A Linear Model for Divorce Rates in the 20^{th} Century

Aidan Dykstal, Edward Hammond, & Zack Hart

December 9^{th}, 2019

1 Introduction

In this report, we examine the "divusa" data set, which originates from the `faraway` package in `R`. We choose to consider these data, which track the divorce rate in the United States between 1920 and 1996, so that we can quantify the event of divorce in the face of social ramifications and the historical context of the 20^{th} Century. This data set contains 77 observations of the response and each of the seven covariates. These observations occur once per year between 1920 and 1996, and provide information regarding:

- The response, which is the number of divorces per 1000 women aged 15 or higher.
- The year, ranging from 1920 to 1996.
- The unemployment rate as a percentage.
- The percent female participation in the labor force aged 16 or higher.
- The number of marriages per 1000 women aged 16 or higher.
- The number of births per 1000 women aged 15 to 44.
- The number of military personnel per 1000 members of the United States population.
- The time period, which is either early or late 20^{th} Century. Here, the early 20^{th} Century spans 1920 to 1949 and the late 20^{th} Century spans 1950 to 1996.

Our primary research objective focuses on identifying and interpreting a well-fitting linear model of the divorce rate as a function of the other variables in the data set, including their transformations and interactions. While we seek to elevate the goodness-of-fit in this model, we also strive to prevent "overfitting" by employing train-test splits to the recorded data.

Considering that the "divusa" data set only consists of 77 observations, the computational intensity of best subsets regression does not significantly hinder our model selection and development. As such, we employ the method of best subsets using a simple iterative scheme in an attempt to produce a model with the most predictive power for estimating the divorce rate in the United States between 1920 and 1996. In doing so, we select a model by optimizing the mean-square prediction error (MSPE) on a series of randomly-selected training data batches. We then check our various model assumptions and analyze the results.

2 The Multiple Regression Approach

2.1 Assumptions

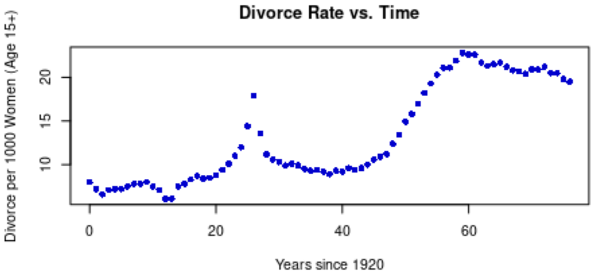
To begin our exploration of the available data, we assume that the covariates can suitably model the response - namely, that the divorce rate can be explained by mathematical manipulations of the covariates (described in Section 1). To check this assumption, we will fit linear models to the divorce rate data and check for significance of the parameters corresponding to some or all of the explanatory variables.

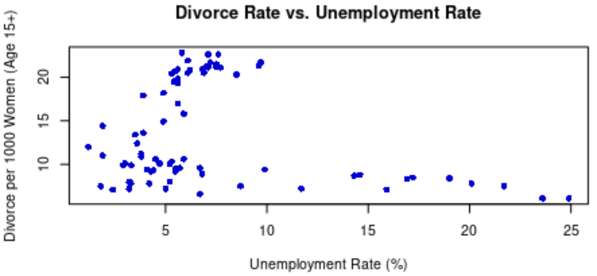
We also assume that the technical machinery involved in linear modeling is sufficient for showing the relationships we suspect. This means we do not consider other potential methods for modeling (such as difference equations, systems of partial differential equations, etc.). Luckily, linear modeling tends to be quite robust, so it is plausible that we can create a convincing representation of the divorce rate in terms of the available explanatory variables. Of course, we will contextualize this assumption when we fit a final linear model to the data.

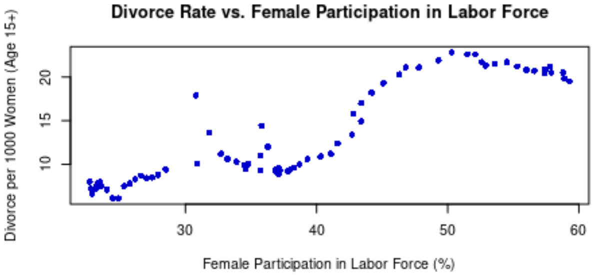
The linear model is $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$ is the response, X is the matrix whose columns are the observed values of the explanatory variables, and $\boldsymbol{\epsilon}$ is the unobservable random error in the response as a function of the covariates. In our case, there are 77 observations in $\boldsymbol{y}$ and X since data is recorded once per year between 1920 and 1996. For each of these observations, we assume that the elements of $\boldsymbol{\epsilon}$ are independent and identically distributed according to the normal distribution with mean 0 and unknown variance σ^2. That is, an observation error for the divorce rate during a particular year in the United States does not influence any other observation error - we will check this assumption rigorously in Section 3 of this report.

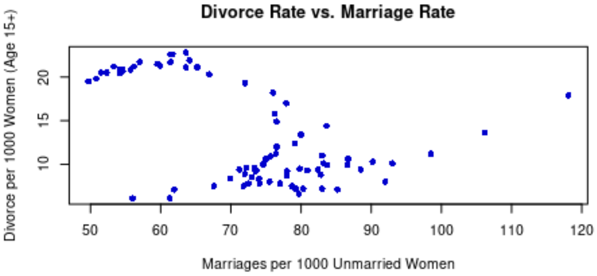
2.2 Model Selection Methods

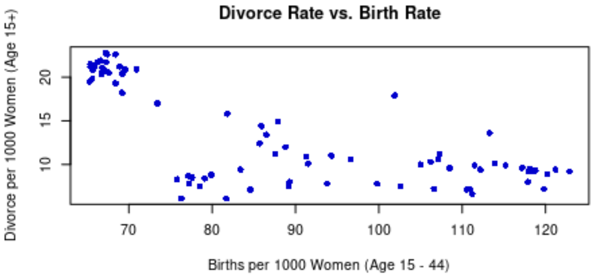
Since out data set is relatively small with a small number of candidate independent variables, the computational expense of the best subsets algorithm is negligible. As such, we employ this method for selecting our model so that we can compare the goodness-of-fit for models with every number of variables betweeen 1 and p, where p is the total number of candidate independent variables considered (excluding the intercept). To identify these possible explanatory variables, including their potentially relevant interactions and transformations, we view the response versus each covariate.

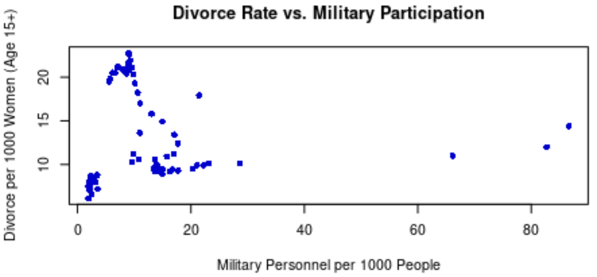












From these plots, we see clear relationships forming. In particular, the divorce rate versus the time (measured in years since 1920), the percent female participation in the labor force, and the marriage rate appear to exhibit roughly cubic behavior in the range of the data. In addition, the the divorce rate versus the birth rate seems to follow a slightly quadradic pattern in the range of the data while the divorce rate versus the unemployment rate seems to follow a weak linear association. Finally, the divorce rate appears to have a weak linear or logarithmic relationship with the military participation rate. These apparent relationships between the response and the covariates justify our choice of potential variable transformations. We will elaborate on our choice of variable transformations momentarily.

In addition, we consider the physical relationships in the data and their interpretations to ensure that our model is both logically and mathematically plausible. To us, it makes sense that evolving social conditions over the course of the 20th Century in the United States can help to explain the changing divorce rate over time. This results in two key model decisions:

i. The number of years since 1920 (with its relevant functional transformations, as described earlier) will be included as a candidate independent variable in the model to act as a proxy for other possible explanatory factors that are not present in our data set.

ii. The interactons between each covariate (other than the number of years since 1920) and the relative time period in the 20th Century will be included as candidate independent variables so that we can consider the changing effect of the recorded covariates with time. The relative time period is recorded as a categorical variable with the levels "early" (1920 - 1949) and "late" (1950 - 1996) so that the effect modifiers separate the periods of time prior to and immediately following the Second World War.

So, with these relationships and the Principle of Parsimony in mind, we consider the following main effects, transformations, and interactions for our linear model of the divorce rate in 20th Century America:

1. The number of years since 1920.
2. The squared number of years since 1920.
3. The cubed number of years since 1920.
4. The percent female participation in the United States labor force aged 16 or higher.
5. The squared percent female participation in the United States labor force aged 16 or higher.
6. The cubed percent female participation in the United States labor force aged 16 or higher.
7. The number of births per 1,000 women aged 15 to 44 in the United States.
8. The squared number of births per 1,000 women aged 15 to 44 in the United States.
9. The percent unemployment in the United States.
10. The number of marriages per 1,000 unmarried women aged 16 or higher in the United States.
11. The squared number of marriages per 1,000 unmarried women aged 16 or higher in the United States.
12. The cubed number of marriages per 1,000 unmarried women aged 16 or higher in the United States.
13. The number of military personnel per 1,000 members of the United States population.
14. The natural log of the number of military personnel per 1,000 members of the United States population.
15. The time period in the 20th Century with categories ”early” [1920 - 1949] and ”late” [1950 - 1996].
16. The product (interaction) of the unemployment rate and the relative time period in the United States.
17. The product of the percent female participation in the labor force and the relative time period in the United States.
18. The product of the marriage rate and the relative time period in the United States.
19. The product of the birth rate and the relative time period in the United States.
20. The product of the military participation rate and the relative time period in the United States.

Thus, we have $p = 20$ total candidate independent variables for our linear model, excluding the intercept.

With this established, we perform 1,000 rounds of best subsets regression with these candidate variables. In each round, the entire data set is randomly split into training (75% of the total data) and testing (25% of the total data) sets before the best subsets algorithm operates on the training batch, selecting models of size $1, 2, \dots, p$. At the conclusion of each round, the selected models with each number of variables are applied to the testing set, wherefrom the MSPE is calculated for that model. After the 1,000 rounds are complete, the result is a set of 1,000 MSPE values for each model size tested by best subsets (that is, a $1000 \times p$ matrix of MSPE values results). To ensure that we choose the model with the number of variables that minimizes the impact of overfitting, we compare the median MSPE for the models of every size considered by best subsets. In particular, the number of variables correponsing to the minimum MSPE, say p_{best}, is chosen to be the number of variables in the final model, provided that it is parsimonious. Then, we apply the best subsets method to the entire divorce data set and use the model selected with p_{best} variables as our final model.

This method yields the following median MSPE values for model evaluations on the testing data for models fit to the training data using each number variables between 1 and p, inclusive:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
10.079	10.160	9.300	8.717	7.633	6.680	6.946	6.591	6.616	6.414	6.413	6.323	6.361	6.425	6.405	6.451	6.465	6.390	6.416	6.417

Figure 1: Median MSPE Values for Each Number of Variables in Tested Linear Models.

As seen in Figure 1 above, the models with twelve variables have the minimum median MSPE in our experiment. As such, we select the best subset with 12 variables fit to the entire data frame. The variables in this subset (with the unemployment rate also included to remain parsimonious) construct the final model, provided they are parsimonious. The 12 variables featured in our final model of divorce rate appear below with their coefficients and standard errors. In addition, the test statistic and p-value for the null hypothesis that the coefficient is zero are included for each variable:

Variable	Coefficient Estimate	Standard Error	Test Statistic	p-Value
Intercept	145.3	27.66	5.255	1.87×10^{-6}
% Female Labor Force	-9.888	1.586	-6.235	4.21×10^{-8}
Squared % Female Labor Force	0.233	0.034	6.778	4.87×10^{-9}
Cubed % Female Labor Force	-0.002	0.0002	-7.031	1.77×10^{-9}
Marriage Rate	1.535	0.626	2.453	1.70×10^{-2}
Squared Marriage Rate	-0.020	0.007	-2.753	7.70×10^{-2}
Cubed Marriage Rate	0.00009	0.00003	3.272	1.74×10^{-3}
Birth Rate	-0.721	0.100	-7.208	8.65×10^{-10}
Squared Birth Rate	0.003	0.0005	6.264	3.75×10^{-8}
Military Participation Rate	-0.037	0.015	-2.531	1.39×10^{-2}
Early 20th Century	-40.43	9.894	-4.086	1.26×10^{-4}
Unemployment Rate	0.069	0.105	0.660	5.12×10^{-1}
Early × Unemployment	-0.174	0.104	-1.669	1.00×10^{-1}
Early × % Female Labor Force	1.244	0.288	4.320	5.63×10^{-5}

Figure 2: Summary of the Explanatory Variables in the Final Linear Model.

Also, it is important that we consider confidence intervals for these coefficients. A table of 95% confidence intervals for the values of the coefficients in our final linear model are given on the following page:

Variable	Coefficient Lower Bound	Coefficient Upper Bound
Intercept	90.068	200.612
% Female Labor Force	-13.057	-6.719
Squared % Female Labor Force	0.164	0.302
Cubed % Female Labor Force	-0.002	-0.001
Marriage Rate	0.284	2.785
Squared Marriage Rate	-0.035	-0.006
Cubed Marriage Rate	0.00003	0.00015
Birth Rate	-0.921	-0.521
Squared Birth Rate	0.002	0.004
Military Participation Rate	-0.067	-0.008
Early 20th Century	-60.201	-20.066
Unemployment Rate	-0.141	0.279
Early × Unemployment	-0.383	0.034
Early × % Female Labor Force	0.669	1.819

Figure 3: Confidence Intervals for the Coefficients in the Final Linear Model.

Because our iterative scheme with best subsets calculates 20,000 potential models before arriving at a final representation of the data, we omit the coefficients and p-values for intermediate models. After all, it would be a cruel injustice toward the dwindling rainforests if we were to append hundreds of additional pages of computer output which the reader would not usefully entertain.

2.3 Application to the United States Divorce Study

Interpreting the coefficients in our final model, we learn that the divorce rate in the United States tends to:

- decrease by roughly 9.8 divorces per 1,000 women for a 1% growth in the female participation in the United States labor force;
- decrease by roughly 0.7 divorces per 1,000 women for each additional birth per 1,000 women aged 15 to 44 in the United States.
- increase by roughly 0.007 divorces per 1,000 women for a 1% increase in the unemployment rate in the United States.
- increase by roughly 1.5 divorces per 1,000 women for each additional marriage per 1,000 unmarried women aged 16 or higher in the United States.
- decrease by roughly 0.004 divorces per 1,000 women for each additional member of the United States military per 1,000 members of the population.
- increase moving from the early to the late 20th Century.
- decrease by an additional amount of about 0.2 divorces per 1,000 women for a 1% increase in the unemployment rate between 1920 and 1949.
- increase by an additional amount of about 1.2 divorces per 1,000 women for a 1% growth in the female participation in the United States labor force between 1920 and 1949.

The transformations of these variables included in the final model adjust these metrics over the range of the observation space, but do not create significant deviations from these relationships between the response and the explanatory variables. So, we can use these observations about the final model to help explain the divorce rate in 20th Century America.

From these results, our linear model suggests a multi-part answer to our research question. On the net, our model implicates that the divorce rate tends to increase with time in the 20th Century. In addition, our model implies that – regardless of the relative time in the 20th Century – the United States divorce rate generally tends to increase as the number of marriages increases and tends to decrease as the number of births and the proportion of military participants increases. But, our final model also suggests a dynamic effect of the female participation in the labor force and the unemployment rate on the divorce rate over time. While this model suggests that divorce rates tend to decrease with increasing female participation in the American labor force for all times in the data, it amplifies this effect over time. Specifically, the impact of female participation in the labor force between 1920 and 1949 is roughly 77% of its impact between 1950 and 1996. Similarly, our model suggests that the unemployment rate only noticeably affects the divorce rate between 1920 and 1949. During this time period, the model communicates that elevated unemployment rates tend to be associated with a decreasing number of divorces per 1,000 women.

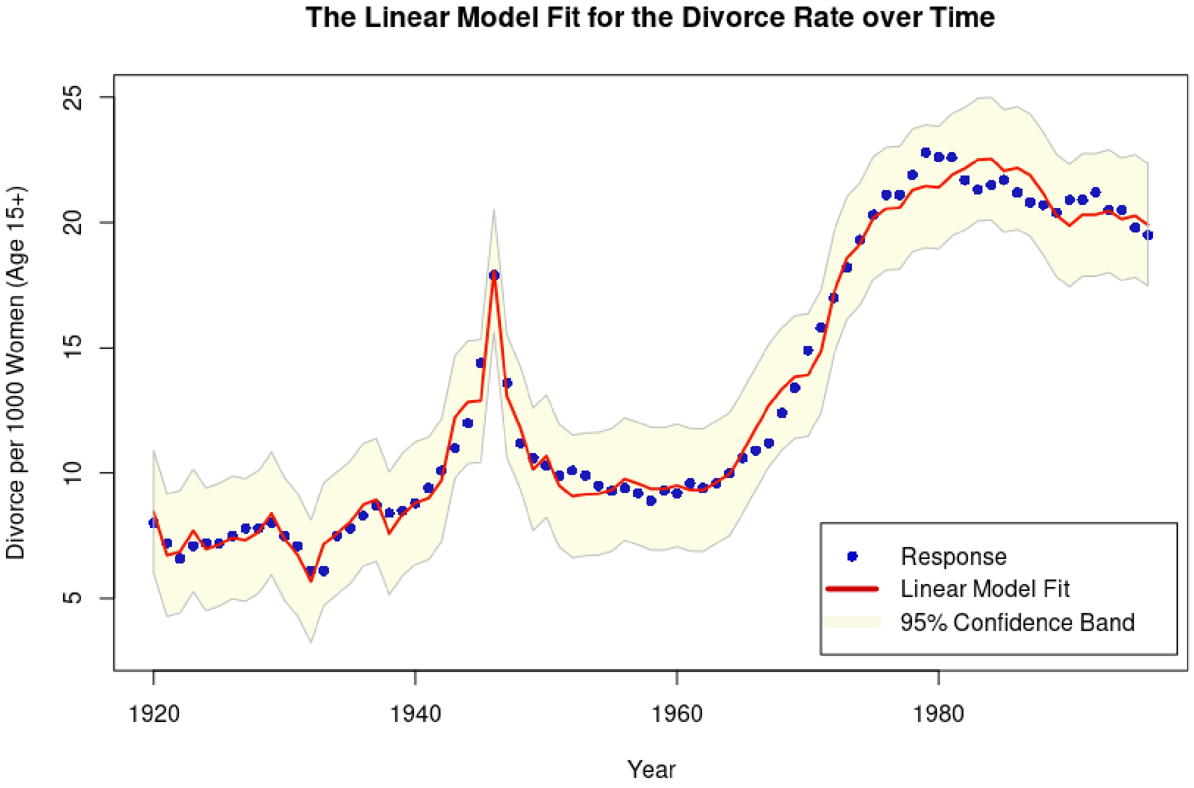
In addition, the 95% confidence intervals for the coefficients in our final model reveal that every effect described above is significant, except perhaps the impact of unemployment. This is because the only confidence intervals that capture zero are those that correspond to the effect of the unemployment rate and the interaction between time and unemployment. So, it possible, but not certain, that unemployment actually has no impact on the divorce rate in 20th Century America. Indeed, the other confidence intervals do not capture zero, so we conclude that every other relationship described above does occur, but the impacts of these relationships may vary in magnitude.

From a historical perspective, these model-based answers to our research question seem plausible. After all, we see that there are multiple factors that influence the divorce rate over time and that the impact of these factors can be modified with the changing social conditions reflected in the passage of time. It seems peculiar that the response is independent of the time since 1920 since there are likely variables outside of our data set that help to explain the dynamic divorce rate in the United States. We will reflect more on this potential shortcoming in Section 4 after conducting a detailed model analysis.

3 Model Analysis

3.1 Overview of the Final Linear Model

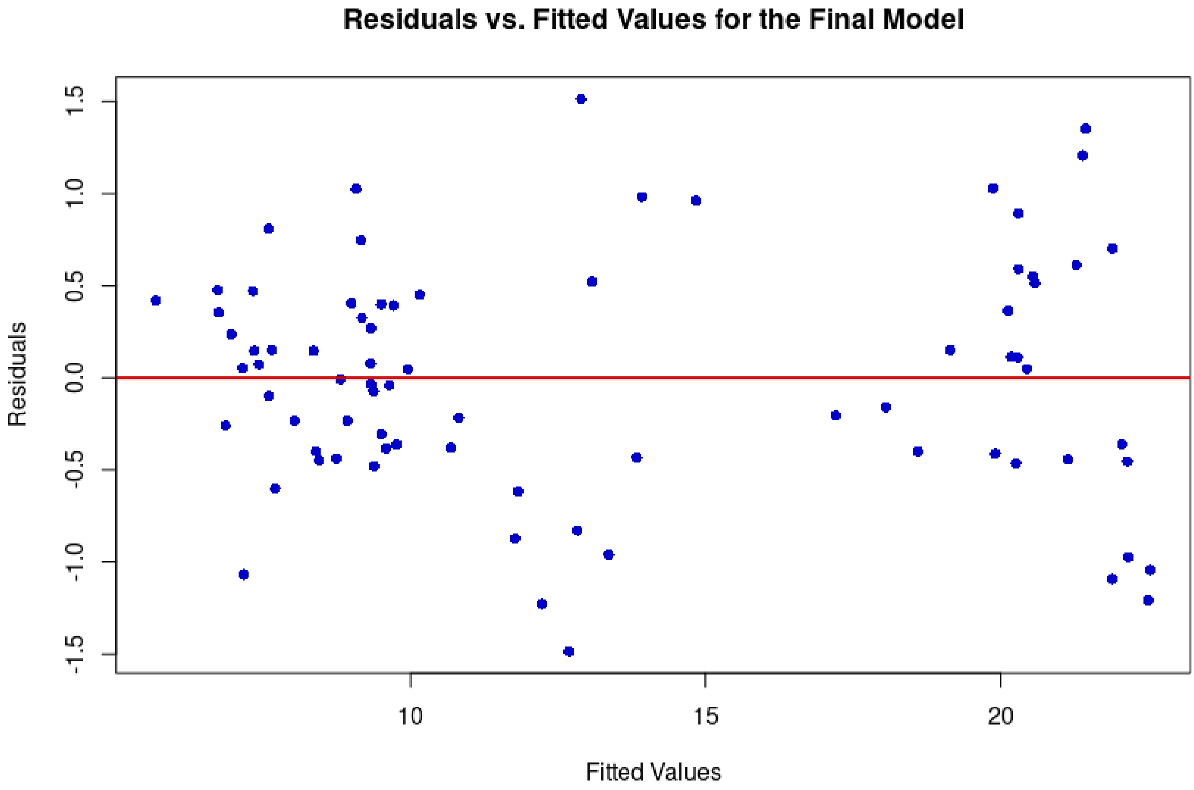
Applying the final linear model to the entire divorce data set, we are able to visualize the predicted values for the number of divorces per 1,000 women in the United States throughout the range of the data. These predicted values, expressed as a curve, are plotted through the raw data points over time along with a 95% confidence band below:



This final model fit appears to approximate the data with reasonable precision for several reasons. First, the 95% confidence band, constructed using a Bonferonni correction, encompasses all of the response values. Next, the predicted curve tends to smooth over the regions of data with more vertical spread in the response; it does not interpolate these points too closely. As such, our iterative scheme targeted at minimizing the MSPE seems to have prevented significant overfitting. Finally, the coefficient of determination for our final model is roughly $R^2 \approx 0.9873$, meaning our multiple-regression model explains about 98.73% of the variability in the United States divorce rate between 1920 and 1996.

3.2 A Plot of the Residuals

Next, we check the assumption of independent and identically distributed observation errors by viewing a plot of the residuals versus the predicted values in our model. This residual plot, with a horizontal line drawn at zero for reference, is featured on the following page:



These residuals do not appear to demonstrate any assumption-violating behavior. We notice that the residuals seem to be distributed nearly equally on either side of the horizontal line at zero, suggesting an absence of any substantial bias in our predictions. Additionally, there does not appear to be major variation in the vertical spread of the residuals over the range of the predicted values; the largest and smallest residuals are not closely clustered near any particular predicted value. Finally, there does not appear to be any obvious curve or pattern in the residuals versus the predicted values. Together, these observations about the residual plot suggest that the assumption of independent and identically distributed observation errors according to the normal distribution with mean 0 and variance σ^2 is plausible.

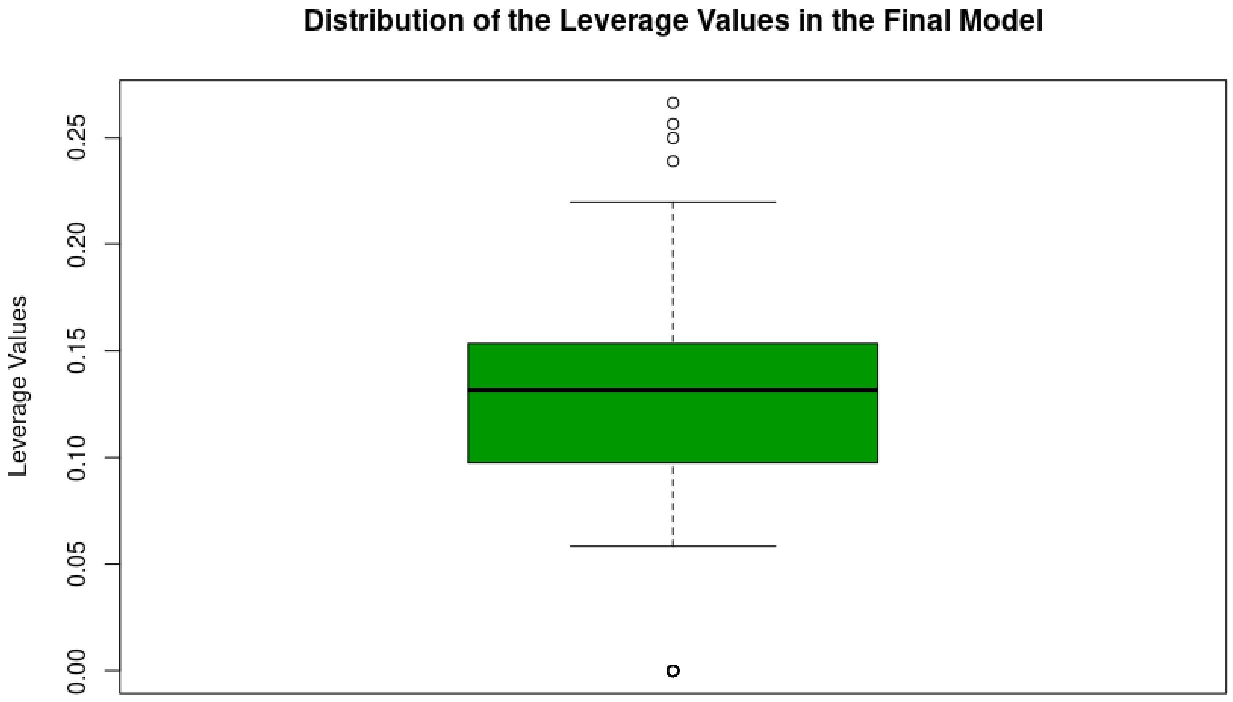
3.3 High-Leverage Points

Next, we identify the high-leverage points in our model. By sorting the set of all leverage values from smallest to largest, we form the following table of the ten largest leverage values and their corresponding observation times:

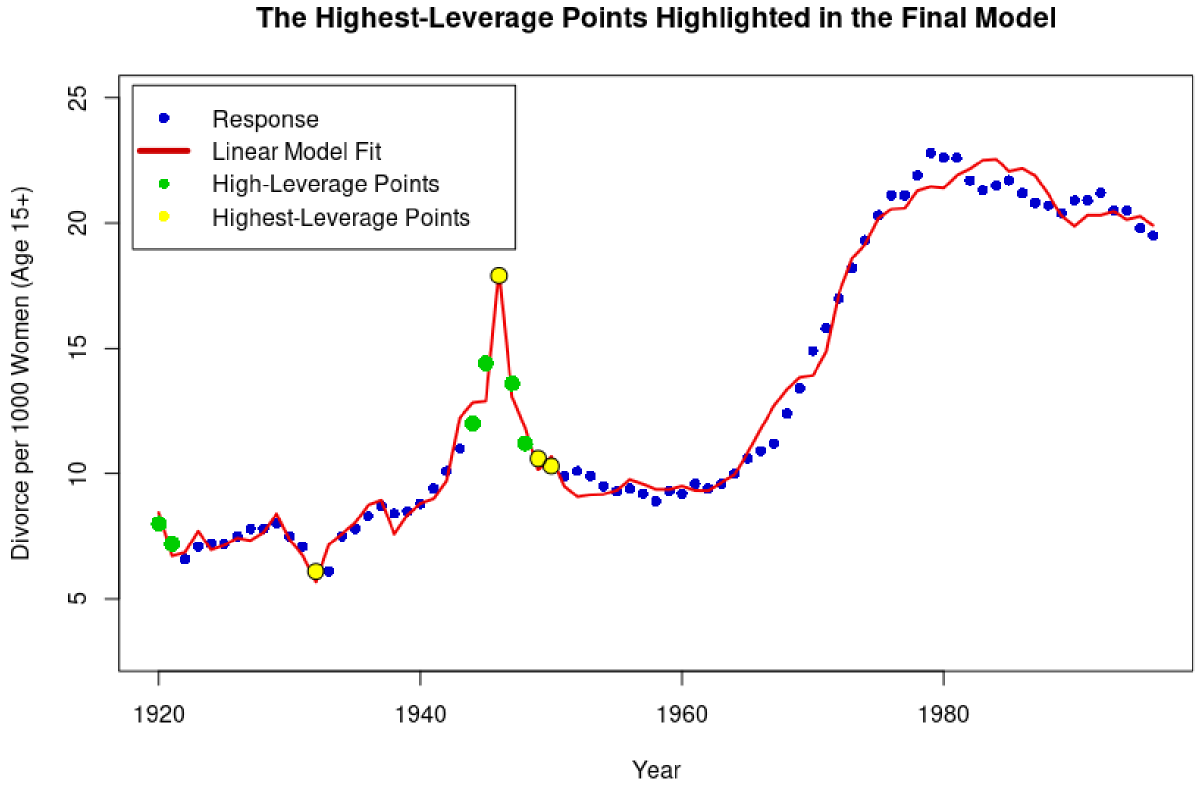
Year	1947	1933	1951	1950	1946	1945	1922	1949	1948	1921
Leverage	0.917	0.571	0.544	0.461	0.427	0.392	0.341	0.339	0.304	0.301

Figure 4: The Ten Highest-Leverage Points in the Final Model.

These ten points with the highest-leverage are compared to the leverage values for every observation in the divorce data set. A boxplot visualizing the distribution of the leverage values for the entire set of data is shown on the following page:



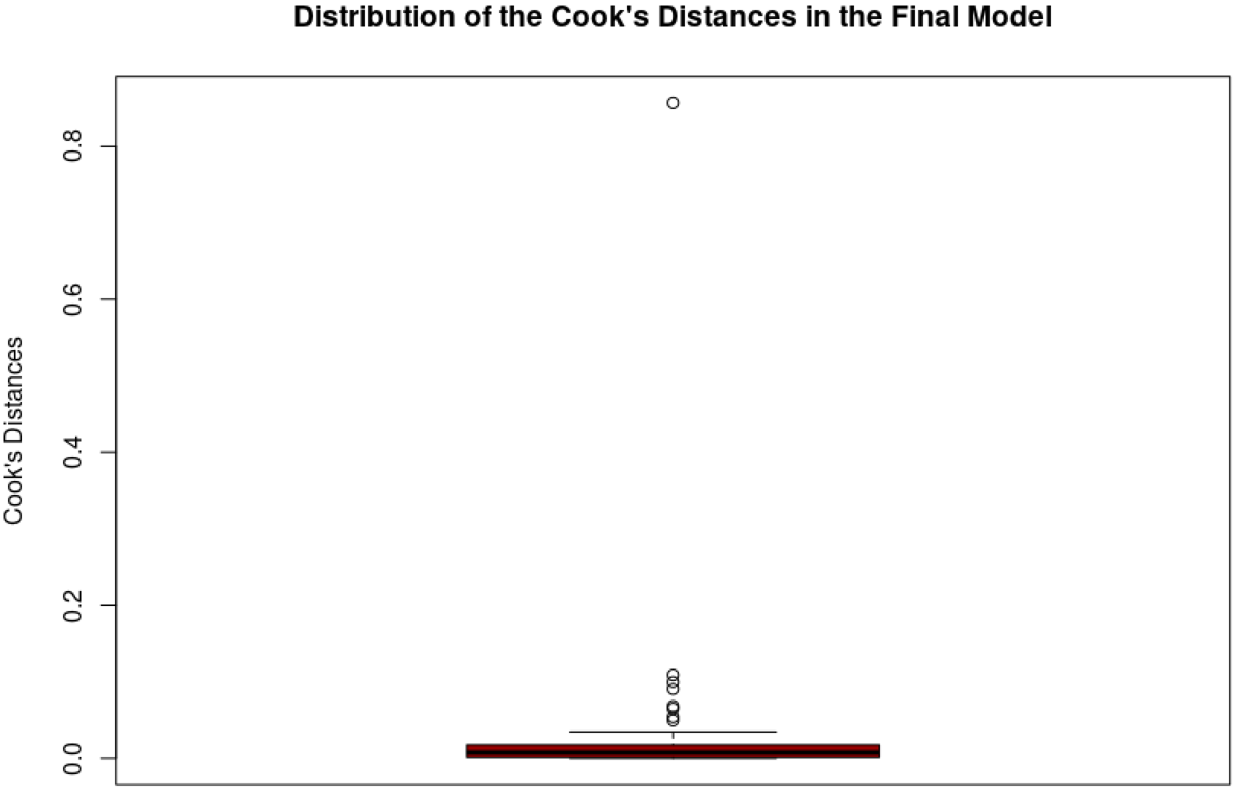
This distribution reveals that most of the leverage values tend not to be very large, but there are roughly four unusually high points of leverage. A plot of the predicted curve through the response with the four highest leverage points in yellow and the six next-highest leverage points in green is featured below:



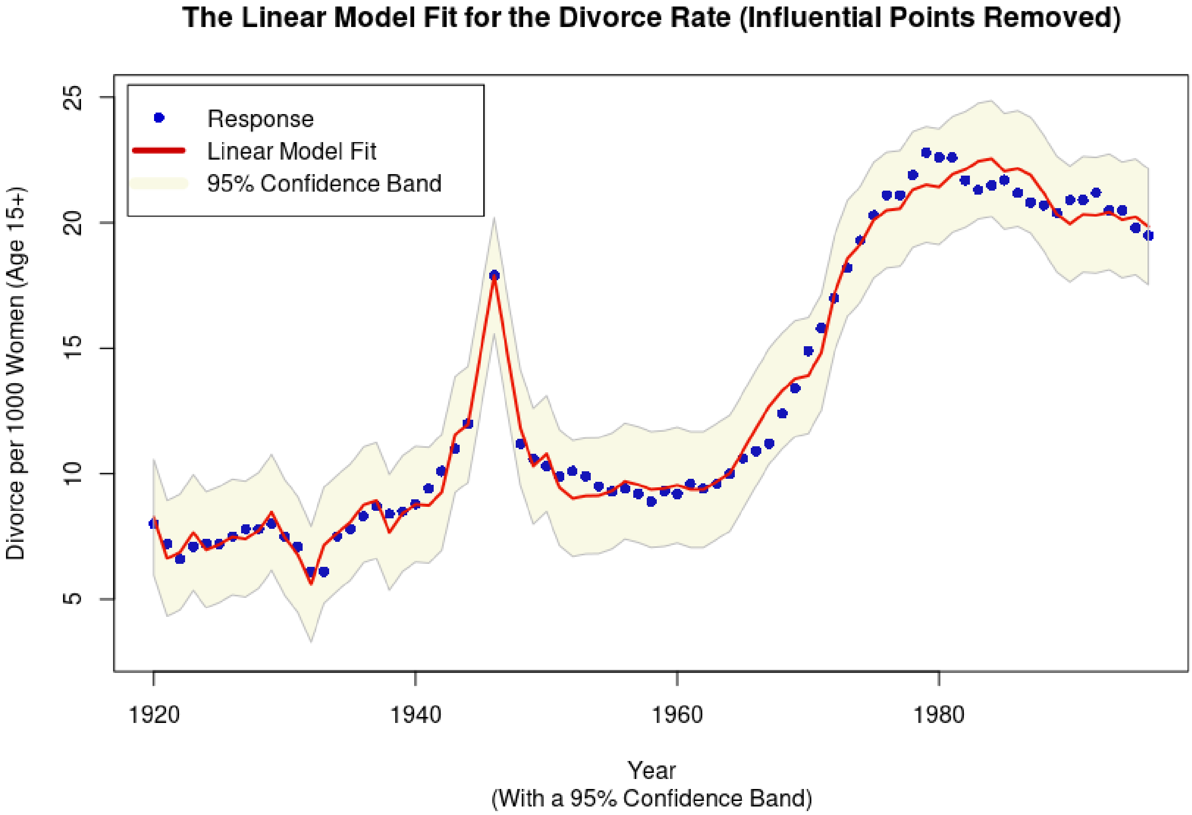
So, this plot now shows the predicted curve from our linear model along with the ten highest-leverage points; the four highest-leverage points are also distinguished. As expected, we see that the linear model fit tends to pass almost directly through the highest-leverage points. Unsurprisingly, the points with the unusually high values of leverage tend to coincide with pivotal events in American history that heavily altered the divorce rate (in this case, these events are the Great Depression and the Second World War).

3.4 High-Influence Points

At this point, we identify the points of high influence in our model. First, we view the distribution of the Cook's distances in the final model to gauge how many points are potentially of high influence. This distribution, visualized with a boxplot, is given below:



Here, we see that there are two points whose Cook's distances are significantly larger than the others. We will consider these two points the only highly influential points in the fit. Ultimately, we find that the point of highest influence occurs at the observation year of 1947 and has a Cook's distance of roughly 0.493 while the other point of high influence is observed in 1946 and has a Cook's distance of about 0.431. Because influential points have the potential to detract from the goodness-of-fit of linear models, we re-fit the same linear model to the divorce data set with the observations in 1946 and 1947 excluded from the observation space. The resulting linear model has a maximum Cook's distance of roughly 0.135 which is not large enough to correspond to an influential point; this second fit of our final model is free of influential points. The resulting predicted curve plotted through the reduced response, again with a 95% confidence band from the Bonferonni correction, is given in the figure atop the following page:



In this model fit without influential points, the response still falls entirely within the 95% confidence band and the predicted curve still appears to fit the data reasonably well. The 12 variables featured in our final model of divorce rate appear below with their coefficients and standard errors in the data set with the influential points removed. In addition, the test statistic and p-value for the null hypothesis that the coefficient is zero are included for each variable in this model on the reduced data set:

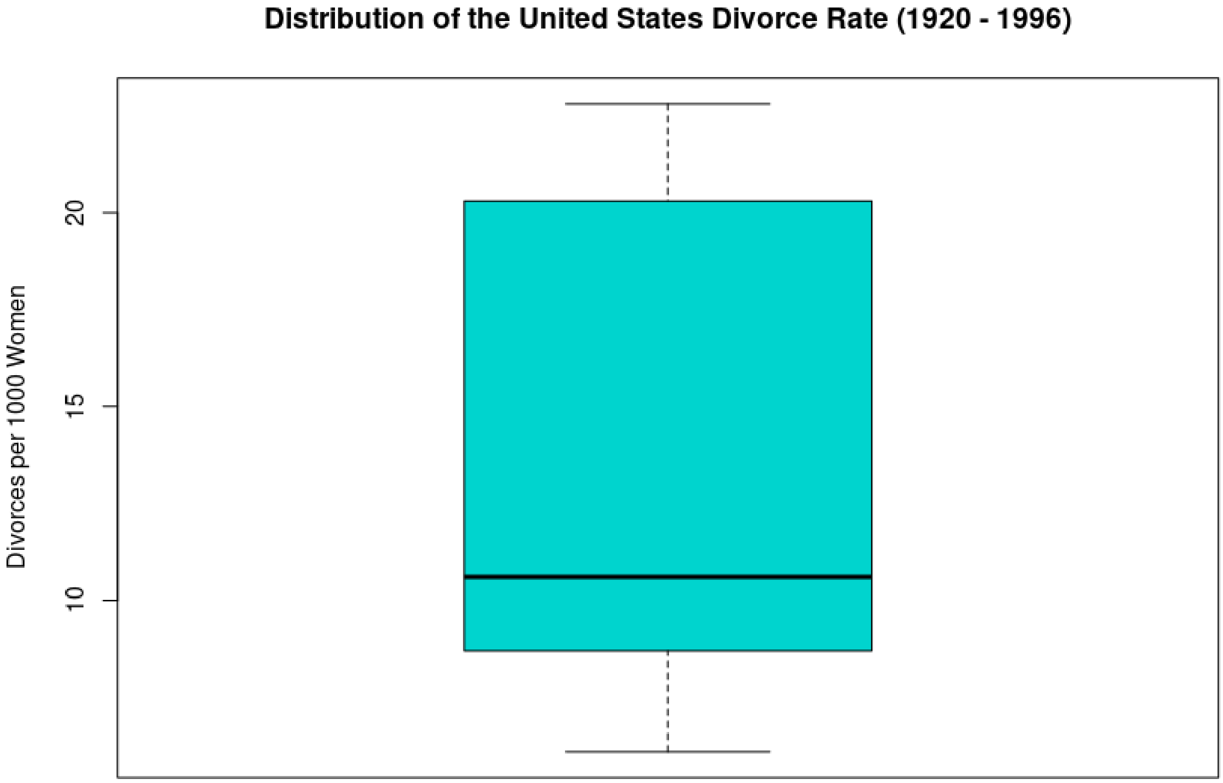
Variable	Coefficient Estimate	Standard Error	Test Statistic	p-Value
Intercept	145.2	26.35	5.512	7.59×10^{-7}
% Female Labor Force	-10.41	1.506	-6.913	3.27×10^{-9}
Squared % Female Labor Force	0.243	0.033	7.456	3.80×10^{-10}
Cubed % Female Labor Force	-0.002	0.0002	-7.683	1.54×10^{-10}
Marriage Rate	1.832	0.609	3.008	3.82×10^{-3}
Squared Marriage Rate	-0.024	0.007	-3.335	1.46×10^{-3}
Cubed Marriage Rate	0.0001	0.00003	3.88	2.59×10^{-4}
Birth Rate	-0.695	0.095	-7.314	6.67×10^{-10}
Squared Birth Rate	0.003	0.0005	6.312	3.47×10^{-8}
Military Participation Rate	-0.057	0.016	-3.624	5.93×10^{-4}
Early 20th Century	-45.15	9.46	-4.773	1.18×10^{-5}
Unemployment Rate	0.039	0.099	0.392	6.96×10^{-1}
Early × Unemployment	-0.145	0.099	-1.464	1.48×10^{-1}
Early × % Female Labor Force	1.381	0.275	5.015	4.85×10^{-6}

Figure 5: Summary of the Explanatory Variables in the Final Model without Influential Points.

In addition, this final model fit to the data without the influential points has a coefficient of determination of roughly $R^2 \approx 0.989$, meaning this modified linear model explains about 98.9% of the variability in the divorce rate between 1920 and 1996. This percentage of explained variance is only slightly (0.17%) larger than in the original model fit with the influential points. In addition, comparing Figures 2 and 4 reveals that the removal of the influential points does not significantly change the values of any model coefficients and does not cause the significance of any variable to change at the $\alpha = 0.05$ level. So, with these metrics considered, there does not appear to be a substantial difference in predictive power nor goodness of fit between the models fit with and without the influential points.

3.5 Outliers in the Response

Now, we check for outliers in the response that could potentially impact our model. First, we view the distribution of the response as a boxplot:



While this boxplot implies that there are no outliers, we use Tukey's criterion to verify that this is the case. So, we find the first and third quantiles of the response to be $Q_1 = 8.7$ and $Q_3 = 20.3$ divorces per 1,000 women, respectively. Then, we calculate the interquartile range to be $IQR = 11.6$ divorces per 1,000 women and construct the "inner fence" threshold with

$$\mathcal{F}_i = Q_1 - 1.5 \times IQR = -8.7 \text{ divorces per 1,000 women}$$

and build the "outer fence" threshold with

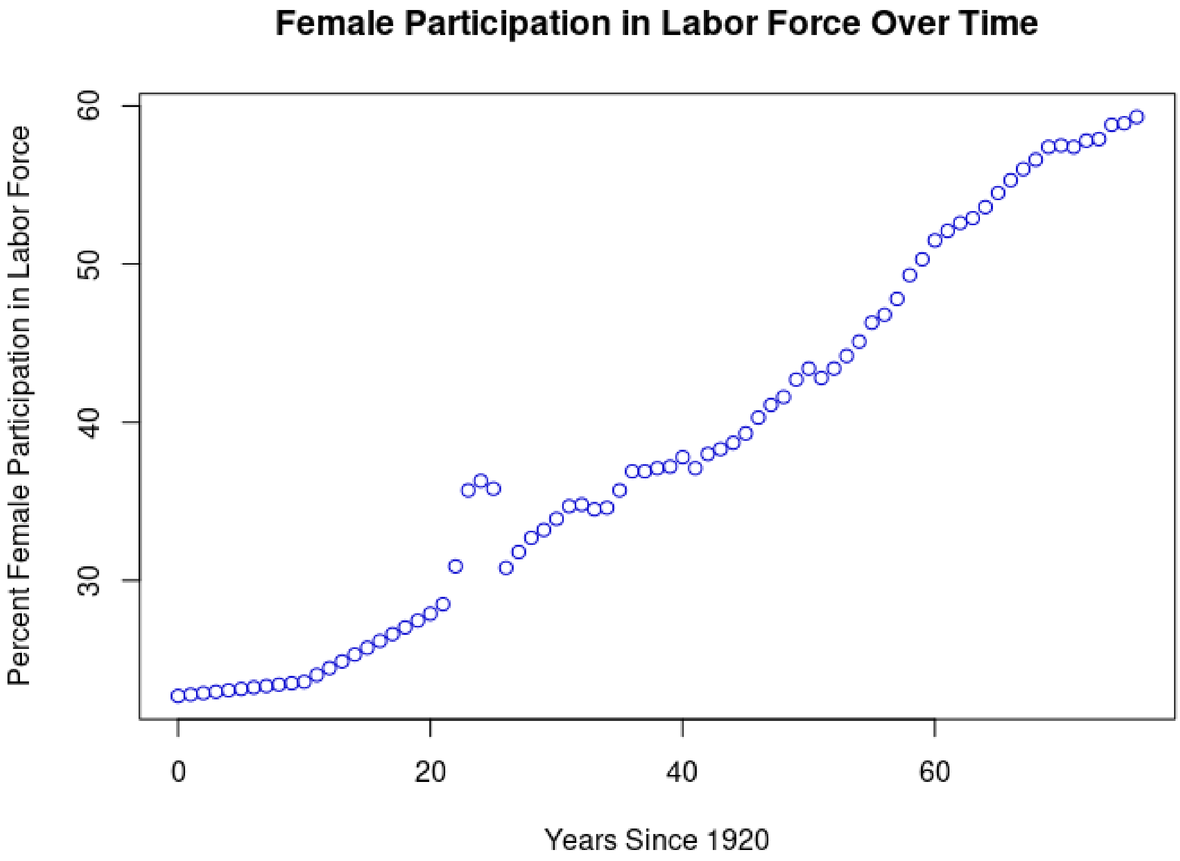
$$\mathcal{F}_o = Q_3 + 1.5 \times IQR = 37.7 \text{ divorces per 1,000 women.}$$

Ultimately, we find that the maximum value of the response is 22.8 divorces per 1,000 women (occurring in 1980) and the minimum value of the response is 6.1 divorces per 1,000 women (occurring in 1933). This means that every observed divorce rate between 1920 and 1996 falls between $\mathcal{F}_i$ and $\mathcal{F}_o$, so we conclude that there are no outliers in our divorce data.

3.6 Limitations of the Multiple Regression Approach

Following the examination of our final linear model for the divorce rate in the United States between 1920 and 1996, a pair of clear limitations become evident in our approach.

First, an examination of the associations between the covariates reveals a collinear relationship between the number of years since 1920 and the female participation in the labor force. A plot of the percent female participation in the labor force versus the number of years since 1920 is given below:



In this plot, it is evident that there is a strong linear association between these variables, noticeably violated only during the period of time enclosing the Second World War. Due to this collinearity, it is possible that our interpretations of the impact of the female participation in the labor force on the divorce rate may be unreliable. After all, it is possible that time (as a proxy for some variable or variables that are not included in the data set) explains the change in the divorce rate in our data set. But, because time and the percent female labor force participation have a strong linear association, it is not possible to distinguish which of these variables possesses a greater ability to explain the divorce rate; we could only make this distinction with access to more data in which these variables are mostly uncorrelated in the range of the data. So, a major weakness of our model is some, or even all, of the impact on the divorce rate attributed to the percent female participation in the labor force could truly be due to some other variable correlated with time.

Another key limitation of our model arises from our small data set. Despite our best efforts to mitigate overfitting to the data, it is still possible that our linear model is too specific to the few data locations we have and could not generalize well to other values in the range of the divorce rate study. Since our data set tracks divorce rates over time, it is not possible to measure new data using exactly the same methods as in this study. So, we can never validate the accuracy of our model at times between 1920 and 1996 that do not correspond to a data observation. We are forced to assume that our linear model constructed from yearly observations of divorce rate and the explanatory variables can sufficiently describe the divorce rate within the range of the data.

4 Conclusions

The multiple regression model for divorce rates from time, unemployment rate, female labor force participation, birth rate, marriage rate, and military participation rate seems to offer well-fitting predictions that explain the vast majority of the variability in the response while meeting the assumptions necessary for validity. With this model, we are able to identify associations between the divorce rate and the explanatory variables with dynamic effect modifications over time. Namely, our linear model features a predicted divorce rate that tends to increase with the marriage rate and decrease with the birth and military participation rates for all values of time. In addition, our final model contains predicted divorce rates that tend to decrease with increasing magnitude over time in 20th Century America and that tend to change with the unemployment rate only during the period between 1920 and 1949. Outside of this time interval, the final model shows no association between unemployment and predicted divorce rates.

Of course, including interactions between explanatory variables and time adds an ability for the changing social conditions throughout time to be reflected in our model. By considering these interactions, our model not only makes mathematical sense following a careful model selection procedure using best subsets regression, but also makes physical sense in the context of American history.

Because model assumptions meet checks, outliers are not present, and points of high leverage and influence do not significantly impact the predictions of our model, it seems plausible that our representation of the United States divorce rate between 1920 and 1996 is reasonably accurate. Of course, our small set of data coupled with a pair of collinear explanatory variables leaves room for error in our multiple regression approach; despite its strengths, we must pay careful attention to the possible impact of these weaknesses in our model when performing analysis.

Appendix

In this appendix, we feature the `R` code used to generate the results and visualizations featured in this report. We will feature two primary sections for code: model selection and model analysis.

Model Selection

This code is responsible for executing the model selection strategy explained in Section 2.2 of this report. Comments in the code highlight specific tasks.

```r
# Get Some Handy R Packages
library(leaps)
library(scales)

# Load Divorce Data; Call it Depression because Divorce is Sad
Depression <- read.table("divusa.txt", header = TRUE)

# Center the Year at 1920; this Gives Years Since 1920
Depression$yearsSince1920 <- Depression$year - 1920

# Change "Early" Covariate to a Dummy Variable
# This is Because it is Categorical
Depression$early <- as.factor(Depression$early)

# Make Some New Variables to Consider Covariate Transformations
Depression$yearsSince1920Squared <- Depression$yearsSince1920^2
Depression$yearsSince1920Cubed <- Depression$yearsSince1920^3
Depression$femlabSquared <- Depression$femlab^2
Depression$femlabCubed <- Depression$femlab^3
Depression$birthSquared <- Depression$birth^2
Depression$marriageSquared <- Depression$marriage^2
Depression$marriageCubed <- Depression$marriage^3
Depression$logMilitary <- log(Depression$military)

# Make a Figure of Response vs. Each Covariate
par(mfrow = c(3, 2))
plot(Depression$yearsSince1920, Depression$divorce,
     col = "blue3", pch = 16,
     xlab = "Years since 1920",
     ylab = "Divorce per 1000 Women (Age 15+)",
     main = "Divorce Rate vs. Time")
plot(Depression$unemployed, Depression$divorce,
     col = "blue3", pch = 16,
     xlab = "Unemployment Rate (%)",
     ylab = "Divorce per 1000 Women (Age 15+)",
     main = "Divorce Rate vs. Unemployment Rate")
plot(Depression$femlab, Depression$divorce,
     col = "blue3", pch = 16,
     xlab = "Female Participation in Labor Force (%)",
     ylab = "Divorce per 1000 Women (Age 15+)",
     main = "Divorce Rate vs. Female Participation in Labor Force")
plot(Depression$marriage, Depression$divorce,
     col = "blue3", pch = 16,
     xlab = "Marriages per 1000 Unmarried Women",
     ylab = "Divorce per 1000 Women (Age 15+)",
     main = "Divorce Rate vs. Marriage Rate")
plot(Depression$birth, Depression$divorce,
     col = "blue3", pch = 16,
     xlab = "Births per 1000 Women (Age 15 - 44)",
     ylab = "Divorce per 1000 Women (Age 15+)",
     main = "Divorce Rate vs. Birth Rate")
plot(Depression$military, Depression$divorce,
     col = "blue3", pch = 16,
     xlab = "Military Personnel per 1000 People",
     ylab = "Divorce per 1000 Women (Age 15+)",
     main = "Divorce Rate vs. Military Participation")
par(mfrow=c(1,1))

# Utilize Best-Subsets on the Training Set
# Perform Best-Subsets 10000 Times and Find the MSPE
N <- 1000
p <- 20

# Make the Model Matrix
X <- model.matrix(divorce ~ yearsSince1920 +
                            yearsSince1920Squared +
                            yearsSince1920Cubed +
                            femlab +
                            femlabSquared +
                            femlabCubed +
                            birth +
                            birthSquared +
                            unemployed +
                            marriage +
                            marriageSquared +
                            marriageCubed +
                            military +
                            logMilitary +
                            early +
                            unemployed * early +
                            femlab * early +
                            marriage * early +
                            birth * early +
                            military * early,
                  data = DepressionTest)

# Find the MSPE for Each Best Subset on 1000 Different Train-Test Splits
MSPEs <- matrix(NA, nrow = N, ncol = p)
for (i in 1:N) {
  # Perform the Test-Train Split
  trainIndex <- sample(1:77, 58)
  DepressionTrain <- Depression[trainIndex,]
  DepressionTest <- Depression[-trainIndex,]

  # Call the Best Subsets Algorithm on this Iteration of the Data
  bestSubsets <- regsubsets(divorce ~ yearsSince1920 +
                                      yearsSince1920Squared +
                                      yearsSince1920Cubed +
                                      femlab +
                                      femlabSquared +
                                      femlabCubed +
                                      birth +
                                      birthSquared +
                                      unemployed +
                                      marriage +
                                      marriageSquared +
                                      marriageCubed +
                                      military +
                                      logMilitary +
                                      early +
                                      unemployed * early +
                                      femlab * early +
                                      marriage * early +
                                      birth * early +
                                      military * early,
                            nvmax = p,
                            data = DepressionTrain)
  bestSubsetsSummary = summary(bestSubsets)

  # Find the MSPE for Each Size Subset
  for (j in 1:p) {
    cHat <- coef(bestSubsets, id = j)
    aParticularlyAptMartix <- X[, names(cHat)]
    yHat <- aParticularlyAptMartix %*% cHat
    MSPEs[i, j] <- mean((DepressionTest$divorce - yHat)^2)
  }
}

# Find the Median MSPE for Each Number of Variables
medianMSPEs <- vector()
for (i in 1:p) {
  medianMSPEs[i] <- median(MSPEs[,i])
}

# Show the Number of Variables Minimizing the Median MSPE
pBest <- which.min(medianMSPEs)
pBest

# Find the Best pBest-Variable Model using the Entire Data Set
bestBestSubsets <- regsubsets(divorce ~ yearsSince1920 +
                                        yearsSince1920Squared +
                                        yearsSince1920Cubed +
                                        femlab +
                                        femlabSquared +
                                        femlabCubed +
                                        birth +
                                        birthSquared +
                                        unemployed +
                                        marriage +
                                        marriageSquared +
                                        marriageCubed +
                                        military +
                                        logMilitary +
                                        early +
                                        unemployed * early +
                                        femlab * early +
                                        marriage * early +
                                        birth * early +
                                        military * early,
                              nvmax = p,
                              data = Depression)

# Fit the Final Model, Showing the Variables and Coefficients in It
bestCHat <- coef(bestBestSubsets, id = pBest)
bestCHat
FinalModel <- lm(divorce ~  femlab  +
                            femlabSquared +
                            femlabCubed +
                            marriage +
                            marriageSquared +
                            marriageCubed +
                            birth +
                            birthSquared +
                            military +
                            early +
                            early * unemployed +
                            early * femlab,
                 data = Depression)

# Plot the Final Model Fit through the Response
plot(Depression$year, Depression$divorce,
     pch = 16, col = "blue3", ylim = c(3, 25),
     xlab = "Year", ylab = "Divorce per 1000 Women (Age 15+)",
     main = "The Linear Model Fit for the Divorce Rate over Time")
lines(Depression$year, FinalModel$fitted.values,
      col = "red2", lwd = 2)

# Compute the Bonferroni Correction for a 95 Confidence Band with 100 Intervals
numIntervals <- 100
zB <- qnorm(.025 / numIntervals, lower.tail = FALSE)

# Find the Estimated Standard Error for Our Linear Model Fit
s <- summary(FinalModel)$s

# Draw the 95 Percent Confidence Band on the Plot of the Linear Model
xG <- c(Depression$year, rev(Depression$year))
yG <- c(FinalModel$fitted.values - zB * s, rev(FinalModel$fitted.values + zB * s))
polygon(xG, yG, border = "grey", col = alpha("yellow2", 0.1))

# Create a Legend to Identify the Linear Model Fit and the Data Points
legend(x = 1969, y = 8.0,
       legend = c("Response", "Linear Model Fit", "95% Confidence Band"),
       col = c("blue3", "red3", alpha("yellow3", .1)),
       pch = c(16, NA, NA), lty = c(NA, 1, 1), lwd = c(NA, 4, 8))

# Confounding Plot
plot(Depression$yearsSince1920,Depression$femlab,
     col = "blue3",
     xlab = "Years Since 1920", ylab = "Percent Female Participation in Labor Force",
     main = "Female Participation in Labor Force Over Time")
```

Model Analysis

This code formulates the results used to analyze and interpret the model in Sections 2.3 and 3. Again, comments in the code reveal specific tasks.

```r
# 1.a. Plot Residuals vs. Fitted Values
# Put a Horizontal Line at Residuals == 0 for Reference
plot(FinalModel$fitted.values, FinalModel$residuals,
     pch = 16, col = "blue3",
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values for the Final Model")
lines(seq(-20,40,length.out = length(FinalModel$fitted.values)),
      seq(0, 0, length.out = length(FinalModel$fitted.values)),
      col = "red3", lwd = 2)

# 1.b. Identify High Leverage Points
leveragePoints <- hatvalues(FinalModel)
topTen <-vector()
topTenIndices <- vector()
for(i in 1:10) {
  topTen[i] <- max(leveragePoints)
  topTenIndices[i] <- which.max(leveragePoints)
  leveragePoints[which.max(leveragePoints)] <- 0
}

# Show the Ten Largest Leverage Values and their Indicies in the Data
topTen
topTenIndices

# Create a Boxplot for the Leverage Values
boxplot(leveragePoints, col = "green4",
        ylab = "Leverage Values",
        main = "Distribution of the Leverage Values in the Final Model")

# Create a Plot of the Model fit with High Leverage Points Highlighted
plot(Depression$year, Depression$divorce,
     pch = 16, col = "blue3", ylim = c(3, 25),
     xlab = "Year", ylab = "Divorce per 1000 Women (Age 15+)",
     main = "The Highest-Leverage Points Highlighted in the Final Model")
lines(Depression$year, FinalModel$fitted.values,
      col = "red2", lwd = 2)
points(Depression$year[topTenIndices], Depression$divorce[topTenIndices],
       pch = 16, cex = 1.5, col = "green3")
points(Depression$year[topTenIndices[1:4]], Depression$divorce[topTenIndices[1:4]],
       pch = 16, cex = 1.5, col = "yellow1")
points(Depression$year[topTenIndices[1:4]], Depression$divorce[topTenIndices[1:4]],
       pch = 1, cex = 1.5, col = "black")

# Create a Legend to Identify the High-Leverage Points vs. Data Points
legend(x = 1918, y = 25.5,
       legend = c("Response", "Linear Model Fit", "High-Leverage Points", "Highest-Leverage Points"),
       col = c("blue3", "red3", "green3", "yellow1"),
       pch = c(16, NA, 16, 16), lty = c(NA, 1, NA, NA), lwd = c(NA, 4, NA, NA))
```

```r
# 1.c. Identify Influential Points (Cook's Distance)
# Show the Maximum and Minimum Values for the Cook's Distance
cooksDistances <- cooks.distance(FinalModel)
max(cooksDistances)
min(cooksDistances)

# Construct a Boxplot for the Distribution of the Cooks Distance
boxplot(cooksDistances, col = "red4",
        ylab = "Cook's Distances",
        main = "Distribution of the Cook's Distances in the Final Model")

# Fit the Model with and without the (Two) Influential Points
sort(cooksDistances)
lessInfluentialDepression <- Depression[-26,]
lessInfluentialDepression <- lessInfluentialDepression[-27,]
lessInfluentialFinalModel <- lm(divorce ~   femlab  +
                                            femlabSquared +
                                            femlabCubed +
                                            marriage +
                                            marriageSquared +
                                            marriageCubed +
                                            birth +
                                            birthSquared +
                                            military +
                                            early +
                                            early * unemployed +
                                            early * femlab,
                                data = lessInfluentialDepression)

# View the Summary of this Model with Influential Points Removed
summary(lessInfluentialFinalModel)

# Check that there are no More Influential Points
max(cooks.distance(lessInfluentialFinalModel))

# Plot the Final Model Fit through the Response without the Influential Points
plot(lessInfluentialDepression$year, lessInfluentialDepression$divorce,
     pch = 16, col = "blue3", ylim = c(3, 25),
     xlab = "Year", ylab = "Divorce per 1000 Women (Age 15+)",
     main = "The Linear Model Fit for the Divorce Rate (Influential Points Removed)",
     sub = "(With a 95% Confidence Band)")
lines(lessInfluentialDepression$year, lessInfluentialFinalModel$fitted.values,
      col = "red2", lwd = 2)

# Compute the Bonferroni Correction for a 95 Confidence Band with 100 Intervals
numIntervals <- 100
zB <- qnorm(.025 / numIntervals, lower.tail = FALSE)

# Find the Estimated Standard Error for Our Linear Model Fit
s <- summary(lessInfluentialFinalModel)$s

# Draw the 95 Percent Confidence Band on the Plot of the Linear Model
xG <- c(lessInfluentialDepression$year, rev(lessInfluentialDepression$year))
```

```
yG <- c(lessInfluentialFinalModel$fitted.values - zB * s,
        rev(lessInfluentialFinalModel$fitted.values + zB * s))
polygon(xG, yG, border = "grey", col = alpha("yellow3", 0.1))

# Create a Legend to Identify the Linear Model Fit and the Data Points
legend(x = 1918, y = 25.5,
       legend = c("Response", "Linear Model Fit", "95% Confidence Band"),
       col = c("blue3", "red3", alpha("yellow3", .1)),
       pch = c(16, NA, NA), lty = c(NA, 1, 1), lwd = c(NA, 4, 8))

# 1.d. Outliers
# Plot the Distribution of the Response
boxplot(Depression$divorce, col = "cyan3",
        ylab = "Divorces per 1000 Women",
        main = "Distribution of the United States Divorce Rate (1920 - 1996)")

# Check for Outliers by Hand by Computing the Upper and Lower Fence
# from the First and Third Quartiles and the Interquartile Range
Q3 <- quantile(Depression$divorce, 0.75)
Q1 <- quantile(Depression$divorce, 0.25)
IQR <- Q3 - Q1
Q1
Q3
IQR
upperFence <- Q3 + IQR * 1.5
lowerFence <- Q1 - IQR * 1.5

# Show that the Response Falls within the Upper and Lower Fences
upperFence
lowerFence
max(Depression$divorce)
min(Depression$divorce)

# 1.e.Confidence Intervals
# Show a Summary of the Final Model, Including the Variables Used,
# p-Values for Tests that Coefficients are Zero, and 95% Confidence
# Intervals for the Coefficients
summary(FinalModel)
confint(FinalModel)
```