In [ ]:

```
import numpy as np
import statsmodels.stats.contingency_tables as tbl

# 这里必须使用np.asarray函数进行转换，否则后续计算可能报错
table = tbl.Table2x2(np.asarray(pd.crosstab(ccss.Ts9, ccss.O1)))
table
```

In [ ]:

```
table.oddsratio
```

In [ ]:

```
table.summary()
```

7.4 实战练习

使用适当的指标表述职业和总信心指数之间的关联性。

使用适当的指标表述职业和汽车拥有情况之间的关联性。

8 线性回归模型入门

8.1 线性回归模型的基本原理

8.1.1 相关与回归的区别和联系

8.1.2 线性回归模型概述

8.1.3 线性回归模型的适用条件

8.1.4 线性回归模型的标准建模步骤

8.2 线性回归模型的Python实现

8.2.1 scipy的实现方式

scipy.stats.linregress(

x, y ： 类数组格式的自变量、因变量，均为一维，也可以直接以k*2的二维数组格式提供
注意：该命令的参数格式是自变量x在前！

)

返回结果：

slope : 回归系数b
intercept : 常数项a
r-value : 两个变量的相关系数
p-value : 回归系数的双侧检验
stderr : 回归系数的标准误

In [ ]:

```
# 建立年龄和总信心指数的回归方程
ss.linregress(ccss.s3, ccss.index1)
```

In [ ]:

```
# 以k*2形式的二维数组提供数据
ss.linregress(ccss.loc[:, ['s3', 'index1']])
```

8.2.2 statsmodels的实现方式

class statsmodels.regression.linear_model.OLS(

endog : 因变量，1维数组格式
exog = None : n*k格式数组，k代表自变量数量
missing = 'none' : 对缺失值的处理方式
'none' : 不做任何检查
'drop' : 发现缺失值后该案例直接删除
'raise' : 检查并抛出错误
hasconst = None : T/F，是否允许用户自定义的常数被纳入方程，模型不默认常数项

)

无缺失值时拟合单自变量模型

In [ ]:

```
# 在数据集中加入常数项
dfreg = ccss.loc[:, ['s3']]
dfreg['cons'] = 1
dfreg.head()
```

In [ ]:

```
from statsmodels.regression.linear_model import OLS

regmodel = OLS(ccss.index1, dfreg[['cons', 's3']]).fit()
```

In [ ]:

```
regmodel.summary()
```

拟合多自变量模型

In [ ]:

```
# 在数据集中加入常数项
dfreg = ccss.loc[:, ['s2', 's3', 'Qs9']]
dfreg['cons'] = 1
dfreg.head()
```

In [ ]:

```
# 将性别转换为数值变量
dfreg.replace(['男', '女'], [1, 2], inplace = True)
```

In [ ]:

```
from statsmodels.regression.linear_model import OLS

regmodel = OLS(ccss.index1, dfreg, missing = 'drop').fit()
```

In [ ]:

```
regmodel.summary()
```

残差分析

In [ ]:

```
resdf = pd.DataFrame({'fit' : regmodel.fittedvalues,
                      'resid' : regmodel.resid,
                      'zresid' : regmodel.resid_pearson})
resdf
```

In [ ]:

```
matplotlib.rcParams['font.sans-serif'] = ['Arial'] # 处理负号显示问题

resdf.resid.plot.hist()
```

In [ ]:

```
resdf.zresid.plot.hist()
```

In [ ]:

```
resdf.plot.scatter('fit', 'resid')
```

In [ ]:

```
resdf.plot.scatter('fit', 'zresid')
```

8.3 实战练习

仿照本章的案例，分别考察性别、年龄、家庭收入等变量对现状指数、预期指数的影响，并对相应的模型进行化简，剔除无统计意义的变量，并完成模型的残差诊断。