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.

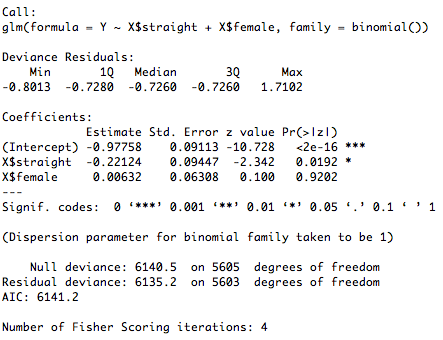
1. **Collecting Data**

Because we are also the ones who Data-Match, [collect collect collect, sql queries monkeys]

1. **Modelling the arrow’s path**
2. **Pair-Wise Models**
3. **Proximity** [Class Year, House]
4. **The Power of Suggestion** [Suggested Compatibility]
5. **The Underlying Truth** [Matching Answers]
6. **Individual-Based Models**
7. **Personality** [As Determined by TheAlgorithm]
8. **Sexuality**[[2]](#footnote-2)[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.

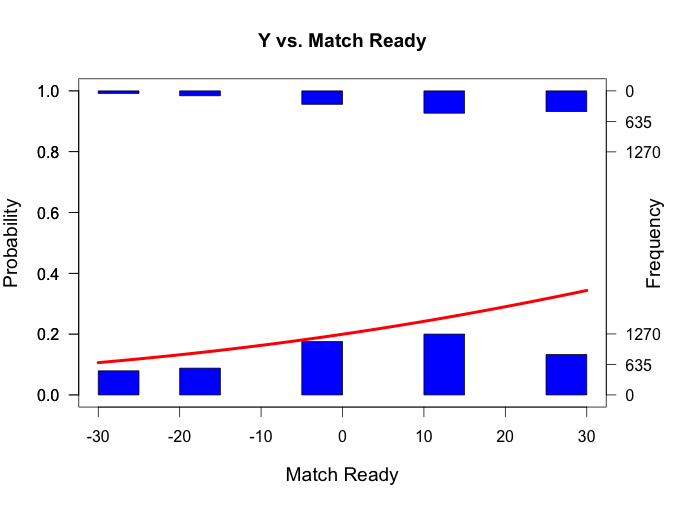
Here is the simple output from R:



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.

1. **Latent Eagerness** [Class Year, Time Responded to Survey]
2. **Self-Delusion** [Prior Willingness, In a Relationship]
3. **Love, what is the real thing?**

Plots:



1. The Crimson Data on Data-Match Article [↑](#footnote-ref-1)
2. See model\_sexuality.R [↑](#footnote-ref-2)