Exploratory Data Analysis of Fair's Affairs

Contents

Introduction

In today's society, a high proportion of marriages end in divorce. Unsurprisingly, one of the leading causes of divorce are extramarital affairs. An extramarital affair is an illicit or sexual relationship outside of marriage. The aim of this report is to analyse the potential predictors of infidelity, and to use logistic regression to see which variables have the highest impact on an individual engaging in an extramarital affair. This will be done by analysing the dataset from Ray Fair in his widely cited "Theory of Extramarital Affairs" (Fair, 1978). The infidelity data, colloquially known as 'Fair's affairs' is data from a cross-sectional survey conducted by Psychology Today in 1969. The dataset consists of 601 observations on nine variables. The variables this report considers are:

1. **Y** = number of extramarital affairs in the past year.
    - A discrete numeric variable denoting how often the individual engaged in extramarital sexual intercourse during the past year. **0** = none, **1** = once, **2** = twice, **3** = 3 times, **7** = 4 to 10 times, **12** = monthly/weekly/daily.
2. $\mathbf{X_1}$ = gender
    - A factor indicating whether individual is male or female
3. $\mathbf{X_2}$ = age,
    - A discrete numerical variable denoting 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
4. $\mathbf{X_3}$ = years married,
    - A discrete numerical variable denoting the number of year 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.
5. $\mathbf{X_4}$ = children
    - A factor denoting whether there are children in the marriage.
6. $\mathbf{X_5}$ = degree of religiousness

- A discreet numeric variable denoting religiousness: **1** = anti, **2** = not at all, **3** = slightly, **4** = somewhat, **5** = very

7. $\mathbf{X_6}$ = education
   - A discrete numerical variable denoting level of education: **9** = primary school, **12** = high school graduate, **14** = some college, **16** = college graduate, **17** = some graduate work, **18** = masters degree, **20** = Ph.D. or other advanced degree.
8. $\mathbf{X_7}$ = occupation
   - A discreet numeric variable denoting occupation according to Hollingshead classification: **7** = higher executives and major professionals, **6** = business managers and lesser professionals, **5** = administrative personnel and minor professionals, **4** = clerical and sales workers, **3** = skilled manual labourers, **2** = machine operators and semi-skilled employees, **1** = unskilled employees.
9. $\mathbf{X_8}$ = self-rating of marriage
   - A discrete numeric variable denoting self-rating of marriage: **1** = very unhappy, **2** = somewhat unhappy, **3** = average, **4** = happier than average, **5** = very happy.

From the dataset we see that there are eight independent variables and one dependent variable. Applying logistic regression, the aim is to find which factors are the highest contributors to infidelity, and consequently the type of individual most likely to engage in an extramarital affair. Before performing logistic regression however, some exploratory data analysis involving descriptive statistics and univariate analyses are needed in order to better understand the data.

Exploratory Data Analysis

Below are some descriptive statistical summaries of the Affairs dataset generated in R using the summary and table functions. From the descriptive statistics below we can see that from the 601

affairs	gender	age	yearsmarried	children	religiousness	education	occupation	rating
Min. : 0.000	female:315	Min. :17.50	Min. : 0.125	no :171	Min. :1.000	Min. : 9.00	Min. :1.000	Min. :1.000
1st Qu.: 0.000	male :286	1st Qu.:27.00	1st Qu.: 4.000	yes:430	1st Qu.:2.000	1st Qu.:14.00	1st Qu.:3.000	1st Qu.:3.000
Median : 0.000		Median :32.00	Median : 7.000		Median :3.000	Median :16.00	Median :5.000	Median :4.000
Mean : 1.456		Mean :32.49	Mean : 8.178		Mean :3.116	Mean :16.17	Mean :4.195	Mean :3.932
3rd Qu.: 0.000		3rd Qu.:37.00	3rd Qu.:15.000		3rd Qu.:4.000	3rd Qu.:18.00	3rd Qu.:6.000	3rd Qu.:5.000
Max. :12.000		Max. :57.00	Max. :15.000		Max. :5.000	Max. :20.00	Max. :7.000	Max. :5.000

Table 1: Summary statistics of numeric variables

respondents, 451 did not engage in an affair, while 150 engaged in one or more. Figure 1 helps visualise the frequency of affairs and shows us the dependant variable is positively skewed with the majority of respondents having not engage in an extramarital affair. In Figure 2 below we can see that the number of affairs between male and females were relatively close. From the summary output we can see that 286 (48%) of the respondents are male and 315 (52%) of respondents were female. The average age of

```
table(Affairs$affairs)
##   0   1   2   3   7  12
## 451  34  17  19  42  38
table(Affairs$age)
## 17.5  22  27  32  37  42  47  52  57
##    6 117 153 115  88  56  23  21  22
table(Affairs$yearsmarried)
## 0.125 0.417  0.75   1.5     4     7    10    15
##    11    10    31    88   105    82    70   204
table(Affairs$religiousness)
##   1   2   3   4   5
##  48 164 129 190  70
```

```
table(Affairs$education)
##   9  12  14  16  17  18  20
##   7  44 154 115  89 112  80
table(Affairs$occupation)
##   1   2   3   4   5   6   7
## 113  13  47  68 204 143  13
table(Affairs$rating)
##   1   2   3   4   5
##  16  66  93 194 232
```

Table 2: Frequency breakdown of each variable

respondents was 32.5. As 'age' was an numeric variable denoting a particular age bracket, the average age of respondents lies between 30-34 years of age. In addition, 430 (72%) of respondents had reported having children. The majority of respondents have been married for more than 4 years and approximately a third (204) of respondents have been married for 15 years. Just 52 respondents have been married for less than a year. Combining those who scored themselves 3 and above in religiousness makes the total number of respondents who are religious to 389 (65%). The majority of respondents had fairly high education levels ranging from 14 to 20 years. 550 (92%) graduated from secondary school with only 7 respondents having only had primary school education. 327 (54%) respondents were happy (4 and 5) with their marriage, while 93 (15%) were somewhat undecided rating their marriage as average. Only 82 respondents (14%) were unhappy with their marriages (16 on 1 and 66 on 22).

	affairs	age	yearsmarried	religiousness	education	occupation	rating
affairs	1						
age	0.09524	1					
yearsmarried	0.18684	0.77755	1				
religiousness	-0.1445	0.19378	0.21826	1			
education	-0.00244	0.1346	0.04	-0.04257	1		
occupation	0.04961	0.16641	0.04459	-0.03972	0.53361	1	
rating	-0.27951	-0.199	-0.24312	0.0243	0.1093	0.01742	1

Table 3: Correlation matrix of discrete numeric variables

The correlation matrix above contains the correlation coefficients between each variable and the others. The two factor variables ('gender' and 'children') have been removed.

From the correlation matrix we can see that the dependent variable (number of affairs) has positive correlations with age, years married and occupation while it has negative correlations with

religiousness, education, and marriage rating. X_2 e) and X_3 (years married) both have negative correlations with X_8 (marriage rating). X_2 has the largest correlation with X_3 (0.7775). X_6 (years of education) and X_7 (occupation) have a correlation of 0.5336.

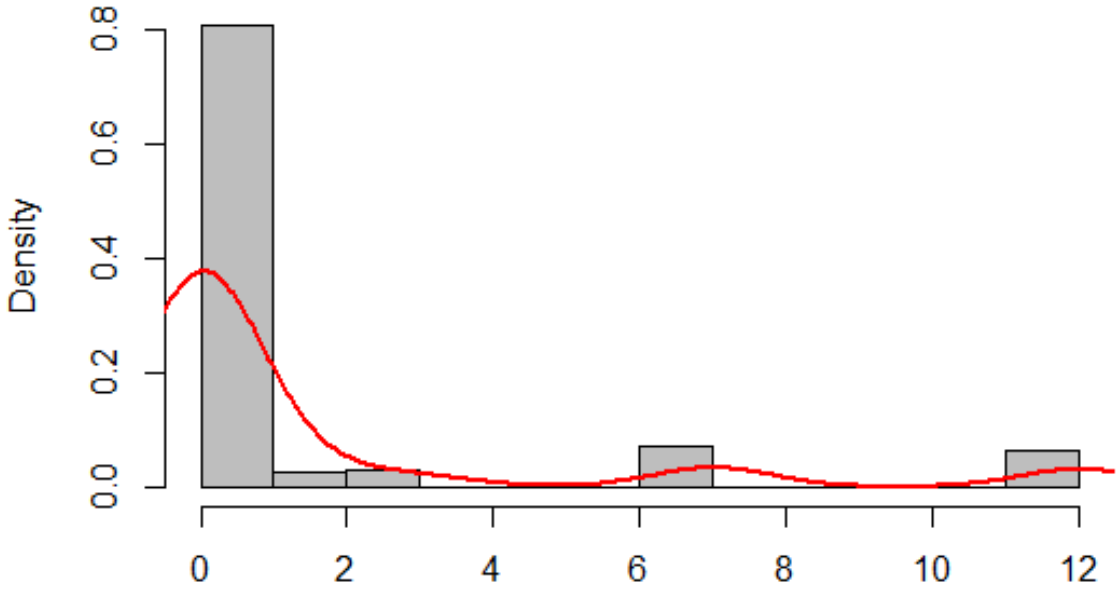

Figure 2: Histogram and plot density function on the distribution of affairs

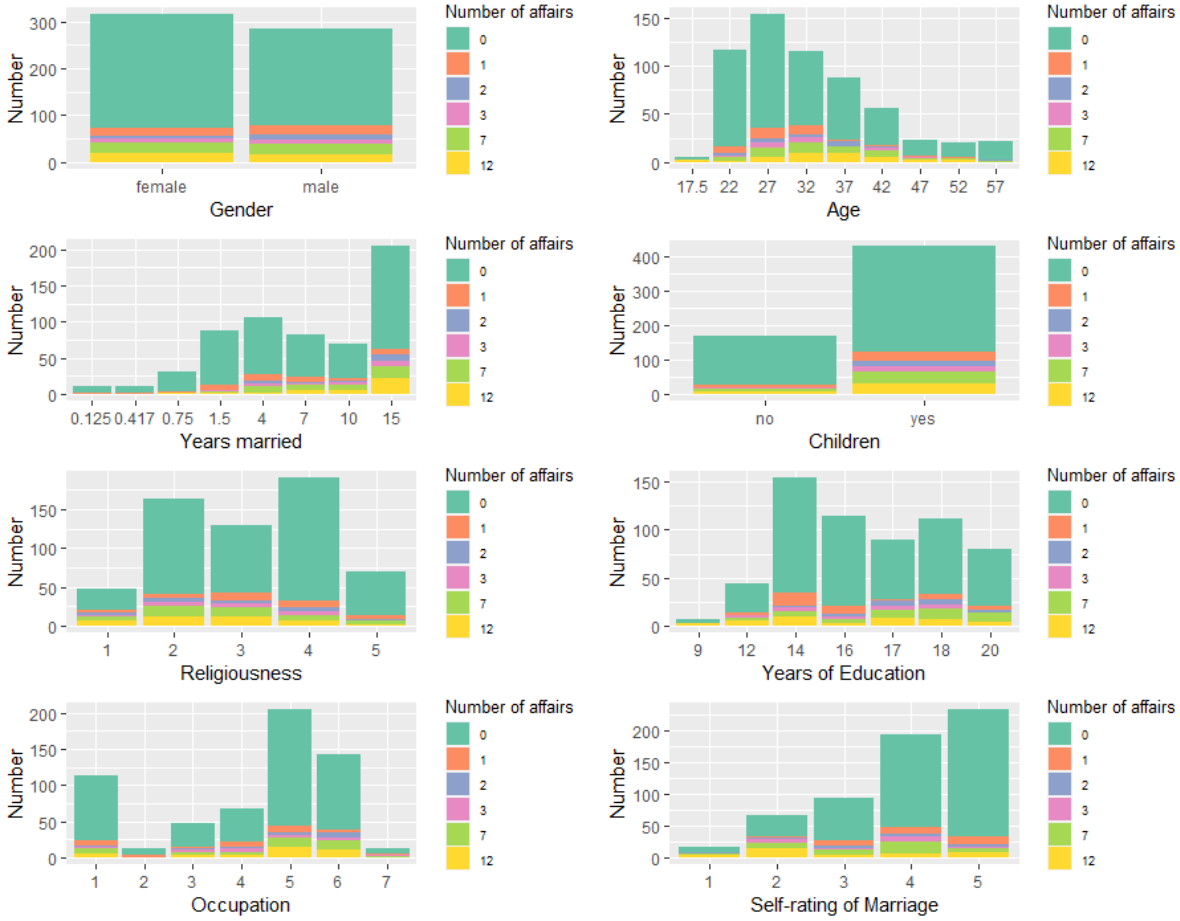

Figure 1: Bar plot of the independent variables against the number of affairs

Logistic Regression

```
Min          1Q  Median          3Q     Max
-1.571  -0.750   -0.569    -0.254    2.519

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.3773    0.8878    1.55   0.12081
gendermale       0.2803    0.2391    1.17   0.24108
age             -0.0443    0.0182   -2.43   0.01530*
yearsmarried     0.0948    0.0322    2.94   0.00326**
childrenyes      0.3977    0.2915    1.36   0.17251
religiousness   -0.3247    0.0898   -3.62   0.00030***
education        0.0211    0.0505    0.42   0.67685
occupation       0.0309    0.0718    0.43   0.66663
rating          -0.4685    0.0909   -5.15   2.6e-07***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 675.38 on 600 degrees of freedom
Residual deviance: 609.51 on 592 degrees of freedom
AIC: 627.5
```

The first part of performing this logistic regression is to create a binary outcome for each respondent (had an affair/did not have an affair). This can then be used as the outcome variable in the logistic regression model. The first model included all the independent variables. By observing the p-values in the regression coefficients to the left, the full model suggests that variables gender, children, education and occupation are not significant variables for the model. From this a reduced model was made including only the statistically significant variables from the previous full model (age *, years married **, religiousness ***, and rating ***). To confirm whether the reduced model truly fits as well as the full model, a chi-square test was employed to compare the two models. The results of the chi-square test suggests that the full model with a non-significant chi-square value ($p = 0.21$), gives further reason to believe that gender, children, education and occupation do not add significantly to the prediction. Therefore the reduced model can be used.

Exponentiating the regression coefficients of the reduced model makes the log(odds) easier to interpret by putting them onto an odds scale. Observing the exponentiated odds indicate that for every one year increase in years married (holding all other variables constant) increases the odds of an extramarital affairs by a factor of 1.106. On the contrary, for every year increase in age, the odds of an extramarital affair are multiplied by a factor of 0.965. This indicates that the chance of having an affair drops -3.5% every time an individual gets older. To summarise thus far, the odds of an affair increase with years married and decrease with age, religiousness and marital rating.

The final part of this analysis is assess the impact of the predictors on the probability of an outcome. This will be done by applying the predict() function to an artificial data set containing the mean values of the predictor variables apart from the specific variable under observation which will vary. Observations can then be made about what varying the levels of the predictor variables would have on the probability of the outcome. This method will be applied to assess the impact of marital rating and age on the probability of having an extramarital affair.

	rating	age	years married	religiousness	prob
1	1	32.5	8.18	3.12	0.530
2	2	32.5	8.18	3.12	0.416
3	3	32.5	8.18	3.12	0.310
4	4	32.5	8.18	3.12	0.220
5	5	32.5	8.18	3.12	0.151

Table 4: Artificial dataset observing impact of varying rating on probability of outcome

The results show that the probability of an extramarital affair decreases from 0.53 when the marriage is rated 1, to 0.15 when the marriage is rated 5. This indicates that an individual unhappy in their marriage are three times more likely to have an extramarital

affair than compared to their content in marriage counterpart. Furthermore, we can see by looking at the output below, that the probability of engaging in an affair decreases as age increases. From 0.34 at 17 to 0.11 at 57, indicates that the older the individual is the less likely they are to engage in an extramarital affair

	rating	age	years married	religiousness	prob
1	3.93	17	8.18	3.12	0.335
2	3.93	27	8.18	3.12	0.262
3	3.93	37	8.18	3.12	0.199
4	3.93	47	8.18	3.12	0.149
5	3.93	57	8.18	3.12	0.109

Table 5: Artificial dataset observing impact of varying age on probability of outcome

Conclusion

The report will conclude with a brief summation of the findings of this exploratory data analysis. The report found that within the Affairs data set, age years married, religiousness and marriage rating were statistically significant on the dependent variable (number of affairs). Using logistic regression we were able to observe the impact of predictors (marriage rating and age) on the outcome of whether an individual would engage in an affair. The results found that the probability of extramarital affair decreased as marriage rating increased as well as the age of an individual.

Appendix

#R CODE#

```
library(dsEssex)
library(AER)
library(tidyverse)
library(na.tools)
library(dplyr)
library(MASS)
library(wesanderson)
library(psych)
library(gridExtra)
library(grid)
library(patchwork)
library(vtable)
library(GGally)
library(tableone)
library(xtable)

data("Affairs")
summary(Affairs)
str(Affairs)
```

```
numaffairs <- Affairs %>%
  select_if(negate(is.factor))


#UNIVARIATE ANALYSIS#

#HISTOGRAM (FIGURE 1)
hist(Affairs$affairs, # histogram
     col="grey", # column color
     border="black",
     prob = TRUE, # show densities instead of frequencies
     xlab = "Number of affairs",
     main = "Histogram plot and density function of number of affairs")
lines(density(Affairs$affairs), # density plot
      lwd = 2, # thickness of line
      col = "red")

#CONVERSION OF VARIABLES TO FACTORS FOR INDEPENDENT BOX PLOTS (FIGURE 2)
Affairs$affairs <- as.factor(Affairs$affairs)
Affairs$religiousness <- as.factor(Affairs$religiousness)
Affairs$occupation <- as.factor(Affairs$occupation)
Affairs$rating <- as.factor(Affairs$rating)
Affairs$age <- as.factor(Affairs$age)
Affairs$education <- as.factor(Affairs$education)
Affairs$yearsmarried <- as.factor(Affairs$yearsmarried)

#GenderBox
B1 <- ggplot(Affairs, aes(fill=affairs, x=gender)) +
  geom_bar(position = "stack") +
  xlab("Gender") + ylab("Number") + guides(fill=guide_legend(title="Number of affairs")) +
  theme(legend.key.size = unit(0.5, 'cm'), #change legend key size
        legend.key.height = unit(0.5, 'cm'), #change legend key height
        legend.key.width = unit(0.5, 'cm'), #change legend key width
        legend.title = element_text(size=10), #change legend title font size
        legend.text = element_text(size=7)) + #change legend text font size
  scale_fill_brewer(palette="Set2")
#AgeBox
B2 <- ggplot(Affairs, aes(fill=affairs, x=age)) +
  geom_bar(position = "stack") +
  xlab("Age") + ylab("Number") + guides(fill=guide_legend(title="Number of affairs")) +
  theme(legend.key.size = unit(0.5, 'cm'), #change legend key size
        legend.key.height = unit(0.5, 'cm'), #change legend key height
        legend.key.width = unit(0.5, 'cm'), #change legend key width
        legend.title = element_text(size=10), #change legend title font size
        legend.text = element_text(size=7)) + #change legend text font size
  scale_fill_brewer(palette="Set2")
```

```
#yearsMarried
B3 <- ggplot(Affairs, aes(fill=affairs, x=yearsmarried)) +
 geom_bar(position = "stack") +
 xlab("Years married") + ylab("Number") + guides(fill=guide_legend(title="Number of affairs")) +
 theme(legend.key.size = unit(0.5, 'cm'), #change legend key size
       legend.key.height = unit(0.5, 'cm'), #change legend key height
       legend.key.width = unit(0.5, 'cm'), #change legend key width
       legend.title = element_text(size=10), #change legend title font size
       legend.text = element_text(size=7)) + #change legend text font size
 scale_fill_brewer(palette="Set2")
#ChildrenBox
B4 <- ggplot(Affairs, aes(fill=affairs, x=children)) +
 geom_bar(position = "stack") +
 xlab("Children") + ylab("Number") + guides(fill=guide_legend(title="Number of affairs")) +
 theme(legend.key.size = unit(0.5, 'cm'), #change legend key size
       legend.key.height = unit(0.5, 'cm'), #change legend key height
       legend.key.width = unit(0.5, 'cm'), #change legend key width
       legend.title = element_text(size=10), #change legend title font size
       legend.text = element_text(size=7)) + #change legend text font size
 scale_fill_brewer(palette="Set2")
#ReligiousnessBox
B5 <- ggplot(Affairs, aes(fill=affairs, x=religiousness)) +
 geom_bar(position = "stack") +
 xlab("Religiousness") + ylab("Number") + guides(fill=guide_legend(title="Number of affairs")) +
 theme(legend.key.size = unit(0.5, 'cm'), #change legend key size
       legend.key.height = unit(0.5, 'cm'), #change legend key height
       legend.key.width = unit(0.5, 'cm'), #change legend key width
       legend.title = element_text(size=10), #change legend title font size
       legend.text = element_text(size=7)) + #change legend text font size
 scale_fill_brewer(palette="Set2")
#EducationBox
B6 <- ggplot(Affairs, aes(fill=affairs, x=education)) +
 geom_bar(position = "stack") +
 xlab("Education") + ylab("Number") + guides(fill=guide_legend(title="Number of affairs")) +
 theme(legend.key.size = unit(0.5, 'cm'), #change legend key size
       legend.key.height = unit(0.5, 'cm'), #change legend key height
       legend.key.width = unit(0.5, 'cm'), #change legend key width
       legend.title = element_text(size=10), #change legend title font size
       legend.text = element_text(size=7)) + #change legend text font size
 scale_fill_brewer(palette="Set2")
#OccupationBox
B7 <- ggplot(Affairs, aes(fill=affairs, x=occupation)) +
 geom_bar(position = "stack") +
 xlab("Occupation") + ylab("Number") + guides(fill=guide_legend(title="Number of affairs")) +
 theme(legend.key.size = unit(0.5, 'cm'), #change legend key size
       legend.key.height = unit(0.5, 'cm'), #change legend key height
```

```
      legend.key.width = unit(0.5, 'cm'), #change legend key width
      legend.title = element_text(size=10), #change legend title font size
      legend.text = element_text(size=7)) + #change legend text font size
  scale_fill_brewer(palette="Set2")
#MarriageRatingBox
B8 <- ggplot(Affairs, aes(fill=affairs, x=rating)) +
  geom_bar(position = "stack") +
  xlab("Self-rating of Marriage") + ylab("Number") + guides(fill=guide_legend(title="Number of affairs")) +
  theme(legend.key.size = unit(0.5, 'cm'), #change legend key size
      legend.key.height = unit(0.5, 'cm'), #change legend key height
      legend.key.width = unit(0.5, 'cm'), #change legend key width
      legend.title = element_text(size=10), #change legend title font size
      legend.text = element_text(size=7)) + #change legend text font size
  scale_fill_brewer(palette="Set2")

#Combining Uni-variate plots (FIGURE 2)
grid.arrange(B1,B2,B3,B4,B5,B6,B7,B8, nrow=4, ncol=2)

#EDA
table(Affairs$affairs)
table(Affairs$age)
table(Affairs$yearsmarried)
table(Affairs$religiousness)
table(Affairs$education)
table(Affairs$occupation)
table(Affairs$rating)

#CORRELATION TABLE
corraffairs <- round(cor(numaffairs),5)
upper<- corraffairs
upper[upper.tri(corraffairs)]<-""
upper<-as.data.frame(upper)
print(xtable(upper), type="html", file = "graph.html")

#LOGISTIC REGRESSION
Affairs$ynaffair[Affairs$affairs >  0] <- 1
Affairs$ynaffair[Affairs$affairs == 0] <- 0
Affairs$ynaffair <- factor(Affairs$ynaffair,levels=c(0,1), labels=c("No","Yes"))
table(Affairs$ynaffair)
#full model
fit.full <- glm(ynaffair ~ gender + age + yearsmarried + children + religiousness + education + occupation +rating, data=Affairs, family=binomial())
summary(fit.full)
#reduced model
fit.reduced <- glm(ynaffair ~ age + yearsmarried + religiousness + rating, data=Affairs, family=binomial())
```

```
summary(fit.reduced)
#Chi squared test
anova(fit.reduced, fit.full, test="Chisq")
coef(fit.reduced)
exp(coef(fit.reduced))

#Impact of self-rating
testdata <- data.frame(rating=c(1, 2, 3, 4, 5), age=mean(Affairs$age),
          yearsmarried=mean(Affairs$yearsmarried),
          religiousness=mean(Affairs$religiousness))

testdata$prob <- predict(fit.reduced, newdata=testdata, type="response")
#Impact of Age
testdata <- data.frame(rating=mean(Affairs$rating), age=seq(17, 57, 10),
                   yearsmarried=mean(Affairs$yearsmarried),
                   religiousness=mean(Affairs$religiousness))
testdata$prob <- predict(fit.reduced, newdata=testdata, type="response")
```