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

Note: While we've made the data public to the course staff for the purpose of this project, please do not publicize the existence of it or use it in any way other than for grading purposes. Thanks.

I. 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] 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] The data on data-match: http://www.thecrimson.com/article/2014/2/19/the-data-on-datamatch/

II. 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 structure it into what we wanted. To spare a reader the boringness from having to follow our path, a summary of what we did to build the data is in our footnotes.[2]

Essentially the end result of collecting all of the data is a data frame called master_frame.RData.

Here is a quick description of all of our variables:

Id: Id of user in our database
Y: 1 if user said yes to waflle-date, 0 otherwise[3]
straight: sexual preference
female: gender
soph, junior, senior, grad, alum, reference (freshman): class year
seconds_after_start: # of seconds that user completes survey after survey release
dating: 1 if in a relationship, 0 otherwise
same.house: 1 if user is in the same house as their potential waffle date
same.class: 1 if user is in the same class as their potential waffle date
same.questions: # questions user and potential waffle date answered the same
compat.score: compatibility score shown to the user as calculated by TheAlgorithm
ambitious-twisted: personality scores as calculated by TheAlgorithm

[2] See writeup/mysql_queries.txt and data/build_master_frame.R

[3] Ultimately, after filling out the Datamatch survey, users have an option of clicking "yes" to go on a waffle date with their top match. If someone else who is not the user's top match has the user as their own top match, the user will also have an option of saying yes to them. This means each user can have multiple "yes buttons" that they may press.

III. Modelling the arrow's path

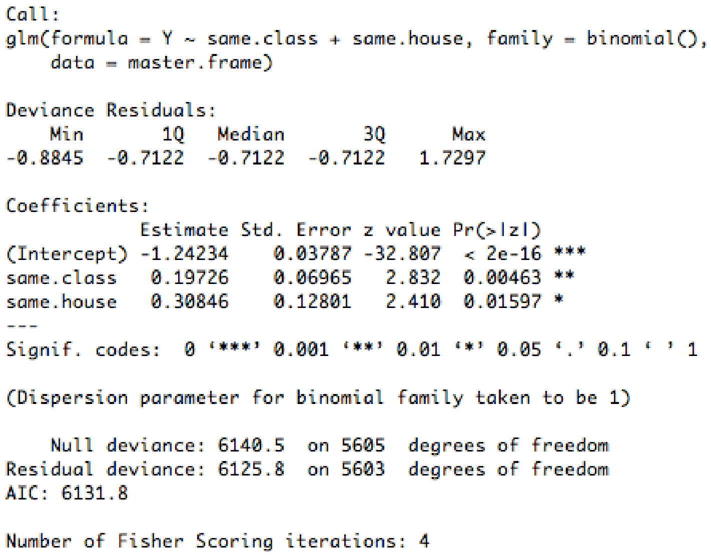
We first approached the problem by first considering model which individually could craft a narrative for the button clicks. In these steps, we did little assumption-checking or cross-validation; the only goal was to throw arrows at a target and see which couples stuck.

A. 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 [4][Class Year, House]

Proximity is one of the best predictors of love[5]. Within the Harvard bubble, quadlings date a disproportionate number of quadlings, Kirkland is incestuous, and Mather lathers in their troves of singles.



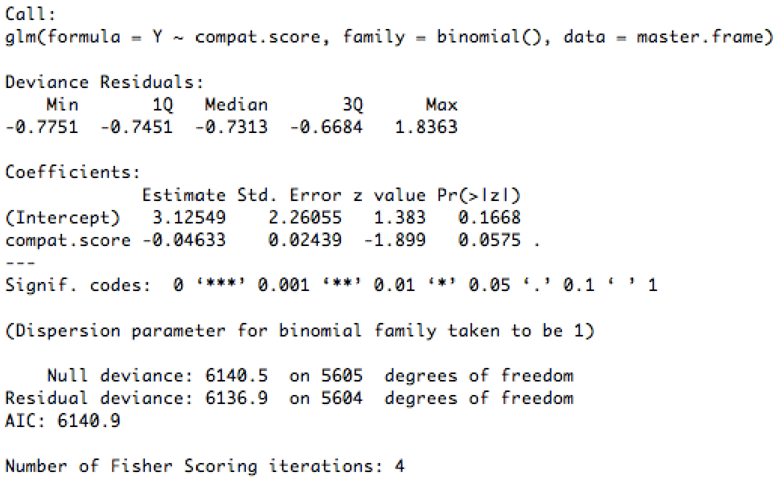
```
Call:
glm(formula = Y ~ same.class + same.house, family = binomial(),
    data = master.frame)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8845  -0.7122  -0.7122  -0.7122   1.7297

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.24234    0.03787 -32.807  < 2e-16 ***
same.class   0.19726    0.06965   2.832  0.00463 **
same.house   0.30846    0.12801   2.410  0.01597 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6140.5  on 5605  degrees of freedom
Residual deviance: 6125.8  on 5603  degrees of freedom
AIC: 6131.8

Number of Fisher Scoring iterations: 4
```

[4] See model_proximity.R

[5] http://www.npr.org/templates/story/story.php?storyId=112330125

2. The Power of Suggestion [6][Suggested Compatibility]

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

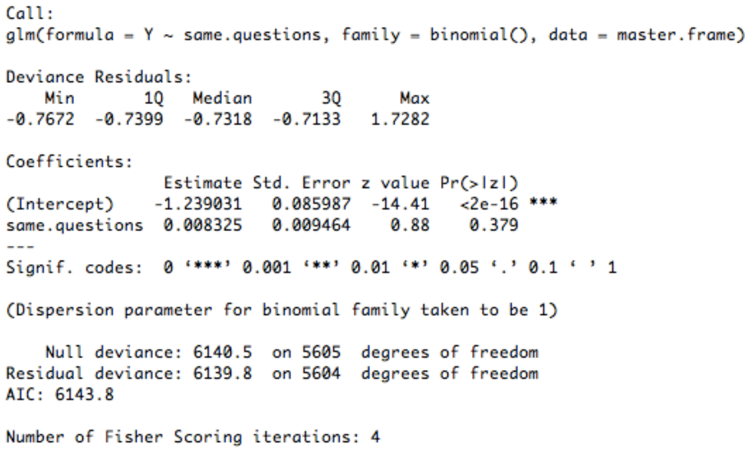


```
Call:
glm(formula = Y ~ compat.score, family = binomial(), data = master.frame)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7751  -0.7451  -0.7313  -0.6684   1.8363

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.12549    2.26055   1.383   0.1668
compat.score -0.04633    0.02439  -1.899   0.0575 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6140.5  on 5605  degrees of freedom
Residual deviance: 6136.9  on 5604  degrees of freedom
AIC: 6140.9

Number of Fisher Scoring iterations: 4
```

Probably not. Lol.

3. The Underlying Truth [7][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:



```
Call:
glm(formula = Y ~ same.questions, family = binomial(), data = master.frame)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7672  -0.7399  -0.7318  -0.7133   1.7282

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.239031   0.085987  -14.41   <2e-16 ***
same.questions  0.008325   0.009464    0.88    0.379
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6140.5  on 5605  degrees of freedom
Residual deviance: 6139.8  on 5604  degrees of freedom
AIC: 6143.8

Number of Fisher Scoring iterations: 4
```

B. Individual-Based Models

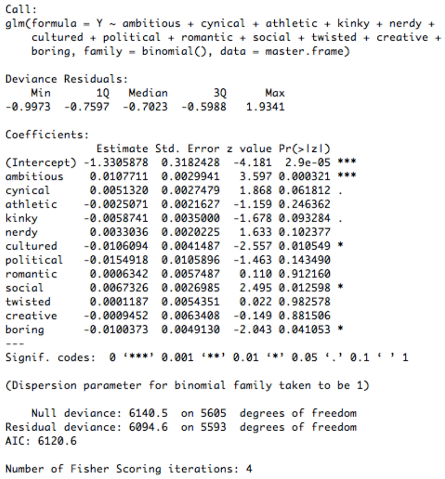
For these set of models, we ignored the interaction between partners, but rather tried to guess whether someone would agree to a waffle date based on their own inherent traits.

[6] See model_suggestion.R

[7] See model_truth.R

1. **Personality** [8][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:



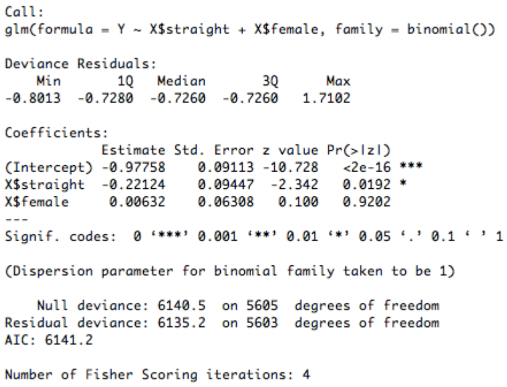
```
Call:
glm(formula = Y ~ ambitious + cynical + athletic + kinky + nerdy +
    cultured + political + romantic + social + twisted + creative +
    boring, family = binomial(), data = master.frame)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9973  -0.7597  -0.7023  -0.5988   1.9341

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.3305878  0.3182428  -4.181  2.9e-05 ***
ambitious    0.0107711  0.0029941   3.597 0.000321 ***
cynical      0.0051320  0.0027479   1.868 0.061812 .
athletic    -0.0025071  0.0021627  -1.159 0.246362
kinky       -0.0058741  0.0035000  -1.678 0.093284 .
nerdy        0.0033036  0.0020225   1.633 0.102377
cultured    -0.0106094  0.0041487  -2.557 0.010549 *
political   -0.0154918  0.0105896  -1.463 0.143490
romantic     0.0006342  0.0057487   0.110 0.912160
social       0.0067326  0.0026985   2.495 0.012598 *
twisted      0.0001187  0.0054351   0.022 0.982578
creative    -0.0009452  0.0063408  -0.149 0.881506
boring      -0.0100373  0.0049130  -2.043 0.041053 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6140.5  on 5605  degrees of freedom
Residual deviance: 6094.6  on 5593  degrees of freedom
AIC: 6120.6

Number of Fisher Scoring iterations: 4
```

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!

2. **Sexuality**[9] [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:



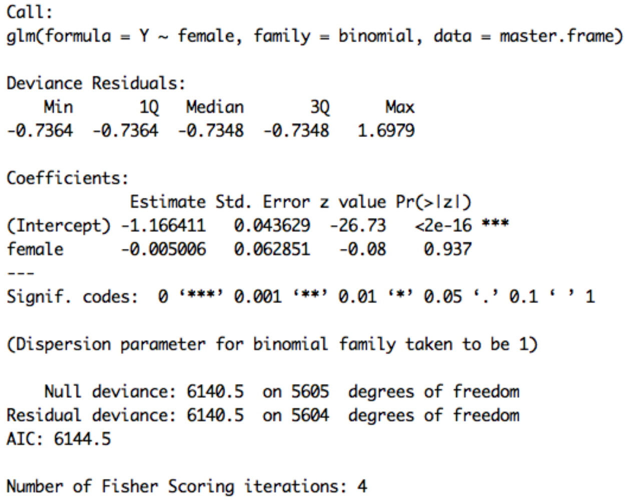
```
Call:
glm(formula = Y ~ X$straight + X$female, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8013  -0.7280  -0.7260  -0.7260   1.7102

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.97758    0.09113 -10.728   <2e-16 ***
X$straight  -0.22124    0.09447  -2.342   0.0192 *
X$female     0.00632    0.06308   0.100   0.9202
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6140.5  on 5605  degrees of freedom
Residual deviance: 6135.2  on 5603  degrees of freedom
AIC: 6141.2

Number of Fisher Scoring iterations: 4
```

[8] See model_personality.R

[9] See model_sexuality.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:

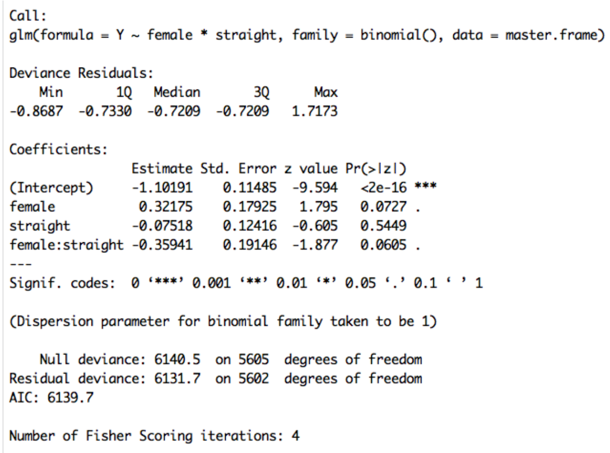


```
Call:
glm(formula = Y ~ female, family = binomial, data = master.frame)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7364  -0.7364  -0.7348  -0.7348   1.6979

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.166411   0.043629  -26.73   <2e-16 ***
female      -0.005006   0.062851   -0.08    0.937
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6140.5  on 5605  degrees of freedom
Residual deviance: 6140.5  on 5604  degrees of freedom
AIC: 6144.5

Number of Fisher Scoring iterations: 4
```

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:



```
Call:
glm(formula = Y ~ female * straight, family = binomial(), data = master.frame)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8687  -0.7330  -0.7209  -0.7209   1.7173

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.10191    0.11485  -9.594   <2e-16 ***
female           0.32175    0.17925   1.795   0.0727 .
straight        -0.07518    0.12416  -0.605   0.5449
female:straight -0.35941    0.19146  -1.877   0.0605 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6140.5  on 5605  degrees of freedom
Residual deviance: 6131.7  on 5602  degrees of freedom
AIC: 6139.7

Number of Fisher Scoring iterations: 4
```

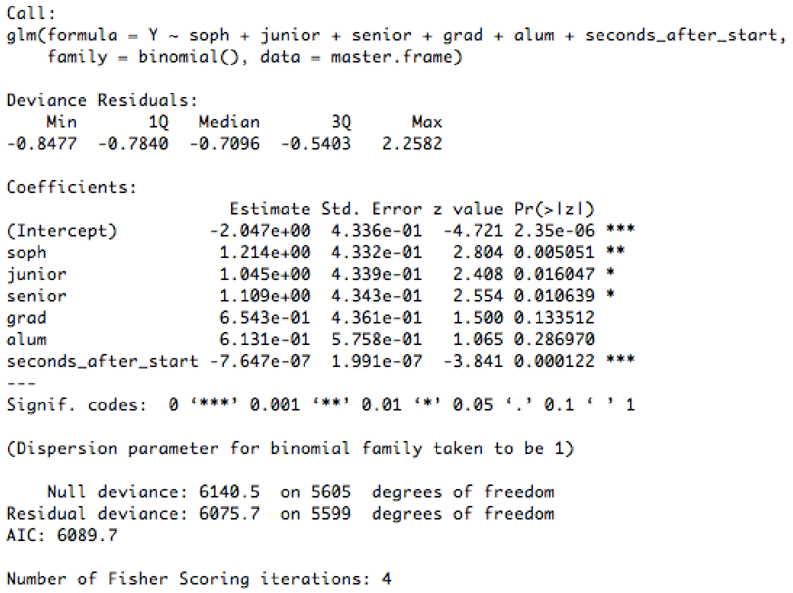
Which corresponds to:

	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.

3. **Latent Eagerness** [10][Class Year, Time Responded to Survey]

We know everyone wants to be seen as the one who cares less,[11] but we know how much you’ve actually been waiting for Datamatch.



```
Call:
glm(formula = Y ~ soph + junior + senior + grad + alum + seconds_after_start,
    family = binomial(), data = master.frame)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8477  -0.7840  -0.7096  -0.5403   2.2582

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.047e+00  4.336e-01  -4.721 2.35e-06 ***
soph                 1.214e+00  4.332e-01   2.804 0.005051 **
junior               1.045e+00  4.339e-01   2.408 0.016047 *
senior               1.109e+00  4.343e-01   2.554 0.010639 *
grad                 6.543e-01  4.361e-01   1.500 0.133512
alum                 6.131e-01  5.758e-01   1.065 0.286970
seconds_after_start -7.647e-07  1.991e-07  -3.841 0.000122 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6140.5  on 5605  degrees of freedom
Residual deviance: 6075.7  on 5599  degrees of freedom
AIC: 6089.7

Number of Fisher Scoring iterations: 4
```

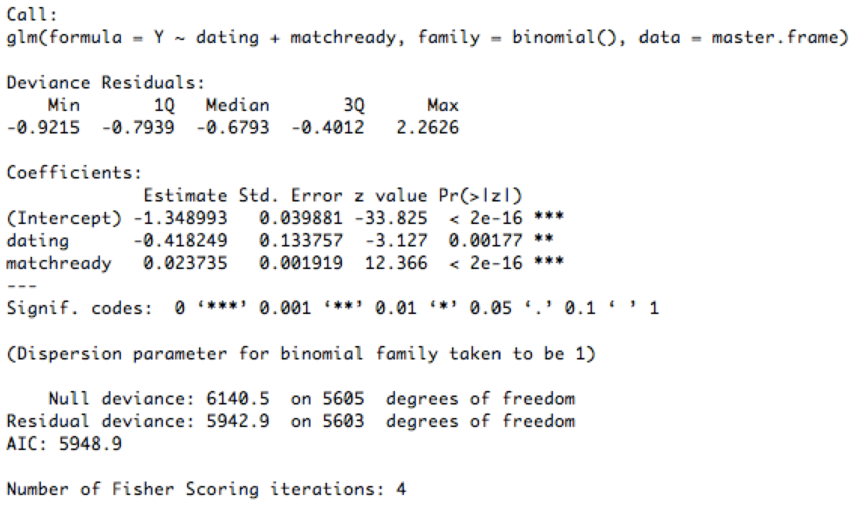
If you’re someone who filled out Datamatch early-on, then you’re quite more likely to go on a waffle date.

[10] See model_latent_eagerness.R

[11] http://www.cosmopolitan.com/sex-love/advice/a5585/college-dating-screwed-up/

4. **Explicit Eagerness** [12][Prior Willingness, In a Relationship]

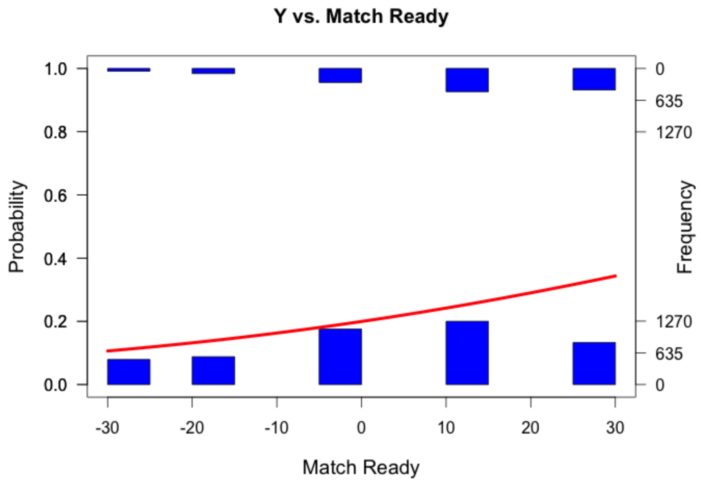
And finally, we can question how eager you are to go on a waffle-date given your explicit eagerness.



```
Call:
glm(formula = Y ~ dating + matchready, family = binomial(), data = master.frame)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9215  -0.7939  -0.6793  -0.4012   2.2626

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.348993   0.039881 -33.825  < 2e-16 ***
dating      -0.418249   0.133757  -3.127  0.00177 **
matchready   0.023735   0.001919  12.366  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6140.5  on 5605  degrees of freedom
Residual deviance: 5942.9  on 5603  degrees of freedom
AIC: 5948.9

Number of Fisher Scoring iterations: 4
```

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:



[12] See model_explicit_eagerness.R

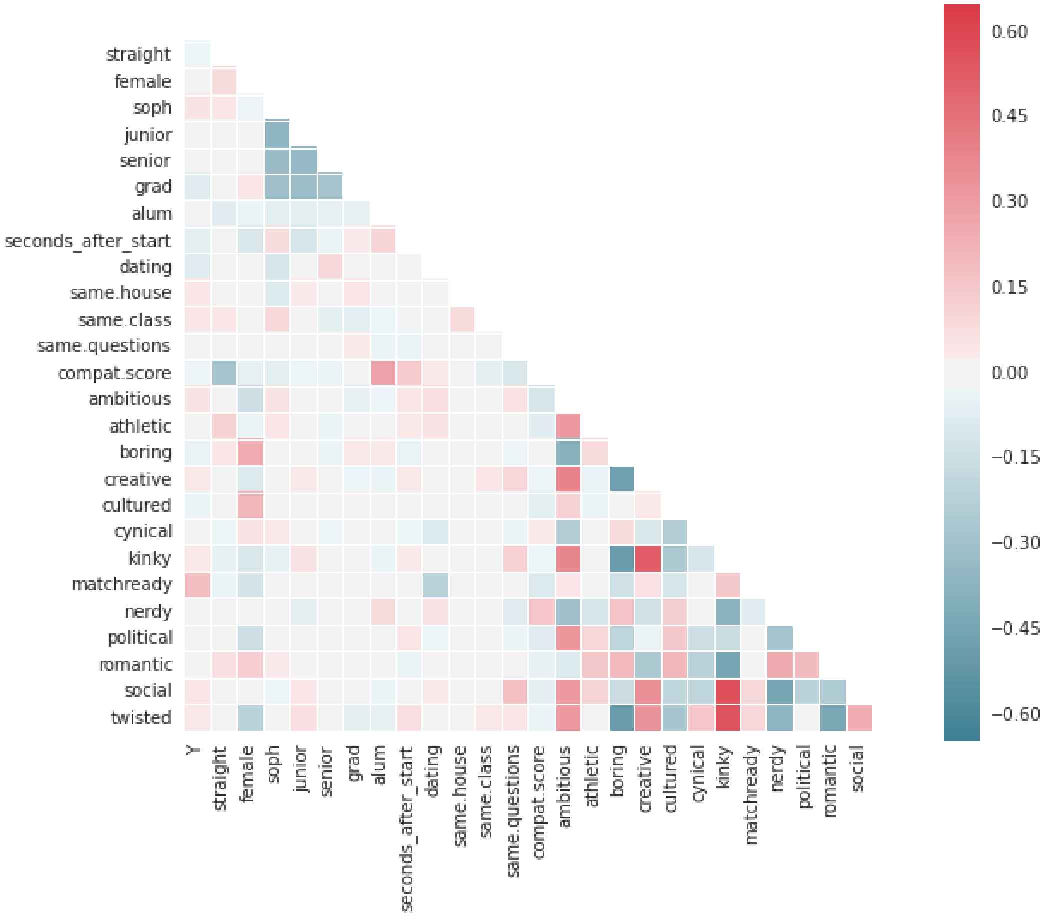
IV. Love, what is the real thing?

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

A. 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.

3. 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.

B. In search of the best model[13].

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:

1. *Generate a model with no interaction terms*
2. *Run forward and backward step function between null model and all interaction terms to minimize AIC*

model.base.BIC

1. *Generate a model with no interaction terms*
2. *Run forward and backward step function between null model and all interaction terms to minimize BIC*

model.base.AIC

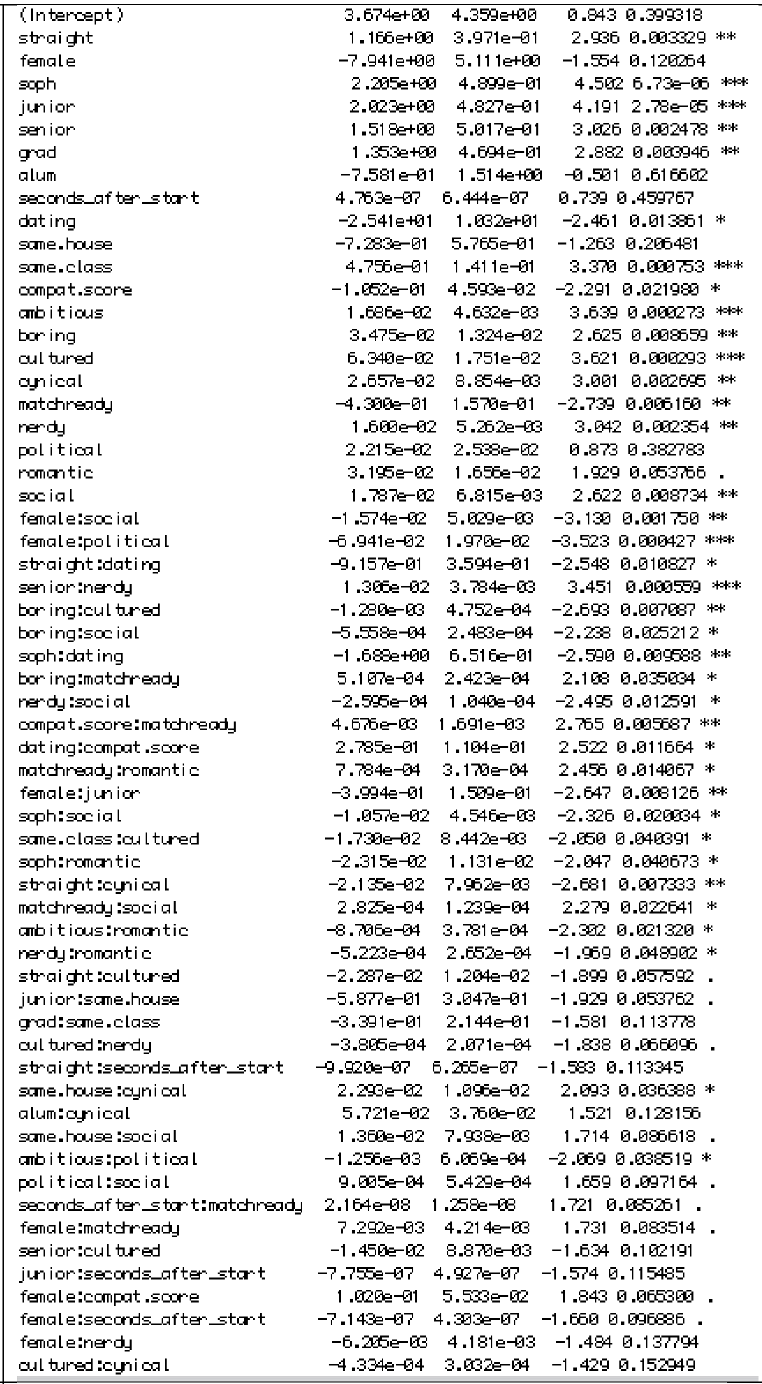
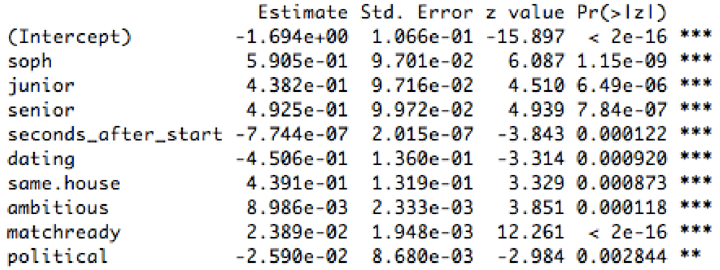
1. *Generate a model with no interaction terms*
2. *Run forward and backward step function between null model and full no − interaction model to minimize AIC.*
3. *Using the no − interaction AIC model, run forward and backward step function to minimize AIC.*

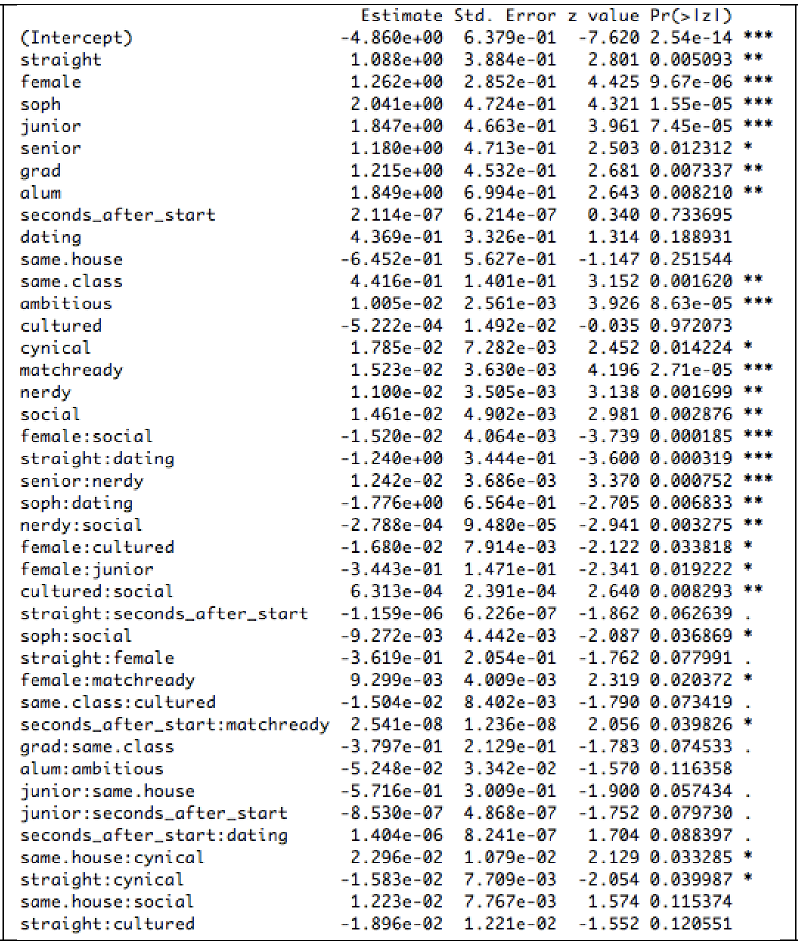
We get the following sets of significant variables in each:

Model	**Variables Selected**	**CV-Score[14]**
Null Model	--	.3613

[13] See build_model_best.R

[14] 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.

<table>
<tr>
<td><b>Best Model</b></td>
<td>
<pre>
(Intercept)   3.674e+00 4.359e+00   0.843 0.399318
straight   1.166e+00 3.971e-01   2.936 0.003329 **
female   -7.941e+00 5.111e+00 -1.554 0.120264
soph   2.205e+00 4.899e-01   4.502 6.73e-06 ***
junior   2.023e+00 4.827e-01   4.191 2.78e-05 ***
senior   1.518e+00 5.017e-01   3.026 0.002478 **
grad   1.353e+00 4.694e-01   2.882 0.003946 **
alum   -7.581e-01 1.514e+00 -0.501 0.616602
seconds_after_start   4.763e-07 6.444e-07   0.739 0.459767
dating   -2.541e+01 1.032e+01 -2.461 0.013861 *
same.house   -7.283e-01 5.765e-01 -1.263 0.206481
same.class   4.756e-01 1.411e-01   3.370 0.000753 ***
compat.score   -1.052e-01 4.593e-02 -2.291 0.021980 *
ambitious   1.686e-02 4.632e-03   3.639 0.000273 ***
boring   3.475e-02 1.324e-02   2.625 0.008659 **
cultured   6.340e-02 1.751e-02   3.621 0.000293 ***
cynical   2.657e-02 8.854e-03   3.001 0.002695 **
matchready   -4.300e-01 1.570e-01 -2.739 0.006160 **
nerdy   1.600e-02 5.262e-03   3.042 0.002354 **
political   2.215e-02 2.538e-02   0.873 0.382783
romantic   3.195e-02 1.656e-02   1.929 0.053766 .
social   1.787e-02 6.815e-03   2.622 0.008734 **
female:social   -1.574e-02 5.029e-03 -3.130 0.001750 **
female:political   -6.941e-02 1.970e-02 -3.523 0.000427 ***
straight:dating   -9.157e-01 3.594e-01 -2.548 0.010827 *
senior:nerdy   1.306e-02 3.784e-03   3.451 0.000559 ***
boring:cultured   -1.280e-03 4.752e-04 -2.693 0.007087 **
boring:social   -5.558e-04 2.483e-04 -2.238 0.025212 *
soph:dating   -1.688e+00 6.516e-01 -2.590 0.009588 **
boring:matchready   5.107e-04 2.423e-04   2.108 0.035034 *
nerdy:social   -2.595e-04 1.040e-04 -2.495 0.012591 *
compat.score:matchready   4.676e-03 1.691e-03   2.765 0.005687 **
dating:compat.score   2.785e-01 1.104e-01   2.522 0.011664 *
matchready:romantic   7.784e-04 3.170e-04   2.456 0.014067 *
female:junior   -3.994e-01 1.509e-01 -2.647 0.008126 **
soph:social   -1.057e-02 4.546e-03 -2.326 0.020034 *
same.class:cultured   -1.730e-02 8.442e-03 -2.050 0.040391 *
soph:romantic   -2.315e-02 1.131e-02 -2.047 0.040673 *
straight:cynical   -2.135e-02 7.962e-03 -2.681 0.007333 **
matchready:social   2.825e-04 1.239e-04   2.279 0.022641 *
ambitious:romantic   -8.706e-04 3.781e-04 -2.302 0.021320 *
nerdy:romantic   -5.223e-04 2.652e-04 -1.969 0.048902 *
straight:cultured   -2.287e-02 1.204e-02 -1.899 0.057592 .
junior:same.house   -5.877e-01 3.047e-01 -1.929 0.053762 .
grad:same.class   -3.391e-01 2.144e-01 -1.581 0.113778
cultured:nerdy   -3.805e-04 2.071e-04 -1.838 0.066096 .
straight:seconds_after_start -9.920e-07 6.265e-07 -1.583 0.113345
same.house:cynical   2.293e-02 1.096e-02   2.093 0.036388 *
alum:cynical   5.721e-02 3.760e-02   1.521 0.128156
same.house:social   1.360e-02 7.938e-03   1.714 0.086618 .
ambitious:political   -1.256e-03 6.069e-04 -2.069 0.038519 *
political:social   9.005e-04 5.429e-04   1.659 0.097164 .
seconds_after_start:matchready 2.164e-08 1.258e-08   1.721 0.085261 .
female:matchready   7.292e-03 4.214e-03   1.731 0.083514 .
senior:cultured   -1.450e-02 8.870e-03 -1.634 0.102191
junior:seconds_after_start   -7.755e-07 4.927e-07 -1.574 0.115485
female:compat.score   1.020e-01 5.533e-02   1.843 0.065300 .
female:seconds_after_start   -7.143e-07 4.303e-07 -1.660 0.096886 .
female:nerdy   -6.205e-03 4.181e-03 -1.484 0.137794
cultured:cynical   -4.334e-04 3.032e-04 -1.429 0.152949
</pre></td>
<td>.3339</td>
</tr>
<tr>
<td><b>Base BIC</b></td>
<td>
<pre>
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.694e+00 1.066e-01 -15.897 < 2e-16 ***
soph   5.905e-01 9.701e-02   6.087 1.15e-09 ***
junior   4.382e-01 9.716e-02   4.510 6.49e-06 ***
senior   4.925e-01 9.972e-02   4.939 7.84e-07 ***
seconds_after_start -7.744e-07 2.015e-07 -3.843 0.000122 ***
dating   -4.506e-01 1.360e-01 -3.314 0.000920 ***
same.house   4.391e-01 1.319e-01   3.329 0.000873 ***
ambitious   8.986e-03 2.333e-03   3.851 0.000118 ***
matchready   2.389e-02 1.948e-03 12.261 < 2e-16 ***
political   -2.590e-02 8.680e-03 -2.984 0.002844 **
</pre></td>
<td>.3439</td>
</tr>
</table>

Base AIC		.3369

```
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 -4.860e+00  6.379e-01  -7.620 2.54e-14 ***
straight                     1.088e+00  3.884e-01   2.801 0.005093 ** 
female                       1.262e+00  2.852e-01   4.425 9.67e-06 ***
soph                         2.041e+00  4.724e-01   4.321 1.55e-05 ***
junior                       1.847e+00  4.663e-01   3.961 7.45e-05 ***
senior                       1.180e+00  4.713e-01   2.503 0.012312 *  
grad                         1.215e+00  4.532e-01   2.681 0.007337 ** 
alum                         1.849e+00  6.994e-01   2.643 0.008210 ** 
seconds_after_start          2.114e-07  6.214e-07   0.340 0.733695    
dating                       4.369e-01  3.326e-01   1.314 0.188931    
same.house                  -6.452e-01  5.627e-01  -1.147 0.251544    
same.class                   4.416e-01  1.401e-01   3.152 0.001620 ** 
ambitious                    1.005e-02  2.561e-03   3.926 8.63e-05 ***
cultured                    -5.222e-04  1.492e-02  -0.035 0.972073    
cynical                      1.785e-02  7.282e-03   2.452 0.014224 *  
matchready                   1.523e-02  3.630e-03   4.196 2.71e-05 ***
nerdy                        1.100e-02  3.505e-03   3.138 0.001699 ** 
social                       1.461e-02  4.902e-03   2.981 0.002876 ** 
female:social               -1.520e-02  4.064e-03  -3.739 0.000185 ***
straight:dating             -1.240e+00  3.444e-01  -3.600 0.000319 ***
senior:nerdy                 1.242e-02  3.686e-03   3.370 0.000752 ***
soph:dating                 -1.776e+00  6.564e-01  -2.705 0.006833 ** 
nerdy:social                -2.788e-04  9.480e-05  -2.941 0.003275 ** 
female:cultured             -1.680e-02  7.914e-03  -2.122 0.033818 *  
female:junior               -3.443e-01  1.471e-01  -2.341 0.019222 *  
cultured:social              6.313e-04  2.391e-04   2.640 0.008293 ** 
straight:seconds_after_start -1.159e-06  6.226e-07  -1.862 0.062639 .  
soph:social                 -9.272e-03  4.442e-03  -2.087 0.036869 *  
straight:female             -3.619e-01  2.054e-01  -1.762 0.077991 .  
female:matchready            9.299e-03  4.009e-03   2.319 0.020372 *  
same.class:cultured         -1.504e-02  8.402e-03  -1.790 0.073419 .  
seconds_after_start:matchready  2.541e-08  1.236e-08   2.056 0.039826 *  
grad:same.class             -3.797e-01  2.129e-01  -1.783 0.074533 .  
alum:ambitious              -5.248e-02  3.342e-02  -1.570 0.116358    
junior:same.house           -5.716e-01  3.009e-01  -1.900 0.057434 .  
junior:seconds_after_start  -8.530e-07  4.868e-07  -1.752 0.079730 .  
seconds_after_start:dating   1.404e-06  8.241e-07   1.704 0.088397 .  
same.house:cynical           2.296e-02  1.079e-02   2.129 0.033285 *  
straight:cynical            -1.583e-02  7.709e-03  -2.054 0.039987 *  
same.house:social            1.223e-02  7.767e-03   1.574 0.115374    
straight:cultured           -1.896e-02  1.221e-02  -1.552 0.120551    
```

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:

straight
female
soph
junior
senior
seconds_after_start
dating
same.house
ambitious
matchready

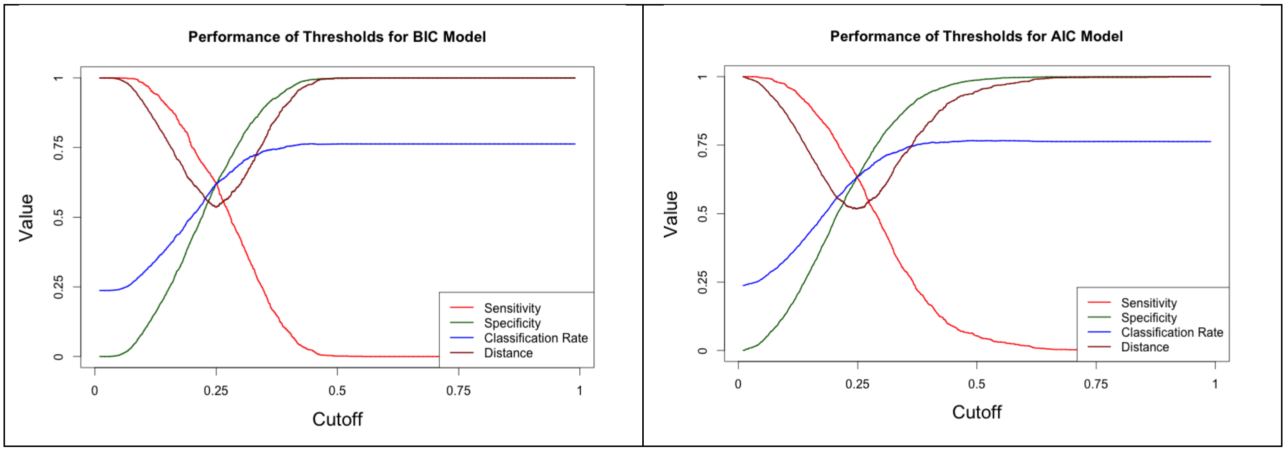
Looking at our BIC model as it's the easiest to interpret, the model can be understood as follows. Holding for all other things constant each variable with a coefficient β_v has a β_v increase in the log-odds with each unit increase of the variable. This results in a e^{β_v} change in the odds of saying yes to going on a waffle-date.

For example, if you're dating someone you have a $e^{-.4506} = .6372$ change in your odds of saying yes to a waffle date.

C. Prediction

Now, suppose that we want to predict which users will say yes or no to a waffle date in order to increase the actual number of waffle dates. To do so, we would need to set a theshold value for p so that we would predict a user would say "yes" to the date if the model predicts a value higher than p, and "no" otherwise.

We can then compute the sensitivity, specificity, and classification rates for various values of p, as shown by the graphs below for our AIC and BIC models:



The sensitivity curve depicts $P(\hat{Y}_i = 1 \,|Y_i = 1)$, or the proportion of "yes's" that we correctly identify. The specificity curve depicts $P(\hat{Y}_i = 0 \,|Y_i = 0)$, or the proportion of "no's" that we correctly identify. The classification rate curve depicts $P(\hat{Y}_i = Y_i)$, or the proportion of responses that we correctly identify.

Hence, depending on which measure that we want to do well on, we would choose the cutoff correspondingly based on the values of the graph above.

As an example, suppose we want to maximize the number of pairs of people who go on waffle dates. This would require matching people who are more likely to say yes to people who are also more likely to say yes, and vice versa. Hence, we would want to jointly optimize both of the sensitivity and specificity rates of our predictions. We could therefore use the Euclidean distance of our sensitivity and specificity rates to (1, 1) as a measure of optimality, as shown below

$$\delta = \sqrt{\left(P\left(\hat{Y}_i = 1 \,\middle|\, Y_i = 1\right) - 1\right)^2 + \left(P\left(\hat{Y}_i = 0 \,\middle|\, Y_i = 0\right) - 1\right)^2}$$

This is also depicted by the brown line in the graph above. We would then find the threshold value that minimizes δ in order to find the optimal cutoff to predict whether someone would say yes or no to a waffle date.

V. The End of Love

We've come a long way from leaving love to Cupid's arrow or hoping for a miracle spawned by St. Valentine in the days of our great grandfathers and grandmothers. The world of love is now thrown against a world of data, matching our primordial emotions against the cold, calculating kinds of Coffee Meets Bagel, Match.com, Tinder, and now, Datamatch.

Our goal was to return the power back to Cupid and back to St. Valentine. We toiled long and hard to come up with models to best predict couples, to give Cupid that nudge he needs to shoot the arrow, or to give St. Valentine the courage he needs to make a match. We tested both a series of models coming from our human intuition, models with easy narratives like sexuality, proximity, eagerness. But, in the end, we left the work to the data itself – coming up with our best model, aptly named: "model.best."

As St. Valentine is in the off-season, waiting for his time to shine again come February, the front-office staff has optimized his work. And, in the end, if we are the cause of people's meeting, if we are the movers of fate and the creators of destiny for a single couple, we will have done our job as St. Valentine's front office staff.