Linear Regression in Finance

This is a simple linear regression equation:

$$\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{x}$$

(1)

We take into account for the error term.

$$y_i = \alpha + \beta x_i + \hat{\epsilon}_i$$

(2)

Rearrange it in terms of the error term and we have this.

$$\hat{\epsilon}_i = y_i - \alpha - \beta x_i$$

Our goal is to effectively minimise the sum of the errors. And to do this we calculate the parameters based on this equation:

(3)

This is the Slope Parameter.

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(4)

This is the intercept.

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

(5)

This is the Slope Parameter in vector form.

$$\hat{\beta} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \cdot \\ \cdot \\ \cdot \\ x_{n-1} - \bar{x} \\ x_n - \bar{x} \end{pmatrix} \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \cdot \\ \cdot \\ \cdot \\ y_{n-1} - \bar{y} \\ y_n - \bar{y} \end{pmatrix}$$

Let

$$C = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(8)

$$\hat{\beta} = C \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ . \\ . \\ . \\ x_{n-1} - \bar{x} \\ x_n - \bar{x} \end{pmatrix} \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ . \\ . \\ . \\ y_{n-1} - \bar{y} \\ y_n - \bar{y} \end{pmatrix}$$

Let:

$$\hat{X} = \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ . \\ . \\ . \\ x_{n-1} - \bar{x} \\ x_n - \bar{x} \end{pmatrix}$$

Now, we know that

$$\hat{X} = \begin{pmatrix} 1 - \bar{x} \\ 2 - \bar{x} \\ . \\ . \\ . \\ n - 1 - \bar{x} \\ n - \bar{x} \end{pmatrix}$$

$$\bar{x} = \frac{(1 + 2 + 3 + 4 + 5... + n - 1 + n)}{n}$$
$$\bar{x} = \frac{n(n+1)}{2}\frac{1}{n}$$
$$\bar{x} = \frac{(n+1)}{2}$$

So:

$$\hat{X} = \begin{pmatrix} 1 - \frac{(n+1)}{2} \\ 2 - \frac{(n+1)}{2} \\ . \\ . \\ . \\ n - 1 - \frac{(n+1)}{2} \\ n - \frac{(n+1)}{2} \end{pmatrix}$$

Note that:

$$1 - \frac{n+1}{2} = -(n - \frac{n+1}{2})$$

Proof: For the LHS

$$\frac{2-n-1}{2} = \frac{1-n}{2}$$

For the RHS

$$\begin{gathered} -(\frac{2n-n-1}{2}) = \frac{-2n+n+1}{2} \\ \frac{-2n+n+1}{2} = \frac{1-n}{2} \end{gathered}$$

Therefore $LHS = RHS$

Similarly in the general case:

$$j - \frac{n+1}{2} = -((n-j+1) - \frac{n+1}{2})$$

$$\begin{gathered} \frac{2j-n-1}{2} = -(\frac{2n-2j+2-n-1}{2}) \\ \frac{2j-n-1}{2} = -(\frac{n-2j+1}{2}) \\ \frac{2j-n-1}{2} = (\frac{-n+2j-1}{2}) \\ \frac{2j-n-1}{2} = \frac{2j-n-1}{2} \end{gathered}$$

And thus we can rearrange $\hat{X}$

$$\hat{X} = \begin{pmatrix} -(n - \frac{(n+1)}{2}) \\ -((n-1) - \frac{(n+1)}{2}) \\ . \\ . \\ . \\ ((n-1) - \frac{(n+1)}{2}) \\ (n - \frac{(n+1)}{2}) \end{pmatrix}$$

So now we have:

(9)

$$\hat{\beta} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} -(n - \frac{(n+1)}{2}) \\ -((n-1) - \frac{(n+1)}{2}) \\ . \\ . \\ . \\ ((n-1) - \frac{(n+1)}{2}) \\ (n - \frac{(n+1)}{2}) \end{pmatrix} \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ . \\ . \\ . \\ y_{n-1} - \bar{y} \\ y_n - \bar{y} \end{pmatrix}$$

Now if we expand the above

$$\hat{\beta} = C(-(n - \frac{(n+1)}{2})(y_1 - \bar{y}) + ... + (n - \frac{(n+1)}{2})(y_n - \bar{y}))$$

The $\bar{y}$ terms end up cancelling out. We can factorise the like terms like this:

(10)

$$\hat{\beta} = C((n - \frac{(n+1)}{2})(y_n - y_1) + ((n-1) - \frac{(n+1)}{2})(y_{n-1} - y_2) + ... + ((\frac{n}{2}+1) - \frac{(n+1)}{2})(y_{\frac{n}{2}+1} - y_{\frac{n}{2}}))$$

We can then further rearrange this into vector form:

(11)

$$\hat{\beta} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} n - \frac{(n+1)}{2} \\ (n-1) - \frac{(n+1)}{2} \\ . \\ . \\ . \\ (\frac{n}{2}+1) - \frac{(n+1)}{2} \end{pmatrix} \begin{pmatrix} y_n - y_1 \\ y_{n-1} - y_2 \\ . \\ . \\ . \\ y_{\frac{n}{2}+1} - y_{\frac{n}{2}} \end{pmatrix}$$

We have reduced the number of terms by HALF and we also notice that the vector on the right hand side is in the form of returns. If we log the prices, then these could be interpreted as log returns. So, essentially the linear regression of Time and Prices is the sum of weighted returns OR the sum of weighted log returns.

Interestingly, the log returns of the edge prices are weighted more, and the weight of each inner returns are weighted less and less.

(12)

$$\hat{\beta} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} n - \frac{(n+1)}{2} \\ (n-1) - \frac{(n+1)}{2} \\ . \\ . \\ . \\ (\frac{n}{2}+1) - \frac{(n+1)}{2} \end{pmatrix} \begin{pmatrix} log_e(\frac{y_n}{y_1}) \\ log_e(\frac{y_{n-1}}{y_2}) \\ . \\ . \\ . \\ log_e(\frac{y_{\frac{n}{2}+1}}{y_{\frac{n}{2}}}) \end{pmatrix}$$

Thus, the slope parameter of the linear regression can also have many combinations of differing log returns

For example.

$$Time = [1, 2, 3, 4]$$
$$Price = [y_1, y_2, y_3, y_4]$$

$$C \times \hat{X} = \begin{pmatrix} 0.3 \\ 0.1 \end{pmatrix}$$

We also have our Log differences, let this be:

$$\hat{L} = \begin{pmatrix} log_e(\frac{y_4}{y_1}) \\ log_e(\frac{y_3}{y_2}) \end{pmatrix}$$

And so on..

$$\hat{\beta} = C \times \hat{X} \cdot \hat{L}$$

Using our example we now have:

$$\hat{\beta} = 0.3(log_e(y_4) - log_e(y_1)) + 0.1(log_e(y_3) - log_e(y_2))$$

Given a fixed/constant $\hat{(\beta)}$ parameter, what sort of combination of prices can be revealed?

We know that $y_i > 0$ and if $\hat{\beta} > 0$ then:

$$\hat{\beta} = 0.3log_e(y_4) - 0.3log_e(y_1) + 0.1log_e(y_3) - 0.1log_e(y_2)$$

As $\hat{\beta} > 0$, then:

$$0.3log_e(y_4) - 0.3log_e(y_1) + 0.1log_e(y_3) - 0.1log_e(y_2) > 0$$

$$3(log_e(y_4) - log_e(y_1)) > (-log_e(y_3) + log_e(y_2))$$

$$3(log_e(y_4) - log_e(y_1)) > (log_e(y_2) - log_e(y_3))$$

The difference between the 4th and 1st price must be greater than the difference between the 2nd and 3rd price.

If All of (1) is greater than 0, then this means that:

$$log_e(y_4) - log_e(y_1) > 0$$

It also means that:

$$(log_e(y_2) - log_e(y_3))$$

Can be positive or negative, but that depends on the magnitude of difference between the 4th and 1st price and the actual slope value.

Thus, looking at the slope parameter may not be sufficient enough in understanding the movements of prices. However, despite having the same slope you can have different combinations of prices with different variances. However, we can account for these differences by looking at the intercept value as:

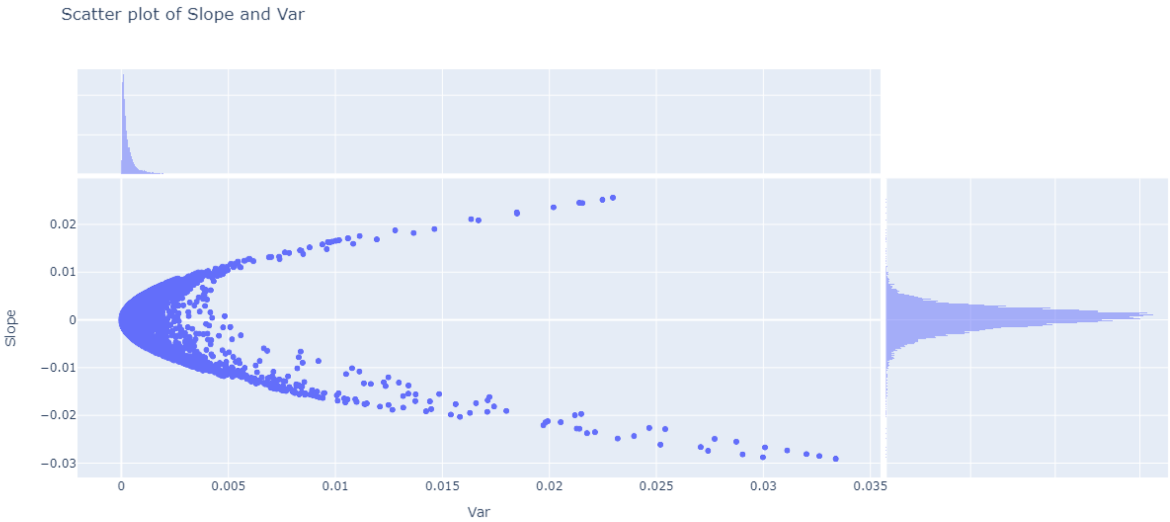
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

So, what would the movement of the intercept look like if the slope parameter stayed the same but the price parameters did not?

Assume we were looking at a random dataset. We recognize that a specific slope value occurs many times. What possible values can we encounter?

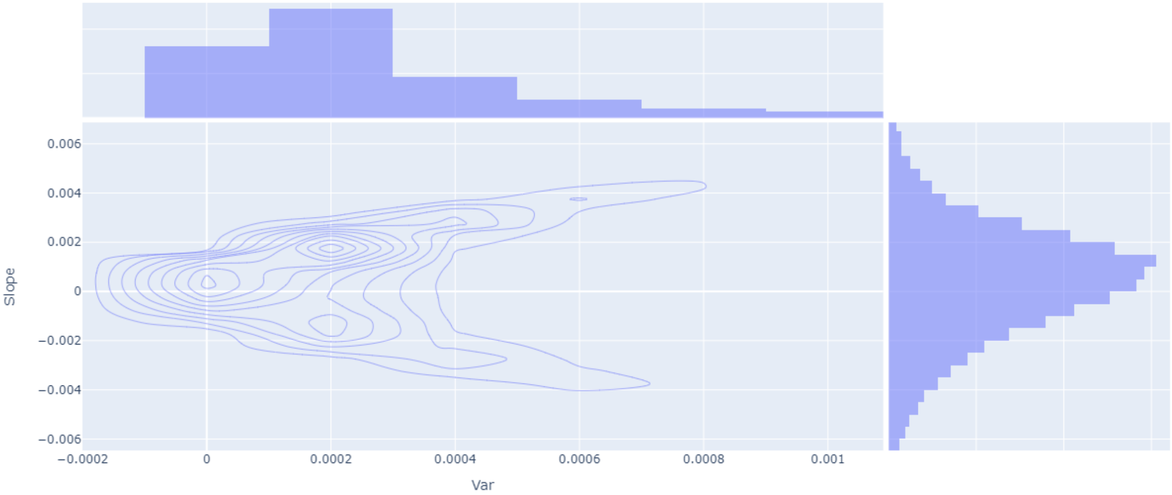
However, in regards to looking at the intercept it is clear that it would be difficult to look at it objectively because the intercept values can change drastically as prices start to change too. Instead of looking at the intercept we can look at the variance of prices along with the slope itself. This will be supplemented with an analysis of how variance can change over time, slopes can change over time and the relationship between the change in variance and the change in slopes will be discussed some where else.

This is the Slope and Variances plotted:



In the above image we look at a scatter plot of the slope and variances, it is hard to tell if whether the proximity of the plots is corresponded with the density of the plots, so we later instead look at the density contours.

This is the Density contour of the Slope and Variances:



Slope Values near the mean are more likely to have differing combinations of prices, there is a greater density of different price variances BUT the price variance themselves though are of a smaller magnitude. Meanwhile slope values further away from the mean are less likely to have differing combinations of prices, there is less density of different price variances for that slope BUT the price variance themselves are a lot larger in magnitude.

There should also be a comparison of looking at these slope values and comparing them to different return measurements, percentage change and log differences.

The code in LinearRegressionAnalysis.py produces a histogram, looking at various features such as mean, variance, kurtosis and skewness.

There is also an analysis of Residuals in the file, however no meaningful features were produced.