Influence of Education and Occupation on Infidelity

Yuzhen Ke, Ruodan Yang, Brittni Echols and Emily M. Whisler

Department of Statistics, Purdue University

1. Introduction

Marriages are regarded as a lifetime commitment in most cases. However, sometimes infidelity does occur in the marriage. There are many variables that are believed to contribute to partners going astray. The dataset used in this study records individual responses and the number of affairs in their marriage and other factors. The original study (Fair, 1978) has used the dataset to determine the time allocation of individual time between spouse and their paramour. While much literature about the mere phenomena of infidelity ranges in scope, it is rather difficult to locate studies focusing on those underlying causes of infidelity.

This study uses statistical measures to investigate the relationship between variables such as education, occupation, religion, and happiness to rates of infidelity. We hypothesize that environmental factors affecting the partners will ultimately have a greater effect on the rate of infidelity in marriage than the personal and behavioral factors. Moreover, our study differs from the original study from 1978 because we are not examining leisure time of the individuals in the study, as we do not have access to that portion of the dataset. However, our study is useful in observing some of the significant variables that are associated with infidelity.

2. Methods

2.1 Description of Data

The data set that will be used on this project is called Extramarital Affairs Data which has a total of 601 observations and 8 variables for the response (affairyes). In order to better analyze our data we made some operations on the original dataset. The detailed information is shown in the chart below.

Table 1

Description of Variables in the Data

Variable Name	Dependent or Independent	Type	Description
affairyes	Dependent	N/A	Does the person have an affair(s)? (0 = No, 1 = Yes)
gendermale	Independent	Personal	Is the person a male? (0 = female, 1 = male)
age	Independent	Personal	Age in years: 17.5 = under 20; 22 = 20 - 24; 27 = 25 - 29; 32 = 30 - 34; 37 = 35 - 39; 42 = 40 - 44; 47 = 45 - 49; 52 = 50 - 54; 57 = 55 or over
yearsmarried	Independent	Personal	Number of years married: 0.125 = 3 months or less; 0.417 = 4 - 6 months; 0.75 = 6 months - 1 year; 1.5 = 1 - 2 years; 4 = 3 - 5 years; 7 = 6 - 8 years; 10 = 9 - 11 years; 15 = 12 or more years
childrenyes	Independent	Environmental	Are there children in the marriage? (0 = No, 1 = Yes)
education	Independent	Environmental	Level of education: 9 = grade school; 12 = high school graduate; 14 = some college; 16 = college graduate; 17 = some graduate work; 18 = master's degree; 20 = Ph.D., M.D., or other advanced degree
occupation	Independent	Environmental	Occupation according to Hollingshead classification (reverse numbering).
religiousness	Independent	Behavioral	Religiousness: 1 = anti; 2 = not at all; 3 = slightly; 4 = somewhat; 5 = very
rating	Independent	Behavioral	Self-rating of marriage: 1 = very unhappy; 2 = somewhat unhappy; 3 = average; 4 = happier than average; 5 = very happy

2.2 Preliminary Exploratory Analysis

According to the correlation matrix (Table 2), it can be inferred that some predictor variables may have multicollinearity issues. For example, the correlation between age and yearmarried is 0.7775 which is significant, so including both variables might be unnecessary. However, we shouldn't hastily exclude any variables in this stage since further testing is needed.

Table 2

Correlation Matrix

	age	yearsmarried	religiousness	education	occupation	rating	affairsyes	gendermale	childrenyes
age	1.0000	0.7775	0.1938	0.1346	0.1664	-0.1990	0.0573	0.1906	0.4219
yearsmarried	0.7775	1.0000	0.2183	0.0400	0.0446	-0.2431	0.1403	0.0303	0.5729
religiousness	0.1938	0.2183	1.0000	-0.0426	-0.0397	0.0243	-0.1301	0.0077	0.1294

education	0.1346	0.0400	-0.0426	1.0000	0.5336	0.1093	0.0193	0.3975	-0.0070
occupation	0.1664	0.0446	-0.0397	0.5336	1.0000	0.0174	0.0376	0.4679	-0.0927
rating	-0.1990	-0.2431	0.0243	0.1093	0.0174	1.0000	-0.2538	-0.0075	-0.1963
affairsyes	0.0573	0.1403	-0.1301	0.0193	0.0376	-0.2538	1.0000	0.0510	0.1336
gendermale	0.1906	0.0303	0.0077	0.3975	0.4679	-0.0075	0.0510	1.0000	0.0692
childrenyes	0.4219	0.5729	0.1294	-0.0070	-0.0927	-0.1963	0.1336	0.0692	1.0000

Before we start fitting a model for our data, we want to check the performance of our variables first. According to the histogram for each variables (Figure 1) in the data, we observe that predictors age is skew to the left, while predictors children, yearmarried, and rating are skew to the right, which means our sample has a great portion of people in a young age and most of them have children and highly rated their marriage. What's more, the histogram of the response variable shows that it's a binary variable which only contains values 0 and 1. For a dataset with binary response variables, we can't fit in a linear regression model since box-cox power transformation (Equation 1) can't work for correcting the problem that our data doesn't follow a normal distribution (Figure 2). Therefore, we decide not to use the linear regression model to deal with our data.

$$Y=\begin{cases}0\\1\end{cases}\Rightarrow \begin{matrix}When\ Y=0,\ Y^{\lambda}=0\\When\ Y=1,\ Y^{\lambda}=1\end{matrix}$$

Equation 1

Figure 1

Histogram of Variables

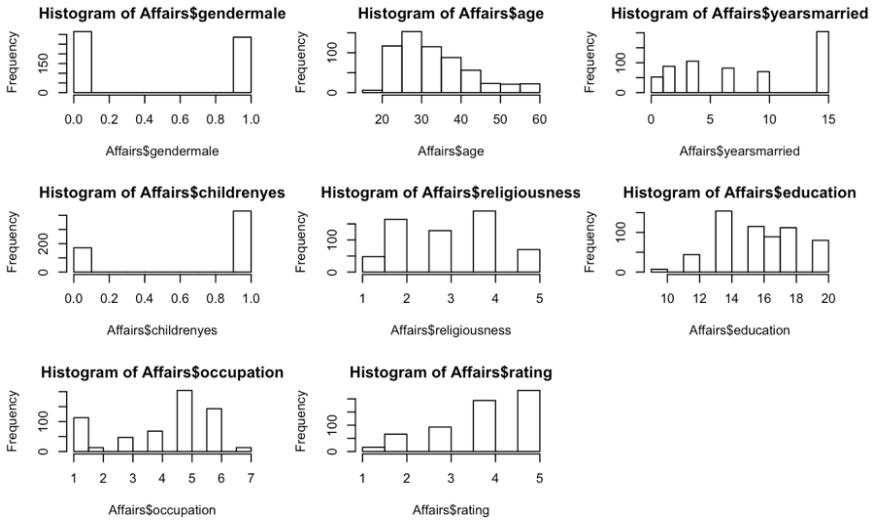
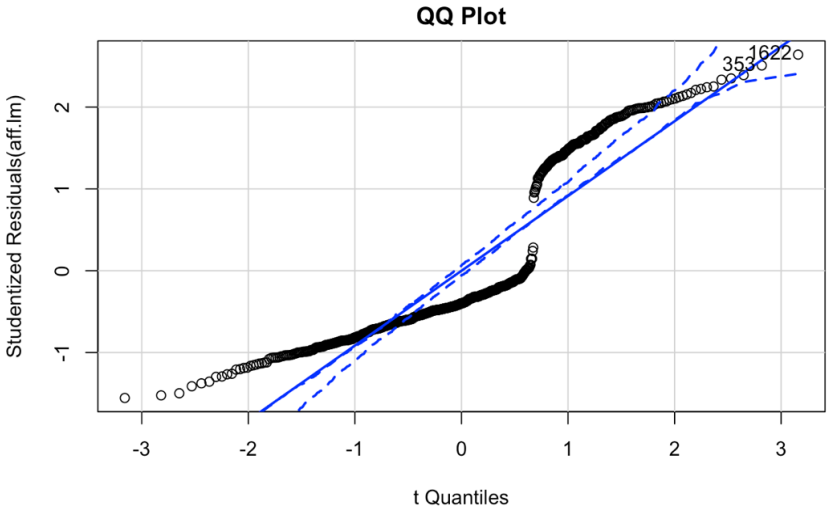


Figure 2

QQ-plot for Linear Regression Model

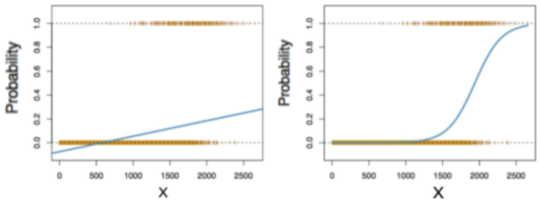


Comparing to linear regression model, logistic regression model will give us outputs, denoted by P(X) = Pr(Y = 1 | X), between 0 and 1 for all X variables . In order to better introduce how logistic regression fits binary variables better. Please see the comparison between linear and logistic regression models shown in the graph below.

From Figure 3, The orange ticks indicate the 0/1 values coded for probability (No or Yes). We observe that the estimated probability in the left plot using linear regression sometimes is negative. However, the predicted probabilities in the right plot using logistic regression all lie between 0 and 1.

Figure 3

Result of Fitting Two Models



Why does logistic regression give us outputs between 0 and 1? Let's take a look at the logistic function (Equation 2).

$$p(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p}} \qquad \text{Equation 2}$$

It's easy to see that no matter what value of β and Xi , the equation (Equation 2) will always yield outputs between 0 and 1. For this reason, the logistic regression model suits better for our data with binary response variables.

2.3 Model Building

We use K-Fold Cross Validation (or K-Fold CV), Stepwise Selection, and Best Subset Selection methods to conduct model selection. Although the 10-Fold CV method selects the full model, predicted errors from different models don't have much difference meaning that the model we finally choose won't make much difference. Combined with the Best Subset Selection according to the Mallow's Cp criteria and the Stepwise Selection according to the AIC criteria, we finally choose the 5-predictor model containing variables of gendermale, age, yearsmarried, religiousness, and rating.

Besides, since we can't find a similar survey to test our model, we simply refer to the predicted errors from 10-Fold CV as our testing error. The MSE of testing data, from Table 3, is 0.1718066 which indicates that our model is good for prediction.

Table 3

K-Fold Cross Validation Table

Full Model	Estimated Prediction Error	Variables Removed
8	0.1708306	none
7	0.0711786	education
6	0.1722147	education, occupation
5	0.1718066	education, occupation, childrenyes
4	0.172333	education, occupation, childrenyes, gendermale
3	0.1725937	education, occupation, childrenyes, gendermale, age
2	0.1748188	education, occupation, childrenyes, gendermale, age, yearsmarried
1	0.1766155	education, occupation, childrenyes, gendermale, age, yearsmarried, rating

In conclusion, the model that we are going to use for inference is following:

$$p(x) = \frac{e^{0.82 + 0.06(gender) - 0.01(age) + 0.02(yearmarried) - 0.05(religiousness) - 0.09(rating)}}{1 + e^{0.82 + 0.06(gender) - 0.01(age) + 0.02(yearmarried) - 0.05(religiousness) - 0.09(rating)}}$$

Equation 3

More detailed analysis about model selection will be further discussed in the Result part.

2.4 Diagnostics

Note that the assumptions of the logistic regression model don't make as many key assumptions as the linear regression model. For example, the logistic regression model does not require a linear relationship between the response and explanatory variables, and the residuals don't have to be normally distributed.

The main 5 assumptions in logistic regression model are: (1) Response variables are qualitative; (2) Observations should be independent of each other; (3) There is no multicollinearity between explanatory variables; (4) Explanatory variables are linearly related to the log odds; (5) The data should have a large sample size which requires 10 cases for each predictor.

Before we check the performance of our model according to these assumptions one by one, we want to first deal with problems of outliers and influential points. Firstly, we test the outlier according to X and Y respectively. We use this formula (Equation 4) to find whether there are any x outliers. R output shows that we have 27 outliers according to X (shown in Figure 4). Also, we use the following equation (Equation 5) to find if there is any Y outlier. Although it seems that there are some possible y outliers for example case 1,294 and case 1,622 in Figure 5, the consequence from Equation turns out to be no significant Y outliers. Secondly, we want to test whether there are influential points in the model.

$$h_{ii} > 2\bar{h}$$ Equation 4

$$|t_i| > t(1 - \frac{a}{2n}, n - 1 - p)$$ Equation 5

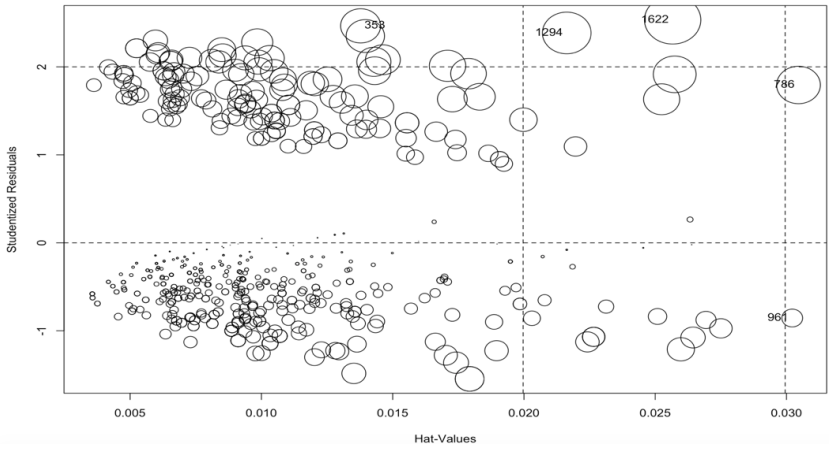
Figure 4

27 x Outliers Output

```
 [1]   16  382  409  491  517  734  751  794  800  876  961
[12]  967 1084 1328 1453 1473 1595 1704 1775 1834  174  786
[23]  858 1294 1622 1669 1782
```

Figure 5

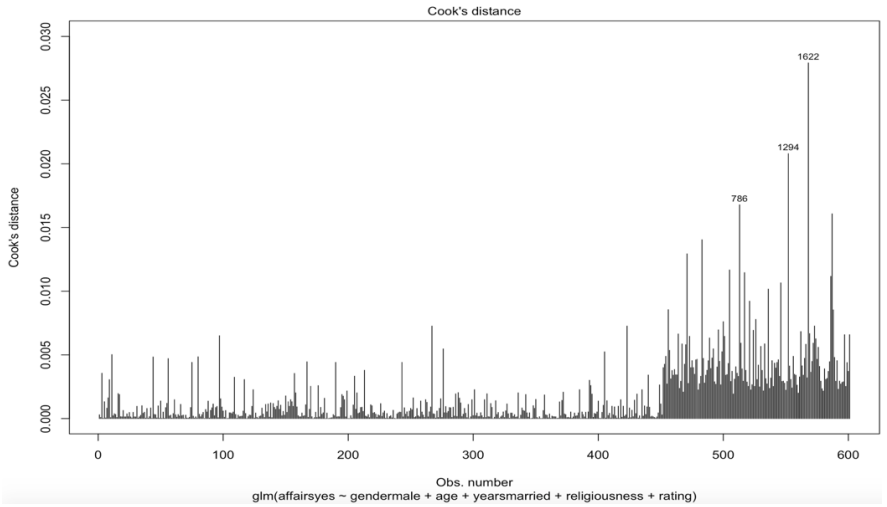
Plot of Hat Values Sudentized Residuals



In Figure 6, we can see that there could be some significant points on the right of the Cook's Distance plot. And then, we use the formula (Equation 6) below to check the influential point. We finally get no influential point. In conclusion, there's 27 outliers according to X, no outliers according to Y, and no influential points. However, since we have a large sample size of 601, and 27 outliers only account for 4.49% which is smaller than 10%, we can ignore these outliers.

Figure 6

Plot of Cook's Distance



$$D_i > F(0.2, p, n - p)$$ Equation 6

Next, for the first and fifth assumptions that require a qualitative Y variable and a large sample size, we can simply conclude that our data well satisfied these two assumptions due to the equation (Equation 7) shown below.

$$(Sample\ Size = 601) > (80 = 8\ Predictors * 10)$$ Equation 7

For the second assumption, we are going to analyze residuals and test the assumptions of residuals. Since we don't have residuals against time, collection order, spatial coordinates and so on, we cannot check the independence of the residuals here. Instead, we can only check the constancy of variables. Figure 7 shows the result of the Brown-Forsythe test under the best model. The P-value is equals to 0.5249 greater than the significance level $\alpha = 0.05$, thus we do not reject the null hypothesis. We can therefore conclude that the residuals in this data have constant variance.

Figure 7

Result of Brown-Forsythe Test

```
  Brown-Forsythe Test
------------------------------------------------------------
  data : residN and groupN

  statistic  : 0.8021913
  num df     : 4
  denom df   : 216.7151
  p.value    : 0.524949

  Result     : Difference is not statistically significant.
------------------------------------------------------------
```

Table 4

Result of VIF Selected Model

	age	yearsmarried	religiousness	rating
gendermale	2.533813	2.620064	1.058834	1.070378
	gendermale	yearsmarried	religiousness	rating
age	1.000919	1.123722	1.057082	1.069966
	age	gendermale	religiousness	rating
yearsmarried	1.129002	1.039854	1.044581	1.046892
	age	gendermale	yearsmarried	rating
religiousness	2.722043	1.077057	2.677272	1.063123
	age	gendermale	yearsmarried	religiousness
rating	2.725711	1.077138	2.654458	1.051737

Then, we want to check the multicollinearity issue. Table 4 shows the values of multicollinearity for the best model, using variance inflation factor. This plot shows us the variance inflation factor for each predictor given other predictors in the mode. For example, in the first line, gendermale is tested while age, yearmarried, religiousness, and rating in the model. The first value, 2.5338 is the VIF for gendermale related to age in this specific model. As all the maximum VIF are smaller than 5, there is no excessive multicollinearity in the best model.

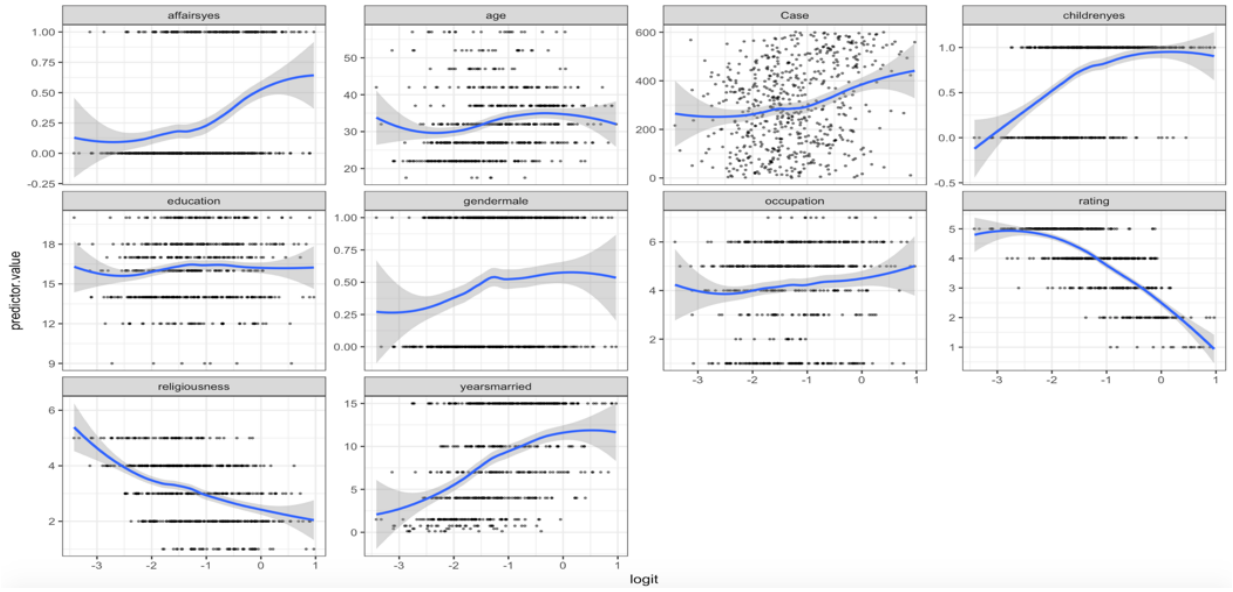
Finally, to check whether explanatory variables are linearly related to the log odds. We need to rearrange the formula to log odds equation (Equation 8) first.

$$log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \qquad \text{Equation 8}$$

We can easily see from the equation that the X variables have a linear relationship with log odds. In R, we use the ggplot function to generate Figure 8 from which we can analyze the linearity between X variables and log odds. From Figure 8, we can see that predictors childrenyes, gendermale, rating, religiousness and yearsmarried are linear to logit scale of y response indicating that for these variables the linearity to log odds assumption are met.

Figure 8

Result of ggplot for Linearity Between x Variables and Log Odds



2.5 Inferential Methods

To answer our research question, we set the hypothesis that "Environmental factors such as education and occupation affecting the partners will ultimately have a greater effect on the rate of infidelity in marriage than the demographics factors associated with the couple." with the significance level of 0.05. After we do the model selection, the summary table of the best subset is given in Figure 9 and the type one and type two anova table Figure 10. From Figure 9 and Figure 10, we observe that 4 predictors (age, yearsmarried, religiousness, and rating) have p-value smaller than the significance level, which means that they have significant impact on response variable, while environmental factors (education, occupation and childrenyes) do not have a significant impact on Y. We also employ a goodness of fit which output shown in Figure 11. The p-value = 0.1444 > 0.05 = α, so we can conclude that our selected model has good fit. In this stage so far, our hypothesis can be rejected and we can conclude that Environmental factors do not ultimately have a greater effect on the rate of infidelity in marriage than the demographics factor associated with the couple.

Figure 9

Summary of Best Selected Predicted Model

```
Call:
glm(formula = affairsyes ~ gendermale + age + yearsmarried +
    religiousness + rating, data = Affairs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6304  -0.2663  -0.1586   0.1077   1.0250

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.821582   0.103300   7.953 9.18e-15 ***
gendermale     0.063607   0.034902   1.822 0.068892 .
age           -0.007397   0.002988  -2.475 0.013586 *
yearsmarried   0.018596   0.004970   3.741 0.000201 ***
religiousness -0.054425   0.014815  -3.674 0.000261 ***
rating        -0.087599   0.015764  -5.557 4.15e-08 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 10

Result of ANOVA Tables

```
             Df Deviance Resid. Df Resid. Dev
NULL                          600     112.56
gendermale    1   0.2923       599     112.27
age           1   0.2647       598     112.00
yearsmarried  1   2.9919       597     109.01
religiousness 1   2.9140       596     106.10
rating        1   5.2347       595     100.86
Analysis of Deviance Table (Type II tests)

Response: affairsyes
             LR Chisq Df Pr(>Chisq)
gendermale      3.3212  1  0.0683905 .
age             6.1276  1  0.0133091 *
yearsmarried   13.9979  1  0.0001830 ***
religiousness  13.4948  1  0.0002392 ***
rating         30.8793  1  2.746e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11

Result of Goodness of Fit Test

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  Affairs$affairsyes, fitted(aff.red)
X-squared = 12.156, df = 8, p-value = 0.1444
```

3. Results

Our first goal was to examine predictors of infidelity in marriage. To address this goal, we considered K-Fold cross validation, stepwise selection, and best subset model selection. Table 3 shows the results of the K-Fold model in 2.3 Model Building. With all eight predictors, the estimated prediction error of 0.1708306 was the smallest value. Notably, the differences between the K-Fold models were marginally different.

The stepwise selection process was implored to determine the best model. Using forward, backward, and both functions in R, we find that education was the first variable to drop from the model with the largest AIC value (625.68). When the model dropped education and occupation, the AIC value lowered to 624.15. After the childrenyes, education, and occupation variables were removed from the model, the best AIC (623.86) corresponded with five predictor model. The method suggested using gendermale, age, yearsmarried, religiousness, and rating as the best model.

Although the stepwise selection model informed the model chosen in the best subset model, we did not solely rely on the stepwise selection process due to the potential to underestimate certain combinations of variables, and the fact that the selection is determined by order. Therefore, the best subset model (Mallow's CP) was the method used to determine the best model for logistic regression. The rss determined eight predictor models as optimum at 100.564 value. The value for adjr2 was 0.09659 corresponding with six predictors. Only three predictor variables were chosen by bic and -32.738 was the value. In Mallow's CP, the results show the five predictor models being a good model to choose. Furthermore, the best subset model predicts the best p-value. This works best with our data because the data already contained very small p-values. Table 5 shows the model results from the best subset model process.

Table 5

Summary of Different Measures for Determining the Best Model

	1	2	3	4	5	6	7	8
rss	105.300	103.581	102.151	101.428	100.865	100.673	100.586	100.564
adjr2	0.06286	0.07671	0.08794	0.09287	0.09639	0.09659	0.09585	0.09452
cp	22.9500	14.7643	8.34278	6.08756	4.77317	5.64328	7.12987	9.00000
bic	-27.200	-30.778	-32.738	-30.607	-27.600	-22.300	-16.422	-10.155

As mentioned earlier, the model generated in this study:

$$p(x) = \frac{e^{0.82 + 0.06(gender) - 0.01(age) + 0.02(yearmarried) - 0.05(religiousness) - 0.09(rating)}}{1 + e^{0.82 + 0.06(gender) - 0.01(age) + 0.02(yearmarried) - 0.05(religiousness) - 0.09(rating)}}$$

Equation 3

Our model results tell us that low happiness (rating) results in an increased probability of engaging in infidelity. We also find that years married have a positive increasing effect on the probability of infidelity. The model suggests both age and religiousness with slightly negative effects on our affairs variable. Those who identify as more religious are less likely to have an affair. The ANOVA and Parameter Estimate tables are provided in Figure 11 for the logistic model used in our study. All parameters except for gender have a significant effect on affairs.

4. Discussion

Our hypothesis that environmental factors have the greatest impact on the probability of infidelity was rejected. Our findings indicated education and occupation as insignificant predictors. The happiness rating, religiousness, age, and years married were the best predictors of infidelity. Past studies corroborate our results as happiness and marital status were significant predictors of infidelity (Davis, 1984; Kaufman and Taniguchi, 2010). Other studies confirm low levels of religiosity and dissatisfaction being significant predictors of increased probability of infidelity (Plack et al, 2010). Although income was not directly measured but somewhat captured in the classification for the occupation variable in our study, we find that other sources indicate that those living in wealthy countries tend to value monogamy because it is less expensive and less emotionally “torturing”.

There were some limitations to our findings. Due to the discrete nature of the entire dataset, there were limited options for analysis with little variation of sources of variation. Additionally, logistic regression requires the observations to be independent of each other. We could not check our indepence assumption because we did not have a time variable. Finally, the original dataset was a result of two magazine questionnaires from 1969. We assume that there was no randomization of the participant selection, making selection bias possible for the study.

Future direction for this study includes adding a time variable to follow subjects over a period of time. This would allow us to check the independence assumption. Finally, if we explore reasons for marriage such as desire for commitment, companionship, financial stability, benefits/legal rights, having children we may find stronger predictors of infidelity. There was not very much literature discussing these relationships with infidelity.

References

Assumptions of Logistic Regression. (n.d.). Retrieved from https://www.statisticssolutions.com/assumptions-of-logistic-regression/

Cohen, Y. (2015). Extramarital Relationships and the Theoretical Rationales for the Joint Property Rules – A New Model. Retrieved from https://scholarship.law.missouri.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=4123&context=mlr

Fair [Data file]. Available from http://www.stern.nyu.edu/~wgreene/Text/Edition7/TableF18-1.csv

Fair, R. C. (1978). A Theory of Extramarital Affairs. Journal of Political Economics, 86(1), 45-61

Fienberg, S., De Veaux, R. D. (n.d.) Springer Texts in Statistics. Retrieved from https://www.springer.com/series/417

KAUFMAN, G., & TANIGUCHI, H. (2010). MARRIAGE AND HAPPINESS IN JAPAN AND THE UNITED STATES. International Journal of Sociology of the Family, 36(1), 25-48. Retrieved from http://www.jstor.org.ezproxy.lib.purdue.edu/stable/23070777