i + \beta_5 skinMedium\_lig$$

where

- $log(\frac{\pi}{1-\pi})_{ij}$ *is the log count of* i^{th} *customer's flags in* j^{th} *population group*
- β_0 *is customer i's log count of flags who is dark skin with minimum age*
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ *are the coefficient for explanatory variables (fixed effects)*
- $b_{1j} \sim N(0, \sigma_b^2)$ *is the random intercept of* j^{th} *population group*
- $\_b_{2i} \sim N(0, \sigma_b^2)$ is the random intercept of i^{th} customer.

Note: In order to avoid issues of scaling, we rescale our explanatory variables by dividing each variable by the differences between maximum and minimum.

Results:Interpretation of Parameter Estimates

Table 6: Estimates of Parameters in model 2

Parameters	Estimate	P-value
β_0	-3.384198	<2e-16
β_1	-1.631590	<2e-16
β_2	-2.390108	<2e-16
β_3	-1.212189	<2e-16
β_4	-0.499539	<2e-16
β_5	-1.613801	<2e-16
β_6	-0.046897	0.0067

Referring to the output of our estimation, we obtain estimates of our parameters among 255 population group of each customer:

- $\beta_0 = -3.384198 =$ *log mean count of flags of a customer with dark skin at age of 18*
- $\beta_1 = -1.631590 =$ *log relative count for customers with default skin*
- $\beta_2 = -2.390108 =$ *log relative count for customers with light skin*
- $\beta_3 = -1.212189 =$ *log relative count for customers with medium color skin*
- $\beta_4 = -0.499539 =$ *log relative count for customers with medium-dark skin*
- $\beta_5 = -1.613801 =$ *log relative count for customers with medium-light skin*
- $\beta_6 = -0.046897 =$ *when a customer's age increases from 18 to 92, the log count of flags decreases by 0.046897*
- All estimates are statistically significant with small p-value.

All analysis for this report was programmed using `R version 4.0.5` with R package Tidyverse, lme4, rvest, polite AND lmtest, plots were generated by ggplot2

Discussion

From the models we constructed, we found that the odds of being new customers are higher for customers who are old and with lower income regardless of the impacts on the household area. The impacts of sex are not statistically significant on the odd. For the second model, we found that the expected count of flags is higher for people with darker skin by reducing the dependence between observations. Therefore, we suggest that new customers are expected to be older and with lower income. Also, the product works poorly on people with darker skin. Customers' concerns might be true.

Strengths and limitations

The strength of this report is that both models reduce the dependence on observations. As a result, the estimates of each parameter can be more credible. Also, questions of MINGAR are explained by models. We are able to construct solid pieces of evidence for MINGAR to develop their products.

The main limitation of the first model is that we generated the skin color of customers by the skin color of the emoji they used on social media. The prediction of the skin color might not be reliable because we cannot ensure that emoji skin colors are the same as customers' skin colors.

Moreover, we included undefined skin color in our model. We cannot identify the true skin color of this part which can produce bias.

Another limitation would be violations of assumptions. We constructed the model by assuming a linear relationship between explanatory variables and transferred variables. From another perspective, we assumed all groups only have random intercepts since we were not able to investigate the true difference between the group to groups.

Consultant information

Consultant profiles

Sihan Li Sihan is a Copy Editor with SPQS,LLC. He specializes in report writing and layout. Sihan earned his Bachelor of Science, Majoring in Mathematics and Physics from University of Toronto in 2023.

Qilong Zeng Qilong is Data Analyst with SPQS,LLC. He specialized in data analysis. Qilong earned his Bachelor of Science, Majoring in Statistic and Economics from University of Toronto in 2023.

Lihang Shen Lihang is Data Analyst with SPQS,LLC. He specialized in data analysis. Lihang earned his Bachelor of Science, Majoring in Statistic and Mathematics from University of Toronto in 2023.

Xiaopu Zhou Xiaopu is Cartographers with SPQS,LLC. He specialized in data visualization. Xiaopu earned his Bachelor of Science, Majoring in Statistic and Mathematics from University of Toronto in 2023.

Code of ethical conduct

- All the results obtained from the analysis of the data are objective and unbiased. In the report, all data analysis processes are transparent and each step is professionally based, not on a personal understanding of bias.
- The consequences of any errors that may occur throughout the report are borne by the four members of our group together.
- This report does not contain or engage in any conflict of interest, and all statements in the report are neutral and do not contain any interests or political leanings.
- Possible errors or missing data are described in the report and explained effectively.
- All the data we used that contains personal information will not occur in our result, which means will not cause an information security issue.
- Any statistician will be able to comment on this report and monitor it effectively.

References

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Hadley Wickham (2021). rvest: Easily Harvest (Scrape) Web Pages. https://rvest.tidyverse.org/, https://github.com/tidyverse/rvest.

Dmytro Perepolkin (2019). polite: Be Nice on the Web. R package version 0.1.1. https://github.com/dmi3kno/polite

Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Ferreira, Fernandes, C. I., Rammal, H. G., & Veiga, P. M. (2021). Wearable technology and consumer interaction: A systematic review and research agenda. Computers in Human Behavior, 118, 106710–. https://doi.org/10.1016/j.chb.2021.106710

Appendix

Web scraping industry data on fitness tracker devices

The industry data is web scrapped from https://fitnesstrackerinfohub.netlify.app/ with a crawl delay of 12 seconds. We provide user information and the reason for scrapping.

Accessing Census data on median household income

We generate median household income from Census data. We access the 2016 census data from the library "census" by keeping median household income, population, and CSDuid (for matching with postcode).

Accessing postcode conversion files

The postcode conversion files from Census data needs permission to access. We used CSDuid to merge postcode with median household income. Then, we merged median household income with customer data provided by MINGAR through postcode.