

婚姻幸福的影响因素

中国传媒大学广告学（计算广告双学士学位复合型人才培养项目）

杨译博

2022/07/07

问题概述

当年轻人们希望自己有一段完美幸福的婚姻，但是有些人或受到他人恐吓，或看到一些影视剧及新闻报道，开始担忧自己未来的婚姻幸福问题，甚至产生了恐婚现象。那么就派生出一个问题，婚姻幸福到底与那些因素有关？我选取了三个方面（个人理念，客观因素，现实状况）他们分别为：“对中国有关性方面的态度”，“个人出生日期”“家庭年总收入”进行研究，进而分析三个大方面对婚姻幸福感的影响。

采用方法主要是 多元线性回归分析

文献综述

郭玲学者有发表《婚姻家庭幸福感报告 揪出幸福婚姻的隐形杀手》的报告，与其他学者类似，基本只涉及到了婚姻中双方的性格，行为等主观因素影响，基本局限于“个人理念”这一方面，而且对于预防的作用不是很大。目前我想到的改进方式就是增加原因种类的分析，并且不是凭空增加，都是在我们相亲时彼此会询问的问题，相处时注重对方性格，家长会看重经济工作情况，另外加入我自己的一个思想，随着时代的发展，我们这一代人和老一辈一代的婚姻，也有着很多不同，以往可能会保留一些传统观念，如今可能更加多元开放，为爱执着，婚姻状况这个东西，是不是也会随着时代的更迭，出现了变数，对此条信息的研究，选用了被研究者的年龄这个指标。综上，我选择加上了这些因素进行研究，希望可以在婚姻幸福这个问题上做到一些小补充。

模型构建或方法

因为是研究相关性，并且自己期望这些因素能够和因变量有一定的相关性，于是打算使用多元线性回归分析这一方法，构建如下模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu$$

并通过 python 的 statsmodels 库对数据进行处理分析，从而进行回归计算。

数据分析

首先是读取数据，数据来源于老师提供的“cgss2015.csv”文件，从中选取”问题概述“中提到的三个小方面，利用 python 进行处理，首先对数据进行提取，经过思考选取了‘a38’，‘a39’，‘a40’，‘a41’，‘a421’，‘a422’，‘a423’，‘a424’，‘a425’，‘a62’，‘a69’，‘a301’几组数据进行分析，并且对前九组数据求和形成关于“个人观念”形象分析的数据“H66”。提取代码如下：

```
1 import os
2 import pandas as pd
3
4 path1 = "D:\\996"
5 path2 = "./data"
6 if not os.path.exists(path2):
7     os.mkdir(path2)
8
9 for filename in os.listdir(path1):
10     if filename.endswith(".csv"):
11         file_path1 = path1 + "/" + filename
12         df1 = pd.read_csv(file_path1, encoding='gbk', low_memory=False)
13         df2 = df1[['a38', 'a39', 'a40', 'a41', 'a421', 'a422', 'a423',
14                 'a424', 'a425', 'a62', 'a69', 'a301']]
15         df2.to_csv(path2 + "/" + filename,
16                 index=False, encoding="gb2312")
```

再根据我们提到的研究方法—多元线性分析，对上述得到的数据进行处理，得到分析所需要的线性回归分析表格，操作代码如下（后文代码与此基本一致）：

```

1 import pandas as pd
2 import numpy as np
3 import csv
4 data = pd.DataFrame()
5 data = pd.DataFrame(pd.read_csv('D:\\996\\grtd.csv', encoding="gbk"))
6
7 y = np.array(data['a69'])
8 x1 = np.array(data['H66'])
9 x2 = np.array(data['a62'])
10 x3 = np.array(data['a301'])
11
12 import scipy.stats as stats
13 r1 = stats.pearsonr(x1,y)[0]
14 r2 = stats.pearsonr(x2,y)[0]
15 r3 = stats.pearsonr(x3,y)[0]
16
17 import pandas.tseries
18 import statsmodels.api as sm
19 import pandas as pd
20 import numpy as np
21
22 vars = ['a69', 'H66', 'a62', 'a301']
23 df = data[vars]
24
25 from patsy import dmatrices
26 y,X = dmatrices('a69~H66+a62+a301', data = df,
27 return_type = 'dataframe')
28
29 model = sm.OLS(y,X)
30 result = model.fit()
31 print(result.summary())

```

分析三组数据得到如下的结果：

```

===== OLS Regression Results =====
=====
Dep. Variable: a69 R-squared: 0.252
Model: OLS Adj. R-squared: 0.251
Method: Least Squares F-statistic: 1225.
Date: Thu, 07 Jul 2022 Prob (F-statistic): 0.00
Time: 16:12:28 Log-Likelihood: -17865.
No. Observations: 10943 AIC: 3.574e+04
Df Residuals: 10939 BIC: 3.577e+04
Df Model: 3
Covariance Type: nonrobust
=====
coef std err t P>|t| [95.0% Conf. Int.]
-----
Intercept 86.4378 1.380 62.616 0.000 83.732 89.144
H66 -0.0026 0.001 -1.969 0.049 -0.005 -1.15e-05
a62 -2.994e-08 4.17e-08 -0.717 0.473 -1.12e-07 5.19e-08
a301 -0.0423 0.001 -60.144 0.000 -0.044 -0.041
=====
Omnibus: 2145.021 Durbin-Watson: 1.969
Prob(Omnibus): 0.000 Jarque-Bera (JB): 3830.770
Skew: 1.253 Prob(JB): 0.00
Kurtosis: 4.456 Cond. No. 3.40e+07
=====

```

从这张得到的表格分析，F 值是较高的 1225，P 值基本就为 0，可见这个模型整体来说还是比较显著的，具有一定的参考性，并且很稳定，虽说决定系数 R 仅为 0.252，修正决定系数为 0.251，并不是像其他基本完全线性相关的模型能够达到特别接近一的零点九几，感觉并没有很符合，但是关于这个问题，要多想几个方面，我们研究的算是一个社会问题，而这种问题有一个很大的特点，就是这个因变量是由成百上千的自变量引起的，所以在此种情况下，想要揪出几个因素来基本上决定性地影响此因变量，几乎是不可能事件，相反在这种情况下，能够达到 0.252 的数值，我认为已经是达到对因变量有很大影响的地步了，下面可以具体分析一下，首先是我们所创建的模型如下：

$$A69 = 86.4378 - 0.0026H66 - (2.994e + 0.8) a62 - 0.0423a301$$

从他们的相关系数来看，相对较小的“出生日期”，“个人观念”，只有绝对值不到 1 的系数，而我们都比较关心的“年总收入”问题则是相对较高的，难道说就像传统意义上说的，金钱问题基本能解决一个家庭的婚姻幸福？只能说有可能，毕竟我们不能通过一个模型得到最准确的结论，这三个因素也不能说只是这三个因素，他们彼此也会牵扯到其他的因素，可能合在一起的效果对他们的影响力产生了指数性的增强。只能说我们在此可以作出一个假设：“年总收入”的影响在选择到的三个因素下高于“出生日期”以及“个人观念”。带着这个假设，继续下面的模型构建。

那么就一个一个来吧，第一个就先构造“个人理念”的模型，结果如下：

```
.....OLS Regression Results.....
=====
Dep. Variable:.....a69.....R-squared:.....0.003
Model:.....OLS.....Adj. R-squared:.....0.003
Method:.....Least Squares.....F-statistic:.....36.36
Date:.....Thu, 07 Jul 2022.....Prob (F-statistic):.....1.69e-09
Time:.....16:52:43.....Log-Likelihood:.....-19433.
No. Observations:.....10944.....AIC:.....3.887e+04
Df Residuals:.....10942.....BIC:.....3.889e+04
Df Model:.....1.....
Covariance Type:.....nonrobust.....
=====
.....coef.....std.err.....t.....P>|t|.....[95.0% Conf. Int.]
-----
Intercept.....3.4213.....0.033.....104.179.....0.000.....3.357.....3.486
H66.....-0.0092.....0.002.....-6.030.....0.000.....-0.012.....-0.006
=====

=====
.....coef.....std.err.....t.....P>|t|.....[95.0% Conf. Int.]
-----
Intercept.....3.4213.....0.033.....104.179.....0.000.....3.357.....3.486
H66.....-0.0092.....0.002.....-6.030.....0.000.....-0.012.....-0.006
=====
Omnibus:.....2690.988.....Durbin-Watson:.....1.967
Prob(Omnibus):.....0.000.....Jarque-Bera (JB):.....5986.767
Skew:.....1.410.....Prob(JB):.....0.00
Kurtosis:.....5.275.....Cond. No.....52.1
=====
```

从表格分析得到，F 值不高，只有 36.36，P 值 $1.36e - 0.9$ 也不是特别接近 0，整体不显著了，而且决定系数 R 值和修正决定系数也均为极低的 0.003。整体的模型为：

$$a36 = 3.4213 - 0.0092H66$$

这个结果在第一个模型下，显得理所当然，在两者下的结果都是表示“个人观念”这个因素对于“婚姻状况”这个因变量几乎无影响。这时一定有人会有疑问，“个人观念”往深处说就是“性格”，性格这种东西怎么可能对婚姻状况的影响这么小呢？这个问题已经从数据上说明了，那现在可以从现实上进行一个小分析，或许不可思议，但是是情理之中的，我们都不否认一个观点，“任何性格的人都有人喜欢”，现在这个数据结果算是大多数都是已婚人士，于是乎就排除了性格不合导致婚姻状态的影响，因为大多数人来说是了解对方后才去结婚的，那么这一点就分析到这里了。

以下进行第二个数据组的分析，先放结果：

```

.....OLS Regression Results.....
=====
Dep. Variable:.....a69.....R-squared:.....0.001
Model:.....OLS.....Adj. R-squared:.....0.001
Method:.....Least Squares.....F-statistic:.....8.127
Date:.....Thu, 07 Jul 2022.....Prob. (F-statistic):.....0.00437
Time:.....17:24:45.....Log-Likelihood:.....-19495.
No. Observations:.....10967.....AIC:.....3.899e+04
Df Residuals:.....10965.....BIC:.....3.901e+04
Df Model:.....1.....
Covariance Type:.....nonrobust.....
=====
.....coef.....std.err.....t.....P>|t|.....[95.0% Conf. Int.]
-----
Intercept.....3.2513.....0.014.....231.590.....0.000.....3.224.....3.279
a62.....-1.374e-07.....4.82e-08.....-2.851.....0.004.....-2.32e-07.....-4.29e-08
=====
Omnibus:.....2712.944.....Durbin-Watson:.....1.966
Prob(Omnibus):.....0.000.....Jarque-Bera (JB):.....6053.427
Skew:.....1.417.....Prob(JB):.....0.00
Kurtosis:.....5.284.....Cond. No. ....2.99e+05
=====

```

这个数据，可能有些出乎意料，但是他确实是“年总收入”这一组数据，不想一一列举了，很低的 R 值 8.127，并不是很接近 0 的 0.00437，以及很低的 R 值 0.001，这真的是第一组总数据里分析得到契合度最高的那一组嘛？难道说是数据出错了？确实不排除这个可能，但是数据摆在这不要去想特殊情况了，一定是因果必有因的，现在细分析这组数据，数据内容是婚姻双方一年的总收入情况，这个东西一定是因人而异，有多有少，但本质上是反应一个家庭的经济情况，这个因素在婚前考虑的人多半不是婚姻受益者双方，而是彼此的家人父母，两人踏入婚姻并且能够在婚姻中体会到幸福，一定是因为彼此相爱，而不是金钱，所以单论收入这个因素对婚姻状况的影响，一定是低的，那为什么在多种数据共同影响下金钱会占据较为主导的因素呢？李云勇在研究杨振宁妻子的文章《翁帆婚姻的附加值》中提到“婚姻附加值就是婚姻另一半的个人能力、家庭财产、政治背景、父母情况等对婚姻将会带来眼前或者长远的各种重要利益。”在婚姻情况大致确定的时候，金钱的附加值是比其他东西要高一些的，就是说对情感的调和幅度较大，所以在第一组模型下，它所体现的影响力要大得多，但是把它单拎出来，并没有那么大的效果。

举个不恰当的例子，收入这个东西类似于化学反应里的“催化剂”，在主反应进行中，能加快主反应的进程，但是如果只有催化剂，基本上是不反应的，就大概类似于这个原理。

马上来到我们最后一组数据，直接放结果了：

```

===== OLS Regression Results =====
Dep. Variable: a69 R-squared: 0.251
Model: OLS Adj. R-squared: 0.251
Method: Least Squares F-statistic: 3676.
Date: Thu, 07 Jul 2022 Prob (F-statistic): 0.00
Time: 19:52:21 Log-Likelihood: -17915.
No. Observations: 10968 AIC: 3.583e+04
Df Residuals: 10966 BIC: 3.585e+04
Df Model: 1
Covariance Type: nonrobust
=====
coef std err t P>|t| [95.0% Conf. Int.]
-----
Intercept 86.6602 1.376 62.981 0.000 83.963 89.357
a301 -0.0425 0.001 -60.627 0.000 -0.044 -0.041
=====
Omnibus: 2150.310 Durbin-Watson: 1.967
Prob(Omnibus): 0.000 Jarque-Bera (JB): 3838.730
Skew: 1.254 Prob(JB): 0.00
Kurtosis: 4.453 Cond. No. 2.28e+05
=====

```

这个数据看起来就心情好一些了，高达 3676 的 F 值，等于 0 的 P 值，模型既显著又稳定，决定系数 R 值和修正决定系数也还不错，这个模型大体上来说是能说明“出生日期”这个因素是对“婚姻状况”有一定的影响力的。模型为：

$$A69 = 86.6602 - 0.0425a301$$

简要分析的话，就来到了我提出问题时的那个猜想，时代在进步，理念也在变化，对于婚姻的理解在变化，古时候包办婚姻，一夫多妻制，那时候的夫妻双方并不像现在如此亲密，甚至好多婚姻关系是为了政治结盟，于是乎婚姻状况也自然不会很好，当然数据里面是没有古代的人的，但是单从这个思想就能够说明问题，年纪稍大一些的人相对来说传统观念大一些，婚礼的进行大多是为了家族的传承，也都是相亲为主，当然不代表婚姻状况不好，反倒是当时物资相对贫乏的年代，两个人之间相互依靠才能撑起一个家，，依赖性更高，所以婚姻状况更稳定，当今的时代，虽说技术发达，思想理念更加先进，男女关系平等，也更加敢爱敢恨，情感也更加复杂，也因为物质资源的丰厚，夫妻之间的依赖性更小了，产生了一定的两极分化，因爱结合在一起的两人，可能因为这种爱长久幸福在一起，也可能因为当代人的自由洒脱，过于注重自我，造成婚姻的破裂，这也是为什么当今离婚率不低的原因。数据已经正视了“出生日期”对于婚姻状况的影响，也提醒我们去改善，既然现状是如此，原因也能抓到，就是我们做出行动的时候了，这也就是作研究的意义。

小结

总结下来，选择分析的三组数据：“出生日期”，“年总收入”，“个人理念”中，在共同作用时，由于婚姻附加值的作用，“年总收入”一项为影响最大的一项，在每个因素单独作用时，影响最大的是“出生日期”。

具有的问题还是很大的，第一点是数据上不是一定的准确；

第二点是数据的种类选择并不多，也没有支持分析数据的数据，文中分析数据时没有理论数据支撑，可靠性不高；

第三点是缺少一定的文献支持，网络上对于“出生日期”这一类东西的研究较少，缺少对比。

最后一点是可能模型用的不是最好的，只是想到了这个模型，就进行了应用，可能有其他模型更适合这个问题的研究。

参考文献

- 1.《婚姻家庭幸福感报告 揪出幸福婚姻的隐形杀手》—郭玲
- 2.《翁帆婚姻的附加值》---李云勇