# Divorce Rate Prediction

Yixuan Xie, Sean Liao, Zhoutian Xu

December 2023

**Abstract**

In this project, we utilized the Marriage and Divorce Dataset[1] to develop a linear regression model, aiming to accurately predict the rate of divorce per one thousand people. By utilizing $sin^{-1}\sqrt{Y}$ transformation on the response variable, followed by variable selection, we revealed that education, income, love, the desire for marriage, and the age of starting to socialize with the opposite sex are most deterministic for divorce rate. The model is then refined by adding polynomial terms and results in a $R^2$ of 0.2328, indicating that the model is not highly accurate. Possible limitations of our model are the use of averaged data, oversimplification of love, and the limited amount of data.

## 1 Introduction

A successful marriage fills the later chapters of our lives with enduring happiness. Therefore, our group plans to use a multi-linear regression model to estimate the divorce rate per 1000 people, which functions as a measurement of the success of marriages. Consequently, the result can not only be used as a pre-marriage counselor that helps prevent high-risk marriages but also helps us understand the underlying factors that affect our intimate relationships.

## 2 Dataset

The Dataset being used is the Marriage and Divorce Dataset[1] provided by Kaggle. This data contains 100 observations with 31 variables including, Age Gap, Education, Economic Similarity, Good Income, Love, Desire to Marry, Start Socializing with the Opposite Sex Age, etc, and the Divorce Probability — rate of divorce per 1000 people — as the response variable.

## 3 Methods

Our group used two methods to construct models, and the final model was obtained by comparing and combining the two models.

### 3.1 Model A

For the first approach, we utilized the backward selection($AIC$) on the ordinary least square regression and obtained. After that, we examined the diagnostic plot and utilized the Box-Cox transformation to address the issue of heteroscedasticity. However, $\lambda \approx 1$ indicated that the issue is light and we consequently tried both the polynomial regression and adding interaction terms which resulted as our first model, Model A, as shown in Figure 1.

### 3.2 Model B

In the second approach, we decided to transform the response variable to address the issue of heteroscedasticity, which we found in the first approach. The $sin^{-1}(\sqrt{Y})$ transformation was used after noticing that the rate of divorce follows a binomial distribution. It is quite surprising that after the variable selection, the same factors were obtained compared to the first approach. Model B, as shown in Figure 2, that we obtain has a more flat line in the Scale-Location plot, as shown in Figure 5.

### 3.3 Final Model

Since Models A and B have the same predictors, we decided to combine the two together as our final model. In specific, we used the $sin^{-1}(\sqrt{Y})$ transformation to the response variable, and then applied the polynomial regression after variable selection. Eventually, we obtained our final model, as shown in Figure 3.

## 4 Results

The final model, as shown in Figure 3, includes six predictors: $Education$, $Good\ Income$, $Love$, $Desire\ to\ marry$, and the first and second terms of $Socializing\ Age\ with\ Opposite\ Sex$.

From the model, it is evident that $Education$ and $Good\ Income$ negatively impact the divorce rate, with coefficients of -0.101577 and -0.002545, respectively. In contrast, $Love$ shows a positive effect on the divorce rate (coefficient 0.003361), aligning with the adage ”Marriage is the grave of love”. This suggests that love alone might not be a sufficient foundation for a lasting marriage.

Moreover, $Desire\ to\ marry$ exhibits a negative coefficient (-0.001990), indicating that a stronger desire to marry correlates with a lower divorce rate, presumably because those more eager to marry are less likely to divorce later.

An intriguing aspect is the impact of $Socializing\ Age\ with\ Opposite\ Sex$. The first-order term's negative coefficient (-0.798453) suggests that starting romantic relationships too early may not favor marital success. On the other hand, the second-order term's positive coefficient (0.427489) hints that delaying social interactions with the opposite sex excessively might also contribute to marital challenges.

# 5 Limitations

There are several limitations of our model, including, Averaging Data in Regression, Oversimplification of Love in Predictive Modeling, and Potential for Overfitting/Limited Sample Size.

## 5.1 Averaging Data in Regression

The current linear regression model relies on average data, which trades a group of data as an individual, to predict the divorce rate per thousand people. Averaging data can lead to a loss of variability and individual differences within the dataset, resulting in a model that may not sensitively reflect the nuances of individual cases. This approach can also underestimate the true variability and error, possibly leading to an overestimated fit of the model. Additionally, the reduction in the number of data points due to averaging can weaken the model's robustness and potentially lower the $R^2$ value, indicating a less effective explanation of the variance in the dependent variable.

## 5.2 Oversimplification of Love in Predictive Modeling

The data represents love as a singular numerical value to depict spousal relationships, potentially oversimplifying the intricate, multi-faceted nature of love. To more accurately capture the essence of intimate relationships, nuanced and detailed questions such as "Do we share common goals?" and "Am I fully aware of my spouse's preferences?" should be incorporated. This approach would enable the model to better comprehend the subtleties of individual relationships, thereby enhancing its capacity to offer personalized insights into the likelihood of divorce.

## 5.3 Potential for Overfitting/Limited Sample Size

The model may be susceptible to overfitting, particularly when a large amount of variables are included. Given the limited size of the raw dataset (100 observations), such overfitting could impair its effectiveness on new data, reducing its real-world predictive accuracy for the rate of divorce. Therefore, it is important to supplement the model with more data for future improvement of the model.

# 6 Contribution

The individual contributions for the project are the following:
**Sean Liao:**

- Searched and selected the raw dataset for our model.
- Performed forward selection on the full model to select appropriate predictors.
- Contributed to the introduction, contribution, and reference sections of the final report.

**Yixuan Xie:**

- Performed linear transformation and $\sin^{-1}(\sqrt{Y})$ transformation on the response to scale it within the range of [0, 1] and mitigate the issue of heteroscedasticity.

- Utilized backward selection on the full model to choose the optimal predictors.
- Fitted the final model and analyzed it.
- Contributed to the method, result, and conclusion sections of the final report.

**Zhoutian Xu:**

- Performed Box-Cox transformation on the response to solve the issue of heteroskedasticity.
- Performed polynomial transformation on the predictors and added interaction terms to improve the fit.
- Proposed ideas regarding how to interpret the final model and coefficients.
- Contributed to the method, result, and limitation sections of the report.

# 7 Conclusion

In conclusion, our group utilized multi-linear regression to analyze the factors influencing the divorce rate, aiming to create a comprehensive pre-marriage counselor that reduces the number of unsuccessful marriages. In our modeling approach, we first applied the $sin^{-1}(\sqrt{Y})$ transformation to the response variable, and then identified key predictors using variable selection. Finally, we construct a comprehensive model that predicts the divorce rate using education, income, love, desire to marry, and the age of socializing with the opposite sex.

Our final model has several limitations that reduce its accuracy. Firstly, the reliance on average data curtails the model's ability to capture individual variability and differences. Additionally, the representation of love as a singular numerical value oversimplifies a complex emotion, potentially skewing the results. Furthermore, the model has many variables but a limited dataset size, raising concerns about overfitting, which further compromises the model's accuracy and predictive reliability.

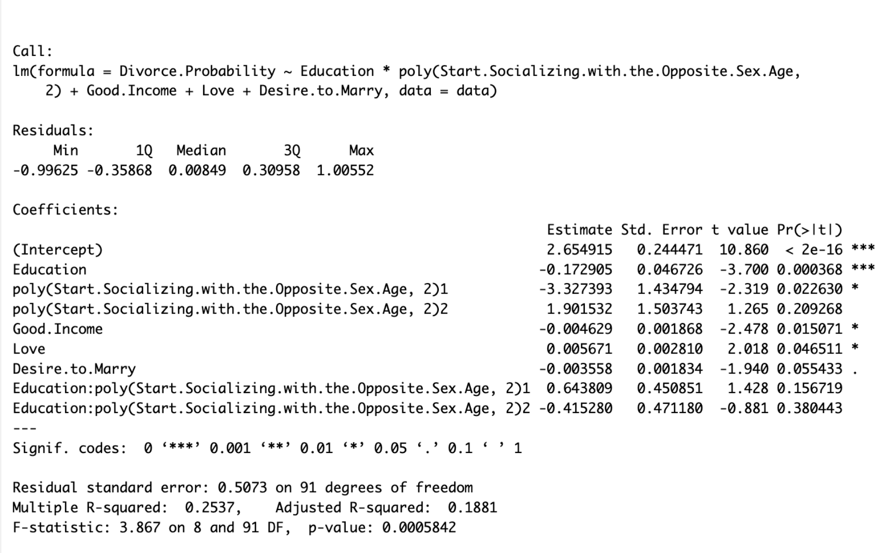
In further study, we will address the current limitations and utilize more detailed and personalized data. We believe that our model will not only provide a deeper understanding of the complexities of modern marital relationships but also have the potential to prevent unsuccessful marriages, contributing to the overall happiness of the population.

# References

[1] Seyed Muhammad Hossein Mousavi. (2022). *Marriage and Divorce Dataset.* Kaggle. Available at: `https://doi.org/10.34740/KAGGLE/DSV/4066976`

# A Appendix

## A.1 Additional Figures and Tables



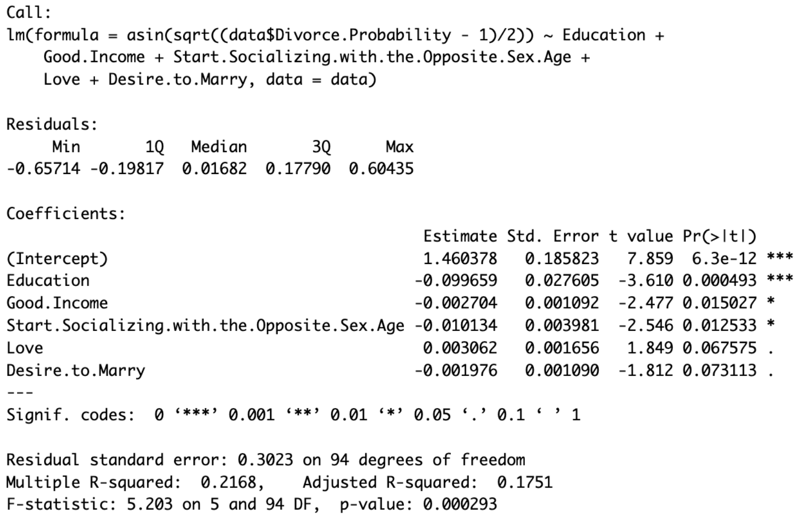
```
Call:
lm(formula = Divorce.Probability ~ Education * poly(Start.Socializing.with.the.Opposite.Sex.Age, 
    2) + Good.Income + Love + Desire.to.Marry, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.99625 -0.35868  0.00849  0.30958  1.00552 

Coefficients:
                                                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                                   2.654915   0.244471  10.860  < 2e-16 ***
Education                                                    -0.172905   0.046726  -3.700 0.000368 ***
poly(Start.Socializing.with.the.Opposite.Sex.Age, 2)1        -3.327393   1.434794  -2.319 0.022630 *  
poly(Start.Socializing.with.the.Opposite.Sex.Age, 2)2         1.901532   1.503743   1.265 0.209268    
Good.Income                                                  -0.004629   0.001868  -2.478 0.015071 *  
Love                                                          0.005671   0.002810   2.018 0.046511 *  
Desire.to.Marry                                              -0.003558   0.001834  -1.940 0.055433 .  
Education:poly(Start.Socializing.with.the.Opposite.Sex.Age, 2)1  0.643809   0.450851   1.428 0.156719    
Education:poly(Start.Socializing.with.the.Opposite.Sex.Age, 2)2 -0.415280   0.471180  -0.881 0.380443    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5073 on 91 degrees of freedom
Multiple R-squared:  0.2537,	Adjusted R-squared:  0.1881 
F-statistic: 3.867 on 8 and 91 DF,  p-value: 0.0005842
```

Figure 1: Model A



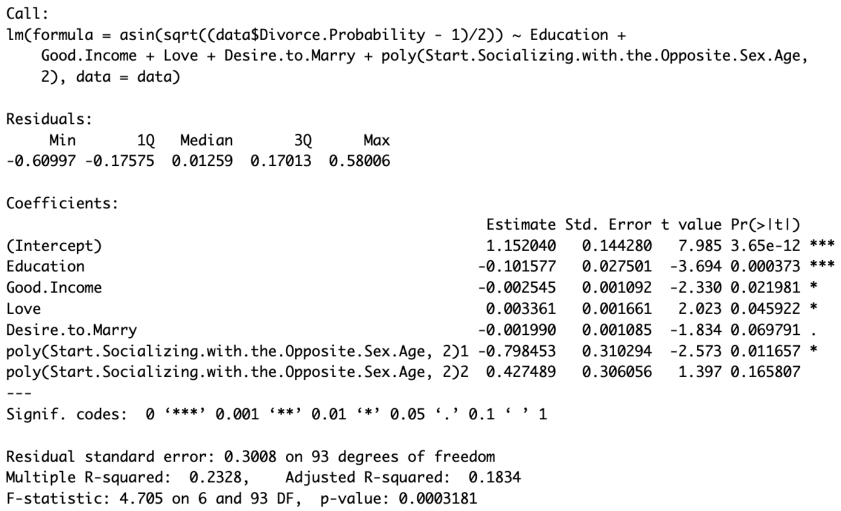
```
Call:
lm(formula = asin(sqrt((data$Divorce.Probability - 1)/2)) ~ Education + 
    Good.Income + Start.Socializing.with.the.Opposite.Sex.Age + 
    Love + Desire.to.Marry, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.65714 -0.19817  0.01682  0.17790  0.60435 

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                  1.460378   0.185823   7.859  6.3e-12 ***
Education                                   -0.099659   0.027605  -3.610 0.000493 ***
Good.Income                                 -0.002704   0.001092  -2.477 0.015027 *  
Start.Socializing.with.the.Opposite.Sex.Age -0.010134   0.003981  -2.546 0.012533 *  
Love                                         0.003062   0.001656   1.849 0.067575 .  
Desire.to.Marry                             -0.001976   0.001090  -1.812 0.073113 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3023 on 94 degrees of freedom
Multiple R-squared:  0.2168,	Adjusted R-squared:  0.1751 
F-statistic: 5.203 on 5 and 94 DF,  p-value: 0.000293
```

Figure 2: Model B



```
Call:
lm(formula = asin(sqrt((data$Divorce.Probability - 1)/2)) ~ Education +
    Good.Income + Love + Desire.to.Marry + poly(Start.Socializing.with.the.Opposite.Sex.Age,
    2), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60997 -0.17575  0.01259  0.17013  0.58006

Coefficients:
                                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                                          1.152040   0.144280   7.985 3.65e-12 ***
Education                                           -0.101577   0.027501  -3.694 0.000373 ***
Good.Income                                         -0.002545   0.001092  -2.330 0.021981 *
Love                                                 0.003361   0.001661   2.023 0.045922 *
Desire.to.Marry                                     -0.001990   0.001085  -1.834 0.069791 .
poly(Start.Socializing.with.the.Opposite.Sex.Age, 2)1 -0.798453   0.310294  -2.573 0.011657 *
poly(Start.Socializing.with.the.Opposite.Sex.Age, 2)2  0.427489   0.306056   1.397 0.165807
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3008 on 93 degrees of freedom
Multiple R-squared:  0.2328,    Adjusted R-squared:  0.1834
F-statistic: 4.705 on 6 and 93 DF,  p-value: 0.0003181
```

Figure 3: Final Model

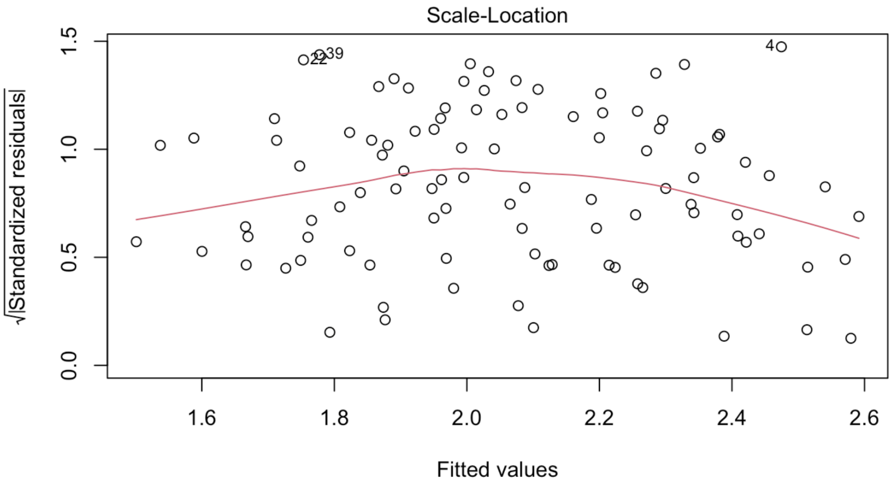


Figure 4: Scale-Location for Model A

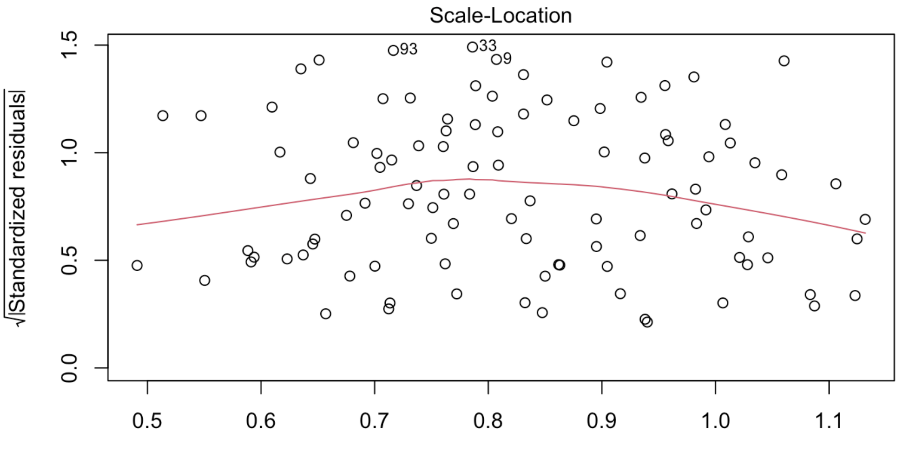


Figure 5: Scale-Location for Model B