Final Report

Jonathan Che, Rene Kooiker, Sarah Teichman

Sunday, May 10, 2015

Introduction

Our data comes from a study conducted between 2009 and 2013, in four waves, by Michael J. Rosenfeld, Reuben J. Thomas, and Maja Falcon, which surveyed 4002 adults in the United States about how they met their romantic partners. It contains 457 variables, including demographic information of the respondent, relationship status, living arrangements with the partner, how and when the respondent met their partner, and self-reported quality of the relationship. Demographics includes such information as religion, level of education, age, race/ethnicity, gender, income, and household composition. Where applicable, corresponding information about the respondent's partner was provided by the respondent. Consecutive waves each followed up with respondents after a year, asking about changes in relationship status and reasons for separation where applicable. The fourth wave adds a self-reported level of attractiveness for the respondent and their partner. The survey research was done by Knowledge Networks, a survey firm. They recruited respondents from an ongoing, representative panel using random digit dial phone survey.

Initially, we tried to predict how couples met with demographic information and found that there weren't any strong correlations. We also looked for any strong associations between various pairs of variables, picking variables that intuitively made sense together and checking to see if the statistical results matched our predictions. This method was not very productive, and we didn't come to any specific conclusions about the dataset. Instead, we came up with the following research question: What variables of the survey are significant predictors of whether a couple will break up? We fit one model with mainly background information, and one with data about where and how the couple met.

It's interesting to study whether different demographic groups have different breakup rates. There might be some disparities, which would indicate an instability of relationships in certain groups. For example, are same-sex couples more or less likely to separate than traditional heterosexual couples? Are there differences in age? If so, is that attributable to the couple having been together for a long time, or cultural and attitudinal differences? Additionally, the way couples meet might change corresponding to different demographic backgrounds. We might expect young people to meet in different ways than older people, which in turn might be predictive of couple dissolution rates. Especially with new developments of online dating, it's interesting to find out whether having met online has any predictive power for whether a couple will break up.

Dataset

The dataset we started with had 4002 observations for 457 variables, although we ended up removing various observations and variables, and creating other variables in our analysis. One unit in this dataset was one person who responded to the survey, and the various columns were the questions asked in the survey. This dataset fulfills the condition of randomness because the experimenters used a random sample of adults in the United States by using a random phone number paradigm. However, these findings can only be generalized to US adults with working phones because that is the population that was sampled from. The observations are independent because the population is very large, and the answers of one participant would not affect the answers of another participant.

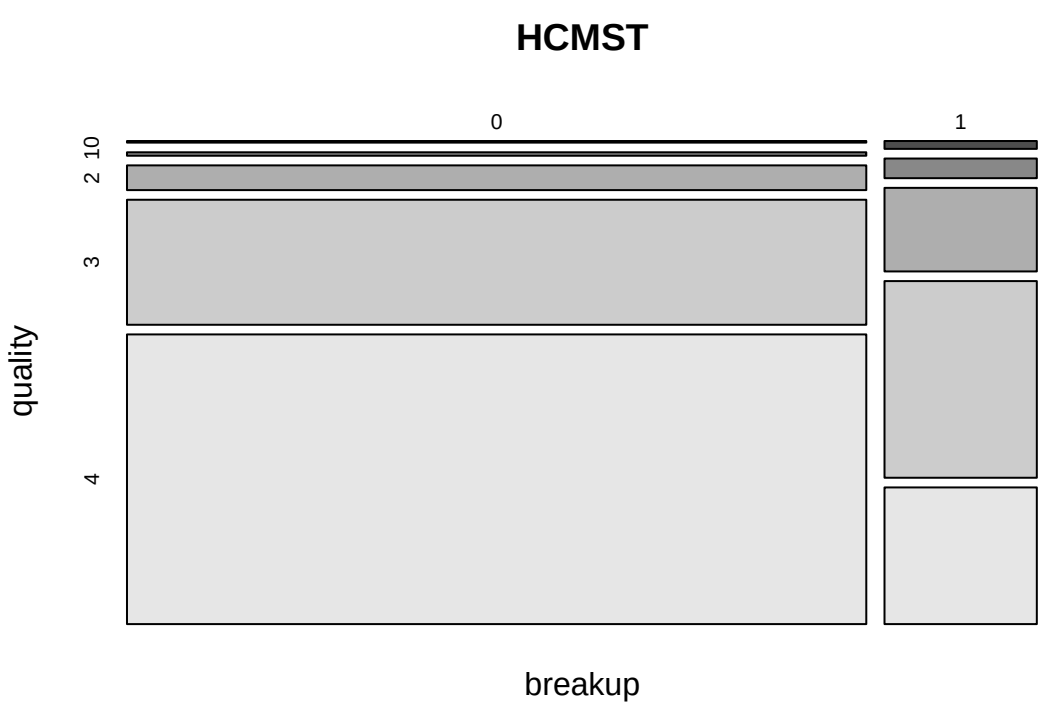
The biggest challenge we faced using this dataset was the size of it, and trying to determine which variables were worth studying without getting lost in the many confusing and repetitive variables the dataset contains. Most of the variables that we used in our models came from the first wave of questions in the survey, because this wave had the most new variables and the highest response rates. The second, third, and fourth wave of the survey mainly asked about whether certain variables: status of the relationship, religion, job, etc. had changed since the first wave.

Another difficulty in this dataset was all of the non responses for various questions. There were many variables that appeared promising, until we checked the codebook and saw that they had very low response rates. This limited the variables that we could put into our models without eliminating too many observations.

We focused on the response variable of breakup. We used the variable for breakup that combined participants who had reported breakups from waves 2, 3, and 4. This had a response rate of 76%, so in using this in our models we immediately eliminated the participants who did not respond to this question. It is a binary variable, with couples who stayed together coded with a 0, and breakups that were reported in wave 2, 3, or 4 coded as a 1. This led to our decision to create models using multiple logistic regression. The variables quality, age category, marital status, age difference, how long since first met, and metropolitan area or not were our main explanatory variables for our first model. Quality was a self-reported measure of quality of the relationship from excellent to very poor. We recoded this as a quantitative variable to use in our analysis as a variable from 0 to 4, with 0 as the worst and 4 the best. Age category was a variable that measured the participant's age in the first wave of the survey, and reported it in one of seven age categories. Marital status is a variable that reports whether the participant is currently married, widowed, divorced, separated, never married, or living with a partner, and these options are mutually exclusive because the sum of the counts for each option equal the number of observations in total. Age difference is the absolute value of difference in age between the participant and his or her partner. How long since first met measured the number of years since the participant had met his or her current partner, and reported it in one of seven categories. Metropolitan area is a binary variable that is coded as a 0 for non-metro areas and a 1 for metro-areas. Each of these variables had only a quarter or less nonresponses, so we felt comfortable using them without eliminating too many observations. Our other model used how people met as the explanatory variable, with 8 indicator variables for different ways of meeting. 4 were significant in our model: meeting at work, at school, at church, and through a dating service. These were coded as 0 if not how the couple met, and 1 if it was how the couple met. This question had a response rate of 75%, so we were comfortable using it without eliminating too much data.

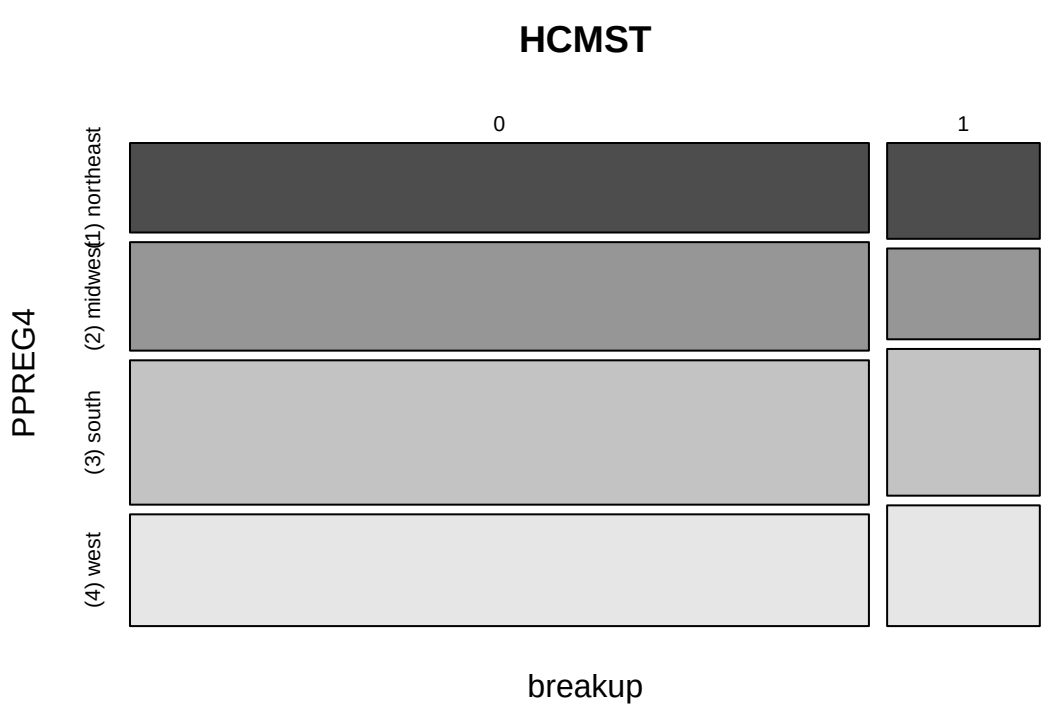
```
load("30103-0001-Data.rda")
HCMST <- da30103.0001
rm(da30103.0001)
HCMST <- transform(HCMST, breakup=as.numeric(W234_COMBO_BREAKUP)-1)
HCMST <- transform(HCMST, quality = 5 - as.numeric(Q34)) #0-4, with 0 being the lowest quality
```

```
mosaicplot(~breakup+quality, data=HCMST, color=T)
```



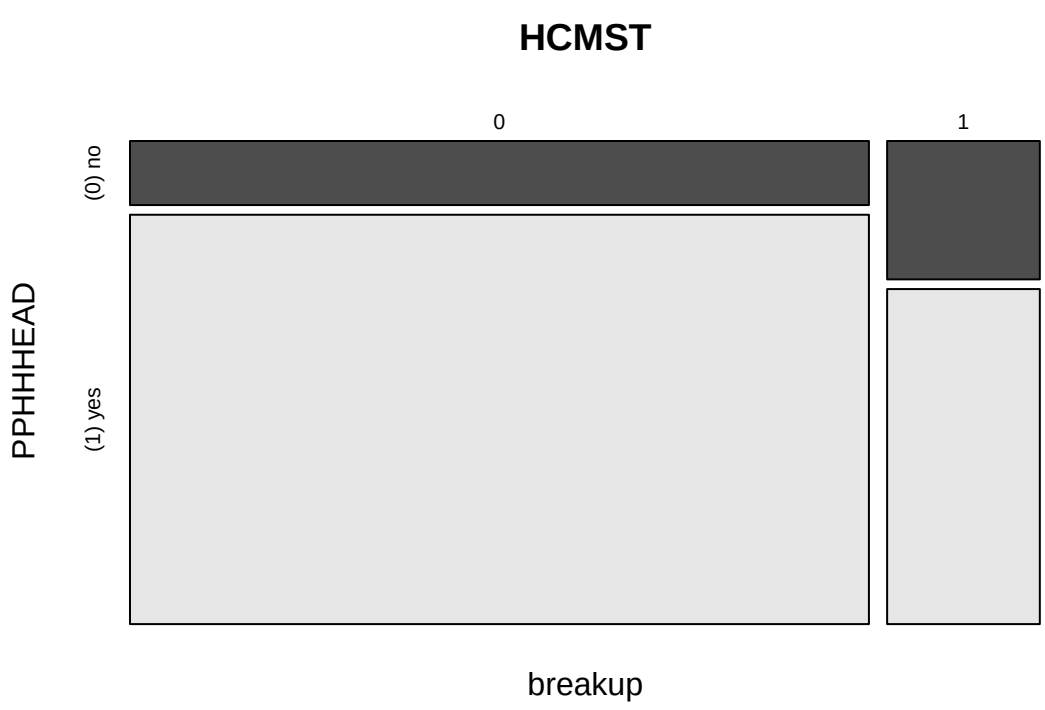
This is a mosiac plot of breakup and quality of relationship. This shows that there is a strong association between the two variables because of the large differences especially for the extremes. Most people who rate their quality of relationship as a 0 are likely to break up, while many more people who rate their quality of relationship as a 5 are likely to stay together.

```
mosaicplot(~breakup+PPREG4, data=HCMST, color=T)
```



This is a mosiac plot of breakup and region of the country. It shows that there does not appear to be a strong association between the two variables, because most of the region cells have similar sizes despite the couple breaking up or staying together. However, there are a few very small differences so it could be useful to keep this variable in the large model and use step-wise regression to see if it is significant.

```
mosaicplot(~breakup+PPHHHEAD, data=HCMST, color=T)
```



This is a mosiac plot of breakup and whether or not the participant is the head of their household. This variable appears to be significant, but after further analysis we found that it wasn't a significant predictor in our model. Because mosaic plots like this one can sometimes be misleading, we decided to make a large model with many variables that could be significant predictors, and then use stepwise regression to make a final model.

Analysis

To begin, we wanted to look at whether we could predict breakups based on information about the respondents' backgrounds and relationships. We started with a full model using all of the variables that both made sense to include in this specific model as predictors of breakups and did not have too many nonresponses. After omitting the N/As from the new dataset, we end up with a dataset of 2,619 observations of the 18 variables that we think could be statistically significant predictors of breakups.

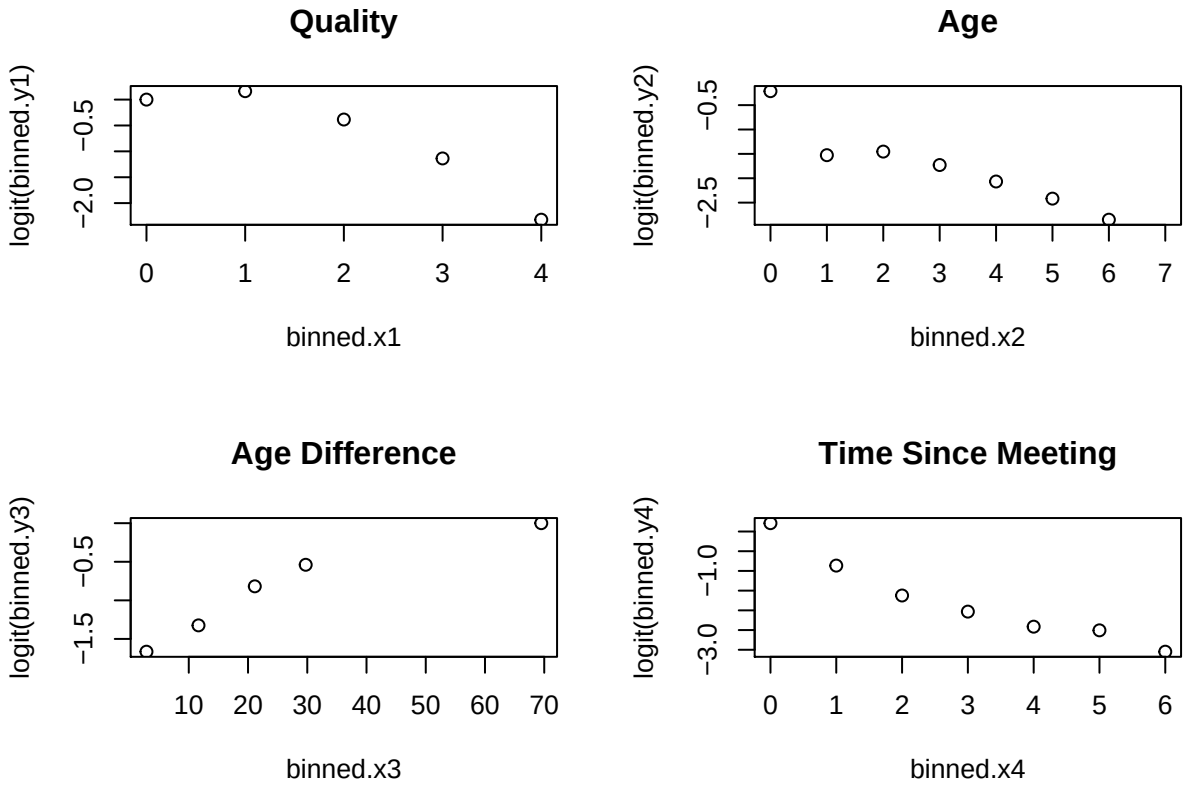
```
HCMST2 <- select(HCMST, breakup, quality,PPAGECAT,PPEDUCAT,PPETHM,PPHHHEAD,PPMARIT,PAPRELIGION,
                 AGE_DIFFERENCE,HOW_LONG_AGO_FIRST_MET_CAT,CHILDREN_IN_HH,
                 PPMSACAT,PPREG4,PPINCIMP,PPHOUSE,PPWORK,PPPARTYID3,PAPGLB_STATUS)
HCMST2 <- na.omit(HCMST2)
```

Next, we run a backwards stepwise regression to determine which of the 18 selected variables are the most significant:

```
mbreak3 <- glm(formula=breakup~quality+PPAGECAT+PPEDUCAT+PPETHM+PPMARIT+PAPRELIGION+
                 AGE_DIFFERENCE+HOW_LONG_AGO_FIRST_MET_CAT+CHILDREN_IN_HH+PPMSACAT+
                 PPREG4+PPINCIMP+PPHOUSE+PPWORK+PPPARTYID3+PAPGLB_STATUS,
                 family="binomial", data=HCMST2)
```

```
mbreak4 <- step(mbreak3,direction="backward")
```

The stepwise regression returns a model that uses quality, PPAGECAT, PPMARIT, AGE DIFFERENCE, HOW LONG AGO FIRST MET CAT, and metro to predict whether a couple will break up. Before making any conclusions about the summary output, we need to check the conditions for logistic regression. As mentioned earlier, we feel that the methodology behind the collection of the dataset allows us to assume both independence and randomness of observations. So, we can move on to check the empirical logit plots for the 6 predictors in the final model. Since metro is a binary predictor, we know that linearity is guaranteed. PPMARIT is effectively a collection of 6 binary predictors, since there is no obvious order in which the 6 categories of marital status should be placed, and the effect of each category can be considered independently of the other 5 categories. As such, we also know that the linearity of those 6 binary predictors is guaranteed. So, we need only to check the empirical logit plots of quality, PPAGECAT, AGE DIFFERENCE, and HOW LONG AGO FIRST MET CAT.



All four of the empirical logit plots quite linear, with the exception of a few unusual points. For the quality plot, we notice that the "Quality = 0" bin is not in line with the other 4. If we look at a tally of quality, however:

```
tally(~quality, HCMST3)
```

```
##
##    0    1    2    3    4
##   16   37  205  807 1554
```

we notice that only 16 people reported a quality of 0. With such a small sample size, we feel okay to proceed with caution in our analysis. Similarly, a tally of the Age Difference categories:

```
tally(~AGE_DIFF_Cat, HCMST3)
```

```
##
## (-0.07,8.75]  (8.75,17.5]  (17.5,26.2]    (26.2,35]    (35,43.8]
##         2188          357           49           19            3
##  (43.8,52.5]  (52.5,61.2]  (61.2,70.1]
##            1            0            2
```

reveals that the age difference categories past 35 years of age difference have extremely low sample sizes. As such, we feel okay to proceed with caution even though the empirical logit plot tails off of linearity toward the higher categories.

Since the conditions for logistic regression are met, we can now interpret the coefficients of the summary output for the model.

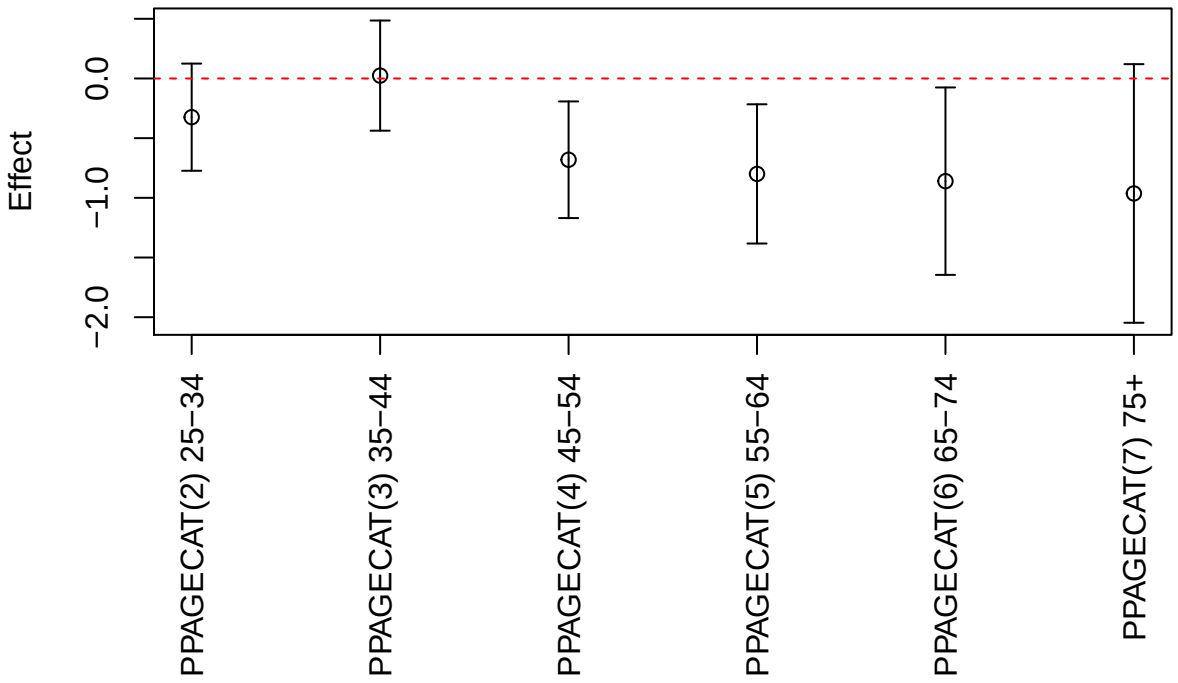
```
summary(mbreak4)$coefficients[,-c(2,3)]
```

```
##                                      Estimate     Pr(>|z|)
## (Intercept)                         0.94958555 2.295211e-02
## quality                            -0.92056308 1.287932e-31
## PPAGECAT(2) 25-34                  -0.32428934 1.568629e-01
## PPAGECAT(3) 35-44                   0.02392449 9.190470e-01
## PPAGECAT(4) 45-54                  -0.68106111 6.303003e-03
## PPAGECAT(5) 55-64                  -0.79956906 7.205028e-03
## PPAGECAT(6) 65-74                  -0.86042833 3.172230e-02
## PPAGECAT(7) 75+                    -0.96338612 8.132784e-02
## PPMARIT(2) widowed                  1.84546995 3.344771e-05
## PPMARIT(3) divorced                 2.80426859 9.477479e-33
## PPMARIT(4) separated                1.52261187 2.805537e-04
## PPMARIT(5) never married            2.17357400 1.448869e-25
## PPMARIT(6) living with partner      1.30527671 5.877988e-11
## AGE_DIFFERENCE                      0.02722283 1.486463e-02
## HOW_LONG_AGO_FIRST_MET_CAT(2) 3-5  -0.64548734 1.471795e-03
## HOW_LONG_AGO_FIRST_MET_CAT(3) 6-10 -0.97216130 2.603968e-06
## HOW_LONG_AGO_FIRST_MET_CAT(4) 11-15 -1.12670420 1.367506e-06
## HOW_LONG_AGO_FIRST_MET_CAT(5) 16-20 -1.53989043 1.751548e-07
## HOW_LONG_AGO_FIRST_MET_CAT(6) 21-30 -1.31777991 2.128824e-06
## HOW_LONG_AGO_FIRST_MET_CAT(7) 31+  -1.56456214 8.790759e-07
## PPMSACAT(1) metro                   0.50802913 1.568336e-02
```

The easiest coefficients to interpret are quality, AGE DIFFERENCE, and metro. At an alpha-level of 0.05, all three of these predictors are significant. The coefficient for quality is negative, so we conclude that greater self-reported quality of relationship is correlated with lesser log odds of breakup, after accounting for the effects of the other variables. The coefficient for AGE DIFFERENCE is positive, so we conclude that greater difference in age between the partners is correlated with greater log odds of breakup increases, after accounting for the effects of the other variables. The coefficient for metro is positive, so we conclude that living in metropolitan areas is correlated with a greater log odds of breakup, after accounting for the effects of the other variables.

To help visualize the interpretation of the other coefficients, we construct confidence-interval plots. First, let's look at the plot for PPAGECAT:

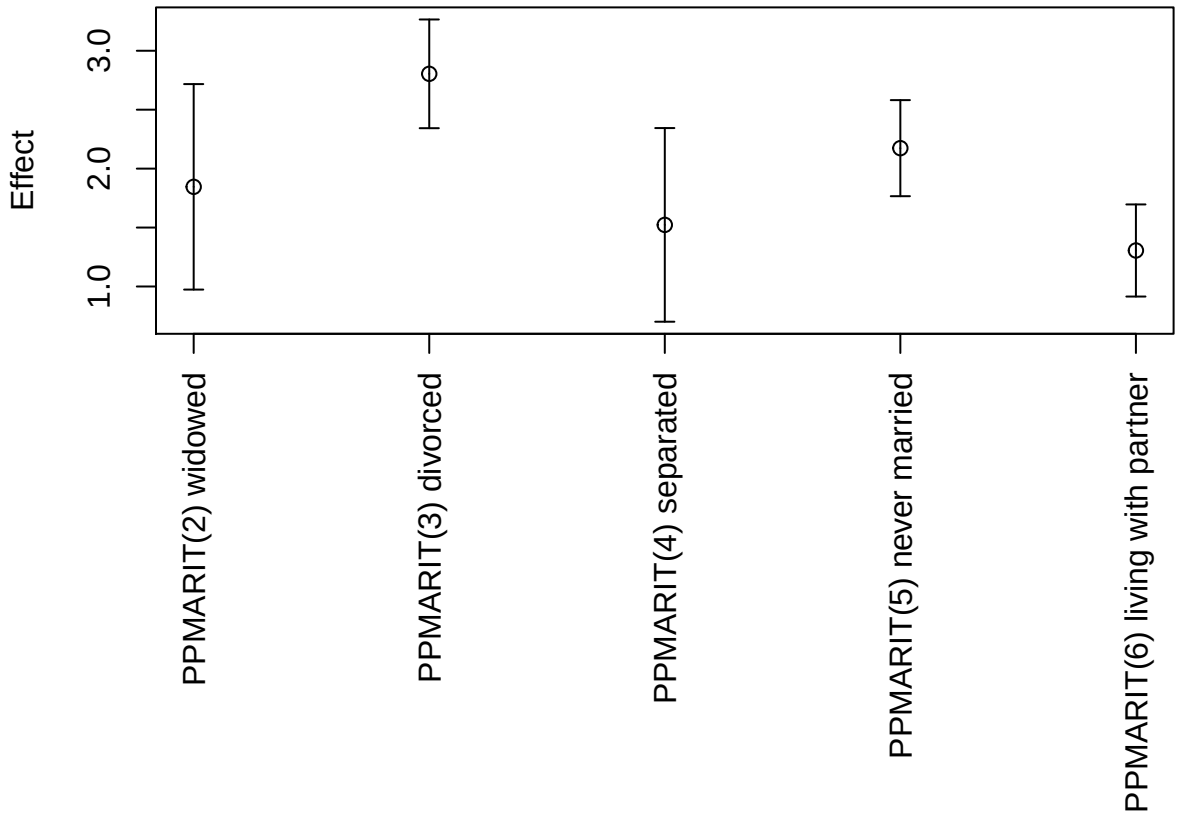
```
s <- data.frame(summary(mbreak4)$coefficients)
s <- bind_cols(data.frame(rownames(s)),s)

par( mfrow = c( 1, 1 ) )

ageCI <- s[3:8,]
plotCI(ageCI,conf.level=0.95)
```



The intervals for categories of 25-34, 35-44, and 75+ intersect 0, so they are not significant at an alpha-level of 0.05. The three categories for 45-74, however, are all significant and less than 0. It seems that, controlling for other variables, those three categories are significantly correlated with lower log odds of breakup than the base group of 18 to 25-year-olds.

Next, let's look at the confidence interval plot for PPMARIT:

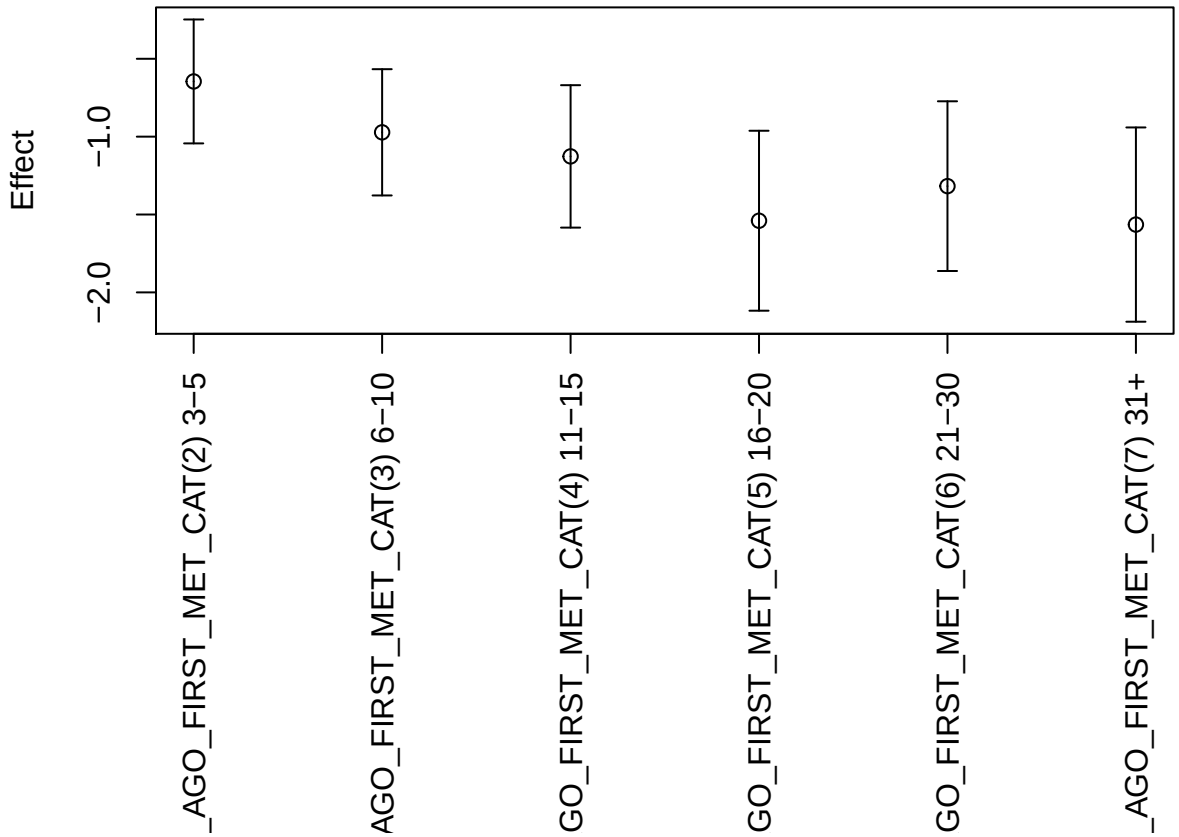
```
maritCI <- s[9:13,]
plotCI(maritCI,conf.level=0.95)
```



The intervals for all of the categories are above 0, so at an alpha-level of 0.05, they all significantly correlate with a greater log odds of breakup than the base group of "married", controlling for the effects of other variables. Also, we note that the confidence intervals for "divorced" and "living with partner" do not overlap, so we can conclude, at an alpha-level of 0.05, that divorced people have a significantly greater log odds of breaking up than people who are living with their partner.

Finally, let's look at the confidence interval plot for HOW LONG AGO FIRST MET CAT:

```
metCI <- s[15:20,]
plotCI(metCI,conf.level=0.95)
```



The intervals for all of the categories are less than 0, so at an alpha-level of 0.05, they all significantly correlate with a less log odds of breakup than the base group of "Met 0-2 years ago", controlling for the effects of other variables. It seems like there is also a trend in trends, that longer times since meeting correlate with even lower log odds of breakup, but the confidence intervals overlap and we cannot make any significant conclusions about that.

After interpreting the model, we want to evaluate it to see how good it is at predicting data. First, we check its Somer's D:

```
Association(mbreak4)
```

```
## $`Concordant Pairs`
## [1] 857466
##
## $`Discordant Pairs`
## [1] 111409
##
## $Tied
## [1] 283
##
## $Pairs
## [1] 969158
```

```
(857466-111409)/969158
```

```
## [1] 0.7697991
```

The Somer's D of this model is 0.77, so it seems that this model is actually quite good at predicting.

Next, we run a 5-fold cross-validation test:

```
cv.glm(HCMST2, mbreak4, K=5)$delta
```

```
## [1] 0.09954639 0.09913240
```

The first component here is the raw cross-validation estimate of prediction error. The second component is the adjusted cross-validation estimate. The adjustment is designed to compensate for the bias introduced by not using leave-one-out cross-validation. So the error rate of this model is 9.8%, which is quite a low error rate. If we look at a tally of where the error lies, though:

```
ests <- predict(mbreak4, newdata=HCMST2, type="response")
ests <- ests > 0.5
tally(~ests+HCMST2$breakup)
```

```
##        HCMST2$breakup
## ests       0    1
##   TRUE   110  190
##   FALSE 2063  256
```

We notice that most of the model's error comes from failing to predict actual breakups. Though this result makes sense, since in the original dataset most couples do in fact stay together, it would still be something to consider if we wish to use this model as a predictive model for breakups.

Overall, the model that we built based on demographic/background information to predict breakups seems like a strong predictive model. Though it somewhat underpredicts breakups, this underprediction is in line with the generally low proportion of couples that do in fact break up over the course of 4 years.

After building the demographic model, we wanted to look at whether the ways in which couples met would be significant predictors of whether they would break up. How couples met was answered in two different questions in the dataset, one of which was free response and coded by the researchers, and one that had the participant pick one or more of nine possible answers. We used the fixed choice variable because the free response coded answers were less organized and could have been influenced by the subjective biases of the coders. In our full model we use eight indicator variables for where couples met (excluding the "other" option). These indicators were meeting through, school, church, personal ads, vacation, bar, social club, and personal party.

```
HCMST4 <- select(HCMST, breakup, met_through_work, met_through_school, met_through_church,
                 met_through_personal_ads, met_through_vacation, met_through_bar,
                 met_through_social_club, met_through_private_party)
HCMST4 <- na.omit(HCMST4)
```

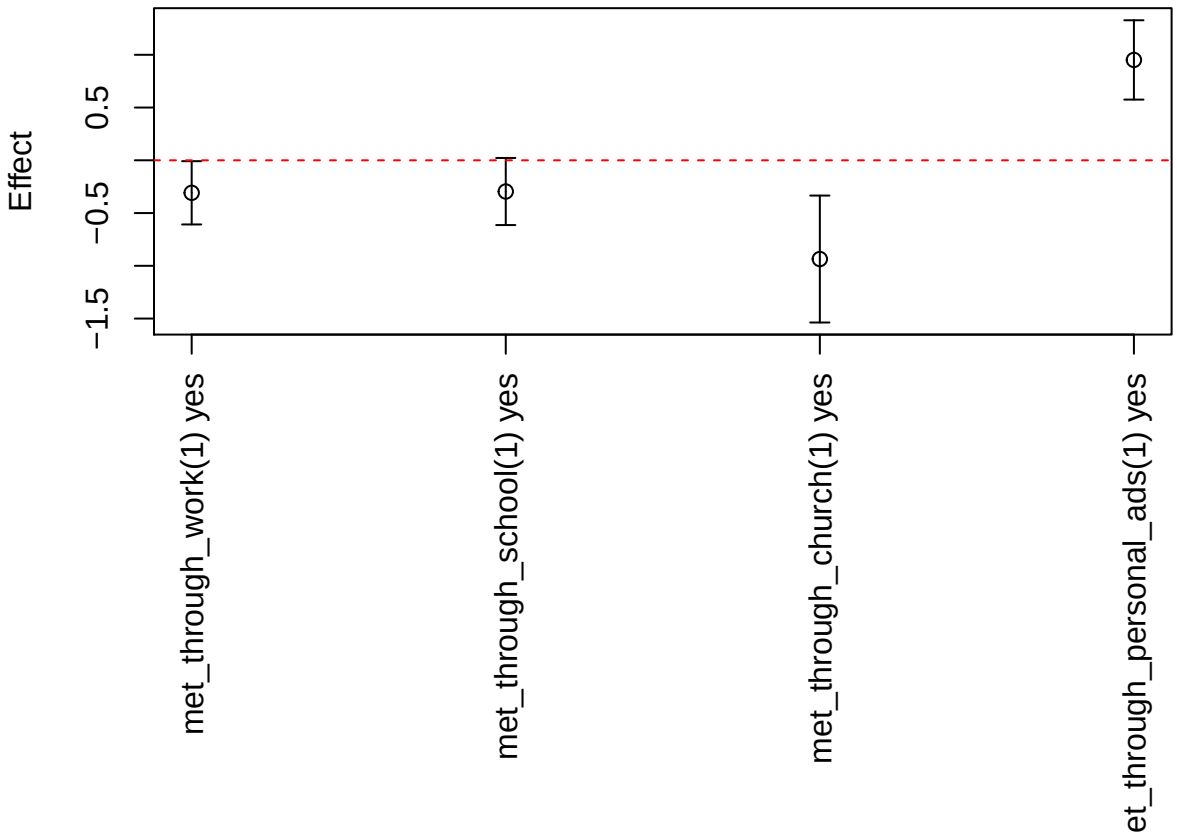
Next, we run a backwards stepwise regression to determine which of the 8 selected variables is the most significant:

```
mbreak5 <- glm(formula=breakup~met_through_work+met_through_school+met_through_church+
                 met_through_personal_ads+met_through_vacation+met_through_bar+
                 met_through_social_club+met_through_private_party,
                 family="binomial", data=HCMST4)
```

```
mbreak6 <- step(mbreak5,direction="backward")
```

The stepwise regression returns a model that uses met through school, met through work, met through church, and met through personal ads to predict whether a couple will break up. Before making any conclusions about the summary output, we need to check the conditions for logistic regression. As mentioned earlier, we feel that the methodology behind the collection of the dataset allows us to assume both independence and randomness of observations. Also, all of our predictors are binary, so linearity is guaranteed. Thus, we can move on to interpret the coefficients. Confidence interval plots will help us visualize the output.

```
t <- data.frame(summary(mbreak6)$coefficients)
t <- bind_cols(data.frame(rownames(t)),t)
meetingCI <- t[2:5,]
plotCI(meetingCI,conf.level=0.95)
```



At an alpha-level of 0.05, not all of the predictors returned by the backwards stepwise regression are significant. The confidence interval for met through school includes zero, so we cannot conclude that it is a significant predictor of breakups. The confidence intervals for met through work and met through church are entirely negative, so we can conclude that there is evidence that couples who meet through work and church have significantly lower odds of breaking up than people who meet through ways that are not considered in this model, controlling for the effects of meeting through school and meeting through personal ads. The confidence interval of met through personal ads is entirely positive, so we can conclude that there is evidence that couples who meet through personal ads have significantly higher odds of breaking up than people who meet through ways that are not considered in this model, controlling for the effects of meeting through work, school, and church.

Next, we move on to evaluate the predictive power of this model:

```
Association(mbreak6)
```

```
## $`Concordant Pairs`
## [1] 367481
##
## $`Discordant Pairs`
## [1] 217118
##
## $Tied
## [1] 403993
##
## $Pairs
## [1] 988592
```

```
(367481-217118)/(988592-403993)
```

```
## [1] 0.2572071
```

Clearly, there are a lot of ties in the association test. This is because the model uses only 4 binary predictors to predict the log odds of breakup. Thus, there is a high chance that two observations will have the same values for all 4 binary predictors, and thus have the same predicted log odds of breakup. So, we use the Goodman-Kruskal Gamma instead of Somer's D. The Goodman-Kruskal Gamma of this model is 0.257, so it seems that this model is not very good at predicting.

When we try to run a k-fold cross-validation test, we realize that the model never predicts a probability of breakup that is greater than 0.5. If we look at the coefficients of the model:

```
summary(mbreak6)
```

```
##
## Call:
## glm(formula = breakup ~ met_through_work + met_through_school +
##     met_through_church + met_through_personal_ads, family = "binomial",
##     data = HCMST4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9479  -0.6294  -0.6294  -0.5463   2.2524
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -1.5184     0.0655 -23.184  < 2e-16 ***
## met_through_work(1) yes          -0.3082     0.1530  -2.014  0.04400 *
## met_through_school(1) yes        -0.2959     0.1622  -1.824  0.06816 .
## met_through_church(1) yes        -0.9359     0.3069  -3.050  0.00229 **
## met_through_personal_ads(1) yes   0.9513     0.1918   4.960 7.05e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2415.2  on 2642  degrees of freedom
```

```
## Residual deviance: 2369.0  on 2638  degrees of freedom
## AIC: 2379
##
## Number of Fisher Scoring iterations: 5
```

We notice that the maximum possible predicted log odds of breakup is (-1.518 + 0.951) = -0.567, which, after some arithmetic, corresponds to a probability of 0.36. Thus, the model will never predict that a couple will break up at a probability greater than 0.5. Though this result makes sense, given the fact that the probability of breakup is so low in general, it means that this model should not be used for prediction puposes. We can still, though, look at the interpretations of its coefficients and draw conclusions from those.

Conclusions

In our project, we came up with two models to predict breakups. One used demographic and relationship information, and the other used indicator variables about how couples met. By using stepwise regression, we found which variables were significant and which ones were not in our model.

Our final model used quality of relationship, age category, age difference, how long since meeting, marital status, and metropolitan area to predict breakups. We found that most of these results made sense with our initial predictions and societal stereotypes. For example, our model shows that greater age differences will lead to more breakups, that someone who has been divorced previously is more likely to have another breakup, and that the longer a couple has known each other, the less likely they are to break up. These are all assumptions that we previously had from societal norms. However, our model also had significant predictors that we hadn't initially thought of, like metropolitan area. We were also surprised by the predictors that we put in the larger model that turned out not to be signficant with the other predictors in the final model. We expected to find significant results for number of children in the household, religion, and education. Our meeting model predicted breakups using whether couples met at work, school, church, or personal ad. It was interesting that three indicators were significant at the 0.05 alpha level (school had a p-value of slightly larger), with work and church decreasing the odds of breakup and personal ads increasing the odds of breakup, but that the other four indicators for the other categories were not significant.

We have concerns about this data because of all the observations that we had to eliminate for nonresponse. We believe that we didn't have to exclude too much of the data, because it is likely that many of the nonresponses were from the same people, who only filled out a minimum amount of the survey. However, if there are systematic differences in people who responded to everything and people who had one or more nonresponses, these differences could affect our analysis and results.

Our analysis could have been supplemented by other data sources. It would have been interesting to look at a US census from the same year as the first wave and to compare some of the demographic data through descriptive statistics. By looking at distributions of races, education, incomes, and other variables, we could have observed how accurate of a sample our dataset is for the entire population of adults in the United States. This could help us determine whether or not the condition of randomness is really met. We could have compared tally tables for most of the demographic variables for the two groups: one from the census and one from the How Couples Meet survey, to see if there were any noticeable differences between the two groups. We could have also used divorce records (if they contained all of the variables that we used) as individual case studies, to test our model by seeing how well it predicted the divorces.

The data from this survey is very extensive and thorough, and the researchers did an impressive job in collecting and sorting it. If we were collecting the same data, we think it might be worthwhile to ask fewer questions, with the objective of having a shorter survey and less nonresponses. If there were fewer questions to answer, the partcipants might feel more motivation to complete the entire survey. However, with all surveys there are questions about subjectivity and biases. It would be interesting to observe a few couples in a small observational study, in order to measure some of the more subjective variables as less biased researchers instead of asking for self reported answers. It would then be interesting to compare the results from the small observational study to the large self-reported data from the survey. However, it would not be a very practical

experimental design because it would be intrusive and would take a large amount of time and effort to collect the data.

Code Appendix

Association Function:

```
# ***modified FUNCTION TO CALCULATE CONCORDANCE AND DISCORDANCE from R blog***#
Association = function(model) {
    Con_Dis_Data = cbind(model$y, model$fitted.values)
    ones = Con_Dis_Data[Con_Dis_Data[, 1] == 1, ]
    zeros = Con_Dis_Data[Con_Dis_Data[, 1] == 0, ]
    conc = matrix(0, dim(zeros)[1], dim(ones)[1])
    disc = matrix(0, dim(zeros)[1], dim(ones)[1])
    ties = matrix(0, dim(zeros)[1], dim(ones)[1])
    for (j in 1:dim(zeros)[1]) {
        for (i in 1:dim(ones)[1]) {
            if (ones[i, 2] > zeros[j, 2]) {
                conc[j, i] = 1
            } else if (ones[i, 2] < zeros[j, 2]) {
                disc[j, i] = 1
            } else if (ones[i, 2] == zeros[j, 2]) {
                ties[j, i] = 1
            }
        }
    }
    Pairs = dim(zeros)[1] * dim(ones)[1]
    Concordance = sum(conc)
    Discordance = sum(disc)
    Tied = sum(ties)
    return(list(`Concordant Pairs` = Concordance, `Discordant Pairs` = Discordance,
        `Tied`= Tied, Pairs = Pairs))
}
```

Plot CI Function

```
# Goal: trying to plot confidence intervals (assuming normality)
# Inputs:
#  - coef: coefficient matrix, first column is a label, second column is estimate,
#        third column is SE
#  - conf.level (optional): the confidence level
# Output: plot of confidence intervals
plotCI <- function(coef, conf.level=0.95) {
  n <- nrow(coef)
  xvec <- 1:n
  est <- coef[,2]
  CIs <- cbind(coef[,2], coef[,2])
  CIs[,1] <- CIs[,1]  - qnorm((1-conf.level)/2)*coef[,3]
  CIs[,2] <- CIs[,2]  + qnorm((1-conf.level)/2)*coef[,3]
  par(mar=c(13,5,1,1))
  plot(est, ylim=range(CIs), xaxt="n", ylab="Effect", xlab="")
  arrows(xvec,est,xvec,CIs[,1],angle=90,length=0.05) ## lower bars
  arrows(xvec,est,xvec,CIs[,2],angle=90,length=0.05) ## upper bars
```

```
  abline(h=0, col="red", lty=2)
  axis(1, at=xvec, labels=coef[,1], las=2)
}
```

Renaming meeting variables

```
HCMST <- rename(HCMST, met_through_work = Q31_1)
HCMST <- rename(HCMST, met_through_school = Q31_2)
HCMST <- rename(HCMST, met_through_church = Q31_3)
HCMST <- rename(HCMST, met_through_personal_ads = Q31_4)
HCMST <- rename(HCMST, met_through_vacation = Q31_5)
HCMST <- rename(HCMST, met_through_bar = Q31_6)
HCMST <- rename(HCMST, met_through_social_club = Q31_7)
HCMST <- rename(HCMST, met_through_private_party = Q31_8)
```

Creating the empirical logit plots

```
binned.x1 <- 0:4
binned.y1 <- mean(breakup ~ quality, data=HCMST2)

binned.x2 <- 0:7
binned.y2 <- mean(breakup ~ PPAGECAT, data=HCMST2)

HCMST3 <- transform(HCMST2, AGE_DIFF_Cat = cut(AGE_DIFFERENCE, 8))
binned.x3 <- mean(AGE_DIFFERENCE ~ AGE_DIFF_Cat, data=HCMST3)
binned.y3 <- mean(breakup ~ AGE_DIFF_Cat, data=HCMST3)

binned.x4 <- 0:6
binned.y4 <- mean(breakup ~ HOW_LONG_AGO_FIRST_MET_CAT, data=HCMST2)
par( mfrow = c( 2, 2 ) )

plot(logit(binned.y1) ~ binned.x1, main = "Quality")
plot(logit(binned.y2) ~ binned.x2, main = "Age")
plot(logit(binned.y3) ~ binned.x3, main = "Age Difference")
plot(logit(binned.y4) ~ binned.x4, main = "Time Since Meeting")
```