Lectures notes: Linear Regressions in Finance

Mathis MOUREY

Lecturer in Finance, The Hague University of Applied Sciences (THUAS)

November 15, 2021

[1]

Contents

[1] *Disclaimer: These lectures notes are from a lecture on financial application for linear regression made at Grenoble IAE (Oct. 2021). The document is not in his final version and might have some errors. Please do not cite.*

1 What is a linear regression?

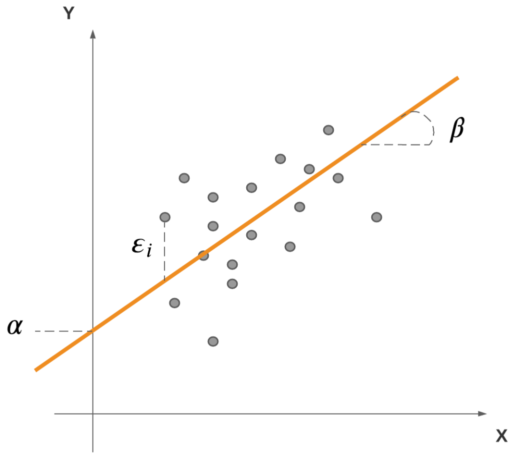
A linear regression is a tool to find a relationship between a variable y (the dependent variable) and multiple other variables $\{x_1, x_2, x_3, ..., x_n\}$, called the explanatory variables. In its simplest form, there is only one explanatory variable. Formally,

$$y = \alpha + \beta x + \varepsilon \tag{1}$$

In such case, the interpretation is quite straightforward. The β coefficient gives the nature of the relationship. For instance, if $\beta = 1$, y and x move in similar manner. However, if $\beta = -1$, they would move in a perfectly opposite fashion. Hence, the β gives you the direction and the extent of the relationship. Figure 1 gives a visual representation of an univariate linear regression.

Figure 1: **Linear Regression.**

Example of a univariate linear regression. The equation of the regression is $y = \alpha + \beta x + \varepsilon$. The intercept corresponds to the value of α. β gives the slope of the regression line. ε is a vector of all errors (the distance between a data point and the regression line), mathematically: $\varepsilon = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]$



A fair amount of cases will be limited to a univariate linear model (only one explanatory variable). For instance, the infamous CAPM (Markowitz, 1954) could identified as a univariate linear regression between the excess financial returns of a firm $(R - R_f)$ and the excess return of the market $(R_m - R_f)$. In this case, we have $y = R - R_f$ and $x = R_m - R_f$. We write the regression as:

$$R - R_f = \alpha + \beta[R_m - R_f] + \varepsilon \tag{2}$$

The CAPM would be a specific case where $\alpha = 0$ and where there is no errors (ε). By looking at the definition of a linear regression, we can say that the CAPM tries to explain the excess return of a firm by the excess return of the market. If the β is strongly positive (superior to 1), it means that the returns of the firm are varying in the same direction as the market but with larger changes. In the other case where $\beta \leq -1$ the firm's returns have going in this opposite direction and with larger changes as well.

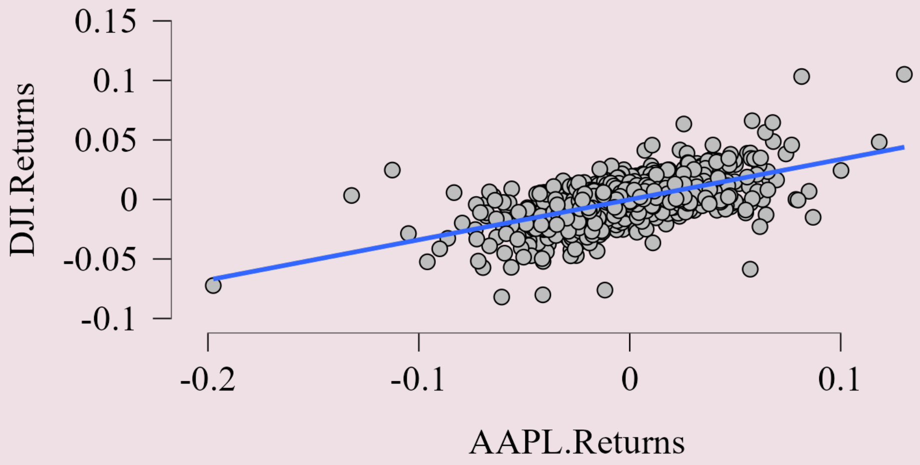
Case Study 1: Apple & Dow Jones Industrial Average Returns

Let us consider the case of Apple stock returns against the Dow Jones Industrial returns (See Figure 2). In order to verify the CAPM, we run the following regression (and for the sake of simplicity, we assume a $R_f = 0$),

$$R_{AAPL} = \alpha + \beta R_{DJI} + \varepsilon$$

Graphically, the results are as follows,

Figure 2: **Linear Regression - Apple's returns and Dow Jones returns.** *Linear regression between the logarithmic returns of Apple (AAPL.Returns) and the Dow Jones Industrial Average (DJI.Returns). The sample consists of daily returns from 2007 to 2019. The blue line shows the linear relationship between the two variables.* $R_{AAPL} = \alpha + \beta R_{DJI} + \varepsilon$



As Figure 2 shows there is a positive relationship between the stock returns of Apple and the ones from the Dow Jones Industrial index. The exact information of that particular regression are given in the appendix (see Table 1).

In order to know whether your regression model is good, you need to look at the *explanatory power* of your model. Explanatory power means: how much does your explanatory variables explain your dependent variable? In a regression, the explanatory power is given by the R^2. Its value ranges from 0 to 1. The higher it is, the better is your model. A simple interpretation could be, if $R^2 = 1$ then you model explains 100% of the variance of your dependent variable. In other terms, you model perfectly explains your dependent variable. In 'hard' sciences as maths or physics, where the stakes are usually high (for instance astrophysics) in order for a model to be valid, the R^2 should be extremely close to 1. In social sciences however, and so in finance, it is almost impossible to get such models, so researchers are usually happy with R^2 approximating 30% or 40%.

You also need to be able to know how much you can trust the value of the parameters given by the regression (βs). For this, you can pay attention to the p-value. Simply put, the lower the p-value the better. We usually consider that a p-value below 5% is significant. Strong significance is usually evidenced by a p-value lower than 0.1%. For instance, if you refer to the Table 1, you can see that the β is strongly significant and with value 1. Indicating that Apple's returns varies almost perfectly as the Dow Jones Index returns.

However, thinking that financial returns are driven by a single factor (single explanatory variable) is a strong, and probably wrong, assumption. It is more likely that more factors (explanatory variables) will be able to explain better financial returns. We would then need a multivariate regression.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon \tag{3}$$

Or equivalently,

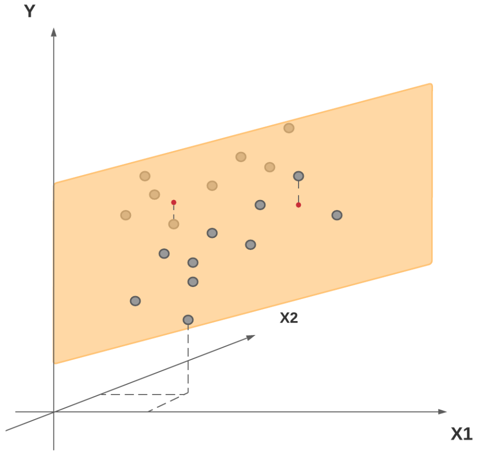
$$y = \alpha + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \tag{4}$$

Figure 3 gives a graphically representation of a multivariate (bivariate) regression.

Multivariate regressions allow for a more complete analysis of what factors are explaining in the dependent variable. There is, however, a number of possible problems you need to be aware of.

Figure 3: **Multivariate Linear Regression.**

Example of a multivariate linear regression. The equation of the regression is $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$*. Since we are looking at two dimensions (two explanatory variables), we obtain a response surface (orange surface) that corresponds to the regression line in the univariate case. The errors are the distances that separate a data point from the surface.*



Some problems with multivariate regressions

- *An illusion of increased explanatory power*

  Because the model becomes more complex, the explanatory power (R^2) will automatically increase. In other words, if would like to study the relationship between the number of birds and the number of trees in multiple areas ($BIRDS = \alpha + \beta TREES + \varepsilon$), my model will be better (or at least as good) regardless of what other explanatory variable I would add to it[2]. In order to deal with that, we have an 'adjusted R^2'. The value of this adjusted R^2 is penalized according to the amount of explanatory variables present in the model. The more you will use, the less explanatory power you will have. It forces you to use a small amount of explanatory variables that explains the best your dependent variable.

- *Multicolinearity: What if explanatory variables are correlated?*

  You will often hear/read this complex idea of 'multicolinearity'. It is however a quite simple idea. What would happen if I would put twice the same explanatory

[2] For instance: $BIRDS = \alpha + \beta_1 TREES + \beta_2 LAKES + \varepsilon$

variable in my model? Let us say that $x_1 = x_2$ in the following models:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{5}$$

Which of β_1 or β_2 would show the real relationship between x_1 (or x_2) and y? Additionally, the presence of two identical explanatory variable cannot improve the explanatory power[3] of my model. Multicolinearity makes the value of your βs and your R^2 wrong. Remember to check if your explanatory variables are correlated before running your regression!

- *Endogeneity or the art of constructing a variable from another*

  Endogeneity is a (barbaric) statistic term used when the explanatory variable (x) is made from another variable in the model (y or other x). For instance, let us consider a regression between assets of firms ($ASSET$) against the net earnings (NE) of firms and their earnings before interest and taxes($EBIT$) such that,

$$ASSETS = \alpha + \beta_1 NE + \beta_2 EBIT + \varepsilon \tag{6}$$

  We know that Net Earnings are directly stemming from EBIT, hence they are almost the same variable. We then fall right back into the case of multicolinearity described above. Each factor (explanatory variable) in your regression model must independent from each other!

In finance, we tend to work with three types of data, and depending on the type of data involved, we will call regressions differently.

2 Cross-sectional regressions: size and profitability of banks

The first type of data is called **cross-sectional**. Cross-sectional data is data associated to a single point in time. For instance, consider the market capitalization of the 60

[3] It cannot because, if it could, I could just keep adding the same explanatory variable until my model is perfect! Fortunately, that is not possible.

largest U.S. banks the 1st of January 2020. You have a vector consisting in the market capitalization of each bank at that precise point in time, such as

$$MC = [MC_1, MC_2, MC_3, ..., MC_{60}] \tag{7}$$

Where MC_1 is the market capitalization of Bank 1 on the 01/01/2020. When a regression is made using cross-sectional data, we call it a cross-sectional regression. For instance, if we would regress these market capitalization with the leverage (LEV) of the same 60 largest U.S. banks, we would get the following cross-sectional (univariate) regression,

$$MC = \alpha + \beta LEV + \varepsilon \tag{8}$$

Usually, cross-sectional regression are used in order to find out relationship between indicators at a certain point in time. Do bigger financial firms have a higher profitability? Do smaller firms tend to be less liquid? These kind of questions relates to finding correspondence across a sample.

Case Study 2: Banks Size and Bank Profitability.

Let us consider the following question: Do larger banks are more profitable than smaller banks? Another way of putting the problem would be to ask whether the is a link between the size of banks and their profitability. First of all, we would need to retrieve data on the size and profitability of banks, such that be obtain two vectors: One for size ($ASSET$), another for profitability ($EBIT$), such as,

$$ASSET = [A_1, A_2, A_3, ..., A_n]$$

$$EBIT = [E_1, E_2, E_3, ..., E_n]$$

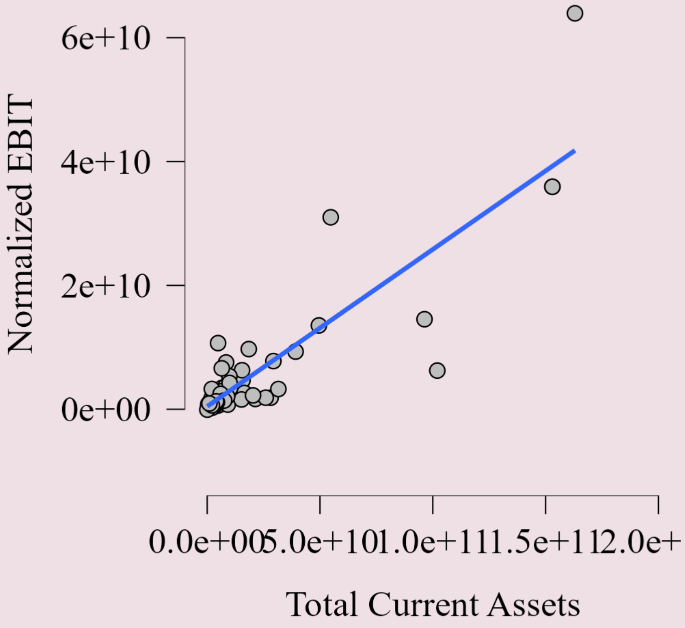
Where A_1 is the total asset value of Bank 1, and E_1 is the Earning Before Interest and Taxes (EBIT) of Bank 1. In order to answer our research question, we would then need to evaluate the following cross-sectional regression,

$$ASSET = \alpha + \beta EBIT + \varepsilon \tag{9}$$

Retrieving empirical data on 82 of the largest U.S. banks, we get the following (graphical) results,

Figure 4: **Linear Regression - Apple's returns and Dow Jones returns.**

Linear regression between the bank total asset (ASSET) and their normalized EBIT (EBIT) . The sample consists of cross-sectional data of 82 of the larest U.S. banks. The blue line shows the linear relationship between the two variables. $ASSET = \alpha + \beta EBIT + \varepsilon$



Of course cross-sectional do not have to be univariate. Most of the time, they are not in fact. We usually want to see how multiple variables are influencing the dependent variable. Let say that you want to know the determinants of profitability of a firm. In finance, profit is intrinsically linked with risk. Hence, we could consider risks indicators and see if they are positively related to profitability (higher risk, higher reward). We could then consider the following model,

$$EBIT = \alpha + \beta_1 ILLIQ + \beta_2 VaR + \beta_3 ASSET + \beta_4 LEV + \varepsilon \qquad (10)$$

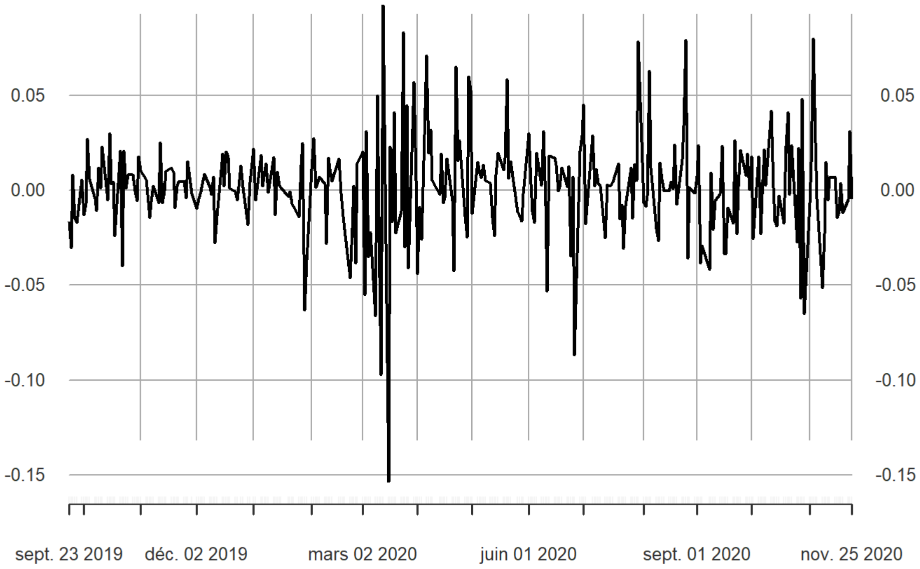
Where $ILLIQ$ is a (il)liquidity indicator, VaR measures loss risk, $ASSET$ is for size and LEV represents the leverage of the firm. We would then be interested in the values of $\{\beta_1, \beta_2, \beta_3, \beta_4\}$ and their significance. The most likely result being that they are all positively linked to profitability.

3 Timeseries regressions: CAPM & Fama-French Factor models

The second, and arguably most prominent, type of financial data are **timeseries**. Oppositely to cross-sectional, timeseries are data related to a single firm over multiple dates. For instance, we could consider the daily financial returns of Facebook (FB) from the 01/01/2019 to 01/01/2021.

Figure 5: **Financial Returns of Facebook**

Daily returns for Facebook (FB) from the 01/01/2019 to 01/01/2021.



Most of the studies in portfolio theory are going to use timeseries data. Event studies are also a specific case that uses timeseries. In particular for computing abnormal returns. We know that abnormal returns are the difference between actual returns and returns predicted by a given model. The simplest case is based on the CAPM, and corresponds to,

$$AR = R - E[R] \tag{11}$$

Where, AR are the abnormal returns, R the actual returns and $E[R]$ returns predicted by the model chosen. Given the model is the CAPM (and consider $R_f = 0$), we will have,

$$E[R] = \alpha + \beta R_m \tag{12}$$

Where R_m are the market's returns. You would need to run a (timeseries) regression in order to estimate α and β. And then simply compute,

$$AR = R - (\alpha + \beta R_m) \tag{13}$$

In order to have a visual representation of the CAPM, see Case Study 1. The CAPM is a simple case and it has been shown that it is not consistent with empirical observation of market data. However, there exists a lot of various models for stock returns. Ones of the most documented of the Fame-French factors models (3 and 5 factors). Their reasoning is that stock returns are not only driven by a risk premium, but by various other effects. The size of the firm will have an impact of its returns, as well as its profitability. The model is as follows,

$$R = \alpha + \beta_1[R_m - R_f] + \beta_2 SMB + \beta_3 HML \tag{14}$$

Where SMB is the size factor and HML is the profitability factor[4].

Case Study 3: Fama-French 3 Factors (FF3 Factors)

Let us use the Fama-French 3 Factors model and compare it to the usual CAPM. Whichever model produces a higher R^2 will be the most effective at explaining stock returns. To be more precise, we will look at the adjusted R^2. Because Fama-French is the same than CAPM but with two additional explanatory variable we are sure to get a better R^2 (or at least as good). In order to make sure than the additional factors are improving the model, we need to look at the adjusted R^2.

Let us consider the case of BNP Paribas between 12/10/2020 and 31/08/2021 (the period is chosen randomly). We retrieve the data from Yahoo finance, and compute the logarithmic return on the Adjusted Price. In order to get the factors (or explanatory variable): $R_m - R_f$, SMB and HML, we go the website of Fama-French and retrieve the factors for the European market. Our regressions are:

CAPM: $R_{BNP} = \alpha + \beta[R_m - R_f] + \varepsilon$

FAMA-FRENCH 3F: $R_{BNP} = \alpha + \beta_1[R_m - R_f] + \beta_2 SMB + \beta_3 HML + \varepsilon$

[4] The factors can be found on the website of Fama & French. They are usually up to date and free to download! Link: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

The results are given in Table 4 and Table 3. By running the CAPM, we find a adjusted R^2 of 0.406 which is an impressive explanatory power already. However, the Fama-French model display an explanatory power of 74.4%. We can see, without doubt that the Fama-French can explain better the results of BNP. Figure 6 and Figure 7 give a nice visual representation of the results.

Figure 6: **CAPM for BNP**

Fit between actual returns of BNP and the predicted returns using the CAPM. The middle line corresponds to a perfect fit (if the model would be perfect, all dots would be lying on the central line).

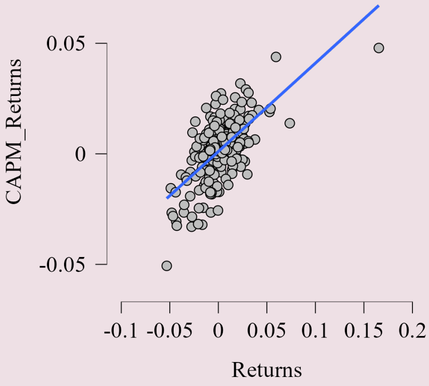
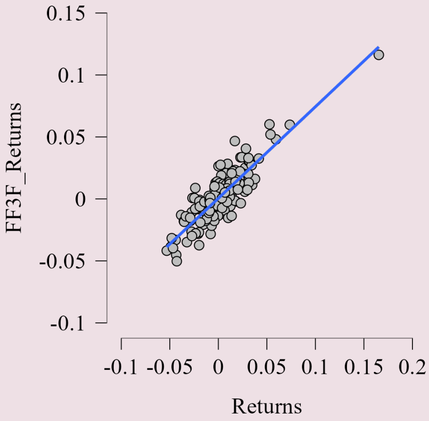


Figure 7: **Fama-French 3 Factor for BNP**

Fit between actual returns of BNP and the predicted returns using the Fama-French 3 Factor. The middle line corresponds to a perfect fit (if the model would be perfect, all dots would be lying on the central line).



It turns out that returns are not only driven by the levels of risk but also by other effects like the size and profitability effects.

4 Intuitive approach to panel data: the Fama-McBeth approach

When you end up having both types of data at the same time (cross-sectional and timeseries), we call your dataset a panel. For example, having the market capitalization of the largest U.S. banks for multiple years is a panel. You have the cross-sectional component: each year you have one value for each bank (there is N banks). And you also have a timeseries dimension: you have data for multiple years (there is T years). Visually,

$$\begin{bmatrix} MC_{T;1} & MC_{T;2} & MC_{T;3} & ... & MC_{T;N} \\ \vdots & \vdots & \vdots & ... & \vdots \\ MC_{2;1} & MC_{2;2} & MC_{2;3} & ... & MC_{2;N} \\ MC_{1;1} & MC_{1;2} & MC_{1;3} & ... & MC_{1;N} \end{bmatrix} \tag{15}$$

Given you want to assess the relationship between the market capitalization of banks and their profitability over multiple years, you would get a difficult equation like,

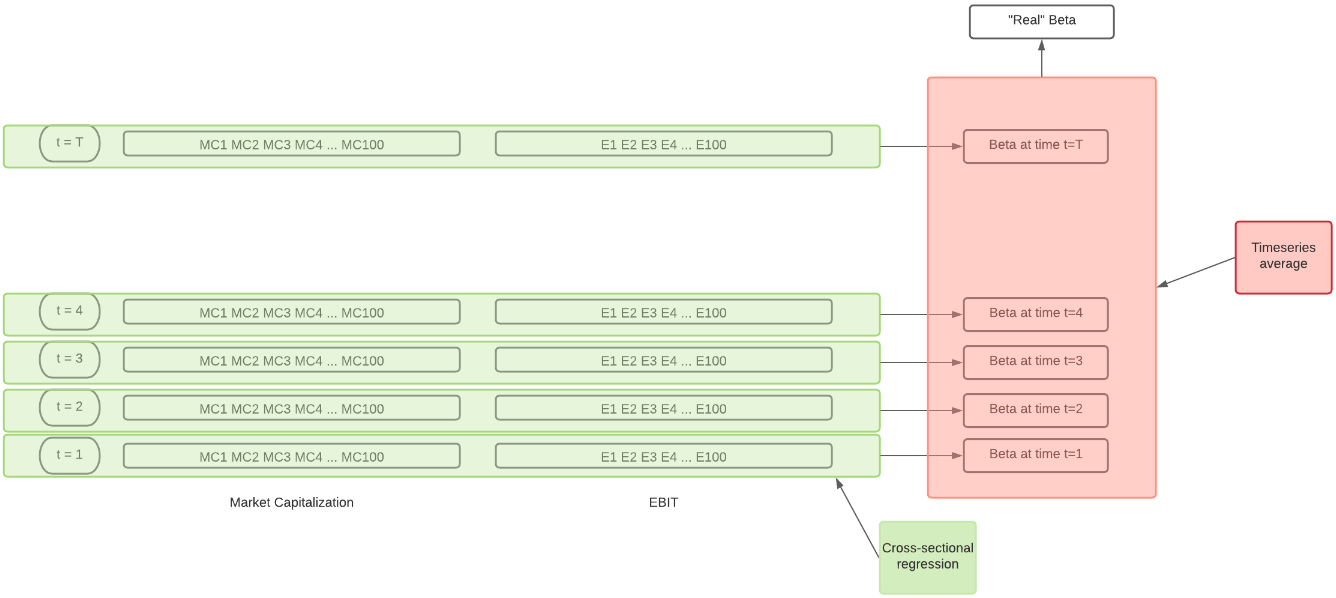
$$\begin{bmatrix} MC_{T;1} & MC_{T;2} & MC_{T;3} & ... & MC_{T;N} \\ \vdots & \vdots & \vdots & ... & \vdots \\ MC_{2;1} & MC_{2;2} & MC_{2;3} & ... & MC_{2;N} \\ MC_{1;1} & MC_{1;2} & MC_{1;3} & ... & MC_{1;N} \end{bmatrix} = \begin{bmatrix} \alpha_T \\ \vdots \\ \alpha_2 \\ \alpha_1 \end{bmatrix} + \begin{bmatrix} \beta_T \\ \vdots \\ \beta_2 \\ \beta_1 \end{bmatrix} * \begin{bmatrix} E_{T;1} & E_{T;2} & E_{T;3} & ... & E_{T;N} \\ \vdots & \vdots & \vdots & ... & \vdots \\ E_{2;1} & E_{2;2} & E_{2;3} & ... & E_{2;N} \\ E_{1;1} & E_{1;2} & E_{1;3} & ... & E_{1;N} \end{bmatrix} + \begin{bmatrix} \varepsilon_T \\ \vdots \\ \varepsilon_2 \\ \varepsilon_1 \end{bmatrix} \tag{16}$$

Where $E_{i,j}$ is the EBIT of bank j at time i. One can easily see that it will be difficult to estimate such a model. Fortunately, two academics came up with a method that only involve cross-sectional regressions. The method is schematically described below (see Figure 8).

The procedure consists in two steps. First, run as many cross-sectional regressions as you have data point (one for each point in time). Such that to estimate β_1, you will run

Figure 8: **Schematic view of a Fama-MacBeth procedure**

Fama-MacBeth is a way to deal with panel data without using panel models. You first use cross-sectional regressions for each point in time (green boxes) in order to get the β coefficents at that time. And then you average your cross-sectional β (red box) in order to get the final value for your β.



the following:

$$MC_{t=1} = \alpha + \beta_1 E_{t=1} + \varepsilon \tag{17}$$

By repeating that for every point in time in your sample you will get a vector of βs, such as

$$[\beta_1, \beta_2, ..., \beta_T] \tag{18}$$

The second and final step is then to average all βs in order to get your 'real' β showing the average relationship between market capitalization and profitability for banks.

$$\beta = \frac{\sum\limits_{i=1}^{T} \beta_i}{T} \tag{19}$$

An issue with Fama-MacBeth is that it ignores timeseries variation. In our example, we want to have a general idea on the relationship between market capitalization and profitability (EBIT) for banks in general. Fama-MacBeth tells us what is the average relationship over the time period selected in our sample. It however overlooks the possibility that the relationship changes over time (which, in some cases, might be a possibility). When applying the procedure you should then be aware of that limitation.

Appendix

Tables

Table 1: **Linear Regression Summary - AAPL - DJI**

Summary statistics for: $R_{AAPL} = \alpha + \beta R_{DJI} + \varepsilon$. *Explanatory power of the regression is presented. Values and significance of the coefficients are included.*

Coefficients	Unstandardized	Standard Error	Standardized	t	p
α (Intercept)	7.458e-4	2.803e-4		2.661	0.008
β (R_{DJI})	1.018	0.025	0.587	41.324	$< .001$

R	R^2	Adjusted R^2	RMSE
0.587	0.344	0.344	0.016

Table 2: **Linear Regression Summary - ASSET - EBIT**

Summary statistics for: $ASSET = \alpha + \beta EBIT + \varepsilon$. *Explanatory power of the regression is presented. Values and significance of the coefficients are included.*

Coefficients	Unstandardized	Standard Error	Standardized	t	p
α (Intercept)	2.157e+9	1.909e+9		1.130	0.262
β (EBIT)	3.056	0.184	0.880	16.598	$< .001$

R	R^2	Adjusted R^2	RMSE
0.880	0.775	0.772	1.560e+10

Table 3: **Fama-French 3 Factors for BNP**

Summary statistics for: $R_{BNP} = \alpha + \beta_1[R_m - R_f] + \beta_2 SMB + \beta_3 HML + \varepsilon$. *Explanatory power of the regression is presented. Values and significance of the coefficients are included.*

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	6.028e-5	7.640e-4		0.079	0.937
β_1 (Rm-Rf)	0.012	8.411e-4	0.515	14.696	$< .001$
β_2 (SMB)	-0.003	0.002	-0.051	-1.434	0.153
β_3 (HML)	0.017	0.001	0.579	16.303	$< .001$

R	R^2	Adjusted R^2	RMSE
0.865	0.748	0.744	0.011

Table 4: **CAPM for BNP**

Summary statistics for: $R_{BNP} = \alpha + \beta[R_m - R_f] + \varepsilon$. *Explanatory power of the regression is presented. Values and significance of the coefficients are included.*

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	3.104e-4	0.001		0.269	0.788
β (Rm-Rf)	0.015	0.001	0.639	12.414	$< .001$

R	R^2	Adjusted R^2	RMSE
0.639	0.409	0.406	0.017