Summer Seminar: Philosophy of Statistics
Lecture Notes 10: Linear Regression Model: a brief introduction

Aris Spanos [SUMMER 2019]

1 Introduction

The Linear Regression (LR) model is arguably the most widely used statistical model in empirical modeling across many disciplines. It provides the exemplar for all regression models as well as several other statistical models referred to as 'regression-like' models, some of which will be discussed briefly in this chapter. The primary objective is to discuss the LR model and its associated statistical inference procedures. Special attention is paid to the model assumptions and how they relate to the sampling distributions of the statistics of interest. The main lesson of this chapter is that when any of the probabilistic assumptions of the LR model are *invalid* for data $\mathbf{z}_0$:=$\{(x_t, y_t)\,,\ t=1,...,n\}$, inferences based on it will be unreliable. The unreliability of inference will often stem from inconsistent estimators and sizeable discrepancies between actual and nominal error probabilities induced by statistical misspecification.

1.1 The Linear Regression model: specification

The traditional specification of the Linear Regression (LR) model for the simple case of one regressor is usually given in terms of the error term $\{(u_t|X_t{=}x_t)\,,\ t{\in}\mathbb{N}\}$ assumptions {1}-{4} in table 1.

Table 1: Linear Regression model (traditional specification)

Statistical GM: $Y_t{=}\beta_0+\beta_1 x_t+u_t,\quad t{\in}\mathbb{N}{:=}(1,2,...,n,...)$

{1}	Normality:	$(u_t\|X_t{=}x_t)\backsim \mathsf{N}(.,.)$,	$t{\in}\mathbb{N}$.
{2}	Zero mean:	$E\,(u_t\|X_t{=}x_t)=0$,	
{3}	Homoskedasticity:	$Var\,(u_t\|X_t{=}x_t)=\sigma^2$,	
{4}	Zero correlation:	$\{(u_t\|X_t{=}x_t)\,,\ t{\in}\mathbb{N}\}$ is uncorrelated,	

When viewed from a purely probabilistic perspective as a parameterization of a stochastic process $\{\mathbf{Z}_t{:=}(Y_t, X_t),\ t{\in}\mathbb{N}\}$ giving rise to data:

$$\mathbf{Z}_0{:=}\{(x_t, y_t),\ t{=}1,2,...,n\},$$

the **complete specification** of the probabilistic assumptions defining the LR model is given in table 2 in terms of the statistical GM and the probabilistic assumptions [1]-[5] in terms of the stochastic process $\{(Y_t|X_t{=}x_t,),\ t{\in}\mathbb{N}\}$, together with the statistical

parameterization for the unknown parameters $(\beta_0, \beta_1, \sigma^2)$.

Table 2: Normal, Linear Regression model

Statistical GM: $Y_t{=}\beta_0+\beta_1 x_t+u_t,\ \ t{\in}\mathbb{N}{:=}(1,2,...,n,...)$

[1]	Normality:	$(Y_t\|X_t{=}x_t)\backsim \mathsf{N}(.,.)$,	$\left.\begin{array}{l}\\ \\ \\ \\ \\ \end{array}\right\} t{\in}\mathbb{N}.$
[2]	Linearity:	$E\left(Y_t\|X_t{=}x_t\right){=}\beta_0+\beta_1 x_t$,	
[3]	Homoskedasticity:	$Var\left(Y_t\|X_t{=}x_t\right){=}\sigma^2$,	
[4]	Independence:	$\{(Y_t\|X_t{=}x_t)\,,\ t{\in}\mathbb{N}\}$ indep. process,	
[5]	t-invariance:	$(\beta_0,\beta_1,\sigma^2)$ are *not* changing with t,	

$$\beta_0{=}\left(\mu_1{-}\beta_1\mu_2\right){\in}\mathbb{R},\ \beta_1{=}\left(\tfrac{\sigma_{12}}{\sigma_{22}}\right){\in}\mathbb{R},\ \sigma^2{=}(\sigma_{11}{-}\tfrac{\sigma_{12}^2}{\sigma_{22}}){\in}\mathbb{R}_+,$$

$\mu_1{=}E(Y_t),\ \mu_2{=}E(X_t),\ \sigma_{11}{=}Var(Y_t),\ \sigma_{22}{=}Var(X_t),\ \sigma_{12}{=}Cov(Y_t,X_t)$

A direct comparison between the two specifications of the LR model in tables 1 and 2 reveals that assumptions {1}-{4} are equivalent to assumptions [1]-[4]; note that when the process $\{(Y_t|X_t{=}x_t,),\ t{\in}\mathbb{N}\}$ is Normal, {4} coincides with [4] since non-correlation is equivalent to independence. Hence, the crucial difference comes in the form of assumption [5], which is missing from the traditional specification. Having said that, it is clear that assumption [5] is implicit in table 1 since the parameters do not have subscripts, say $(\beta_{0t},\beta_{1t},\sigma_t^2)$. This, however, blurs the distinction between a t-varying $(Var\left(u_t|X_t{=}x_t\right){=}\sigma^2(t))$ and a heteroskedastic $(Var\left(u_t|X_t{=}x_t\right){=}g(x_t))$ conditional variance; these are very different assumptions.

The last noticeable difference between the LR specifications in tables 1 and 2 is the explicit **statistical parameterization** for the unknown parameters in table 2:

$$\beta_0{=}\left(\mu_1{-}\beta_1\mu_2\right){\in}\mathbb{R},\ \beta_1{=}\left(\frac{\sigma_{12}}{\sigma_{22}}\right){\in}\mathbb{R},\ \sigma^2{=}(\sigma_{11}{-}\frac{\sigma_{12}^2}{\sigma_{22}}){\in}\mathbb{R}_+$$

It can be shown, however, this statistical parameterization is also implicit in the specification in table 1 in the sense that (McGuirk and Spanos, 2009):

$$[\{2\}\text{-}\{4\}]\longrightarrow \beta_0{=}\left(E(Y_1){-}\beta_1E(X_t)\,,\ \beta_1{=}\left(\frac{Cov(Y_t,X_t)}{Var(X_t)}\right),\ \sigma^2{=}Var(X_t){-}\frac{\left[Cov(Y_t,X_t)\right]^2}{Var(X_t)}\right.$$

Therefore, the specification in table 2 can be viewed as bringing out all the probabilistic assumptions explicitly or implicitly made in table 1 and transforms the error term assumptions into assumptions pertaining to the **observable** stochastic process $\{(Y_t|X_t{=}x_t,),\ t{\in}\mathbb{N}\}$ underlying the data $\mathbf{Z}_0{:=}(\mathbf{y}_0,\mathbf{x}_0)$. In another sense, the specification in table 2 goes much further than that. Let us unpack this claim.

The specification in table 2 provides a purely probabilistic interpretation of the LR model, that can be viewed as a particular parameterization of the vector stochastic process $\{\mathbf{Z}_t{:=}(Y_t,X_t),\ t{\in}\mathbb{N}\}$, assumed to be Normal, Independent and Identically Distributed (NIID). The NIID assumptions enable one to relate the process

$\{\mathbf{Z}_t{:=}(Y_t, X_t),\ t{\in}\mathbb{N}\}$ to the conditional process $\{(Y_t|X_t{=}x_t)\,,\ t{\in}\mathbb{N}\}$ in terms of which the model assumptions [1]-[5] are specified. Formally the connection between the two processes comes in the form of the following reduction:

$$f(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_n;\boldsymbol{\phi})\overset{\text{I}}{=}\prod_{t=1}^{n}f_t(x_t,y_t;\boldsymbol{\varphi}_t)\overset{\text{IID}}{=}\prod_{t=1}^{n}f(x_t,y_t;\boldsymbol{\varphi}){=}\prod_{t=1}^{n}f(y_t|x_t;\boldsymbol{\varphi}_1)f(x_t;\boldsymbol{\varphi}_2), \tag{1}$$

$\forall(x_t,y_t){\in}\mathbb{R}^2$. At a formal level this reduction:

(a) ensures the internal consistency of assumptions [1]-[5], and brings out their interdependence (table 3):

Table 3: Reduction and Model assumptions

$\{\mathbf{Z}_t{:=}(X_t,Y_t),\ t{\in}\mathbb{N}\}$		$\{(Y_t\|X_t{=}x_t,),\ t{\in}\mathbb{N}\}$
N	$\rightarrow$	[1]-[3]
I	$\rightarrow$	[4]
ID	$\rightarrow$	[5]

(b) assigns "a well-defined meaning" to the model parameters $\boldsymbol{\theta}{=}(\beta_0,\beta_1,\sigma^2)$ via their explicit parameterization, and

(c) brings out the restrictions (probabilistic assumptions) on the process $\{\mathbf{Z}_t,\ t{\in}\mathbb{N}\}$ needed for the model parameterization in table 2 to hold.

At a practical level this reduction provides a direct link between the NIID reduction and the model assumptions, as shown in table 3. This link enables one to use data plots associated with the original data $\mathbf{Z}_0{:=}(\mathbf{y}_0,\mathbf{X}_0)$ to evaluate (indirectly) the model assumptions [1]-[5]. This link shown in table 3, provides an indirect way for informed conjectures about which model assumptions are likely to be invalid when any departures from NIID are detected using t-plots and scatterplots of the data $\mathbf{Z}_0$.

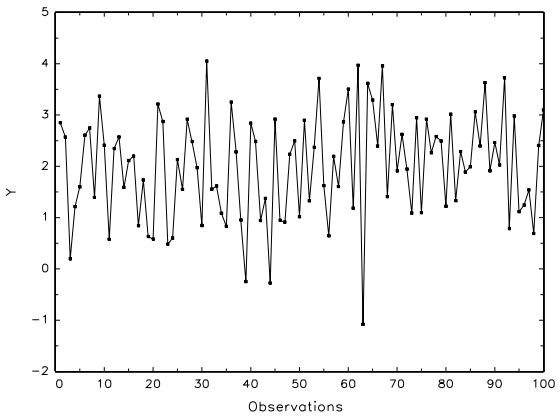

Fig. 1: t-plot of y_t

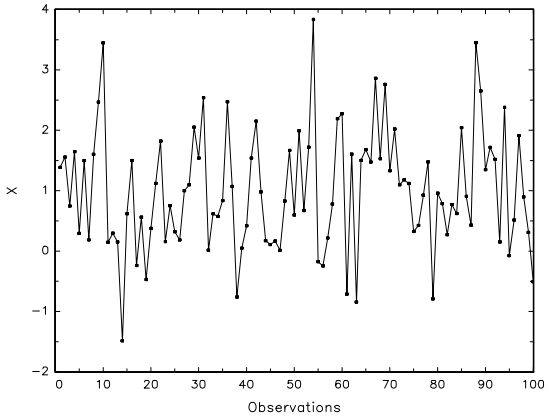

Fig. 2: t-plot of x_t

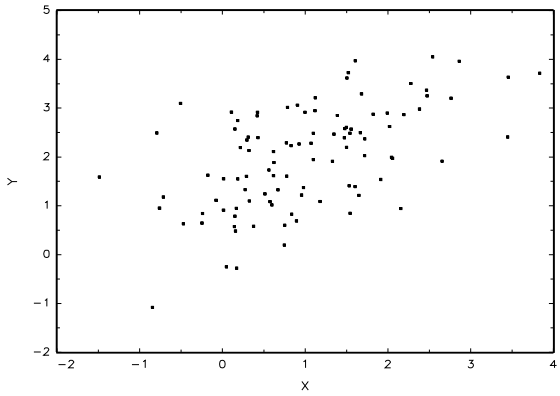

Fig. 3: Scatter-plot of (x_t, y_t)

It turns out that this link renders well-informed, not only the specification of a statistical model, but also its M-S testing and respecification when certain model assumptions are found wanting.

1.2 Estimation

Assumption [1]-[5] imply that the Likelihood (LF) and log-likelihood functions are:

$$L(\beta_0,\beta_1,\sigma^2;\mathbf{y}) \;\propto (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\tfrac{1}{2\sigma^2}\textstyle\sum_{t=1}^{n}(Y_t-\beta_0-\beta_1x_t)^2\right\},$$

$$\ln L(\boldsymbol{\theta};\mathbf{y})= \;\text{const}-\tfrac{n}{2}\ln\sigma^2-\tfrac{1}{2\sigma^2}\textstyle\sum_{t=1}^{n}(Y_t-\beta_0-\beta_1x_t)^2.$$

The first-order conditions take the form:

$$\begin{aligned}
&\text{(i)}\ \tfrac{\partial \ln L(\boldsymbol{\theta};\mathbf{y})}{\partial\beta_0}= -\tfrac{1}{2\sigma^2}(-2)\textstyle\sum_{t=1}^{n}(Y_t-\beta_0-\beta_1x_t)=\tfrac{1}{\sigma^2}\sum_{t=1}^{n}u_t=0,\\
&\text{(ii)}\ \tfrac{\partial \ln L(\boldsymbol{\theta};\mathbf{y})}{\partial\beta_1}= -\tfrac{1}{2\sigma^2}(-2)\textstyle\sum_{t=1}^{n}(Y_t-\beta_0-\beta_1x_t)x_t=\tfrac{1}{\sigma^2}\sum_{t=1}^{n}x_tu_t=0, \qquad (2)\\
&\text{(iii)}\ \tfrac{\partial \ln L(\boldsymbol{\theta};\mathbf{y})}{\partial\sigma^2}=\text{-}\tfrac{n}{2\sigma^2}+\tfrac{1}{2\sigma^4}\textstyle\sum_{t=1}^{n}(Y_t-\beta_0-\beta_1x_t)^2=\text{-}\tfrac{n}{2\sigma^2}+\tfrac{1}{2\sigma^4}\sum_{t=1}^{n}u_t^2=0.
\end{aligned}$$

where $u_t=(Y_t-\beta_0-\beta_1x_t)$. Solving (i)-(ii) for (β_0,β_1) yields the MLEs:

$$\widehat{\beta}_0=\overline{Y}-\widehat{\beta}_1\overline{x},\ \widehat{\beta}_1=\tfrac{\sum_{t=1}^{n}(Y_t-\overline{Y})(x_t-\overline{x})}{\sum_{t=1}^{n}(x_t-\overline{x})^2},\ \overline{Y}=\tfrac{1}{n}\textstyle\sum_{t=1}^{n}Y_t,\ \overline{x}=\tfrac{1}{n}\sum_{t=1}^{n}x_t. \qquad (3)$$

Solving condition (iii) for σ^2 and using $(\widehat{\beta}_0,\widehat{\beta}_1)$ yields the MLE:

$$\widehat{\sigma}^2_{ML}=\tfrac{1}{n}\textstyle\sum_{t=1}^{n}(Y_t-\widehat{\beta}_0-\widehat{\beta}_1x_t)^2=\tfrac{1}{n}\sum_{t=1}^{n}\widehat{u}_t^2, \qquad (4)$$

where $\{\widehat{u}_t=(Y_t-\widehat{\beta}_0-\widehat{\beta}_1x_t),\ t=1,2,...,n\}$ denotes the residuals.

Sampling distributions of MLEs $(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\sigma}^2_{ML})$. The finite sampling distributions of $(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\sigma}^2_{ML})$ have a family resemblance to those of the simple Normal model:

$$\widehat{\beta}_0\backsim \mathsf{N}(\beta_0,\sigma^2\left(\tfrac{1}{n}+\varphi_x\overline{x}^2\right)),\quad \widehat{\beta}_1\backsim \mathsf{N}(\beta_1,\sigma^2\varphi_x),\quad \tfrac{n\widehat{\sigma}^2_{ML}}{\sigma^2}\backsim\chi^2(n-2), \qquad (5)$$

where $\varphi_x=[\sum_{t=1}^n(x_t-\overline{x})^2]^{-1}$. Note that $(\widehat{\beta}_0,\widehat{\beta}_1)$ are dependent since $Cov(\widehat{\beta}_0,\widehat{\beta}_1)=-\sigma^2(\varphi_x\overline{x})$.

Finite sample properties of MLEs $(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\sigma}^2_{ML})$.

Unbiasedness. It is clear from (5) that the MLEs $(\widehat{\beta}_0,\widehat{\beta}_1)$ are unbiased estimators of (β_0,β_1). In contrast, $\widehat{\sigma}^2_{ML}$ is a biased estimator of σ^2 since $E(\frac{n\widehat{\sigma}^2_{ML}}{\sigma^2})=(n-2)$ which implies that $E(\widehat{\sigma}^2_{ML})=\frac{(n-2)\sigma^2}{n}\neq\sigma^2$, but by the same token the estimator $s^2=\frac{1}{n-2}\sum_{t=1}^n(Y_t-\widehat{\beta}_0-\widehat{\beta}_1x_t)^2$ is unbiased.

Full efficiency. The Fisher information matrix is:

$$\mathcal{I}_n(\beta_0,\beta_1,\sigma^2)=\begin{pmatrix}\frac{n}{\sigma^2} & \frac{1}{\sigma^2}\sum_{t=1}^n x_t & 0\\ \frac{1}{\sigma^2}\sum_{t=1}^n x_t & \frac{1}{\sigma^2}\sum_{t=1}^n x_t^2 & 0\\ 0 & 0 & \frac{n}{2\sigma^4}\end{pmatrix}, \tag{6}$$

and the Cramer-Rao (C-R) lower bound is:

$$\text{C-R}(\boldsymbol{\theta})=\mathcal{I}_n^{-1}(\boldsymbol{\theta})=\begin{pmatrix}\sigma^2\left(\frac{1}{n}+\varphi_x\overline{x}^2\right) & -\sigma^2(\varphi_x\overline{x}) & 0\\ -\sigma^2(\varphi_x\overline{x}) & \sigma^2\varphi_x & 0\\ 0 & 0 & \frac{2\sigma^4}{n}\end{pmatrix},$$

Hence, $(\widehat{\beta}_0,\widehat{\beta}_1)$ are fully efficient.

Example 14.2. Consider the data in table 1 in Appendix 14.C (Mendhall and Sinchich, 1996, p. 184), where z_{1t}-the auction final price and z_{2t}-the age of an antique grandfather clock in a sequence of n=32 such transactions. The assumed regression model is of the form given in table 14.1, where $Y_t=\ln(Z_{1t})$ and $X_t=\ln(Z_{2t})$. The idea is to account for the final auction price using the age of the antique clock as the explanatory variable. The estimated linear regression using the data in table 1 (Appendix 14.C) yields:

$$\underset{}{Y_t}=\underset{(.966)}{1.312}+\underset{(.195)}{1.177}x_t+\widehat{u},\quad s=.208,\ n=32, \tag{7}$$

Note that all the above estimates and their standard errors stem from five numbers, the first two sample moments steaming from data $\mathbf{z}_0$:$=\{(x_t,y_t),\ t=1,2,...,n\}$:

$$\begin{array}{l}\overline{Y}=7.148,\ \overline{x}=4.958,\ \widehat{Var}(Y_t)=\frac{1}{n}\sum_{t=1}^n(Y_t-\overline{Y})^2=.0905,\\ \widehat{Var}(X_t)=\frac{1}{n}\sum_{t=1}^n(x_t-\overline{x})^2=.0359,\ \widehat{Cov}(Y_t,X_t)=\frac{1}{n}\sum_{t=1}^n(Y_t-\overline{Y})(x_t-\overline{x})=.0423.\end{array} \tag{8}$$

The estimated coefficients seem reasonable on substantive grounds because an economist expects the value of the antique clock to increase with its age. Having said that, a modeler should exercise caution concerning such evaluations, before the validity of the model assumptions [1]-[5] in table 14.1 is established.

One can use informed conjectures based on the link between reduction and model assumptions, shown in table 14.3, to provide informed conjectures about which model

assumptions are likely to be invalid when certain potential departures from NIID are gleaned. A glance at the four figures 14.1-4 does not indicate any serious departures from the NIID assumptions, suggesting that [1]-[5] are likely to be valid for this data.

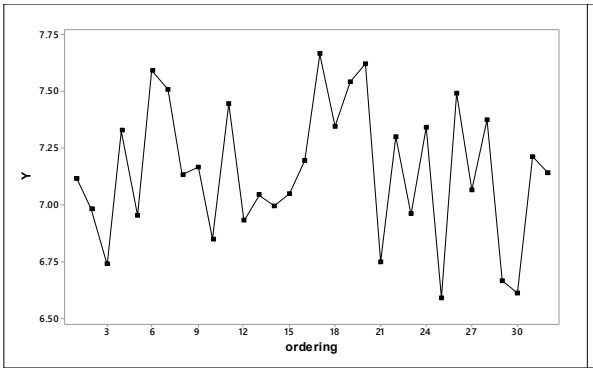

Fig. 14.1: t-plot of y_t

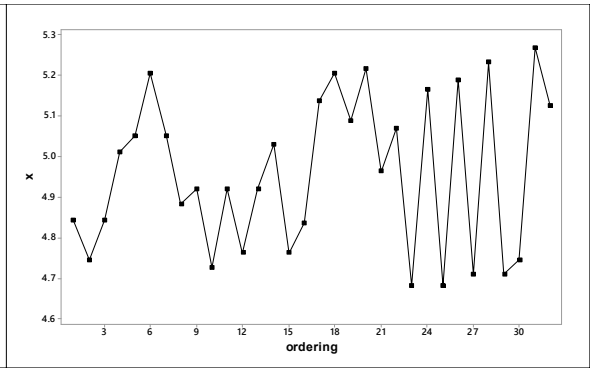

Fig. 14.2: t-plot of x_t

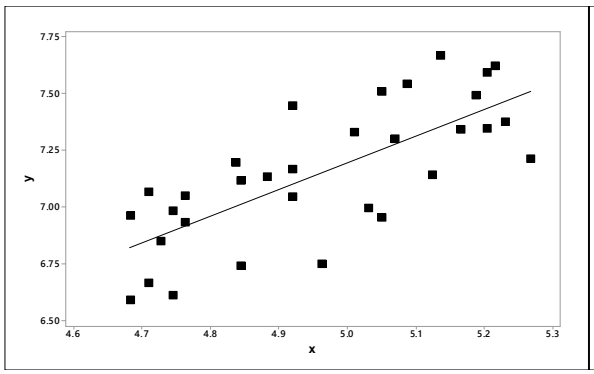

Fig. 14.3: Scatterplot of (x_t, y_t), t=1, ..., n

Fig. 14.4: t-plot of the residuals from (31)

CAUTION: it is important to emphasize that to establish statistical adequacy formally, requires one to apply comprehensive Mis-Specification (M-S) testing to evaluate the validity of the model assumptions thoroughly; see chapter 15. The sample size in this case is rather small to allow thorough M-S testing.

1.2.1 Sample moments can be highly misleading; plot the data!

As argued above, all the inference results associated with the Linear Regression model (table 14.1) discussed so far depend on the first two sample moments (means, variance-covariance) of the observable stochastic process $\{\mathbf{Z}_t := (X_t, Y_t),\ t \in \mathbb{N}\}$; see (8) for the estimated model in (31). Going directly to the sample moments, however, is a very bad strategy because these moments might turn out to be statistically misspecified. To illustrate that, let us consider the following example.

Example 14.3: Anscombe's (1973) data. Table 2 in Appendix 14.C reports the data contrived by Anscombe (1973) to emphasize how important it is to look at data plots before any modeling is attempted. The data comprise 4 pairs of variables (Y_{it}, x_{it}), i=1, 2, 3, 4, t=1, 2, ..., 11. The sample moments of each pair are identical

for i=1, .., 4:

$$\begin{pmatrix} \overline{Y}_i=7.501 \\ \overline{x}_i=9.00 \end{pmatrix}, \quad \begin{pmatrix} \frac{1}{n}\sum_{t=1}^{n}\left(Y_{it}-\overline{Y}_i\right)^2=4.127 & \frac{1}{n}\sum_{t=1}^{n}(Y_{it}-\overline{Y}_i)(x_{it}-\overline{x}_i)=5.5 \\ \frac{1}{n}\sum_{t=1}^{n}(x_{it}-\overline{x}_i)(Y_{it}-\overline{Y}_i)=5.5 & \frac{1}{n}\sum_{t=1}^{n}(x_{it}-\overline{x}_i)^2=11.0 \end{pmatrix}$$

giving rise to numerically identical estimated regression results:

$$\underset{(1.12)}{Y_{it}=3.00}+\underset{(.118)}{.500}x_{it}+\widehat{u}_{it},\ R_i^2=.667,\ s_i=1.236,\ n=11,\ i=1,..,4,$$

giving rise to identical inferences results pertaining to $(\beta_0, \beta_1, \sigma^2)$. The scatterplots of the 4 estimated linear regressions (figures 14.5-8), however, reveal a very different story about the trustworthiness of these inference results.

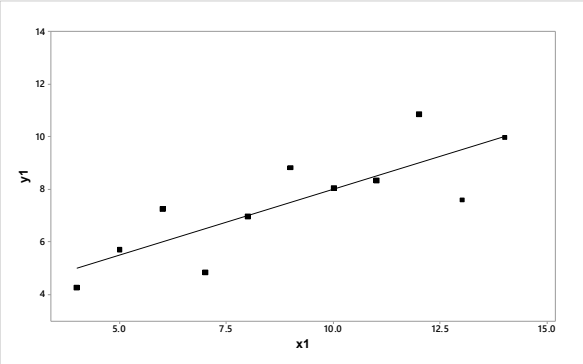

Fig. 14.5: Scatterplot of (x_{1t}, y_{1t}), t=1, ..., n

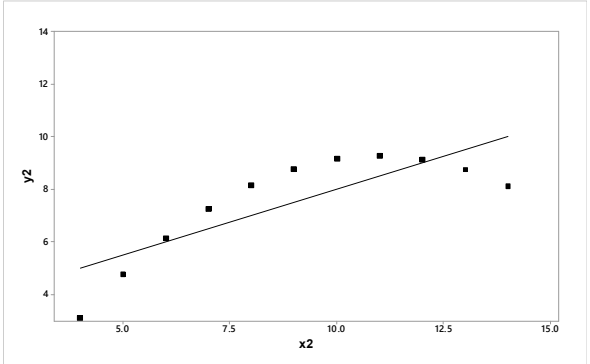

Fig. 14.6: Scatterplot of (x_{2t}, y_{2t}), t=1, ..., n

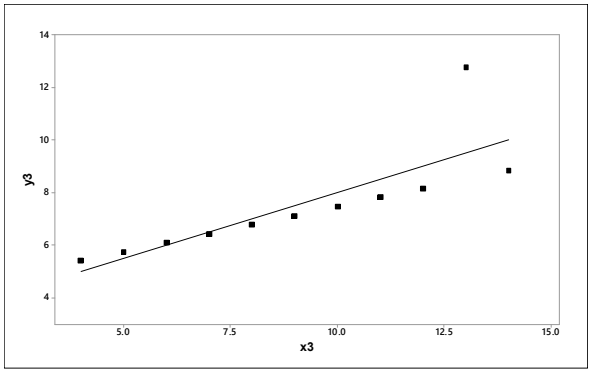

Fig. 14.7: Scatterplot of (x_{3t}, y_{3t}), t=1, ..., n

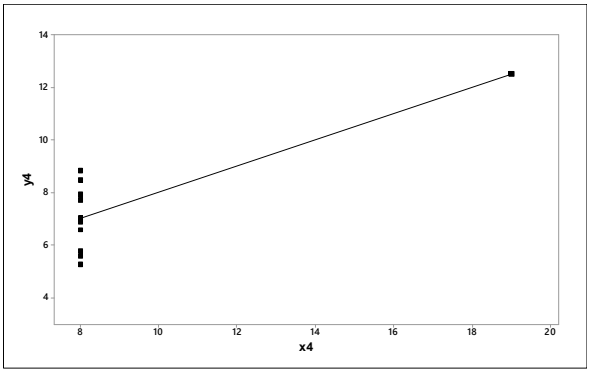

Fig. 14.8: Scatterplot of (x_{4t}, y_{4t}), t=1, ..., n

Only one estimated regression, based on the data in figure 14.5 is seemingly statistically adequate. The scatterplots of the rest indicate clearly that they are statistically misspecified. In light of the small sample size n=11, no formal M-S testing is possible, but as argued in chapter 1, when n is too small for a comprehensive M-S testing, it should be considered too small for inference purposes.

1.3 Fitted values and residuals

The LR regression *fitted values* and *residuals* are defined by:

$$\widehat{Y}_t=\widehat{\beta}_0+\widehat{\beta}_1 x_t, \quad \text{and } \widehat{u}_t=Y_t-\widehat{\beta}_0-\widehat{\beta}_1 x_t, \quad t=1,2,...,n,$$

and satisfy the restrictions:

$$\text{(a)}^* \ \tfrac{1}{n}\textstyle\sum_{t=1}^{n} \widehat{u}_t=0, \quad \text{(b)}^* \ \tfrac{1}{n}\textstyle\sum_{t=1}^{n} \widehat{Y}_t\widehat{u}_t=0.$$

By adding and subtracting the fitted values $\widehat{Y}_t$ in $\left(Y_t-\overline{Y}\right)$ yields the variance decomposition:

$$\textstyle\sum_{t=1}^{n}(Y_t-\overline{Y})^2= \ \ \sum_{t=1}^{n}(\widehat{Y}_t-\overline{Y})^2 \ \ +\sum_{t=1}^{n} \widehat{u}_t^2, \tag{9}$$

where $\mathsf{TSS}=\sum_{t=1}^{n}\left(Y_t-\overline{Y}\right)^2$ is known as the Total Sum of Squares,
$\mathsf{ESS}=\sum_{t=1}^{n}(\widehat{Y}_t-\overline{Y})^2=\widehat{\beta}_1^2\sum_{t=1}^{n}(x_t-\overline{x})^2$ is known as the Explained Sum of Squares, and
$\mathsf{RSS}=\sum_{t=1}^{n}(\widehat{Y}_t-\overline{Y})^2=\sum_{t=1}^{n}\widehat{u}_t^2$ is known as the Residual Sum of Squares.

1.4 Goodness-of-fit measures

The decomposition in (9) can be used to define a goodness-of-fit measure:

$$R^2=\tfrac{\mathsf{ESS}}{\mathsf{TSS}}=\tfrac{\sum_{t=1}^{n}(\widehat{Y}_t-\overline{Y})^2}{\sum_{t=1}^{n}(Y_t-\overline{Y})^2}=1-\tfrac{\mathsf{RSS}}{\mathsf{TSS}}=1-\tfrac{\sum_{t=1}^{n}\widehat{u}_t^2}{\sum_{t=1}^{n}(Y_t-\overline{Y})^2}, \ 0\leq R^2\leq 1, \tag{10}$$

which, in this case of a single regressor (X_t) coincides with $Corr(Y_t, X_t)$:

$$R^2=\frac{\widehat{\beta}_1^2\sum_{t=1}^{n}(x_t-\overline{x})^2}{\sum_{t=1}^{n}\left(Y_t-\overline{Y}\right)^2}=\frac{\left[\sum_{t=1}^{n}(Y_t-\overline{Y})(x_t-\overline{x})\right]^2}{\left[\sum_{t=1}^{n}(x_t-\overline{x})^2\right]\left[\sum_{t=1}^{n}\left(Y_t-\overline{Y}\right)^2\right]}.$$

The R^2 can also be expressed in terms of an F-statistic for testing H_0: $\beta_1=0$ and H_1: $\beta_1\neq 0$:

$$F(\mathbf{y})=\tfrac{[(\mathsf{TSS}-\mathsf{RSS})/1]}{(\mathsf{RSS}/(n-2))}=\left(\tfrac{R^2}{(1-R^2)/(n-2)}\right) \overset{H_0}{\backsim} \mathsf{F}(1,n-2), \tag{11}$$

where $\mathsf{F}(1,n-2)$ denotes the F-distribution with 1 and $n-2$ degrees of freedom.

Example 14.2 (continued). Returning to the estimated linear regression model in (31), the $R^2=1-\frac{1.2655}{2.8069}=.549$, where the variance decomposition in (9) is given in table 14.4.

Table 14.4: Variance Decomposition

Source	Sum of Squares	df	Mean Square	F(1,30)
ESS	1.5415	1	1.5415	36.54
RSS	1.2655	30	$\frac{1.2655}{30}=.0422$	
TSS	2.8069	31		

It is important to emphasize that the above statistics are meaningful only in the case where the estimated linear regression model in (31) is statistically adequate.

2 Testing a substantive model against the data

As argued in the previous chapters, behind every structural (substantive) model:

$$\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})=\{f(\mathbf{z};\boldsymbol{\varphi}),\ \boldsymbol{\varphi}\in\Phi\},\ \mathbf{z}\in\mathbb{R}^n_Z,\ \text{for } \boldsymbol{\varphi}\in\Phi\subset\mathbb{R}^p,\ p<n,$$

being subjected to statistical analysis, there is always a statistical model (often implicit):

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})=\{f(\mathbf{z};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{z}\in\mathbb{R}^n_Z,\ \text{for } \boldsymbol{\theta}\in\Theta\subset\mathbb{R}^m,\ p<m<n,$$

and the two are related via certain nesting restrictions $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$, where $\boldsymbol{\theta}$ denotes the statistical parameters and $\boldsymbol{\varphi}$ the substantive parameters of interest.

The Capital Asset Pricing Model (CAPM). Consider the following example:

$$\begin{aligned}&\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})\text{: } (Y_t-x_{2t})=\alpha_1(x_{1t}-x_{2t})+\varepsilon_t,\ \varepsilon_t\backsim\mathsf{NIID}(0,\sigma^2_\varepsilon),\ t=1,\ldots,n,\ldots,\\&\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})\text{: } Y_t=\beta_0+\beta_1x_{1t}+\beta_2x_{2t}+u_t,\ u_t\backsim\mathsf{NIID}(0,\sigma^2_u),\ t=1,2,\ldots,n,\ldots\end{aligned}\tag{12}$$

where the substantive model represents the CAPM, with y_t- returns of a particular asset or portfolio of assets, x_{1t}- market returns, and x_{2t}- returns of a risk free asset. The structural model has 3 parameters $\boldsymbol{\varphi}:=(\alpha_1,\sigma^2_\varepsilon)$ and the statistical has 4 parameters $\boldsymbol{\theta}:=(\beta_0,\beta_1,\beta_2,\sigma^2_u)$. The important point is that the substantive model is a special case of the statistical, and in practice the former is part of probing substantive adequacy is to test the restrictions relating the statistical with the substantive parameters. In this case these restrictions are:

$$\beta_0=0,\ \beta_1+\beta_2=1.\tag{13}$$

Before these restrictions can be tested using reliable testing procedures, it is important to secure the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. Without it, there is no reason to believe the testing results because they likely to be unreliable.

2.1 Statistical misspecification and untrustworthy empirical evidence

What goes wrong when an estimated model is statistically misspecified? For the estimators of the model parameters, certain departures often imply that the estimates are highly unreliable; the estimators are likely to be inconsistent in the presence of mean t-heterogeneity. For hypothesis testing, the *nominal* (assumed) error probabilities, like the type I and II and the coverage probability, are very different from the *actual* ones! Applying a .05 significance level test, when the actual type I error is .95, will lead an inference astray! This is because instead of rejecting a true null hypothesis 5% of the time, as assumed, one is actually rejecting it 95% of the time! WARNING: when looking at published empirical results in prestigious journals, keep in mind that a very high proportion, say 99%, are of questionable trustworthiness!

Example 14.8. Lai and Xing (2008), pp. 71-81, illustrate the CAPM using *monthly data* for the period Aug. 2000 to Oct. 2005 (n=64); see Appendix 5.C. For

simplicity, let us focus on one of their equations where: y_t is excess (log) returns of Intel, x_t is the market excess (log) returns based on the SP500 index; the risk free returns is based on the 3-month Treasury bill rate. Estimation of the statistical (LR) model that nests the CAPM when the constant is zero yields:

$$\underset{(.009)}{Y_t=.020}+\underset{(.237)}{1.996}x_t+\widehat{u}_t,\ R^2=.536,\ s=.0498,\ n=64, \tag{14}$$

where the standard errors are given in parentheses.

On the basis of the estimated model in (14), the authors proceeded to draw the following inferences providing strong evidence *for* the CAPM:

(a) the signs and magnitudes of the estimated (β_0,β_1) corroborate the CAPM:

(i) the beta coefficient β_1 is statistically significant: $\tau_1(\mathbf{y}_0)=\frac{1.996}{.237}=8.422[.000]$.

(ii) the restriction $\beta_0=0$ is *accepted* at $\alpha=.025$ since $\tau_0(\mathbf{y}_0)=\frac{.019}{.009}=2.111[.039]$,

(b) the goodness-of-fit ($R^2=.536$) is high enough to provide additional evidence for the CAPM.

The problem with these inferences is that their trustworthiness depends crucially on the estimated Linear Regression in (14) being statistically adequate. But is it?

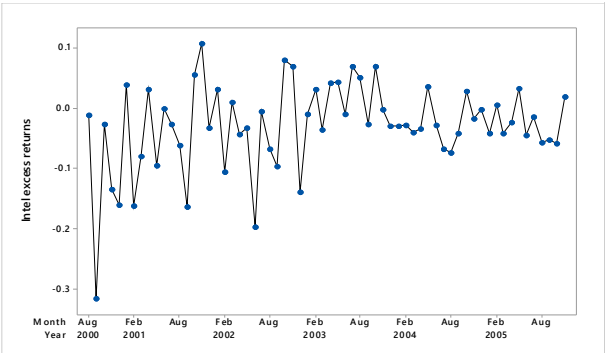

Fig. 14.9: Intel Corp. excess returns

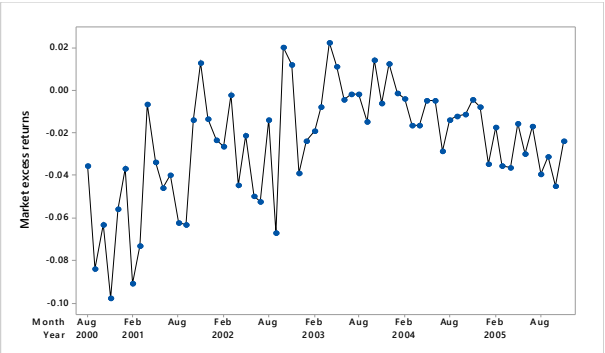

Fig. 14.10: Market excess returns

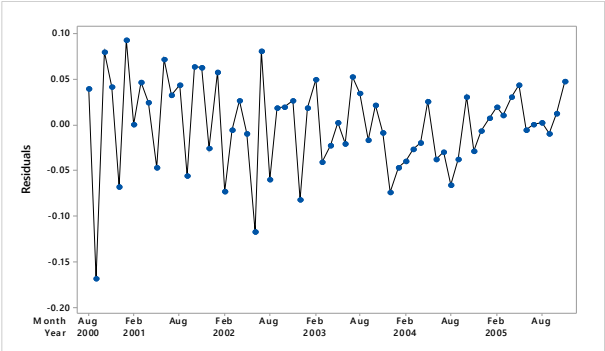

Fig. 14.11: t-plot of the residuals from (14)

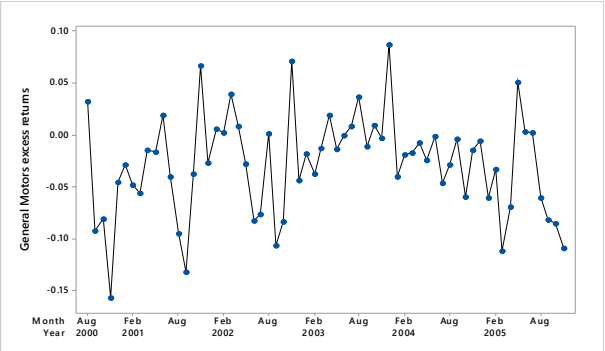

Fig. 14.12: GM excess returns

The first hint that some of the model assumptions [1]-[5] (table 14.4) are most probably invalid for the particular data comes from basic data plots. A glance at

the t-plots of the data $\{(y_t, x_t),\ t=1,2,...,n\}$ (fig. 14.9-10), suggests that the data exhibit very distinct time cycles and trends in the mean, and a shift in the variance after observation t=30. Using the link between reduction and model assumptions in table 14.2, it can be conjectured that assumptions [4]-[5] are likely to be invalid. The residuals from the estimated equation in (14), shown in fig. 14.11, corroborate this.

A formal introduction to Mis-Specification (M-S) testing is given in chapter 15, but as a prelude to that discussion let us consider two auxiliary regressions aiming to bring out certain forms of departure from the model assumptions [2]-[5] as indicated in (15)-(16). It suffices at this stage to interpret such auxiliary regressions as an attempt to probe for the presence of statistical systematic information in the residuals $(\widehat{u}_t)$ and their squares $(\widehat{u}_t^2)$, where the $\widehat{u}_t$-equation in (15) probes for departures pertaining to assumptions about $E(Y_t|X_t{=}x)$ and the $\widehat{u}_t^2$-equation in (16) for departures from $Var(y_t|X_t{=}x)$.

The misspecifications conjectured above are confirmed by the following auxiliary regressions:

$$\widehat{u}_t = \underset{(.018)}{.047} + \underset{(.474)}{.761}x_t + \overbrace{\underset{(6.24)}{7.69}x_t^2}^{[2]} - \overbrace{\underset{(.044)}{.192}D_2}^{[5]} - \overbrace{\underset{(.338)}{1.11}t^2 + \underset{(.054)}{.175}t^3}^{[5]} - \overbrace{\underset{(.109)}{.320}\widehat{u}_{t-1}}^{[4]} + \widehat{v}_{1t}, \tag{15}$$

where D_2 is a dummy variable (takes the value one for t=2 and zeroes for $t{\neq}2$):

$$\widehat{u}_t^2 = \underset{(.0008)}{.004} + \overbrace{\underset{(.0025)}{.025}D_2}^{[5]} - \overbrace{\underset{(.002)}{.006}t}^{[5]} - \overbrace{\underset{(.179)}{.053}x_t^2}^{[3]} + \widehat{v}_{2t}, \quad R^2{=}.19,\ n{=}64 \tag{16}$$

Linearity [2]: $\tau(56){=}\frac{7.69}{6.24}{=}.123[.223]$,

Homoskedasticity [3]: $\tau(56){=}\frac{.053}{.179}{=}.30[.768]$

Dependence [4]: $\tau(56){=}\frac{.320}{.109}{=}2.93[.005]^*$

Mean-heterogeneity [5]: $\tau_D(56){=}\frac{.192}{.044}{=}4.32[.0000]^*$,
$F(2;57){=}\frac{(.035671)/2}{(.093467)/56}{=}10.686[.0002]^*$,

Variance-heterogeneity [5]: $\tau_D(56){=}\frac{.025}{.0025}{=}9.75[.0000]^*$,
$\tau(56){=}\frac{.006}{.002}{=}3.04[.004]^*$.

Testing Normality test using the original model residuals is a good strategy only when all the other model assumptions [2]-[5] are shown to be valid. This is because the current Normality tests assume that the residuals are IID, i.e. assumptions [2]-[5] are valid, which is not the case above. Hence, the above M-S testing results indicate clearly that no reliable inferences can be drawn on the basis of the estimated model in (14) since assumptions [4]-[5] **are invalid**.

2.2 Probing for substantive adequacy?

2.2.1 Case 1: the statistical model is misspecified

To illustrate the perils of a misspecified model, consider posing the question: is z_{t-1}-last period's excess returns of General Motors (see fig. 14.12) an omitted variable in (14)? Adding z_{t-1} to the estimated equation (14) gives rise to:

$$Y_t = \underset{(.009)}{.013} + \underset{(.232)}{2.082}x_t - \underset{(.129)}{.296}z_{t-1} + \widehat{\epsilon}_t,\ R^2=.577,\ s=.0483,\ n=63. \tag{17}$$

Taking (17) at face value, the t-statistic: $\tau(60)=\frac{.296}{.129}=2.29[.026]$,
suggests that z_{t-1} is a relevant omitted variable, which is a highly misleading inference. The truth is that any variable which picks up the unmodeled trend, will misleadingly appear to be statistically significant. Indeed, a simple respecification of the original model, such as adding trends and lags to account for the detected departures based on the above M-S testing:

$$Y_t = \underset{(.02)}{.049} - \underset{(.045)}{.175}D_2 + \underset{(.272)}{2.307}x_t - \underset{(.101)}{.755}t^2 + \underset{(.057)}{.119}t^3 - \underset{(.090)}{.205}Y_{t-1} + \underset{(.138)}{.032}z_{t-1} + \widehat{\varepsilon}_t,$$
$$R^2=.704,\ s=.0418,\ n=63,$$

renders z_{t-1} insignificant since its t-statistic is: $\tau(56)=\frac{.032}{.138}=.023[.818]$. The moral of this example is that one should never probe for substantive adequacy when the underlying statistical model is misspecified. In such a case, the inference procedures used to decided whether a new variable is relevant are unreliable!

2.2.2 Case 2: the statistical model is statistically adequate

Example 14.9. Let us return to the example 14.2 where the original structural model coincided with a LR model with $Y_t=\ln(Z_{1t})$ and $X_t=\ln(Z_{2t})$, z_{1t}-the auction final price and z_{2t}-the age of an antique grandfather clock in a sequence of $n=32$ such transactions. Let us assume that during a presentation of the estimated regression model:

$$Y_t = \underset{(.966)}{1.312} + \underset{(.195)}{1.177}x_t + \widehat{u}, \quad s=.208,\ R^2=.549,\ n=32, \tag{18}$$

an economist in the audience raised the possibility that (18) is substantively inadequate because a crucial explanatory variable, z_{2t}-the number of bidders has been omitted. The modeler decides to evaluate this and re-estimates (18) with the additional regressor $x_{2t}=\ln z_{2t}$:

$$Y_t = -\underset{(.382)}{1.316} + \underset{(.070)}{1.418}x_{1t} + \underset{(.044)}{.649}x_{2t} + \widehat{u}_t,\ s=.0718,\ R^2=.947,\ n=32. \tag{19}$$

and its residuals are plotted in figure 14.13.

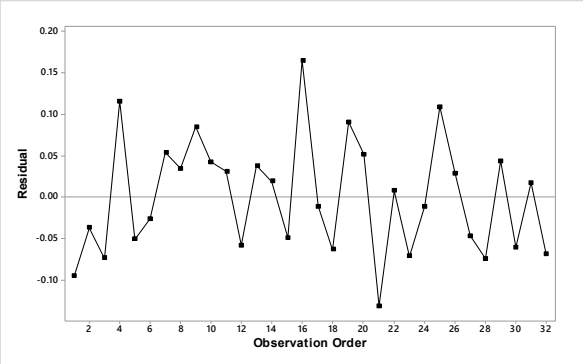

Fig. 14.13: t-plot of the residuals of (19)

In light of the fact that (18) is statistically adequate, one can trust the t-test for the significance of X_{2t}:

$$\tau(\mathbf{y})=\frac{.649}{.044}=14.75[.000],$$

to infer that indeed, X_{2t} is indeed a relevant explanatory variable that enhances the substantive adequacy of the original model in (18). The t-plots of the residuals from (19) confirm that the respecified structural model has retained the statistical adequacy.

2.2.3 Statistical vs. substantive adequacy: the tale of two error terms

To distinguish between statistical and substantive adequacy, it is important to bring out the different interpretations for the two error terms associated with the statistical and substantive models, and how the probing of potential departures from these error terms differs in crucial respects.

The **statistical error term** u_t is assumed to represent the *non-systematic* statistical information in data $\mathbf{Z}_0{:=}(\mathbf{y},\mathbf{X})$, left behind by the systematic component $m(t){=}E(Y_t|X_t{=}x_t)$. This is true when assumptions [1]-[5] are valid. Hence, the statistical error term u_t is:

[i] *Derived* ($u_t{=}Y_t{-}\beta_0{-}\beta_1x_{1t}{-}\beta_2x_{2t}$) in the sense that the probabilistic structure of $\{(u_t|X_t{=}x_t),\ t{\in}\mathbb{N}\}$ is completely determined by that of the observable process $\{(Y_t|X_t{=}x_t),\ t{\in}\mathbb{N}\}$; assumptions [1]-[5]. This implies that when any of the assumptions [1]-[5] are invalid, u_t will include the systematic statistical information in $\mathbf{Z}_0$ left unaccounted for by $m(t)$.

[ii] *Local,* in the sense that the validity of its probabilistic structure revolves around how adequately the statistical model accounts for the statistical systematic information in data $\mathbf{Z}_0$. That is, when probing to establish statistical adequacy the potential errors pertain only to statistical systematic information in $\mathbf{Z}_0$ that might have been overlooked by the statistical model, e.g. departures from assumptions [1]-[5].

In contrast, the **structural error term** ε_t is assumed to represent the *non-systematic* substantive information left behind by the structural (substantive) model. In this sense, ε_t is:

[i]* *Autonomous* in the sense that, unlike u_t, the probabilistic structure of ε_t is not entirely determined by the substantive information framed by the structural model and its variables. It also depends on other relevant substantive information that might have been overlooked. This includes omitted variables, confounding factors, external shocks, systematic errors of measurement and approximation, etc. This implies that when the structural model does not account adequately (describes, explains, predicts) for the phenomenon of interest, ε_t will include such neglected substantive information.

[ii]* *Global,* in the sense that the validity of its probabilistic structure revolves around how adequately the structural model accounts for the phenomenon of interest. Hence, when probing to establish substantive adequacy one needs to consider the different ways the structural model might deviate from the actual data generating mechanism that gave rise to the phenomenon of interest; not just the part that generated data $\mathbf{Z}_0$.

3 Linear regression and least-squares

3.1 Mathematical approximation and statistical curve-fitting

As argued in chapter 12, the *principle of least-squares* has its origins in mathematical approximation using linear (in parameters) functions, originally proposed by Legendre in 1805. The basic idea involves the approximation of an *unknown* function:

$$y{=}h(\mathbf{x}),\ (\mathbf{x},y){\in}\left(\mathbb{R}_X^p{\times}\mathbb{R}_Y\right),$$

where $h(.)$: $\mathbb{R}_X^p \to \mathbb{R}_Y$, by selecting an *approximating* function, say:

$$g(\mathbf{x}){=}\alpha_0{+}\textstyle\sum_{i=1}^m \alpha_i x_i,\ (\mathbf{x},y){\in}\left(\mathbb{R}_X^p{\times}\mathbb{R}_Y\right),$$

and using data $\mathbf{z}_0{:=}\{(\mathbf{x}_t,y_t),\ t{=}1,2,..,n\}$ to get the curve of best fit:

$$\widehat{y}_t{=}\widehat{g}(\mathbf{x}_t){=}\widehat{\alpha}_0{+}\textstyle\sum_{i=1}^p \widehat{\alpha}_i x_{it}.$$

Gauss in 1809 transformed this mathematical approximation problem into a statistical estimation procedure by adding a probabilistic structure via a generic error term:

$$Y_t{=}g(\mathbf{x}_t)+\epsilon_t,\ \epsilon_t{\backsim}\mathsf{NIID}\left(0,\sigma^2\right),\ t{=}1,2,\ldots,n, \tag{20}$$

that gave rise to the *Gauss Linear (GL) model* in table 14.5; see Seal (1967) and Plackett (1965) for its historical development. In this book, the GL model is not viewed as a variation on the LR model, because the two are very different as statistical models, even though they share certain apparently common procedures where the matrix notation blurs the differences. As argued in chapter 12, the GL is primarily motivated by the curve-fitting perspective in contrast to the LR model which is the quintessential statistical model, with clear statistical parameterizations.

Let us bring out some of the differences between the Linear Regression (LR) model specified in table 14.1 and the GL model. Apart from the fact that $\{\mathbf{x}_t\}_{t=1}^n$ is viewed as a sequence of given numbers, and not the observations associated with a random variable X_t, the linearity assumption in table 14.5 that matters is in terms of the

parameters. That is, the statistical GM of the Gauss Linear model can be extended to a polynomial (preferably orthogonal) in x_t, say (Seber and Lee, 2002):

$$Y_t{=}\alpha_0+\sum_{k=1}^{p}\alpha_k x_t^k+\epsilon_t,\ \epsilon_t\backsim \mathsf{NIID}\left(0,\sigma^2\right),\ t{\in}\mathbb{N},$$

without changing the probabilistic structure of the model. In this sense, the GL model is more appropriately viewed in the context of a curve-fitting framework. Indeed, the Gauss Linear model is not usually specified as in table 14.5, but as in table 14.6 (Kennedy, 2008), where the parameters $\boldsymbol{\theta}{:=}\left(\boldsymbol{\alpha},\sigma^2\right)$, $\boldsymbol{\alpha}{:=}(\alpha_0,\boldsymbol{\alpha}_1)$ are often assigned a substantive interpretation, ignoring the statistical interpretation and the functional form is selected on goodness-of-fit grounds; see Spanos (2010).

Table 14.5: Gauss Linear (GL) model

$$Y_t{=}\alpha_0+\alpha_1^{\top}\mathbf{x}_t+\epsilon_t,\ t{\in}\mathbb{N}{:=}(1,2,...,n,...)$$

[1]	Normality:	$Y_t\backsim \mathsf{N}(.,.)$,	$t{\in}\mathbb{N}$.
[2]	Linearity:	$E\left(Y_t\right){=}\alpha_0+\alpha_1^{\top}\mathbf{x}_t$,	
[3]	Homoskedasticity:	$Var\left(Y_t\right){=}\sigma^2$,	
[4]	Independence:	$\{Y_t,\ t{\in}\mathbb{N}\}$ is independent,	
[5]	t-invariance:	$(\alpha_0,\boldsymbol{\alpha}_1,\sigma^2)$ are *not* changing with t,	

Table 14.6: Traditional Gauss Linear model

$$Y_t{=}\alpha_0+\alpha_1^{\top}\mathbf{x}_t+\epsilon_t,\ t{\in}\mathbb{N}{:=}(1,2,...,n,...)$$

{1}	Zero mean:	$E\left(\epsilon_t\right){=}0,\ \forall t{\in}\mathbb{N}$,	$t{\in}\mathbb{N}$.
{2}	Constant variance:	$E\left(\epsilon_t^2\right){=}\sigma^2>0,\ \forall t{\in}\mathbb{N}$,	
{3}	Zero covariance:	$E\left(\epsilon_t\epsilon_s\right){=}0,\ t{\neq}s,\ t,s{\in}\mathbb{N}$,	
{4}	Fixed $\{\mathbf{x}_t\}_{t=1}^n$:	$\{\mathbf{x}_t\}_{t=1}^n$ is fixed in repeated samples,	
	$\mathbf{X}{:=}\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)^{\top}$	such that $\mathsf{rank}(\mathbf{X}){=}p,\ p<n$.	

The curve-fitting differs from the statistical modeling perspective on a number of different dimensions the most important of which is their primary objective. In curve-fitting the functional form $g(.)$ is determined on mathematical approximation grounds and the primary criterion in selecting the fittest curve $\widehat{Y}_t{=}\widehat{g}(\mathbf{x}_t)$ is the 'smallness' of the residuals $\{\widehat{\epsilon}_t{=}\left(Y_t{-}\widehat{Y}_t\right),\ t{=}1,2,...,n\}$ measured in terms of the prespecified objective function. Least Squares (LS) amounts to minimizing:

$$\ell(a_0,\boldsymbol{\alpha}_1){=}\sum_{t=1}^{n}\left(Y_t{-}\alpha_0{-}\alpha_1^{\top}\mathbf{x}_t\right)^2\rightarrow\ \min_{(\boldsymbol{\alpha})}\ell(a_0,\boldsymbol{\alpha}_1){=}\sum_{t=1}^{n}\widehat{\epsilon}_t^2. \tag{21}$$

It should come as no surprise that this perspective gave rise to several new curve-fitting procedures, widely used in the *statistical learning* literature (Murphy, 2012), including the following.

(a) **Ridge regression**, whose objective function is (Tibshirani, 1996):

$$\widetilde{\boldsymbol{\alpha}}_{\text{ridge}}=\min_{\boldsymbol{\alpha}}\left[\sum_{t=1}^{n}\left(Y_t-\alpha_0-\alpha_1^{\top}\mathbf{x}_t\right)^2+\lambda\sum_{i=1}^{p}\alpha_i^2\right]\rightarrow\widetilde{\boldsymbol{\alpha}}_{\text{ridge}}=\left(\mathbf{X}^{\intercal}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\intercal}\mathbf{y}.$$

(b) **LASSO** (Least Absolute Shrinkage and Selection Operator) **regression**, whose objective function is:

$$\widetilde{\boldsymbol{\alpha}}_{\text{LASSO}}=\min_{\boldsymbol{\alpha}}\left[\sum_{t=1}^{n}\left(Y_t-\alpha_0-\alpha_1^{\top}\mathbf{x}_t\right)^2+\lambda\sum_{i=1}^{p}|\alpha_i|\right].$$

(c) **Elastic Net regression**, whose objective function is (Zou and Hastie, 2005):

$$\widetilde{\boldsymbol{\alpha}}_{\text{EN}}=\min_{\boldsymbol{\alpha}}\left[\sum_{t=1}^{n}\left(Y_t-\alpha_0-\alpha_1^{\top}\mathbf{x}_t\right)^2+\lambda_1\sum_{i=1}^{p}|\alpha_i|+\lambda_2\sum_{i=1}^{p}\alpha_i^2\right].$$

What all these curve-fitting procedures have in common is that their ultimate objective is to minimize the Mean Square Error of the resulting estimators $E(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha})^2$ because the constraints give rise to biased estimators; see chapter 11.

The above curve-fitting procedures should be contrasted with the statistical modeling perspective where $g(.)$ stems from the probabilistic structure of the stochastic process $\{\mathbf{Z}_t{:=}(\mathbf{X}_t,y_t),\ \ t{\in}\mathbb{N}\}$, the statistical parameterization $\boldsymbol{\theta}$ is of paramount importance, and the statistical model is selected on statistical adequacy grounds; it's the 'non-systematicity' of the residuals that matters and not their 'smallness' measured by an arbitrary distance function. Moreover, the statistical parameters need to be related to the substantive parameters stemming from a structural model.

Although the probabilistic assumptions {1}-{3} are directly related to assumptions [2]-[4] in table 14.5, the specification of the GL model in table 14.6 and that of the LR model in table 14.1 differ in a number of important respects:

(i) {1}-{3} pertain to the probabilistic structure of $\{Y_t,\ \ t{\in}\mathbb{N}\}$ as opposed to $\{(Y_t|\mathbf{X}_t{=}\mathbf{x}_t)\,,\ t{\in}\mathbb{N}\}$,

(ii) {1} pertains to the linearity in parameters (not the linearity in x_t),

(iii) [5] is clearly implicit in the specification in table 14.6, and

(iv) {4} pertains to the numbers $\{\mathbf{x}_t\}_{t=1}^{n}$ and the condition $\mathsf{rank}(\mathbf{X}){=}p$ is primarily a numerical issue; see section 5.

The traditional econometric literature considers the omission of the Normality assumption [1] as a major advantage of the specification in table 14.6 over that of table 14.5. The idea is that the assumptions {1}-{4} are weaker than [1]-[5] and thus the inference results are (i) more general as well as (ii) less vulnerable to statistical misspecification; a highly questionable claim (Spanos, 2018).

3.2 Gauss-Markov theorem

Under assumptions {1}-{4} in table 14.6, the Ordinary Least Squares (OLS) estimators:

$$\widehat{\alpha}_0{=}\overline{Y}-\hat{\beta}_1\bar{x},\quad\widehat{\alpha}_1{=}\frac{\sum_{t=1}^{n}(Y_t-\overline{Y})(x_t-\bar{x})}{\sum_{t=1}^{n}(x_t-\bar{x})^2},$$

of (α_0,α_1) are Best (minimum variance), among the class of Linear and Unbiased Estimators (BLUE).

Table 14.7: Gauss-Markov theorem conclusions

(i)	Best (relatively efficiency): $Var(\widehat{a}_i) \leq Var(\widetilde{\alpha}_i), i{=}0,1,$ for any $\widetilde{\alpha}_i$, $i{=}0,1$, that also satisfy (ii)-(iii) below:
(ii)	Linear: $\widetilde{\alpha}_0{=}\sum_{t=1}^{n} w_{0t}Y_t$, $\widehat{\alpha}_1{=}\sum_{t=1}^{n} w_{1t}Y_t$, where w_{it}, $t{=}1,..,n$, $i{=}0,1$ are constants,
(iii)	Unbiased: $E(\widetilde{a}_i){=}a_i$, $i{=}0,1$.

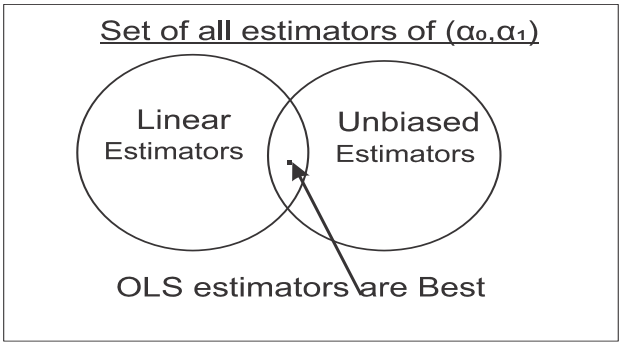

Fig. 14.13: Gauss-Markov theorem

That is, the OLS estimators $(\widehat{\alpha}_0, \widehat{\alpha}_1)$ are best (relatively more efficient) (figure 14.13):

$$Var(\widehat{\alpha}_i) \leq Var(\widetilde{\alpha}_i), \text{ for any other estimators } \widetilde{\alpha}_i,\ i{=}0,1,$$

within the class of Linear and Unbiased (LU) estimators of (α_0, α_1); see table 14.7 and fig. 14.13. Note that the OLS estimators are linear functions of $\mathbf{y}$ since:

$$\widehat{\alpha}_1{=}\sum_{t=1}^{n} \xi_t(Y_t{-}\overline{Y}),\quad \widehat{\alpha}_0{=}\overline{y}{-}\overline{x}\left(\sum_{t=1}^{n} \xi_t(Y_t{-}\overline{Y})\right),$$

where $\xi_t{=}[(x_t{-}\bar{x})/\sum_{t=1}^{n}(x_t{-}\bar{x})^2]$, $t{=}1,2,..,n$.

Although this is often celebrated as a major result in econometrics, a closer examination reveals that this is not anything to write home about. *First*, as argued in chapter 11, relative efficiency means very little if one restricts unnaturally the class of competing estimators; recall that I am the best econometrician in my family! The class of unbiased estimators, although interesting, it is too restrictive because numerous unbiased estimators are often useless unless they satisfy more potent optimal properties, such as full efficiency. Restricting the class of unbiased estimators further to only *linear* functions of $\mathbf{y}$ is entirely artificial because there is no intrinsic value in the linearity of an estimator. Linearity is a beguiling irrelevancy since without which the relative efficiency does *not* hold. *Second*, there is a degree of arbitrariness in minimizing:

$$\ell(\alpha_0, \alpha_1){=}\sum_{t=1}^{n} \left(Y_t{-}\alpha_0{-}\alpha_1 x_t\right)^2, \tag{22}$$

with respect to (α_0, α_1) that could be at odds with the probabilistic premises for statistical inference. The equivalence of the least-squares minimization to the *maximization* of the log-likelihood function:

$$\ln L(\boldsymbol{\theta}; \mathbf{y}){=}\quad \text{const}{-}\tfrac{n}{2}\ln\sigma^2{-}\tfrac{1}{2\sigma^2}\sum_{t=1}^{n}(Y_t{-}\alpha_0{-}\alpha_1 x_t)^2, \tag{23}$$

brings out an intrinsic affinity between (22) and the Normality assumption; ignored by the Gauss-Markov theorem. As argued by Pearson (1920), Normality is the only real justification for minimizing the squares of the errors:

"Theoretically therefore to have justification for using the method of least squares to fit a line or a plane to a swarm of points we must assume the arrays to follow a normal distribution. If they do not, we may defend least squares as likely to give a fairly good result but we cannot demonstrate its accuracy. Hence, in disregarding normal distributions and claiming great generality for our correlation by merely using the principle of least squares, we are really depriving that principle of the basis of its theoretical accuracy, and the apparent generalization has been gained merely at the expense of theoretical validity. " (ibid. (1920), p. 45)

In that sense the least-squares method will be very difficult to justify when the underlying distribution is non-Normal, since the Gauss-Markov theorem holds for all the distributions of the error in table 14.8.

Table 14.8: G-M theorem and admissible error distributions

(i) **Normal**:	$f(\epsilon_t)=\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{\epsilon_t^2}{2\sigma^2}),\ \epsilon_t\in\mathbb{R},\ Var(\epsilon_t)=\sigma^2,$
(ii) **Laplace**:	$f(\epsilon_t)=\frac{1}{2\sigma}\exp(-\frac{\lvert\epsilon_t\rvert}{\sigma}),\ \epsilon_t\in\mathbb{R},\ Var(\epsilon_t)=2\sigma^2,$
(iii) **Uniform**:	$f(\epsilon_t)=\frac{1}{2\sigma},\ -\sigma<\epsilon_t<\sigma,\ Var(\epsilon_t)=\frac{\sigma^2}{3},$
(iv) **Euler**:	$f(\epsilon_t)=\frac{3}{4\sigma^3}(\sigma^2-\epsilon_t^2),\ -\epsilon_t<\sigma<\epsilon_t,\ Var(\epsilon_t)=\frac{\sigma^2}{5}.$

However, the only case for which minimizing the least squares objective function $\ell(\alpha_0,\alpha_1)$ in (22) can be formally justified on statistical grounds is for (i) Normal, because it is equivalent to maximizing the log-likelihood function! On the other hand, minimizing $\ell(\alpha_0,\alpha_1)$ in (22) is in conflict with the MLE estimator because the implicit objective function is different in each case.

For instance, in the case of the *Laplace distribution* the natural objective function is not (22) but the Least Absolute Deviation (LAD) function:

$$\ell_1(\beta_0,\beta_1)=\textstyle\sum_{t=1}^{n}\left|Y_t-\beta_0-\beta_1x_t\right|. \tag{24}$$

Analogously, in the case of the *Uniform distribution*, the natural objective function to minimize is:

$$\ell_\infty(\beta_0,\beta_1)=\sup_{\epsilon_t\in[-\sigma,\sigma]}\left|Y_t-\beta_0-\beta_1x_t\right|. \tag{25}$$

As argued by Welsh (1996): "It [the Gauss-Markov theorem] omits to point out that, except at distributions which are very close to Gaussian, all linear unbiased estimators are poor ..." (p. 286).

The third and most crucial weakness of the Gauss-Markov theorem is that it has little to no value for inference purposes. It asserts that:

$$\widehat{\alpha}_0\backsim\overset{?}{\mathsf{D}}_0(\alpha_0,\sigma^2\left(\tfrac{1}{n}+\varphi_x\overline{x}^2\right)),\quad \widehat{\alpha}_1\backsim\overset{?}{\mathsf{D}}_1(\alpha_1,\sigma^2\varphi_x), \tag{26}$$

where $\varphi_x=[\sum_{t=1}^{n}(x_t-\overline{x})^2]^{-1}$ and $\overset{?}{\mathsf{D}}(.)$, $i=0,1$, denote the *unknown* (finite) sampling distributions. The only possible way to draw any inferences pertaining to (α_0,α_1) would be to use inequalities that use the first two moments, such as Chebyshev's:

$$\mathbb{P}(|\widehat{\alpha}_i-\alpha_i|>\varepsilon)\leq\tfrac{Var(\widehat{\alpha}_i)}{\varepsilon^2},\ i=0,1,\ \text{for any } \varepsilon>0,$$

to provide upper (or lower) bounds for the relevant error probabilities. As shown in chapter 9, however, these inequalities provide very crude approximations to the relevant error probabilities, leading to highly imprecise inferences in both interval estimation and testing.

HISTORICAL GOSSIP. Gauss (1821) abandoned the Normality assumption and proved what is called today the Gauss-Markov theorem. It is interesting to note that the credit to Andrei Markov (1856–1922) is misplaced. The paper that initially misnamed the theorem, was David and Neyman (1938). As Plackett (1949), however, argues: "Markoff, who refers to Gauss's work, may perhaps have clarified assumptions implicit there, but proved nothing new." (p. 460) Neyman (1952), corrected his mistake for the false attribution: "the theorem that I ascribed to Markoff was discovered by Gauss." (p. 228). Unfortunately, this retraction went unnoticed by the textbooks.

4 The LR model; prediction (forecasting)

The term *prediction* is used as synomymous to *forecasting* in what follows. The term forecasting has a temporal connotation that might not be relevant in cases where the ordering of interest $t\in\mathbb{N}$ does not refer to time.

Table 6: Normal, Linear Regression (LR) Model

Statistical GM: $Y_t=\beta_0+\boldsymbol{\beta}_1^\top\mathbf{x}_t+u_t,\ t\in\mathbb{N}$,

[1]	Normality:	$(Y_t\vert\mathbf{X}_t=\mathbf{x}_t)\backsim\mathsf{N}(.,.)$,	$t\in\mathbb{N}$.
[2]	Linearity:	$E(Y_t\vert\mathbf{X}_t=\mathbf{x}_t)=\beta_0+\boldsymbol{\beta}_1^\top\mathbf{x}_t$,	
[3]	Homosk/city:	$Var(Y_t\vert\mathbf{X}_t=\mathbf{x}_t)=\sigma^2$,	
[4]	Independence:	$\{(Y_t\vert\mathbf{X}_t=\mathbf{x}_t),\ t\in\mathbb{N}\}$ indep. process,	
[5]	t-invariance:	$\boldsymbol{\theta}:=(\beta_0,\boldsymbol{\beta}_1,\sigma^2)$ are *not* changing with t,	

$\boldsymbol{\beta}:=(\beta_0,\boldsymbol{\beta}_1)^\top,\ \beta_0=E(Y_t)-\boldsymbol{\beta}_1^\top E(\mathbf{X}_t),\ \boldsymbol{\beta}_1=[Cov(\mathbf{X}_t)]^{-1}Cov(\mathbf{X}_t,Y_t)$,
$\sigma^2=Var(Y_t)-Cov(\mathbf{X}_t,Y_t)^\top[Cov(\mathbf{X}_t)]^{-1}Cov(\mathbf{X}_t,Y_t)$

The statistical GM of the LR model (table 6) is supposed to represent the generating mechanism that could have given rise to data:

$$\mathbf{Z}_0:=\{(\mathbf{x}_t,y_t),\ t=1,2,...,n\}.$$

When the statistical adequacy of prespecified statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{y})=\{f(\mathbf{y}\vert\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\subset\mathbb{R}^m\},\ \mathbf{y}\in\mathbb{R}_Y^n,\ m<n, \tag{27}$$

is established, the clause 'it could have' is transformed into 'it has'. Indeed, one can use the estimated parameters $\boldsymbol{\theta}:=(\beta_0,\boldsymbol{\beta}_1,\sigma^2)$, say $\widehat{\boldsymbol{\theta}}:=(\widehat{\beta}_0,\widehat{\boldsymbol{\beta}}_1,s^2)$ to generate additional sampe realizations (simulated data) at will.

In light of the above, it makes sense to use the estimated statistical GM to predict values of Y_t beyond the observation period $t=1,2,...,n$, say Y_{n+k}:

$$Y_{n+k}=\beta_0+\boldsymbol{\beta}_1^{\top}\mathbf{x}_{1,n+k}+u_{n+k},\ k=1,2,...,h. \tag{28}$$

This suggests that an obvious way to predict y_{n+k} is to use the estimated statistical GM using the observations for $t=1,2,...,n$ is:

$$\widehat{Y}_{n+k}=\widehat{\beta}_0+\boldsymbol{\beta}_1^{\top}\mathbf{x}_{1,n+k}=\mathbf{x}_{n+k}^{\top}\widehat{\boldsymbol{\beta}},\ \text{for any } k\geq 1,$$

where $\widehat{\boldsymbol{\beta}}^{\top}:=(\widehat{\beta}_0,\widehat{\boldsymbol{\beta}}_1^{\top})$, $\widehat{\boldsymbol{\beta}}=(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{y}$ and $\mathbf{x}_{n+1}:=(1,\mathbf{x}_{1n+1})$, assuming that $\mathbf{x}_{n+k}$ has been observed.

Given that $\widehat{Y}_{n+1}$ is a random variable with a sampling distribution:

$$(\widehat{Y}_{n+k}|\mathbf{X}_t=\mathbf{x}_t)\backsim \mathsf{N}\left(\mathbf{x}_{n+k}^{\top}\boldsymbol{\beta},\ \sigma^2\left(\mathbf{x}_{n+k}^{\intercal}\left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\mathbf{x}_{n+k}\right)\right),$$

one can use the equality:

$$Y_{n+k}=\widehat{Y}_{n+k}+\widehat{u}_{n+k},\ \text{where } \widehat{Y}_{n+k}\perp\widehat{u}_{n+k}\ \text{for any } k\geq 1,$$

to define the prediction error $\widehat{u}_{n+k}$ to be:

$$\widehat{u}_{n+k}=(Y_{n+k}-\widehat{Y}_{n+k})=\mathbf{x}_{n+k}^{\top}\boldsymbol{\beta}+u_{n+k}-\mathbf{x}_{n+k}^{\top}\widehat{\boldsymbol{\beta}}=\mathbf{x}_{n+k}^{\top}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})+u_{n+k}.$$

This indicates that there are two sources of error in prediction. The first stems from u_{n+k}, the unknown σ^2, and the second from replacing the unknown coefficients $\boldsymbol{\beta}$ with their estimators $\widehat{\boldsymbol{\beta}}$. The sampling distribution of

$$(\widehat{u}_{n+k}|\mathbf{X}_t=\mathbf{x}_t)=(Y_{n+k}-\widehat{Y}_{n+k}|\mathbf{X}_t=\mathbf{x}_t)\backsim \mathsf{N}\left(0,\sigma^2\left(1+\mathbf{x}_{n+k}^{\intercal}\left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\mathbf{x}_{n+k}\right)\right),$$

since $Var(\widehat{u}_{n+k})=\sigma^2+\sigma^2\left(\mathbf{x}_{n+k}^{\intercal}\left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\mathbf{x}_{n+k}\right)$. A popular way to evaluate the size of the prediction error is to use the Root Mean Square Prediction error:

$$\text{RMSPE}\left(\widehat{u}_{n+k}\right)=\sqrt{E(Y_{n+k}-\widehat{Y}_{n+k})^2}=\sigma\sqrt{1+\mathbf{x}_{n+k}^{\intercal}\left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\mathbf{x}_{n+k}}.$$

This can be used to construct the pivotal quantity (Y_{n+k} is unobserved):

$$\frac{\left(Y_{n+k}-\widehat{Y}_{n+k}\right)}{s\sqrt{1+\mathbf{x}_{n+k}^{\intercal}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{x}_{n+k}}}\overset{\boldsymbol{\theta}=\boldsymbol{\theta}^*}{\backsim}\mathsf{St}(n-p), \tag{29}$$

where $s^2=\frac{1}{n-p}\sum_{t=1}^{n}\widehat{u}_t^2$ is the estimator of σ^2. This gives rise to a $(1-\alpha)$ two-sided Prediction Interval (PI) for Y_{n+k}:

$$\mathbb{P}\left(\widehat{Y}_{n+k}-c_{\frac{\alpha}{2}}s\sqrt{1+\mathbf{x}_{n+k}^{\intercal}\left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\mathbf{x}_{n+k}}\leq Y_{n+1}<\widehat{Y}_{n+k}+c_{\frac{\alpha}{2}}s\sqrt{1+\mathbf{x}_{n+k}^{\intercal}\left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\mathbf{x}_{n+k}}\right)=1-\alpha. \tag{30}$$

Example 14.2. Consider the data in table 1 in Appendix 14.C (Mendhall and Sinchich, 1996, p. 184), where z_{1t}-the auction final price and z_{2t}-the age of an antique grandfather clock in a sequence of n=32 such transactions. The assumed regression model is of the form given in table 14.1, where Y_t=ln(Z_{1t}) and X_t=ln(Z_{2t}). The idea is to account for the final auction price using the age of the antique clock as the explanatory variable. The estimated linear regression using the data in table 1 (Appendix 14.C) yields:

$$Y_t = \underset{(.966)}{1.312} + \underset{(.195)}{1.177} x_t + \widehat{u}, \quad s=.208, \ n=32. \tag{31}$$

The data on the age of the clock range from 108 to 194 years old. One could use the estimated regression in (31) to construct a Prediction Interval (PI) for the value $z_{2,n+1}$=100 year old antique clock, with α=.05 ($c_{\frac{\alpha}{2}}$=2.042). The observed PI is:

$$\widehat{Y}_{n+1} \pm 2.042(.208)\sqrt{\left(1+\tfrac{1}{32}+.870\,(4.605-4.958)^2\right)} = 6.732 \pm .453,$$

which is reasonably precise. NOTE that the predicted price for a $z_{2,n+1}$=100 year old antique clock is: $z_{1,n+1}$=exp(6.732)=838.82 $\pm$ 1.573=exp(.453).

4.1 Several post-data predictions

In evaluating the predictive capacity of an estimated LR model (28) is used to predict h periods ahead, one can use:

$$\text{MSPE}\,(h) = \textstyle\sum_{k=1}^{h} E(\widehat{u}_{n+k}^2),$$

which can be estimated using:

$$\widehat{\text{MSPE}}\,(h) = \tfrac{1}{h} \textstyle\sum_{k=1}^{h} \widehat{u}_{n+k}^2. \tag{32}$$

Although the sampling distribution of this statistic is rather complicated, it is often approximated, using:

$$\sum_{k=1}^{h} \frac{(Y_{n+k}-\widehat{Y}_{n+k})^2}{s^2\left(1+\mathbf{x}_{n+k}^{\intercal}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{x}_{n+k}\right)} \underset{n\to\infty}{\overset{\boldsymbol{\theta}=\boldsymbol{\theta}^*}{\backsim}} \chi^2(h).$$

The above evaluating measures are often used to compare the predictive ability of different models.

CAUTION: comparing statistically misspecified statistical models on prediction grounds makes little sense because the above statistics assume the statistical adequacy of such models at the outset.

4.2 The traditional approach

Traditional goodness-of-prediction for the post-sample period, $n+1, n+2, ..., n+h$, is often evaluated using some form of (32). The justification used that the MSPE stems from minimizing the expected value of a quadratic loss function of the form:

$$E[L_2(Y_{n+k}, \widehat{Y}_{n+k})]=E(Y_{n+k}-\widehat{Y}_{n+k}|\mathbf{X}))^2,$$

and by invoking **CE4. The best least-squares prediction property:**

$$E\left[Y-E(Y|\sigma(X))\right]^2 \leq E\left[Y-g(X)\right]^2 \text{ for any Borel functions } g(.).$$

Hence, $\widehat{Y}_{n+k}=\mathbf{x}_{n+k}^{\top}\widehat{\boldsymbol{\beta}}$ represents the estimated regression function $E(Y|\sigma(X))$.

This result is often detached from the statistical model and the Normality assumption underlying its likelihood function and traditional goodness-of-prediction is evaluated relative to **loss functions** stemming from information extraneous to the data and the underlying probabilistic structure of the statistical model in question. As argued by Elliot and Zimermann (2016): **"We regard loss functions a realistic exposition of the forecaster's objectives, and consider the specification of the loss function as an integral part of the forecaster's decision problem."** (p. 4). Examples of loss functions include the ones in table 2.

Table 2: Loss functions for prediction evaluation

Square:	$L_2(e) = ae^2,\ a>0,$
Absolute value:	$L_1(e) = a\|e\|,\ a>0,$
Pairwise linear:	$L(e)=\begin{cases} -a(1-\gamma)e & \text{if } e \leq 0 \\ a\gamma e & \text{if } e > 0 \end{cases},\ a>0,$
Linex:	$L_\ell(e)=a_1(\exp(a_2e)-(a_2e)-1),\ a_1>0,\ a_2\neq 0,$
Asymmetric square:	$L_2(e)=\begin{cases} (1-a)e & \text{if } e \leq 0 \\ ae^2 & \text{if } e > 0 \end{cases}.$

Moreover, prediction (forecasting) is traditionally viewed as primarily **a decision theoretic problem** and not a model-based frequentist inference. Lehmann (1984) warned about the perils of arbitrary loss functions:

"It is argued that the choice of a loss function, while less crucial than that of the model, exerts an important influence on the nature of the solution of a statistical decision problem, and that an arbitrary choice such as squared error may be baldly misleading as to the relative desirability of the competing procedures." (p. 425).

For further discussion on this issue see Spanos, A. (2017) "Why the Decision-Theoretic PerspectiveMisrepresents Frequentist Inference", chapter 1, pp. 3-28, *Advances in Statistical Methodologies and Their Applications to Real Problems*, ISBN 978-953-51-4962-0.

4.3 The Normal, Autoregressive model

The above results on prediction can be extended to the AR(1) model with minor modifications.

Table 7: Normal, AutoRegressive [AR(1)] model

Statistical GM:		$Y_t=\alpha_0+\alpha_1 Y_{t-1}+u_t,\ t\in\mathbb{N}.$	
[1]	Normality:	$(Y_t,Y_{t-1})\backsim \mathsf{N}(.,.),$	$\left.\begin{array}{l}\\\\\\\\\end{array}\right\} t\in\mathbb{N}.$
[2]	Linearity:	$E\left(Y_t\vert\sigma(Y_{t-1})\right)=\alpha_0+\alpha_1 Y_{t-1},$	
[3]	Homoskedasticity:	$Var\left(Y_t\vert\sigma(Y_{t-1})\right)=\sigma_0^2,$	
[4]	Markov:	$\{Y_t,\ t\in\mathbb{N}\}$ is a Markov process,	
[5]	t-invariance:	$(\alpha_0,\alpha_1,\sigma_0^2)$ are *not* changing with t,	
$\alpha_0=\mu(1-\alpha_1)\in\mathbb{R},\ \alpha_1=\frac{\sigma(1)}{\sigma(0)}\in(-1,1),\ \sigma_0^2:=\sigma(0)(1-\alpha_1^2)\in\mathbb{R}_+.$			

Using the statistical GM on can argue that:

$$Y_{n+k}=\alpha_0+\alpha_1 Y_{n+k-1}+u_{n+k},\ k=1,2,...,h.$$

Hence, the natural predictor of Y_{n+k} takes the form:

$$\widehat{Y}_{n+k}=\widehat{\alpha}_0+\widehat{\alpha}_1 Y_{n+k-1},\ k=1,2,...,h,$$

which gives rise to the prediction errors:

$$\widehat{u}_{n+k}=(Y_{n+k}-\widehat{Y}_{n+k})=[(\alpha_0-\widehat{\alpha}_0)+(\alpha_1-\widehat{\alpha}_1)Y_{n+k-1}]+u_{n+k},\ k=1,2,...,h,$$

associated with the estimation of the coefficients and the unknown σ_0^2, respecti ely.

The key difference with the LR model is that for $k=1$, y_n belongs to the observation period, allowing for the possibility that one can forecast Y_{n+1} with data for $t=1,2,...,n$. In turn, this allows for the possibility that one can use the predicted value $\widehat{Y}_{n+1}$ in conjunction with the statistical GM to predict Y_{n+k} for $k>1$:

$$\widehat{Y}_{n+k}=\widehat{\alpha}_0+\widehat{\alpha}_1\widehat{Y}_{n+k-1},\ k=2,3,...,h.$$

This is known as recursive prediction that is particularly useful for post-estimation prediction because no additional observations beyond y_t, $t=1,2,...,n$, are needed.

4.3.1 How should predictive capacity be evaluated?

It is important to emphasize that, like all other forms of inference, prediction depends crucially on the statistical adequacy of the statistical model in question. In cases where the estimated statistical model:

$$\mathcal{M}_{\widehat{\boldsymbol{\theta}}}(\mathbf{z})=\{f(\mathbf{z};\widehat{\boldsymbol{\theta}})\},\ \mathbf{z}\in\mathbb{R}_Z^n,$$

accounts for all the systematic information in the data, represents an adequate description of the stochastic mechanism that gave rise to the particular data $\mathbf{z}_0$. Hence, $\mathcal{M}_{\widehat{\boldsymbol{\theta}}}(\mathbf{z})$ should give rise to prediction errors that are non-systematic in the same way its residuals are non-systematic for the observation period, unless the actual generating mechanism has changed after n.

The traditional **goodness-of-prediction** evaluation based on loss functions ignores the fact that the MSPE is a reliable measure of predictive accuracy only when the prediction errors $\widehat{u}_{n+k},\ k=1,2,...,h$, are non-systematic. Statistically, non-systematic prediction errors $\widehat{u}_{n+k},\ k=1,2,...,h$, are viewed as a realization of a martingale difference process. When the estimated statistical model over-predicts or under-predicts systematically, it should be considered inaccurate on prediction grounds. the terms 'checked and not tested is used to allow for the possibility that the prediction period is often too small for proper testing and one uses the plot of the prediction errors to assess non-systematicity. In this sense, goodness-of-prediction is not 'small' in some loss function sense but statistically non-systematic errors. In practice, it is expected that $\mathcal{M}_{\widehat{\boldsymbol{\theta}}}(\mathbf{z})$ represents an adequate description of the stochastic generating mechanism, not only for the within sample data, but also beyond the sample period.

4.4 Comparing statistical models on predictive grounds

When prediction (forecasting) is a crucial inference for a statistical model, one might want to compare several models for their prediction accuracy. In such cases the modeler might decide to leave out of the sample data the last few observations to assess the prediction accuracy of different models. For instance, for n=120, leave out the last 20 to use for evaluating the prediction reliability of different models.

CAUTION: in the traditional literature empirical modeling is often viewed as **curve-fitting guided by goodness-of-fit criteria** and thus it is often the case that several different models are estimated that are more or less equivalent on goodness-of-fit grounds and one would like to make a choice among these estimated models on other grounds, such as predictive accuracy. As argued above, excellent goodness-of-fit is neither necessary nor sufficient for statistical adequacy. Hence, comparisons among statistically misspecified models to choose the 'best' on prediction grounds should be avoided because it can easily give rise to untrustworthy evidence. For instance, Enders (2004), p. 83, suggests comparing an ARMA(2,1) with AR(1) model on prediction grounds without checking their statistical adequacy first.

Are there any cases where one has to choose between two statistically adequate models on other grounds? The answer is affirmative because there are cases where two different parameterizations of the same underlying stochastic process can be equally acceptable on statistical adequacy grounds. An example of this are the following two estimated statistical models that represent two parameterizations of a Normal,

Markov but separable heterogeneous process $\{y_t,\ t\in\mathbb{N}\}$:

$$\text{UR(1): } \Delta y_t = \underset{(.250)}{.799}+\widetilde{\varepsilon}_t,\ \widetilde{\sigma}^2=1.472,\ n=100, \tag{33}$$

$$\text{AR(1): } y_t = \underset{(.605)}{.846}+\underset{(.041)}{.084}t+\underset{(.056)}{.900}y_{t-1}+\widehat{u}_t,\ \widehat{\sigma}^2=1.3957,\ n=100. \tag{34}$$

(33) is a UR(1) parameterization with only a constant and (34) is an AR(1) parameterization with a constant and a trend. These two estimated models can be compared on prediction grounds using their respective MSPE:

$$\widehat{\text{MSPE}}_1(h)=\tfrac{1}{h}\sum_{k=1}^{h}\widetilde{\varepsilon}_{n+k}^2,\quad \widehat{\text{MSPE}}_2(h)=\tfrac{1}{h}\sum_{k=1}^{h}\widehat{u}_{n+k}^2.$$

The most widely used formal comparison between the two models is based on the F-test:

$$F(\mathbf{y})=\frac{\sum_{k=1}^{h}\widehat{u}_{n+k}^2}{\sum_{k=1}^{h}\widetilde{\varepsilon}_{n+k}^2}\overset{H_0}{\backsim}\mathsf{F}(h,h),\ C_1(\alpha)=\{\mathbf{y}\text{: } F(\mathbf{y})>c_\alpha\},\ \alpha=\int_{c_\alpha}^{\infty}f(v)dv, \tag{35}$$

where the larger of the two $\widehat{\text{MSPE}}$'s is selected to be the numerator and null hypothesis is that the two models are equally good on prediction grounds; see Enders (2004), p. 84.

CAUTION: a closer look at the F-test in (35) reveals that the prediction capacity of the two models is in terms of their post data variances:

$$H_0\text{: } \sigma_u^2=\sigma_\varepsilon^2 \text{ vs. } H_0\text{: } \sigma_u^2>\sigma_\varepsilon^2,$$

based on $\widehat{\sigma}_u^2=\frac{1}{h}\sum_{k=1}^{h}\widehat{u}_{n+k}^2$ and $\widehat{\sigma}_\varepsilon^2=\frac{1}{h}\sum_{k=1}^{h}\widetilde{\varepsilon}_{n+k}^2$. For these two estimators to be accurate enough on needs a large enough h and $(\widehat{u}_{n+k},\ \widetilde{\varepsilon}_{n+k})$ need to be statistically non-systematic, i.e. they represent realizations of martingale difference processes. When these prediction errors exhibit over or under prediction for several periods, this assumption is likely to be invalid.

5 Conclusions

This chapter brings out three important distinctions that can help to elucidate the modeling and inference for the Linear Regression (LR) model: (a) the statistical vs. the substantive information/model, (b) the modeling vs. the inference stages, and (c) statistical modeling vs. curve-fitting.

A statistical model aims to account for the chance regularities in the data, and a substantive (structural) model aims to explain the phenomenon of interest giving rise to this data. The two models are related via the parameterization chosen for the statistical model. The two models are ontologically distinct and have very different objectives, rendering the criteria for evaluating their adequacy completely different; see Spanos and Mayo (2015). However, probing for substantive adequacy requires statistical adequacy to ensure the reliability of the statistical procedures employed. The main reason why the distinction between the two models is often blurred is because a theory-driven empirical modeling considers statistical modeling as curve-fitting: (i) selecting a substantive model (a family of curves), and (ii) choose the estimated model that best fits the data. The traditional viewpoint is: "Econometrics is concerned with the estimation of relationships suggested by economic theory" (Harvey, 1990, p. 1). As argued above, the curve-fitting mathematical framework provides an inadequate basis for reliable *inductive inference*. For that one needs to recast the approximation problem into one of modeling the 'systematic information' in the data, i.e. embed the substantive model into *a statistical model*. This enables one to address *statistical adequacy* first with a view to secure the reliability of inference pertaining to substantive adequacy. The basic conflict between the statistical and curve-fitting perspectives is that the former requires the residuals from an estimated model to be *non-systematic* (a martingale difference), i.e. the estimated model accounts for all the statistical information in the data (statistical adequacy), but the latter selects a fitted curve based on *small* residuals; best-fit criterion.

Using the above distinctions, the discussion compared and contrasted the Linear Regression (LR) with the Gauss Linear (GL) model, emphasizing the fact that what matters for inductive inference purposes are the inductive premises (the probabilistic assumptions) underlying the two models, and not the algebra and the formulae for estimators and tests. A case is made that the scope of the celebrated Gauss-Markov theorem is much too narrow, and its conclusions are of little value for inference purposes. When the primary objective is learning from data, there is no premium for weaker but non-testable premises. Such a strategy sacrifices the reliability and precision of inference at the altar of wishful thinking that relies on non-validated asymptotics ($n\to\infty$).