View metadata, cita

HOW MENSTRUAL PRODUCT USERS DIFFER - A LOGISTIC REGRESSION MODEL

JOHANNA HELLDÉN

Bachelor's thesis
2016:K1

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

Johanna Helldén
Sammanfattning av kandidatuppsats i matematisk statistik
Lunds Universitet
2016-01-13

SKILLNADER HOS MENSSKYDDSANVÄNDARE

- en logistisk regressionsmodell

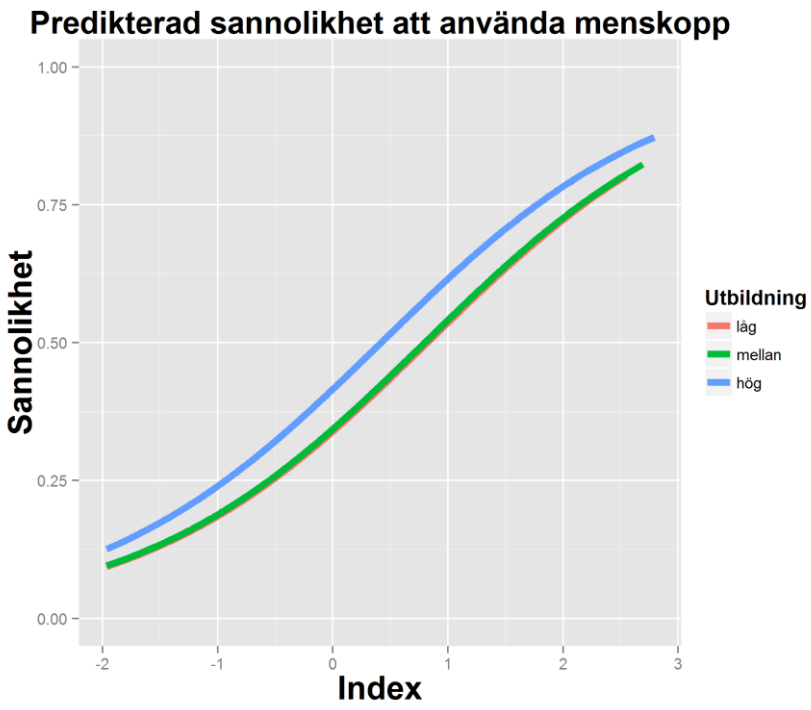
De två vanligaste mensskyddssorterna är bindor och tamponger. De senaste åren har ett nytt mensskydd dykt upp och börjat tävla om användarna, nämligen menskoppen. Menskoppen uppfanns samtidigt som tampongerna men har först nu letat sig in i svenska affärer.

Men hur skiljer sig menskoppsanvändarna åt från de andra användarna? Att använda menskopp kräver en viss närhet till den egna kroppen och menstruationen. Är detta något som visar sig hos personerna i fråga? Går det att förutse vem som är en menskoppsanvändare utifrån viss fakta om en person?

För att svara på frågorna gjordes en enkät om inställningar till mens och mensskydd, enkäten spreds på sociala medier under julen 2014 och fick över 3000 svar. Med hjälp av svaren i enkäten gjordes sedan en statistisk modell över hur menskoppsanvändarna skiljer sig åt från de andra användarna.

Menskoppsanvändarna visade sig då ha en mer positiv inställning till mens än de andra användarna. Ju mer positiv inställning en person hade desto större var sannolikheten att hen använde en menskopp istället för tamponger och bindor.

Bilden visar hur sannolikheten att vara en menskoppsanvändare ökar vart eftersom inställningen till mens (Index) blir mer positiv. Linjen i bilden är delad i tre delar, som alla representerar personer med olika hög utbildningsnivå. Vi kan se att de med hög utbildning visade sig ha högre sannolikhet att använda menskopp än de med lägre utbildning eftersom den blå linjen hela tiden ligger en bit över de andra två linjerna.



Slutligen kan kritik framföras mot studien i och med att svaren samlades in på ett sådant sätt att det är svårt att veta ifall modellen kan appliceras på befolkningen i stort eller inte. Det kan hända att de som svarade på enkäten svarade på ett sätt som är väldigt olikt hur befolkningen i stort hade svarat.

Abstract

The two most popular menstrual products, tampon and menstrual pads are challenged by a product that is on the rise; the menstrual cup. How do the menstrual cup users differ from the other users? Do menstrual cup users have a more positive feeling towards menstruation?

The results from a survey about menstruation that collected over 3000 answers are used to answer this. Users of three different menstrual products, menstrual cup, menstrual pad and tampon are analysed depending on their feeling towards menstruation, their age and their education. Multinomial and binomial logistic regression models are used, and different versions of the models are compared and assessed with cross validation.

The menstrual pad and tampon users are very similar to each other but the menstrual cup users differ and are found to have a more positive feeling towards menstruation.

The sample is however non-representative and the results should not be used to draw conclusions about the main population.

Keywords: Multinomial logistic regression, statistical modelling, logistic regression, cross validation, menstruation, menstrual cup, tampon, menstrual pad

Contents

List of Tables

List of Figures

Acknowledgements

First of all, I would like to give my thanks to the generous persons who gave me of their time and experiences when they answered the survey; without you this thesis would not have been written.

Secondly, all who helped me spread the survey are also a big part of what made this work possible. Thank you for giving me the usage of your extended networks.

I want to thank my supervisor Anna Lindgren for help with finding the way in the choice of statistical methods and models.

Thank you, friends and family who helped me with comments, encouragements and hints. Martin, who has been involved and committed to my project from the beginning when I constructed the first version of the survey in his kitchen. Claire, who gave valuable help as a native speaker of the English language. Dad, who read from an outside perspective. Henrik, who knows more than me about R and Latex and who are not afraid of giving help to someone in need. Sebastian, who put up with me talking about the project through endless lunches and who gave me insight into his own thesis process.

Chapter 1

Introduction

The menstrual product market is dominated by two products; the tampon and the menstrual pad. But there is one product that seems to be up and coming, namely the menstrual cup. There are menstrual cup users that sound almost like missionaries when talking about the change from tampon/menstrual pad to menstrual cup. It also seems (from a limited sample of private empirical data) as if menstrual cup users have a more positive attitude towards menstruation than other menstruators. Whether the menstrual cup is as liberating as they say is not easy to measure. But it is possible to find out if menstrual cup users have a different view on menstruation. The aim of this thesis is to use data collected in a survey about menstruation and see if a multinomial logistic regression can predict the use of menstrual products depending on the view the person has on menstruation. And if prediction is possible, how big are the effects of different characteristics on product choice?

1.1 Background

Previous studies show that the main products used by menstruators are menstrual pads and tampons [6, 13]. These two kinds of products are one use products with a slight different usage, where pads are for external use and tampons internal use. The menstrual cup, which is not as widely used as the first two (even though it nowadays is possible to by it in Swedish pharmacies) is an internal product just as the tampon. The menstrual cup is however not one use, and "[t]he menstrual cup requires more comfort and direct contact with the body and menstrual blood than most mainstream disposable products (e.g., tampons with an applicator)" [8]. There is reason to believe that there is something that differs between menstrual cup users and other users since their product

experience differ.

When social psychologists [8] studied perceptions of the menstrual cup in non users, the results showed that overall there was a negative attitude towards the menstrual cup, and the persons being more positive had heard about the menstrual cup beforehand. The researchers also found that people with more positive attitudes towards menstruation were more positive towards the use of a menstrual cup.

1.2 Aim

There is reason to believe that menstrual cup users differ from other menstruators. This could be an effect of the usage or something that singles out the users even before they switch to menstrual cups. [1]

This thesis will examine if and how menstrual cup users differ. The focus will be on measuring the difference in how menstruators feel about menstruation. This could be done in two different directions where the first is to use product usage as an explanatory variable and see how it effects the feeling towards menstruation. The second direction is to use the feeling towards menstruation as an explanatory variable and product usage as the outcome of the feeling. The focus in this thesis will be on the second version, with product usage as an outcome.

The results of this paper can be of use to a wide range of people, for example medical staff, sex education teachers and companies producing menstrual products, as well as the menstruators themselves.

1.2.1 Research questions

- Do feelings towards menstruation effect the choice of menstrual product?
- Is it possible to predict usage of menstruation products depending on feelings towards menstruation?

[1] Previous research show that mensturators almost always have menstrual pads as their first product [1]

Chapter 2

Theory

Predictions and studies of the effect a variable has on another, can be made with the help of linear regression. In linear regression, there is an outcome variable Y and one or more explanatory variables x. A model for the connection between Y and x when there is only one explanatory variable is

$$Y = \beta_0 + \beta_1 * x$$

There is always some errors in measuring and some natural variation, therefore each observation is assumed to have the following form

$$Y_i = \beta_0 + \beta_1 * x_i + e_i$$

where e_i is presumed to be normally distributed with mean zero and variation 1, N(0,1). Given observations of Y and x, β-coefficients that fits the data best can be computed.

The regression in this thesis will be built on generalisations of linear regression models and the theory behind it will be presented in this chapter.

2.1 Generalized linear models

The Generalized Linear Models (GLM) can be specified by the fact that they all have a *random component*, a *systematic component* and a *link function*.

Firstly there is the response variable Y, which is identified by the *random component*. If we have $(Y_1, Y_2, ...Y_n)$ as observations of Y then these can be said to come from a specific distribution, for example Normal or Binomial distribution. In linear regression, normal distribution is used.

The *systematic component* is a linear combination of the explanatory variables denoted by x_j, and the combination of x_j with coefficients β_j is called the linear predictor:

$$\beta_0 + \beta_1 x_1 + ...\beta_k x_k$$

The last component in the GLM is the *link function* $g()$, which links the linear predictor to the mean of Y, $E(Y) = \mu$. We have the following formula

$$g(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

Linear regression is a kind of GLM with $g(\mu) = \mu$. Two other common link functions are $g(\mu) = \log(\mu)$ which is called the log-linear link, and $g(\mu) = \log[\mu/(1-\mu)]$ which is called the logit link, the latter is what is used in a logistic regression model.

Just as with linear regression, with observations of Y and x, β_j can be estimated.

2.2 Logistic regression

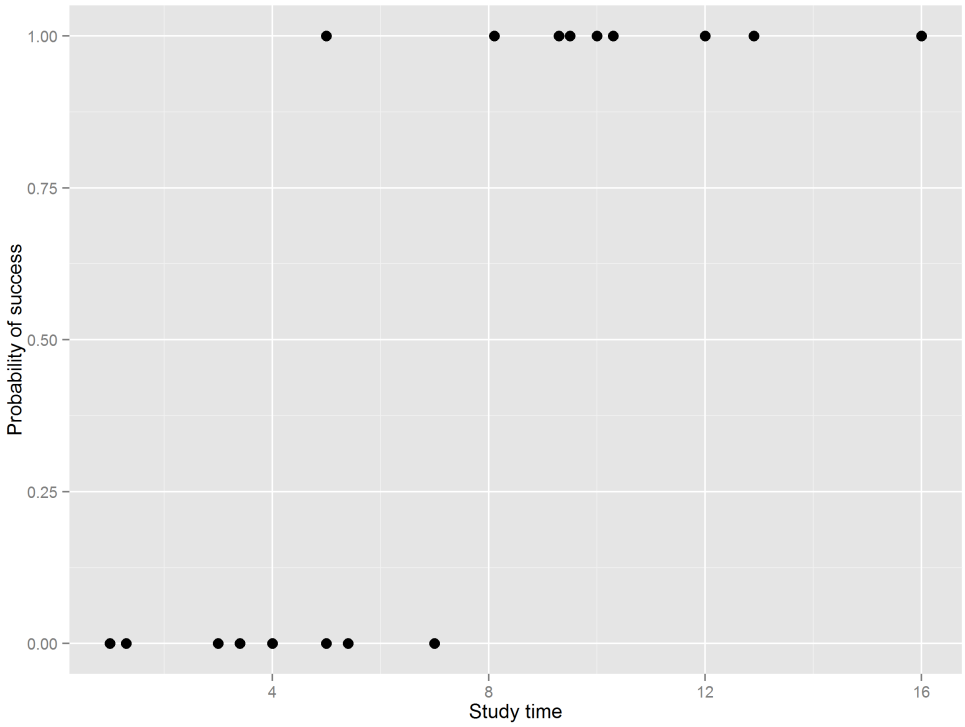
In the case of a categorical response variable Y, with two outcomes that can be thought of as success and failure, it is not advisable to use standard linear regression even if it is possible.

Say we have observations of a phenomenon, maybe success or failure on a test and time spent studying. Figure 2.1 shows study time and success or failure for 17 made up students, the x-axis is study time and the y-axis is success or failure with 1 as success and 0 as failure. Each dot is a student.

If a linear regression model is used to model the example-data despite earlier advice, with *success* given the value 1 and failure value 0 and the outcome used as a continuous variable, the predictions will be values that could be thought of as probabilities. The predictions can be defined as belonging in one or the other group with a cut-off at value 0.5.

Figure 2.2 shows a linear model applied on the example data. The result is that all student that studied more than eight hours are predicted as being more probable to have a success and the ones who studied less than eight hours are predicted as being probable to fail. In this way we get predicted values outside of [0,1], which makes no sense since a probability cannot be higher than 1 or lower than 0. In real life there is only success and failure, but some students are predicted to more than pass the test or be worse than a failure. The probability is also handled as if it would be linear in this example whereas it is better to think of probabilities as having a bell shape.

Figure 2.1: Example data, studying time and success on a test



It is not as usual to get a probability close to 0 and 1 as it is to get one that is close to 0.5. [2, p. 120]

The most popular model for data with binomial outcomes is logistic regression [3]. If we have k explanatory variables $x_1, ..., x_k$ and call $\pi(\mathbf{x})$=P(Y=success|$\mathbf{x}$) the probability of success given values $\mathbf{x} = [x_1, ..., x_k]$. The model has the form:

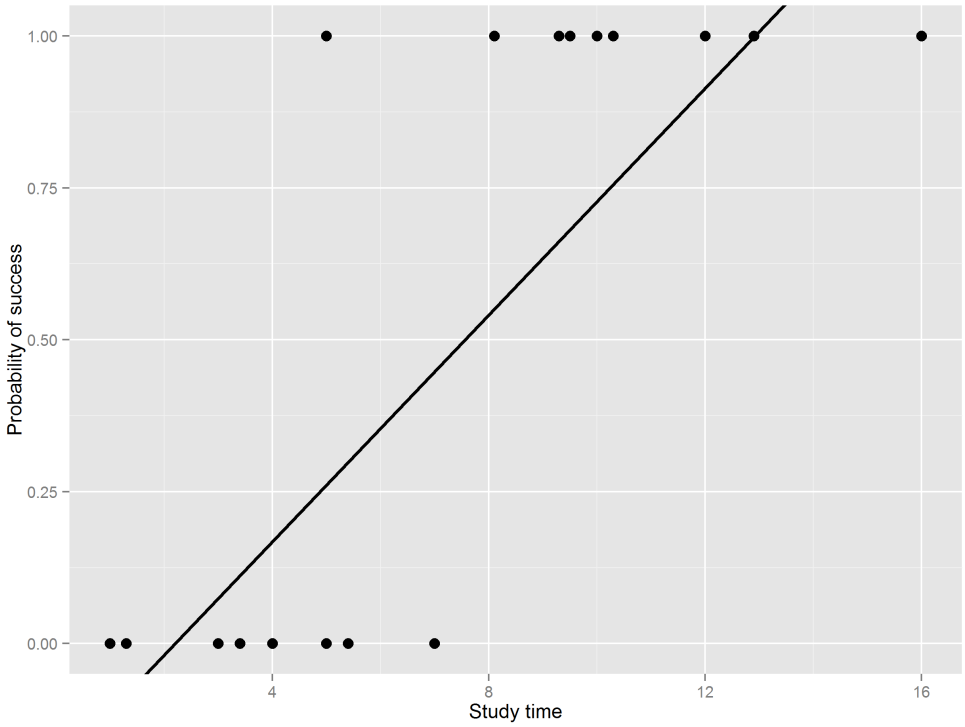
$$\text{logit}[\pi(\mathbf{x})] = \log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

This is a Generalized linear model with Bernoulli distribution for Y, and with the logit function used as link function. The mean of a Bernoulli(π) is π. With observations of Y and x, β_j can be estimated. Figure 2.3 shows a logistic model for the data on students and test results. The dotted line is the linear model and the regular line is the logistic model. In this model, no one is predicted to more than pass or more than fail.

Usually when interpreting logistic regression models, the odds and odds ratios are used. The odds of *success* for the model is

$$\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k) \tag{2.1}$$

Figure 2.2: Example data, Linear model



The *odds* for success is the probability for success divided by the probability of failure. An odds ratio on the other hand, is the multiplicative change in odds.

To find the probability for a *success* given a value of $\mathbf{x}$, the equation is solved for $\pi(\mathbf{x})$, giving

$$\pi(\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}$$

Interpreting the β-coefficients from a logistic regression model is slightly different from interpretation in a linear regression model. In equation (2.1) we see that a 1 on a categorical binary variable x_i is represented by (2.1) being multiplied by e^{β_i}. Strictly speaking, this means that the β-coefficients should be interpreted as odds-ratios.

A positive β-coefficient on a continuous variable means that with growth in this variable, the probability of *success* will rise. A negative β-coefficient is equal to lessening of the probability of *success*.

A positive β-coefficient on a categorical variable means that the person who has a 1 on this variable has a higher probability of *success* than a person who does not belong in the group.

The sizes of the coefficients give the translations that can be seen in table 2.1. A β-coefficient

Figure 2.3: Example data, Logistic model

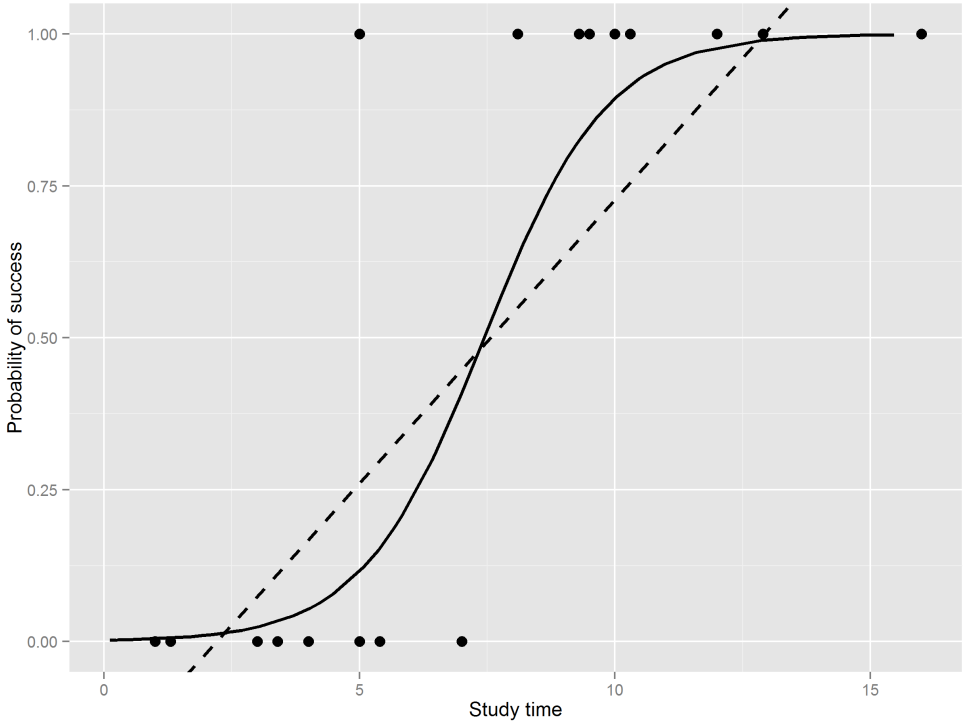


Table 2.1: Magnitudes of coefficients in logistic regression

β	$\exp(\beta)$	$\frac{\exp(\beta)}{1+\exp(\beta)} = \pi$
∞	∞	1
0	1	0.5
-∞	0	0

of plus infinity corresponds to a probability 1 and a β-coefficient of 0 means probability 0.5 which means that both outcomes are as likely. A β-coefficient of minus infinity corresponds to probability zero.

2.3 Multinomial logistic regression

When the response variable has more than two categories, ordinary logistic regression does not work and instead multinomial logistic regression has to be used. Say we have J categories of Y and $\mathbf{x} = [x_1, ..., x_k]$. Let $\pi_j(\mathbf{x}) = P(Y = j|\mathbf{x})$, the probability of getting j given x, with $\sum_j \pi_j(\mathbf{x}) = 1$. Then choose a *baseline category* J, which is often the last category or the most usual one [2, p. 268]. Then the model is for j = 1, ..., J - 1

$$\log \frac{\pi(\mathbf{x})_j}{\pi(\mathbf{x})_J} = \beta_{j,0} + \beta_{j,1}x_1 + ... + \beta_{j,k}x_k \tag{2.2}$$

The result is J-1 sub-models which are logistic models with β that are interpreted as the effect of $\mathbf{x}$ on the odds of being in category j rather than category J. The probability of being in a category given $\mathbf{x}$ is found by solving equation (2.2) for $\pi(\mathbf{x})_j$ and using $\sum_j \pi_j(\mathbf{x}) = 1$, it is then for j = 1,...,J and with all $\beta = 0$ for j = J

$$\pi(\mathbf{x})_j = \frac{\exp(\beta_{0,j} + \beta_{j,1}x_1 + ... + \beta_{j,k}x_k)}{\sum_{h=1}^{J} \exp(\beta_{h,0} + \beta_{h,1}x_1 + ... + \beta_{h,k}x_k)}$$

Chapter 3

Method

This chapter contains description of the used data. After that, the different methods of choosing a model will be described. Lastly the implementation will be presented.

3.1 The Data

The data that is analysed in this study was collected with a survey about menstruation and menstrual products (Original name: "Enkät om inställningar till mens och mensskydd") consisting of 26 questions. The survey was constructed and sent out on social media platforms during Christmas 2014. 3 195 persons answered the questionnaire and of these, 2 827 persons answered all the questions that are used in this thesis. All respondents needed to answer that they have been menstruating during the previous year to be able to participate in the survey. The sample is non-representative, since it is collected in a non-representative way. The demographic questions included in the survey shows that the respondents are more educated and younger than the group of menstruating Swedish persons which further shows that the sample is not representative. [9]

The questions from the survey that are used in this thesis are included in Appendix A.

3.1.1 Variables

The variables that will be used to build the models in this thesis are the following:

Main menstrual product: Has three answers - menstrual pad, menstrual cup and tampon. People who answered differently than these three were excluded from the sample.

Feeling towards menstruation: Normalised index created from seven survey questions. The questions, which are presented in Appendix A (question F6), ask whether sentences apply to the

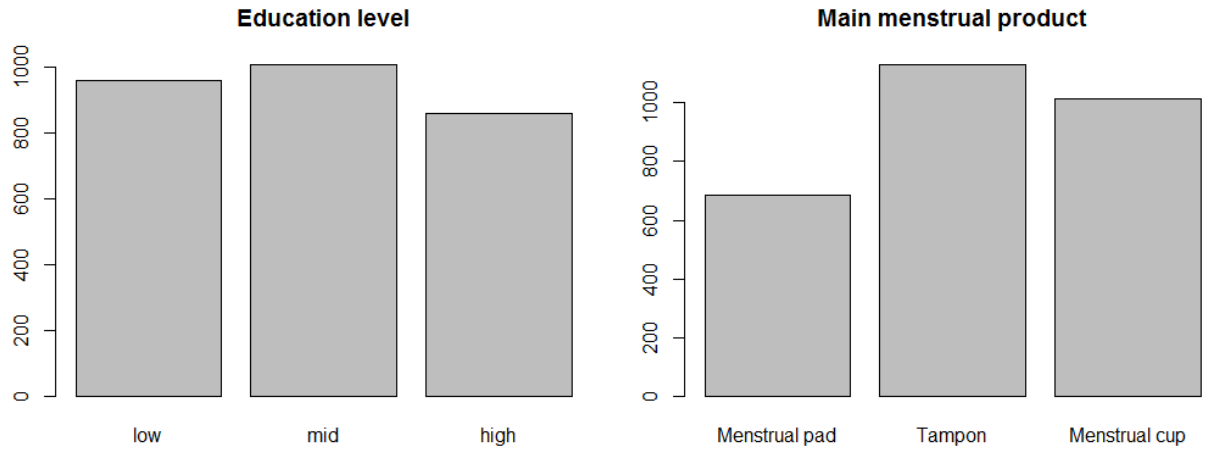
respondent and lets the respondent use a scale as to answer how much it applies. The index was created with a factor analysis and by making a weighted sum of the seven questions. The weights are presented in Appendix B. A detailed description of the creation of the index can be found in [9]. A score above zero on the index means more positive feeling than the average, negative score means less positive feeling. This variable will sometimes be referred to as the index.

Age: Age of respondent in years.

Education level: Has three levels, high, mid and low. Low equals studies below university level, high is graduated from university level and mid equals started studies at university level.

3.1.2 Descriptive statistics

Figure 3.1: Barcharts for categorical variables



(a) (b)

The education variable is divided into three groups of similar size (see figure 3.1a). The division is due to the creation of the education levels which was made with the intent of creating groups of similar sizes.

The response variable main menstrual product has three groups where tampon holds the highest amount of people and menstrual pad the lowest (see figure 3.1b).

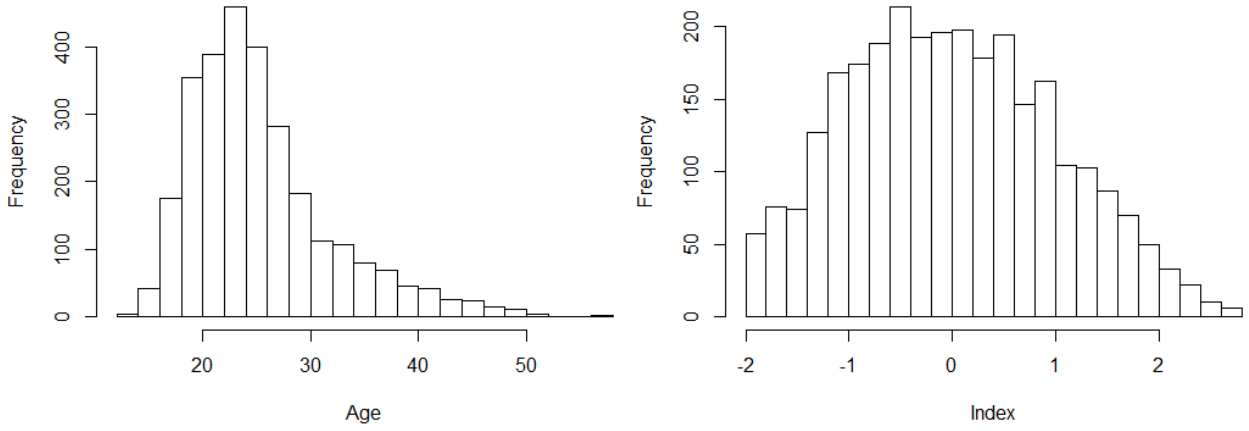
The ages of the respondents vary from 13 to 57. The mean is 25.85 years (see table 3.1 and figure 3.2a).

Table 3.1: Description table of continuous variables

Statistic	N	Mean	St. Dev.	Min	Max
Age	2 827	25.85	6.66	13	57
Index	2 827	0.00	1.00	−1.97	2.79

The index measuring *feeling towards menstruation* is standardised, that means that it has mean 0 and standard deviation 1 (see table 3.1 and figure 3.2b).

Figure 3.2: Histogram over numerical variables



(a) (b)

There are respondents of all education levels in all the groups of menstrual product users (see table 3.2).

The age variable and the index has a low correlation, as well as index and education. But education and age are more correlated, which can be expected since a young person does not have the possibility to have a high education (see table 3.3).

Table 3.2: Menstrual product usage and education level

		Main product			
		Pad	Tamp	Cup	Sum
Education	Low	237	430	292	959
	Mid	225	408	375	1008
	High	223	291	346	860
	Sum	685	1129	1013	2827

Table 3.3: Correlation matrix for independent variables

	Education	Age	Index
Education	1	0.46	0.13
Age	0.46	1	0.11
Index	0.13	0.11	1

3.2 The IIA-assumption

An assumption for multinomial logistic regression models is that Independence of Irrelevant Alternatives (IIA) is true. The IIA-assumption is usually explained with a transportation example.

Say that we want to model whether people in a community want to take the bus or a car to work. In this particular community the buses are red. Say that the proportions taking each of these are 1/2 and 1/2. The IIA- assumption would then suggest that if we include one other category, the proportions between the two categories should be the same as before. That is, if we included another transportation choice, the categories car and bus should still be 1/3 each (or some other proportion where red bus and car has the same probability and the third choice has the remaining proportion). But if the new category is another bus, but a bus that is blue, then 1/2 of the population would probably still choose a car and 1/2 would choose bus, since the colour of the bus does not matter to them. The 1/2 of the population who chooses bus would be split between

the two buses, red and blue. Say we have 1/2 choosing car, 1/4 choosing red bus and 1/4 choosing blue bus. This means that the IIA-assumption does not hold for the transportation example since the proportion choosing car and red bus are no longer equivalent. The blue bus is an irrelevant alternative that is to similar to the red bus and therefore the proportions are not independent. [12]

The are tests that are said to test this assumption and one of these are the Hausman-McFadden test [11]. However, using testing to check the IIA-assumption is criticised and might not work since the IIA is a characteristic of the way the variables or choices are defined, rather than something measurable [4]. The model can be switched to for example a multinomial probit model if the assumption is violated. However, even if the multinomial probit model strictly speaking is better when the IIA-assumption is violated, the model will in the end be almost indistinguishable from a multinomial logistic model [7].

The model in this project could be violating the IIA-assumption of multinomial logistic models since we do not know how the respondents would choose between products if one of the alternatives did not exist. But since the probit model usually takes extra work without giving especially different estimates [7], the multinomial logistic model will be used. Note that the model chosen in the end of this paper is a binomial model and thus the IIA-assumption is irrelevant for the final model.

3.3 Choice of model

For the choice of model the data will be randomly ordered and then divided in two parts; one model-building and one final testing part. 10/11 of the data will go into the model-building and 1/11 to the final testing. Cross validation, AIC-values and significance testing will be made on the 10/11 of the data and will help choosing the model. When the final model is chosen, it will be taken to a finalised test where it is checked how well it can predict the values of the last 1/11 of the data. This kind of division is possible thanks to the large amount of data (N = 2827).

To choose which versions of models to test, different variables will be excluded from a larger model to see whether this can be done without spoiling the explanatory effect too much.

3.3.1 Cross validation

Cross validation is dividing the data into different subgroups where the model is specified or "trained" on one group and tested on another. This can be done in different ways. One way is to use K-fold-cross validation which is what will be used here. In K-fold cross validation, the data is

divided into K parts. For each of these K groups, the other K-1 groups are used to make a model which will be used to predict the last part. In this way an observation is never "creating" its own model. In the choice of model in this thesis, 10 folds will be used in the cross validation. This since the data is divided into 11 parts where 10 parts is used for the testing.

The precision of the predictions can be measured in different ways. One is to use accuracy:

$$\text{ACC} = \frac{\text{Number of correctly predicted observations}}{\text{Number of observations}}$$

If there are three categories and a large number of observations, a random assignation of groups would lead to an accuracy of approximately 1/3. The number of 1/3 can be used as a point to measure against.

The cross validation is usually presented with a table (confusion matrix), showing real and predicted group for all the observations. A confusion matrix table can look like table 3.4, where the columns are the predicted groups and the rows are the real groups. When making this table,

Table 3.4: Example of confusion matrix for three variables

		Predicted group			
		1	2	3	Correct
Real group	1	A	B	C	A/(A+B+C)
	2	D	E	F	E/(D+E+F)
	3	G	H	I	I/(G+H+I)

$$\text{ACC} = \frac{\text{A+E+I}}{\text{A+B+C+D+E+F+F+H+I}}$$

all cases are forced to "choose" a group. Each observation gets a probability to be in each group, and then it is put in the group for which it has the highest probability. To avoid the roughness of this measure (some cases might fit almost equally well in each group) a modified version of the accuracy (ACC2) will also be presented in this thesis. In the modified version, only cases who has probability higher than 0.5 to be in a specific group get predicted. The ones who do not fulfil this is deemed unpredictable, and counted as missing (NA). Then the proportion NA's and the modified accuracy for the predicted cases can be measured.

3.3.2 Deviance

Two models M_1 and M_0 where M_0 is nested in M_1, where nested means that M_0 can be created by excluding variables from M_1, can be compared using deviance difference:

$$\text{Deviance Difference} = -2(\text{Log likelihood for } M_0 - \text{Log likelihood for } M_1)$$

The deviance difference has an approximate χ^2 distribution with df = the difference between numbers of parameters. If the deviance difference is smaller than the value for $\chi^2_{df}(0.05)$, the simpler model is kept since this means that it does not explain the observations significantly worse than the bigger model. If the deviance difference is larger than the cut-off value, the more complex model is kept since the simpler one is significantly worse at explaining the observations.

3.3.3 AIC and BIC

Models can also be compared using the Akaike information criterion (AIC). The AIC is a measure on how well the model fits the data. The best model has the lowest AIC which is defined as

$$\text{AIC} = \text{-2(Log likelihood - number of parameters)}$$

The AIC penalises the log-likelihood on the number of parameters, giving bigger models a lower value. Another similar measure is the Bayesian information criterion (BIC)

$$\text{BIC} = \text{-(2*Log likelihood - number of parameters * ln(number of observations))}$$

The lowest BIC is the best. The BIC penalises a model for the numbers of parameters more severely than the AIC. Since AIC is criticised on its tendency to choose models with larger numbers of independent variables, BIC can be used as an alternative or complement [14].

In this paper, both AIC and BIC will be used.

3.4 Implementation

The modelling is made in R, with the help of four packages. Nnet [15] provides the function for estimating the multinomial model. Mlogit [5] provides extra information on the model. Stargazer [10] provides tables from R-output, even though all tables in this thesis are somewhat altered from the originals. GGplot [16] provides the plots of the final models.

Chapter 4

Results

The first section in this chapter presents eight different multinomial logistic regression models and chooses which one is the best. After that, the section also presents eight binomial logistic models and chooses which one of these that is the best. These two best models are then tested on the test data.

The second section in this chapter presents the two chosen models more thoroughly and interprets the parameters.

4.1 Choice of model

4.1.1 Multinomial logistic model

Eight multinomial models were selected and tested and can be seen in table 4.1. Model 1 has all three independent variables and also interactions between them. From model 1, three models (2a-c) were made where one interaction term is excluded in each. The deviance difference is largest (38.06 (df=4)) when the interaction between age and education is excluded in model 2b. From 2a-c a new model (number 3) was created where the common ground between 2a and 2c was kept, that is the three independent variables and the interaction term between age and education. The deviance difference between model 3 and the first model is 8.75 (df=6), which is not very high. The three last models (4a-c) tests what happens if one of the variables is excluded from the model. The deviance difference is high in all these three cases. All in all, deviance difference helped with choosing which models to test.

The AIC proposes to choose model 2a, but model 1, 2b and 3 has almost as low AIC as 2a. The

accuracy (ACC) is very similar for all models. The BIC proposes that 4b is the best model. BIC is known for penalising big models more. However, the next best BIC and next best AIC coincide in model 3. This mean that according to both criteria it is an okay model.

The other measures, ACC, ACC2 and NA are not taken into account any more than to say that they do not differ much between the models. Only model 4c seems exceptionally bad to use for predictions.

Table 4.1: Model choice 1

NR	Predictors	Dev	edf	AIC	BIC	Compar.*	Dev Diff	ACC**	ACC2***	NA***
1	I + A + E + A*E + I*A + I*E	5102.81	20	5142.81	5259.84	-		0.50	0.55	0.50
2a	I + A + E + A*E + I*A	5106.58	16	5138.58	5232.20	2a - 1	3.77 (df=4)	0.50	0.56	0.54
2b	I + A + E + A*E + I*E	5105.23	18	5141.23	5246.56	2b - 1	2.42 (df=2)	0.50	0.55	0.51
2c	I + A + E + I*A + I*E	5140.87	16	5172.87	5266.50	2c - 1	38.06 (df=4)	0.50	0.56	0.54
3	I + A + E + A*E	5111.56	14	5139.56	5221.48	3 - 1	8.75 (df=6)	0.50	0.56	0.50
4a	I + E + I*E	5184.90	12	5208.90	5279.12	4a - 1	82.09 (df=8)	0.50	0.56	0.55
4b	I + A + I*A	5149.04	8	5165.04	5211.85	4b - 1	46.23 (df=12)	0.51	0.57	0.54
4c	A + E + A*E	5434.67	12	5458.67	5528.89	4c - 1	331.86 (df=8)	0.43	0.45	0.93

Note: I= Index, A = Age, E = Education

N = 2570

* States which models that are compared on deviance

** Created with 10-fold cross validation

*** Only predicting for cases with $p > 0.5$, ACC2 should be maximised and NA should minimised

Model 3 was chosen as the best of the eight multinomial models. Table 4.2 shows the confusion matrix for model 3. The wrong predictions are made up of a lot of menstrual pad users predicted to use tampons. Only ten percent of the menstrual pad users are predicted in the right group by model 3 (see table 4.2). 60 percent of the menstrual pad users were confused with tampon users. For the tampon and menstrual cup users however, a majority were predicted in the right group.

Table 4.2: Confusion matrix for model 3

		Predicted group			
		Pad	Tamp	Cup	Correct
Real group	Pad	65	376	184	0.10
	Tamp	60	678	286	0.66
	Cup	27	352	542	0.59

$$ACC = 0.50$$

To correct for the problem with predicting menstrual pad users, a new kind of model was created where menstrual pad users and tampon users are seen as the same group. When looking at all variables and especially the index, it can be seen that tampon and menstrual pad users do not differ much, which is why the merging can be made without losing much information.

The menstrual pad users and the tampon users can be seen as the people using a red or a blue bus in the example explaining IIA earlier. The choice between tampon and menstrual pad is not different enough, and a solution to this is to merge the categories.

The new models with combined groups are presented in the next subsection.

4.1.2 Binomial logistic model

In the models presented in this section, the outcome variable only has two categories which are menstrual cup and not menstrual cup. The same independent variables as were used for the multinomial models were used for model testing with the binomial models. When reducing the model, deviance difference were used to see whether it was sensible to remove variables. In table 4.3 it can be seen that the AIC and BIC suggests the same model as the one chosen for the multinomial models, that is; number 3. Model 2a and 2b also gets suggested, and as in the last section much

extra information is not gotten from the ACC, except that the last model seems worse than the others. Notably, ACC2 is not presented in table 4.3, since the binomial models do not have as much problems with predictions as the multinomial models do.

Model 3 is chosen as the best model.

Table 4.3: Model choice 2

NR	Predictors	Dev	edf	AIC	BIC	Compar.*	Dev Diff	ACC**
1	I + A + E + A*E + I*A + I*E	2976.11	10	2996.11	3054.63	-		0.69
2a	I + A + E + A*E + I*A	2979.41	8	2995.41	3042.23	2a - 1	3.30 (df=2)	0.69
2b	I + A + E + A*E + I*E	2976.77	9	2994.77	3047.43	2b - 1	0.65 (df=1)	0.69
2c	I + A + E + I*A + I*E	3005.99	8	3021.99	3068.80	2c - 1	29.88 (df=2)	0.68
3	I + A + E + A*E	2980.38	7	2994.38	3035.34	3 - 2b	3.62 (df=2)	0.69
4a	I + E + I*E	3007.74	6	3019.74	3054.85	4a - 1	31.62 (df=4)	0.69
4b	I + A + I*A	3013.16	4	3021.16	3044.57	4b - 1	37.05 (df=6)	0.69
4c	A + E + A*E	3303.93	6	3315.93	3351.04	4c - 1	327.81 (df=4)	0.64

Note: I= Index, A = Age, E = Education

N = 2570

* States which models that are compared on deviance

** Created with 10-fold cross validation

4.1.3 Testing the models

In the last two sections, two models were chosen, model 3 multinomial and model 3 binomial. In this section, these models are tested by letting the models predict the groups for the test data that was left out when the models were created. There are 257 observations that will be predicted by the two models. This data has not been part of the model building.

The multinomial model has an accuracy of 0.53 (see table 4.4). As seen before, the pad users are not predicted in the right group. Only about 3 percent of the menstrual pad users are rightly predicted. As for the tampon and menstrual cup users, a majority of the cases are predicted correctly.

Table 4.4: Confusion matrix for test-data, Multinomial model

		Predicted group			
		Pad	Tamp	Cup	Correct
Real group	Pad	2	38	20	0.03
	Tamp	4	74	27	0.70
	Cup	0	32	60	0.65

ACC = 0.53

The binomial model has a higher accuracy than the multinomial model (see table 4.4 and 4.5). This is expected due to the fact that there are less groups to be placed in. Approximately 70 percent of the cases are placed in the right group by the binomial model. Of these rightly predicted cases, the tampon/pads are more accurately predicted than the menstrual cup users.

Table 4.5: Confusion matrix for test-data, Binomial model

		Predicted group		
		Cup	Pad/Tamp	Correct
Real group	Cup	41	51	0.45
	Pad/tamp	23	142	0.86

ACC = 0.71

4.2 The final model(s)

In this section, the two final models from the last section are presented more thoroughly.

4.2.1 Multinomial model

Table 4.6 contains the multinomial models where the first part is the model for the probability to choose tampon instead of menstrual pad and the second part is the model for the probability to choose menstrual cup instead of menstrual pad.

Table 4.6: Final multinomial logistic regression model

	Dependent variable:					
	Tampon (ref = pad)			Menstrual Cup (ref = pad)		
	β	e^{β}	e^{I_β}	β	e^{β}	e^{I_β}
Constant	1.20***	3.31	[1.75, 6.27]	−0.62	0.54	[0.27, 1.08]
	(0.33)			(0.35)		
Index	0.07	1.08	[0.97, 1.20]	0.86***	2.36	[2.11, 2.64]
	(0.05)			(0.06)		
Age	−0.03	0.98	[0.95, 1.00]	0.04*	1.04	[1.01, 1.07]
	(0.01)			(0.01)		
Education_mid	0.37	1.44	[0.51, 4.05]	1.79**	6.01	[2.00, 18.08]
	(0.53)			(0.56)		
Education_high	1.56**	4.74	[1.63, 13.77]	3.70***	40.42	[13.30, 122.87]
	(0.54)			(0.57)		
Age*Education_mid	−0.01	0.99	[0.95,1.03]	−0.07**	0.93	[0.89,0.98]
	(0.02)			(0.02)		
Age*Education_high	−0.05**	0.95	[0.91, 0.98]	−0.13***	0.81	[0.85, 0.92]
	(0.02)			(0.02)		
Akaike Inf. Crit.	5 655.37					
Bayesian Inf. Crit.	5 738.63					
McFadden R^2	0.08					
edf	14					
N	2827					

Note: *p<0.5; **p<0.01; ***p<0.001
Standard errors in parenthesis
Confidence intervals for p = 0.05

The coefficients for high education for different age groups are presented in table 4.7. The highest age of a respondent is 57 and the lowest is 13. From table 4.6 and 4.7 we can deduce that for a person above age 29, high education as opposed to a low education level gives a lower probability of using tampons. Whereas for a person below 29, higher education gives a higher probability of using tampons instead of menstrual pads.

Table 4.7: Examples of β_{Tamp}-coefficients for education

	Age		
	13	29	57
β_{High}	0.89	-0.01	-1.52

Age 29 chosen as it is the point where the coefficient switches sign

No other coefficients than the one for high education and high education combined with age, were significantly ($p < 0.05$) different from zero and they can therefore be said to have no effect on the choice of product. That is, feeling towards menstruation (Index), age and mid level of education does not change the probability to use tampons instead of menstrual pads.

As for the second part of the model, where menstrual cup usages is measured against menstrual pad usage, all variables has a significant effect on the probability.

The main variable of interest, feeling towards menstruation (Index) has a positive coefficient which means that a positive index number gives a higher probability to use a menstrual cup, and a negative number gives a lower probability to use a menstrual cup rather than a menstrual pad.

The age coefficient must be interpreted together with its interaction terms, giving three age coefficients that are shown in in table 4.8.

For persons with low education, higher age gives a higher probability to use a menstrual cup instead of menstrual pads. But for people with mid or high education, the relation is reversed, with higher age giving lower probability to use a menstrual cup.

The variables for education must be interpreted together with age. The coefficients can be seen in table 4.9.

Table 4.8: β_{Cup}-coefficients for age

	Education level		
	Low	Mid	High
β_{Age}	0.04	-0.03	-0.09

Table 4.9: Examples of β_{Cup}-coefficients for education

	Age		
	13	26	57
β_{Mid}	0.91	0.03	-2.08
	13	29	57
β_{High}	2.05	0.02	-3.54

Age 26/29 chosen as it is the point where the coefficient switches sign

A person below age 26 with mid education has a higher probability to use a menstrual cup than use menstrual pads than a person of the same age with a low education. For a person above age 26 the effect is reversed.

For a person below age 29, it is more probable that they use a menstrual cup than menstrual pads if they have a high education. If they instead are above age 29, they are more inclined to use menstrual pads if they are highly educated.

In these cases we must bear in mind that there are no highly educated 13-year-olds. And so these coefficients for the end points of the scale must be interpreted with caution.

Figure 4.1 shows how predicted probability to use each product changes with change on the index for a person of age 13, figure 4.2 shows the same thing but with age set to mean age and figure 4.3 shows this but with age as 57. 13 and 57 is the youngest and oldest respondents. For age

= mean (25.85) and 57, the plots are divided by education level, but since there are no university educated 13 year olds, the plot for 13-year-olds only has education level low shown.

Figure 4.1: Multinomial model, Prediction plot 1

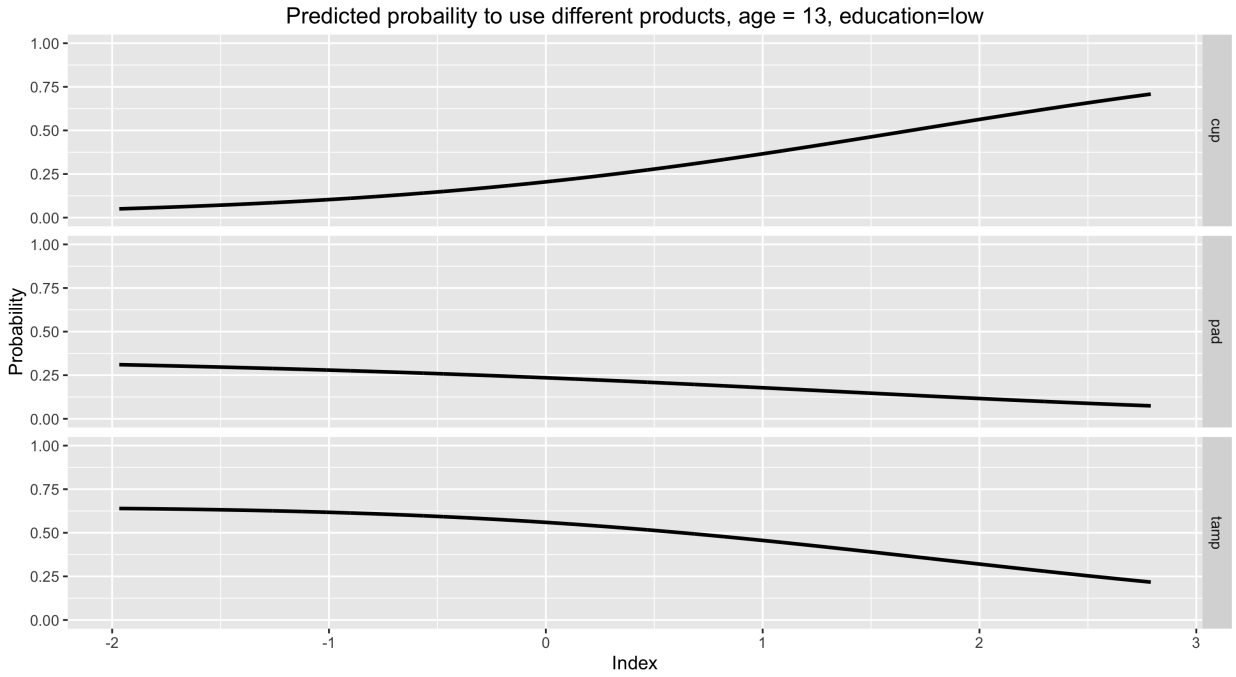


Figure 4.2: Multinomial model, Prediction plot 2

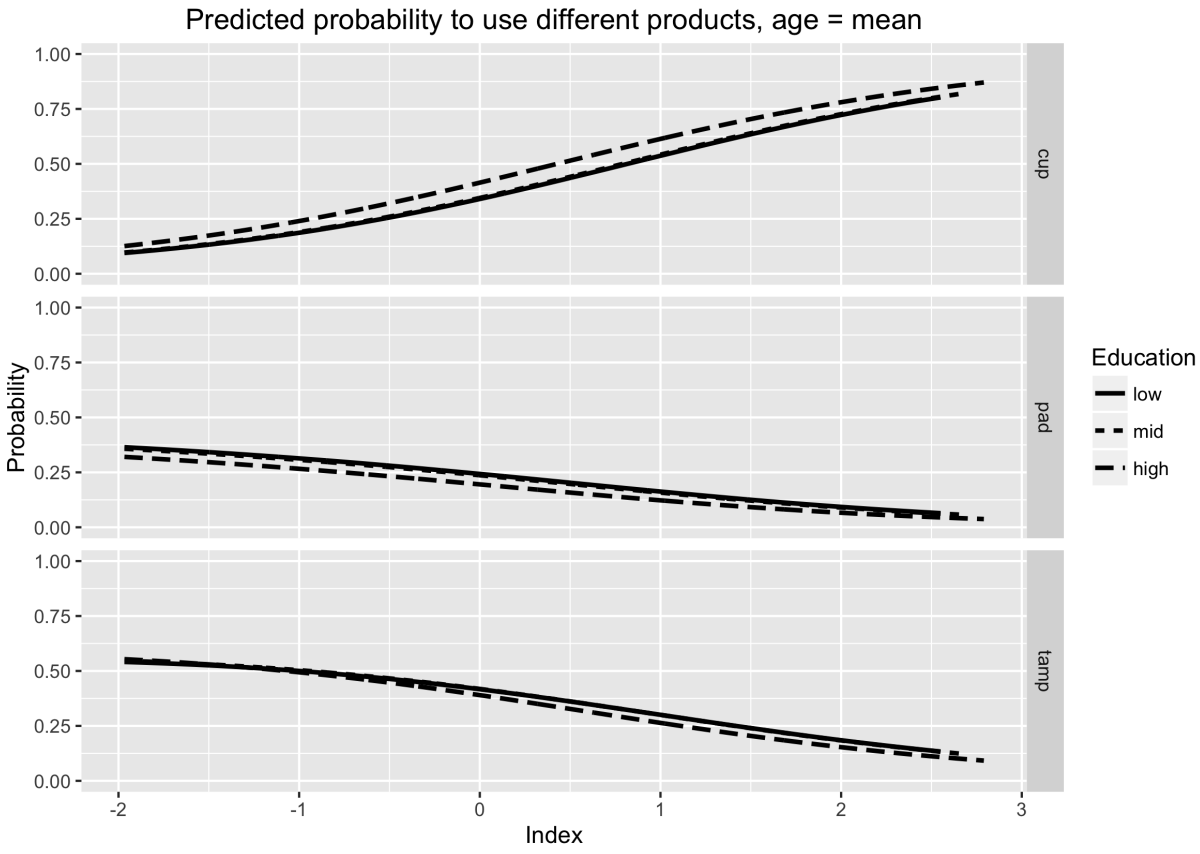
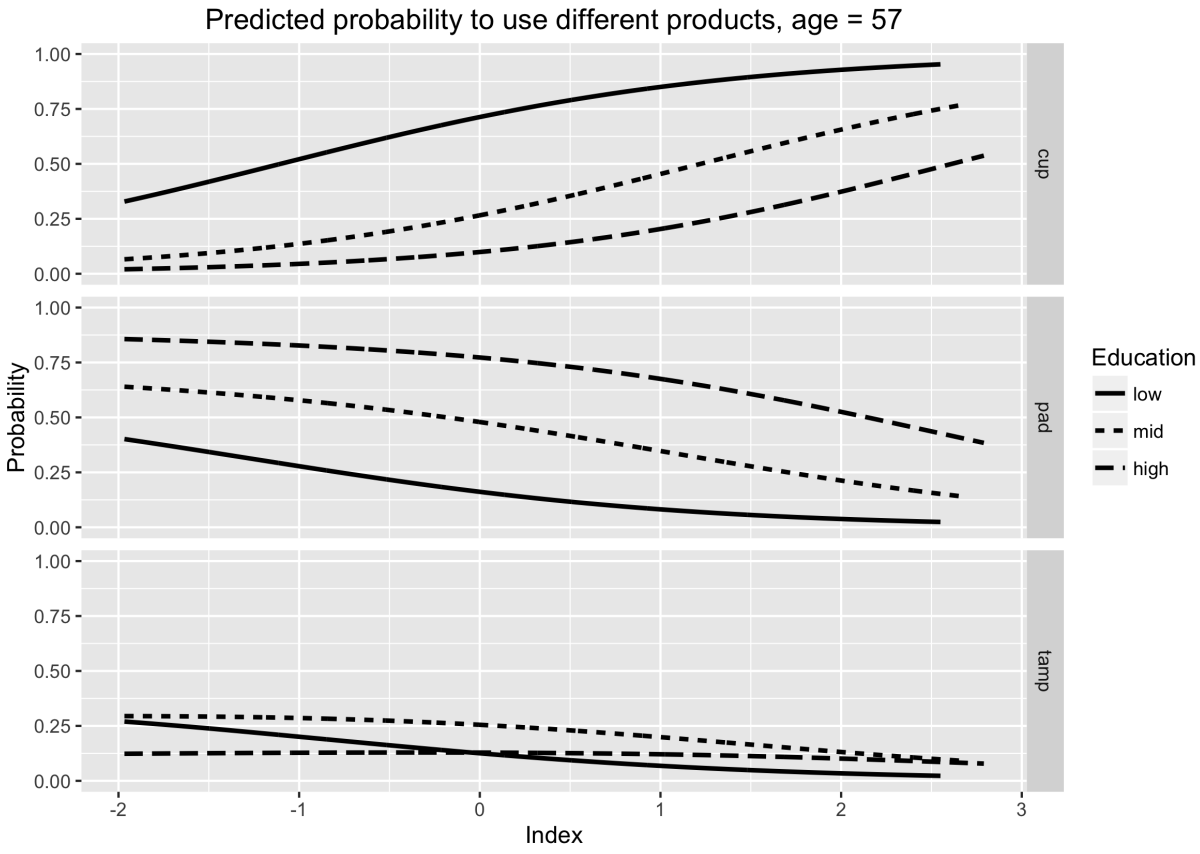


Figure 4.3: Multinomial model, Prediction plot 3



4.2.2 Binomial model

Since menstrual pad users and tampon users are highly confused when predicting (see table 4.4), a model with better prediction precision was created with menstrual pad and tampon in the same category. This model is a regular logistic regression model.

The difference from the model in table 4.6 is that menstrual cup now is measured against not using menstrual cup, which is equivalent to measuring it against menstrual pads and tampons together.

The model has the same variables as the multinomial model. The variables of the models are not chosen with the intent of having the same model. It just happened to be the best model in both cases (see section 4.1.2).

Table 4.10: Final binomial logistic regression model

	Dependent variable:		
	Menstrual cup (ref=pad/tamp)		
	β	e^{β}	e^{I_β}
Constant	-2.03^{***}	0.13	[0.07, 0.23]
	(0.29)		
Index	0.81^{***}	2.25	[2.06,2.47]
	(0.05)		
Age	0.05^{***}	1.05	[1.03, 1.08]
	(0.01)		
Education_mid	1.57^{***}	4.81	[1.93, 11.97]
	(0.47)		
Education_high	2.92^{***}	18.61	[7.39, 46.84]
	(0.47)		
Age*Education_mid	-0.06^{**}	0.94	[0.91, 0.98]
	(0.02)		
Age*Education_high	-0.10^{***}	0.90	[0.88, 0.94]
	(0.02)		
Akaike Inf. Crit.	3 291.56		
Bayesian Inf. Crit.	3 333.18		
McFadden R^2	0.11		
edf	7		
N	2827		
Note:	*p<0.05; **p<0.01; ***p<0.001		
	Standard errors in parenthesis		

The main variable, feeling towards menstruation (index) has a positive coefficient, which means that a higher score on the index means a higher probability to be a menstrual cup user. A negative score on the index gives a lower probability to use a menstrual cup.

The age variable has three coefficients which can be seen in table 4.11. This suggests that for a

Table 4.11: β-coefficients for age

	Education level		
	Low	Mid	High
β_{Age}	0.05	-0.01	-0.05

person with low education, higher age gives a higher probability to use a menstrual cup. Whereas for a person with high education, the effect is reversed with lower probability to use a menstrual cup when a person is older.

The education variable must also be considered together with the age variable and is presented with three different example ages in table 4.12. 13 and 57 are the end points of the age-scale, i.e. the oldest and the youngest persons to answer the survey.

Table 4.12: Examples of β-coefficients for education

	Age		
	13	26	57
β_{Mid}	0.79	0.01	-1.85
	13	29	57
β_{High}	1.62	0.02	-2.78

Age 26/29 chosen as it is the point where the coefficient switches sign

We see that for a person below age 26, a mid level education gives a higher probability to use a

menstrual cup than what the probability is for someone with low education. Whereas for a person above age 26, a mid education level lessens the probability in comparison with a person who has low education.

For person of age below 29, a high education gives a higher probability of using a menstrual cup. Whereas for a person above age 29, a high education gives a lower probability to use a menstrual cup.

As argued in section 4.2.1, these numbers should be interpreted with caution since there are no highly educated 13-year-olds.

Figure 4.4 shows how predicted probabilities to use menstrual cup changes with change on the index with age set to 13. Figure 4.5 shows the same but with age set to the mean age, and figure 4.6 shows this with age set to 57. For age = mean (25.85) and 57, the plots are divided by education level, but since there are no university educated 13-year-olds, the plot for them only has education level low. We see that for all age groups and educations, a higher score on the index gives a higher probability to use a menstrual cup.

Figure 4.4: Binomial model, Prediction plot 4

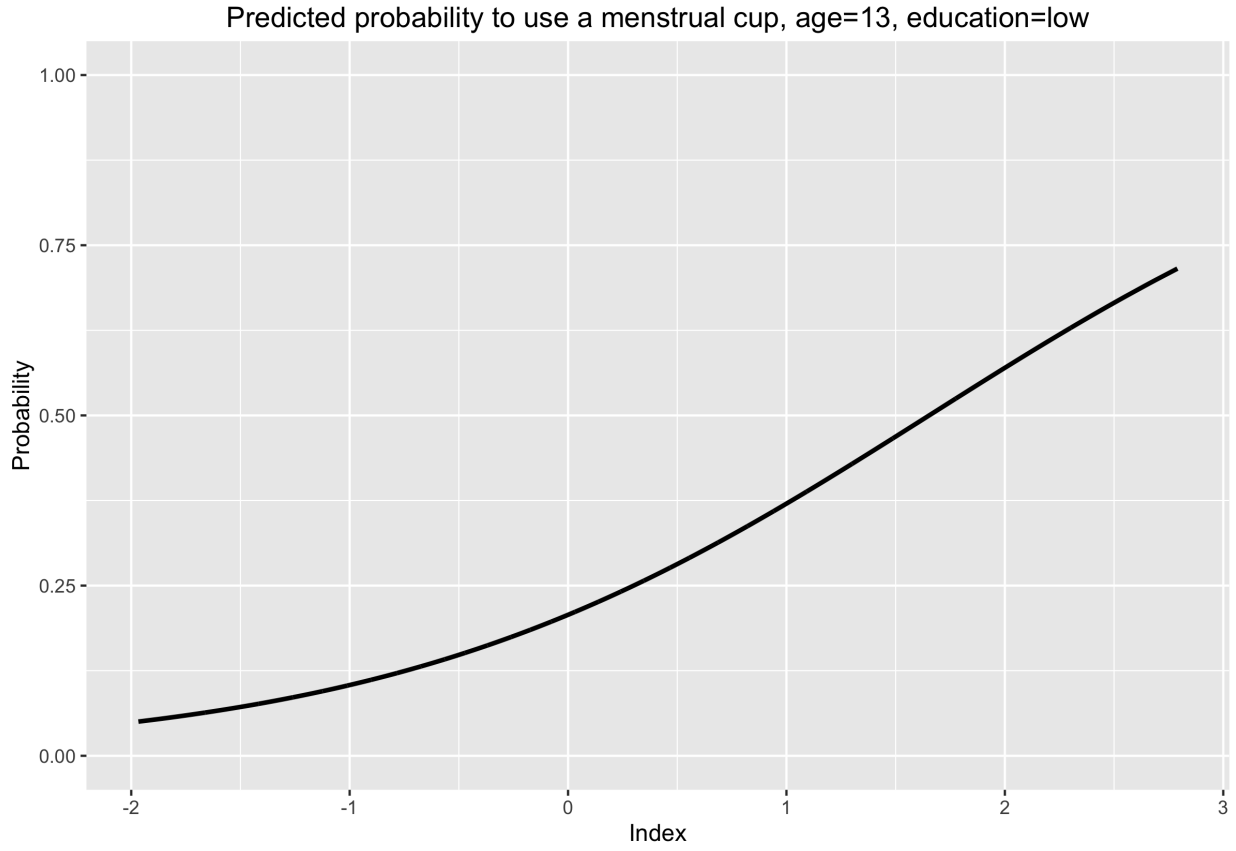


Figure 4.5: Binomial model, Prediction plot 5

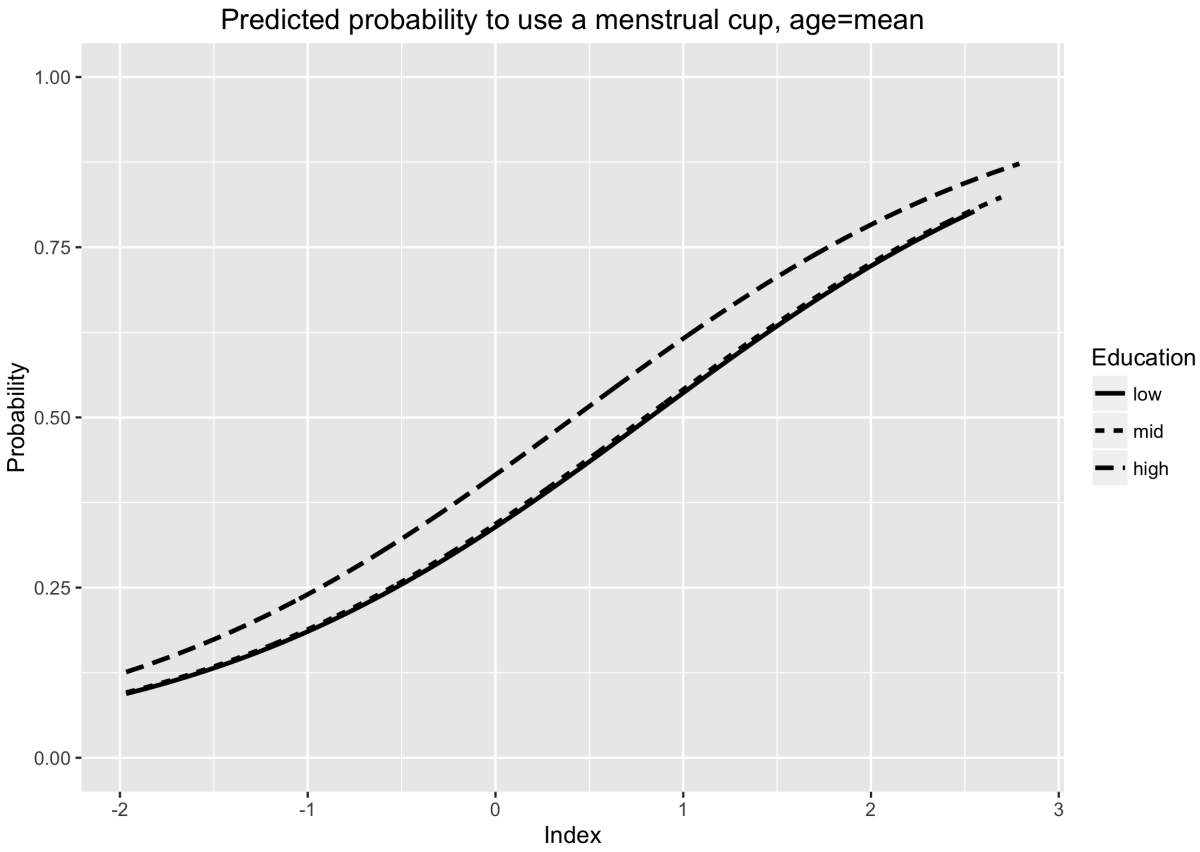
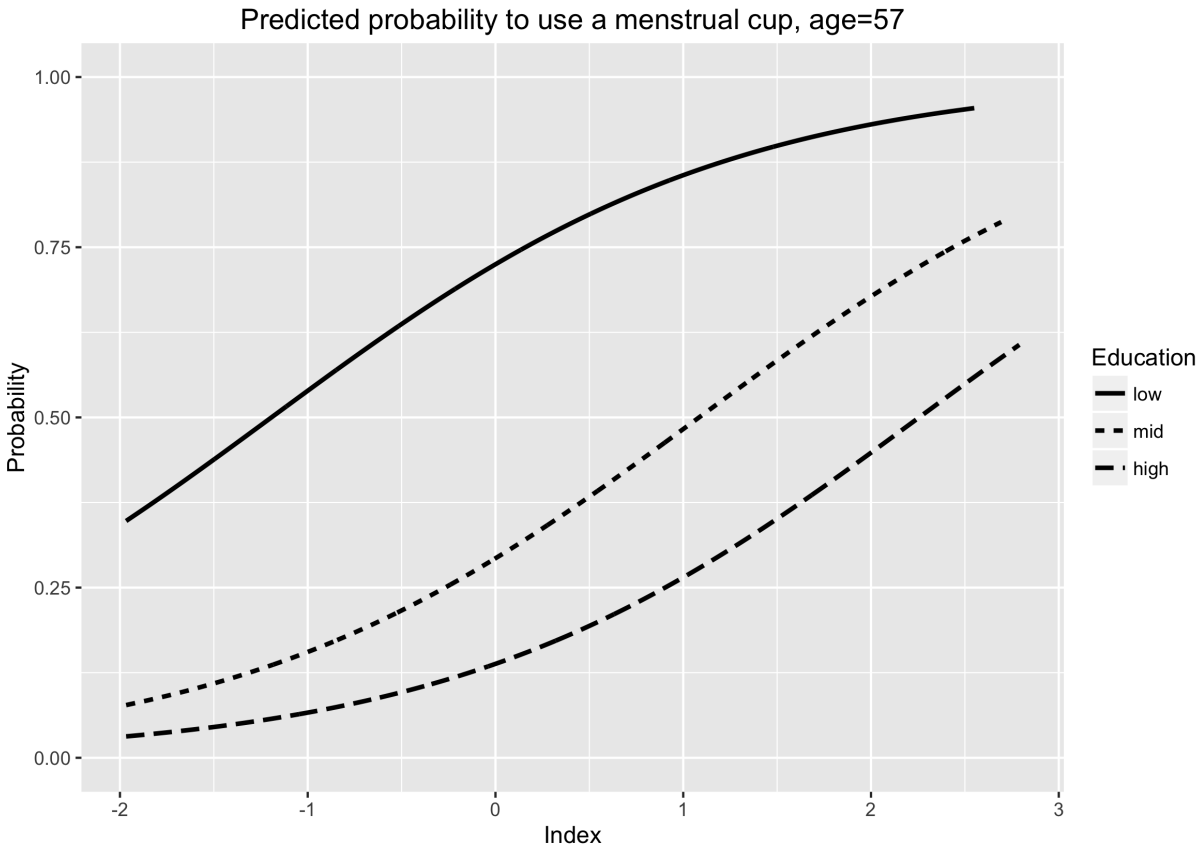


Figure 4.6: Binomial model, Prediction plot 6



Chapter 5

Discussion

Do menstrual pad, tampon and menstrual cup users differ in their attitude towards menstruation? In this thesis, to answer this, 16 different models have been tested to see which best fits the data from a survey about menstruation. The two best models were then presented more thoroughly and interpreted. In this chapter, the final chosen models will be analysed and criticised. And lastly some ideas about possible continuation of the project are presented.

5.1 Analysis

Any analysis of the results must bear in mind that the data does not represent the population in general. The collection of data is not representative due to snowball sampling as a method. For example the respondents had higher education level than than the population in general.

When predicting usage of products, menstrual pads and tampons were confused. The users of these two products seemed to be very similar in the variables used in this thesis. As a result of that they were in the end used as the same variable. Other researchers have shown that it is common to use a combination of menstrual pads and tampons [6, 13] and that could be one of the reasons as to why the users seemed to be so alike.

The feeling towards menstruation changes the probability to use a menstrual cup rather than a menstrual pad, with higher scores comes a higher probability to use a cup. The feeling also change the probability to use a menstrual cup rather than a menstrual pad and a tampon.

Education levels and age has an impact on the choice of products, where the effect is reversed in the older/younger population.

The model could predict which products a person uses, but not perfectly, with tampon usage being most correctly predicted.

As stated earlier, a menstrual cup user must be more comfortable with menstruation to use the product, or have to become more comfortable. The comfortableness could be a reason for the differences among the users. Whether the comfortableness comes before the choice or after cannot be answered by this model. This means that we can predict who already is a menstrual cup user, but we do not have enough knowledge to deduce who is a *potential* user since we do not know if the positiveness towards menstruation comes before or after the choice of product.

5.2 Critique

The amount of variables in the model is small. The model for example lacks variables concerning the characteristics of the respondents' menstruation. It could very well be the fact that the pain or amount of bleeding a person experiences has an impact on both choice of product and feeling towards menstruation.

Another critique is of course of the selection of the respondents. Since the group is a homogeneous one that probably found the survey due to their connection or likeness to each other, we do not know if the correlations found in this thesis exists in the main population, or if the effect is bigger in the whole population.

The sample has a very high percentage of menstrual cup users, which is probably a drastically higher percentage than the amount in the whole population. The problems with the sample can also be a part of the advantage of the sample; a representative sample could have too small a part made up of menstrual cup users. This could be solved in the sampling process, but still it makes the sample in this thesis interesting since we get to know more about menstrual cup users.

5.3 Further research

An obvious extension of this thesis is to do a similar project but with a more representative sample. To do that, a new collection of data would have to be made.

A new version of the project with more variables included, for example more characteristics of the menstruation would also further develop the findings of this thesis.

One final suggestion is to include a time factor where users of the menstrual cup are measured

both before they change the product and then again after a time of usage. This could show if the change in perception of menstruation is in the persons or in the usage.

5.4 Conclusion

It is possible to predict usage of menstruation products depending of feeling towards menstruation, but it is hard to distinguish between menstrual pad and tampon users. The menstrual cup users differ from the other two to a higher extent.

The choice of product is correlated with feeling towards menstruation, with menstrual cup users having a more positive feeling. Other variables also play a role, but it does not take away the effect of the index.

Bibliography

[1] S. Abraham, C. Knight, M. Mira, I. Fraser, D. McNeil, and D. Llewellyn-Jones. Menstrual protection. young women's knowledge, practice and attitudes. *Journal of Psychosomatic Obstetrics & Gynecology*, 4(4):229–236, 1985. URL: http://www.tandfonline.com/doi/abs/10.3109/01674828509016725.

[2] A. Agresti. *Categorical data analysis*. Wiley-Interscience, Hoboken, New Jersey, 2. ed. edition, 2002.

[3] A. Agresti. *An introduction to categorical data analysis*. Wiley, Hoboken, N. J., 2. ed. edition, 2007.

[4] S. Cheng and J. S. Long. Testing for iia in the multinomial logit model. *Sociological Methods & Research*, 35(4):583–600, 2007. URL: http://smr.sagepub.com/content/35/4/583.abstract.

[5] Y. Croissant. *mlogit: multinomial logit model*, 2013. R package version 0.2-4, URL: http://CRAN.R-project.org/package=mlogit.

[6] H. Cronjé and I. Kritzinger. Menstruation: symptoms, management and attitudes in university students. *International Journal of Gynecology and Obstetrics*, 35(2):147–150, 1991. URL: http://www.sciencedirect.com/science/article/pii/002072929190818P.

[7] J. K. Dow and J. W. Endersby. Multinomial probit and multinomial logit: a comparison of choice models for voting research. *Electoral Studies*, 23(1):107 – 122, 2004. URL: http://www.sciencedirect.com/science/article/pii/S0261379403000404.

[8] R. Grose and S. Grabe. Sociocultural attitudes surrounding menstruation and alternative menstrual products: The explanatory role of self-objectification. *Health Care for Women International*, 35(6):677–694, 2014. URL: http://www.tandfonline.com/doi/abs/10.1080/07399332.2014.888721.

[9] J. Helldén. *Blood and Disgust, A Survey about Menstruation.* Student project, Dept. of Political Science, University of Gothenburg, 2015. URL: http://johannahellden.se/wp-content/uploads/2015/01/Rapport.pdf.

[10] M. Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables.* Harvard University, Cambridge, USA, 2015. R package version 5.2, URL: http://CRAN.R-project.org/package=stargazer.

[11] C. Kwak and A. Clayton-Matthews. Multinomial logistic regression. *Nursing Research*, 51(6):404–410, 2002.

[12] T. F. Liao. *Interpreting probability models. : logit, probit and other generalized linear models.* Quantitative applications in the social sciences: 101. Thousand Oaks, [Calif.] ; London : SAGE, cop. 1994., 1994.

[13] H. Omar, S. Aggarwal, and K. Perkins. Tampon use in young women. *Journal of Pediatric and Adolescent Gynecology*, 11(3):143–146, 1998. URL: http://www.sciencedirect.com/science/article/pii/S1083318898701342.

[14] J. O. Rawlings, S. G. Pantula, and D. A. Dickey. *Applied regression analysis, a research tool.* Springer, New York, 2nd ed. edition, 1998.

[15] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[16] H. Wickham. *ggplot2: elegant graphics for data analysis.* Springer New York, 2009. URL: http://had.co.nz/ggplot2/book.

Appendix A

Included survey questions

A.1 Original Swedish version

F1 Vilket är ditt huvudsakliga mensskydd?

Om du använder olika mensskydd under din menstruation ska du svara det som du använder mest

- Bindor, engångsprodukt
- Tygbindor
- Tamponger
- Menskopp
- Annat:

F6 I vilken utsträckning stämmer följande påståenden in på dig?

Om något av påståendena inte gäller dig, t.ex om du inte har sex med andra personer eller aldrig badar, så kan du hoppa över den frågan.

(1: Stämmer helt - 7: Stämmer inte alls)

- Jag undviker att ha sex med andra personer när jag har mens
- När jag har mens känner jag mig mindre fräsch
- Jag tycker att det känns obehagligt att få mens på mina händer vid byte av mensskydd
- Jag tycker att mens luktar otrevligt

- Jag tycker att det är mer jobbigt att sova borta när jag har mens
- Jag tycker inte om att ha mens
- Jag skulle skämmas ifall jag läckte mensblod på mina kläder

F7 Hur gammal är du?

F13 Vilken skolutbildning har du?

Om du ännu inte avslutat din utbildning, markera den du genomgår för närvarande.

- Ej fullgjort grundskola (eller motsvarande obligatorisk skola)
- Grundskola (eller motsvarande obligatorisk skola)
- Studier vid gymnasium, folkhögskola (eller motsvarande)
- Examen från gymnasium, folkhögskola (eller motsvarande)
- Eftergymnasial utbildning, ej högskola/universitet
- Studier vid högskola/universitet
- Examen från högskola/universitet
- Examen från/studier vid forskarutbildning

A.2 Translated English version

F1 What is your main menstrual product?

If you use different products, answer the one that you use the most

- Menstrual pads, one use
- Cloth menstrual pad
- Tampons
- Menstrual cup
- Other:

F6 To which extent does these sentences apply to you?

If a sentence does not fit at all, i.e. if you do not have sex with other people or never go swimming, you can skip it.

(1: Applies completely - 7: Does not fit at all)

- I avoid having sex with others when I'm on my period
- I feel less fresh when I'm on my period
- I feel uncomfortable getting menstrual blood on my hands when changing a product
- I think menstruation blood smells unpleasant
- I think it is more of a problem to sleep over when I have my period
- I don't like having my period
- I would feel ashamed if i got menstrual blood on my clothes

F7 How old are you?

F13 What education do you have?

If you have not yet completed your education, choose the one you are currently undergoing.

- Not completed primary school (or equivalent)
- Primary school (or equivalent)
- Studies at gymnasium, folk high school
- Graduated from gymnasium, folk high school (or equivalent)
- Tertiary education, not university/university college
- Studies at university/university college
- Graduated from university/university college
- Graduated/Undergoing PhD

Appendix B

The Index

The index - feeling towards menstruation	Weight
Jag undviker att ha sex med andra när jag har mens	
I avoid having sex with others when I'm on my period	.570
När jag har mens känner jag mig mindre fräsch	
I feel less fresh when I'm on my period	.774
Jag tycker att det känns obehagligt att få mens på mina händer vid byte av mensskydd	
I feel uncomfortable getting menstrual blood on my hands when changing a product	.683
Jag tycker att mens luktar otrevligt	
I think menstruation blood smells unpleasant	.709
Jag tycker att det är mer jobbigt att sova borta när jag har mens	
I think it is more of a problem to sleep over when I have my period	.660
Jag tycker inte om att ha mens	
I don't like having my period	.620
Jag skulle skämmas ifall jag läckte mensblod på mina kläder	
I would feel ashamed if i got menstrual blood on my clothes	.607

Bachelor's Theses in Mathematical Sciences 2016:K1
ISSN 1654-6229

LUNFMS-4015-2016

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/