Contents

Chapter 3:Linear Models of Financial Returns.

1. Econometric Modelling of Financial Returns: a general framework

Financial data are mostly observational data: they are not generated by well-designed experiment to test hypothesis, they are given to the econometrician. These data can be used to construct non-causal predictive models and to evaluate treatment effects. The second exercise involves a deeper understanding of causation while the implementation of non-causal predictive modelling requires understanding conditional expectations. Econometric models of financial returns specify the distribution of a vector of variables $\mathbf{y}_t$ conditional upon other variables $\mathbf{z}_t$that are helpful in predicting them. The mapping between $\mathbf{y}_t$ and $\mathbf{z}_t$ is determined by some functional relation and some unknown parameters. All the relevant variables are stochastic and they are therefore characterized by a density function. Linear Econometric Models specify conditional means of the $\mathbf{y}_t$ as linear functions of the $\mathbf{z}_t$. . Think, for example, of the typical estimated equation derived from the Capital Asset Pricing Model[1]:

$$\begin{aligned} \left(r_t^i - r_t^{rf}\right) &= \beta_{0,i} + \beta_{1,i}\left(r_t^m - r_t^{rf}\right) + u_{i,t} \\ u_{i,t} &\sim n.i.d.\left(0, \sigma_i^2\right) \end{aligned} \tag{1}$$

The theoretical model predicts that the excess return on each risky asset on the riskfree asset is a linear function of the excess return of the market portfolio on the riskfree asset, the linear relationship has a slope equal to β_i and a intercept equal to zero. The excess returns of assets are determined by their exposure to a single common risk factor

[1] A more extensive discussion of the CAPM and its derivation will be provided in the next chapter. Note that strictly speacking the validity of the CAPM requires $\alpha_i = 0$.

(captured by the excess returns on the market portfolio) and an hydiosincratic risk factor which is captured by a noise term with zero mean and a constant (in time but not across different assets) variance σ_i^2. The CAPM is a specific case of linear multivariate model that relates the vector of excess returns to a single regressor, the market excess returns. The linear model feature some restrictions on the parameters, i.e. the intercept is zero. In the special case in which only one asset is considered, the CAPM delivers a univariate regression model.

Technically the CAPM asset pricing regression is a special case of a specification of the conditional density of a vector of variables $\mathbf{y}_t$. The conditioning is upon a vector of regressors $\mathbf{z}_t$, an information set available at time $t-1$, that contains past observations of regressors and dependent variables $\mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}$ and a set of parameters $\boldsymbol{\beta}_1$:

$$D\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}_1\right)$$

This conditional density is best interpreted as the outcome of a reduction process that allows a simplified representation of reality. Of course such a simplified representation omits an enormous amount of information. The validity of the model adopted is crucially affected by the importance of the omitted information in determining the density of $\mathbf{y}_t$.To understand the reduction process partiton the set of all variables into three types of variables:

$$\mathbf{x}_t = \left(\mathbf{w}_t, \mathbf{y}_t, \mathbf{z}_t\right),$$

$\mathbf{w}_t$ identifies variables which are ignored in the specification of the econometric model. Their exclusion might be motivated by a number of reasons, they could be unobservable or irrelevant to the problem investigated according to the specific theory that inspires the specification adopted by the econometrician. In practice these variables are ignored, in theory such a result is obtained by factorizing the joint density and integrating it with respect to $\mathbf{w}_t$.In formal terms, we have no information loss only if

$$D\left(\mathbf{y}_t, \mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}\right) = D\left(\mathbf{y}_t, \mathbf{z}_t, \mathbf{w}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \mathbf{W}_{t-1,}\boldsymbol{\theta}\right).$$

This is the statistical description of the model considered by the econometrician, this is technically called i.e. the reduced form of the structure of interest. In general this reduced form is a more general model than the one estimated. It is constructed by parameterizing $E\left(\mathbf{y}_t, \mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}\right)$ and by deriving a vector of innovations from the

difference between the vector of observed variables and the vector of their means. In the case of the CAPM the specification of the reduced form is the following one:

$$\begin{aligned} \left(r_t^i - r_t^{rf}\right) &= \mu_i + \beta_i u_{m,t} + u_{i,t} \\ \left(r_t^m - r_t^{rf}\right) &= \mu_m + u_{m,t} \\ \begin{pmatrix} u_{i,t} \\ u_{m,t} \end{pmatrix} &\sim n.i.d. \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{ii} & \sigma_{im} \\ \sigma_{im} & \sigma_{mm} \end{pmatrix} \right] \end{aligned} \tag{2}$$

The reduced form is the description of the data used (often implicitly) by the econometrician. This description of the data omits a lot of information, it is very important to make sure that no loss of relevant information occurs when concentrating on such a simplified statistical model. Visual inspection of the data is a very important first step in this direction, further insight can be gained by analyzing the residuals of the specification and by making sure that they do indeed possess the properties that the econometrician attributes to them. In the case of our model we can see that the maintained assumption behind its specification is the so called constant expected (excess) returns model: excess returns fluctuate randomly around a constant and no variables helps in predicting them. Visual inspection of the relevant time series (which are both the dependent variables and the regressors in the CAPM) and statistical analysis of the residuals (which in the case at hand are the de-meaned variables) should be implemented to validate the statistical model.

Unfortunately reduced forms and their validation are not the common first stage of the analysis as researchers tend to estimate directly a structure consistent with the theoretical model of interest.As a matter of fact in the case of CAPM the equation that it is estimated is (1) , which does not describe $D\left(\mathbf{y}_t, \mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}\right)$ but $D\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}_1\right)$. In this case the relevant problem to the researcher is inference on a subset $\boldsymbol{\beta}_1$ of the parameters $\boldsymbol{\beta}$ determining the conditional density of $\mathbf{y}_t$ given $\mathbf{z}_t$. The intereresting question here is about the possible loss of information in concentrating on $D\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}_1\right)$ rather than on $D\left(\mathbf{y}_t, \mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}\right)$.

Note that in general it is always possible to re-write $D\left(\mathbf{y}_t, \mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}\right)$ as follows:

$$\begin{aligned} & D\left(\mathbf{y}_t, \mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}\right) \\ = \; & D\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right) D\left(\mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right). \end{aligned} \tag{3}$$

The general case admits as a specific case the existence of a 'sequential cut', which we

represent as follows:

$$\begin{aligned} & D\left(\mathbf{y}_t, \mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}\right) \\ = \; & D\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}_1\right) D\left(\mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}_2\right). \end{aligned} \tag{4}$$

If the sequential cut is admissible and if the set on which the parameters $\boldsymbol{\beta}_1$are defined is totally independent from the set on which the parameters $\boldsymbol{\beta}_2$ are defined ($\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are *variation free*) then inference on $\boldsymbol{\beta}_1$can be performed by concentrating only on the conditional density for $\mathbf{y}_t$, without explicitly treating the marginal density for $\mathbf{z}_t$.

Let us go back to our CAPM example, what does weak exogeneity mean ? The estimated CAPM equation implies that

$$E\left(\left(r_t^i - r_t^{rf}\right) \mid \left(r_t^m - r_t^{rf}\right), \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \beta_i\right) = \alpha_i + \beta_i \left(r_t^m - r_t^{rf}\right) \tag{5}$$

Consider now the full system (1), weak exogeneity is satisfied if the conditional mean of $\left(r_t^i - r_t^{rf}\right)$ derived by considering the full system coincides with (5). To derive the conditional mean from the full system consider the following theorem:

Theorem 1 *For any* $\mathbf{x} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, given any* $(m \times n)$ $\mathbf{B}$ *matrix and any* $(m \times 1)$ *vector,* $\mathbf{d}$*, if* $\mathbf{y} = \mathbf{B}\mathbf{x} + \mathbf{d}$*, this implies* $\mathbf{y} \sim \mathbf{N}(\mathbf{B}\boldsymbol{\mu} + \mathbf{d}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$.

Consider a partitioning of an n-variate normal vector in two sub-vectors of dimensions n_1 and $n - n_1$:

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim \mathbf{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right).$$

we then have:

1. $\mathbf{x}_1 \sim \mathbf{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$;
2. $(\mathbf{x}_1 \mid \mathbf{x}_2) \sim \mathbf{N}\left(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$,

By applying 2. to (1) we have:

$$\begin{aligned} E\left(\left(r_t^i - r_t^{rf}\right) \mid \left(r_t^m - r_t^{rf}\right), \mathbf{I}_{t-1}, \beta_i\right) &= \mu_i + \left(\frac{\beta_i \sigma_{mm} + \sigma_{im}}{\sigma_{mm}}\right)\left(r_t^m - r_t^{rf} - \mu_m\right) \\ &= \left(\mu_i - \left(\beta_i + \frac{\sigma_{im}}{\sigma_{mm}}\right)\mu_m\right) + \left(\beta_i + \frac{\sigma_{im}}{\sigma_{mm}}\right)\left(r_t^m - r_t^{rf}\right) \end{aligned}$$

so weak exogeneity requires $\sigma_{im} = 0$. In the case weak exogeneity is note satisfied the data will deliver estimated parameters that reflect the full system, so the estimated slope of the CAPM will be the more distant from β_i the higher the absolute value of σ_{im}.

The interpretation of this condition is pretty simple: weak exogeneity is satisfied in the market excess returns do not respond to idiosyncratic shocks to the excess returns to asset i. To put it in different words, if the weak exogeneity condition is satisfied, than the structural model that is compatible with the reduced form system is unique. In this case we say that the structural model of interest is *identified.* Identification is a crucial condition for the validity of the interpretation of the estimated parameters. If a model is not identified its estimation is pointless. Weak exogeneity is a condition for identification and for validity of estimation. However estimation of parameters is not the only interesting econometrics exercise, after estiamtion of a model we might be interest in simulating it to produce forecasts or to assess the impact of shokcs on the variables included in the model. The validity of a model for the purpose of simulation and for the evaluation of treatment effects requires a different condition from weak exogeneity: strong exogeneity. We have strong exogeneity when the joint density can be factorized as follows:

$$\begin{aligned} & D\left(\mathbf{y}_t, \mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}\right) \\ = \; & D\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}_1\right) D\left(\mathbf{z}_t \mid \mathbf{Z}_{t-1}, \boldsymbol{\beta}_2\right). \end{aligned} \tag{6}$$

In this case the there is no feedback from the past history of $\mathbf{y}_t$on $\mathbf{z}_t$ and the conditional model can be validly simulated.

To sum up[2]

1. if the objective of the analysis is inference on the $\boldsymbol{\beta}_1$ parameters, then the joint density can be reduced to a conditional model if $\mathbf{z}_t$ is weakly exogenous for the estimation of the parameters of interest;

2. if the objective of the analysis is dynamic simulation, then the joint density can be reduced to a conditional model if $\mathbf{z}_t$ satisfies the conditions for strong exogeneity;

2. **From theory to data: the CAPM**

Estimation of parameters in the CAPM equation is important and can be useful for many purposes. Do you want to evaluate fund managers ? Estimate the parameters

[2]for an extensive discussion of exogeneity see Engle, Hendry (1996)

in the CAPM equations for the excess returns of the fund they manage and look at their alphas (Jensen(1968)). In fact, Jensen's alpha are useful also to test the CAPM model because the theory predicts that they should all be zero. Estimating parameters in the CAPM could also be useful to practitioners for estimating the cost of equity. To illustrate how the CAPM can be put at work we will consider time series data of monthly observations of different portfolios made available by Ken French from his website: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

HOME
BIOGRAPHY
CURRICULUM VITAE
WORKING PAPERS
DATA LIBRARY
• U.S. RESEARCH RETURNS
• BENCHMARKS
• U.S. RESEARCH BREAKPOINTS
• U.S. BOOK EQUITY DATA
• INTERNATIONAL RESEARCH RETURNS
• DEVELOPED MARKET FACTORS AND RETURNS NEW!
CONSULTING RELATIONSHIPS
FAMA / FRENCH FORUM
CONTACT INFORMATION

Current Research Returns

We have revised the market return used to measure Rm-Rf in the US. It is now the value-weight return of all CRSP firms incorporated in the US and listed on the NYSE, AMEX, or NASDAQ that have a CRSP share code of 10 or 11 at the beginning of month t, good shares and price data at the beginning of t, and good return data for t. Previously we used the CRSP NYSE/AMEX/NASDAQ Value-Weighted Market Index as the proxy for the market return. The set of firms in the new series is more consistent with the universe used to compute the other US returns.

	June 2014	Last 3 Months	Last 12 Months
Fama/French Research Factors			
Rm-Rf	2.59	4.50	25.14
SMB	2.98	-3.38	0.17
HML	-0.63	0.76	0.86
Fama/French Research Portfolios			
Small Value	4.53	1.67	25.12
Small Neutral	4.63	2.57	27.91
Small Growth	7.39	-0.02	22.11
Big Value	3.38	4.39	24.40
Big Neutral	2.45	5.41	24.55
Big Growth	1.77	4.56	25.69

The data are available in EXCEL format in the file FF_Data_CH3.xls. This files combines 3 data files from the website to make available monthly observation on 30 time series from July 1927 onward.

The first four time series come from the file Fama-French factors that contains, EXRET_MKT, Rf, SMB, and HML.

The excess return on the market, value-weight return of all CRSP firms incorporated in the US and listed on the NYSE, AMEX, or NASDAQ that have a CRSP share code of 10 or 11 at the beginning of month t, good shares and price data at the beginning of t, and good return data for t minus the one-month Treasury bill rate (from Ibbotson Associates).

The Fama/French factors are constructed using the 6 value-weighted portfolios formed on size and book-to-market.

SMB (Small Minus Big) is the average return on the three small portfolios minus the average return on the three big portfolios,

$$SMB = 1/3(SmallValue+SmallNeutral+SmallGrowth)-1/3(BigValue+BigNeutral+BigGrowth)$$

HML (High Minus Low) is the average return on the two value portfolios minus the average return on the two growth portfolios,

$$HML = 1/2(SmallValue + BigValue) - 1/2(SmallGrowth + BigGrowth)$$

The fifth time series is the momentum factor, MOM, tha comes from the file Momentum factor. It contains a momentum factor, constructed from six value-weight portfolios formed using independent sorts on size and prior return of NYSE, AMEX, and NASDAQ stocks. MOM is the average of the returns on two (big and small) high prior return portfolios minus the average of the returns on two low prior return portfolios. The portfolios are constructed monthly. Big means a firm is above the median market cap on the NYSE at the end of the previous month; small firms are below the median NYSE market cap. Prior return is measured from month -12 to - 2. Firms in the low prior return portfolio are below the 30th NYSE percentile. Those in the high portfolio are above the 70th NYSE percentile.

The last 25 time series come from the file 25_portfolios_5x5. This file contains value- and equal-weighted returns for the intersections of 5 ME portfolios and 5 BE/ME portfolios. The 25 time series we consider are to equal weighted returns.

The portfolios are constructed at the end of Jun. ME is market cap at the end of Jun. BE/ME is book equity at the last fiscal year end of the prior calendar year divided by ME as of 6 months before formation. Firms with negative BE are not included in any portfolio. PR11 are the returns on the portfolio made with the smallest firm and the lowest book equity, PR12 are the returns on the portfolio made with the smallest firm and the second lower book to equity and so on until PR55, which are the returns on the portfolio of the largest firms with the highest book to equity.

3. **Graphical and Descriptive Data Analysis**

Before doing any econometrics make sure that you know your data and you know them well. Descriptive statistics are helpful but the the best way to know your data is to apply

graphical analysis. If you are interested in running CAPM regressions a good point to start is the time-series plots of the returns you are interested into. We know that the CAPM is consistent with a Constant Expected Returns view of the world. So the first feature that returns (and excess returns) should show is that they fluctuate randomly around a constant. Time-series plot are univariate graphs in which only one variable at the time is considered. Further information gain be gained by multivariate graphs in which multiple variables are put together in a graph. Here the obvious graph would be a cross-plot, also called scatter-plot, of the excess returns on asset i and the excess return on the market portfolio. In case the CAPM is applied to many returns a scatter-plot matrix, that contains the set of all possible bivariate scatter-plots, could be an interesting and informative way of summarizing the data.

To illustrate how Graphical Data Analysis can be implemented consider the case of the application of the CAPM to two of the FF portfolios:PR15 and PR51. These are respectively the portfolios made by the smallest firms with the highest book-to-market and b the largest firm with the lowest book to market. We begin by plotting the excess returns on the risk free fromthese two portofolios and from the market portfolio over a sample of monthly data from 1962 to 2014:

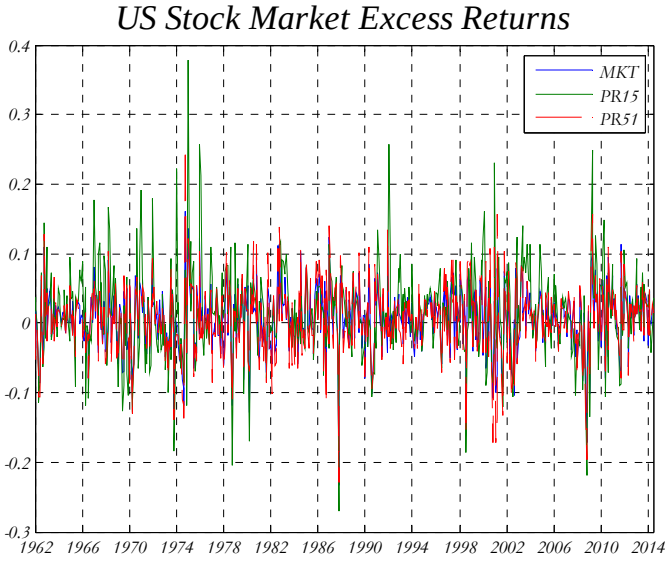


Figure 1

Figure 1 illustrates several interesting features of graphical analysis. First, the eyes are informative, excess returns are visually fluctuating randomly around a constant illustrating that the CER model good consistent with the CAPM, could be a good description

of the data. Second, the eyes could be deceptive, no strong systematic pattern of difference between the excess returns considered seems to emerge from the Figure. Consider now looking at the same data from a different angle: invest one euro at the beginning of the period and track the over time the value of the one euro invested respectively in the market portfolio in portfolio 15, in portfolio 51 and in the risk-free. Figure 2 illustrates that portfolio 15 strongly dominates the other in terms of total returns and that small differences in returns can turn out into huge differences in cumulative total returns.

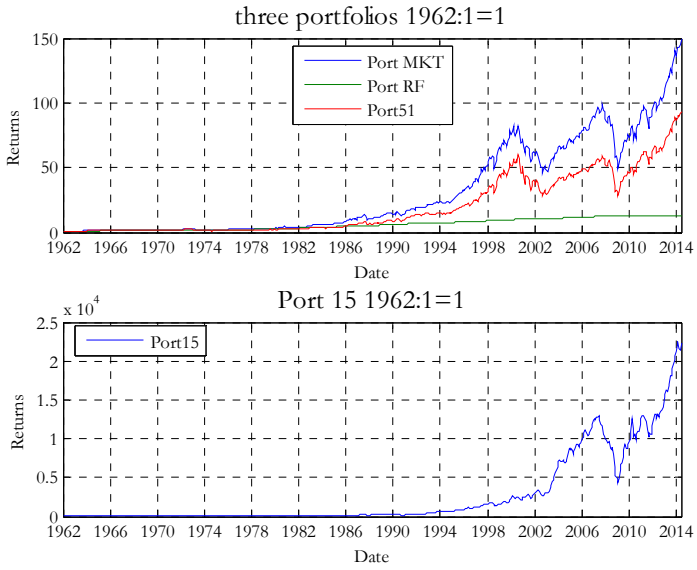


Figure 2

As a last first look at the data let us aggregate monthly returns into annual and have a look at them

Figure 3 reports annual and monthly returns on the market portfolio oserved at monthly frequencies:

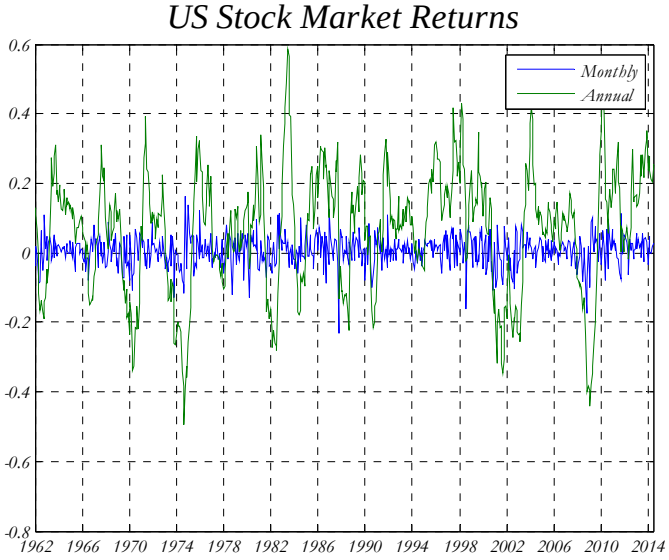


Figure 3

Apparently the CER model does much worse when applied to annual returns as they show a much stronger degree of persistence. Before rushing to conclusions it is important to keep in many that such persistence is directly caused by time-aggregation. To illustrate the point consider the case in which log monthly returns are correctly described by the CER model

$$\begin{aligned} r_{i,t} &= \alpha_i + u_{i,t} \\ u_{i,t} &\sim n.i.d.\left(0, \sigma_i^2\right) \end{aligned}$$

by definition annual log-returns are constructed as

$$r^A_{i,t,t+12} = \sum_{j=0}^{11} r_{i,t+j} = 12\alpha_i + \sum_{j=0}^{11} u_{i,t+j}$$

and autocorrelation up to the eleventh lag (with decreasing size) is automatically generated by the construction of returns. Consider now "decimated" annual returns when after construction of the series by aggregation of monthly returns we consider one observation per year (i.e. we consider observation separated by 12 months), the pattern of random fluctuations around a constant seem to re-emerge again in Figure 4

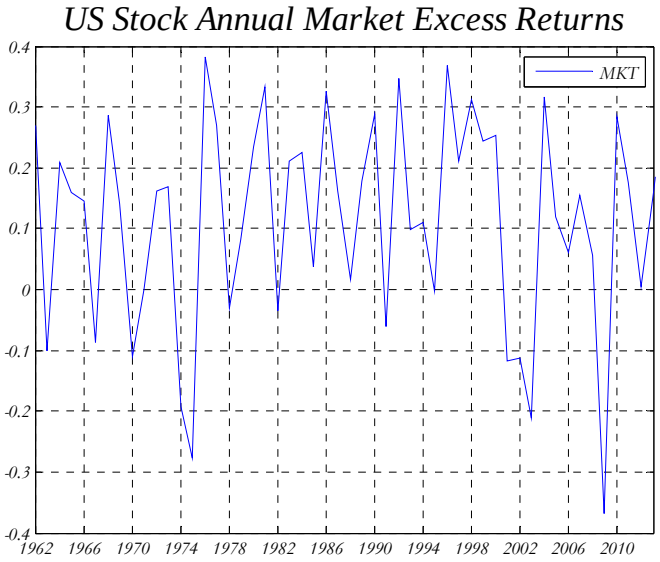


Figure 4

4. **Estimation Problem: Ordinary Least Squares**

To illustrate how estimation can be performed to derive conditional expectations , consider the following general representation of the model of interest:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & . & . & x_{1k} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{N1} & x_{N2} & . & . & x_{Nk} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_N \end{pmatrix}.$$

The vector $\mathbf{y}$ contains N observations on the dependent variable, while matrix $\mathbf{X}$ (NxK)contains N observations on the K regressor. In the case of our data set we have observations on monthly returns from 1927 to 2013 and the estimation of the CAPM for each portfolio included in the data-set implies filling $\mathbf{y}$ with the observation on the relevant

portfolio and filling $\mathbf{X}$ $(Nx2)$ with a first column made of ones and a second column containing the observations on the market portfolio. The vector $\boldsymbol{\beta}$ contains therefore two parameters to be estimated : a constant and the loading on the the uniuq factor considered as a potential determnant of excess returns.

The simplest way to derive estimates of the parameters of interest is the ordinary least squares (OLS) method. Such a method chooses values for the unknown parameters to minimize the magnitude of the non-observable components. In our simple bivariate case this amount to choosing a line that goes through the scatterplot of excess returns on each asset on the market excess returns such that it provides the best fit. The best fit is obtained by minimizing the sum of squared vertical deviations of the data points from the fitted line. Define the following quantity:

$$\mathbf{e}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{e}(\boldsymbol{\beta})$ is a $(n \times 1)$ vector. If we treat $\mathbf{X}\boldsymbol{\beta}$, as a (conditional) prediction for $\mathbf{y}$, then we can consider $\mathbf{e}(\boldsymbol{\beta})$ as a forecasting error. The sum of the squared errors is then

$$\mathbf{S}(\boldsymbol{\beta}) = \mathbf{e}(\boldsymbol{\beta})' \mathbf{e}(\boldsymbol{\beta}).$$

The OLS method produces an estimator of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$, defined as follows:

$$\mathbf{S}\left(\widehat{\boldsymbol{\beta}}\right) = \min_{\boldsymbol{\beta}} \mathbf{e}(\boldsymbol{\beta})' \mathbf{e}(\boldsymbol{\beta}).$$

Given $\widehat{\boldsymbol{\beta}}$, we can define an associated vector of residual $\widehat{\epsilon}$ as $\widehat{\epsilon} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$. The OLS estimator is derived by considering the necessary and sufficient conditions for $\widehat{\boldsymbol{\beta}}$ to be a unique minimum for **S:**

1. $\mathbf{X}'\widehat{\epsilon} = 0$;
2. $\text{rank}(\mathbf{X}) = k$.

Condition 1 imposes orthogonality between the right-hand side variables on the OLS residuals, and ensures that residuals have an average of zero when a constant is included among the regressors. Condition 2 requires that the columns of the $\mathbf{X}$ matrix are linearly independent: no variable in $\mathbf{X}$ can be expressed as a linear combination of the other variables in $\mathbf{X}$.

From 1 we derive an expression for the OLS estimates:

$$\begin{aligned} \mathbf{X}'\widehat{\epsilon} &= \mathbf{X}'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = 0, \\ \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned}$$

4.1. *Properties of the OLS estimates*

We have derived the OLS estimator without any assumption on the statistical structure of the data. However, the statistical structure of the data is needed to define the properties of the estimator. To illustrate them, we refer to the basic concepts of mean and variance of vector variables.

Given a generic vector of variables, $\mathbf{x}$,

$$\mathbf{x} = \begin{pmatrix} x_1 \\ . \\ . \\ . \\ x_n \end{pmatrix},$$

we define the mean vector $E(\mathbf{x})$ and the mean matrix of outer products $E(\mathbf{x}\mathbf{x}')$ as:

$$\begin{aligned} E(\mathbf{x}) &= \begin{pmatrix} E(x_1) \\ . \\ . \\ . \\ E(x_n) \end{pmatrix}, \\ E(\mathbf{x}\mathbf{x}') &= E\begin{pmatrix} x_1^2 & x_1x_2 & . & . & x_1x_n \\ . & x_2^2 & . & . & x_2x_n \\ . & . & . & . & . \\ . & . & . & . & . \\ x_nx_1 & x_nx_2 & . & . & x_n^2 \end{pmatrix} \\ &= \begin{pmatrix} E(x_1^2) & E(x_1x_2) & . & . & E(x_1x_n) \\ . & E(x_2^2) & . & . & E(x_2x_n) \\ . & . & . & . & . \\ . & . & . & . & . \\ E(x_nx_1) & E(x_nx_2) & . & . & E(x_n^2) \end{pmatrix}. \end{aligned}$$

The variance-covariance matrix of $\mathbf{x}$ is the defined as:

$$\begin{aligned} var(\mathbf{x}) &= E(\mathbf{x}-E(\mathbf{x}))\,E(\mathbf{x}-E(\mathbf{x}))' \\ &= E(\mathbf{x}\mathbf{x}') - E(\mathbf{x})\,E(\mathbf{x})'. \end{aligned}$$

Note that the variance-covariance matrix is symmetric and positive definite, by construction. Given an arbitrary $\mathbf{A}$ vector of dimension n, we have:

$$var(\mathbf{A}'\mathbf{x}) = \mathbf{A}'var(\mathbf{x})\,\mathbf{A}.$$

The first relevant hypothesis for the derivation of the statistical properties of OLS regards the relationship between disturbances and regressors in the estimated equation. This hypothesis is constructed in two parts: first we assume that $E(\mathbf{y}_i \mid \mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$, ruling out the contemporaneous correlation between residuals and regressors (note that assuming the validity of this hypothesis implies that there are no omitted variables correlated with the regressors), second we assume that the components of the available sample are independently drawn. The second assumption guarantees the equivalence between $E(\mathbf{y}_i \mid \mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$ and $E(\mathbf{y}_i \mid \mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_n) = \mathbf{x}_i'\boldsymbol{\beta}$. Using vector notation, we have:

$$E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta},$$

which is equivalent to

$$E(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}. \tag{7}$$

Note that hypothesis (7) is very demanding. It implies that

$$E(\boldsymbol{\epsilon}_i \mid \mathbf{x}_1, ...\mathbf{x}_i, ...\mathbf{x}_n) = 0 \quad (i = 1, ...n).$$

The conditional mean is, in general, a non-linear function of $(\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_n)$ and (7) requires that such a function is a constant of zero. Note that (7) requires that each regressor is orthogonal not only to the error term associated with the same observation ($E(x_{ik}\varepsilon_i) = 0$ for all k), but also to the error tem associated with each other observation ($E(x_{jk}\varepsilon_i) = 0$ for all $j \neq k$). This statement is proved by using the properties of conditional expectations.

Since $E(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}$ implies, from the law of iterated expectations, that $E(\boldsymbol{\epsilon}) = \mathbf{0}$, we have

$$E(\varepsilon_i \mid x_{jk}) = E[E(\varepsilon_i \mid \mathbf{x}) \mid x_{jk}] = 0. \tag{8}$$

Then

$$\begin{aligned} E(\varepsilon_i x_{jk}) &= E[E(\varepsilon_i x_{jk} \mid x_{jk})] && (9)\\ &= E[x_{jk}E(\varepsilon_i \mid x_{jk})] && (10)\\ &= 0. && (11) \end{aligned}$$

Such a hypothesis is clearly false in any time-series model when the time-series shows some degree of persistence. If the CER model is valid such property is satisfied by the CAPM regressors as the CER implies that no variable should help in predicting returns, including their past history. The graphic analysis conducted prior to regression can also

be informative on this issue by providingsome visual evidence on the persistence of the variable entering the relevant empirical model. This property, satified by monthly returns, is rarely satisfied when considering macro and financial time-series. Think of a simplest time-series model for a generic variable y:

$$y_t = a_0 + a_1 y_{t-1} + u_t.$$

Clearly, if $a_1 \neq 0$, then, $E\left(u_{t-1} \mid y_{t-1}\right) \neq 0$ although it is still true that $E\left(u_t \mid y_{t-1}\right) = 0$, and (7) breaks down, without any omitted variable problem.

This is why we use a monthle returns in this introductory chapter. We shall complicate the framework to deal properly with time-series observations in the next chapters.

The second hypothesis defines the constancy of the conditional variance of shocks:

$$E\left(\boldsymbol{\epsilon}'\boldsymbol{\epsilon} \mid \mathbf{X}\right) = \sigma^2 I, \tag{12}$$

where σ^2 is a constant independent from $\mathbf{X}$. In the case of our data, this is a strong assumption unlikely to be met in practice. Model of time-varying volatility are of crucial importnace in finance and we shall reconsider this issue at a later stage of the book.

The third hypothesis is the one already introduced, which guarantees that the OLS estimator can be derived:

$$rank\left(\mathbf{X}\right) = k. \tag{13}$$

Under hypotheses $(7) - (13)$ we can derive the properties of the OLS estimator.

Property 1: unbiasedness

The conditional expectation (with respect to **X)** of the OLS estimates is the vector of unknown parameters $\boldsymbol{\beta}$:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\ E\left(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}\right) &= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\epsilon} \mid \mathbf{X}) \\ &= \boldsymbol{\beta}, \end{aligned}$$

by hypothesis (7).

Property 2: variance of OLS

The conditional variance of the OLS estimator is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$:

$$\begin{aligned} var\left(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}\right) &= E\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)' \mid \mathbf{X}\right) \\ &= E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mid \mathbf{X}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 I\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Property 3: Gauss-Markov theorem

The OLS estimator is the most efficient in the class of linear unbiased estimators.

Consider the class of linear estimators:

$$\boldsymbol{\beta}_L = \mathbf{L}\mathbf{y}.$$

This class is defined by the set of matrices $(k \times n)$ $\mathbf{L}$, which are fixed when conditioning upon $\mathbf{X}$. $\mathbf{L}$ does not depend on $\mathbf{y}$. Therefore we have:

$$\begin{aligned} E(\boldsymbol{\beta}_L \mid \mathbf{X}) &= E(\mathbf{L}\mathbf{X}\boldsymbol{\beta}+\mathbf{L}\boldsymbol{\varepsilon} \mid \mathbf{X}) \\ &= \mathbf{L}\mathbf{X}\boldsymbol{\beta}, \end{aligned}$$

and $\mathbf{L}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$ only if $\mathbf{L}\mathbf{X} = \mathbf{I}_k$. Such a condition is obviously satisfied by the OLS estimator, which is obtained by setting $\mathbf{L} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The variance of the general estimator in the class of linear unbiased estimators is readily obtained as:

$$\begin{aligned} var(\boldsymbol{\beta}_L \mid \mathbf{X}) &= E(\mathbf{L}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{L}' \mid \mathbf{X}) \\ &= \sigma^2\mathbf{L}\mathbf{L}'. \end{aligned}$$

To show that the OLS estimator is the most efficient within this class we have to show that the variance of the OLS estimator differs from the variance of the generic estimator in the class by a positive semidefinite matrix.

To this aim define $\mathbf{D} = \mathbf{L} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$; $\mathbf{LX} = \mathbf{I}$ requires $\mathbf{DX} = \mathbf{0}$.

$$\begin{aligned}
\mathbf{LL}' &= \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}\right)\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{D}'\right) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}' + \\
&\quad + \mathbf{DX}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{DD}' \\
&= (\mathbf{X}'\mathbf{X})^{-1} + \mathbf{DD}',
\end{aligned}$$

from which we have that

$$var(\boldsymbol{\beta}_L \mid \mathbf{X}) = var\left(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}\right) + \sigma^2\mathbf{DD}',$$

which proves the point. For any given matrix $\mathbf{D}$, (not necessarily square), the symmetric matrix $\mathbf{DD}'$ is positive semidefinite.

4.2. *Residual Analysis*

Consider the following representation:

$$\begin{aligned}
\widehat{\boldsymbol{\epsilon}} &= \mathbf{y} - \mathbf{X}\widehat{\beta} \\
&= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{My},
\end{aligned}$$

where $\mathbf{M} = \mathbf{I}_n - \mathbf{Q}$, and $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The $(n \times n)$ matrices $\mathbf{M}$ and $\mathbf{Q}$, have the following properties:

1. they are symmetric: $\mathbf{M}' = \mathbf{M}, \mathbf{Q}' = \mathbf{Q}$;
2. they are idempotent: $\mathbf{QQ} = \mathbf{Q}, \mathbf{MM} = \mathbf{M}$;
3. $\mathbf{MX} = \mathbf{0}, \mathbf{MQ} = \mathbf{0}, \mathbf{QX} = \mathbf{X}$.

Note that the OLS projection for $\mathbf{y}$ can be written as $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{Qy}$, and that $\widehat{\boldsymbol{\epsilon}} = \mathbf{My}$, from which we have the known result of orthogonality between the OLS residuals and regressors. We also have $\mathbf{My} = \mathbf{MX}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\epsilon} = \mathbf{M}\boldsymbol{\epsilon}$, given that $\mathbf{MX} = \mathbf{0}$. Therefore we have a very well-specified relation between the OLS residuals and the errors in the model $\widehat{\boldsymbol{\epsilon}} = \mathbf{M}\boldsymbol{\epsilon}$, which cannot be used to derive the errors given the residuals, since the $\mathbf{M}$ matrix is not invertible.

We can re-write the sum of squared residuals as:

$$S\left(\widehat{\boldsymbol{\beta}}\right) = \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}'\mathbf{M}'\mathbf{M}\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}.$$

$S\left(\widehat{\boldsymbol{\beta}}\right)$ is an obvious candidate for the construction of an estimate for σ^2. To derive an estimate of σ^2 from $S\left(\widehat{\boldsymbol{\beta}}\right)$, we introduce the concept of trace. The trace of a square matrix is the sum of all elements on its principal diagonal. The following properties are relevant:

1. given any two square matrices $\mathbf{A}$ and $\mathbf{B}$, $tr\left(\mathbf{A}+\mathbf{B}\right) = tr\mathbf{A} + tr\mathbf{B}$;
2. given any two matrices $\mathbf{A}$ and $\mathbf{B}$, $tr\left(\mathbf{AB}\right) = tr\left(\mathbf{BA}\right)$;
3. the rank of an idempotent matrix is equal to its trace.

Using property 2 together with the fact that a scalar coincides with its trace, we have:

$$\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon} = tr\left(\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}\right) = tr\left(\mathbf{M}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right).$$

Now we analyse the expected value of $S\left(\widehat{\boldsymbol{\beta}}\right)$, conditional upon $\mathbf{X}$:

$$\begin{aligned} E\left(S\left(\widehat{\boldsymbol{\beta}}\right) \mid \mathbf{X}\right) &= E\left(tr\mathbf{M}\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \mathbf{X}\right) \\ &= trE\left(\mathbf{M}\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \mathbf{X}\right) \\ &= tr\mathbf{M}\left(E\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \mathbf{X}\right) \\ &= \sigma^2 tr\mathbf{M}. \end{aligned}$$

From properties 1 and 2 we have:

$$\begin{aligned} tr\mathbf{M} &= tr\mathbf{I}_n - tr\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right) \\ &= n - tr\left(\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right) \\ &= n - k. \end{aligned}$$

Therefore, an unbiased estimate of σ^2 is given by $s^2 = S\left(\widehat{\boldsymbol{\beta}}\right)/\left(n-k\right)$.

Using the result of orthogonality between the OLS projections and residuals, we can write:

$$var\left(\mathbf{y}\right) = var\left(\widehat{\mathbf{y}}\right) + var\left(\widehat{\boldsymbol{\epsilon}}\right),$$

from which we can derive the following residual-based indicator of the goodness of fit:

$$R^2 = \frac{var\left(\widehat{\mathbf{y}}\right)}{var\left(\mathbf{y}\right)} = 1 - \frac{var\left(\widehat{\boldsymbol{\epsilon}}\right)}{var\left(\mathbf{y}\right)}.$$

The information contained in R^2 is associated with the information contained in the standard error of the regression, which is the square root of the estimated variance of OLS residuals.

Note that, when a model is estimated in logarithms, residuals and, consequently, the standard error of the regression do not depend on the unit of measure in which the variables are expressed. In fact, we have:

$$\begin{aligned} \widehat{\boldsymbol{\epsilon}} &= \log(\mathbf{y}) - \log(\widehat{\mathbf{y}}) \\ &= \log\left(\frac{\mathbf{y}}{\widehat{\mathbf{y}}}\right) = \log\left(1 + \frac{\mathbf{y} - \widehat{\mathbf{y}}}{\widehat{\mathbf{y}}}\right) \simeq \frac{\mathbf{y} - \widehat{\mathbf{y}}}{\widehat{\mathbf{y}}}. \end{aligned}$$

When the model is not specified in logarithms, standard errors are usually intepreted by dividing them by the mean of the dependent variable.

5. **Interpreting Regression Results**

Intepreting regression results is not a simple exercise. We propose to split these procedure in three steps.

First, understand the relevance of our regression independently from inference on the parameters. There is an easy way to do this: suppose all parameters in the model are known and identical to the estimated values and learn how to read these.

Second, introduce a measure of sampling variability and evaluate again what you know taking into account that parameters are estimeted and there is uncertainty surroinding your point estimates.

Third, remember that each regression is run after a reduction process has been, explicitly or implicitly implemented. The relevant question is what happens if something went wrong in the reduction process? What are the consequences of omitting relevant information or of including irrelevant one in your specification?

5.1. *The R^2 as a measure of relevance of a regression*

Relevance of a regression is different form statistical significance of the estimated parameters. In fact, confusing statistical significance of the estimated parameter describing the effect of a regressor on the dependent variable with practical relevance of that effect is a rather common mistake in the use of the linear model. Statistical inference is a tool for estimating parameters in a probability model and assessing the amount of sampling variability. Statistics gives us indication on what we can say about the values of the parameters in the model on the basis of our sample.

The relevance of a regression is determined by the share of the unconditional variance of $\mathbf{y}$ that is explained by the variance of $E(\mathbf{y} \mid \mathbf{X})$. Measuring how large is the share of the unconditional variance of $\mathbf{y}$ explained by the regression function is the fundamental role of R^2.

To illustrate the point let us consider two specific cases of applications of the CAPM, in which we simulate the data for the excess returns on the market portfolio $\left(r_t^m - r_t^{rf}\right)$ and the excess returns on two hypothetical assets $\left(r_t^1 - r_t^{rf}\right), \left(r_t^2 - r_t^{rf}\right)$. We assume that the Data Generating Process implicit in the CAPM is the same for the two assets but we calibrate the two processes differently as follows:

$$\begin{aligned}
\left(r_t^i - r_t^{rf}\right) &= 0.8\sigma_m u_{m,t} + \sigma_i u_{i,t} \qquad (14)\\
\left(r_t^m - r_t^{rf}\right) &= \mu_m + \sigma_m u_{m,t}\\
\begin{pmatrix} u_{i,t} \\ u_{m,t} \end{pmatrix} &\sim n.i.d.\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]\\
\mu_m &= 0.0065, \sigma_m = 0.054, \sigma_1 = 0.09, \sigma_2 = 0.005 \qquad (15)
\end{aligned}$$

We simulate an artificial sample of 1056 (same length with the sample July 1926-June2014) observations for each process. μ_m and σ_m are calibrated to match the first two moments of the market portfolio excess returns over the sample 1926:7-2014:7. While the standard errors of the two excess returns are calibrated to deliver R^2 in the CAPM regression of respectively about .22 and .98. By running the two CAPM regressions on the artificial sample we obtain the following results:

TABLE 3.1: The estimation of the CAPM on artificial data

Dependent Variable $\left(r_t^1 - r_t^{rf}\right)$				
Regressor	Coefficient	Std. Error	t-ratio	Prob.
$\left(r_t^m - r_t^{rf}\right)$	0.875		17.48	0.000
R^2 0.22	S.E. of regression 0.0076			

Dependent Variable $\left(r_t^2 - r_t^{rf}\right)$				
Regressor	Coefficient	Std. Error	t-ratio	Prob.
$\left(r_t^m - r_t^{rf}\right)$	0.793		201.86	0.000
R^2 0.972	S.E. of regression 0.0000			

In both cases the estimated beta are statistically significant and very close to their true value of 0.8. Consider now the following experiment, simulate again the processes but for introduce at some point a temporary shift of two per cent in the excess returns in the market portfolio. Look first at the simulated process for $\left(r_t^1 - r_t^{rf}\right)$ can you tell when the shift happened ? Consider now the same thought experiment by examining $\left(r_t^2 - r_t^{rf}\right)$.

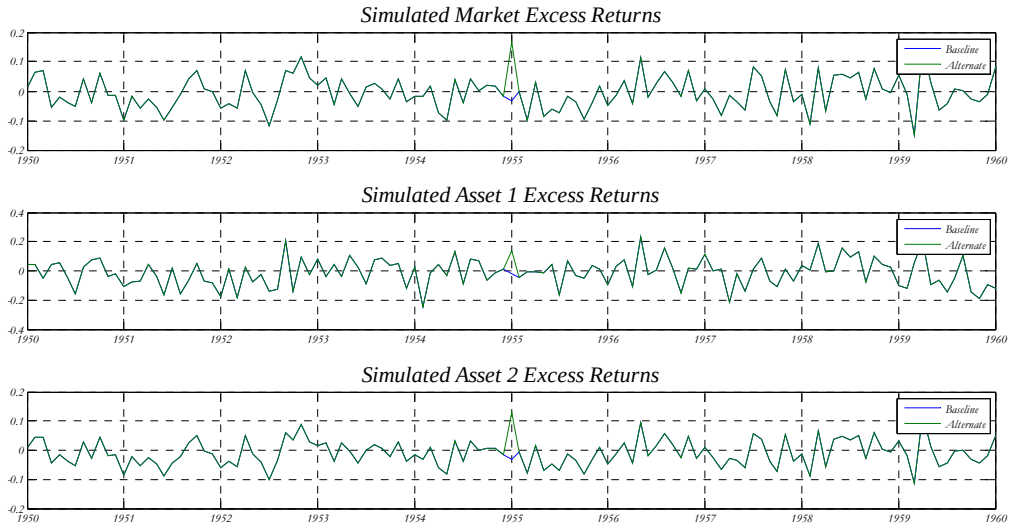


Figure 5

Figure 5 tells us that the shift in market excess returns happened in 1955:1,this shift can be easily traced in the Excess Returns for Asset 2 but it cannot be traced in the Excess returns for assset 1. The CAPM model for excess returns on asset 2 is practically relevant while the evidence in this direction is much more limited for excess returns on asset 1.

In both experiments the conditional expectation changes of the same amount but the share of the unconditional variance of $\mathbf{y}$ explained by the regression function is very different, as different are the $R^2 s$. In one case the change implied by the conditional expectation is "drowned" in the "noise" " and the change is not identifiable, while in the other the effect is clearly visible. This super-simple example shows that the actual meaning and relevance of the same vector of coefficients is very heterogenous and it is strongly affected affected by the corresponding R^2. It also shows that simulation is much more important than estimation to understand the properties of any given econometric model and the empirical relevance of the results.

5.2. *Inference in the Linear Regression Model*

Users of eeconometric models in finance attributes high priority to the concept of "statistical significance" of their estimates. In the standard statistical jargon an estimate of a parameter is "statistical significant" if its estimated value, compared with its sampling standard deviation makes it unlikely that in other samples the estimate may change of sign.

In the linear regression model the statistical index mostly used is the t-ratio and an estimated parameter has a significance which is usually measured in terms of its P-value, the probability with which that coefficient is equal to zero.

In the previous section we have illustrated how statistical significance is different from relevance: even if a parameter is known and is different from zero (so that its P-value is exactly 0) the actual relevance of the corresponding regressor could be absolutely negligible. To further illustrate the idea that statistical significance is a measure of the quality of the estimate think of an iid sample, if the size of the sample is sufficiently large, then all parameters become "statistically significant". The sampling standard deviation decreases at speed of sqaure root of the size of the sample and even a practically negligible effect can be estimated with enough precision to allow us to distinguish it from zero. In the previous section we have shown with a sample example that the common confusion between "statistical significance" and "relevance" quiclly disappear when models are used for forecasting or simulation analysis.

In this section we illustrate the basic principles that allow us to evaluate statistical significance and to perfortest of relevant hypothesis on the estimated coefficient in a linear model.

5.2.1. **Elements of distribution theory**

We consider the distribution of a generic n-dimensional vector $\mathbf{z}$, together with the derived distribution of the vector $\mathbf{x} = g(\mathbf{z})$ which admits the inverse $\mathbf{z} = h(\mathbf{x})$, with $h = g^{-1}$. If $prob(z_1 < z < z_2) = \int_{z_1}^{z_2} f(z)\, dz$, and $prob(x_1 < x < x_2) = \int_{x_1}^{x_2} f^*(x)\, dx$, then:

$$f^*(x) = f(h(x))\, J,$$

where $J = \begin{vmatrix} \frac{\partial h_1}{\partial x_1} & \dots & \frac{\partial h_n}{\partial x_1} \\ \dots & \dots & \dots \\ \frac{\partial h_1}{\partial x_n} & \dots & \frac{\partial h_n}{\partial x_n} \end{vmatrix} = \left|\frac{\partial \mathbf{h}}{\partial \mathbf{x}'}\right|.$

The normal distribution The standardized normal univariate has the following distribution:

$$\begin{aligned} f(z) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right), \\ E(z) &= 0, \qquad var(z) = 1. \end{aligned}$$

By considering the transformation $x = \sigma z + \mu$, we derive the distribution of the univariate normal as:

$$\begin{aligned} f(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \\ E(x) &= \mu, \qquad var(x) = \sigma^2. \end{aligned}$$

Consider now the vector $\mathbf{z} = (z_1, z_2, ..., z_n)$, such that

$$f(\mathbf{z}) = \prod_{i=1}^{n} f(z_i) = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\mathbf{z}'\mathbf{z}\right).$$

$\mathbf{z}$ is, by construction, a vector of normal independent variables with zero mean and identity variance covariance matrix. The conventional notation is $\mathbf{z} \sim \mathbf{N}(0, I_n)$.

Consider now the linear transformation,

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu},$$

where $\mathbf{A}$ is an $(n \times n)$ invertible matrix. We consider the following transformation $\mathbf{z} = \mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ with Jacobian $J = |\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$. By applying the formula for the transformation of variables, we have:

$$f(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \left|\mathbf{A}^{-1}\right| \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \mathbf{A}^{-1\prime}\mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right),$$

which, by defining the positive definite matrix $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$, equals

$$f(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \left|\boldsymbol{\Sigma}^{-\frac{1}{2}}\right| \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \tag{16}$$

The conventional notation for the multivariate normal is $\mathbf{x} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The formula of the trasnformation of variable allows us to better understand the theorem introduced in a previous section of this chapter. ,

Theorem 2 *For any* $\mathbf{x} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, given any* $(m \times n)$ $\mathbf{B}$ *matrix and any* $(m \times 1)$ *vector,* $\mathbf{d}$*, if* $\mathbf{y} = \mathbf{B}\mathbf{x} + \mathbf{d}$*, this implies* $\mathbf{y} \sim \mathbf{N}(\mathbf{B}\boldsymbol{\mu} + \mathbf{d}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$.

Consider a partitioning of an n-variate normal vector in two sub-vectors of dimensions n_1 and $n - n_1$:

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim \mathbf{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

By applying the formula for the transofrmation of variables, we obtain two results:

1. $\mathbf{x}_1 \sim \mathbf{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, which follows from applying the general formula in the case $\mathbf{d} = \mathbf{0}$, $\mathbf{B} = (I_{n_1} \ \mathbf{0})$;

2. $(\mathbf{x}_1 \mid \mathbf{x}_2) \sim \mathbf{N}\left(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$, which is obtained by applying the general formula to the case $\mathbf{d} = \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}}\mathbf{x}_2, \mathbf{B} = \left(I_{n_1} \ -\Sigma_{12}\Sigma_{22}^{-1}\right)$.

Result 2 shows clearly that the absence of correlation is equivalent to independence within the framework of a multivariate normal. This result is justified by the fact that the normal distribution is entirely described by its first two moments.

Distributions derived from the normal Consider $\mathbf{z} \sim \mathbf{N}(0, I_n)$, an n-variate standard normal. The distribution of $\boldsymbol{\omega} = \mathbf{z}'\mathbf{z}$ is defined as a $\chi^2(n)$ distribution with n degrees of freedom. Consider two vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ of dimensions n_1 and n_2 respectively, with the following distribution:

$$\begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} \sim \mathbf{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} I_{n_1} & 0 \\ 0 & I_{n_2} \end{pmatrix} \right).$$

We have $\boldsymbol{\omega}_1 = \mathbf{z}_1'\mathbf{z}_1 \sim \chi^2(n_1)$, $\boldsymbol{\omega}_2 = \mathbf{z}_2'\mathbf{z}_2 \sim \chi^2(n_2)$, and $\boldsymbol{\omega}_1 + \boldsymbol{\omega}_2 = \mathbf{z}_1'\mathbf{z}_1 + \mathbf{z}_2'\mathbf{z}_2 \sim \chi^2(n_1 + n_2)$. In general, the sum of two independent $\chi^2(n)$ distributions is in itself distributed as χ^2 with a number of degrees of freedom equal to the sum of the degrees of freedom of the two χ^2.

Our discussion of the multivariate normal concludes that if $\mathbf{x} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(n)$.

A related result establishes that if $\mathbf{z} \sim \mathbf{N}(0, I_n)$ and $\mathbf{M}$ is a symmetric idempotent $(n \times n)$ matrix of rank r, then $\mathbf{z}'\mathbf{M}\mathbf{z} \sim \chi^2(r)$.

Another distribution related to the normal is the F-distribution. The F-distribution is obtained as the ratio of two independent χ^2 divided by the respective degrees of freedom. Given $\boldsymbol{\omega}_1 \sim \chi^2(n_1)$, and $\boldsymbol{\omega}_2 \sim \chi^2(n_2)$, we have:

$$\frac{\boldsymbol{\omega}_1/n_1}{\boldsymbol{\omega}_2/n_2} \sim F(n_1, n_2).$$

The Student's $t-$distribution is then defined as:

$$t_n = \sqrt{F(1,n)}.$$

Another useful result establishes that two quadratic forms in the standard multivariate normal, $\mathbf{z}'\mathbf{M}\mathbf{z}$ and $\mathbf{z}'\mathbf{Q}\mathbf{z}$, are independent if $\mathbf{MQ} = \mathbf{0}$. We can finally state the following theorem, which is fundamental to the statistical inference in the linear model:

Theorem 3 *If* $\mathbf{z} \sim \mathbf{N}(0, I_n)$, $\mathbf{M}$ *and* $\mathbf{Q}$ *are symmetric and idempotent matrices of ranks* r *and* s *respectively and* $\mathbf{MQ} = \mathbf{0}$, *then* $\frac{\mathbf{z}'\mathbf{Q}\mathbf{z}}{\mathbf{z}'\mathbf{M}\mathbf{z}} \frac{r}{s} \sim \mathbf{F}(s, r)$.

5.2.2. **The conditional distribution** $y \mid \mathbf{X}$

To perform inference in the linear regression model, we need a further hypothesis to specify the distribution of $\mathbf{y}$ conditional upon $\mathbf{X}$**:**

$$y \mid \mathbf{X} \sim \mathbf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I\right), \tag{17}$$

or, equivalently

$$\epsilon \mid \mathbf{X} \sim \mathbf{N}\left(\mathbf{0}, \sigma^2 I\right). \tag{18}$$

Given (17) we can immediately derive the distribution of $\left(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}\right)$ which, being a linear combination of a normal distribution, is also normal:

$$\left(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}\right) \sim \mathbf{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right). \tag{19}$$

Equation (19) constitutes the basis to construct confidence intervals and to perform hypothesis testing in the linear regression model. Consider the following expression:

$$\begin{aligned} \frac{\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mathbf{X}'\mathbf{X} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{\sigma^2} &= \frac{\boldsymbol{\epsilon}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\epsilon}}{\sigma^2} \\ &= \frac{\boldsymbol{\epsilon}'\mathbf{Q}\boldsymbol{\epsilon}}{\sigma^2}, \end{aligned}$$

and, applying the results derived in the previous section, we know that

$$\frac{\boldsymbol{\epsilon}'\mathbf{Q}\boldsymbol{\epsilon}}{\sigma^2} \mid \mathbf{X} \sim \chi^2(k). \tag{20}$$

Equation (20) is not useful in practice, as we do not know σ^2. However, we know that

$$\frac{S\left(\widehat{\boldsymbol{\beta}}\right) \mid \mathbf{X}}{\sigma^2} = \frac{\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}}{\sigma^2} \mid \mathbf{X} \sim \chi^2(T-k). \tag{21}$$

Since $\mathbf{MQ} = \mathbf{0}$, we know the distribution of the ratio of (20) and (21); moreover, taking the ratio, we get rid of the unknown term σ^2:

$$\frac{\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mathbf{X}'\mathbf{X} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) / \sigma^2}{s^2/\sigma^2} = \frac{\boldsymbol{\epsilon}'\mathbf{Q}\boldsymbol{\epsilon}}{\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}} (T-k) \sim kF(k, T-k). \tag{22}$$

We use result (22) to obtain from the tables of the F-distribution the critical value $F^*_\alpha(k, T-k)$ such that

$$prob\left[F(k, T-k) > F^*_\alpha(k, T-k)\right] = \alpha, \qquad 0 < \alpha < 1,$$

for different values of α we are in the position of evaluating exactly an inequality of the following form:

$$prob\left\{\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mathbf{X}'\mathbf{X} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \leq ks^2 F^*_\alpha(k, T-k)\right\} = 1 - \alpha,$$

which defines confidence intervals for $\boldsymbol{\beta}$ centred upon $\widehat{\boldsymbol{\beta}}$. Hypothesis testing is strictly linked to the derivation of confidence intervals. When testing the hypothesis, we aim at rejecting the validity of restrictions imposed on the model on the basis of the sample evidence. Within this framework, $(7) - (19)$ are the maintained hypothesis and the restricted version of the model is identified with the null hypothesis H_0. Following the Neyman−Pearson approach to hypothesis testing, one derives a statistic with known distribution under the null. Then the probability of the first-type error (rejecting H_0 when it is true) is fixed at α. For example, we use a test at the level α of the null hypothesis $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, based on the F-statistic, when we do not reject the null H_0 if $\boldsymbol{\beta}_0$ lies within the confidence interval associated with the probability $1 - \alpha$. However, in practice, this is not a useful way of proceeding, as the economic hypotheses of interest rarely involve a number of restrictions equal to the number of estimated parameters. Think of the CAPM for example, testing its validity given a bivariate regression of any given portfolio excess returns on the market portfolio excess returns requires testing the restriction that a subset of the estimted coefficients (the constant) is eqaul to zero.

The general case of interest is therefore the one when we have r restrictions on the vector of parameters with $r < k$. If we limit our interest to the class of linear restrictions, we can express them as

$$H_0 = \mathbf{R}\boldsymbol{\beta} = \mathbf{r},$$

where $\mathbf{R}$ is an $(r \times k)$ matrix of parameters with rank k and $\mathbf{r}$ is an $(r \times 1)$ vector of parameters. To illustrate how $\mathbf{R}$ and $\mathbf{r}$ are constructed, we consider the baseline case of the

CAPM model; we want to impose the restriction $\beta_1 = -\beta_2$ on the following specification:

$$\begin{aligned} \left(r_t^i - r_t^{rf}\right) &= \beta_{0,i} + \beta_{1,i}\left(r_t^m - r_t^{rf}\right) + u_{i,t}, \qquad (23) \\ \mathbf{R}\boldsymbol{\beta} &= \mathbf{r}, \\ \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_{0,i} \\ \beta_{1,i} \end{pmatrix} &= (0). \end{aligned}$$

The distribution of a known statistic under the null is derived by applying known results.

If $\left(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}\right) \sim \mathbf{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$, then:

$$\left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r} \mid \mathbf{X}\right) \sim \mathbf{N}\left(\mathbf{R}\boldsymbol{\beta} - \mathbf{r}, \sigma^2 \mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right). \qquad (24)$$

The test is constructed by deriving the distribution of (24) under the null $\mathbf{R}\boldsymbol{\beta} - \mathbf{r} = \mathbf{0}$.

Given that

$$\left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r} \mid \mathbf{X}\right) = \mathbf{R}\boldsymbol{\beta} - \mathbf{r} + \mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u},$$

under H_0, we have:

$$\begin{aligned} & \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)' \left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)^{-1} \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right) \\ = \; & \boldsymbol{\epsilon}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}' \left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)^{-1} \mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\ = \; & \boldsymbol{\epsilon}'\mathbf{P}\boldsymbol{\epsilon}. \end{aligned}$$

where $\mathbf{P}$ is a symmetric idempotent matrix of rank r, orthogonal to $\mathbf{M}$.

Then

$$\frac{\left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)' \left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)^{-1} \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)}{s^2} \sim \mathbf{rF}\left(r, T-k\right), \quad \text{under } H_0,$$

which can be used to test the relevant hypothesis.

Testing the significance of the subset of coefficients In the general framework to test linear restrictions we set $\mathbf{r} = \mathbf{0}$, $\mathbf{R} = \begin{bmatrix} I_r & 0 \end{bmatrix}$, and partition $\boldsymbol{\beta}$ in a corresponding way into $\begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 \end{bmatrix}$. In this case the restriction $\mathbf{R}\boldsymbol{\beta} - \mathbf{r} = \mathbf{0}$ is equivalent to $\boldsymbol{\beta}_1 = 0$ in the partitioned regression model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

in which partitioning creates two blocks of dimension r and $k - r$.

Before proceeding to the discussion of hypothesis testing, it is useful to derive the formula for the OLS estimator in the partitioned regression model. To obtain such results we partition the 'normal equations' $\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^{'}\mathbf{y}$ as:

$$\begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix} \mathbf{y},$$

or, equivalently,

$$\begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{pmatrix}. \tag{25}$$

System (25) can be resolved in two stages by first deriving an expression $\widehat{\boldsymbol{\beta}}_2$ as:

$$\widehat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1} \left(\mathbf{X}_2'\mathbf{y} - \mathbf{X}_2'\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1\right),$$

and then by substituting it in the first equation of (25) to obtain

$$\mathbf{X}_1'\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 + \mathbf{X}_1'\mathbf{X}_2 (\mathbf{X}_2'\mathbf{X}_2)^{-1} \left(\mathbf{X}_2'\mathbf{y} - \mathbf{X}_2'\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1\right) = \mathbf{X}_1'\mathbf{y},$$

from which:[3]

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_1 &= (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1} \mathbf{X}_1'\mathbf{M}_2\mathbf{y} \\ \mathbf{M}_2 &= \left(\mathbf{I} - \mathbf{X}_2 (\mathbf{X}_2'\mathbf{X}_2)^{-1} \mathbf{X}_2'\right). \end{aligned}$$

Note that, as $\mathbf{M}_2$ is idempotent, we can also write:

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{M}_2'\mathbf{M}_2\mathbf{X}_1)^{-1} \mathbf{X}_1'\mathbf{M}_2'\mathbf{M}_2\mathbf{y},$$

and $\widehat{\boldsymbol{\beta}}_1$can be interpreted as the vector of OLS coefficients of the regression of $\mathbf{y}$ on the matrix of residuals of the regression of $\mathbf{X}_1$ on $\mathbf{X}_2$. Thus, an OLS regression on two regressors is equivalent to two OLS regressions on a single regressor (Frisch-Waugh theorem).

Finally, consider the residuals of the partitioned model:

$$\begin{aligned} \widehat{\boldsymbol{\epsilon}} &= \mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2, \\ \widehat{\boldsymbol{\epsilon}} &= \mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}} - \mathbf{X}_2 (\mathbf{X}_2'\mathbf{X}_2)^{-1} \left(\mathbf{X}_2'\mathbf{y} - \mathbf{X}_2'\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1\right), \\ \widehat{\boldsymbol{\epsilon}} &= \mathbf{M}_2\mathbf{y} - \mathbf{M_2}\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 \\ &= \mathbf{M}_2\mathbf{y} - \mathbf{M_2}\mathbf{X}_1 (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1} \mathbf{X}_1'\mathbf{M}_2\mathbf{y} \\ &= \left(\mathbf{M}_2 - \mathbf{M_2}\mathbf{X}_1 (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1} \mathbf{X}_1'\mathbf{M}_2\right)\mathbf{y}, \end{aligned}$$

[3] Note that the expression for the estimator can be obtained by applying the formula of the partitioned inverse directly on the normal equations:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} E & -EBD^{-1} \\ -D^{-1}CE & D^{-1} + D^{-1}CEBD^{-1} \end{pmatrix}, E = (A - BD^{-1}C)^{-1}.$$

however, we already know that $\widehat{\boldsymbol{\epsilon}} = \mathbf{My}$, therefore,

$$\mathbf{M} = \left(\mathbf{M}_2 - \mathbf{M_2}\mathbf{X}_1 \left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{-1} \mathbf{X}_1'\mathbf{M}_2\right). \tag{26}$$

Now reconsider testing for our null of interest. Under H_0, $\mathbf{X}_1$ has no additional explicatory power for $\mathbf{y}$ with respect to $\mathbf{X}_2$**,** therefore:

$$H_0\text{: } \mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \qquad (\boldsymbol{\epsilon} \mid \mathbf{X}_1, \mathbf{X}_2) \sim N\left(0, \sigma^2 I\right).$$

Note that the statement

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\gamma}_2 + \boldsymbol{\epsilon}, \qquad (\boldsymbol{\epsilon} \mid \mathbf{X}_2) \sim N\left(0, \sigma^2 I\right),$$

is always true under our maintained hypotheses. However, in general $\boldsymbol{\gamma}_2 \neq \boldsymbol{\beta}_2$.To derive a statistic to test H_0 remember that the general matrix $\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'$ is the upper left block of $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$, which we can now write as $\left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{-1}$. The statistic then takes the form

$$\frac{\widehat{\boldsymbol{\beta}}_1' \left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)\widehat{\boldsymbol{\beta}}_1}{rs^2} = \frac{\mathbf{y}'\mathbf{M_2}\mathbf{X}_1 \left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y}}{\mathbf{y}'\mathbf{My}}\frac{T-k}{r} \sim F\left(T-k, r\right).$$

Given (26), (25) can be re-written as:

$$\frac{\mathbf{y}'\mathbf{M_2}\mathbf{y} - \mathbf{y}'\mathbf{My}}{\mathbf{y}'\mathbf{My}}\frac{T-k}{r} \sim F\left(T-k, r\right), \tag{27}$$

where the denominator is the sum of the squared residuals in the unconstrained model, while the numerator is the difference between the sum of residuals in the constrained model and the sum of residuals in the unconstrained model.

Consider the limit case $r = 1$ and β_1 is a scalar. The F-statistic takes the form

$$\frac{\widehat{\boldsymbol{\beta}}_1^2}{s^2\left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)} \sim F\left(T-k, r\right), \text{ under } H_0,$$

where $\left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{-1}$ is element $(1,1)$ of the matrix $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

Using the result on the relation between the F and the Student's t-distribution:

$$\frac{\widehat{\boldsymbol{\beta}}_1}{s\left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{1/2}} \sim t\left(T-k\right) \text{ under } H_0.$$

Therefore, an immediate test of significance of the coefficient can be performed, as it is in Table 1.1, by taking the ratio of each estimated coefficient and the associated standard error.

5.2.3. **The partial regression theorem**

The Frisch-Waugh Theorem described above is worth more consideration.

The theorem tells us than any given regression coefficient in the model $E(y \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ can be computed in two different but exactly equivalent ways: 1) by regressing y on all the columns of $\mathbf{X}$, 2) by first regressing the j-th column of $\mathbf{X}$ on all the other columns of $\mathbf{X}$, computing the residuals of this regression and then by regressing y on these residuals.

This result is relevant in that it clarifies that the relationships pinned down by the estimated parameters in a linear model do not describe the connections between the regressand and each regressor but the connection between the part of each regressor that is not explained by the other ones and the regressand.

This is importnat in the case the regression that a regression analysis is a first step in a "what if" analysis. The relevant question in this case becomes "how much shall y change if I change $\mathbf{X}_i$?"

The estimation of a single equation linear model does not allow to anser that question, for a number of reasons.

First, estimated parameters in a linear model can only answer the question how much shall $E(y \mid \mathbf{X})$ if I change $\mathbf{X}$? We have seen that the two questions are very different if the R^2 of the regression is low, in this case a change in $E(y \mid \mathbf{X})$ may not effect any visible and relevant effect on y.

Second, a regression model is a conditional expected value GIVEN $\mathbf{X}$. In this sense there is no space for "changing" the value of any element in $\mathbf{X}$. Any statement involving such a change requires some assumption on how the conditional expectation of y changes if $\mathbf{X}$ changes and a correct analysis of this requires an assumption on the joint distribution of y and $\mathbf{X}$. Simulation might require the use of the multivariate joint model even when valid estimation can be performed concentrating only on the conditional model. Strong exogeneity is stronger than weak exogeneity for the estimation of the parameters of interest.

Think of a linear model with know parameters

$$y = \beta_1 x_1 + \beta_2 x_2$$

What is in this model the effect of on y of changing x_1 by one unit while keeping x_2 constant ? Easy β_1.

Now think of the estimated linear model:

$$y = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{u}$$

Now y is different from $E(y \mid \mathbf{X})$ and the question "what is in this model the effect of on $E(y \mid \mathbf{X})$ of changing x_1 by one unit while keeping x_2 constant ?" does not in general make sense. Changing x_1 keeping x_2 implies that there is zero correlation among this variables. But the estimates $\hat{\beta}_1$and $\hat{\beta}_2$ are obained by using data in which in general there is some correlation between x_1 and x_2. Data in which fluctuations in x_1 do not have any effect on x_2 would have most likely generated different estimates from those obtained in the estimation sample. The only valid question quetion that can be answered using the coefficients in linear regression is "What is the effect on $E(y \mid \mathbf{X})$ of changing the part of each regressors that is orthogonal to the other ones". Only in the case regressors are orthogonal to each other the parameters estimates to describe the effect of changing the regressor of one unit.

Historically regression methods were simply not conceived with a "what if" analysis in mind, they were conceived to assess the overall regressive dependence of a y on a set $\mathbf{X}$, and, to a lesser extent, to evaluate the marginal contribution of each $x_i$to this overall dependence. "What if" analysis requires simulation and in most cases a low level ofredution thatn that used for regression analysis.

Relevance and Significance again: the semi-partial $\mathbf{R}^2$ OK, thedifferenc between significance of parameters and relevance of a regression is well understood. Is there something that we can do if we aare interested in measuring the "relevance" of single columns of $\mathbf{X}$, say by partitioning the "overall explaining power" (better: the variance of the regression function) in a regressor by regressor evaluation?

When the columns of $\mathbf{X}$ are orthogonal to eache other the total R^2 can be exactly decomposed in the sum of the partial R^2 due to each regressor x_i (the partial R^2 of a regressor i is defined as the R^2 of the regression of y on x_i).

This is in general not the case in applications with non experimental data: columns of $\mathbf{X}$ are correlated and a (often large) part of the overall R^2 does depend on the joint behaviour of the columns of $\mathbf{X}$. However, it is always possible to compute the marginal contribution to the overall R^2 due to each regressor x_i, defined as the difference between the overall R^2 and the R^2 ot the regression that inlcudes all columns $\mathbf{X}$ except x_i. This is called the semi-partial R^2.

Interestingly, the the semi-partial R^2 is a simple tranformation of the t-ratio:

$$spR_i^2 = \frac{t_{\beta_i}^2 (1 - R^2)}{(T - k)}$$

This result has two interesting implications.

First, a quantity which we considered as just a measure of statistical reliability, can lead to a measure of relevance when combined with the overall R^2 of the regression.

Second, we can re-iterate the difference between statistical significance and relevance. Suppose you have a sample size of 10000 and you have 10 columns in $\mathbf{X}$ and the t-ratio on a coefficient β_i is of about 4 with an associate P-value of the order .01: "very" statistical significant! The derivation of the semi-partial R^2 tells us that the contribution of this variable to the overall R2 is at most approximately 16/(10000-10) that is: less than two thousands.

To put it differently, in the case here described we need a t-ratio on β_i of the order of eleven to have a marginal contribution of the regressor to the overall R^2 of the 10%.

5.3. *The effects of mis-specification*

The third important element to consider when interpeting regression results are the consequences of adopting a "wrong" model. Each specification can be interpreted of the result of a reduction process, what happens if the reduction process that has generated $E(y \mid \mathbf{X})$ omits some relevant information. We shall consider three general cases of mis-specification . We take first the case of under-parameterization (the estimated model omits variables included in the DGP) to move on to over-parameterization (the estimated model includes more variables than the DGP). Finally, we consider mis-specification deriving from ignoring the existence of constraints on an estimated parameter.

5.3.1. **Under-parameterization**

Given the DGP:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{28}$$

for which hypotheses $(7) - (17)$ hold, the following model is estimated:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\nu}. \tag{29}$$

The OLS estimates are given by the following expression:

$$\widehat{\boldsymbol{\beta}}_1^{up} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}, \tag{30}$$

while the OLS estimates which are obtained by estimation of the DGP, are:

$$\widehat{\boldsymbol{\beta}}_1 = \left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y}. \tag{31}$$

The estimates in (31) are best linear unbiased estimators (BLUE) by construction, while the estimates in (30) are biased unless $\mathbf{X}_1$and $\mathbf{X}_2$ are uncorrelated. To show this, consider:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_1 &= \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1'\mathbf{y} - \mathbf{X}_1'\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2\right) && (32)\\ &= \widehat{\boldsymbol{\beta}}_1^{up} + \widehat{\mathbf{D}}\widehat{\boldsymbol{\beta}}_2, && (33)\end{aligned}$$

where $\widehat{\mathbf{D}}$ is the vector of coefficients in the regression of $\mathbf{X}_2$ on $\mathbf{X}_1$ and $\widehat{\boldsymbol{\beta}}_2$ is the OLS estimator obtained by fitting the DGP.

To provide further interpretation of these results, note that if

$$\begin{aligned} E\left(\mathbf{y} \mid \mathbf{X}_1, \mathbf{X}_2\right) &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \\ E\left(\mathbf{X}_1 \mid \mathbf{X}_2\right) &= \mathbf{X}_1\mathbf{D}, \end{aligned}$$

then,

$$E\left(\mathbf{y} \mid \mathbf{X}_1\right) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X_1}\mathbf{D}\boldsymbol{\beta}_2 = \mathbf{X}_1\boldsymbol{\alpha}.$$

Therefore the OLS estimator in the under-parameterized model is a biased estimator of $\boldsymbol{\beta}_1$, but an unbiased estimator of $\boldsymbol{\alpha}$. Then, if the objective of the model is forecasting and $\mathbf{X}_1$ is more easily observed than $\mathbf{X}_2$, the under-parameterized model can be safely used. On the other hand, if the objective of the model is to test specific predictions on parameters, the use of the under-parameterized model delivers biased results. When we are interested in the effect of $\mathbf{X}_1$ on $\mathbf{y}$, independently from other factors, it is crucial to control the effects of omitted variables.

5.3.2. **Over-parameterization**

Given the DGP,

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}, \tag{34}$$

for which hypotheses $(7) - (17)$ hold, the following model is estimated:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{v}. \tag{35}$$

The OLS estimator of the over-parameterized model is

$$\widehat{\boldsymbol{\beta}}_1^{op} = \left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y}, \tag{36}$$

while, by estimating the DGP, we obtain:

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}. \tag{37}$$

By substituting $\mathbf{y}$ from the DGP, one finds that both estimators are unbiased and the difference is now made by the variance. In fact we have:

$$var\left(\widehat{\boldsymbol{\beta}}_1^{op} \mid \mathbf{X}_1, \mathbf{X}_2\right) = \sigma^2 (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}, \tag{38}$$

$$var\left(\widehat{\boldsymbol{\beta}}_1 \mid \mathbf{X}_1, \mathbf{X}_2\right) = \sigma^2 (\mathbf{X}_1'\mathbf{X}_1)^{-1}. \tag{39}$$

One can show that the estimator derived from the correct model is more efficient. The difference between the two variance-covariance matrices is a positive semidefinite matrix. To show this, remember that if two matrices $\mathbf{A}$ and $\mathbf{B}$ are positive definite and $\mathbf{A}-\mathbf{B}$ is positive semidefinite, then also the matrix $\mathbf{B}^{-1}-\mathbf{A}^{-1}$ is positive semidefinite. We have to show that $\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1$ is a positive semidefinite matrix. Such a result is almost immediate:

$$\begin{aligned}\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1 &= \mathbf{X}_1'(\mathbf{I}-\mathbf{M}_2)\mathbf{X}_1 \\ &= \mathbf{X}_1'\mathbf{Q}_2\mathbf{X}_1 = \mathbf{X}_1'\mathbf{Q}_2\mathbf{Q}_2\mathbf{X}_1.\end{aligned}$$

We conclude that over-parameterization impacts on the efficiency of estimators and the power of the tests of hypotheses.

5.3.3. **Estimation under linear constraints**

In this section we analyse the impact on the OLS estimator of a mis-specification deriving from ignoring the existence of constraints on an estimated parameter. To analyse the mis-specification, we introduce the difference between the estimated model and the data generating process (DGP).

The estimated model is the linear model analysed up to now:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

while the DGP is instead:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad subject\ to\ \mathbf{R}\boldsymbol{\beta} - \mathbf{r} = \mathbf{0},$$

where the constraints are expressed using the so called implicit form. A useful alternative way of expressing constraints, known as the 'explicit form' has been expressed by Sargan (1988):

$$\boldsymbol{\beta} = \mathbf{S}\boldsymbol{\theta} + \mathbf{s},$$

where $\mathbf{S}$ is a $(k \times (k-r))$ matrix of rank $k-r$ and $\mathbf{s}$ is a $k \times 1$ vector.

To show how constraints are specified in the two alternatives let us consider the case of $\beta_1 = -\beta_2$ on the following specification:

$$\ln y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i. \tag{40}$$

Using $\mathbf{R}\boldsymbol{\beta} - \mathbf{r} = \mathbf{0}$:

$$\begin{pmatrix} 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = (0)\,,$$

while using $\boldsymbol{\beta} = \mathbf{S}\boldsymbol{\theta} + \mathbf{s}$:

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

In practice the constraints in the explicit form are written by considering $\boldsymbol{\theta}$ as the vector of free parameters. Note that there is no unique way of expressing constraints in the explicit form, in our case the same constraint can be imposed as:

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

As the two alternatives are indifferent, $\mathbf{R}\boldsymbol{\beta} - \mathbf{r} = \mathbf{0}$ and $\mathbf{RS}\boldsymbol{\theta} + \mathbf{Rs} - \mathbf{r} = \mathbf{0}$ are equivalent**,** which implies:

1. $\mathbf{RS} = \mathbf{0}$**;**
2. $\mathbf{Rs} - \mathbf{r} = \mathbf{0}$**.**

We use the explicit form of imposing constraints to derive the restricted least squares (RLS) estimators, and to evaluate the consistency and relative efficiency of OLS and RLS.

The restricted least squares (RLS) estimator To construct RLS, substitute the constraint in the original model to obtain:

$$\mathbf{y} - \mathbf{Xs} = \mathbf{XS}\boldsymbol{\theta} + \boldsymbol{\epsilon}. \tag{41}$$

Equation (41) is equivalent to:

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{42}$$

where $\mathbf{y}^* = \mathbf{y} - \mathbf{X}\mathbf{s}$, $\mathbf{X}^* = \mathbf{X}\mathbf{S}$.

Note that the transformed model features the same residuals with the original model; therefore, if hypotheses $(7) - (17)$ hold for the original model, they also hold for the transformed. We apply OLS to the transformed model to obtain:

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{y}^* \\ &= \left(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S}\right)^{-1}\mathbf{S}'\mathbf{X}'\left(\mathbf{y} - \mathbf{X}\mathbf{s}\right). \end{aligned} \tag{43}$$

From (43) the RLS estimation is easily obtained by applying the transformation $\widehat{\boldsymbol{\beta}}^{rls} = \mathbf{S}\widehat{\boldsymbol{\theta}} + \mathbf{s}$. Similarly, the variance of the RLS estimator is easily obtained as:

$$\begin{aligned} var\left(\widehat{\boldsymbol{\theta}} \mid \mathbf{X}\right) &= \sigma^2\left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1} = \sigma^2\left(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S}\right)^{-1}, \\ var\left(\widehat{\boldsymbol{\beta}}^{rls} \mid \mathbf{X}\right) &= var\left(\mathbf{S}\widehat{\boldsymbol{\theta}} + \mathbf{s} \mid \mathbf{X}\right) \\ &= \mathbf{S}\, var\left(\widehat{\boldsymbol{\theta}} \mid \mathbf{X}\right)\mathbf{S}' \\ &= \sigma^2\mathbf{S}\left(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S}\right)^{-1}\mathbf{S}'. \end{aligned}$$

We can now discuss the properties of OLS and RLS in the case of a DGP with constraints.

Unbiasedness

Under the assumed DGP, both estimators are unbiased, since such properties depend on the validity of hypotheses $(7) - (17)$, which is not affected by the imposition of constraints on parameters.

Efficiency

Obviously, if we interpret RLS as the OLS estimator on the transformed model (43) we immediately derive the results that the RLS is the most efficient estimator, as the hypotheses for the validity of the Gauss Markov theorem are satisfied when OLS is applied to (43). Note that by posing $\mathbf{L} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ in the context of the transformed model, we do not generally obtain OLS but an estimator whose conditional variance with respect to $\mathbf{X}$, coincides with the conditional variance of the OLS estimator.

We support this intuition with a formal argument by showing that the difference between the variance of the OLS estimator and the variance of the RLS estimator is a positive semidefinite matrix.

$$var\left(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}\right) - var\left(\widehat{\boldsymbol{\beta}}^{rls} \mid \mathbf{X}\right) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} - \sigma^2\mathbf{S}\left(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S}\right)^{-1}\mathbf{S}'.$$

Define $\mathbf{A}$ as:

$$\mathbf{A} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} - \mathbf{S}\left(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S}\right)^{-1}\mathbf{S}'.$$

Given that

$$\begin{aligned}
\mathbf{AX'XA} &= \left((\mathbf{X'X})^{-1} - \mathbf{S}\left(\mathbf{S'X'XS}\right)^{-1}\mathbf{S'}\right)\mathbf{X'X}\left((\mathbf{X'X})^{-1} - \mathbf{S}\left(\mathbf{S'X'XS}\right)^{-1}\mathbf{S'}\right) \\
&= (\mathbf{X'X})^{-1} - 2\mathbf{S}\left(\mathbf{S'X'XS}\right)^{-1}\mathbf{S'} + \mathbf{S}\left(\mathbf{S'X'XS}\right)^{-1}\mathbf{S'S}\left(\mathbf{S'X'XS}\right)^{-1}\mathbf{S'} \\
&= (\mathbf{X'X})^{-1} - \mathbf{S}\left(\mathbf{S'X'XS}\right)^{-1}\mathbf{S'} \\
&= \mathbf{A},
\end{aligned}$$

A is positive semidefinite, being the product of a matrix and its transpose.

The OLS estimator ignores available information and therefore is less efficient than the RLS estimator. However, there is no difference between the two estimators in terms of unbiasedness.

So far we have evaluated the gains of imposing true restrictions. A related interesting exercise is the evaluation of the loss from imposing false restrictions.

6. **Econometrics in action: From the CAPM to Fama and French Factors**

Let us run some CAPM regression on real data. Table 3.2 reports the results of running the CAPM regression on returns from portfolios 15 and 51 in the FF data over a sample on monthly observations from 1962:1-2014:6.

TABLE 3.2: The Estimation of the CAPM on the FF data

Dependent Variable $\left(r_t^{15} - r_t^{rf}\right)$				
Variable	Coefficient	Std. Error	t-ratio	Prob.
C	0.009	0.0018	5.13	0.0000
$\left(r_t^m - r_t^{rf}\right)$	1.016	0.038	26.11	0.0000
R^2 0.52	S.E. of reg 0.0438	S.E. dep.var 0.063	1962:1-2014:6	
F-statistic C=0, F(1,628)=26.33(0.0000)				
Dependent Variable $\left(r_t^{51} - r_t^{rf}\right)$				
Variable	Coefficient	Std. Error	t-ratio	Prob.
C	-0.00096	0.0006	-1.52	0.13
$\left(r_t^m - r_t^{rf}\right)$	1.110027	0.0139	79.92	0.0000
R^2 0.91	S.E. of reg 0.0156	S.E. dep.var 0.052	1962:1-2014:6	
F-statistic C=0, F(1,628)=2.328(0.13)				

The results of the regressions show a β_1 significantly different from zero but not significantly different from 1 and a significantly positive β_0 for portfolio 15 and a β_1 significantly

different from zero and from 1 and a not-significantly negative β_0 for portfolio 51. Note that the test of significance on β_0' is strictly speaking a test ofthe CAPM, and it can be equivalently conducted using the t-statistic or constructing the F-statistic for restriction on the relevant set of coefficients (Table 3.1 reports both tests).

To assess the poitential effect of omitted variables we consider augmenting the CAPM regression with the Fama-French Factors, SMB and HML, and the momentum factor MOM.

TABLE 3.3: The Estimation of the CAPM on the FF data

Dependent Variable $\left(r_t^{15} - r_t^{rf}\right)$				
Variable	Coefficient	Std. Error	t-ratio	Prob.
C	0.006	0.0010	5.83	0.0000
$\left(r_t^m - r_t^{rf}\right)$	0.857	0.024	35.02	0.0000
HML	0.568	0.038	15.02	0.0000
SMB	1.154	0.034	33.61	0.0000
MOM	-0.166	0.024	-6.86	0.0000
R^2 0.85	S.E. of reg 0.0249	S.E. dep.var 0.063	1962:1-2014:6	
F-statistic $\beta_2 = \beta_3 = \beta_4 = 0$, F(3,625)=439.02(0.0000)				
Dependent Variable $\left(r_t^{51} - r_t^{rf}\right)$				
Variable	Coefficient	Std. Error	t-ratio	Prob.
C	0.002	0.00048	4.18	0.0000
$\left(r_t^m - r_t^{rf}\right)$	1.044	0.0112	92.64	0.0000
HML	-0.378	0.017	-21.73	0.0000
SMB	-0.107	0.0158	-6.765	0.0000
MOM	-0.129	0.011	-11.565	0.0000
R^2 0.91	S.E. of reg 0.0156	S.E. dep.var 0.052	1962:1-2014:6	
F-statistic $\beta_2 = \beta_3 = \beta_4 = 0$, F(3,625)=179.85(0.0000)				

All the three added factors are strongly significant in all regressions showing that the CAPM equations suffer from omitted variables problem. Note also that the coefficients on the market protfolio excess returns change when the extended specification is adopted. We have therefore evidence that the augmenting factors are not orthogonal to excess returns on market portfolios.

6.1. *Fama-French Factors and the Fama-MacBeth procedure*

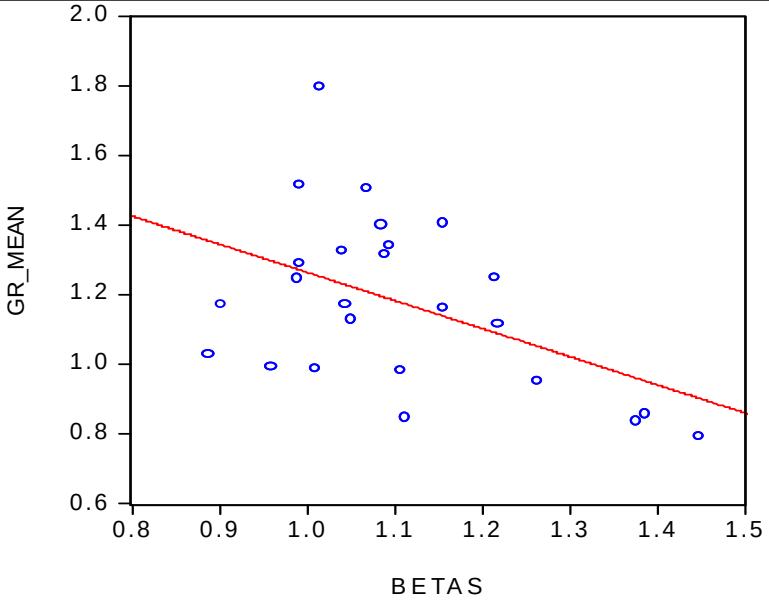
Alternative evidence on the validity of the CAPM can be provided by analyzing the cross-section of returns (Fama-French(1992,1993), FamaMacBeth(1973)). To illustrate how this can be done consider the 25 portfolios and run for each of them the CAPM regression over the sample 1962:1 2014:6. These regressions deliver 25 betas. Take now a second-step regression in which the cross-section of the average (over the sample 1962:1-2014:6) monthly returns on the 25 portfolios are projected on the 25 betas:

$$r_i = \gamma_0 + \gamma_1 \beta_i + u_i$$

Under the null of the CAPM i) residuals should be randomly distributed around the regression line, ii) $\gamma_0 = E\left(r^f\right), \gamma_1 = E\left(r^m - r^f\right)$

TABLE 3.4: CAPM in the cross-section of 25 portfolios

Dependent Variable r_i $(i = 1, ...25)$					
Variable	Coefficient	Std. Error	t-ratio	HAC Std.Err.	Prob.
C	2.07	0.35	5.94	0.451	0.3069
β_i	-0.80	0.31	-2.57	0.364	0.0169
R^2 0.22	S.E. of reg 0.22	S.E. dep.var 0.25	1926:7-2014:6		
$E\left(r^f\right) = 0.40$, $E\left(r^m - r^f\right) = 0.89$					



The cross-sectional regression strongly rejects the CAPM. Note however that this regression is affected by an inference problem caused by the correlation of residuals in the cross-section regression. Fama and MacBeth (1973) address this problem by estimating

month-by-month cross-section regressions of monthly returns on the betas obtained on the full sample. The time series means of the monthly slopes and intercepts, along with the standard errors of the means, are then used to test whether the average premium for beta is positive and whether the average return on assets uncorrelated with the market is equal to the average riskfree interest rate. In this approach, the standard errors of the average intercept and slope are determined by the month-to-month variation in the regression coefficients, which fully captures the effects of residual correlation on variation in the regression coefficients. The application of the Fama-MacBeth on the sample 1962:1 2014:6 delivers the following results

TABLE 3.5: Statistics on the distribution of coefficients from Fama-MacBeth

	γ_0	γ_1
Mean	2.07	-0.807
St.Dev	9.56	10.06
Obs	630	630
t-stat	5.437	-2.01

An alternative route would be to construct Heteroscedasticity and Autocorrelation Consistent(HAC) estimators. We illustrate how do this in the following section, Table 3.4 includes a colum containing the HAC estimators of the standard errors of the regression parameters,

7. **Heteroscedasticity, Autocorrelation, and the GLS estimator**

In the second step of the Fama-French Regression, the dependent variable is the cross-section of returns on different portfolios. When these returns are projected on the factor loadings it is very unlikely that the variance-covariance matrix of residuals is diagonal. Let us reconsider the single equation model and generalize it to the case in which the hypotheses of diagonality and constancy of the conditional variances-covariance matrix of the residuals do not hold:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \\ \boldsymbol{\epsilon} &\sim n.d.\left(\mathbf{0}, \boldsymbol{\sigma}^2\boldsymbol{\Omega}\right), \end{aligned} \tag{44}$$

where the vector $\mathbf{y}$ contains T observations on the dependent variables, $\mathbf{X}$ contains $(T \times K)$ observations on the K explanatory variables exogenous for the estimation of $(K \times 1)$ the vector $\boldsymbol{\beta}$, and $\boldsymbol{\Omega}$ is a $(T \times T)$ symmetric and positive definite matrix. When the OLS method is applied to model (44), it delivers estimators which are consistent but not

efficient; moreover, the traditional formula for the variance-covariance matrix of the OLS estimators, $\boldsymbol{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}$, is wrong and leads to an incorrect inference. Using the standard algebra, it can be shown that the correct formula for the variance-covariance matrix of the OLS estimator is:

$$\boldsymbol{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\Omega}\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}.$$

To find a general solution to this problem, remember that the inverse of a symmetric definite positive matrix is also symmetric and definite positive and that for a given matrix $\boldsymbol{\Omega}$, symmetric and definite positive, there always exists a $(T \times T)$ non-singular matrix $\mathbf{K}$, such that $\mathbf{K}'\mathbf{K} = \boldsymbol{\Omega}^{-1}$ and $\mathbf{K}\boldsymbol{\Omega}\mathbf{K}' = \mathbf{I}_T$.

To find the solution, consider the regression model obtained by pre-multiplying both the right-hand and the left-hand sides of (44) by $\mathbf{K}$:

$$\begin{aligned} \mathbf{K}\mathbf{y} &= \mathbf{K}\mathbf{X}\boldsymbol{\beta} + \mathbf{K}\boldsymbol{\epsilon}, \\ \mathbf{K}\boldsymbol{\epsilon} &\sim n.d.\left(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I}_T\right). \end{aligned} \tag{45}$$

The OLS estimator of the parameters of the transformed model (45) satisfies all the conditions for the applications of the Gauss−Markov theorem; therefore, the estimator

$$\begin{aligned} \widehat{\beta}_{GLS} &= (\mathbf{X}'\mathbf{K}'\mathbf{K}\mathbf{X})^{-1} \mathbf{X}'\mathbf{K}'\mathbf{K}\mathbf{y} \\ &= \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}, \end{aligned}$$

known as the generalised least squares (GLS) estimator, is BLUE. The variance of the GLS estimator, conditional upon $\mathbf{X}$, becomes

$$Var\left(\widehat{\beta}_{GLS} \mid \mathbf{X}\right) = \Sigma_\beta = \sigma^2 \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}.$$

Note that, from the application of the Gauss−Markov theorem, it follows immediately that the variance of the GLS estimator is equal to the sum of the variance of any other linear estimator and a positive semidefinite matrix. Consider, for example, the variances of the OLS and the GLS estimators. Using the fact that if $\mathbf{A}$ and $\mathbf{B}$ are positive definite and $\mathbf{A} - \mathbf{B}$ is positive semidefinite, then $\mathbf{B}^{-1} - \mathbf{A}^{-1}$ is also positive semidefinite, we have:

$$\begin{aligned} & \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right) - (\mathbf{X}'\mathbf{X}) (\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1} (\mathbf{X}'\mathbf{X}) \\ = \; & \mathbf{X}'\mathbf{K}'\mathbf{K}\mathbf{X} - (\mathbf{X}'\mathbf{X}) \left(\mathbf{X}'\mathbf{K}^{-1} (\mathbf{K}')^{-1} \mathbf{X}\right)^{-1} (\mathbf{X}'\mathbf{X}) \\ = \; & \mathbf{X}'\mathbf{K}' \left(\mathbf{I} - (\mathbf{K}')^{-1} \mathbf{X} \left(\mathbf{X}'\mathbf{K}^{-1} (\mathbf{K}')^{-1} \mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{K}^{-1}\right) \mathbf{K}\mathbf{X} \\ = \; & \mathbf{X}'\mathbf{K}'\mathbf{M}_W'\mathbf{M}_W\mathbf{K}\mathbf{X}, \end{aligned}$$

where

$$\mathbf{M}_W = \left(\mathbf{I} - \mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\right), \tag{46}$$

$$\mathbf{W} = \left(\mathbf{K}'\right)^{-1}\mathbf{X}. \tag{47}$$

The applicability of the GLS estimator requires an empirical specification for the matrix $\mathbf{K}$. We consider here three specific applications where the appropriate choice of such a matrix leads to the solution of the problems in the OLS estimator generated, respectively, by the presence of first-order serial correlation in the residuals, by the presence of heteroscedasticity in the residuals and by the preseece of both of them.

7.1. *Correction for Serial Correlation (Cochrane-Orcutt)*

Consider first the case of first-order serial correlation in the residuals. We have the following model:

$$\begin{aligned} y_t &= \mathbf{x}_t'\boldsymbol{\beta}+u_t, \\ u_t &= \rho u_{t-1}+\epsilon_t, \\ \epsilon_t &\sim n.i.d.\left(0,\sigma_\epsilon^2\right), \end{aligned}$$

which, using our general notation, can be re-written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}, \tag{48}$$

$$\boldsymbol{\epsilon} \sim n.d.\left(\mathbf{0},\boldsymbol{\sigma}^2\boldsymbol{\Omega}\right),$$

$$\boldsymbol{\sigma}^2 = \frac{\sigma_\epsilon^2}{1-\rho^2}, \tag{49}$$

$$\boldsymbol{\Omega} = \begin{bmatrix} 1 & \rho & \rho^2 & . & . & \rho^{T-1} \\ \rho & 1 & \rho & . & . & \rho^{T-2} \\ \rho^2 & . & 1 & . & . & . \\ . & . & . & . & . & . \\ \rho^{T-2} & . & . & \rho & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & . & . & \rho & 1 \end{bmatrix}.$$

In this case, the knowledge of the parameter ρ allows the empirical implementation of the GLS estimator. An intuitive procedure to implement the GLS estimator can then be the following:

1. estimate the vector $\boldsymbol{\beta}$ by OLS and save the vector of residuals $\widehat{u}_t$;

2. regress $\widehat{u}_t$ on $\widehat{u}_{t-1}$ to obtain an estimate $\widehat{\rho}$ of ρ;

3. construct the transformed model and regress $(y_t - \widehat{\rho} y_{t-1})$ on $(\mathbf{x}_t - \widehat{\rho}\mathbf{x}_{t-1})$ to obtain the GLS estimator of the vector of parameters of interest.

Note that the above procedure, known as the Cochrane−Orcutt procedure, can be repeated until convergence.

7.2. *Correction for Heteroscedasticity (White)*

In the case of heteroscedasticity, our general model becomes

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \\ \boldsymbol{\epsilon} &\sim n.d.\left(\mathbf{0}, \boldsymbol{\Omega}\right), \\ \boldsymbol{\Omega} &= \begin{bmatrix} \sigma_1^2 & 0 & 0 & . & . & 0 \\ 0 & \sigma_2^2 & 0 & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & . & . & 0 & \sigma_{T-1}^2 & 0 \\ 0 & 0 & . & . & 0 & \sigma_T^2 \end{bmatrix}. \end{aligned}$$

In this case, to construct the GLS estimator, we need to model heteroscedasticity choosing appropriately the $\mathbf{K}$ matrix. White (1980) proposes a specification based on the consideration that in the case of heteroscedasticity the variance-covariance matrix of the OLS estimator takes the form:

$$\boldsymbol{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\boldsymbol{\Omega}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

which can be used for inference, once an estimator for $\boldsymbol{\Omega}$ is available. The following unbiased estimator of $\boldsymbol{\Omega}$ is proposed:

$$\widehat{\boldsymbol{\Omega}} = \begin{bmatrix} \widehat{u}_1^2 & 0 & 0 & . & . & 0 \\ 0 & \widehat{u}_2^2 & 0 & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & . & . & 0 & \widehat{u}_{T-1}^2 & 0 \\ 0 & 0 & . & . & 0 & \widehat{u}_T^2 \end{bmatrix}.$$

This choice for $\widehat{\boldsymbol{\Omega}}$ leads to the following degrees of freedom corrected heteroscedasticity consistent parameters' covariance ince matrix estimator:

$$\Sigma_{\beta}^{W} = \frac{T}{T-k}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{t=1}^{T}\widehat{u}_t^2\mathbf{X}_t\mathbf{X}_t'\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

This estimator corrects for the OLS for the presence of heteroscedasticity in the residuals without modelling in that, later in the book we shall consider alternative models for heteroscedasticity, known as ARCH (autoregressive conditional heteroscedasticity) processes, useful for high-frequency financial series, and based upon simultaneous modelling of the first two moments of time-series processes. These models have been been proposed by Engle (1980) and Bollerslev (1986).

7.3. *Correction for heteroscedasticity and serial correlation (Newey-West)*

The White covariance matrix assumes serially uncorrelated residuals. Newey and West(1987) have proposed a more general covariance estimator that is robust to heteroscedasticity and autocorrelation of the residuals of unknow form. This HAC (heteroscedasticity and autocorrelation consistent) coefficient covariance estimators is given by:

$$\Sigma_{\beta}^{NW} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} T\hat{\Omega}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

where $\hat{\Omega}$ is a long-run covariance estimators

$$\begin{aligned}\widehat{\Omega} &= \widehat{\Gamma}(0) + \sum_{j=1}^{p}\left[1-\frac{j}{p+1}\right]\left[\widehat{\Gamma}(j)+\widehat{\Gamma}(-j)\right], \qquad (50)\\ \widehat{\Gamma}(j) &= \left(\sum_{t=1}^{T}\widehat{u}_t\widehat{u}_{t-j}\mathbf{X}_t\mathbf{X}_{t-j}'\right)\frac{1}{T}\end{aligned}$$

note that in absence of serial correlation $\widehat{\Omega} = \widehat{\Gamma}(0)$ and we are back to the White Estimator. Implementation of the estimator requires a choice of p, which is the maximum lagat whcih correlation is still present. The weighting scheme adopted guarantees a positive definite estimated covaraince matrix by multiplying the sequence of the $\widehat{\Gamma}(j)$'s by a sequence of weights that decreases as $|j|$ increases.

8. **References**

Bollerslev, T. (1986). 'Generalized Autoregressive Conditional Heteroscedasticity'. *Journal of Econometrics*, 31: 307-327.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation.

Engle et al., (1983) R.F. Engle, D.F. Hendry, and J.F. Richard. Exogeneity. Econometrica, 51:277-304

Fama, E.F. and K.R. French(1992) "The Cross-Section of Expected Stock Returns", Journal of Finance, 47:2, 427-465

Fama, E.F. and K.R. French(1993) "Common Risk Factors in the Returns on Stock and Bonds" Journal of Financial Economics 33:1, 3-56

Fama, E.F. and J.D. MacBeth(1973) "Risk, Return and Equilibrium: Empirical Tests" Journal of Political Economy. 81:3, 607-636

Hendry, D. F. (1996). *Dynamic Econometrics*. Oxford: Oxford University Press.

Jensen, Michael C. 1968. "The Performance of Mutual Funds in the Period 1945-1964." Journal of Finance. 23:2, pp. 389-416.

Klein, L. (1983). *Lectures in Econometrics*. Amsterdam: North-Holland.

Newey, Whitney K; West, Kenneth D (1987). "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix". Econometrica 55 (3): 703–708.

White, A. (1980). 'A heteroscedastic consistent covariance matrix estimator and a direct test for heteroscedasticity'. *Econometrica*, 48: 817-838.