Fair's Affairs Data -- A Logistic Analysis

---Chengcheng Liu

## Abstract

This project used the Logistic regression to analyse the relationship between factors such as . Result show

# 1. Objective

In Fair’s paper (Fair, Ray. 1978. “A Theory of Extramarital Affairs,” Journal of Political Economy`, February, 45-61.), he came up with an economic theory of an individual’s decision to allocate time between work, time spend with spouse, and time spend in affairs in order to maximize his utility. And the decision is based on the persons wage rate, price level, and a set of other potential factors.

In this paper, I want to test his conclusion, and moreover, visualize the effect of factors that can influence a person’s decision to have an affair using Python. In the end, I will suggest a model that can predict a person’s decision of having an affair using the logistic model.

# 2. Data Preparation

The dataset I chose is the Affairs dataset that comes with a standard machine learning data set "Statsmodels". The dataset was derived from a survey conducted in 1974 by Redbook Magazine (RT), in which the participants are married woman. The survey include questionnaire of their participation in extramarital affairs, and their information including the quality of marriage (self-rated), years of marriage, numbers of children, education, the couple's occupation information.

The number of observations is 6366, and the 9 variables included are:

|  |  |
| --- | --- |
| rate\_marriage | woman's rating of her marriage (1 = very poor, 5 = very good) |
| age | woman's age |
| yrs\_married | number of years married |
| children | number of children |
| religious | woman's rating of how religious she is  (1 = not religious, 4 = strongly religious) |
| educ | level of education. (9 = grade school, 12 = high school, 14 = some college,  16 = college graduate, 17 = some graduate school, 20 = advanced degree) |
| occupation | woman's occupation (1 = student, 2 = farming/semi-skilled/unskilled,  3 = "white collar", 4 = teacher/nurse/writer/technician/skilled,  5 = managerial/business, 6 = professional with advanced degree) |
| occupation\_husb | husband's occupation (same coding as woman's occupation) |
| affairs | time spent in extra-marital affairs |

In this survey, a questionnaire on sex was published and readers were asked to mail in the answers. The source data is well structured and do not have null values or outliers in it, so preprocessing is relatively simple. However, it is likely that the survey is not random selected survey, therefore it can potentially cause bias in the following results.

A few summery statistics show the distribution of data through histogram, and we can see after the binary transformation of the affair data, almost 33% individual have at least one affair in their marriage. Grouped summary statistics show that chance of having an affair decrease with rate of marriage (happiness), and increase with years of marriage(but whether the effect is statistically significant still yet to be tested.

# 3. Model

## 3.0. Base Model

In this project, I am using a logistic regression model to identify various factors that may influence a married female’s chance of having an affair. In the raw data, we have data of a person’s time allocated in affair, however in this case I will transform this time to a binary as our dependent variable, to indicate whether this individual have an affair or not.

Summary statistics showed that the chance of individual’s chance of having affair might depend on rate of marriage, years of marriage, numbers of children and other factors. Original data have occupation of both wife and husband as numeric attributes, I will here transform them into categorical dummies as occu\_1, occu\_2, …, , occuh\_6, and omit the first category of both husband and wife to avoid multicollinearity.

After getting a transformation of data I ran regression of these chosen independent variables against “affair”. From the result we can see:

1. Rate of marriage, as expected, has a statistical significant negative effect on probability of having an affair

2. Result show that age has a negative effect while year of marriage has a positive effect (and both effects are statistically significant). However, obviously there is a positive correlation between age and years of marriage (the larger the age, the longer years of marriage). Therefore in the following regression I want to drop the age column and only keep the years of marriage.

3. Having a religious will have a statistically significant negative effect on the probability.

4. Number of child and education effects are not statistically significant.

5. Among the wifes’ occupation and husbands' occupation, husbands' occupation individually does not show statistical significance, I will test join effect (F test) next (See Result 1). However, occupation of wife have a positive influence on probability of having an affair, and F test I believe will show statistical significance (See Result 2).

Logit Regression Results

==============================================================================

Dep. Variable: affair No. Observations: 6366

Model: Logit Df Residuals: 6349

Method: MLE Df Model: 16

Date: Sun, 09 Apr 2017 Pseudo R-squ.: 0.1365

Time: 06:48:10 Log-Likelihood: -3456.2

converged: True LL-Null: -4002.5

LLR p-value: 1.534e-222

=================================================================================

coef std err z P>|z| [95.0% Conf. Int.]

---------------------------------------------------------------------------------

rate\_marriage -0.7102 0.031 -22.560 0.000 -0.772 -0.649

age -0.0613 0.010 -5.936 0.000 -0.082 -0.041

yrs\_married 0.1080 0.011 9.836 0.000 0.086 0.129

children 0.0156 0.032 0.488 0.625 -0.047 0.078

religious -0.3754 0.035 -10.766 0.000 -0.444 -0.307

educ -0.0017 0.017 -0.099 0.921 -0.036 0.032

occu\_2.0 0.3902 0.448 0.872 0.383 -0.487 1.267

occu\_3.0 0.7027 0.441 1.592 0.111 -0.163 1.568

occu\_4.0 0.4714 0.443 1.065 0.287 -0.396 1.339

occu\_5.0 1.0542 0.447 2.360 0.018 0.179 1.930

occu\_6.0 1.1080 0.494 2.242 0.025 0.139 2.077

occuh\_2.0 0.1704 0.186 0.916 0.360 -0.194 0.535

occuh\_3.0 0.2842 0.202 1.406 0.160 -0.112 0.680

occuh\_4.0 0.1428 0.181 0.789 0.430 -0.212 0.498

occuh\_5.0 0.1723 0.183 0.944 0.345 -0.186 0.530

occuh\_6.0 0.1828 0.204 0.897 0.369 -0.216 0.582

intercept 2.9708 0.572 5.192 0.000 1.849 4.092

=================================================================================

Python code 1: print result\_0.f\_test([0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,0] )

Result 1: <F test: F=array([[ 4.34278311]]), p=0.0372060120949, …

Python code 2: print result\_0.f\_test([0,0,0,0,0,1,1,1,1,1,1,0,0,0,0,0,0] )

Result 2:<F test: F=array([[ 2.8446414]]), p=0.0917281747694, …

F test show effect of husband's occupation jointly does have a statistical significantly effect on an individual’s chance of having an affair (significant at 95% confidence level), so does wife’s occupation (statistical significance at 90% level). However, I suspect there is a positive relationship between wife’s occupation and husband’s occupation (the better the husband’s occupation, the better the wife’s occupation), therefore I suspect the insignificance of husband’s occupation can be caused by inter-correlation between variables. Therefore I decided to drop the husband’s occupations in the final regression.

## 3.1. Final Model

After dropping number of children, age and education, we include only years of marriage, rate of marriage, and wife’s occupation to predict individual’s chance of having an affair.

Logit Regression Results

==============================================================================

Dep. Variable: affair No. Observations: 6366

Model: Logit Df Residuals: 6357

Method: MLE Df Model: 8

Date: Sun, 09 Apr 2017 Pseudo R-squ.: 0.1314

Time: 06:13:50 Log-Likelihood: -3476.5

converged: True LL-Null: -4002.5

LLR p-value: 8.257e-222

=================================================================================

coef std err z P>|z| [95.0% Conf. Int.]

---------------------------------------------------------------------------------

rate\_marriage -0.7071 0.031 -22.608 0.000 -0.768 -0.646

yrs\_married 0.0592 0.004 14.599 0.000 0.051 0.067

religious -0.3782 0.035 -10.925 0.000 -0.446 -0.310

occu\_2 0.4091 0.442 0.925 0.355 -0.458 1.276

occu\_3 0.7004 0.437 1.603 0.109 -0.156 1.557

occu\_4 0.4001 0.438 0.912 0.362 -0.459 1.260

occu\_5 1.0117 0.442 2.287 0.022 0.145 1.879

occu\_6 0.9537 0.486 1.963 0.050 0.002 1.906

intercept 1.8194 0.458 3.969 0.000 0.921 2.718

=================================================================================

One thing to note is that, the Pseudo R-square of the base model is 0.1365, while the final model’s Pseudo R-square is 0.1314. When analyzing data with a logistic regression, it is to maximize the likelihood estimates instead of minimizing variance as in the case with OLS regression. Therefore a equivalent R-square statistic does not exist in the Logistic Regression. However the Pseudo R-square is a similar estimate (also range from 0 to 1), and it has a higher values with a better model fit. However, they cannot be interpreted as the one in an OLS regression and with different kind of Pseudo R-square comes different values, ie. Efron’s R-square, Mcfadden’s R-square, etc (Source: UCLA IDRE Stats, <http://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/>). However, we do can compare between the base model, and we can see that, after removing age, husband’s occupation and a couple of other variables, the Pseudo R-square actually did not decrease much.

In the final model, I also calculated the Odds Ratio (Column OR in the chart below) of the logistic model in order to further interpret the model. The odds ratio generated by taking the exponential of each of the coefficients. And it tells how a 1 unit increase or decrease or change in a variable affects the odds of having an affair. For example, we can expect the odds of having an affair to decrease by about 50% if rate of marriage increase by 1.

|  |
| --- |
| Odds Ratio |
| 2.5% 97.5% OR  rate\_marriage 0.463765 0.524254 0.493083  yrs\_married 1.052576 1.069437 1.060973  religious 0.640188 0.733222 0.685128  occu\_2.0 0.632646 3.582452 1.505465  occu\_3.0 0.855705 4.742491 2.014491  occu\_4.0 0.631703 3.523814 1.491980  occu\_5.0 1.155821 6.544633 2.750350  occu\_6.0 1.001604 6.724139 2.595173  intercept 2.511716 15.146620 6.167983 |

We also did the same calculations using the coefficients estimated using the 95% confidence interval to get a better picture of how uncertainty in variables can impart the chance of having an affair (shown in the 2.5%-97.5% column in the chart above). Note that the odds ratio of occupations are comparing to the first category which is student. For example, having an occupation in the second category, comparing to having individual having an occupation in the first category (Student), increase the chance of affair by a probability between 63% and 358% using a 95% confidence interval. (More detail from Odds Ratio at UCLA IDRE <http://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/>)

However, odds ratio does not depend on where the value is held at, I want to further interpret the model using visualization, so we can see how different effects affects the chance of affair at different levels.

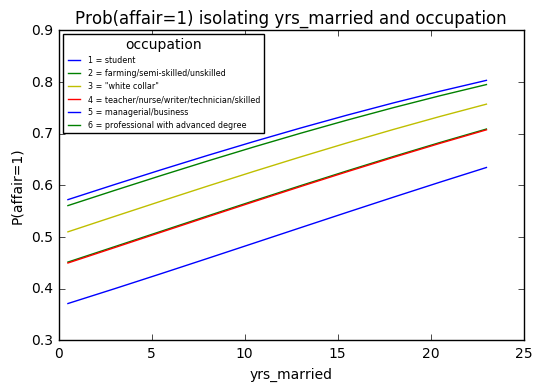
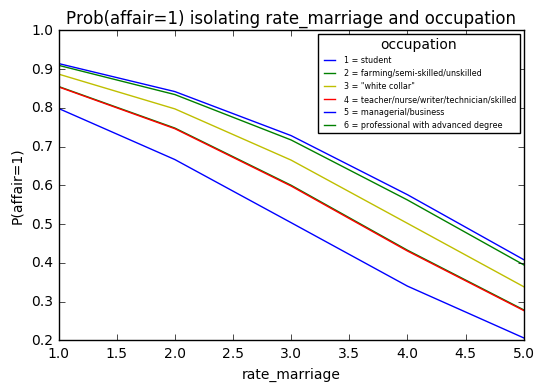
## 3.2, Visualization

Digging a little deeper, as a way to evaluate our dummy transformation, we are going to recreate the dataset with every logical combination of input values, this will allow us to see how the predicted probability of having an affair increases/decreases across different variables.

First I generated all the combinations using a helper function called cartesisian which I found here (<http://stackoverflow.com/questions/1208118/using-numpy-to-build-an-array-of-all-combinations-of-two-arrays>). Then we are going to use np.linspace to create a range of values for rate of marriage and years of marriage. This creates a range of evenly spaced ranges of 10 values from a specified min and max value of observed values.

After generating all the combinations, we generated predictions of these combinations using the estimator in the final model. Using the helper function isolate\_and\_plot, I can compare a given variable with the different occupation and the mean probability for that combination. To isolate the occupation and other variables we used a pivot\_table which allows to easily aggregate the data.

The resulting plots show how rate of marriage and years of marriage affect the chance of affair. We can see how the probability gradually increase/decrease as these variables increase. And that different occupation yield drastic probabilities of affair.

T

After visualization, clearly we can see that as the rate of marriage increase, the probability of having an affair dramatically decreased about 50%, and as years of marriage increase from 0 to 25, probability of affair can increase about 22% percent.

# 4. Conclusion

In Fair’s paper, he came out with a model based on the economic theory people make decision of allocation of time based on the Lagrinian derived from maximizing utility function subject to time constraints and budget constraints, and

Further possible improvements of the model can be, adding interaction items in the model to incorporate cross effects/ using alternative modeling approaches such as Tobit and Decision Tree model, and using training and testing schema to test the model.

3. Analysis or Model:

If you’re conducting an inference test explain the analysis you performed clearly and include well-labelled diagrams to make your points. If you chose to do a predictive model, explain the model, how you trained and tested it, and how well it works. How did you confirm that the data met the requirements for the test or modeling technique to be valid?

rate\_marriage 0.493083

yrs\_married 1.060973

religious 0.685128

occu\_2 1.505465

occu\_3 2.014491

occu\_4 1.491980

occu\_5 2.750350

occu\_6 2.595173

intercept 6.167983

dtype: float64

Discrete Choice Models

[Link to Notebook GitHub](https://github.com/statsmodels/statsmodels/blob/master/examples/notebooks/discrete_choice_example.ipynb)

Fair's Affair data

A survey of women only was conducted in 1974 by *Redbook* asking about extramarital affairs.

In [1]:

**from** **\_\_future\_\_** **import** print\_function

**import** **numpy** **as** **np**

**from** **scipy** **import** stats

**import** **matplotlib.pyplot** **as** **plt**

**import** **statsmodels.api** **as** **sm**

**from** **statsmodels.formula.api** **import** logit, probit, poisson, ols

In [2]:

**print**(sm.datasets.fair.SOURCE)

Fair, Ray. 1978. "A Theory of Extramarital Affairs," `Journal of Political

Economy`, February, 45-61.

The data is available at http://fairmodel.econ.yale.edu/rayfair/pdf/2011b.htm

In [3]:

**print**( sm.datasets.fair.NOTE)

::

Number of observations: 6366

Number of variables: 9

Variable name definitions:

rate\_marriage : How rate marriage, 1 = very poor, 2 = poor, 3 = fair,

4 = good, 5 = very good

age : Age

yrs\_married : No. years married. Interval approximations. See

original paper for detailed explanation.

children : No. children

religious : How relgious, 1 = not, 2 = mildly, 3 = fairly,

4 = strongly

educ : Level of education, 9 = grade school, 12 = high

school, 14 = some college, 16 = college graduate,

17 = some graduate school, 20 = advanced degree

occupation : 1 = student, 2 = farming, agriculture; semi-skilled,

or unskilled worker; 3 = white-colloar; 4 = teacher

counselor social worker, nurse; artist, writers;

technician, skilled worker, 5 = managerial,

administrative, business, 6 = professional with

advanced degree

occupation\_husb : Husband's occupation. Same as occupation.

affairs : measure of time spent in extramarital affairs

See the original paper for more details.

In [4]:

dta = sm.datasets.fair.load\_pandas().data

In [5]:

dta['affair'] = (dta['affairs'] > 0).astype(float)

**print**(dta.head(10))

rate\_marriage age yrs\_married children religious educ occupation occupation\_husb \

0 3 32 9.0 3.0 3 17 2 5

1 3 27 13.0 3.0 1 14 3 4

2 4 22 2.5 0.0 1 16 3 5

3 4 37 16.5 4.0 3 16 5 5

4 5 27 9.0 1.0 1 14 3 4

5 4 27 9.0 0.0 2 14 3 4

6 5 37 23.0 5.5 2 12 5 4

7 5 37 23.0 5.5 2 12 2 3

8 3 22 2.5 0.0 2 12 3 3

9 3 27 6.0 0.0 1 16 3 5

affairs affair

0 0.111111 1

1 3.230769 1

2 1.400000 1

3 0.727273 1

4 4.666666 1

5 4.666666 1

6 0.852174 1

7 1.826086 1

8 4.799999 1

9 1.333333 1

[10 rows x 10 columns]

In [6]:

**print**(dta.describe())

rate\_marriage age yrs\_married children religious educ \

count 6366.000000 6366.000000 6366.000000 6366.000000 6366.000000 6366.000000

mean 4.109645 29.082862 9.009425 1.396874 2.426170 14.209865

std 0.961430 6.847882 7.280120 1.433471 0.878369 2.178003

min 1.000000 17.500000 0.500000 0.000000 1.000000 9.000000

25% 4.000000 22.000000 2.500000 0.000000 2.000000 12.000000

50% 4.000000 27.000000 6.000000 1.000000 2.000000 14.000000

75% 5.000000 32.000000 16.500000 2.000000 3.000000 16.000000

max 5.000000 42.000000 23.000000 5.500000 4.000000 20.000000

occupation occupation\_husb affairs affair

count 6366.000000 6366.000000 6366.000000 6366.000000

mean 3.424128 3.850141 0.705374 0.322495

std 0.942399 1.346435 2.203374 0.467468

min 1.000000 1.000000 0.000000 0.000000

25% 3.000000 3.000000 0.000000 0.000000

50% 3.000000 4.000000 0.000000 0.000000

75% 4.000000 5.000000 0.484848 1.000000

max 6.000000 6.000000 57.599991 1.000000

[8 rows x 10 columns]

In [7]:

affair\_mod = logit("affair ~ occupation + educ + occupation\_husb"

"+ rate\_marriage + age + yrs\_married + children"

" + religious", dta).fit()

Optimization terminated successfully.

Current function value: 0.545314

Iterations 6

In [8]:

**print**(affair\_mod.summary())

Logit Regression Results

==============================================================================

Dep. Variable: affair No. Observations: 6366

Model: Logit Df Residuals: 6357

Method: MLE Df Model: 8

Date: Sun, 01 Feb 2015 Pseudo R-squ.: 0.1327

Time: 09:32:42 Log-Likelihood: -3471.5

converged: True LL-Null: -4002.5

LLR p-value: 5.807e-224

===================================================================================

coef std err z P>|z| [95.0% Conf. Int.]

-----------------------------------------------------------------------------------

Intercept 3.7257 0.299 12.470 0.000 3.140 4.311

occupation 0.1602 0.034 4.717 0.000 0.094 0.227

educ -0.0392 0.015 -2.533 0.011 -0.070 -0.009

occupation\_husb 0.0124 0.023 0.541 0.589 -0.033 0.057

rate\_marriage -0.7161 0.031 -22.784 0.000 -0.778 -0.655

age -0.0605 0.010 -5.885 0.000 -0.081 -0.040

yrs\_married 0.1100 0.011 10.054 0.000 0.089 0.131

children -0.0042 0.032 -0.134 0.893 -0.066 0.058

religious -0.3752 0.035 -10.792 0.000 -0.443 -0.307

===================================================================================

How well are we predicting?

In [9]:

affair\_mod.pred\_table()

Out[9]:

array([[ 3882., 431.],

[ 1326., 727.]])

The coefficients of the discrete choice model do not tell us much. What we're after is marginal effects.

In [10]:

mfx = affair\_mod.get\_margeff()

**print**(mfx.summary())

Logit Marginal Effects

=====================================

Dep. Variable: affair

Method: dydx

At: overall

===================================================================================

dy/dx std err z P>|z| [95.0% Conf. Int.]

-----------------------------------------------------------------------------------

occupation 0.0293 0.006 4.744 0.000 0.017 0.041

educ -0.0072 0.003 -2.538 0.011 -0.013 -0.002

occupation\_husb 0.0023 0.004 0.541 0.589 -0.006 0.010

rate\_marriage -0.1308 0.005 -26.891 0.000 -0.140 -0.121

age -0.0110 0.002 -5.937 0.000 -0.015 -0.007

yrs\_married 0.0201 0.002 10.327 0.000 0.016 0.024

children -0.0008 0.006 -0.134 0.893 -0.012 0.011

religious -0.0685 0.006 -11.119 0.000 -0.081 -0.056

===================================================================================

In [11]:

respondent1000 = dta.ix[1000]

**print**(respondent1000)

rate\_marriage 4.000000

age 37.000000

yrs\_married 23.000000

children 3.000000

religious 3.000000

educ 12.000000

occupation 3.000000

occupation\_husb 4.000000

affairs 0.521739

affair 1.000000

Name: 1000, dtype: float64

In [12]:

resp = dict(zip(range(1,9), respondent1000[["occupation", "educ",

"occupation\_husb", "rate\_marriage",

"age", "yrs\_married", "children",

"religious"]].tolist()))

resp.update({0 : 1})

**print**(resp)

{0: 1, 1: 3.0, 2: 12.0, 3: 4.0, 4: 4.0, 5: 37.0, 6: 23.0, 7: 3.0, 8: 3.0}

In [13]:

mfx = affair\_mod.get\_margeff(atexog=resp)

**print**(mfx.summary())

Logit Marginal Effects

=====================================

Dep. Variable: affair

Method: dydx

At: overall

===================================================================================

dy/dx std err z P>|z| [95.0% Conf. Int.]

-----------------------------------------------------------------------------------

occupation 0.0400 0.008 4.711 0.000 0.023 0.057

educ -0.0098 0.004 -2.537 0.011 -0.017 -0.002

occupation\_husb 0.0031 0.006 0.541 0.589 -0.008 0.014

rate\_marriage -0.1788 0.008 -22.743 0.000 -0.194 -0.163

age -0.0151 0.003 -5.928 0.000 -0.020 -0.010

yrs\_married 0.0275 0.003 10.256 0.000 0.022 0.033

children -0.0011 0.008 -0.134 0.893 -0.017 0.014

religious -0.0937 0.009 -10.722 0.000 -0.111 -0.077

===================================================================================

In [14]:

affair\_mod.predict(respondent1000)

Out[14]:

array([ 0.5188])

In [15]:

affair\_mod.fittedvalues[1000]

Out[15]:

0.075161592850634618

In [16]:

affair\_mod.model.cdf(affair\_mod.fittedvalues[1000])

Out[16]:

0.51878155721216501

The "correct" model here is likely the Tobit model. We have an work in progress branch "tobit-model" on github, if anyone is interested in censored regression models.