Statistics 139 Final Project

St. Valentine’s Front-Office - An Analysis of HCS Datamatch 2015

*Frederick Widjaja, Kevin Eskici, Willy Xiao*

To view this file and any code, visit: <http://github.com/willyxiao/stat139-datamatch>

1. **A Memo from the Front-Office Staff**

As members of HCS board past and present, we reserve a special place in our heart-of-hearts for HCS’s biggest event of the year: Datamatch. Exactly one week before Valentine’s Day we release a 30-question survey ranging from sex to favorite classes at Harvard to more sex in an attempt to pair 3,000 Harvard undergraduates in a flurry of last-minute Valentine’s day love-finding, the kind of deep romantic-type love you only see in Disney animations. While we may love our matching-algorithm like any mother loves an ugly child, we understand that using the results of TheAlgorithm[[1]](#footnote-1) itself as some Y variable probably won’t give us any meaningful insights into this complex symphony of human emotion. Luckily, a closer predictor of love emerges a few days after St. Valentine’s has done his deed – whether or not top-paired Datamatch couples agree to go on an HCS sponsored Waffle Date at Zinneken’s.

As witnesses to the game of love, bystanders will only see the whistling passage of Cupid’s arrow and the sparks of passion that come thereafter. But long after St. Valentine’s has retired to the locker-room, we – the metaphorical front-office staff – are still working hard pouring over the data underlying every match. Our singular purpose is to maximize matching-making efficiency between Harvard undergraduates, and we’ll do it by discovering which people are most likely to go on a Waffle Date.

We are the sabermetrics of match-making. We the latent factor underlying love. We are St. Valentine’s Front-Office.

1. **Collecting Data**

Because we are also the ones who own Datamatch, data is easy to come-by. Of course, while the data is somewhat structured in a relational sql database, a lot still must be done to [collect collect collect, sql queries monkeys]

1. **Modelling the arrow’s path**

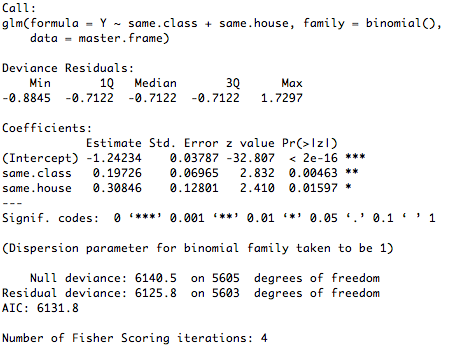
We first approached the problem by first considering model which individually could craft a narrative for the button clicks. In these steps, we did little assumption-checking or cross-validation; the only goal was to throw arrows at a target and see which couples stuck.

1. **Pair-Wise Models**

For pair-wise models, we considered whether or not a data-matchee would click a button based on their interaction with who their matched with. If two people are in the same house or the same class year, does that make a difference?

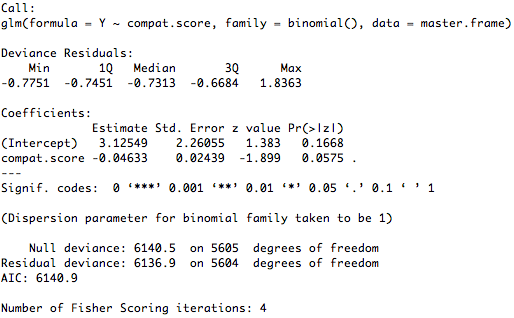
1. **Proximity** [[2]](#footnote-2)[Class Year, House]

Proximity is one of the best predictors of love[[3]](#footnote-3). Within the Harvard bubble, quadlings date a disproportionate number of quadlings, Kirkland is incestuous, and Mather lathers in their troves of singles.



1. **The Power of Suggestion** [[4]](#footnote-4)[Suggested Compatibility]

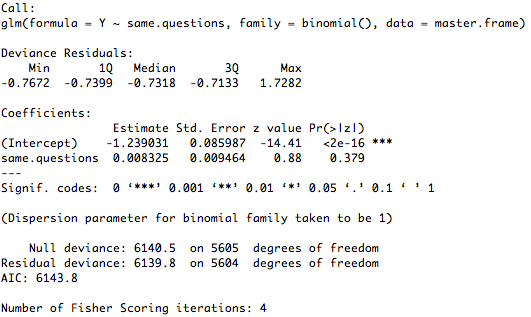
If you’re told you’d be compatible with someone, does of power of suggestion compel you to try it out?



Probably not. Lol.

1. **The Underlying Truth** [[5]](#footnote-5)[Matching Answers]

If there exists any underlying, latent truth in the questions that HCS asks, then we ought to be able to see it here:

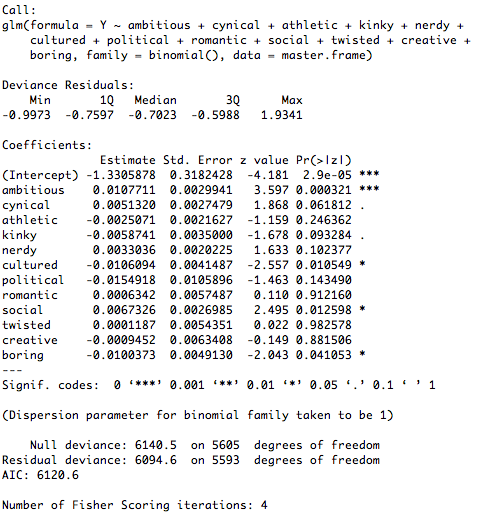


1. **Individual-Based Models**

For these set of models, we ignored the interaction between partners, but rather tried to guess whether someone would agree to a waffle date based on their own inherent traits.

1. **Personality** [[6]](#footnote-6)[As Determined by TheAlgorithm]

TheAlgorithm calculates 12 personality traits from the answers people give to us. Herein lies the magic of TheAlg and here, we hope to find something:

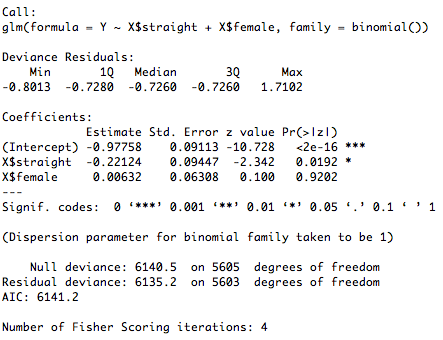


A satisfying result emerges from this model: Ambitious people are statistically significantly more likely to go on a waffle date! Wow, good for them! And…as validation to our data-match questions, good for us!

1. **Sexuality**[[7]](#footnote-7)[Sexuality, Gender]

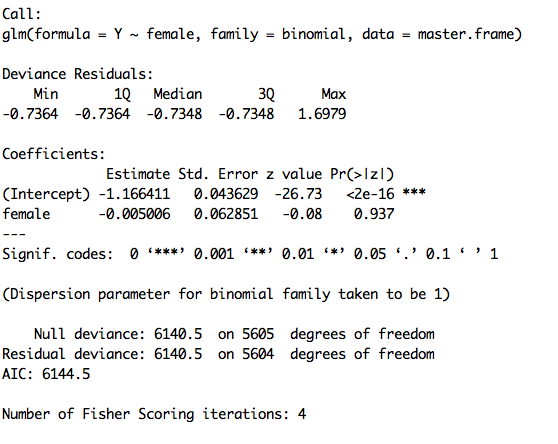
Sexuality so often comes to statisticians as clear binaries and easily-discernable (read: discriminatory) categories. While we may condemn this simplistic classification in our public discourse, as statisticians, we laud it. For its simplicity and ease of implementation, we did this first.

Here is the simple output from R:



What have we found? Significance, the arbitrary p-value kind! In the variable called *straight*. Interpreting the variable gives us a conclusion that we thought might’ve been true given our daily perceptions: if you’re not straight, you’re much more likely to agree to a Waffle Date.

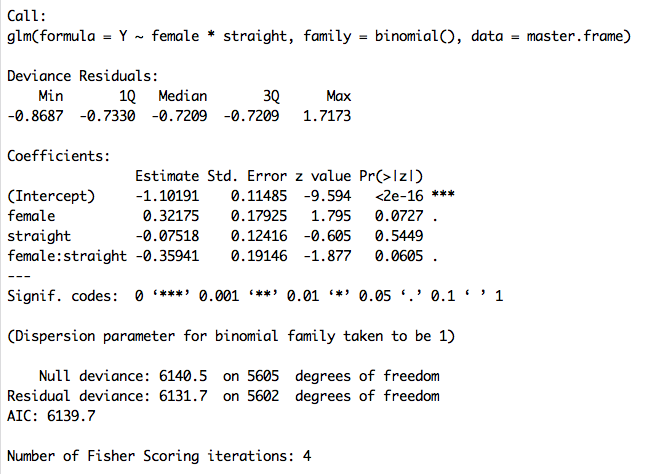
We also find something interesting in the interaction terms. Modelling our data just with a female variable, we see:



Which corresponds to the following probabilities of saying yes to a waffle date, and we see that non-females are slightly more likely than females to say yes (though not by much and not to a significant amount).

|  |  |
| --- | --- |
| Non-Female | Female |
| 0.2375043 | 0.236599 |

However, when we add in the interaction term with female and straight, we see this:



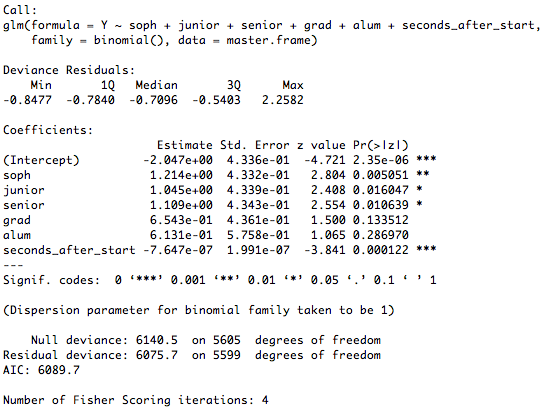
Which corresponds to:

|  |  |  |
| --- | --- | --- |
|  | Non-Female | Female |
| Straight | .23558 | .22886 |
| Not-Straight | .24938 | .31428 |

This is an interesting scenario in which the interaction terms actually matter. A non-straight female is the most likely out of these four groups to agree to a waffle date even though females overall are less likely to agree to one.

1. **Latent Eagerness** [[8]](#footnote-8)[Class Year, Time Responded to Survey]

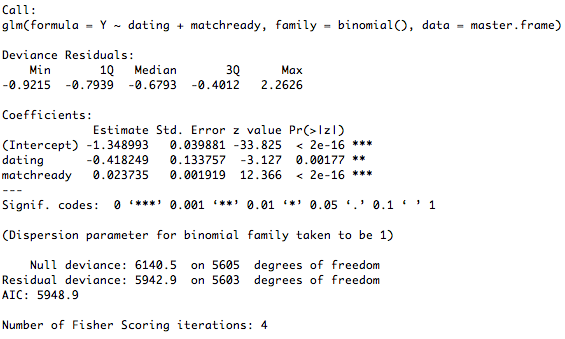
We know everyone wants to be seen as the one who cares less,[[9]](#footnote-9) but we know how much you’ve actually been waiting for Datamatch.



If you’re someone who filled out Datamatch early-on, then you’re quite more likely to go on a waffle date.

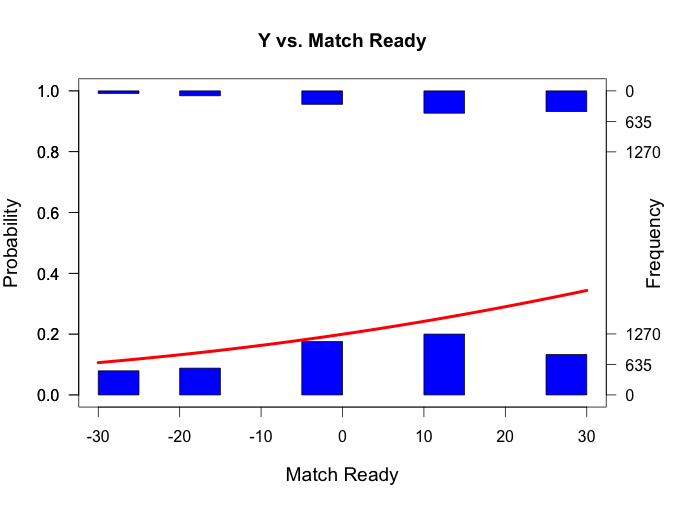
1. **Explicit Eagerness** [[10]](#footnote-10)[Prior Willingness, In a Relationship]

And finally, we can question how eager you are to go on a waffle-date given your explicit eagerness.



In many ways, this model by itself should give us a lot of information; we literally ask the question of how likely are you to go on a date with your data-match before people fill out the survey, this should give us some information. In fact, this ought to be the best predictor that we have.

Isolating the match-ready variable, we can see the correlation between match-ready and yes to waffle-dates:



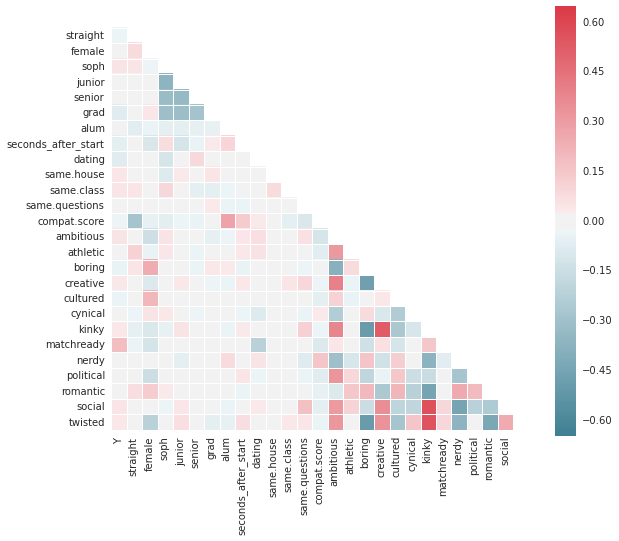
1. **Love, what is the real thing?**

Now that we’ve played around with all of the individual models, we can go about constructing the best model for Cupid’s use.

1. **Model Assumptions**

A few model-assumptions ought to be made for logistic regression:

1. Independence of Y variable. We know for a fact that this assumption is broken. This is because in constructing our data the same person can exist multiple times in our data-set. In many instances, there may even be a negative correlation between individuals.
2. Independence of X variables. Plotting the correlations between each of these x-variables we get the following graph:



A couple of interesting factors appear here. First, we see a block of highly-negatively correlated variables between soph, junior, senior, grad, and alum. This makes sense because if someone’s a sophomore, they can’t also be any of the other class years.

On the other hand, we have highly positively correlated variables between social, kinky, twisted, creative. This makes sense because TheAlgorithm fundamentally correlates these personality traits when creating the personality scores. Thus, when we began looking for Cupid’s final model, we ignored the variables: kinky, twisted, and creative.

1. The log-odds is linear with respect to the X variables. This is very hard to check because of the number of X variables that we have. We can also probably assume that this is NOT true at some point in our data. This is because unlike the CLT leading to normality or the law of large numbers leading to a poisson, nothing fundamental in statistics suggests that a variable ought to be linear with respect to the log-odds of another.
2. **In search of the best model[[11]](#footnote-11).**

We took three different but related approaches to producing our master model. These are described as model.best, model.base.BIC, and model.base.AIC. The process for each is as follows:

model.best:

model.base.BIC

model.base.AIC

We get the following sets of significant variables in each:

|  |  |  |
| --- | --- | --- |
| **Model** | **Variables Selected** | **CV-Score[[12]](#footnote-12)** |
| **Null Model** | ­-- | .3613 |
| **Best Model** |  | .3339 |
| **Base BIC** |  | .3439 |
| **Base AIC** |  | .3369 |

From this table, we can see that model.best produces the model with the most number of predictors. In cross-validation tests, it also seems to be the model that performs the best. Yet the base model of the BIC gives us the most easily interpretable results. Either way throughout all of the chosen models, we see a couple of variables that consistently appear to be significant:

straight

female

soph

junior

senior

seconds\_after\_start

dating

same.house

ambitious

matchready

Looking at our best-model,

Ignoring our best model for a moment, we can also take a look at the performance thresholds the AIC and the BIC:

|  |  |
| --- | --- |
|  |  |

­[Fred write something here]

1. The data on data-match: http://www.thecrimson.com/article/2014/2/19/the-data-on-datamatch/ [↑](#footnote-ref-1)
2. See model\_proximity.R [↑](#footnote-ref-2)
3. http://www.npr.org/templates/story/story.php?storyId=112330125 [↑](#footnote-ref-3)
4. See model\_suggestion.R [↑](#footnote-ref-4)
5. See model\_truth.R [↑](#footnote-ref-5)
6. See model\_personality.R [↑](#footnote-ref-6)
7. See model\_sexuality.R [↑](#footnote-ref-7)
8. See model\_latent\_eagerness.R [↑](#footnote-ref-8)
9. http://www.cosmopolitan.com/sex-love/advice/a5585/college-dating-screwed-up/ [↑](#footnote-ref-9)
10. See model\_explicit\_eagerness.R [↑](#footnote-ref-10)
11. See build\_model\_best.R [↑](#footnote-ref-11)
12. CV-Score found in cross\_val.R. This score is the RMSE of the predicted probability and the true values with a training set of size 4000 and done 100 times. The minimum of the RMSE ought to be the model with the best prediction. [↑](#footnote-ref-12)