# The Realized Beta

## 1 Basic Setting

Consider two random variables $X_1$ and $X_2$ with zero mean. The theoretical linear regression of $X_2$ on $X_1$ is:

$$X_2 = \beta X_1 + \varepsilon$$

where:

$$\beta = \frac{\text{Cov}\left(X_1, X_2\right)}{\text{Var}\left(X_1\right)}$$

Given observations of the random variables:

$$\{X_{1,i}, X_{2,i}\}_{i=1}^{n}$$

we can estimate $\beta$ with the estimator:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_{1,i} X_{2,i}}{\sum_{i=1}^{n} X_{1,i}^2}$$

Under some regularity conditions we know that $\hat{\beta}$ is a consistent estimator of $\beta$ with a well defined asymptotic distribution.

The theoretical regression is a statistical construction that essentially always exists so long as the moments exist, but it does not always make economic sense. If we just grab any two variables, then we likely get a nonsense regression of no economic interest. If we use economics to select the variables, then the regression can be a reasonable interpretation. In our case, $X_2$ is the return on a particular asset or portfolio and $X_1$ is the return on the market, and $\beta$ can be interpreted as the beta of the Capital Asset Pricing Model.

Next we will see how to use this same idea in the high-frequency setting.

## 2 High-Frequency Setting

Consider two processes $X_1$ and $X_2$. For example, $X_1$ is the log-price of a market index (like the S&P500 index) and $X_2$ is the log-price of some portfolio or stock. Define the vector:

$$X \equiv \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

Just as in the previous lectures, we assume that $X_t$ follows a semimartingale (the usual model for financial prices). The difference is that $X$ is a 2 dimensional process. The dynamics of $X$ are given by:

$$dX_t = \sqrt{c_t}dW_t + dJ_t \text{ for } 0 \leq t \leq T$$

where $W_t$ is a 2-dimensional Wiener process and $J_t$ is a compound Poisson process.

The variance process $c_t$ is now a matrix:

$$c_t = \begin{bmatrix} c_{11,t} & c_{12,t} \\ c_{21,t} & c_{22,t} \end{bmatrix}$$

The variance matrix is symmetric and we denote the symmetric square root of $c_t$ by $\sqrt{c_t}$. The matrix $c_t$ is referred to as the local variance matrix.

We can decompose the process $X$ into continuous and discontinuous parts:

$$X_t = X_t^c + X_t^d$$

Define the continuous returns as:

$$r_{1,t,i}^c \equiv \Delta_{i+(t-1)n}^n X_1^c$$
$$r_{2,t,i}^c \equiv \Delta_{i+(t-1)n}^n X_2^c$$

From our earlier work, we know how to use the continuous returns to estimate the integrated variance of the processes:

$$TV_{1,t} \equiv \sum_{i=1}^n (r_{1,t,i}^c)^2 \to \int_{t-1}^t c_{11,s} ds$$
$$TV_{2,t} \equiv \sum_{i=1}^n (r_{2,t,i}^c)^2 \to \int_{t-1}^t c_{22,s} ds$$

The continuous returns can also be used to estimate the integrated covariance:

$$\widehat{RCov}_t \equiv \sum_{i=1}^n r_{1,t,i}^c r_{2,t,i}^c \to \int_{t-1}^t c_{12,s} ds =: RCov_t$$

The regression $\beta$ in the first part was given by the covariance between the two processes divided by the variance of the first. This motivates the definition of an analogous $\beta$ in the high-frequency setting: the realized beta.

$$R\beta_t \equiv \frac{\int_{t-1}^t c_{12,s} ds}{\int_{t-1}^t c_{11,s} ds}$$

The definition also suggests the estimator for $R\beta$:

$$\widehat{R\beta}_t \equiv \frac{\widehat{RCov}_t}{TV_{1,t}} = \frac{\sum_{i=1}^n r_{1,t,i}^c r_{2,t,i}^c}{\sum_{i=1}^n (r_{1,t,i}^c)^2} \to R\beta_t$$

# 3 Inference for the Realized Beta

To construct confidence intervals for the realized beta one can study the asymptotic distribution of the realized beta estimator. Results on that front can be found in Ait-Sahalia and Jacod (2014). However, we will rely on a bootstrap approximation to conduct inference for $R\beta$.

The bootstrap scheme is analogous to the bootstrap for the truncated variance. First, for each day divide the observations into $M$ intervals of length $k_n$ (this implies that $M = \lfloor \frac{n}{k_n} \rfloor$). Second, for each interval $j$ (for $j = 1, 2, \ldots, M$), draw $k_n$ independent continuous returns $(\tilde{r}_1^c, \tilde{r}_2^c)$ by sampling randomly with replacement from the continuous returns in the j-th interval. Third, use the new sample to compute the realized beta estimator:

$$\widetilde{R\beta}_t = \frac{\sum_{i=1}^n \tilde{r}_{1,t,i}^c \tilde{r}_{2,t,i}^c}{\sum_{i=1}^n \tilde{r}_{1,t,i}^2}$$

Fourth, repeat the three steps above to obtain several different estimates of the realized beta. Denote those estimates for day $t$ by:

$$\left\{\widetilde{R\beta}_{t,1}, \widetilde{R\beta}_{t,2}, \ldots, \widetilde{R\beta}_{t,10000}\right\}$$

Compute the 2.5% and 97.5% quantiles of the set above to form a 95% confidence interval for the realized beta for day $t$. Repeat the above for of the days $t = 1, 2, \ldots, T$.

When implementing the bootstrap and debugging your code, set the bootstrap repetitions to 100 per day. Then debug the code and only when you are sure it is working increase the number of bootstrap repetitions to a high number (like 10000).