**Picking the Portfolio Dataset**

Predicting stock returns is always a challenging task. This dataset contains weekly stock returns that spans about 10 years (494 weeks). In each week, a clustering algorithm combine stocks in New York Stock Exchanges into 30 groups and compute several features based on the pricing theory. These features are the loadings of key pricing factors such as excess market return, small-stock-minus-big-stock (SMB) return, high book-to-market ratio minus low book-to-market ratio return (HML). Based on the pricing theory, these features should be able to predict the dynamics of stock returns. In the following homework assignments and mini projects, you are going to explore the data and construct a prediction model that is able to profit from the stock market.

This dataset is divided into a training and a testing sets. The training set contains records from week 1 to week 370. The testing set contains records from week 371 to week 494. The following table lists the sample records in the training set:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| week\_  index | group | alpha | beta\_mkt | beta\_hml | beta\_smb | sigma | week\_return1 | week\_return4 | week\_return13 |
| 1 | 1 | 0.116891 | 2.051794 | 1.593424 | 3.217965 | 0.271982 | 0.892801 | 0.900921 | 0.895643 |
| 1 | 2 | 0.033864 | 3.903705 | 3.1536 | 5.176311 | 0.099884 | 0.886809 | 0.875238 | 0.886578 |
| 1 | 3 | 0.006197 | 1.664317 | 1.515463 | -1.03395 | 0.033168 | 1.013198 | 1.030153 | 1.02992 |
| 1 | 4 | 0.010429 | 2.791942 | 0.975734 | 3.03268 | 0.068875 | 0.855544 | 0.842337 | 0.846865 |
| 1 | 5 | 0.005181 | 2.382342 | 3.754779 | 1.944673 | 0.034964 | 1.029054 | 1.036371 | 1.022844 |
| 1 | 6 | 0.015829 | 0.125074 | 1.126783 | 0.802143 | 0.108948 | 0.977428 | 0.964134 | 0.997845 |
| 1 | 7 | 0.005273 | 0.725278 | 1.851199 | 0.988507 | 0.046013 | 0.961787 | 0.952552 | 0.97047 |

The column descriptions are:

* week\_index: runs from 1 to 494.
* group: group id; runs from 1 to 30. Note that group ID across different week has no relations. It is unreasonable to assume that group 1 in week 1 is related to group I in week 2.
* alpha: factor loading for the constant term.
* beta\_mkt: MKT factor loading.
* beta\_hml: HML factor loading.
* beta\_smb: SMB factor loading.
* sigma: variance of white noise in this group.
* week\_return1: future (gross) return in the following week. A return larger than one means the portfolio is making money. A return less than one means the portfolio is losing money.
* week\_return4: future (gross) return in the following 4 weeks.
* week\_return13: future (gross) return in the following 13 weeks.