Swipe Right:
A Statistical Analysis of Meeting Online and Relationship Length

Miro Furtado and Tejal Patwardhan

Final Project for Statistics 139

13 December 2017

1 Introduction

Searching online for love is increasingly popular. As of 2016, over 15% of adults have used some form of online dating site, and the online dating industry is worth over $2B (Pew Research Center, Lilien). This is accompanied by increased societal acceptance of online dating as a legitimate means to find a stable romantic partner (Slater). Given the current cultural context regarding online matchups, we are interested in whether meeting online affects the length of a relationship. Perhaps searching online increases the field of potential partners, decreasing tendency to settle for incompatible partners and improving stability of relationships. Perhaps online matchmaking platforms do a good job optimizing compatibility so that relationships last longer. Perhaps meeting in person through shared hobbies or mutual friends makes it more difficult to break up, resulting in longer relationships. These are just a few of the many theories about the effects of meeting online on relationship length (Slater). We are interested in whether we can find a significant association between meeting online and relationship length after performing a more quantitative analysis of this phenomenon. We hypothesize that meeting online is significantly associated with relationship length. In the following sections, we (1) determine whether there is an association between meeting online and relationship length, (2) determine whether this association holds up when controlling for other factors, and (3) build a best predictive model for relationship length.

2 Data

All data was self-reported through the Stanford Social Science Data Collection Initiative survey *How Couples Meet and Stay Together*, an online questionnaire for English-literate adults in the United States. Documentation and data can be accessed at https://data.stanford.edu/hcmst.

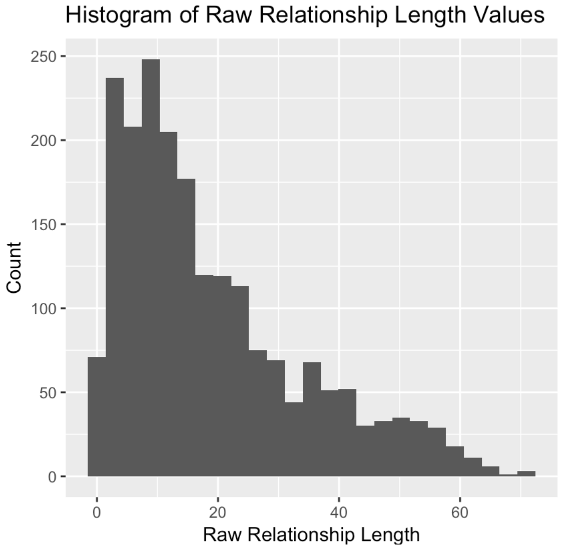
The original dataset included values for 387 questions, but we only wanted to focus on a few variables here. Beyond relationship length and whether the couple met online, we chose additional predictors that we thought were interesting and/or which could be good predictors of relationship length. We will control for these other variables in our analysis to get a better estimate of the relationship between meeting online and relationship length. We included all individuals that had values for these variables in our analysis, and excluded the few that did not. This is valid because if deleting missing cases still leaves a large dataset to analyze, the deletion strategy is satisfactory and avoids the complexity and potential biases introduced by more sophisticated methods to adjust for missing data (Faraway). We have 2056 individuals in our analysis. The variables we included are listed below, with more detailed explanation in the appendix:

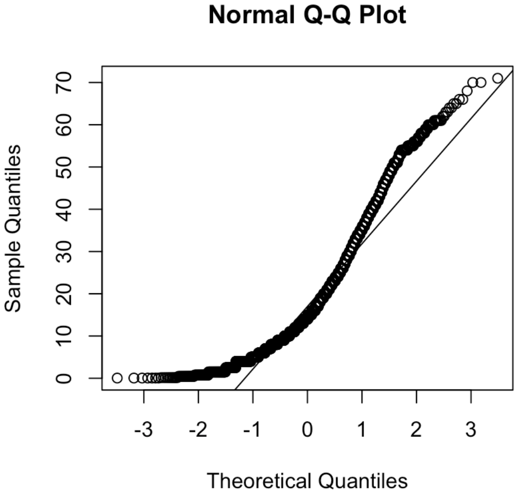
- rlength: relationship length, in years
- online: met online? (online includes via social networking, online matchmaking like eHarmony, internet classified sites like craigslist, and internet chat rooms).

- ppage: age, in years
- ppgender: gender
- ppethm: ethnicity
- ppeduc: level of education, on a numeric scale
- hhinc: household income, in dollars
- ppreg4: region of residence
- pppartyid3: political party
- paplgbstatus: identify as LGBTQ?
- paplgbfriend: any friends or relatives identify as LGBTQ?
- pphhhead: is individual the household head?
- ppnet: internet access at home?
- duration: interview duration in minutes, rounded down

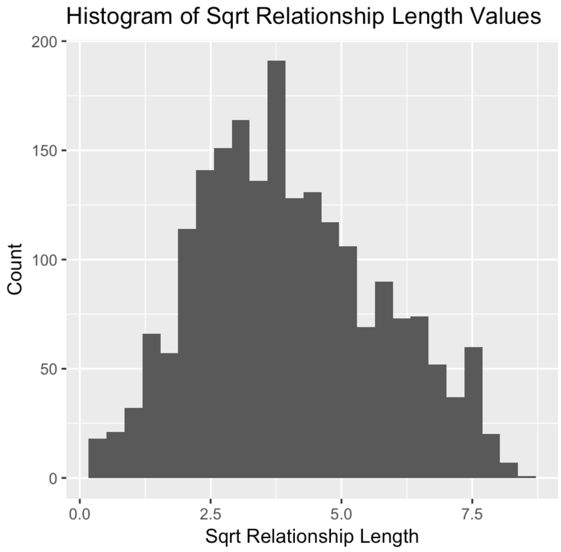
3 Transformations

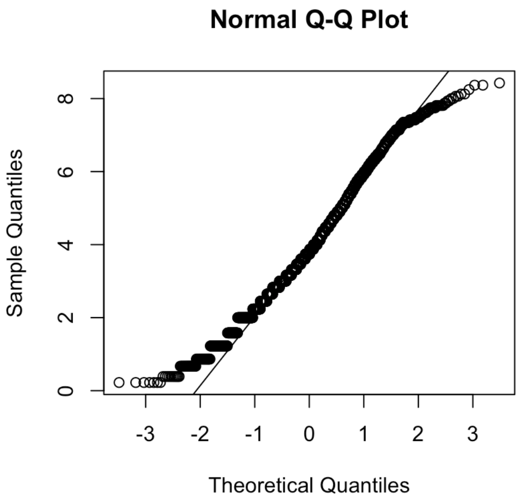
From the histogram and Q-Q plot of relationship length,





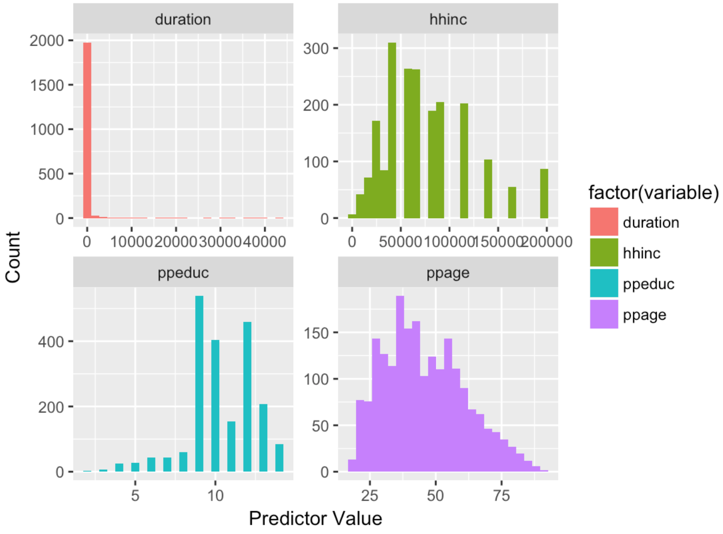
there is a substantial right skew in the raw relationship length variable that warrants a transformation. We tried a few transformations, and the one that worked best on the right tail was the square root transform:



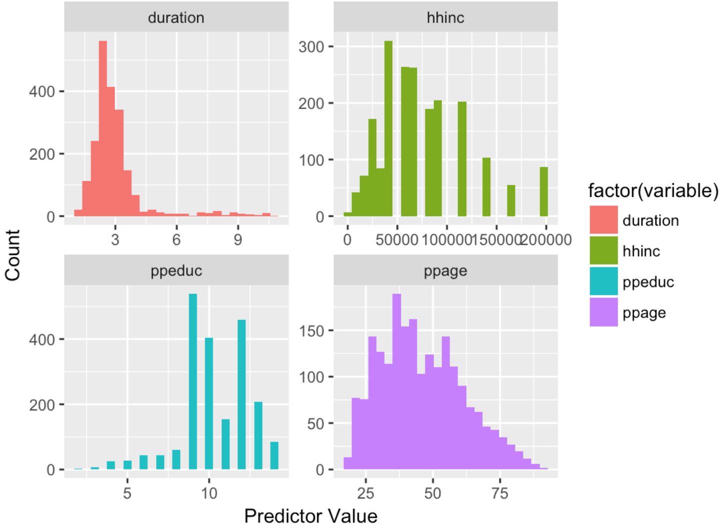


This results in a more bell-shaped histogram and a less tail-heavy Q-Q plot.

Let's now check if we need to transform any predictors. The binary and categorical predictors need not be transformed. For our quantitative predictors, looking at the raw histograms,



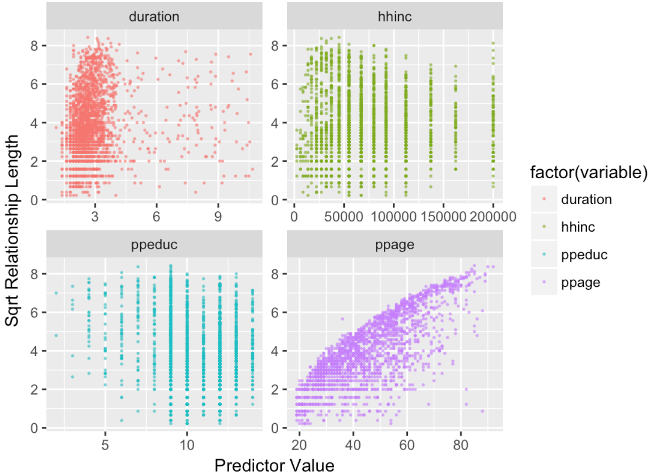
they all seem to be fairly normal, except for duration, which has quite a long right tail and could benefit from a log transform. This results in

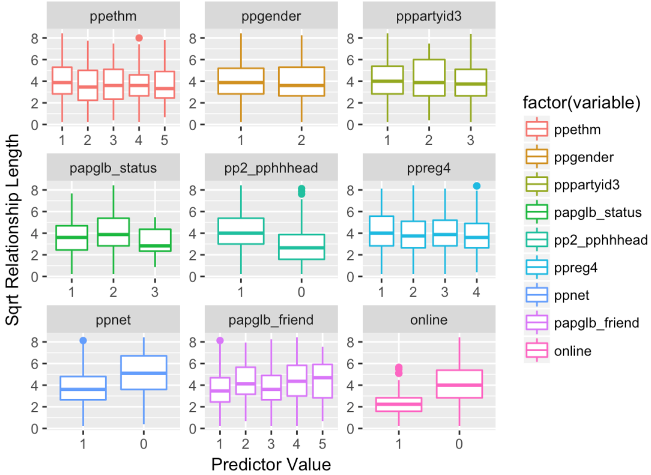


where duration is much more normally distributed. Everything else looks good.

4 Relationships with Response

Let's see how each of our predictors is associated with the response.

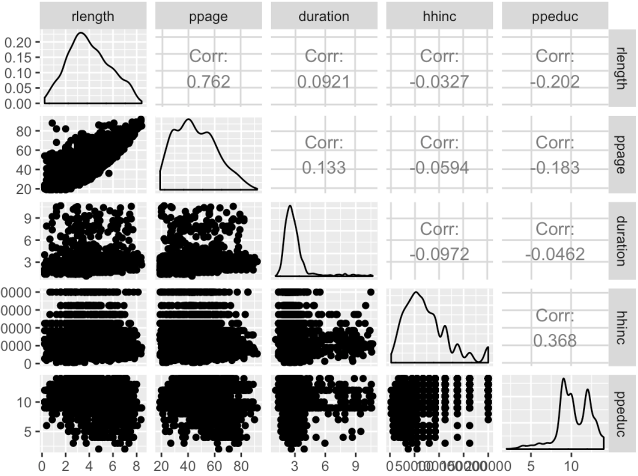




Plotting each of the predictors against the outcome variable, sqrt relationship length, shows no obvious non-linear patterns. There is a strong correlation (0.762) between age and relationship length, which makes sense, because your relationship length by definition cannot exceed your age. All other relationships are a bit more unclear, but there are no signs of non-linearity. The boxplot in the bottom-right, which shows the relationship between meeting online and sqrt relationship length, looks promising.

5 Collinearity

There doesn't seem to be much collinearity between the quantitative predictors:



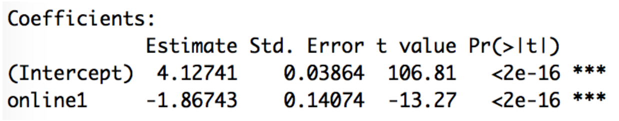
Collinearity between the categorical variables is harder to analyze, but is also much less of a problem. We chose not to include it in this paper because we can safely ignore the collinearity of categorical variables with multiple categories (Allison). A larger 14x14 plot of all variable relationships can be referenced from the appendix if necessary.

Now onto the regression to actually quantify these associations.

6 Analyzing Meeting Online

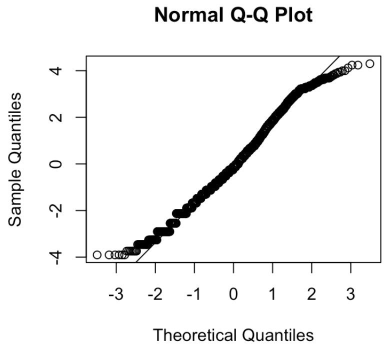
6.1 Model 1: Association between Meeting Online and Relationship Length

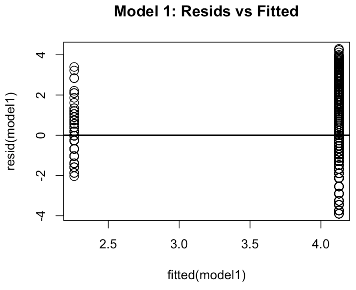
Let's start with the simplest model, which is just an Ordinary Least Squares regression with 1 predictor: online. We call this **Model 1**. This yields the following coefficients:



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.12741    0.03864  106.81   <2e-16 ***
online1     -1.86743    0.14074  -13.27   <2e-16 ***
```

Let's check our assumptions. Below are the residual Q-Q plot and the residual vs fitted scatterplot for model 1.



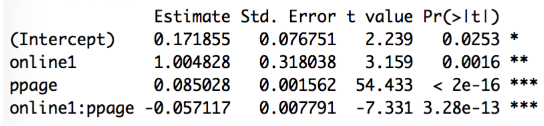


We check the independence assumption based on sampling method, and from the methodology there is no reason to suspect dependence between the observations since one romantic relationship should not affect another. The Normality assumption is satisfied, because the distribution of residuals is normally distributed, as corroborated by the above histogram and Q-Q plots. The linearity assumption is satisfied, as shown in the scatterplot of residuals versus fitted values being centered around the line. The constant variance assumption may potentially be violated, due to the seemingly increased spread on the right in that residuals versus fitted scatterplot. However, the rule of thumb is that we are fine if the maximum variance is ≤ 4x the minimum variance, or in some books, ≤ 3x the minimum variance (Ford, Van Belle). The residual variances here satisfy both thresholds, with a minimum variance of 1.023 and maximum of 2.985. The code for this calculation is in the appendix.

This model tells us that without controlling for other factors, meeting online is associated with an average decrease in sqrt relationship length of about 1.87.

6.2 Model 1b: Adding in Age

As shown in the sections 4 and 5 of this paper, there is a strong positive correlation between age and relationship length. But since widespread access to the internet is a relatively recent phenomenon, it could be that our seemingly negative association between meeting online and relationship length has more to do with the fact that meeting online is associated with youth, and younger people tend to have shorter relationship lengths. To get a better understanding of this effect, we made a modification to model 1, calling it **Model 1b**, which included the predictors online and ppage, as well as their interaction term. This yielded the following coefficients:



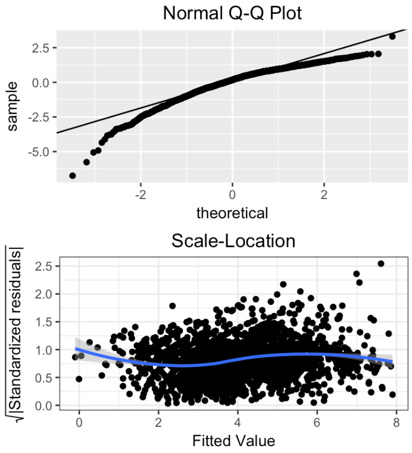
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.171855   0.076751   2.239   0.0253 *
online1        1.004828   0.318038   3.159   0.0016 **
ppage          0.085028   0.001562  54.433  < 2e-16 ***
online1:ppage -0.057117   0.007791  -7.331 3.28e-13 ***
```

Let's interpret the model in the context of meeting online. The coefficient for online is positive, which one naively may think suggests that meeting online is now associated with an increase in relationship length. However, we should take into account the interaction term between age and online, which indicates that the effect of meeting online on relationship length varies by age. Here, with our negative interaction term, we see that meeting online still translates into an overall negative association with relationship length for anyone older than 17.55.

6.3 Model 2: Full Model of Main Effects

Now let's examine the association between meeting online and relationship while controlling for all other factors. To do this, we constructed **Model 2**, which includes the main effects of all the predictors we described in the introduction of this paper. We excluded the full coefficient table from this write up due to length constraints, but it can be found in the appendix. Controlling for age, duration of interview, household income, education level, ethnicity, gender, political party, LGBTQ status, LGBTQ friends or family, household head status, geographic region, and internet access at home, meeting via an online platform was associated with an average decrease in sqrt relationship length of 1.036. We will analyze this coefficient in a moment, but first, let's check our assumptions.

Below are some summary diagnostic plots for this model:



We check the independence assumption based on sampling method, and from the methodology there is no reason to suspect dependence between the observations since one couple should not affect another. The Normality assumption is satisfied, because the distribution of residuals is fairly normally distributed, as corroborated by the Q-Q plot. The linearity assumption is satisfied, as shown in the scatterplot of residuals versus fitted values being centered around the line. Finally, the constant variance assumption is satisfied, because there is no evidence of heteroskedasticity in Scale-Location plot.

We are curious about whether the full Model 2 is significantly better than a reduced Model 2 without the online variable. Let's conduct a hypothesis test to see if the coefficient for online is significant enough to add to a model that already incorporates the other predictors. We start with our hypotheses: $H_0 : \beta_{online} = 0$, and $H_a : \beta_{online} \neq 0$. We can use an Extra-Sum-of-Squares F-test to determine if the extra predictor (online) is worthwhile. To calculate F, we have

$$F = \frac{(SSR_{Reduced} - SSR_{Full})/(df_{Reduced} - df_{Full})}{SSR_{Full}/df_{Full}}.$$

This was calculated in R, with code in the appendix. We have $F = 118.37$ and $p < 0.001$, which is below our threshold of $\alpha = 0.05$, so we reject the null hypothesis. There is evidence of a nonzero coefficient between meeting online and relationship length. In other words, it is worthwhile to include the online variable in our model even after including all other main effects.

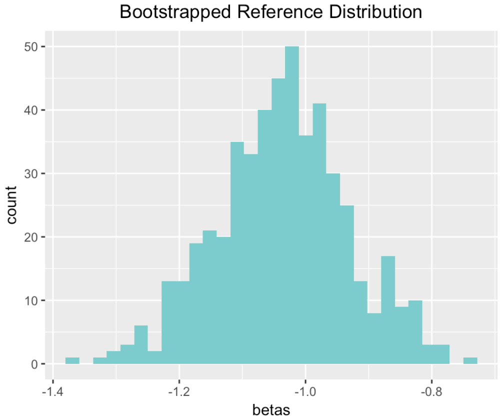
6.3.1 Distribution of β_{online}

Under our assumptions for the linear model, β_{online} is normally distributed. We can construct a 95% confidence interval for β_{online} by looking at our estimator for β_{online} and its standard error. This gets us

$$CI : \hat{\beta}_{online} \pm 1.96 \times \hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{[j+1,j+1]}}$$

$$CI : (-1.22267, -.8493)$$

Due to potential violation of the normality assumption, we also constructed bootstrapped confidence interval estimates, which are robust to deviation from an expected reference distribution. We resampled observations 500 times and fit a model to each iteration, recording the $\hat{\beta}$. This yielded the following reference distribution, with histogram below:

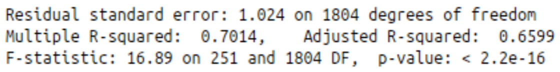


The boostrapped confidence interval for β is $(-1.246, -0.832)$. This interval does not include 0, so it is consistent with the association between meeting online and relationship length being negative.

7 Extension: Broader Analysis of Relationship Length

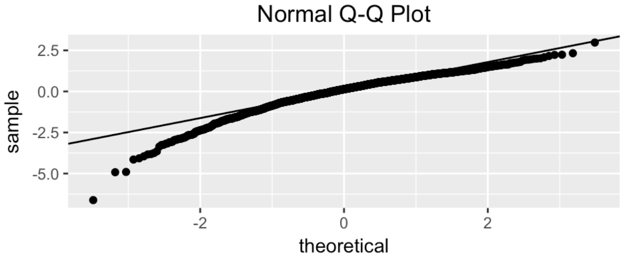
7.1 Model 3: Full Model with Interactions

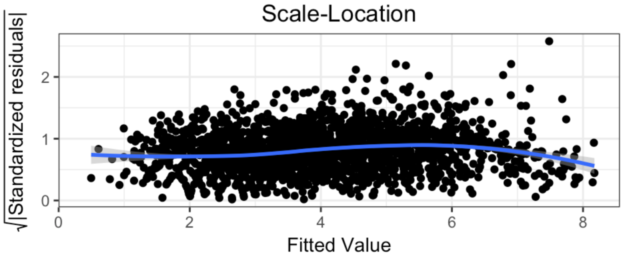
We will start building our best predictive model of relationship length by fitting a model that includes all of our possible predictors and all of their interaction terms, which we call **Model 3**. This model has 371 different coefficients, many of which are not statistically insignificant, and may be highly collinear. This model is unlikely to be a particularly effective one, and is prone towards overfitting. Again, we have excluded the full coefficient table from this write up due to length constraints, but it can be found in the associated appendix.



```
Residual standard error: 1.024 on 1804 degrees of freedom
Multiple R-squared:  0.7014,    Adjusted R-squared:  0.6599
F-statistic: 16.89 on 251 and 1804 DF,  p-value: < 2.2e-16
```

This model will be a useful point of comparison in the cross-validation section. Before continuing, let's check our assumptions.

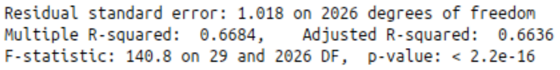




The independence assumptions is satisfied due to the sampling method, as explained before. The residuals are generally normally distributed, with some skewing of the left tail as shown in the Q-Q plot. Linear regression is somewhat robust to normality violations in the residuals so this should be fine given our very large sample size. The Scale-Location plot shows roughly constant variance and no clear patterns that might indicate non-linearity.

7.2 Model 4: Stepwise Model Selection

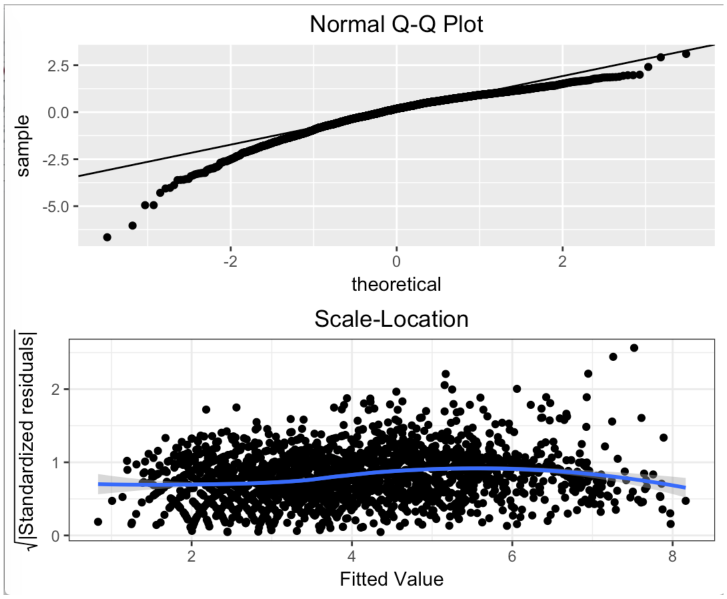
Now let's fit a model, called **Model 4**, using stepwise variable selection in both directions. Starting with model 2, our stepwise procedure adds or removes the variable that best improves AIC. The scope of the stepwise procedure is from the intercept-only model on the lower end to the full model with interactions (model 3).



```
Residual standard error: 1.018 on 2026 degrees of freedom
Multiple R-squared:  0.6684,    Adjusted R-squared:  0.6636
F-statistic: 140.8 on 29 and 2026 DF,  p-value: < 2.2e-16
```

Comparing the R^2 of model 3 and model 4, we see that model 3, the full model, has a higher R^2, suggesting that more of the variation in the response is explained by model 3. It's important to not fall into the trap of assuming that the higher R^2 value means that model 3 is actually the better predictive model. Instead, this illustrates the problem of using R^2 in model selection: there is no penalty for overfitting/inclusion of extra terms. Model 3 will always have the lower R^2 because it has more predictors. We will see in cross-validation how that isn't always a good thing.

Before we continue to model comparison, let's check our assumptions for this model.



Again, we satisfy independence due to sampling method. The residuals are very close to normally distributed save for a touch of curvature at the left tail, as shown in the Q-Q plot, but nothing to worry about given our large sample size and robustness of regression to slight non-normality. The residuals are linearly distributed around the line of fit, satisfying the linearity assumption, and the variance is relatively constant, as shown in the Scale-Location plot.

7.3 Initial Cross-Validation

We now perform cross-validation for each model to determine which model is best for predicting new relationship lengths. We fit the models on a training data set consisting of a random, without replacement, sample of $\frac{3}{4}$ of the total observations.

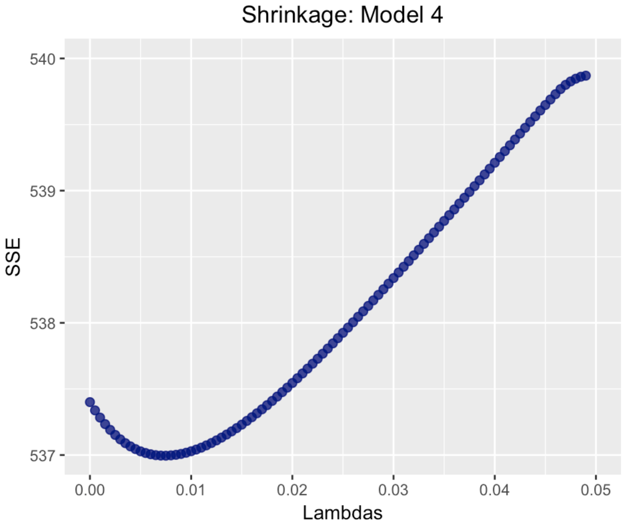
We then record the square sum of residuals when each model is fit on the withheld test data. Doing so yields the following mean residuals:

Model	SSR
1	1457.701
2	565.3533
3	16834.27
4	537.5159

As we can see (and as expected), model 3 (the full model) is the worst model, substantially overfitting the training data and yielding a much higher mean SSR over the test data. The first model performs quite well despite having just one predictor (meeting online), with a residual of 1457. As expected, model 4, the model created using the stepwise procedure, performs the best out of all four with a SSR of ∼537. The high residual from model 3 may be a product of the fact that when fitting on the training values, some of the coefficient values for indicator interactions are NA, due to lack of data that matches those specific terms. Here, it appears that there are no Hispanic respondents who refused to disclose their sexuality, yielding an NA for that specific interaction term. The *lm* package does not properly handle the NA value when predicting, yielding large residuals. Removing the sexuality predictor from the model reduces SSR to 658, still worse than model 2 and model 4, but substantially better than before.

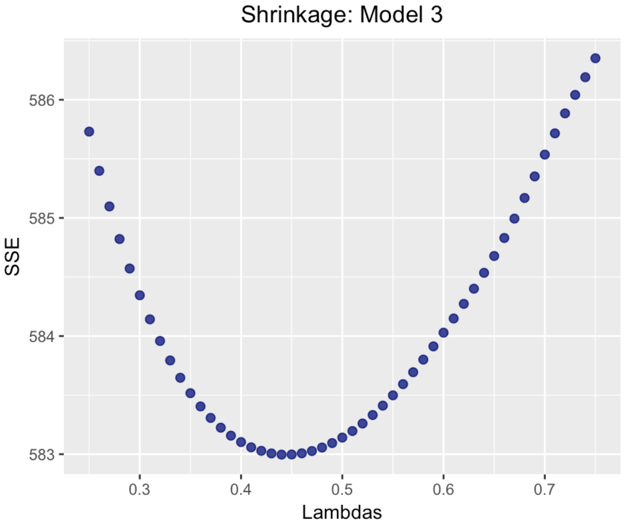
7.3.1 Shrinkage with Ridge

Now, we explore shrinkage methods, specifically the **ridge** method. Generally, performing shrinkage methods on stepwise models isn't super useful (gains in SSR tend to be marginal at best). This plot demonstrates the marginal gains in SSR from the ridge regression on model 4.



As λ increases, the curve arcs slightly down, marking a small decrease in SSR, and then increases, reaching a minimum at a λ of 0.0075.

We also perform a ridge regression on the full model with interactions, model 3. We use cross-validation to determine the optimal shrinkage parameter λ to minimize SSR. Model 3 performed very badly in the first CV, but with ridge shrinkage it performs substantially better, although still subpar to model 2 and model 4. The new SSR of 583 with a lambda around 0.45 is much more in line with the result of model 2 and model 4. However, it is worth noting that *glmnet* is likely not completing convergence to the correct estimate (given that there are so many betas). When $\lambda = 0$, we would expect SSR to be the same as original, but it is only 677. This could be caused by the *glmnet* package treating NA coefficient values differently than *lm*.



We see a curve for $SSE(\lambda)$ that reaches a minimum at ~ 583.

We can now update our table of best models.

Model	SSR	λ
1	1457.701	
2	565.3533	
3	16834.27	
4	537.5159	
Ridge of 3	582.9966	0.44
Ridge of 4	536.9958	0.0075

From the CV results, our best model is ridge shrinkage of model 4 (the stepwise model), although the gains in SSR are only marginal over the original model 4 values.

8 Limitations and Challenges

There are a few limitations to our analysis that are important to consider when drawing broader conclusions. First, the response variable (relationship length) is measured by asking people in the survey who are currently in relationships how long they have been romantically involved with their current partner. This would tend to undersample shorter relationships, since people are in them for shorter periods of time and they are less likely to fall in the surveyed group by chance. Additionally, this would tend to underestimate the length of relationships in general, because except for people who broke up right after this survey, the total length of the relationship would be longer than the length stated at the time of the survey. This could be addressed by asking people about previous relationships, for example, but the data we used measured how long people in relationships had been in them, so this is what we had to work with. Second, although people asked to respond to this survey were randomly selected, this was a voluntary survey, so its generalizability to the larger population is limited by the fact that it was not a truly random sample of the American population. Third, because subjects were not randomly assigned to the meet online versus meet in real life groups, we cannot determine any causality from this analysis, only association. Fourth, the longest possible relationship with online dating is limited by how long meeting online has been possible. Since online dating has been around for less than thirty years, it should correlate with shorter relationships simply for that reason, and controlling for age alone is insufficient to adjust for this effect. It could be that the negative effect on relationship length is way less pronounced than indicated by the data, simply because we have not had online dating for long enough to see many really long, stable relationships born out of it.

9 Conclusion

We found that there is a significant association between meeting online and relationship length, specifically, a negative association between meeting online and relationship length. We found that this association was significant even when controlling for other predictors. We think this decrease in relationship length is interesting to consider given the current prevalence of online matchmaking and dating, but a randomized trial in the future is required before we can draw any causative conclusions. Additionally, we created a best predictive model for relationship length, which also shed light on what factors were most associated with relationship length, which could be useful in further analyses on relationship length.

References

Allison, Paul. (2012). When Can You Safely Ignore Multicollinearity? Retrieved from: https://statisticalhorizons.com/multicollinearity

Faraway, J. (2014). Linear models with R, Second Edition (Texts in statistical science; v. 63). Boca Raton: Chapman & Hall/CRC. 200.

Ford, C. (2013). A Rule of Thumb for Unequal Variances. University of Virgina Library Research Data Services. Retrieved from: http://data.library.virginia.edu/a-rule-of-thumb-for-unequal-variances/

Kalyanaraman, S. (2007). Online Relationships. 636-638.

Lilien, J. (2016). Of Love and Money: The Rise of the Online Dating Industry. Nasdaq. Retrieved from: http://www.nasdaq.com/article/of-love-and-money-the-rise-of-the-online-dating-industry-cm579616

Pew Research Center. (2016). 15% of American Adults Have Used Online Dating Sites or Mobile Dating Apps. Retrieved from: http://scholar.aci.info/view/14bd17773a1000 e0009/152d1dd7ec400014c11.

Slater, D. (2013). Love in the time of algorithms: What technology does to meeting and mating. New York: Current.

Van Belle, G. (2008). Statistical rules of thumb (2nd ed.). Hoboken, N.J.: Wiley.

Wolak, Mitchell, & Finkelhor. (2003). Escaping or connecting? Characteristics of youth who form close online relationships. *Journal of Adolescence*, 26(1), 105-119.