Jonathan Eng

Professor Kapelner

Math 3904

2020 April 24

What makes a Marriage Successful is not Understood, and Will Never be Understood

It is strange that for millions of years, humans have been successfully finding mates, yet even marriage counselors struggle to agree on specific requirements for a successful marriage. Majority of the advice is to improve yourself and increase communication with your partner, but what does that even mean? The advices given are blanket statements that are meaningless and open to various interpretations. If the steps needed for a successful marriage were known and understood, there'd be a list with specific steps that everyone would follow, and all marriages would be successful.

Why does marriage exist in the first place? In terms of natural selection, the current accepted theory of evolution where the main goal is to pass superior genes to the next generation; marriage, or social monogamy in the animal kingdom, appears to go against the grain. Social monogamy is viewed as a prolonged association and exclusive mating relationship between one male and one female. Approximately 85-90% of primates have the opposite, a polygamous relationship, where males mate with several females as it ensures the genes with the highest survival chance are passed on (Fuentes, 2008, p. 1). When females have a preference for a trait in males because it aids survival, (ex: large canines that increases success in fighting and hunting), it is best for the females to mate with the males who have the superior features. This way their offspring's will have the features that maximize their survival and reproduction chance (O'Donald, 1974, p. 2). Males without the desired feature(s) would eventually die without their genes passed on. Given enough generations, all males will exhibit the superior features, improving the survival rate of the overall community. Marriage is essentially a method of finding another person who mutually benefits from the sharing of resources and production of offspring.

Models are approximations of the true phenomenon. The two main goals of a model are to see if the model can predict a future event of the phenomenon under examination, and to show how a phenomenon works. Anything can be a model as long it has a measurable result. Models in general can be ambiguous, so creating a mathematical model with numerical definitions would be less ambiguous and therefore a more useful way to show the approximation of reality.

This paper models the whether a marriage will be successful based on specified variables. For this model, a successful marriage would simply be defined as the marriage not resulting in a divorce from the date of marriage to one of the spouse's deaths. The responses will either be a 1 or 0, indicating that the couple will be successfully married (1), or divorce (0).

The true outcome for a successful marriage based on data will be denoted, "*Y*" which is equal to either 0 or 1, indicating divorced or successfully married respectively. The true causal features that affect a successful marriage, denoted "z," revolve around uncountably infinite variables. For example, there is appearance, beliefs, personality, and background of an individual. Then we have preferences for each individual as well, such as a preference for age, race, religion, appearance, knowledge, occupation, salary, and so on. There are also biological reasons that are not yet understood behind why individuals fall in love, though we do know love plays a role in a successful marriage. Love itself can be further broken down into happiness, opportunity, chemicals, and more. Every true feature "*z*" can be narrowed to more values such as appearance being specified as height, facial symmetry, type of fashion, etc., however there are infinite ways to narrow down the broad "*z*" variables, making it impossible to account for everything. We are limited by knowledge on the subject itself. These true causes are denoted "$z_1, z_2, z_3, \ldots, z_t$", where "$t$" is the amount of true causal features. The true causes interact together in function "$t$", to obtain the measurements of a successful marriage where the interaction is also unknown. The overall formula for true reality is "$Y = t(z_1, z_2, z_3, \ldots, z_t)$."

Due to the impossibility of knowing the values of reality due to its infinite possibilities, the best we can do is create a model that approximates reality. The best guess we can make to measure a successful marriage is denoted "$\hat{y}$". The best approximation for the true causes of reality is denoted "$x_1, x_2, x_3, \ldots, x_p$" where "$p$" is the number of features thought to influence the output. The interaction of the approximated inputs is denoted "$f$". The overall formula for the model is "$y = f(x_1, x_2, x_3, \ldots, x_p) + \delta$," where "$\delta$" is the error due to ignorance since the true causal features are unknown.

Function "$f$" is not an analytical solution and is never truly realized, but rather it is estimated by learning from previous historical data. The data is used to come up with an estimate for function "$f$" in a process known as supervised learning. Supervised learning requires training data where both the inputs and outputs are known, denoted "$D$", and a set of functions used to determine the relationship between input "$x$" and output "$y$" denoted "$H$", and an algorithm denoted "$A$", that takes in data "$D$" and function set "$H$" to generate a model "$g$" such that "$g = A(D, H)$", where model "$g^*$" is the best candidate function that the algorithm can produce for the model. We are feeding both the input and output data to artificial intelligence, where the data is processed and tested onto multiple models until the algorithm finds the one model that gives the smallest error when tested against new data. We will use supervised learning by obtaining previous studies on what makes marriages successful and how impactful certain variables are on the marriage's success. This will also provide a suggestion for variables and act as a guide on how to break down the blanket statements offered in marriage studies into more measurable variables.

According to a census done by the Center for Disease Control (CDC) and National Vital Statistics System (NVSS), in 2018 there was about 2 million marriages and 800,000 divorces in the United States out of a sample population of 300 million people, or about 0.7% of the population is married and 0.3% of the population is divorced each year (CDC & NVSS, 2018, p. 1-2). For every 2 million people that are married, about 800,000 end up result in a divorce. Each year, more people are married than divorced, indicating the mode is married. It can be assumed with no variables, denoted "$p = 0$" where "$p$" is the number of variables being tested, that a marriage would be successful. This makes our null model, denoted "$g_0$", equal to the mode, "$g_0 = 1$", where 1 indicates successful marriage, and 0 would indicate a divorce. Our model must be better than the null model, otherwise we are better off assuming the mode. Before going more into our own theoretical model, we should look at data-driven modes first to guide our variable choice.

Happiness of a couple plays into whether or not the couple would divorce. Happiness is a broad variable, so it must become more measurable. Using data from the US General Social Survey (US GSS), Chapman and Guven analyzed self-reported data on the happiness of the husband and wife. There are 43,000 observations, denoted "$n$". Chapman and Guven took note of whether or not the couple owned a house or rented, employment status for the husband and wife, their education, age, and number of children. In this model, 8 variables are used in this model, making "$p = 8$". Each variable is denoted "$x_1, x_2, x_3, \ldots x_8$". Renting, age, and number of children has a negative correlation towards happiness. Owning a house, employment, and years of education all have a positive correlation towards happiness (Chapman & Guven, 2014, p. 544).

The model used has potential flaws. It may be built upon false or irrelevant data due to it being attained through a self-reporting survey where individuals may lie or have differing opinions on how impactful certain aspects of their lives affect the success of their marriage.

What one expects from a marriage plays a huge role in marriage as well. Between the 1900's to present day, different marital roles surfaced, mainly for women. Women gradually went from not working and being caretakers to working professional jobs, and finally, working while having to take care of children. Ross, Mirowsky, and Huber, attempted to model what factors would make those involved in a marriage become depressed, using variables attempting to explain gender roles in a relationship. There is a total of 680 wives and 680 husbands in the sample, resulting in a sample size of 1360 participants. Marriage is broken down into four types. In "Type 1", only the husband works because the husband and wife both believe the husband should work and the wife should stay at home to do housework and childcare. In "Type 2", the husband and wife work but both the spouses feel the husband should be work and the wife should stay home. In "Type 3", both the husband and wife work and feel like they both should work, however housework and childcare still remain the wife's duty. Lastly in "Type 4", both the husband and wife work and feel like they both should work and household labor is divided (Ross, Mirowsky, & Huber, 1983, p. 811). It is important to note that the husband is working in each type, only the opinion of who should be working alters. This creates four variables, wife's employment status, wife employment preference, husband's preference for wife's employment, and whether the husband splits the housework. They also included additional variables such as race (binary: white or non-white), age, religion (binary: Jewish or non-Jewish), education, family income, and number of children. Currently there are 10 variables, "$p = 10$", in this model. This model runs into the same problems as the previous study, where the information is obtained through verbal surveys. The participations are subject to lying and differing weights for how impactful certain aspects of their life are towards the success of their marriage.

The first regression run with this model determines the chance that the wife is depressed. White (race), age, non-Jewish, education, family income, number of children, husband helping at home, wife's employment preference, and wife's employment status are all inversely correlated to depression. Jewish, and employment, and husband's preference for wife's employment, are positively corelated to depression. The second regression determines the chance the husband is depressed. White, age, non-Jewish, education, family income, number of children, husband helping at home, husband's preference for the wife to work, and (wife's employment * husband's preference for wife to work) are all inversely correlated to depression. Jewish, and wife's employment are all positively correlated to depression (Ross, Mirowsky, & Huber, 1983, p. 818). The data is nearly parallel between the causes of depression for both the husband and wife, where age, non-Jewish, education, family income, number of children, husband helping at home, and matching employment status to employment preference all were negatively correlated to depression, or positively correlated to happiness.

Comparing the model on depression done by Ross, Mirowsky, and Huber to the model on happiness by Chapman and Guven, we can negate the data on depression, so it becomes in terms of happiness to be easier to compare. Interestingly, age and number of children are positively correlated to happiness in Ross, Mirowsky, and Huber's model, but negatively correlated in the model done by Chapman and Guven. This is because of the vagueness of happiness leaves lots of influential variables unaccounted for. Leaving out essential variables can affect the weights of existing variables and add more noise in the model.

To create our own model to determine the whether a marriage would be successful, we will use the unique variables from the studies, currently having "$p = 16$". Some of the categorical values were altered to become more specific, such as race going beyond white and non-white.

$X$ = {house owned (binary: Owned (1) | Not Owned (0)),

employment status husband (binary: employed (1) | unemployed (0)),

employment status wife (binary),

husband's employment preference for wife (binary: work (1) | should not work (0)),

wife's employment preference (binary),

husband helps at home (binary: helps (1) | does not help (0)),

education level husband (categorical: less than high school, high school, some college, Associates, Bachelors, Masters, Doctoral),

education level wife (categorical),

race of husband (categorical: White, Black, Asian, American Indian, Pacific Islander),

race of wife (categorical),

religion of husband (categorical: Catholic, Jewish, Muslim, Hindu, Buddhist, other, atheist),

religion of wife(categorical),

family income (numeric: In thousands of dollars),

age of husband (numeric),

age of wife (numeric),

number of children (numeric)}

$Y$ = {binary: 1 Married | 0 Divorced}; where $Y$ = whether or not a couple will divorce before the death of a spouse.

With infinite possibilities to define what makes a marriage successful, "$p = 16$" appears small. To add more variables, we can look towards the primitive model for marriage to see what variables may be missing. The previous models did not include physical influences on the marriage such as appearance. Though it cannot be assumed that people are happier when married to more attractive people, part of our model is influenced by primitive behavior, therefore attractiveness of the individual, or sexual selection in the animal kingdom, will be included as variables. Attractiveness is another vague variable that must be broken down.

Males typically had to protect their mates and offspring from predators and other male competition. By doing so, only the strongest males would be able to reproduce, as they would monopolize access to most females. Males would often vary in their ability to provide for the females, so females were expected to compete for the male with the most benefits. Males would often be attracted to females with youthful appearances due to its indication of fertility and likely the innate desire to protect those with infant-like features. Female traits tend to lead towards expressing fertility and submissiveness, such as high-pitched voices, reduced facial hair, and slender faces. Male traits tend to lean towards ornaments or weapons, such as deep voice, large and muscular body size, and facial hair (Puts, 2010, p. 168-169). This study adds 8 features to our model such as,

$x_i$ = {male voice pitch (categorical: high, medium, low),

female voice pitch (categorical),

female fertility (binary: fertile (1) | not fertile (0)),

male fertility (binary),

female facial hair (binary: hair (1) | no hair (0)),

male facial hair (binary),

male height (numeric: in inches),

male mass (numeric: in kilograms)}

Now with "*p = 24*" is it likely the model is not close to approximating the true causal variables. The error due to ignorance will be high because of the difficultly of measuring the factors that influence the success of a marriage. The "*x's*" are not reasonably measured because many of the variables are opinionated. For example, a husband may feel his voice is high and state high on his survey, however the wife may feel the voice is medium. Both are targeting the same variable yet yielding different results making the data inconsistent. It could also be as simple as spouses not wanting to state their correct age or wanting to exaggerate the success of their marriage. Another way to receive data is through a government census which contains most of the information used to measure marriage success. This is likely more truthful due to the authority of the government, however there will be some missing data that can only be achieved through surveys. Opinionated self-reported data will never be reliable and puts the model at a large disadvantage due to having no reasonable way to obtain accurate "*x*" variables. Our "*x's*" do not come close to estimating the true causal "z" variables because there are plenty of unaccounted for variables and lost in the noise.

With "*p=26*" variables in the model, many of which are categorical, we need a large sample size to reduce estimation error, avoid overfitting the model, and decrease the likelihood of having bias data. The sample must be large enough to have sufficient information for each level of the categories, each binary value, and a large range for the continuous numeric variables. Because most of the data can be obtained from government census forms, we can easily obtain a large amount a data, typically in the millions. The data available through the forms is also organized by states and countries in the event we choose to only focus in on a single population, which will be beneficial to avoid cultural conflicts affecting the data. For example, New York may value income much more than Florida. Having both states in the sample data can potentially cancel each other out and provide the average income weight on successful marriage that represents neither state. Obtaining data though government data is the best method for modeling a successful marriage because it gives a much larger sample size compared to obtaining our own data and it will be more truthful. However, additional data can be obtained through random surveys with online forms such as google forms or survey monkey. In person data collection is another method, but it risks having bias data unless the data is collected from multiple locations with different populations. Obtaining a large sample size using this method can be expensive, difficult, and less reliable than government data.

With a sample size in the millions, finding new data that is not within the range of the historical data, or extrapolations, is rare, but still possible. Most new data would be within the data set, or interpolations. The model will be able to perform better when predicting on interpolations as to extrapolations, so it is beneficial to have such a large sample size.

Comparing our model to the other models used in the mentioned studies, this model will perform better. Other studies tend to focus in on one specific "*x*" and attempt to explain its effect on marriage success, such as an entire model on happiness, which is broken down to family size, employment beliefs, etc. This model incorporates multiple studies to include the specifics of multiple "*x*" values. Our model includes a break down for happiness, finances, and appearance. Combing the studies of others allows for a more accurate estimation, displaying features that were hidden in the noise of other studies.

We can use the ordinary least squares ("OLS") algorithm to run our model, where if the output is below a certain threshold, it is predicted the couple would have a divorce "$\hat{y} = 0$" and if the output is above the threshold it is predicted the couple would have a successful marriage "$\hat{y} = 1$." The function for a binary "OLS" model is "$\hat{y} = \{\mathbf{1}_{b0 + b1x1+b2x2+....bpxp > 0}\}$" where "$b_0$" is the intercept, or initial value, and "$b_{1...p}$" are the weights of each variable to indicate how influential they are towards the output. The indicator function "$\mathbf{1}_{b0 + b1x1+b2x2+....bpxp > 0}$" returns a 1 if the function is true, otherwise it returns a 0, representing successfully married and divorced respectively. An "OLS" model is a linear modeling strategy that aims to minimize the sum of the squares of the difference between the observed and predicted values. A reason for choosing "OLS" over other model types is because "OLS" will converge to the true population output as the sample size increases.

Our model contains categorical variables represented by words as to numeric values, therefore they cannot be numerically computed. To make it computable, categorical variables would become dummy variables. For example, "race" (categorical: White, Black, Asian, American Indian, Pacific Islander), would become dummy variables, each having their own respective column, where a 1 would be present in the column of the individuals race, and a 0 in the remaining race columns. By making the values 1 and 0, we now have computable numeric values for calculations. One dummy variable is left out to be used for a reference and its value is held in the intercept "$b_0$". The resulting coefficients for the dummy variables would represent their relative relation to the reference variable.

| | | | Dummy Variables | | | |
|---|---|---|---|---|---|---|
| Person | Race | | Black | Asian | American Indian | Pacific Islander |
| $X_1$ | White | | 0 | 0 | 0 | 0 |

To test the error of our model, we can measure the misclassification and accuracy of the model. Misclassification error can be estimated using the amount of incorrect predictions divided by the total amount of predictions denoted "$misclassification\ error = \frac{\#\ of\ incorrect\ predictions}{total\ \#\ of\ predictions}$," where misclassification is a percentage between 0 and 1 where a better model yields a smaller misclassification error. Misclassification error will increase if the model predicts a successful marriage and the couple ends up in a divorce, or if the model predicts a divorce and the couple stays married, otherwise misclassification will decrease. Accuracy would be the amount of correct predictions divided by the total amount of predictions denoted "$accuracy = \frac{\#\ of\ correct\ predictions}{total\ \#\ of\ predictions}$," where accuracy is a percentage between 0 and 1 where a better model yields a higher accuracy. Accuracy will increase if the model predicts a successful marriage and the couple stays married, or if the models predicts divorce and the couple divorces, otherwise accuracy will decrease. Though, if the values are approximately to 0 and 1 respectively, there is likely an issue with the data being tested ("$D_{test}$"), as reality is filled with noise that will cause some predictions to be wrong. The test data may be relatively small, bias, or tampered with.

Historical data is used to train the model. Currently, the model is simple, having only a degree one polynomial function. The functions for the model, contained in set "*H*", can be expanded to include polynomial regressions to the "$n^{th}$" degree, because it is likely the data is not linear, however increasing the degree comes with rewards and risks. Increasing the degree can fit more functions onto the model and avoid an underfit model, but also risks overfitting the model onto noise, eventually yielding an error of 0.

Our model is more subject to underfitting as to overfitting because the true causal values are more complex and interactive with one another, requiring a higher degree polynomial to have a well fit model. The model can be increased to degree two polynomial complexity, as it is likely some variables will have a diminishing and eventually negative correlation to a marriage's success in the future, such as the number of kids. Although having children is seen to increase the success in a marriage, there is a lack of data for families with a large number of kids, indicating it may have an inverse relationship to a successful marriage after a certain amount of kids, so society is taught to avoid such numbers. We can also increase complexity by multiply related variables, such as "*husband's preference for wife to work"* and "*wife's employment*." Income can also be altered to log of income to decrease the effectiveness of higher incomes. This will avoid cases that have a high enough income that counteract all the negatively correlated variables that affect a successful marriage. With different ways to model the data, how can we choose which model is the "best?"

There are infinite models with differing complexities and having to choose the best one introduces a fundamental problem in machine learning. We can test every model and choose the one with the lowest standard error because it would be the most precise, but that does not necessarily mean "best" and there is also the issue of overfitting the model on the data. Overfitting data would mean the model is too tailored to the training data, causing inaccurate future predictions because the training data, "$D_{train}$", does not properly represent reality.

To ensure an honest validation on the model we partition part of the training data "$D_{train}$" and call it "$D_{test}$". The purpose of separating part of the data is so that we can have an honest validation by using data not involved in the training, called out of sample data, which is essentially new data. The unremoved portion of "$D_{train}$" will be used to make a model and prediction, in hopes of having a close estimate to the out of sample data. The out of sample error between "$D_{test}$" and the predicted values from the model would explain the model's performance. This does cross validation, however, suffers from a large variance because the validation could drastically change depending on which data values are put into “$D_{test}$”. Additionally, removing data from “$D_{train}$” leaves the data more vulnerable to underfitting because “$n$” is decreasing in “$D_{test}$”.

Cross validation using the “K-fold method” is similar but solves the issue of high variation. “$D_{train}$” is divided into “$K$” subsets then the simple validation method is performed where one of the “$K$” subsets are “$D_{test}$” and the remaining data becomes “$D_{train}$”. A model “g” is fit onto “$D_{train}$” and the out of sample error is recorded for each “$g$” from “$g_1$” to “$g_K$”. The error between each model is averaged to validate the accuracy of the model. This method reduces potential bias from using one sided data in “$D_{test}$” and reduces variance if a reasonable “$K$” is chosen. This model is still vulnerable to overfitting because the same data is repeatedly used to train the model, resulting in a dishonest validation.

When creating a model, we always assume that the relationship between the phenomenon we are trying to predict and the causal inputs are stationary, defined as never changing regardless of the time frame. Non-stationary data is unpredictable and cannot be modeled through supervised machine learning. If the relationship does change, the model created is no longer valid. The means of a successful marriage is definitely not stationary, however does takes a long time to change. For example, the idea of what role a husband and wife should play has altered over the past century. The role for each spouse has become so important that the personal belief towards roles is included as a variable in this model. Additionally, more extrapolations will appear in the data set as new trends are introduced into society. Given enough time, it is possible for people to alter in preference completely and the model will become invalid.

The model itself is inaccurate due to the difficulty of estimating marriage success due to its opinionated nature. Due to the inaccuracy the model is likely not concrete enough to be useful with high confidence. Additionally, because the model is not stationary in the long run, this model is only good as long as societal values remain unchanged. If society changes, the model must too.

Marriage has become so solidified into the modern world that we should definitely investigate what makes a marriage successful. As seen through the process of creating a model to measure the what goes into a successful marriage such as happiness of the spouses, are too opinionated and broad to be accurately measured. As the broad statements are broken down into more specific and measurable variables, there will always be more immeasurable variables around the corner because all these variables revolve around feelings. Feelings are not and will not be understood. Emotions stem from the lower subcortical systems, along with other instinctive behavior. In order for emotions to be expressed through language, it must interact with higher brain processes (Panksepp, 2011, p. 1797). This indicates emotions cannot be expressed directly through higher brain processes such as language, and therefore all metrics of self-surveying emotion would be false data. Physical studies on the brain are near impossible as well, due to the ethical issues it may cause (Panksepp, 2011, p. 1798). Due to the inability to thoroughly study the relation between human emotion and the brain, feelings such as happiness, cannot be truly understood. There are infinite variables to account for emotions, and our ignorance towards the topic makes it impossible to approximate most of the correct casual inputs.  What makes a marriage successful is impossible to understand because a successful marriage is built on emotions that cannot be directly studied or measured.

# Bibliography

CDC, and NCHS National Vital Statistics System. "Provisional Number of Marriages and Marriage Rate: United States 2000-2018." (2018, January 13)

Chapman, Bruce, and Cahit Guven. "Revisiting the Relationship Between Marriage and Wellbeing: Does Marriage Quality Matter?" Journal of Happiness Studies, vol. 17, no. 2, 2014, pp. 533–551.

Fuentes, Agustin. "Re-Evaluating Primate Monogamy." American Anthropologist, vol. 100, no. 4, 1998, pp. 890–907.

O'Donald, P. (1974). Polymorphisms maintained by sexual selection in monogamous species of birds. Heredity, 32(1), 1–10.

Panksepp, Jaak. "The Basic Emotional Circuits of Mammalian Brains: Do Animals Have Affective Lives?" Neuroscience & Biobehavioral Reviews, vol. 35, no. 9, 2011, pp. 1791–1804.

Puts, David A. "Beauty and the Beast: Mechanisms of Sexual Selection in Humans." Evolution and Human Behavior, vol. 31, no. 3, 2010, pp. 157–175.

Ross, Catherine E., et al. "Dividing Work, Sharing Work, and In-Between: Marriage Patterns and Depression." American Sociological Review, vol. 48, no. 6, 1983, p. 809.