# Linear Models

Carlo Favero

# Econometric Modelling of Financial Returns

- 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.

# Non experimental data

- What do we do with non-experimental data ?
- When you try and explain returns you do not control how the data are generated
- Think of estimating a CAPM model

# Non experimental data: CAPM

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

- you have a time-series on excess returns of a given stock on the safe asset and on the market on the safe asset (you cannot control these data)
- you specify a linear model in which the excess returns on asset i are function of the excess returns on the market, and the excess returns on the market follow a CER.
- you introduce hypothesis on the error terms, to estimate parameters, to implement tests, and to use the model for

# Econometric Modelling of Financial Returns

- 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 $\mathbf{z}_t$.

# Econometric Modelling of Financial Returns

- the data

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

- a general multivariate model

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

- decomposing a multivariate into conditional and marginal

$$D\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}_1\right) D\left(\mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta}_2\right)$$

- a general linear univariate conditional model

$$\begin{aligned} y_t &= \boldsymbol{\beta}' \mathbf{z}_t + u_{1t} \\ \mathbf{z}_t &= \boldsymbol{\beta}_2 \mathbf{z}_{t-1} + u_{2t} \end{aligned}$$

# Econometric Modelling of Financial Returns

- There are many ways in which the CAPM can go wrong:
  - other factors beyond the market are relevant in determining excess returns on asset i
  - the excess returns on the market do depend on excess returns on asset i
  - the model is non-linear
  - the residuals are non-normal and their variance-covariance matrix does not fit the assumptions made

# Reduction Process

- Conditional densities are 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 $\mathbf{y}_t$.

# Reduction Process

- To understand the reduction process partiton the set of all variables into three types of variables:

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

  $\mathbf{w}_t$ identifies variables which are ignored in the specification of the econometric model.
- Exclusion 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

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

This is the statistical 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 $E(\mathbf{y}_t, \mathbf{z}_t \mid \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}, \boldsymbol{\beta})$ and by deriving a vector of innovations from the difference between the vector of observed variables and the vector of their means.

# The CAPM Reduced Form

In the case of the CAPM the general specification of the reduced form is the following one:

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

# From the Statistical Model to the Conditional Model: the CAPM

Statistical model

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

Estimated Equation

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

if $\sigma_{im} = 0$,then the estimated equation is a valid approximation to the statistical model for the estimation of $\beta_i$

# OLS

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

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

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

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

# OLS

- 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

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

# OLS

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

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

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

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

# 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, $\mathbf{x}$,and its mean vector $E(\mathbf{x})$

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

# Properties of the OLS estimates

We define the mean matrix of outer products $E\left(\mathbf{x}\mathbf{x}'\right)$ as:

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

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

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

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

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

# The first hypothesis

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

$$E(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}. \tag{1}$$

# The first hypothesis

Note that hypothesis $(1)$ is very demanding. It implies that

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

The conditional mean is, in general, a non-linear function of $(\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_n)$ and $(1)$ requires that such a function is a constant of zero. Note that $(1)$ 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 $\left(E\left(x_{jk}\varepsilon_i\right) = 0 \text{ for all } j \neq k\right)$. This statement is proved by using the properties of conditional expectations.

# The first hypothesis

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

$$E(\varepsilon_i \mid x_{jk}) = E\left[E(\varepsilon_i \mid \mathbf{x}) \mid x_{jk}\right] = 0. \tag{2}$$

Then

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

# The second and third hypothesis

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

$$E\left(\boldsymbol{\epsilon}'\boldsymbol{\epsilon} \mid \mathbf{X}\right) = \sigma^2 I, \tag{3}$$

where $\sigma^2$ is a constant independent from $\mathbf{X}$. In the case of our data, this is a strong assumption unlikely to be met in practice.
The third hypothesis is the one already introduced, which guarantees that the OLS estimator can be derived:

$$rank\left(\mathbf{X}\right) = k. \tag{4}$$

Under hypotheses $(1) - (4)$ 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 $\beta$:

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

by hypothesis (1).

# Property 2: variance of OLS

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

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

# 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:

$$E\left(\boldsymbol{\beta}_L \mid \mathbf{X}\right) = E\left(\mathbf{L}\mathbf{X}\boldsymbol{\beta} + \mathbf{L}\boldsymbol{\varepsilon} \mid \mathbf{X}\right) = \mathbf{L}\mathbf{X}\boldsymbol{\beta},$$

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 $\mathbf{L} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$. The variance of the general estimator in the class of linear unbiased estimators is readily obtained as:

$$var\left(\boldsymbol{\beta}_L \mid \mathbf{X}\right) = E\left(\mathbf{L}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{L}' \mid \mathbf{X}\right) = \sigma^2\mathbf{L}\mathbf{L}'.$$

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.

# Residual Analysis

Consider the following representation:

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

where $\mathbf{M} = \mathbf{I}_n - \mathbf{Q}$, and $\mathbf{Q} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$. 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. $\mathbf{MX} = \mathbf{0}, \mathbf{MQ} = \mathbf{0}, \mathbf{QX} = \mathbf{X}$.

# Residual Analysis

Note that the OLS projection for $\mathbf{y}$ can be written as $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{Q}\mathbf{y}$, and that $\widehat{\boldsymbol{\epsilon}} = \mathbf{M}\mathbf{y}$, from which we have the known result of orthogonality between the OLS residuals and regressors. We also have $\mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\epsilon} = \mathbf{M}\boldsymbol{\epsilon}$, given that $\mathbf{M}\mathbf{X} = \mathbf{0}$. 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 $\sigma^2$.

# Residual Analysis

To derive an estimate of $\sigma^2$ from $S\left(\widehat{\beta}\right)$, 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.

# Residual Analysis

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 $\mathbf{X}$:

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

# Residual Analysis

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 $\sigma^2$ is given by $s^2 = S\left(\widehat{\boldsymbol{\beta}}\right) / (n-k)$.

# Residual Analysis:the R-squared

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.