Handout for "...and the Cross Section of Expected Returns"

Xiang Fang

April 25, 2015

1 Motivation and Background

In the past 40 years, there are hundreds of papers exploring into finding factors to explain the cross section of asset returns. With so many factors discovered, the single hypothesis testing procedure will generate many spurious factors. The paper proposes a multiple testing framework, providing more stringent criteria for statistical test, which vary over the total number of factors existing.

Some background knowledge of cross section asset returns

Now suppose the current standard factor includes only market return (we are in the CAPM era), now we want to test whether size is a factor. There are two ways to test a factor:

Method1: We sort all firms according to firm size (measured by market cap), and form into 10 portfolios, from smallest to biggest. We take the short-long strategy: short the biggest firm portfolio and long the smallest firm portfolio, which can give us a time series of return. We run the time series regression: $r_t^{SMB} = \alpha + \beta r_t^{market} + \epsilon_t$. If the α in this regression is significant, we say we find a new factor, which has not been priced by the existing factors.

Method2: We sort firms according to size into 10 portfolios, and get SMB (small minus big) return time series. Now we use it as a new factor to price the stylized portfolio (such as Fama-French portfolios), with regression $\bar{r}_i^{FF} = \alpha + \lambda_M \beta_i^M + \lambda_{size} \beta_i^{size} + \epsilon_i$, where βs are obtained from time series regressions. If λ_{size} is significant, we say we find a new factor.

This part is laid out in footnote 35 in the paper.

2 A Multiple Testing Framework

Multiple testing: We are not focusing on the validity of any single hypothesis, but on a synthesis of a class of hypotheses. Suppose we now have 10 existing factors, and find a new one. The multiple testing framework cares about the validity of all 11 hypotheses, not only the newly discovered one.

"Validity": In a single hypothesis testing framework, we have type-I and type-II error. There are counterparts of the two types of error in the multiple testing framework.

Type-I error (false discoveries, or false rejections of the null):

Definition 1 (FWER): Probability of having at least one false discoveries in all the tests (including the existing and the newly proposed one).

Definition 2 (FDR): Probability of the false discovery rate in all the tests exceeding a certain

ratio.

Obviously, FWER is more stringent than FDR. When the number of tests is small, we may prefer FWER. When the number of tests is large, we don't care too much of making a few mistakes, so that we may prefer FDR.

Type-II error(false acceptances of the null): There is conflict of Type-I and Type-II error. The true Type-II error is hard to assess, so that in practice we push Type-I error to a prespecified level, thus not make the test much too stringent.

How to control FWER and FDR? p-value adjustment

Adjustment 1(Bonferroni): $p_i^B = min\{Mp_i, 1\}$, p_i is the p-value of the coefficient, M is the number of hypotheses tested. The critical value is $\frac{\alpha_i}{M}$. If p_i^B is smaller than the critical value, we reject the null (successful discovery of a new factor).

Adjustment 2(Holm): Sort the p-value of different factors $p_{(1)} \leq ... \leq p_{(M)}$, the corresponding hypothesis $H_{(1)}, ..., H_{(M)}$. We find the minimum k such that $p_{(k)} \geq \frac{\alpha_w}{M+1-k}$, where α_w is the pre-specified significance level. We reject the null for $H_{(1)}, ..., H_{(k-1)}$, the accept the rest.

Adjustment 3(BHY): Do the same sort as Holm, and find the maximum k such that $p_{(k)} \leq \frac{k}{Mc(M)}\alpha_d$, where α_d is the prespecified significance level, and c(M) is a function of the number of tests. Usually we take $c(M) = \sum_{j=1}^{M} \frac{1}{j}$.

Relationship between the three adjustments

Holm is less (weakly) stringent than Bonferroni, which means Holm adjustment can have at least as many discoveries as Bonferroni adjustment.

Holm controls FWER regardless of the dependence structure of p values. So do BHY.

3 Discussion

Hidden insignificant tests

Truncated model to deal with hidden test issue. Idea: Assume all factor with t stat larger than 2.57 are observable. Fit the histogram of the observables with a exponential distribution. The fraction of missing observations can be inferred.

71% of the factors are estimated to be missing. In this case, the t ratio criterion is slightly higher than the case of full observations, 4.01 and 3.96 respectively under Bonferroni and Holm.

Correlated hypotheses

Holm and BHY are independent of the structure, but they do not utilize the information. Can we do better if we use the dependence structure of the p-values?

We can impose some contemporaneous correlation structure of the long-short strategy returns of different factors, simulate the returns and t statistics to match the data quantiles. From the simulation exercise we can know the critical t statistic with a certain dependence structure.

Asset pricing factors revisit

With the multiple testing framework and a resulting t criterion of 3.0, almost half of the factors are spurious discoveries. The number of factors is still far more than what we think is true.