

0.1 Statistical Models

Statsmodels is probably the most common Python package for statistical models. Statsmodels(<http://www.statsmodels.org/stable/>) is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator. The results are tested against existing statistical packages to ensure that they are correct.

0.1.1 Simple OLS

在调用相关的package之前，我们需要先理解模型的意思，比如下面这个多元回归方程：

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

明白模型的意思，在于，我们要输入给模型的是什么类型的数据？想要的结果是什么？

通常而言，我们要做的，或者说常见的最小二乘法OLS想做的内容是拟合出一个最优化的模型。这个模型能够较好的表达几个变量之间的关系。我们是采用如下步骤来使用Statsmodels:

1. 处理数据 X 和 $Y(s)$ 为适合的数据类型。
2. 选择模型.如有需要，用“`add_constant()`”添加constant.
3. 对模型做拟合得到模型的结果 “`model.fit()`”
4. 获得模型的summary， “`results.summary()`”
5. 如有需要从 `results` 从获取各个参数

具体做法可以参考：<https://www.statsmodels.org/dev/examples/index.html>。里面包含了主要的统计模型的例子，包括time series的。至于2/3SLS等模型无非是简单OLS的多次运行或变形，具体操作是相同的。

0.2 Machine Learning Models

Sci-kit learn 是Python主要的机器学习package。与统计模型相类似的，甚至是更为重要的是，在跑模型之前，一定要先想清楚要选择的是什么模型？这个模型的基本要求是什么，输入和输出的数据类型是什么？

一般来说，具体做法如下：

1. 选择一个合适的模型
2. 设置模型的相关参数
3. 把数据整合为一个feature matrix和一个target vector. Scikit-Learn通常只接受 NumPy arrays, Pandas DataFrames, SciPy sparse matrices 几种数据类型
4. 调用fit()模型拟合
5. 把模型应用到新的数据上
 - 对于有监督学习来说，我们一般调用predict()方法
 - 对无监督学习，我们则调用transform()或predict()

具体实例，可以参考：https://scikit-learn.org/stable/auto_examples/index.html