PHIL 6334/Econ6614 - Probability/Statistics
Lecture Notes 8: Mis-Specification (M-S) Testing

Aris Spanos [Spring 2019]

1 Introduction

The primary objective of **empirical modeling** is 'to learn from data' about observable stochastic phenomena of interest using a **statistical model**:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,\ \text{for } \boldsymbol{\theta}\in\Theta\subset\mathbb{R}^m,\ m<n, \tag{1}$$

where $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$ denotes the (joint) distribution of the *sample* $\mathbf{X}:=(X_1,...,X_n)$ that encapsulates the prespecified *probabilistic structure* of the underlying stochastic process $\{X_t,\ t\in\mathbb{N}:=(1,2,...,n,...)\}$. The **link** between $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and the phenomenon of interest comes in the form of viewing data $\mathbf{x}_0:=(x_1,x_2,...,x_n)$ as a *truly typical realization* of the process $\{X_k,\ k\in\mathbb{N}\}$. The 'typicality' of $\mathbf{x}_0$ can – and should – be assessed using trenchant **Mis-Specification (M-S) testing**. Appraising the validity of the probabilistic assumptions of the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ vis-a-vis data $\mathbf{x}_0$ is of paramount importance in practice, because without it the *reliability of inference* is at best dubious.

What distinguishes the frequentist approach to statistical modeling and inference from other approaches like the Bayesian and Decision theoretic are the following features:

[a] Mathematical probability is interpreted in terms of stable relative frequencies that underlie the *frequency interpretation*, anchored on the Strong Law of Large Numbers (SLLN).

[b] The chance regularities exhibited by data $\mathbf{x}_0$ constitute the *only relevant statistical information* for selecting the probabilistic structure of the stochastic process $\{X_t,\ t\in\mathbb{N}\}$ underlying the statistical model. *Substantive* information plays a role in selecting the parametrization of the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, as well as placing theoretical restrictions on statistical parameters $\boldsymbol{\theta}$, by relating the statistical ($\boldsymbol{\theta}$) and substantive ($\boldsymbol{\varphi}$) parameters, say $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$. In practice, these substantive restrictions need to be tested before imposed.

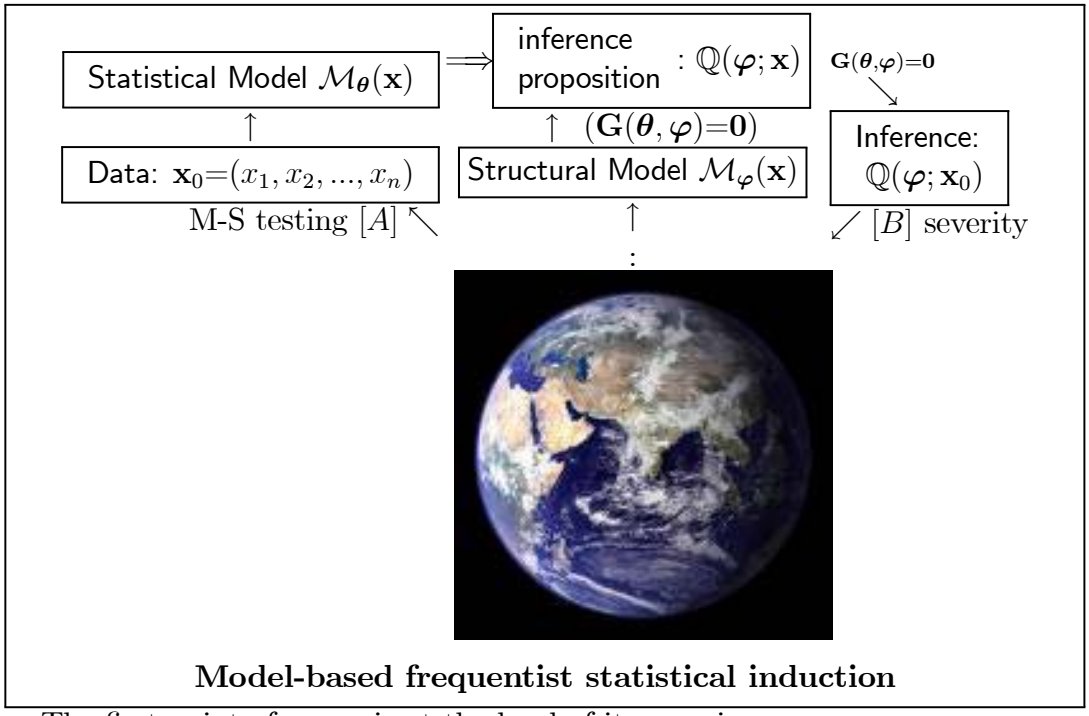
[c] The primary aim of the frequentist approach is to learn from data $\mathbf{x}_0$ about the true statistical Data-Generating Mechanism (DGM):

$$\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\},\ \mathbf{x}\in\mathbb{R}^n_X.$$

This is achieved by employing reliable inference procedures, framed in terms of the unknown parameter(s) $\boldsymbol{\theta}$, that are based on **ascertainable error probabilities**.

1.1 Statistical induction in a nutshell

Probability theory provides the mathematical foundations and the overarching framework for modeling observable stochastic phenomena of interest. The *modus operandi* of this modeling is the concept of a **statistical model** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ that mediates between the data $\mathbf{x}_0=(x_1, x_2, ..., x_n)$ and the real-world phenomenon of interest. A statistical model links up to the real-world phenomenon of interest at two different levels [A] and [B].



Model-based frequentist statistical induction

The first point of nexus is at the level of its premises:

[A] **From a phenomenon of interest to a statistically adequate model**; see fig. 1.7. The probabilistic assumptions specifying the statistical model (its premises) are chosen so that the observed data $\mathbf{x}_0$ constitute a **truly typical realization** of the stochastic process $\{X_t,\ t\in\mathbb{N}\}$ defined by these assumptions. Assessing the appropriateness of this choice takes the form of trenchant **M-S testing**. The validity of these assumptions secures the soundness of the model's premises and renders the second level of contact, the inference, germane to **learning from data** $\mathbf{x}_0$ about the real-world phenomenon of interest.

[B] **From the inference results to the substantive questions of interest**; see figure. This nexus point raises issues like statistical vs. substantive significance and how one assesses substantive adequacy; how adequate the structural (substantive) model is in shedding light (explain, illuminate, predict) the phenomenon of interest. As shown in chapter 14, these issues can be addressed using the post-data severity evaluation of the accept/reject rules of testing by establishing the discrepancy from the null warranted by data $\mathbf{x}_0$ and test T_α.

Statistical inference is often viewed as the quintessential form of **inductive inference**: learning form a particular set of data $\mathbf{x}_0$ about the stochastic phenomenon that gave rise to the data. However, it is often insufficiently recognized that this inductive procedure is embedded in a fundamentally **deductive argument**:

$$\text{If } \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}), \quad \text{then} \quad \mathbb{Q}(\mathbf{x})$$

The procedure from $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ (the premise) to $\mathbb{Q}(\mathbf{x})$ (the inference propositions – estimation, testing, prediction, policy simulation) is *deductive*; no data are used to derive results on the optimality of estimators, tests etc.; estimators and tests are pronounced *optimal* based on a purely deductive reasoning: if $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ then $\mathbb{Q}(\mathbf{x})$. In this sense, the reliability (soundness) of statistical inference depends crucially on **the validity of the premises**.

On the basis of this premise $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ we proceed to derive statistical inference results $\mathbb{Q}(\mathbf{x}_0)$ using a deductively valid argument ensuring that **if the premises are valid**, the conclusions are necessarily (statistically) reliable. To secure the **soundness** of such results one need to establish the adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ vis-a-vis data $\mathbf{x}_0$. By the same token, if $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is misspecified the inference results $\mathbb{Q}(\mathbf{x}_0)$ are generally untrustworthy. Hence, the problem of securing statistical adequacy if of paramount importance. Indeed, the ampliative (going beyond the premises) dimension of statistical induction relies on statistical adequacy to render the specific information in the form of data $\mathbf{x}_0$ pertinent to the stochastic phenomenon of interest; it is the cornerstone of inductive reasoning. Often the substantive questions of interest are framed in the context of the substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ that is parametrically nested within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ via the restrictions $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$. When the substantive parameters $\boldsymbol{\varphi}$ are uniquely defined as functions of $\boldsymbol{\theta}$, one can proceed to derive inferential propositions pertaining to $\boldsymbol{\varphi}$, including the sampling distribution of the estimator of $\boldsymbol{\varphi}$, and the error probabilities associated with different procedures. Such inferential results can be used to test substantive questions of interest, including the empirical validity of particular theories.

1.2 Statistical adequacy

When any of the model assumptions are invalid, the unreliability of inference might take several forms, including **inconsistent estimators**, and **sizeable discrepancies** between the *nominal (assumed) and actual error probabilities*.

► Rejecting a null hypothesis at a nominal α=.05, when the actual type I error probability is closer to .90, provides the surest way for an erroneous inference!

▼ It is important to note that all statistical methods (**frequentist, Bayesian, nonparametric**) rely on an underlying statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, and thus they are equally vulnerable to statistical misspecification.

What goes wrong when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is statistically misspecified? Since the likelihood function is defined via the distribution of the sample:

$$L(\theta;\mathbf{z}_0) \propto f(\mathbf{x}_0;\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta,$$

invalid $f(\mathbf{z};\boldsymbol{\theta}) \rightarrow$ invalid $L(\theta;\mathbf{z}_0) \Rightarrow$ { **Frequentist inference**: incorrect error probabilities, incorrect goodness-of-fit, inconsistent estimators. **Bayesian inference**: erroneous posterior distributions: $\pi(\theta|\mathbf{z}_0) \propto \pi(\theta)L(\theta;\mathbf{z}_0)$

What about nonparametric statistics? is equally vulnerable to statistical misspecification because their inferences also revolve around distribution free statistical models that include highly restrictive dependence and heterogeneity assumptions, often IID! The only difference with parametric statistics is that they replace testable distribution assumptions with untestable indirect distributional assumptions.

What about Akaike-type model selection procedures? They are even more vulnerable to statistical misspecification because they rely on the estimated likelihood function as a goodness-of-fit measure, but when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is misspecified, one is using an erroneous likelihood function!

CAUTION: do not confuse *model selection* with *model validation*!

Error statistics, viewed as a refinement and extension of the Fisher-Neyman-Pearson frequentist statistics proposes a methodology on how to specify (**Specification**) and validate statistical models by probing model assumptions (**Mis-Specification (M-S) testing**), isolate the sources of departures, and account for them in a respecified model (**Respecification**) with a view to secure statistical adequacy. Such a model is then used to probe the **substantive hypotheses** of interest.

Model validation based on M-S testing plays a pivotal role in providing an ***objective scrutiny*** of the reliability of inductive procedures and the trustworthiness of the resulting evidence.

1.3 Misspecification and the unreliability of inference

Before we discuss M-S testing, it is important to see how particular departures from the model assumptions can affect the reliability of inference by distorting the nominal error probabilities and rendering them non-ascertainable.

Table 1 - The simple (one parameter) **Normal model**

Statistical GM:		$X_t=\mu+u_t,\ t\in\mathbb{N}$,
[1]	Normal:	$X_t \backsim \mathsf{N}(.,.)$,
[2]	Constant mean:	$E(X_t)=\mu$, for all $t\in\mathbb{N}$,
[3]	Constant variance:	$Var(X_t)=\sigma^2$-*known*, for all $t\in\mathbb{N}$,
[4]	Independence:	$\{X_t,\ t\in\mathbb{N}\}$ - independent process.

To simplify the discussion that follows, let us focus on the **simple Normal** (one parameter) **model** (table 1).

It was shown above that for testing the *hypotheses*:

$$H_0\text{: } \mu{=}\mu_0 \quad \text{vs.} \quad H_1\text{: } \mu > \mu_0, \tag{2}$$

there is an α-level UMP defined by: $T_\alpha{:=}\{d(\mathbf{X}), C_1(\alpha)\}$:

$$\boxed{d(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}, \qquad C_1(\alpha){=}\{\mathbf{x}\text{: } d(\mathbf{x}) > c_\alpha\},} \tag{3}$$

where $\overline{X}_n{=}\frac{1}{n}\sum_{i=1}^n X_i$, c_α is the threshold rejection value. Given that:

$$\text{(i) } d(\mathbf{X}){=}\tfrac{\sqrt{n}\left(\overline{X}_n-\mu_0\right)}{\sigma} \overset{\mu=\mu_0}{\backsim} \mathsf{N}(0,1), \tag{4}$$

one can evaluate the *type I error probability* (significance level) c_α using:

$$\mathbb{P}(d(\mathbf{X}) > c_\alpha;\ H_0 \text{ true}){=}\alpha,$$

where α is the type I error. To evaluate the *type II error probability* and the power one needs to know the sampling distribution of $d(\mathbf{X})$ when H_0 is false. However, since H_0 is false refers to $H_1 : \mu > \mu_0$, this evaluation will involve all values of μ greater than μ_0 (i.e. $\mu_1{>}\mu_0$) :

$$\left.\begin{array}{ll} \beta(\mu_1) & {=}\mathbb{P}(d(\mathbf{X}) \leq c_\alpha;\ \mu{=}\mu_1), \\ \pi(\mu_1) & {=}1{-}\beta(\mu_1){=}\mathbb{P}(d(\mathbf{X}) > c_\alpha;\ \mu{=}\mu_1) \end{array}\right\} \forall(\mu_1{>}\mu_0)$$

The relevant sampling distribution takes the form:

$$\text{(ii) } d(\mathbf{X}){=}\tfrac{\sqrt{n}\left(\overline{X}_n-\mu_0\right)}{\sigma} \overset{\mu=\mu_1}{\backsim} \mathsf{N}(\delta_1,1),\ \delta_1{=}\tfrac{\sqrt{n}(\mu_1-\mu_0)}{\sigma},\ \forall\mu_1{>}\mu_0. \tag{5}$$

What is often insufficiently emphasized in statistics textbooks is that the above **nominal error probabilities**, i.e. the significance α, as well as the power of test T_α, will be different from the **actual error probabilities** when any of the assumptions [1]-[4] are invalid for data $\mathbf{x}_0$. Indeed, such departures are likely to create significant discrepancies between the *nominal* and *actual* error probabilities that often render inferences based on (3) **unreliable**.

To illustrate how the nominal and actual error probabilities can differ when any of the assumptions [1]-[4] are invalid, let us take the case where the *independence assumption* [4] is false for the underlying process $\{X_t,\ t{\in}\mathbb{N}\}$, and instead:

$$Corr(X_i, X_j){=}\rho,\ 0{<}\rho{<}1,\ \text{for all } i{\neq}j,\ i,j{=}1,...n. \tag{6}$$

How does such a misspecification affect the reliability of test T_α?

When (6) holds, the sampling distribution of $\overline{X}_n$ will be affected in the sense that although $E(\overline{X}_n){=}\mu$, its variance will change because of the presence of correlation among the X_i's ($Cov(X_i, X_j){=}\rho\sigma^2,\ i{\neq}j,\ i,j{=}1,...n$). In particular:

$$Var(\overline{X}_n) \ {=}\tfrac{1}{n^2}\left(\sum_{i=1}^n Var(X_i){+}2\sum_{j=i+1}^n Cov(X_i,X_j)\right){=}\tfrac{\sigma^2 d_n(\rho)}{n},$$

$$d_n(\rho)=(1+(n-1)\rho)>1 \text{ for } 0<\rho<1 \text{ and } n>1.$$

That is, the sampling distribution of $\overline{X}_n$ when (6) holds is:

$$\overline{X}_n \overset{0<\rho<1}{\backsim} \mathsf{N}\left(\mu, \tfrac{\sigma^2 d_n(\rho)}{n}\right), \text{ for } n>1.$$

Hence, the *actual* distributions of $d(\mathbf{X})$ under H_0 and H_1 are:

$$\begin{aligned} &\text{(i)* } d(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma} \overset{\mu=\mu_0}{\backsim} \mathsf{N}\left(0, d_n(\rho)\right), \\ &\text{(ii)* } d(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma} \overset{\mu=\mu_1}{\backsim} \mathsf{N}\left(\tfrac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}, d_n(\rho)\right) \end{aligned} \tag{7}$$

How does this change affect the relevant error probabilities?

Example 1. Consider the case: α=.05 (c_α=1.645), σ=1 and n=100. To find the **actual *type I error probability*** we need to evaluate the tail area of the distribution in (i)* beyond c_α=1.645:

$$\alpha^*=\mathbb{P}(d(\mathbf{X})>c_\alpha; H_0)=\mathbb{P}(Z>\tfrac{1.645}{\sqrt{d_n(\rho)}}; \mu=\mu_0),$$

where $Z \backsim \mathsf{N}(0,1)$. The results in table 2 for different values of ρ indicate that test T_α has now become 'unreliable' because $\alpha^*>\alpha$. One will apply test T_α thinking that it will reject a true H_0 only 5% of the time, when, in fact it is much higher.

Table 2 - Type I error of T_α when $Corr(X_i, X_j)$=ρ

ρ	.0	.05	.1	.2	.3	.5	.75	.8	.9
α^*	.05	.249	.309	.359	.383	.408	.425	.427	.431

The **actual power** should now be evaluated using:

$$\pi^*(\mu_1) \quad =\mathbb{P}(Z>(1/\sqrt{d_n(\rho)})\left[c_\alpha-\tfrac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}\right]; \mu=\mu_1),$$

giving rise to the results in table 3.

Table 3 - Power $\pi^*(\mu_1)$ **of** T_α when $Corr(X_i, X_j)$=ρ

ρ	$\pi^*(.01)$	$\pi^*(.02)$	$\pi^*(.05)$	$\pi^*(.1)$	$\pi^*(.2)$	$\pi^*(.3)$	$\pi^*(.4)$
.0	.061	.074	.121	.258	.637	.911	.991
.05	.262	.276	.318	.395	.557	.710	.832
.1	.319	.330	.364	.422	.542	.659	.762
.3	.390	.397	.418	.453	.525	.596	.664
.5	.414	.419	.436	.464	.520	.575	.630
.8	.431	.436	.449	.471	.515	.560	.603
.9	.435	.439	.452	.473	.514	.556	.598

For small values of μ_1 (.01, .02, .05.1), the power *increases* as $\rho \to 1$, but for larger values of μ_1 (.2, .3, .4), the power *decreases*, ruining the 'probativeness' of a test! It has become like a *defective* smoke alarm which has the tendency to go off when burning toast, but it will not be triggered by real smoke until the house is fully ablaze; Mayo (1996).

1.4 On the reluctance to validate statistical models

The key reason why model validation is extremely important is that No **trustworthy evidence *for*** or ***against*** a substantive claim (or theory) can be secured on the basis of a statistically misspecified model.

In light of this, 'why has model validation been neglected? There are several reasons, including the following.

(1) Inadequate appreciation of the serious implications of statistical misspecification for the reliability of inference.

(2) Inadequate understanding of how one can secure statistical adequacy using thorough M-S testing.

(3) Inadequate understanding M-S testing and confusion with N-P testing render it vulnerable to charges like: (i) infinite regress and circularity, and (ii) illicit double-use of data.

(4) Undue reliance on asymptotic arguments as $n \to \infty$. limit theorems invoked by Consistent and Asymptotically Normal (CAN) estimators and associated tests, also rely on probabilistic assumptions that are often ignored, rendering the reliability of the resulting inferences dubious at best. Indeed, the truth of the matter is that all inference results will rely exclusively on the n available data points $\mathbf{x}_0$ and nothing more. As argued by Le Cam (1986, p. xiv):

"... limit theorems "as n tends to infinity" are logically devoid of content about what happens at any particular n."

Asymptotic theory based on '$n \to \infty$' relate to the 'capacity' of inference procedures to pinpoint μ^*, the 'true' μ, as data information accrues $\{x_k\}_{k=1}^{\infty}:=(x_1, x_2, ..., x_n, ...)$ approaching the limit at ∞. In that sense, asymptotic properties are useful for their value in excluding potentially unreliable estimators and tests, but they do not guarantee the reliability of inference procedures for a given data $\mathbf{x}_0$. For instance, an inconsistent estimator is likely to give rise to unreliable inference, but a consistent one does not guarantee the trustworthiness of the inference results.

(5) There is an erroneous impression that statistical misspecification is inevitable since modeling involves abstraction, simplification and approximation. Hence, the slogan "All models are wrong, but some are useful" is used as the excuse for neglecting model validation.

This aphorism is especially pernicious because confuses two different aspects of empirical modeling:

(i) the adequacy of the substantive (structural) model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ (substantive adequacy), vis-a-vis the phenomenon of interest,

(ii) the validity of the (implicit) statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ (statistical adequacy) vis-a-vis the data $\mathbf{z}_0$.

It's one thing to claim that the structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ is wrong in the sense that it is not realistic enough in a substantive sense, and quite another to claim that it is OK to impose invalid probabilistic assumptions on data $\mathbf{z}_0$. In cases where we may

arrive at statistically adequate models, we can learn true things even with idealized and somewhat unrealistic models.

What is wrong with the traditional approach. When empirical modeling is viewed as curve-fitting one imposes the substantive information (theory) on data $\mathbf{z}_0$ at the outset, by estimating $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$. The end result is often a statistically and substantively misspecified model, but one has no way to delineate the two sources of error:

(a) the inductive premises are invalid, or

(b) the substantive information is inadequate,

and apportion blame with a view to address the unreliability of inference problem.

The key to circumventing this *Duhemian ambiguity* is to find a way to disentangle the statistical $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ from the substantive premises $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$; see Mayo (1996). What is often insufficiently appreciated is the fact that behind every substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ there is (often implicit) a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ which provides the inductive premises for the reliability of statistical inference based on data $\mathbf{z}_0$. The latter is just a set of probabilistic assumptions pertaining to the chance regularities in data $\mathbf{z}_0$. Statistical adequacy ensures error reliability in the sense that the actual error probabilities approximately closely the nominal ones.

2 M-S testing: a first encounter

To get some idea of what M-S testing is all about, let us focus on a few simple tests to assess assumptions [1]-[4] of the simple Normal model (table 4).

Table 4 - The simple Normal model

Statistical GM:		$X_t=\mu+u_t,\ t\in\mathbb{N}$,
[1]	Normal:	$X_t \backsim \mathsf{N}(.,.)$,
[2]	Constant mean:	$E(X_t)=\mu$, for all $t\in\mathbb{N}$,
[3]	Constant variance:	$Var(X_t)=\sigma^2$, for all $t\in\mathbb{N}$,
[4]	Independence:	$\{X_t,\ t\in\mathbb{N}\}$ - independent process.

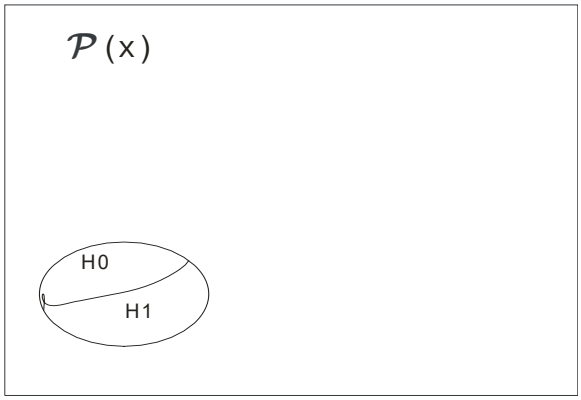


Fig. 1: N-P testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$

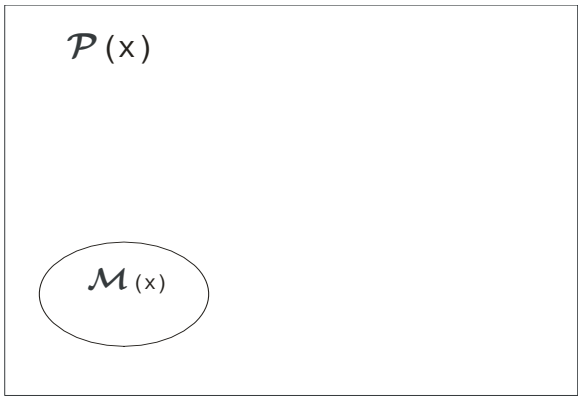


Fig. 2: M-S testing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$

Mis-Specification (M-S) testing differs from Neyman-Pearson (N-P) testing in several respects, the most important of which is that the latter is testing within

boundaries of the assumed statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, but the former is testing outside those boundaries. N-P testing partitions the assumed model using the parameters as an index. Conceptually, M-S testing partitions the set $\mathcal{P}(\mathbf{x})$ of all possible statistical models that could have given rise to data $\mathbf{x}_0$ into $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and its compliment $\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. However, $\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ cannot be expressed in a parametric form and thus M-S testing is more open-ended than N-P testing.

2.1 Omnibus (nonparametric) M-S tests

2.1.1 The 'Runs M-S test' for the IID assumptions [2]-[4]

The hypothesis of interest concerns **the ordering** of the sample $\mathbf{X}:=(X_1, X_2, ..., X_n)$ in the sense that the distribution of the sample remains the same under for any random *reordering* of $\mathbf{X}$, i.e.

$$H_0\text{: } f(x_1, x_2, ..., x_n; \boldsymbol{\theta})=f(x_{i_1}, x_{i_2}, ..., x_{i_n}; \boldsymbol{\theta}),$$

for any permutation $(i_1, i_2, ..., i_m)$ of the index $(i=1, 2, ..., n)$.

Step 1: transform data $\mathbf{x}_0:=(x_1, x_2, ..., x_n)$ into a sequence of differences $(x_t - x_{t-1})$, $t=2, 3, ..., n$.

Step 2: replace each $(x_t-x_{t-1})>0$ with '+' and each $(x_t-x_{t-1})<0$ with '-'. A 'run' is a segment of the sequence consisting of adjacent identical elements which are followed and proceeded by a different symbol.

The transformation takes the form:

$$(x_1, ..., x_n) \to \{(x_t-x_{t-1}),\ t=2, ..., n\} \to (+\,+\,-\,+\cdots+\,-\,-\,+) \tag{8}$$

Step 3: count the number of runs.

Example:

$$\underbrace{++}_{1}\ \underbrace{-}_{2}\ \underbrace{+++}_{3}\underbrace{--}_{4}\ \underbrace{++}_{5}\underbrace{-}_{6}\underbrace{--}_{7}\ \underbrace{+}_{8}\ \underbrace{-}_{9}\ \underbrace{+}_{10}\ \underbrace{--}_{11}\underbrace{+++++}_{12}\ \underbrace{-}\cdots$$

consists of 12 runs; the first is a run of 2 positive signs, the second a run of 1 negative sign, etc.

Runs test. One of the simplest *runs test* is based on comparing the *actual* number of runs R with the number of *expected* runs assuming that the data represent a realization of an IID process $\{X_t,\ t\in\mathbb{N}\}$. The test takes the form:

$$\boxed{d_R(\mathbf{X})=\tfrac{[R-E(R)]}{\sqrt{Var(R)}},\ C_1(\alpha)=\{\mathbf{x}\text{: } |d_R(\mathbf{x})|>c_{\frac{\alpha}{2}}\}}$$

Using simple combinatorics with a sample size n, one can derive:

$$E(R)=\left(\tfrac{2n-1}{3}\right),\qquad Var(R)=\tfrac{16n-29}{90},$$

and show that the distribution of $d_R(\mathbf{X})$ for $n\geq 40$ is:

$$d_R(\mathbf{X})=[R-E(R)]\,/\sqrt{Var(R)}\overset{\text{IID}}{\approx}\mathsf{N}(0,1).$$

Note that this test is *insensitive* to departures from Normality because all distributional information has been lost in the transformation (8).

Example-exam scores. Consider the exam scores data, shown in table 5..

Table 5: Test scores - alphabetical order

98	43	77	51	93	85	76	56	59	62	67	79	66	98	57	80	73	68
71	74	83	75	70	76	56	84	80	53	70	67	100	78	65	77	88	81
66	72	65	58	45	63	57	87	51	40	70	56	75	92	73	59	81	85
62	93	84	68	76	62	65	84	59	60	76	81	69	95	66	87		

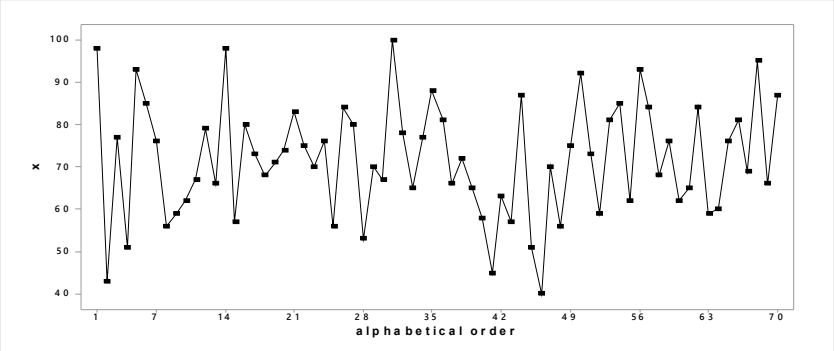


Fig. 3: Exam scores data in alphabetical order

Case 1. Consider the data given in the table 5 that refer to the test scores (y-axis) in a multiple choice exam on Principles of Economics, reported in alphabetical order using the students' surnames (x-axis). The exam scores when arranged in **alphabetical order** (fig, 3) exhibit the following runs:

$$\begin{aligned} &\{1,1,4,1,1,3,1,1,1,1,2,1,1,1,1,1,1,2,2,1,1,2,3,1,1\}^{+}, \\ &\{1,1,3,1,1,2,2,1,2,1,2,2,3,1,2,1,2,1,2,1,2,1,1,1,1\}^{-}. \end{aligned} \tag{9}$$

Hence, the actual number of runs is 50, which is close to the number of runs expected under IID: $(2(70)-1)/3 \simeq 46$. Applying the above runs test yields:

$$d_R(\mathbf{x}_0)=50-\left(\tfrac{2(70)-1}{3}\right)/\sqrt{\tfrac{16(70)-29}{90}}=1.053[.292],$$

where the p-value is in square brackets. This indicates no departure from the IID ([2]-[4]) assumptions.

Case 2. Consider the scores data ordered according to the **sitting order** in figure 4. This data exhibit cycles which yield the following runs up and down:

$$\{3,2,4,4,1,4,3,6,1,4\}^{+}, \quad \{2,2,2,4,3,3,7,4,6,1,3\}^{-}. \tag{10}$$

The difference between the patterns in (9) and (10) is in that there is more clustering and thus fewer runs in the latter case.

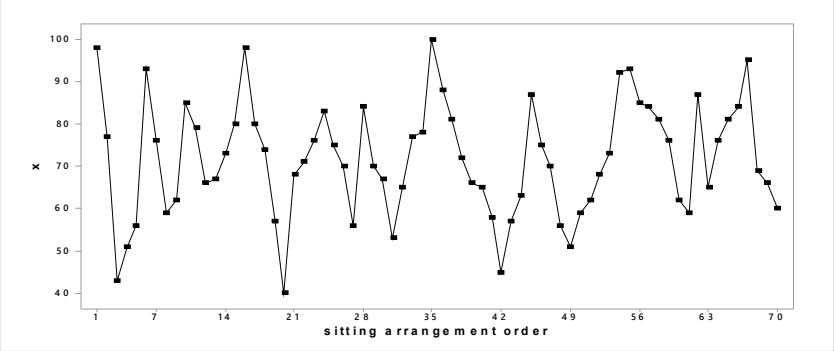


Fig. 4: Exam scores data in sitting order

The actual number of runs is 21; less than half of what were expected under IID.

$$d_R(\mathbf{x}_0)=21-\left(\frac{2(70)-1}{3}\right)/\sqrt{\frac{16(70)-29}{90}}=-7.276[.0000],$$

which clearly indicates strong departures from the IID ([2]-[4]) assumptions.

2.1.2 Kolmogorov's M-S test for Normality ([1])

The Kolmogorov M-S test for assessing the validity of a distributional assumption under two key conditions:

(i) the data $\mathbf{x}_0$:=$(x_1, x_2, ..., x_n)$ can be viewed as a realization of a random (IID) sample $\mathbf{X}$:=$(X_1, X_2, ..., X_n)$, and

(ii) the random variables $X_1, X_2, ..., X_n$ are continuous (not discrete).

The test relies on the empirical cumulative distribution function (ecdf):

$$\widehat{F}_n(x)=\frac{[\text{no of } (x_1,x_2,....,x_n) \text{ that do not exceed } x]}{n},\ \forall x\in\mathbb{R}.$$

Under (i)-(ii), the ecdf is a *strongly consistent* estimator of the cumulative distribution function (cdf): $F(x)=P(X\leq x)$, $\forall x\in\mathbb{R}$.

The generic hypothesis being tested takes the form:

$$H_0\text{: } F^*(x)=F_0(x),\ x\in\mathbb{R}, \tag{11}$$

where $F^*(x)$ denotes the true cdf, and $F_0(x)$ the cdf assumed by the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

Kolmogorov (1933) proposed the distance function:

$$\Delta_n(\mathbf{X})=\sup_{x\in\mathbb{R}}|\widehat{F}_n(x)-F_0(x)|,$$

and proved that under (i)-(ii):

$$\lim_{n\to\infty}\mathbb{P}(\sqrt{n}\Delta_n(\mathbf{X})\leq x)=F_K(x)\simeq 1-2\exp(-2x^2),\ \text{for } x>0. \tag{12}$$

One can define a M-S test based on the test statistic $K_n(\mathbf{X})=\sqrt{n}\Delta_n(\mathbf{X})$, giving rise to the p-value:

$$\mathbb{P}(K_n(\mathbf{X})>K_n(\mathbf{x}_0);\ H_0)=p(\mathbf{x}_0).$$

Example. Applying the Kolmogorov test to the scores data in fig. 3 yielded:

$$\mathbb{P}(K_n(\mathbf{X}) > .039;\ H_0)=.15,$$

which does not indicate any serious departures from the Normality assumption. The graph below provides a pictorial depiction of what this test is measuring in terms of the discrepancies from the line to the observed points.

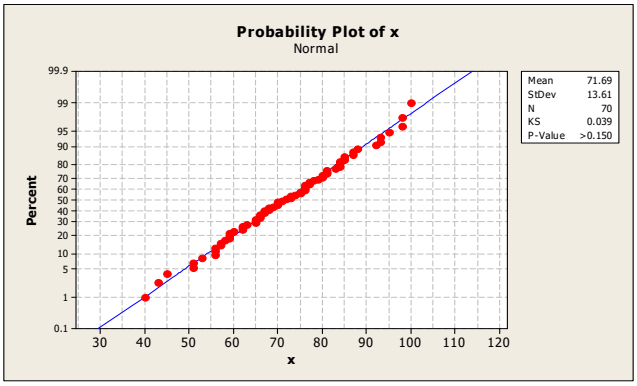


fig. 5: P-P Normality plot

Note that this particular test might be too sensitive to outliers because it picks up only the biggest distance!

2.1.3 The role for omnibus M-S tests

The **key advantage** of the above omnibus tests is that they probe more broadly around the $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ than directional (parametric) M-S tests at the expense of lower power. However, tests with low power are useful in M-S testing because when they detect a departure, they provide better evidence for its presence than a test with very high power!

A **key weakness** of the above omnibus tests is that when the null hypothesis is rejected, the test does not provide any information as to the direction of departure. Such information is needed for the next stage of modeling, that of respecifying the original model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ with a view to account for the systematic information not accounted for by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

2.2 Directional (parametric) M-S tests

2.2.1 A parametric M-S test for independence ([4])

A general approach to deriving M-S tests is to return to the original probabilistic assumptions of the process $\{X_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{x}_0:=(x_1, x_2, ..., x_n)$, and replace one or more assumptions with more general ones and derive relevant distance functions using the two statistical Generating Mechanisms (GMs).

In the case of the simple Normal model, the process $\{X_t,\ t\in\mathbb{N}\}$ is assumed to be NIID. Let us relax the IID assumptions to Markov dependence and stationarity,

which gives rise to the AutoRegressive (AR(1)), model based on $f(x_t|x_{t-1};\boldsymbol{\theta})$, whose statistical GM is:

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + \varepsilon_t,\ \varepsilon_t \backsim \mathsf{N}(0,\sigma_0^2),\ t\in\mathbb{N}, \tag{13}$$

where $\alpha_0{=}\mu(1{-}\alpha_1){\in}\mathbb{R},\ \alpha_1{=}\frac{\sigma(1)}{\sigma(0)}{\in}(-1,1),\ \sigma_0^2{=}\sigma(0)(1{-}\alpha_1^2){\in}\mathbb{R}_+$;

$\mu{=}E(X_t),\ \sigma(0){=}Var(X_t),\ \sigma(1){=}Cov(X_t,X_{t-1}),\ t{=}1,...,n,...$

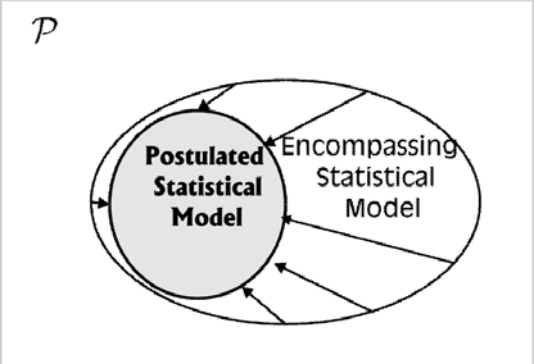


Fig. 6: M-S testing by encompassing

The AR(1) parametrically nests (includes as a special case) the simple Normal model because when $\alpha_1{=}0$:

$$\alpha_0{=}\ \mu(1{-}\alpha_1)\big|_{\alpha_1=0}{=}\mu,\ \sigma_0^2{=}\ \sigma(0)(1{-}\alpha_1^2)\big|_{\alpha_1=0}{=}\sigma(0)$$

the AR(1) reduces to the simple Normal:

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + \varepsilon_t \ \overset{\alpha_1=0}{\rightarrow}\ X_t{=}\mu + u_t,\ t\in\mathbb{N}.$$

This suggests that a way to assess assumption [4] (table 4) is to test the hypotheses:

$$H_0\text{: } \alpha_1{=}0 \text{ vs. } H_1\text{: } \alpha_1 \neq 0, \tag{14}$$

in the context of the AR(1) model. This will give rise to a **t-type** test $T_\alpha{:=}\{\tau(\mathbf{X}),C_1(\alpha)\}$:

$$\tau(\mathbf{X}){=}\frac{(\widehat{\alpha}_1-0)}{\sqrt{Var(\widehat{\alpha}_1)}} \overset{H_0}{\approx} \mathrm{St}(n{-}2),\quad C_1(\alpha){=}\{\mathbf{x}\text{: } |\tau(\mathbf{x})|>c_\alpha\},$$

$\widehat{\alpha}_1{=}\frac{\sum_{t=1}^n(X_t-\overline{X})(X_{t-1}-\overline{X})}{\sum_{t=1}^n(X_{t-1}-\overline{X})^2},\ Var(\widehat{\alpha}_1){=}\frac{s^2}{\sum_{t=1}^n(X_{t-1}-\overline{X})^2},\ s^2{=}\frac{1}{n-2}\sum_{t=1}^n(X_t{-}\widehat{\alpha}_0{-}\widehat{\alpha}_1X_{t-1})^2,\ \widehat{\alpha}_0{=}(1{-}\widehat{\alpha}_1)\overline{X}.$

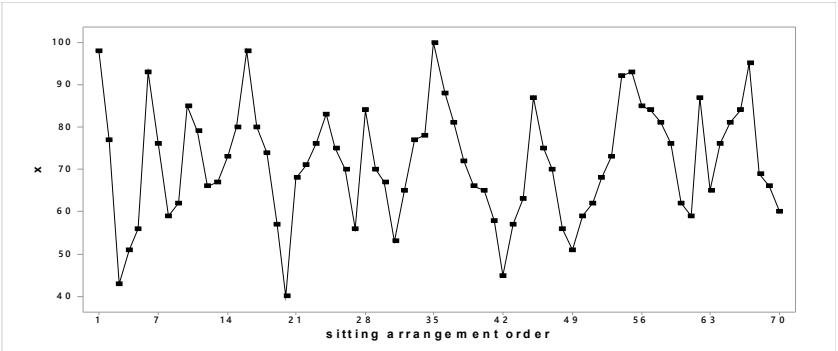


Fig. 4: Exam scores data in sitting order

Example. For the data in figure 4, (13) yields:

$$X_t=\underset{(7.790)}{39.593}+\underset{(0.106)}{0.441}X_{t-1}+\widehat{\varepsilon}_t,\quad R^2=.2,\ s^2=143.42,\ n=69,$$

The M-S t-test for (14) yields:

$$\tau(\mathbf{x}_0)=\left(\tfrac{.441}{.106}\right)=4.160,\ p(\mathbf{x}_0)=.0000,$$

indicating a clear departure from assumption [4].

It is straightforward to extend the above test to Markov(m) by estimating the auxiliary regression:

$$X_t=\alpha_0+\textstyle\sum_{i=1}^{m}\alpha_iX_{t-i}+\varepsilon_t,\ t\in\mathbb{N}, \tag{15}$$

and testing the coefficient restrictions:

$$H_0\text{: }\alpha_1=\alpha_2=\cdots=\alpha_m=0,\ \text{for } m<(n-1). \tag{16}$$

This gives rise to an F-type test, analogous to Ljung and Box (1978) test, with one big difference: the estimated coefficients in (15) can also be assessed individually using t-tests in order to avoid the large m problem raised above. For the case $m=2$, the auxiliary regression is:

$$X_t=\alpha_0+\alpha_1X_{t-1}+\alpha_2X_{t-2}+\varepsilon_t,\ t\in\mathbb{N}, \tag{17}$$

and the F-test for the joint significance of α_1 and α_2 will take the form:

$$F(\mathbf{x})=\tfrac{RRSS-URSS}{URSS}\left(\tfrac{n-3}{2}\right)\overset{H_0}{\backsim}F(2,n-3)$$

where $RRSS=\sum_{t=1}^{n}(X_t-\overline{X}_n)^2$, denote the *Restricted [restrictions* $\alpha_1=\alpha_2=0$ *imposed]Residuals Sum of Squares*, and $URSS=\sum_{t=1}^{n}\widehat{\varepsilon}_t^2$, $\widehat{\varepsilon}_t=X_t-\widehat{\alpha}_0-\widehat{\alpha}_1X_{t-1}-\widehat{\alpha}_2X_{t-2}$ the *Unrestricted Residuals Sum of Squares,* $F(2,n-3)$ denotes the F distribution with 2 and $n-3$ degrees of freedom.

One of the key advantages of this approach is that it can easily be extended to derive joint M-S tests that assess more than one assumption.

2.2.2 A parametric M-S test for IID ([2]-[3])

The above t-type parametric test based on the auxiliary Autoregression (13) can be extended to provide a joint test for assumptions [2] and [4], by replacing the stationarity assumption of $\{X_t,\ t\in\mathbb{N}\}$ with mean non-stationarity, gives rise to a heterogeneous AR(1) model with a statistical GM:

$$X_t=\delta_0+\overbrace{\delta_1t}^{\overline{[2]}}+\overbrace{\alpha_1X_{t-1}}^{\overline{[4]}}+\varepsilon_t,\ t\in\mathbb{N}, \tag{18}$$

$\delta_0=\mu+\alpha_1(\gamma_1\text{-}\mu)$, $\delta_1=(1\text{-}\alpha_1)\gamma_1t$, $\alpha_1=(\sigma(1)/\sigma(0))$, $\sigma_0^2=\sigma(0)(1\text{-}\alpha_1^2)$.

The AR(1) with a trend nests the simple Normal model:

$$X_t=\delta_0+\delta_1t+\alpha_1X_{t-1}+\varepsilon_t\ \overset{\substack{\delta_1=0\\ \alpha_1=0}}{\longrightarrow}\ X_t=\mu+u_t,\ t\in\mathbb{N}.$$

This suggests that a way to assess assumptions [2]&[4] (table 4) jointly is to test the hypotheses:

$$H_0\text{: } \alpha_1{=}0 \text{ and } \delta_1{=}0 \text{ vs. } H_1\text{: } \alpha_1{\neq}0 \text{ or } \delta_1{\neq}0.$$

This will give rise to a **F-type** test $T_\alpha{:=}\{\tau(\mathbf{X}), C_1(\alpha)\}$:

$$\boxed{F(\mathbf{X}){=}\tfrac{\text{RRSS-URSS}}{\text{URSS}}\left(\tfrac{n-3}{2}\right)\overset{H_0}{\approx}\mathsf{F}(2,n{-}3),\quad C_1(\alpha){=}\{\mathbf{x}\text{: } F(\mathbf{x}){>}c_\alpha\},}$$

$$\text{URSS}{=}\textstyle\sum_{t=1}^{n}(X_t{-}\widehat{\delta}_0{-}\widehat{\delta}_1 t{-}\widehat{\alpha}_1 X_{t-1})^2,\ \text{RRSS}{=}\sum_{t=1}^{n}(X_t{-}\overline{X})^2,$$

where URSS and RRSS denote the Unrestricted and Restricted Residual Sum of Squares, respectively, and F(2,n-3) denotes the F distribution with 2 and $n{-}3$ degrees of freedom.

Example. For the data in figure 4, the restricted and unrestricted models yielded, respectively:

$$X_t{=}\underset{(1.631)}{71.69}+\widehat{u}_t,\quad s^2{=}185.23,\ n{=}69,$$

$$X_t{=}\underset{(8.034)}{38.156}+\underset{(.073)}{.055}t+\underset{(0.107)}{0.434}X_{t-1}+\widehat{\varepsilon}_t,\quad s^2{=}144.34,\ n{=}69, \tag{19}$$

where RRSS=2.6845, URSS=2.1543, yielding:

$$F(\mathbf{x}_0){=}\left(\tfrac{2.6845-2.1543}{2.1543}\right)\left(\tfrac{67}{2}\right){=}8.245,\ p(\mathbf{x}_0){=}.0006,$$

indicating a clear departure from the null ([2]&[4]).

What is particularly notable about the auxiliary autoregression (19) is that a closer look at the t-ratios indicates that the source of the problem is dependence and *not* t-heterogeneity. The t-ratio of the coefficient of t is statistically insignificant:

$$\tau(\mathbf{x}_0){=}\left(\tfrac{.055}{.073}\right){=}.753,\ p(\mathbf{x}_0){=}.226,$$

but the coefficient of X_{t-1} is statistically significant:

$$\tau(\mathbf{x}_0){=}\left(\tfrac{.434}{.107}\right){=}4.056,\ p(\mathbf{x}_0){=}.0000,$$

indicating a clear departure from assumption [4], but not from [2]. This information that enables one to apportion blame cannot be gleaned from the runs test.

An alternative, and more preferable, way to specify the above auxiliary regressions is in terms of the *residuals*:

$$\widehat{u}_t{=}(X_t-\overline{x}_n){=}(X_t-71.69),\ t{=}1,2,...,n,$$

in the sense that the auxiliary regression:

$$\widehat{u}_t{=}\underset{(8.034)}{-33.534}+\underset{(.073)}{.055}t+\underset{(0.107)}{.434}X_{t-1}+\widehat{\varepsilon}_t,\quad s^2{=}144.34,\ n{=}69, \tag{20}$$

is a mirror image of (19):

$$X_t{=}\underset{(8.034)}{38.156}+\underset{(.073)}{.055}t+\underset{(0.107)}{.434}X_{t-1}+\widehat{\varepsilon}_t,\quad s^2{=}144.34,\ n{=}69, \tag{21}$$

with identical parameter estimates, apart from the constant ($-33.534{=}38.156{-}71.6$), which is irrelevant for M-S testing purposes.

2.2.3 A parametric M-S test for assumptions [3]-[4]

In light of the fact that $\sigma^2=Var(X_t)=E(u_t^2)$ one can test the variance constancy [3] and independence [4] assumptions using the residuals squared in the context of the auxiliary regression:

$$\widehat{u}_t^2=\gamma_0+\overbrace{\gamma_1 t}^{[3]}+\overbrace{\gamma_2 X_{t-1}^2}^{[4]}+v_t,\ t=1,2,...,n.$$

Using the above data, this gives rise to:

$$\widehat{u}_t^2=\underset{(89.43)}{295.26}-\underset{(1.353)}{1.035}t-\underset{(.014)}{.016}X_{t-1}^2+\widehat{v}_t. \tag{22}$$

The non-significance of the coefficients γ_1 and γ_2 indicate no departures from assumptions [3] and [4].

2.2.4 Extending the above auxiliary regression

The auxiliary regression (18), providing the basis of the joint test for assumptions [2]-[4] can be easily extended to include higher order trends (up to order $m \geq 1$) and additional lags ($\ell \geq 1$):

$$X_t=\delta_0+\textstyle\sum_{k=1}^{m}\delta_k t^k+\sum_{i=1}^{\ell}\alpha_i X_{t-i}+\varepsilon_t,\ t\in\mathbb{N}. \tag{23}$$

2.2.5 A parametric M-S test for Normality ([1])

An alternative way to test Normality is to use parametric tests relying on key features of the distribution. An example of this type of test is the Skewness-Kurtosis test.

A key feature of the Pearson family is that it is specified using the first four moments. Within this family we can characterize several distributions using the skewness and kurtosis coefficients:

$$\alpha_3=\frac{E(X-E(X))^3}{\left(\sqrt{Var(X)}\right)^3},\qquad \alpha_4=\frac{E(X-E(X))^4}{\left(\sqrt{Var(X)}\right)^4}.$$

The skewness is the standardized third central moment and provides a measure of asymmetry of $f(x)$, and the kurtosis is the standardized fourth central moment and is a measure of the peakness in relation to the tails of $f(x)$.

The Normal distribution is characterized within the Pearson family via the restrictions:

$$(\alpha_3=0,\ \alpha_4=3)\Rightarrow f^*(x)=\phi(x),\text{ for all } x\in\mathbb{R},$$

where $f^*(x)$ and $\phi(x)$ denote the true density and the Normal density, respectively.

These moments can be used to derive a M-S test for the Normality assumption [1] (table 4), using the hypotheses:

$$H_0\text{: }\alpha_3=0 \text{ and } \alpha_4=3 \text{ vs. } H_1\text{: }\alpha_3\neq 0 \text{ or } \alpha_4\neq 3.$$

The **Skewness-Kurtosis test** is given by:

$$\boxed{\begin{array}{c} SK(\mathbf{X})=\frac{n}{6}\widehat{\alpha}_3^2+\frac{n}{24}\left(\widehat{\alpha}_4-3\right)^2 \underset{\alpha}{\overset{H_0}{\backsim}} \chi^2(2), \\ \mathbb{P}(SK(\mathbf{X})>SK(\mathbf{x}_0);\ H_0)=p(\mathbf{x}_0), \end{array}} \tag{24}$$

where $\chi^2(2)$ denotes the chi-square distribution with 2 degrees of freedom, and:

$$\widehat{\alpha}_3=\frac{\frac{1}{n}\sum_{k=1}^n(X_k-\overline{X})^3}{\left(\sqrt{\frac{1}{n}\sum_{k=1}^n(X_k-\overline{X})^2}\right)^3}, \qquad \widehat{\alpha}_4=\frac{\frac{1}{n}\sum_{k=1}^n(X_k-\overline{X})^4}{\left(\sqrt{\frac{1}{n}\sum_{k=1}^n(X_k-\overline{X})^2}\right)^4}.$$

Example. For the scores data in fig. 3: $\widehat{\alpha}_3=-.03$, $\widehat{\alpha}_4=2.62$:

$$SK(\mathbf{x}_0)=\frac{70}{6}(-.03)^2+\frac{70}{24}(-.38)^2=.432, \quad p(\mathbf{x}_0)=.806,$$

indicating no departure from the Normality assumption [1].

How is this test different from Kolmogorov's nonparametric test? Depending on whether $\widehat{\alpha}_3\neq 0$ or $\widehat{\alpha}_4\neq 3$, one can conclude whether the underlying distribution $f(x)$ is non-symmetric or leptokurtic and that information can be useful at the respecfication stage.

2.3 t-plots of typical realizations of data

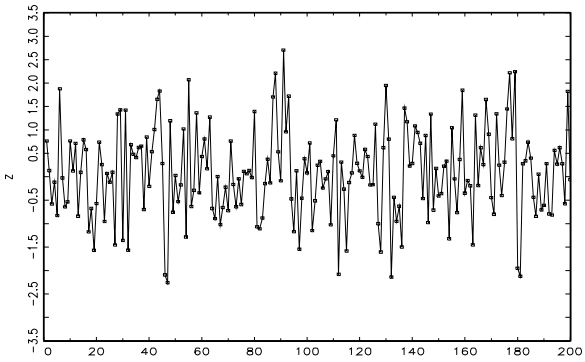


Fig. 5.3: Typical realization of NIID data

Departures from Independence.

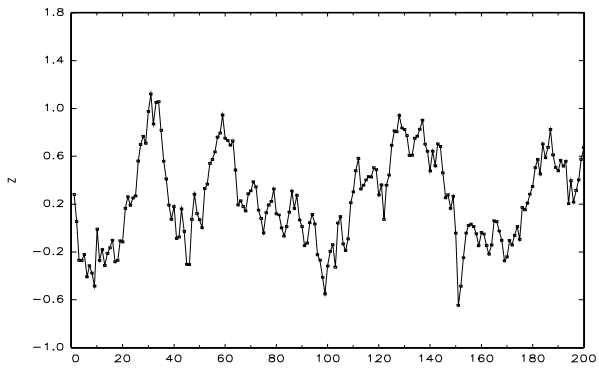


Fig. 5.5: NID positively dependent data

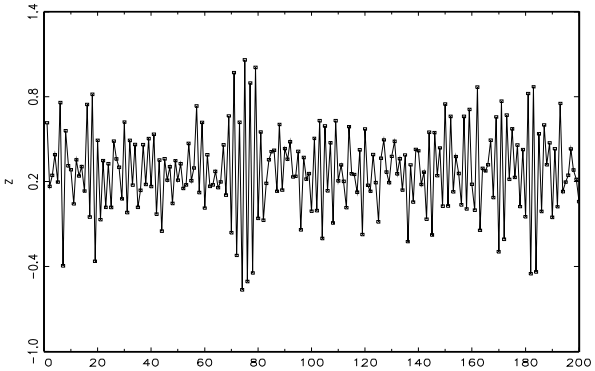


Fig. 5.6: NID negatively dependent data

Departures from Identically Distributed.

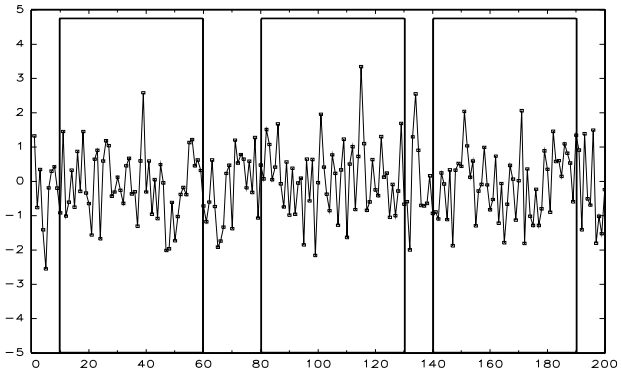


Fig. 5.7: Assessing t-homogeneity using the window thought experiment

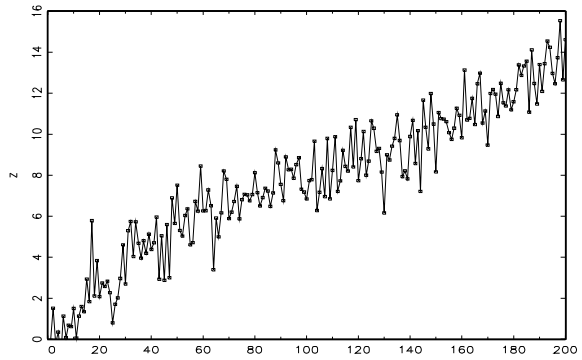


Fig. 5.8: Simulated NI mean (trend) heterogeneous data

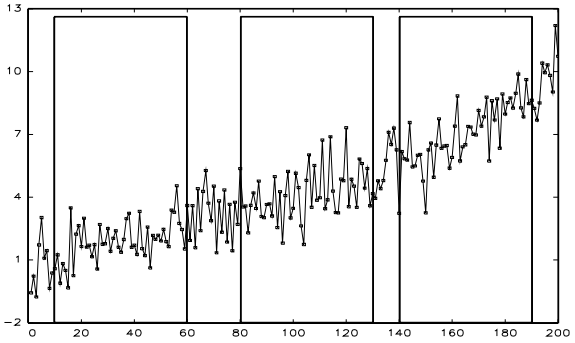


Fig. 5.9: Window experiment to assess t-homogeneity

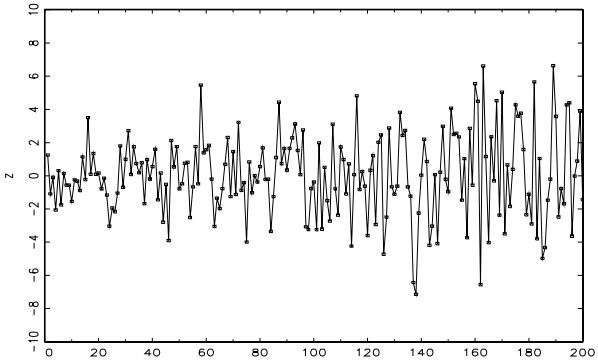


Fig. 5.10: NI variance-trending data

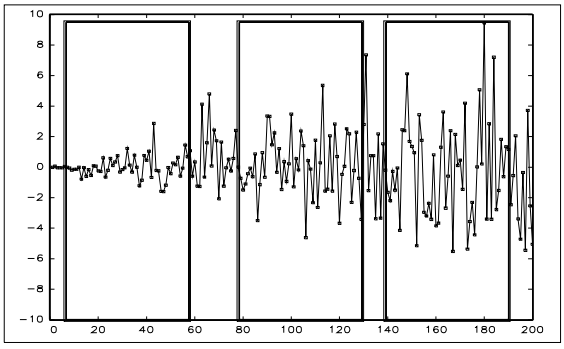


Fig. 5.11: Window experiment-fig. 5.10 data

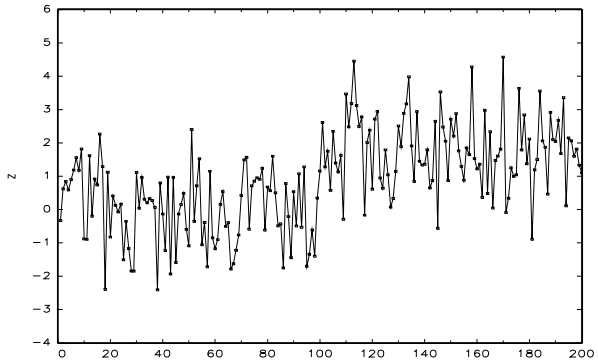


Fig. 5.12: NI mean-shift data

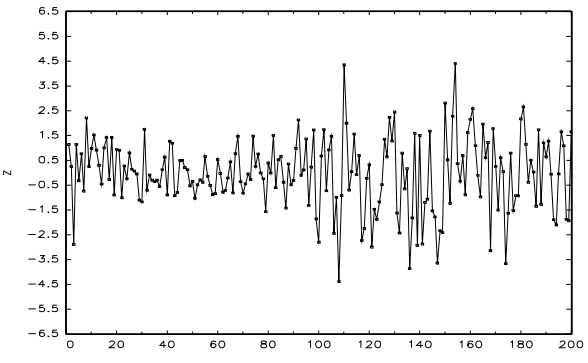


Fig. 5.13: NI variance-shift data

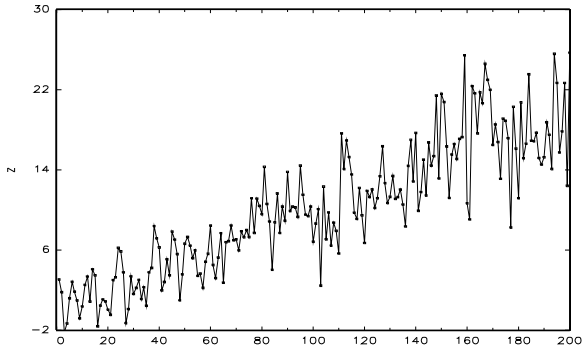


Fig. 5.14: NI mean/variance trending data

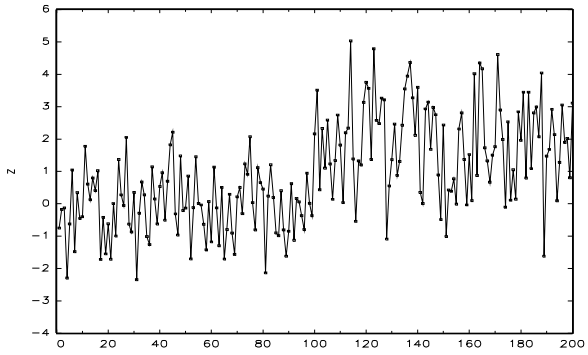


Fig. 5.15: NI mean/variance-shift data

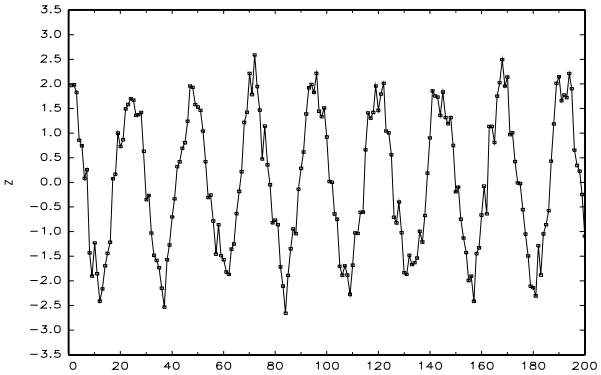


Fig. 5.16: NI mean seasonal data

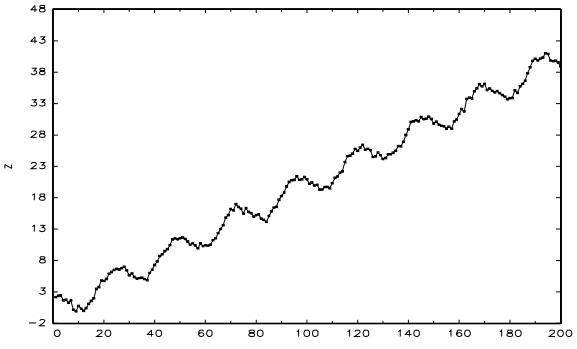


Fig. 5.17: NI mean seasonal and trend data

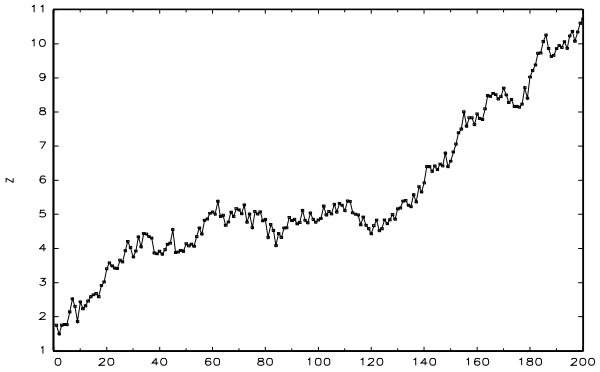


Fig. 5.18: Normal, dependent and trending data

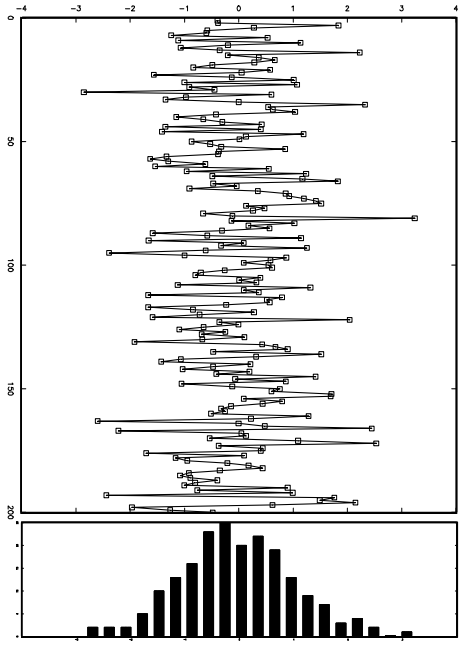


Fig. 5.20: t-plot and histogram of NIID data

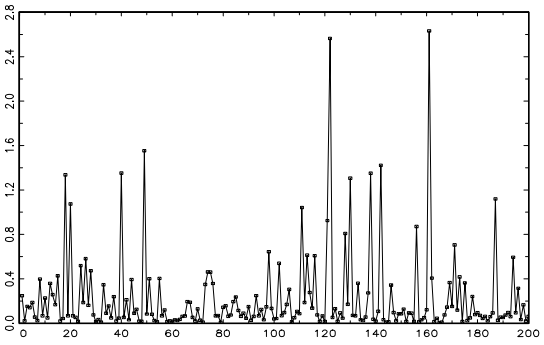


Fig. 5.21: Log-Normal IID data

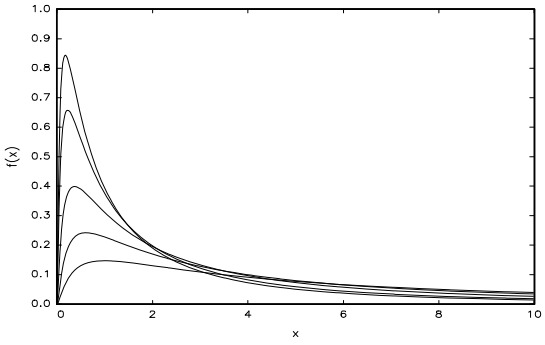


Fig. 5.22: Log-Normal family of densities

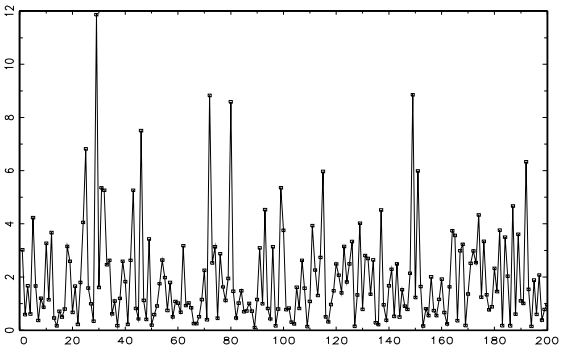


Fig. 5.24: Exponential IID data

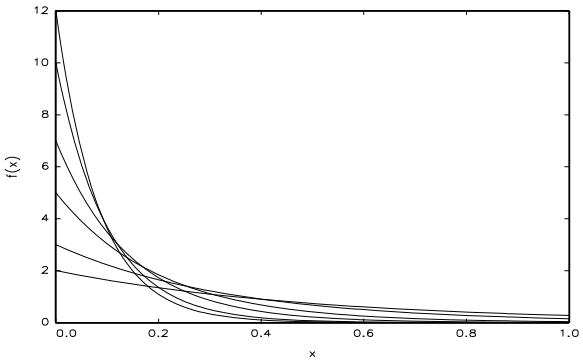


Fig. 5.25: Exponential family of densities

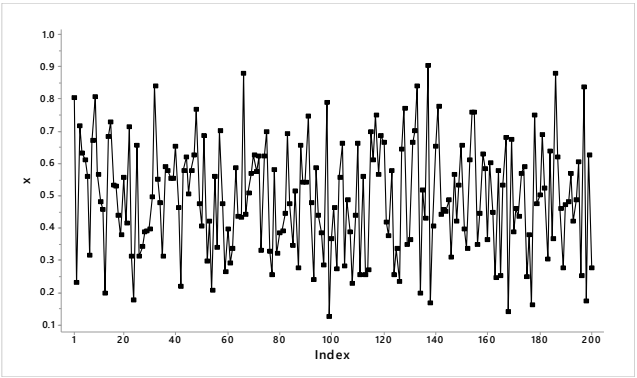


Fig. 5.28: Beta(4,4) IID data

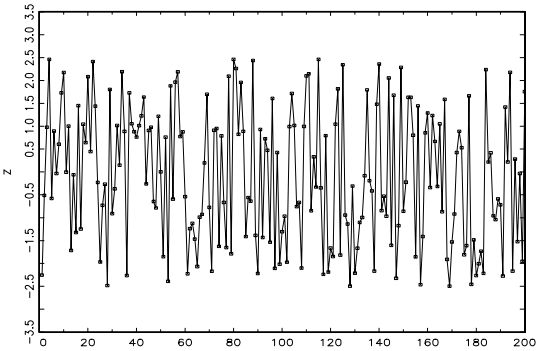


Fig. 5.29: Uniform IID data

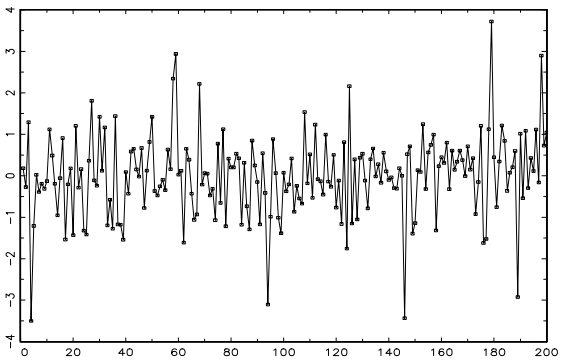


Fig. 5.30: Student's t IID data

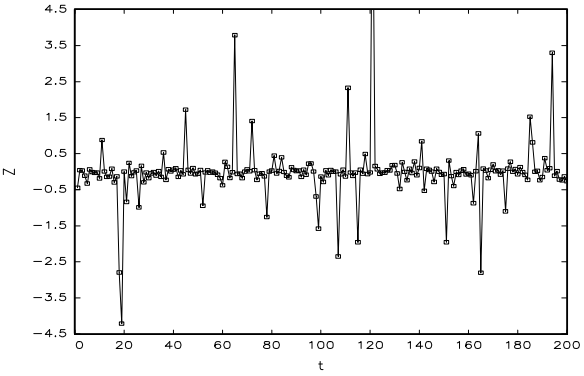


Fig. 5.31: Cauchy IID data

CAUTION: almost very book or paper plotting the N(0,1) vs. Student's t density is wrong. The correct graph is given in 3.23 but the one seen everywhere else is fig. 3.24.

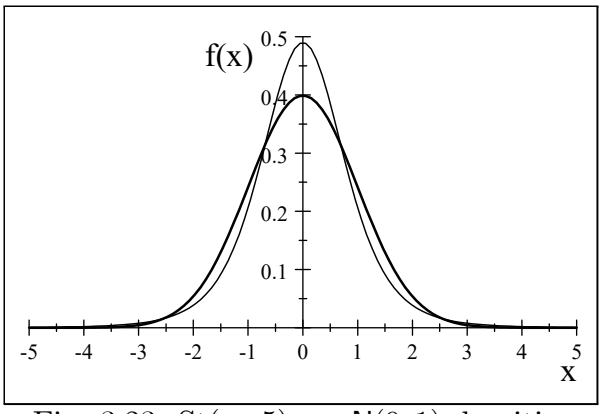


Fig. 3.23: St(ν=5) vs. $\mathsf{N}(0,1)$ densities both scaled to have $\sqrt{Var(X)}$=1.

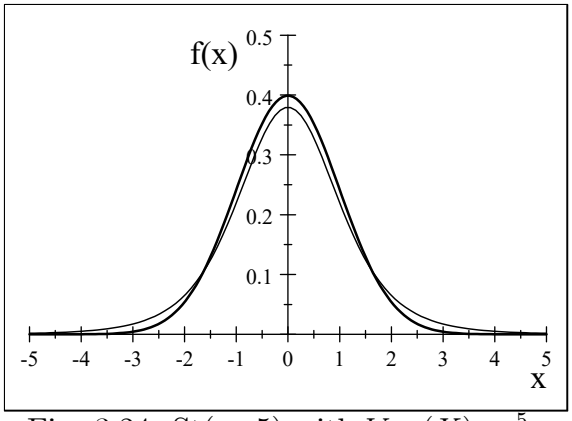


Fig. 3.24: St(ν=5) with $Var(X)=\frac{5}{5-2}$ vs. $\mathsf{N}(0,1)$ with $Var(X)$=1.

2.4 Simple Normal model: a summary of M-S testing

The first auxiliary regression specifies how departures from different assumptions might affect the mean:

$$\text{(i) } \widehat{u}_t=\gamma_{10}+\overbrace{\gamma_{11}t+\gamma_{12}t^2}^{[2]}+\overbrace{\gamma_{13}X_{t-1}+\gamma_{14}X_{t-2}}^{[4]}+\varepsilon_{1t},$$

$$H_0\text{: } \gamma_{11}=\gamma_{12}=\gamma_{13}=\gamma_{14}=0 \text{ vs. } H_1\text{: } \gamma_{11}\neq 0 \text{ or } \gamma_{12}\neq 0 \text{ or } \gamma_{13}\neq 0 \text{ or } \gamma_{14}\neq 0$$

In case where n is small enough one might consider replacing (X_{t-1}, X_{t-2}) with $\widehat{u}_{t-1}$.

The second auxiliary regression specifies how departures from different assumptions might affect the variance:

$$\text{(ii) } \widehat{u}_t^2=\gamma_{20}+\overbrace{\gamma_{21}t+\gamma_{22}t^2}^{[3]}+\overbrace{\gamma_{23}X_{t-1}^2+\gamma_{24}X_{t-2}^2}^{[4]}+\varepsilon_{2t},$$

$$H_0\text{: } \gamma_{21}=\gamma_{22}=\gamma_{23}=\gamma_{24}=0 \text{ vs. } H_1\text{: } \gamma_{21}\neq 0 \text{ or } \gamma_{22}\neq 0 \text{ or } \gamma_{23}\neq 0 \text{ or } \gamma_{24}\neq 0$$

NOTE that the above choices of the various terms for the auxiliary regressions are **only indicative** of the direction of departure from the model assumptions!

Intuition. At the intuitive level the above auxiliary regressions can be viewed as probing the residuals with a view to find systematic statistical information (chance regularities) indicating that the original model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ did not account for. Departures from assumptions [2]-[4] that rightfully belongs to the systematic component and not the error term. More formally, the above auxiliary regressions include terms that represent potential systematic information in data $\mathbf{x}_0$ that might have been disregarded in error by the model assumptions [1]-[4].

When NO departures from assumptions [2]-[4] are detected one can proceed to test the Normality assumption using tests like the skewnesss-kurtosis, the Kolmogorov or

the Anderson-Darling. Otherwise, one uses the residuals from auxiliary regression (i) as a basis for a Normality test.

Example. Consider the casting of two dice data in table 6. Evaluating the sample mean, variance, skewness and kurtosis for the dice data yields:

$$\overline{x}_n=\frac{1}{n}\sum_{k=1}^{n}x_k=7.080,\ \ s^2=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\overline{x}_n)^2=5.993,$$
$$\widehat{\alpha}_3=-.035,\quad \widehat{\alpha}_4=2.362$$

Table 6 - Observed data on dice casting

3	10	11	5	6	7	10	8	5	11	2	9	9	6	8	4	7	6	5	12
7	8	5	4	6	11	7	10	5	8	7	5	9	8	10	2	7	3	8	10
11	8	9	5	7	3	4	9	10	4	7	4	6	9	7	6	12	8	11	9
10	3	6	9	7	5	8	6	2	9	6	4	7	8	10	5	8	7	9	6
5	7	7	6	12	9	10	4	8	6	5	4	7	8	6	7	11	7	8	3

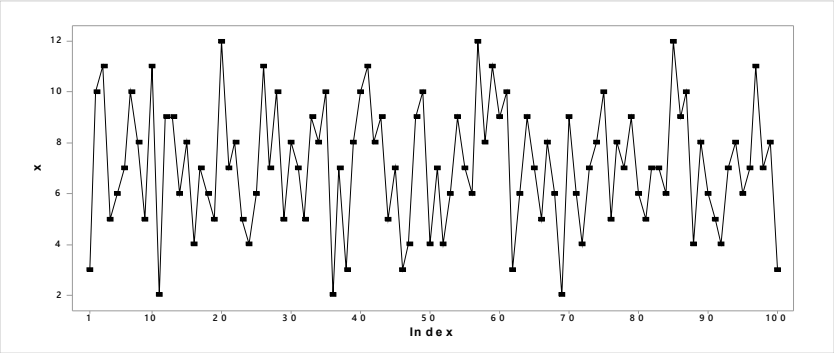


Fig. 7: t-plot of the dice data

(a) Testing assumptions [2]-[4] using the runs test, requires the counting of runs:

+ + - + + + - - + - + + - + - + - - + - + - - + + - + - + - + - - + - + - + - + + + - + - + - + + + -
+ - + + + - - + - + - + - + + - - + - - + - - + + - + - + - - + + - + - + - + - + - - - + + - + + - + -

For $n \geq 40$, the type I error probability evaluation is based on:

$$Z_R(\mathbf{X})=\frac{R-([2n-1]/3)}{\sqrt{[16n-29]/90}}\overset{[1]\text{-}[4]}{\backsim}\mathsf{N}(0,1).$$

For the above data: n=100, R=50:

$$E(R)=(200-1)/3=66.333,\ Var(R)=(16(100)-29)/90=17.456,$$
$$Z_R(\mathbf{X})=\frac{72-66.333}{\sqrt{17.456}}=1.356,\quad \mathbb{P}(|Z_R(\mathbf{X})|>1.356;\ H)=.175.$$

This does not indicate any departure from the IID assumptions.

(b) Test the independence assumption [4] using the auxiliary regression based on $\widehat{u}_t=X_t-\overline{x}_n$:

$$\widehat{u}_t=\alpha_0+\alpha_1X_{t-1}+\varepsilon_t,\ t=1,2,...,n,$$
$$\widehat{u}_t=\underset{(.759)}{0.776}-\underset{(.101)}{.103}X_{t-1}+\underset{(2.425)}{\widehat{\varepsilon}_t},$$

and the t-test for the significance of α_1 yields: $\tau(\mathbf{x})=\frac{.103}{.101}=1.021[.310]$, where the p-value in square brackets indicates no clear departure from the Independence assumption; see Spanos (1999), p. 774.

(c) Test the identically distributed assumptions [2]-[3] using the auxiliary regression:

$$\widehat{u}_t = \underset{(.496)}{0.113} - \underset{(.008)}{.002}t + \underset{(2.460)}{\widehat{\varepsilon}_t},$$

and the t-test for the significance of γ_1 yields: $\tau(\mathbf{x})=\frac{.0022}{.0085}=.259[.793]$, where the p-value indicates no departure from the ID assumption; see Spanos (1999), p. 774.

(d) One can test the IID assumptions [2]-[4] jointly using the auxiliary regression:

$$\widehat{u}_t=\beta_0+\beta_1 t+\beta_2 X_{t-1}+\varepsilon_t,\ t=1,2,...,n,$$

$$\widehat{u}_t = \underset{(.877)}{1.020} - \underset{(.0086)}{.0048}t - \underset{(.101)}{.103}X_{t-1} + \underset{(2.434)}{\widehat{\varepsilon}_t},$$

where the F-test for the joint significance of β_1 and β_2 i.e.

$$H_0\text{: } \beta_1=\beta_2=0, \text{ vs. } H_1\text{: } \beta_1\neq 0, \text{ or } \beta_2\neq 0,$$

$$F(\mathbf{x})=\frac{RRSS-URSS}{URSS}\left(\frac{n-3}{m}\right)=\frac{576.6-568.540}{568.540}\left(\frac{96}{2}\right)=.680[.511],$$

where $RRSS=\sum_{t=1}^{n}(X_t-\overline{X}_n)^2$, denote the *Restricted Residuals Sum of Squares* [the sum of squares of the residuals with the restrictions imposed], and $URSS=\sum_{t=1}^{n}\widehat{\varepsilon}_t^2$, the *Unrestricted Residuals Sum of Squares* [the sum of squares of the residuals without the restrictions], respectively; NOTE that $(RRSS-URSS)$ is often called the *Explained Sum of Squares* (ESS). The p-value in square brackets indicates no departure from the IID assumptions, confirming the previous M-S testing results.

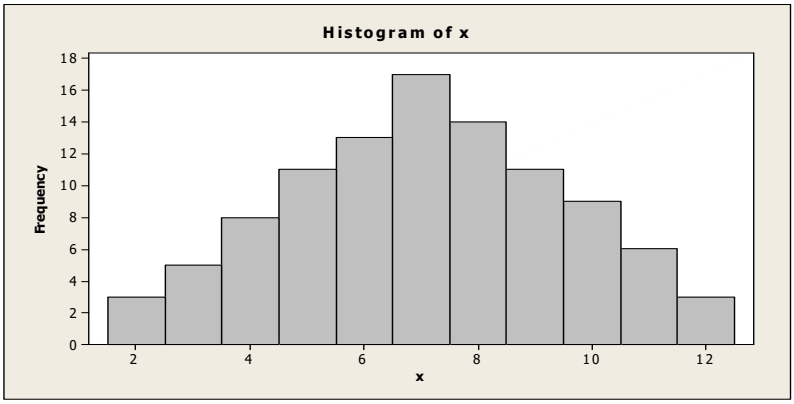


Fig. 8: Histogram of the dice data

(e) Testing the Normality assumption [1] using the SK test yields:

$$SK(\mathbf{x}_0)=\frac{100}{6}(-0.035)^2+\frac{100}{24}(2.362-3)^2=1.716[.424]$$

The p-value indicates no departure from the Normality assumption, but as shown in Spanos (1999), p. 775, this does not mean that the assumption is valid; the test has very low power. This is to be expected because the data come from a discrete triangular distribution with values from 2 to 12, as shown by the histogram (fig. 8).

Using the more powerful Anderson and Darling (1952) test, which for the ordered $\mathbf{X}$ sample simplifies to:

$$\boxed{A\text{-}D(\mathbf{X})=-n-\frac{1}{n}\sum_{k=1}^{n}\left\{(2k-1)\left[\ln Z_{[k]}-\ln(1-\ln Z_{[n+1-k]}\right]\right\}.}$$

however, provides evidence against Normality:

$$A\text{-}D(\mathbf{x}_0)=.772[.041].$$

In light of the M-S results in (a)-(e) one needs to replace the Normality assumption with a triangular discrete distribution in order to get a more adequate statistical model.

3 Mis-Specification (M-S) testing: a formalization

3.1 The nature of M-S testing

The basic question posed by M-S testing is whether or not the particular data $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$ constitute a '*truly typical realization*' of the stochastic process $\{X_t,\ t{\in}\mathbb{N}\}$ underlying the (predesignated) statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}){=}\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}{\in}\Theta\},\ \mathbf{x}{\in}\mathbb{R}^n_X.$$

Parametric M-S tests come in two forms. The first particularizes $\overline{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})}{=}[\mathcal{P}(\mathbf{x}){-}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ by choosing a broader model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})\subset[\mathcal{P}(\mathbf{x}){-}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ that encompasses $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ parametrically (fig. 15.6), and tests the nesting restrictions: $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\psi}){=}\mathbf{0}, \boldsymbol{\theta}{\in}\boldsymbol{\Theta},\ \boldsymbol{\psi}{\in}\boldsymbol{\Psi}$. The second particularizes $\overline{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})}$ in the form of several directions of departure from specific assumptions using auxiliary regressions (fig. 15.7); see section 5.3.

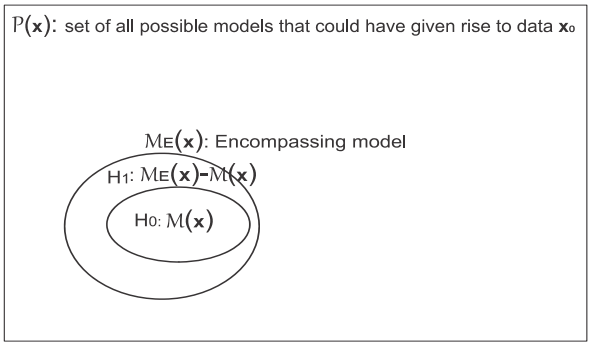


Fig. 9: M-S testing by encompassing

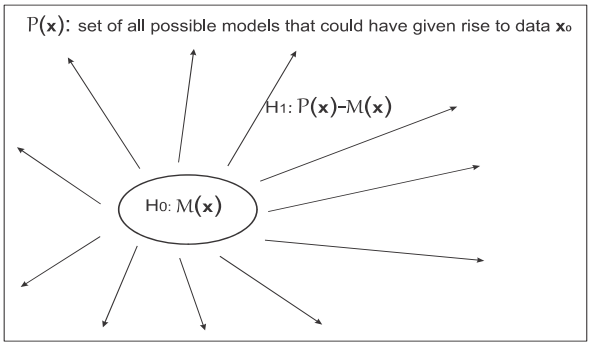


Fig. 10: M-S testing: directions of departures

Hence, **the primary role** of M-S testing is to probe, vis-a-vis data $\mathbf{x}_0$, for possible departures from $\mathcal{M}_\theta(\mathbf{x})$ beyond its boundaries, but within $\mathcal{P}(\mathbf{x})$, the set of all possible statistical models that could have given rise to $\mathbf{x}_0$. In this sense, the generic form of M-S testing is probing **outside** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$:

$$H_0{:}\ f^*(\mathbf{x}){\in}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\quad\text{vs.}\quad \overline{H}_0{:}\ f^*(\mathbf{x}){\in}\left[\mathcal{P}(\mathbf{x}){-}\mathcal{M}_\theta(\mathbf{x})\right],$$

where $f^*(\mathbf{x}){=}f(\mathbf{x};\boldsymbol{\theta}^*)$ denotes the 'true' distribution of the sample. The fact that M-S testing is probing $[\mathcal{P}(\mathbf{x}){-}\mathcal{M}_\theta(\mathbf{x})]$ raises certain technical and conceptual problems pertaining to how one can **operationalize** such investigating. In practice, one needs to replace the broad $\overline{H}_0$ with a more specific operational H_1. This operationalization has a very wide scope, extending from vague **omnibus** (local), to specific **directional** (broader) **alternatives**, like the tests based on the auxiliary autoregressions and the Skewness-Kurtosis test. In all cases, however, H_1 does *not* span $\overline{H}_0$, and that raises additional issues, including:

(a) The higher vulnerability of M-S testing to the fallacy of rejection: (mis)interpreting reject H_0 [evidence against H_0] as evidence for the specific H_1. Rejecting the null in

a M-S test provides evidence *against* the original model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, but that does *not* imply good evidence *for* the particular alternative H_1. Hence in practice one should **never** accept H_1 without further probing because that will be a classic example of the fallacy of rejection.

(b) In M-S testing the type II error [accepting the null when false] is often the more serious of the two errors. This is because for the type I error [rejecting the null when true] one will have another chance to correct the error at the respecification stage. When one, after a battery of M-S tests, erroneously concludes that $\mathcal{M}_{\theta}(\mathbf{x})$ is statistically adequate, one will proceed to draw inferences oblivious to the fact that the actual error probabilities might be very different from the nominal (assumed) ones.

(c) In M-S testing the objective is to **probe** $[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\theta}(\mathbf{x})]$ **as exhaustively as possible**, using a combination of omnibus M-S tests whose probing is more broad but have low power and directional M-S tests whose probing in narrower but goes much further and have higher power.

(d) Applying several M-S tests in probing the validity of one or a combination of assumptions does *not* necessarily increase the relevant type I error probability because the framing of the hypotheses of interest renders them different from the *multiple hypothesis testing problem* as construed in the N-P framework.

3.2 M-S testing and the Linear Regression model

The Normal, Linear Regression (LR) is undoubtedly the quintessential statistical model (table 15.6) in most applied fields, including econometrics. For this reason we will consider the question of M-S testing for this statistical model in more detail. The LR model can be viewed as a parametrization of a vector process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$, where $\mathbf{Z}_t{:=}(Y_t,\mathbf{X}_t)$, is assumed to be NIID. At the specification stage, evaluating whether the model assumptions [1]-[5] are likely to be valid for a particular data is non-trivial since all these assumptions pertain to the conditional process $\{(Y_t|\mathbf{X}_t{=}\mathbf{x}_t)\,,\ t\in\mathbb{N}\}$ which is not directly observable! One can indirectly assess the validity of [1]-[5], via the observable process $\{\mathbf{Z}_t{:=}(Y_t,\mathbf{X}_t),\ t\in\mathbb{N}\}$.

Table 6: 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:	$(Y_t\|X_t{=}x_t)\backsim \mathsf{N}(.,.)$,	$t\in\mathbb{N}$.
[2]	Linearity:	$E\,(Y_t\|X_t{=}x_t){=}\beta_0+\beta_1 x_t$,	
[3]	Homoskedasticity:	$Var\,(Y_t\|X_t{=}x_t){=}\sigma^2$,	
[4]	Independence:	$\{(Y_t\|X_t{=}x_t)\,,\ t\in\mathbb{N}\}$ indep. process,	
[5]	t-invariance:	$\boldsymbol{\theta}{:=}(\beta_0,\beta_1,\sigma^2)$ are *not* changing with t,	
$\beta_0{=}(\mu_1{-}\beta_1\mu_2)\in\mathbb{R}$, $\beta_1{=}(\sigma_{21}/\sigma_{22})\in\mathbb{R}$, $\sigma^2{=}(\sigma_{11}{-}(\sigma_{21})^2/\sigma_{22})\in\mathbb{R}_+$.			

3.2.1 Maximum Likelihood estimators (MLEs):

$$\widehat{\beta}_0=\overline{Y}-\widehat{\beta}_1\overline{x},\ \widehat{\beta}_1=\frac{\sum_{t=1}^n(Y_t-\overline{Y})(x_t-\overline{x})}{\sum_{t=1}^n(x_t-\overline{x})^2},\ \overline{Y}=\frac{1}{n}\sum_{t=1}^n Y_t,\ \overline{x}=\frac{1}{n}\sum_{t=1}^n x_t, \tag{25}$$

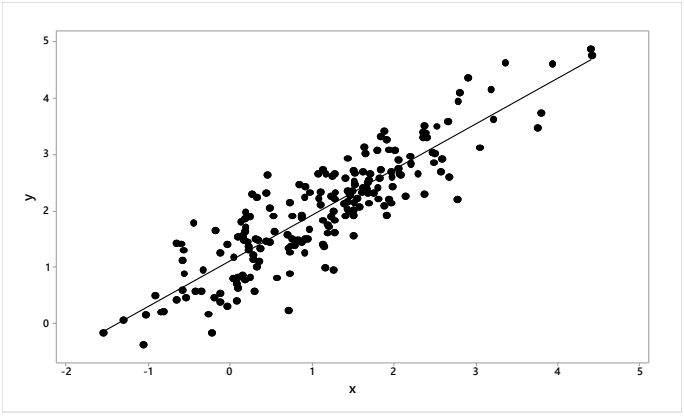
$$\widehat{\sigma}^2_{ML}=\frac{1}{n}\sum_{t=1}^n(Y_t-\widehat{\beta}_0-\widehat{\beta}_1x_t)^2=\frac{1}{n}\sum_{t=1}^n\widehat{u}_t^2, \tag{26}$$

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

Sampling distributions of MLEs $(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\sigma}^2_{ML})$. Given that X_t enters the NLR model as a conditioning variable, the above estimators are function of its observed values $(x_1,x_2,\ldots,x_n)$ and random variables $(Y_1,Y_2,\ldots,Y_n)$. In particular:

$$\widehat{\beta}_0 \backsim \mathsf{N}(\beta_0,\sigma^2\left(\tfrac{1}{n}+\varphi_x\overline{x}^2\right)),\ \ \widehat{\beta}_1 \backsim \mathsf{N}(\beta_1,\sigma^2\varphi_x),\ \ \tfrac{n\widehat{\sigma}^2_{ML}}{\sigma^2} \backsim \chi^2(n-2). \tag{27}$$

When assumptions [1]-[5] are valid, $(\widehat{\beta}_0,\widehat{\beta}_1)$ are unbiased, fully efficient, sufficient, strongly consistent. $\widehat{\sigma}^2_{ML}$ is biased, strongly consistent and asymptotically efficient; $s^2=\frac{1}{n-2}\sum_{t=1}^n(Y_t-\widehat{\beta}_0-\widehat{\beta}_1x_t)^2$ *is unbiased.*



Estimated regression line

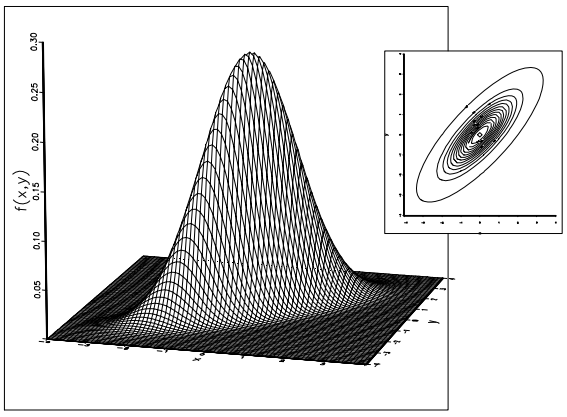


Fig.7.3: Bivariate Normal density with the equal probability contours

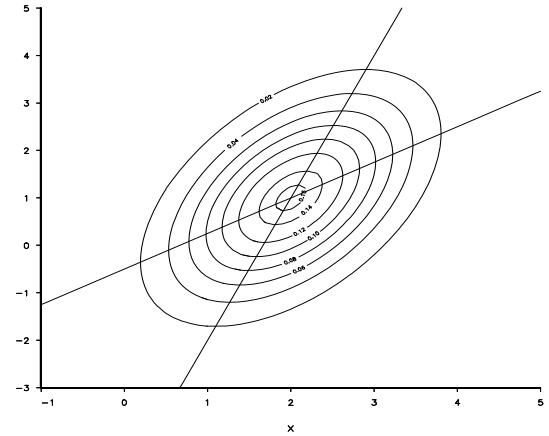


Fig. 7.4: Bivariate Normal contours, and regr. lines: $E(Y|X=x),\ E(X|Y=y)$

3.3 'Ideal data' for the LR model

The simulated data in the t-plots (fig. 7.11-7.12) and the scatterplot (fig. 7.13) provide the ideal graphs for the Normal, Linear Regression model; compare figure 7.13 with 7.3 and 7.4.

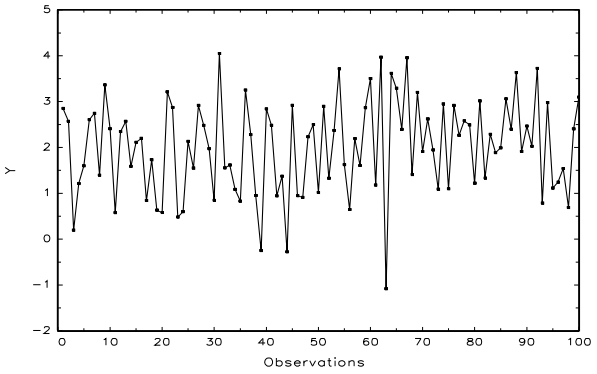


Fig. 7.11: t-plot of y_t

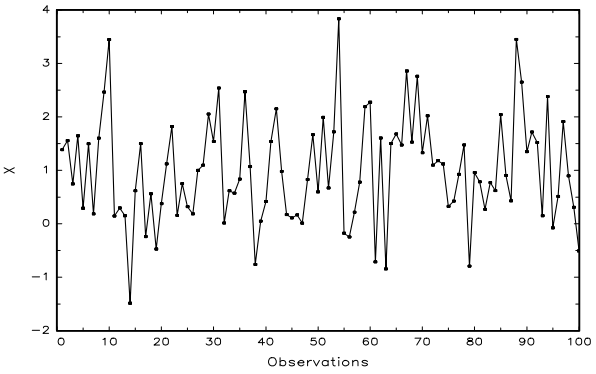


Fig. 7.12: t-plot of x_t

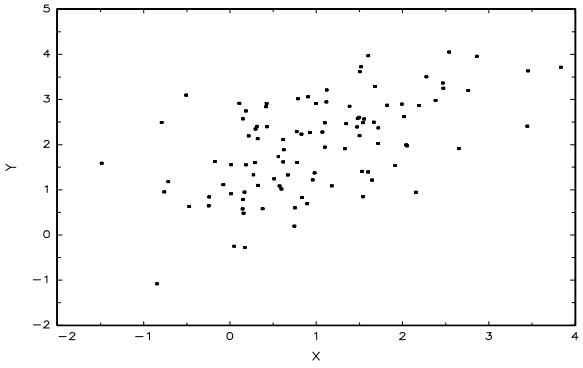


Fig. 7.13: Scatter-plot of (x_t, y_t)

Example 7.9. Consider the t-plots and scatterplot given in figures 7.14-7.16.

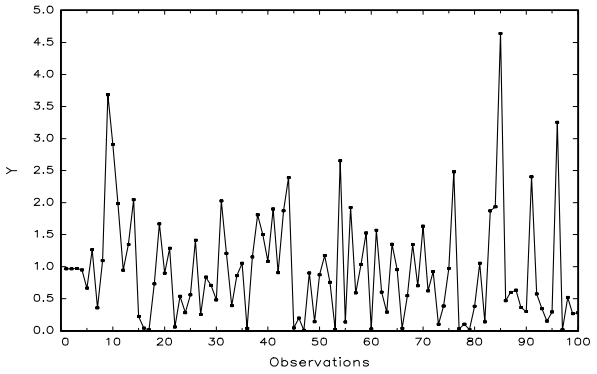


Fig. 7.14: t-plot of y_t

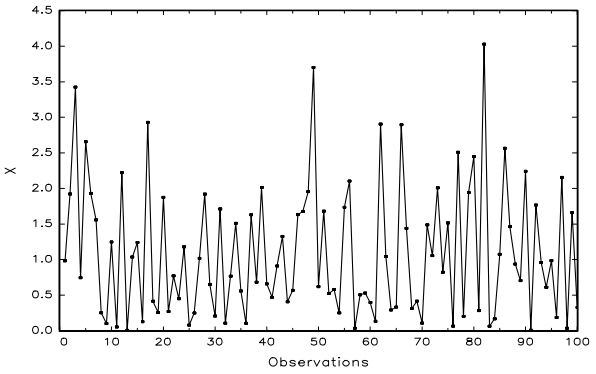


Fig. 7.15: t-plot of x_t

The t-plots exhibit no dependence or heterogeneity and thus one can proceed to 'read'

the scatterplot for chance regularity patterns suggesting a particular distribution.

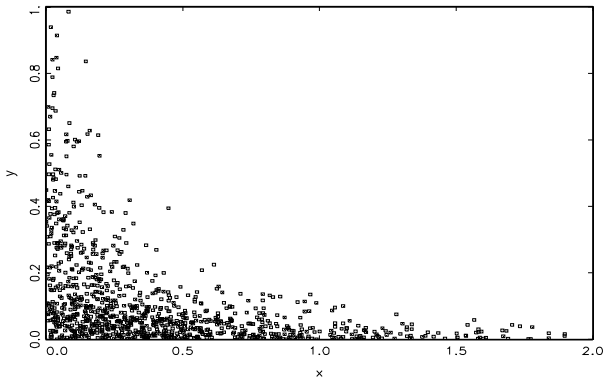


Fig. 7.16: Scatter-plot of (x_t, y_t)

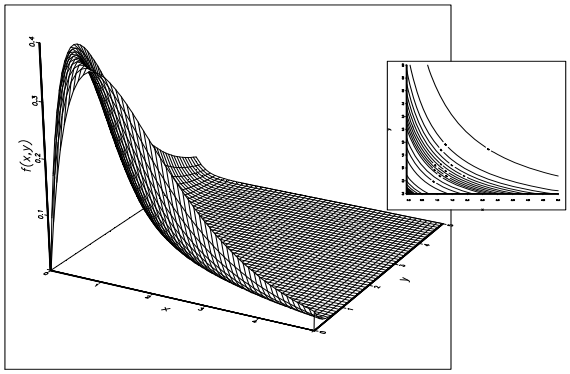


Fig. 7.17: Bivariate Exponential

3.4 Auxiliary regressions for M-S testing

Indicative *auxiliary regressions* for the simple *1-regressor case* can be used to *test jointly* the model assumptions [2]-[5] as different misspecifications might affect the first two conditional moments:

$$E\left(Y_t \mid X_t{=}x_t\right){=}\beta_0{+}\beta_1 x_t,\ Var\left(Y_t \mid X_t{=}x_t\right){=}\sigma^2. \tag{28}$$

The first auxiliary regression specifies how departures from different assumptions might affect the conditional mean:

$$\widehat{u}_t{=}\delta_0+\delta_1 x_t+\overbrace{\delta_2 t}^{\overline{[5]}}+\overbrace{\delta_3 x_t^2}^{\overline{[2]}}+\overbrace{\delta_4 x_{t-1}+\delta_5 Y_{t-1}}^{\overline{[4]}}+v_{1t}, \tag{29}$$

H_0: $\delta_1{=}\delta_2{=}\delta_3{=}\delta_4{=}\delta_5{=}0$ vs. H_1: $\delta_1{\neq}0$ or $\delta_2{\neq}0$ or $\delta_3{\neq}0$ or $\delta_4{\neq}0$ or $\delta_5{\neq}0$.

The second auxiliary regression specifies how departures from different assumptions might affect the constancy of conditional variance:

$$\widehat{u}_t^2{=}\gamma_0+\overbrace{\gamma_2 t}^{\overline{[5]}}+\overbrace{\gamma_1 x_t+\gamma_3 x_t^2}^{\overline{[3]}}+\overbrace{\gamma_4 x_{t-1}^2+\gamma_5 Y_{t-1}^2}^{\overline{[4]}}+v_{2t}, \tag{30}$$

H_0: $\gamma_1{=}\gamma_2{=}\gamma_3{=}\gamma_4{=}\gamma_5{=}0$ vs. H_1: $\gamma_1{\neq}0$ or $\gamma_2{\neq}0$ or $\gamma_3{\neq}0$ or $\gamma_4{\neq}0$ or $\gamma_5{\neq}0$.

Intuitively, the above auxiliary regressions should be viewed as attempts to probe the residuals $\{\widehat{u}_t,\ t{=}1,2,...,n\}$ for any remaining systematic information that has been overlooked by the specification of the regression and skedastic functions in (28) in terms of assumptions [1]-[5]. More formally, the extra terms in (29) and (30) will be zero since they should be orthogonal to $\widehat{u}_t$ and $\widehat{u}_t^2$ when assumptions [1]-[5] are valid for data $\mathbf{Z}_0$. As argued in Spanos (2010b), it is no accident that M-S tests are often specified in terms of the residuals; they often constitute a maximal ancillary statistic.

3.4.1 An empirical example: how to avoid blunders!

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:

$$\underset{(.009)}{Y_t=.020}+\underset{(.237)}{1.996x_t}+\widehat{u}_t,\ R^2=.536,\ s=.0498,\ n=64, \tag{31}$$

where the standard errors are given in parentheses.

On the basis of the estimated model in (31), 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 (31) being statistically adequate. But is it?

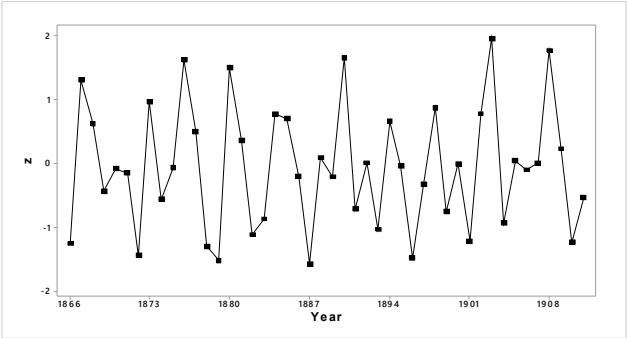


Fig. 15.14: t-plot of NIID data

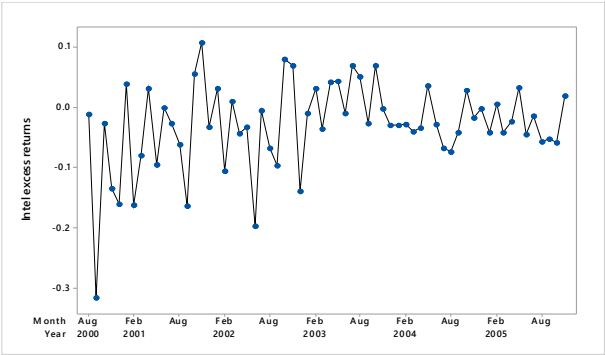


Fig. 14.9: Intel Corp. excess returns

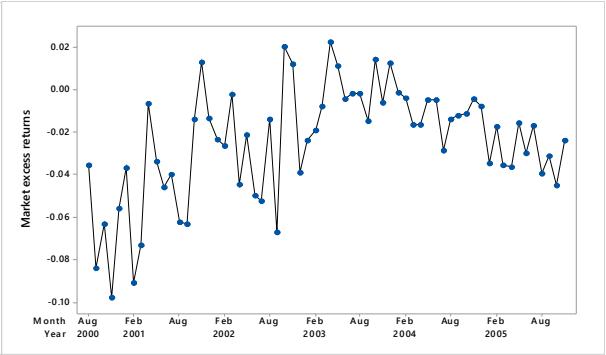


Fig. 14.10: Market excess returns

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 (31), 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 (32)-(33). 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 (32) probes for departures pertaining to assumptions about $E(Y_t|X_t{=}x)$ and the $\widehat{u}_t^2$-equation in (33) for departures from $Var(y_t|X_t{=}x)$.

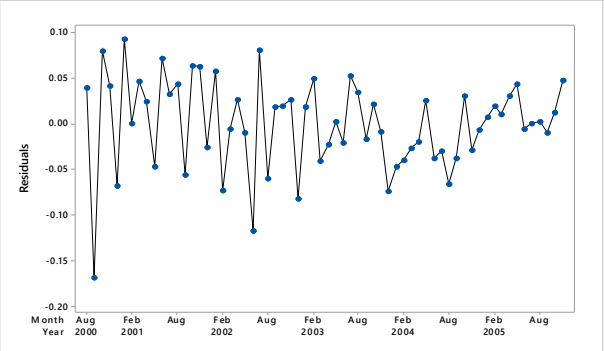


Fig. 14.11: t-plot of the residuals from (31)

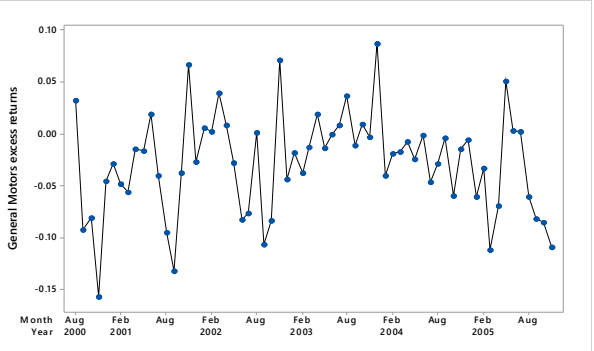


Fig. 14.12: GM excess returns

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

$$\widehat{u}_t = \underset{(.018)}{.047} + \underset{(.474)}{.761}x_t + \overbrace{\underset{(6.24)}{7.69}x_t^2}^{[2]} - \overbrace{\underset{(.044)}{.192}D_2}^{[5]} - \overbrace{\underset{(.338)}{1.11}t^2 + \underset{(.054)}{.175}t^3}^{[5]} - \overbrace{\underset{(.109)}{.320}\widehat{u}_{t-1}}^{[4]} + \widehat{v}_{1t}, \tag{32}$$

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

$$\widehat{u}_t^2 = \underset{(.0008)}{.004} + \overbrace{\underset{(.0025)}{.025}D_2}^{[5]} - \overbrace{\underset{(.002)}{.006}t}^{[5]} - \overbrace{\underset{(.179)}{.053}x_t^2}^{[3]} + \widehat{v}_{2t}, \quad R^2{=}.19,\ n{=}64 \tag{33}$$

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

As argued in chapter 5, applying a Normality test using the original model residuals is a good strategy 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. One can, however, get some idea of the validity of assumption [1] by applying such tests using the post-whitened residuals from the auxiliary equation (32). Hence, the above M-S testing results indicate clearly that no reliable inferences can be drawn on the basis of the estimated model in (31) since assumptions [4]-[5] are invalid.

3.4.2 Probing for substantive adequacy?

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 (31)? Adding z_{t-1} to the estimated equation (31) gives rise to:

$$Y_t=\underset{(.009)}{.013}+\underset{(.232)}{2.082}x_t-\underset{(.129)}{.296}z_{t-1}+\widehat{\epsilon}_t,\ R^2=.577,\ s=.0483,\ n=63. \tag{34}$$

Taking (34) 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!

3.4.3 Antique grandfather clock

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:

$$Y_t=\underset{(.966)}{1.312}+\underset{(.195)}{1.177}x_t+\widehat{u},\quad s=.208,\ n=32, \tag{35}$$

where the estimates of the unknown parameters are evaluated via:

$$\widehat{\beta}_0=\overline{Y}-\widehat{\beta}_1\overline{x}=7.148-1.177(4.958)=1.312,\ \widehat{\beta}_1=\frac{\frac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})(x_t-\overline{x})}{\frac{1}{n}\sum_{t=1}^{n}(x_t-\overline{x})^2}=\frac{.04225}{.0359}=1.177,$$

$$\widehat{\sigma}^2_{ML}=\frac{1}{n}\sum_{t=1}^{n}\widehat{u}_t^2=\frac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})^2-\frac{\left(\frac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})(x_t-\overline{x})\right)^2}{\frac{1}{n}\sum_{t=1}^{n}(x_t-\overline{x})^2}=.0905-\frac{(.0423)^2}{.0359}=.04065,$$

$$s^2=\frac{n}{(n-2)}\widehat{\sigma}^2_{ML}=\frac{32}{30}\,(.04065)=.04336,\ s=\sqrt{.04336}=.208, \tag{36}$$

and the numbers in brackets underneath the estimates denote their estimated standard errors stemming from (27):

$$SE(\widehat{\beta}_0)=s\sqrt{\left(\tfrac{1}{n}+\varphi_x\overline{x}^2\right)}=.966,\quad SE(\widehat{\beta}_1)=s\sqrt{\varphi_x}=.195.$$

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

$$\begin{array}{l}\overline{Y}=7.148,\ \overline{x}=4.958,\ \widehat{Var}(Y_t)=\frac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})^2=.0905,\\ \widehat{Var}(X_t)=\frac{1}{n}\sum_{t=1}^{n}(x_t-\overline{x})^2=.0359,\ \widehat{Cov}\,(Y_t,X_t)=\frac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})(x_t-\overline{x})=.0423.\end{array} \tag{37}$$

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.

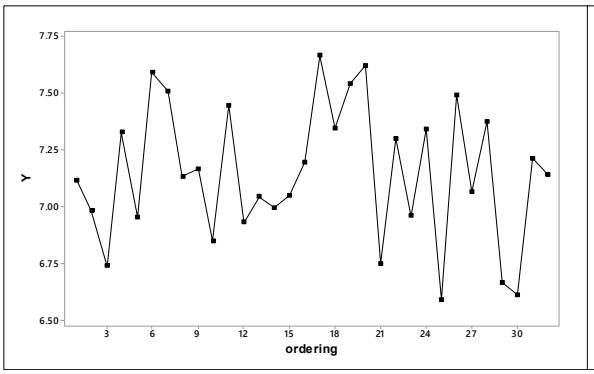


Fig. 14.1: t-plot of y_t

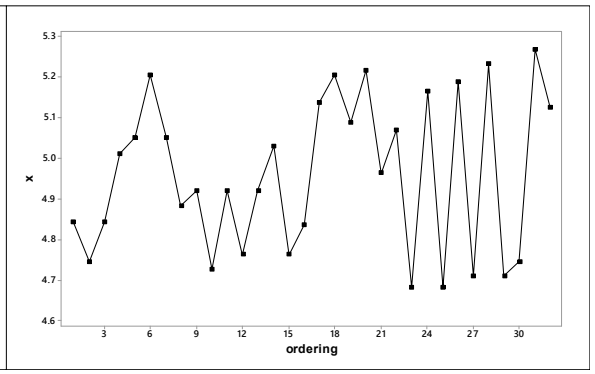


Fig. 14.2: t-plot of x_t

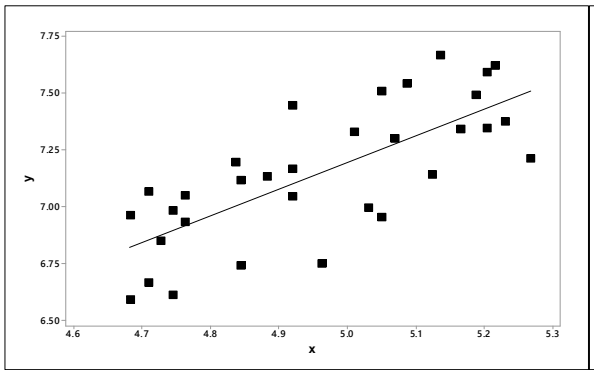


Fig. 14.3: Scatterplot of (x_t,y_t), t=1, ..., n

Fig. 14.4: t-plot of the residuals from (35)

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.

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.

Example 14.2 (continued).During a presentation of the estimated regression model:

$$Y_t = \underset{(.966)}{1.312} + \underset{(.195)}{1.177} x_t + \widehat{u}, \quad s=.208,\ n=32, \tag{38}$$

an economist in the audience raised the possibility that (38) is *substantively inadequate* because a crucial explanatory variable, x_{2t}-the number of bidders has been omitted.

3.4.4 Probing for substantive adequacy for real!

The modeler decides to evaluate this and re-estimates (38) with the additional regressor:

$$Y_t = -\underset{(.382)}{1.316} + \underset{(.070)}{1.418} x_{1t} + \underset{(.044)}{.649} x_{2t} + \widehat{u}_t,\ s=.0718,\ R^2=.947,\ n=32. \tag{39}$$

and its residuals are plotted in figure 14.13.

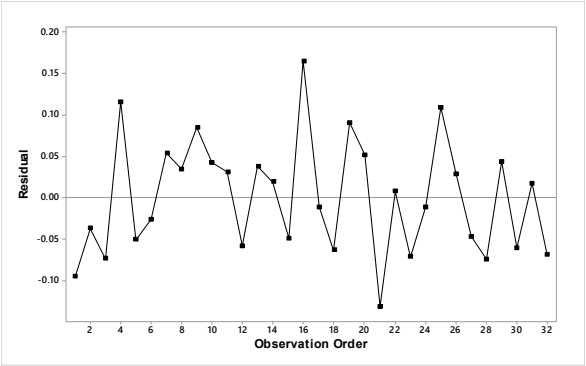


Fig. 14.13: t-plot of the residuals of (39)

In light of the fact that (38) 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 (38). The t-plots of the residuals from (39) confirm that the respecified structural model has retained the statistical adequacy.

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

Example 14.3: Anscombe's (1973) data. Anscombe (1973) contrived four pairs of data series (Y_{it}, x_{it}), i=1, 2, 3, 4, t=1, 2, ..., 11, to ensure that each pair had identical sample moments:

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

giving rise to numerically identical estimated regression results:

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

yielding 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.

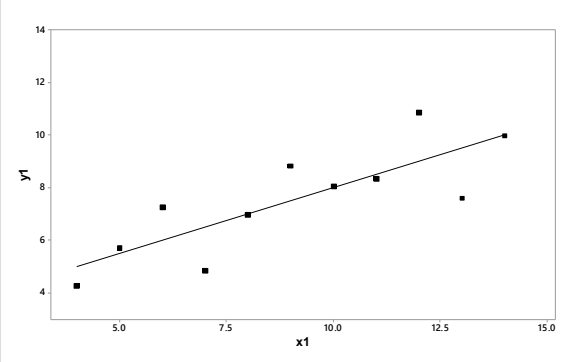


Fig. 14.5: Scatterplot of (x_{1t}, y_{1t}), t=1, ..., n

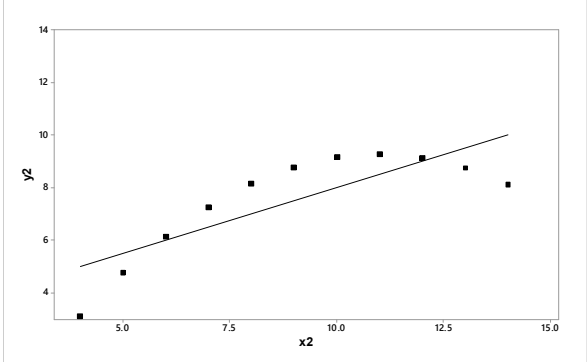


Fig. 14.6: Scatterplot of (x_{2t}, y_{2t}), t=1, ..., n

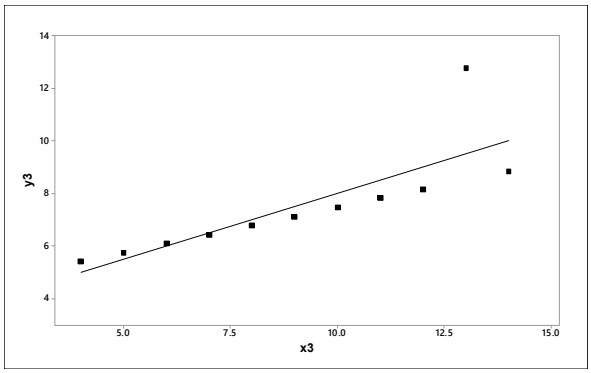


Fig. 14.7: Scatterplot of (x_{3t}, y_{3t}), t=1, ..., n

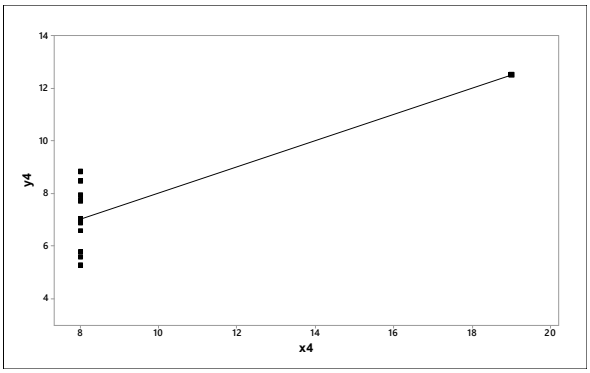


Fig. 14.8: Scatterplot of (x_{4t}, y_{4t}), t=1, ..., n

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.

4 An empirical example from economics

4.1 The traditional curve-fitting perspective

The *structural model* underlying Keynes' Absolute Income Hypothesis (AIH):

$$C{=}\alpha+\beta Y^D,\ \alpha>0,\ 0<\beta<1,$$

is often transformed into a statistical model by attaching the error term $\{u_t,\ t{\in}\mathbb{N}\}$:

$$C_t{=}\alpha+\beta Y_t^D+u_t,\ u_t\backsim \mathsf{NIID}(0,\sigma^2),\ t{=}1,2,...,n. \tag{40}$$

The commonly used justification for the error term is that it represents errors of approximation, omitted effects and anything what is left is non-systematic error! The implicit statistical model underlying (40) is the Linear Regression (LR) model (table 15.4). That is, the statistical and substantive models are assumed to coincide. The relevant data $\mathbf{z}_0{:=}\{(y_1,x_1),...,(y_n,x_n)\}$, are annual USA time series data for the period 1947-1998: y_t-real consumer's expenditure and x_t-personal disposable income.

Example. Estimating (40) yields:

$$\underset{(16.930)}{y_t{=}-45.279}\ \underset{(.007)}{+.936x_t}+\widehat{u}_t,\ R^2{=}.997,\ s{=}49.422,\ n{=}52. \tag{41}$$

The goodness of fit ($R^2{=}.997$) seems 'excellent' and the coefficients appear to be 'highly significant':

$$\tau_0(y){=}\tfrac{45.279}{16.930}{=}2.675[.004],\quad \tau_1(y){=}\tfrac{.936}{.007}{=}133.71[.000]. \tag{42}$$

Looking at the above estimators and tests from a substantive perspective, $\widehat{\beta}{=}.936$ appears to have the correct sign and magnitude ($0<\beta<1$)! Does the estimated model (41) provide *evidence for* the AIH? Before the probabilistic assumptions of the underlying statistical model are tested, one cannot draw any trustworthy evidence from (41).

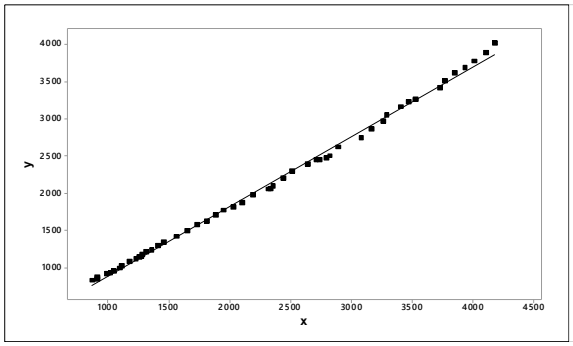


Fig. 15.19: Scatter-plot and fitted line

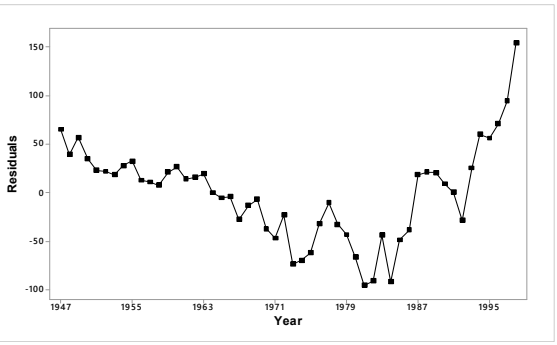


Fig. 15.20: Residuals form the fitted line

Strong hints about the serious statistical misspecifications associated with the estimated LR model in (41) are given by the t-plot of its residuals in fig. 15.18, which look very different from a realization of a NIID process, exhibiting both a non-linear trend and cycles. The scatterplot in fig. 15.19 is clearly misleading because the two

data series mean-heterogeneity as well as irregular cycles [see figures 15.20-25], but it is often used in macroeconomic textbooks as evidence for the appropriateness of linearity in the LR model.

Formal M-S testing confirms that almost all probabilistic assumptions of the Normal, LR model are *invalid* for this data, as the M-S testing results in table 15.8 indicate.

Table 15.8: Mis-Specification (M-S) tests	
Normality:	S-K=1.803[.406]?
Linearity:	$F(1,45)$=3.529[.0005]
Homoskedasticity:	$F(2,45)$=16.318[.000005]
Independence:	$F(2,45)$=11.45[.00006]
t-invariance:	$F(1,45)$=4.235[.023]

4.2 The Probabilistic Reduction (PR) approach

How does the PR approach address the statistical adequacy problem? Using informed specification, M-S testing and respecification guided by data plots.

4.2.1 Specification

What are the probabilistic assumptions pertaining to the process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ for the LR model? NIID! Are they appropriate for the consumption function data?

(D) **Distribution**	(M) **Dependence**	(H) **Heterogeneity**
Normal	Independent	Identically Distributed

It is clear from the t-plots (figures 15.20-21) that both data series are mean-trending and exhibit cycles.

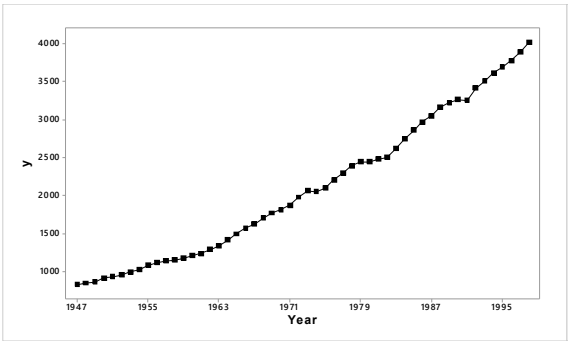


Fig. 15.21: t-plot of y_t

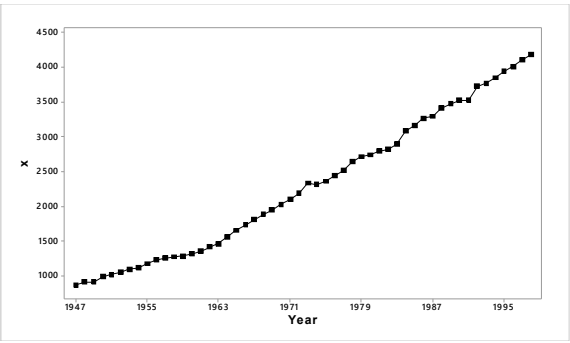


Fig. 15.22: t-plot of x_t

To get a better view of the latter let us subtract the trend using the auxiliary regression:

$$z_t=\delta_0+\delta_1 t+\delta_2 t^2+v_t,\ t=1,2,..,n, \tag{43}$$

and take the residuals: $\{\widehat{v}_t=(z_t-\widehat{\delta}_0-\widehat{\delta}_1 t-\widehat{\delta}_2 t^2)$. This exercise corresponds to the philosopher's counterfactual (what if) reasoning! The residuals from (43) for the

two series (detrended) are plotted in figures 15.18-19.

(D) **Distribution**	(M) **Dependence**	(H) **Heterogeneity**
Normal?	Independent?	mean-heterogeneous

It is clear from figures 15.22–23 that both series exhibit Markov-type temporal dependence.

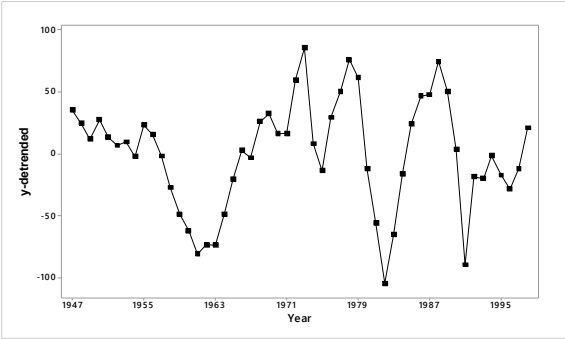


Fig. 15.23: t-plot of y_t-detrened

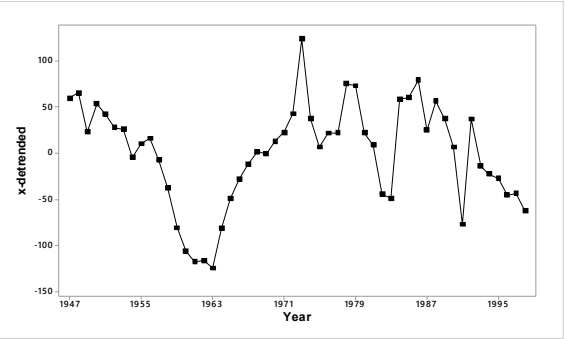


Fig. 15.24: t-plot of x_t-detrened

To assess the underlying distribution we need to subtract that dependence as well, which we can generically do using the extended auxiliary regression:

$$z_t = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \gamma_3 z_{t-1} + \gamma_4 z_{t-2} + \epsilon_t, \;\; t = 1, 2, .., n, \tag{44}$$

and plot the residuals which we call detrended and dememorized data series (see figures 15.24-25).

(D) **Distribution**	(M) **Dependence**	(H) **Heterogeneity**
Normal?	Markov	mean-heterogeneous

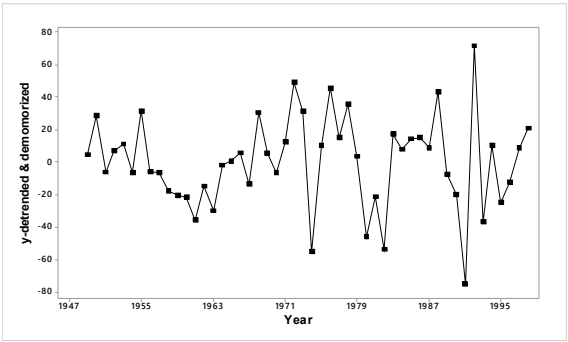


Fig. 15.25: y_t-detrended and dememorized

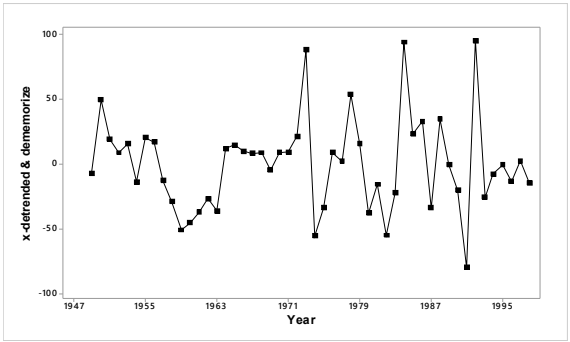


Fig. 15.26: x_t-detrended and dememorized

The t-plots in figures 15.24-25 indicate a trending variance; the variation around the mean increases with t. In addition, the scatter-plot of the two series in figure 15.26 indicates clear departures from the elliptically shaped plot associated with a bivariate Normal distribution.

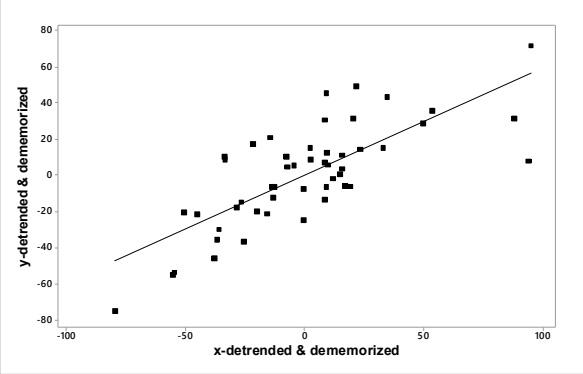


Fig. 15.27: Scatterplot of detrended and demomorized data

(D) **Distribution**	(M) **Dependence**	(H) **Heterogeneity**
Normal?	Markov	mean-heterogeneous
Non-symmetric		variance-heterogeneous?

It is important to note that non-Normality leads to drastic respecifications because both the regression and skedastic functions need to be re-considered.

4.2.2 Mis-Specification (M-S) testing

A. Joint Mis-Specification (M-S) tests for model assumptions [1]-[5]

Regression function tests. In view of the chance regularity patterns exhibited by the data in figures 15.20-26, the test that suggests itself would be based on the auxiliary regression:

$$\widehat{u}_t=\gamma_0+\gamma_1 x_t+\overbrace{\gamma_2 t}^{\overline{[5]}}+\overbrace{\gamma_3 x_t^2}^{\overline{[2]}}+\overbrace{\gamma_4 x_{t-1}+\gamma_5 x_{t-1}}^{\overline{[4]}}+v_{1t},$$

$$H_0\text{: } \gamma_2=\gamma_3=\gamma_4=\gamma_5=0.$$

$$\underset{(52.3)}{\widehat{u}_t=205.5}-\underset{(.082)}{.389}x_t+\underset{(2.61)}{5.37}t-\underset{(.0085)}{.03}\,x_t^2+\underset{(.128)}{.594}y_{t-1}-\underset{(.117)}{.384}x_{t-1}+\widehat{v}_t, \qquad (45)$$

$$R^2=.86,\ s=19.1,\ n=51$$

Note that the scaling of variables, such as x_t^2 is crucial in practice to avoid large approximation errors. The F test for the joint significance of the terms $t, x_t^2, y_{t-1}, x_{t-1}$, yields:

$$F(4,45)=\left(\tfrac{117764-16421}{16421}\right)\left(\tfrac{45}{5}\right)=55.544[.0000000],$$

indicating clearly that this estimated regression is badly misspecified. To get a better idea as to departures from the individual assumptions, let us consider the significance of the relevant coefficients for each assumption separately:

Mean-heterogeneity $\overline{[5]}$ (γ_2=0): $\tau_2(\mathbf{y};45)=\left(\tfrac{5.37}{2.61}\right)=2.058[.023]$,

Non-Linearity $\overline{[2]}$ (γ_3=0): $\tau_3(\mathbf{y};45)=\left(\tfrac{.03}{.0085}\right)=3.529[.0005]$,

Dependence $\overline{[4]}$ (γ_4=γ_5=0): $F(\mathbf{y};2,45)=\left(\tfrac{24778-16421}{16421}\right)\left(\tfrac{45}{2}\right)=11.45[.00006]$

$(\gamma_4=0)$: $\tau_4(\mathbf{y};45)=(\frac{.594}{.128})=4.641[.000015]$, $(\gamma_5=0)$: $\tau_5(\mathbf{y};45)=(\frac{.384}{.117})=3.282[.0001]$

It is important to NOTE that $[\tau_i(\mathbf{y};45)]^2=F(\mathbf{y};1,45),\ i=2,3.$

Skedastic function tests. The auxiliary regression that suggests itself is:

$$(\widehat{u}_t/s)^2=\delta_0+\overbrace{\delta_1 t}^{\overline{[5]}}+\overbrace{\delta_2 x_t^2}^{\overline{[3]}}+\overbrace{\delta_3(\widehat{u}_{t-1}/s)^2}^{\overline{[4]}}+v_{2t},$$
$$H_0:\ \delta_1=\delta_2=\delta_3=0.$$

$$(\widehat{u}_t/s)^2=\underset{(83.83)}{110.2}-\underset{(.043)}{.057}t-\underset{(.127)}{.252}x_t^2+\underset{(.174)}{.874}(\widehat{u}_{t-1}/s)^2+\widehat{v}_{2t}, \tag{46}$$

The F test for the joint significance of the terms $t,\ x_t^2$ and $\widehat{u}_{t-1}^2$ yields:

$$F(3,45)=\tfrac{134.883-66.484}{66.484}\left(\tfrac{45}{3}\right)=15.432[.00000],$$

indicating clearly that some of the model assumptions pertaining to the conditional variance are misspecified. To shed additional light on which assumptions are to blame for the small p-value, let us consider the significance of the relevant coefficients for each assumption separately:

Variance heterogeneity: $\overline{[5]}$ $(\delta_1=0)$: $\tau_1(\mathbf{y};45)=\left(\frac{.057}{.043}\right)=1.326[.194]$,

Heteroskedasticity: $\overline{[3]}$ $\delta_2=\delta_3=0$: $F(2,45)=\left(\frac{114.7-66.484}{66.484}\right)\left(\frac{45}{2}\right)=16.318[.000005]$,

where the latter indicates the presence of heteroskedasticity!

CAUTION. If one were to use the auxiliary regression:

$$(\tfrac{\widehat{u}_t}{s})^2=\delta_0+\delta_1 t+v_{2t}^*,\quad (\tfrac{\widehat{u}_t}{s})^2=-\underset{(273.6)}{82.2}+\underset{(.014)}{.042}t+v_{2t}^*,$$

one would have *erroneously* concluded that [5] is invalid since $\tau_1(\mathbf{y};45)=(\frac{.042}{.014})=3.01[.004]$. This brings out the importance of joint M-S testing to avoid misdiagnosis!

In summary, the M-S testing based on the auxiliary regressions (45)-(46) indicates that there are clear departures from assumptions [2]-[5]. If one were to ignore that and proceed to test the Normality assumption [1], the testing result is likely to be unreliable because as mentioned above, all current M-S tests for Normality assume the validity of assumptions [2]-[5]. To see this, let us use the skewness-kurtosis:

$$SK(\mathbf{x}_0)=\tfrac{52}{6}(.031)^2+\tfrac{52}{24}(3.91-3)^2=1.803[.406],$$

which indicates no departures from [1], but is that a reliable diagnosis? No, see below!

Traditional respecification: embracing the fallacy of rejection. At this point it will be interesting to follow the traditional respecification of misspecified models by embracing the fallacy of rejection and simply adding the additional terms found

to be significant in the above M-S testing based on auxiliary regressions. Estimating an extended regression equation yields:

$$\begin{array}{c} y_t = \underset{(52.0)}{163.7} + \underset{(3.18)}{7.75}t - \underset{(.122)}{.159}t^2 + \underset{(.105)}{.462}x_t + \underset{(.022)}{.057}x_t^2 + \underset{(.130)}{.556}y_{t-1} - \underset{(.132)}{.302}x_{t-1} + \widehat{\varepsilon}_t, \\ R^2 = .9997,\ s = 18.957,\ n = 51, \end{array} \tag{47}$$

Apart from the obvious fact that (47) makes no statistical sense since the specification is both ad hoc and internally inconsistent (Spanos, 1995b), a glance at the residuals from this estimated equations raises serious issues of statistical misspecification stemming from a t-varying conditional variance; see figure 15.27.

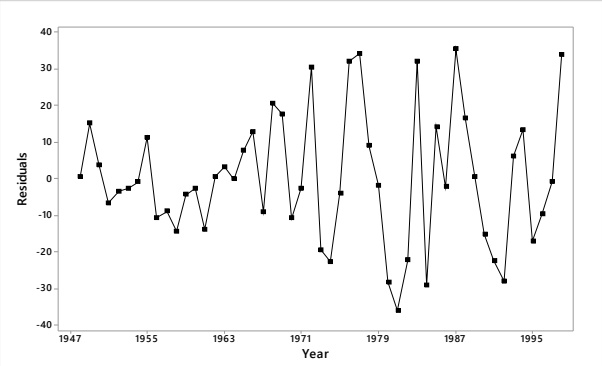


Fig. 15.28: Residuals from (47)

4.2.3 Probabilistic Reduction Respecification

The combination of M-S testing and graphical techniques suggest the following probabilist structure for the process $\{\ln \mathbf{Z}_t,\ t \in \mathbb{N}\}$:

(D) **Distribution**	(M) **Dependence**	(H) **Heterogeneity**
Log-Normal	Markov	mean-heterogeneous

where the logarithm is used as a variance stabilizing transformation; see Spanos (1986). Let us take the logs of the original data series.

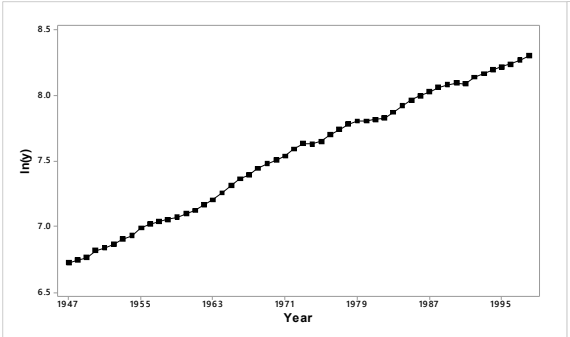


Fig. 15.29: t-plot of $\ln y_t$

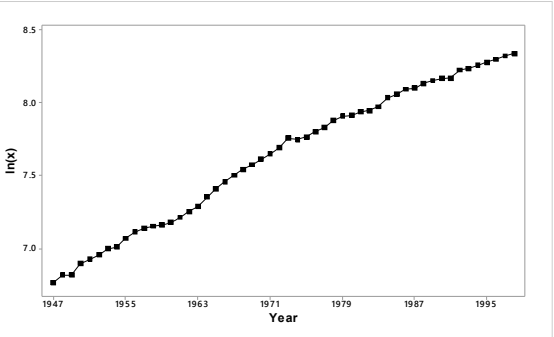


Fig. 15.30: t-plot of $\ln x_t$

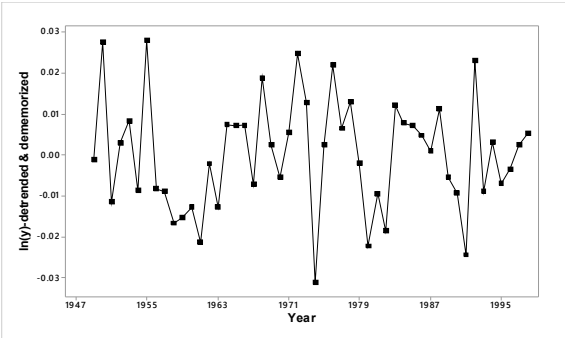


Fig. 15.31: t-plot of $\ln y_t$-detrended & dememorized

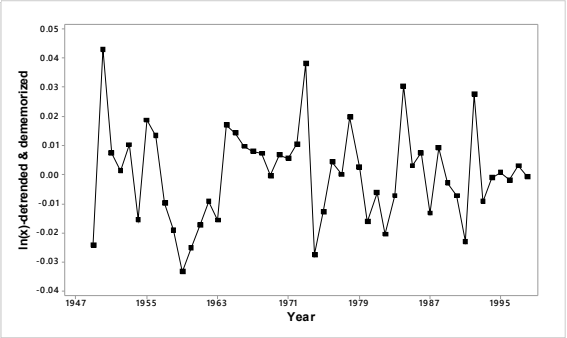


Fig. 15.32: t-plot of $\ln x_t$-detrended & dememorized

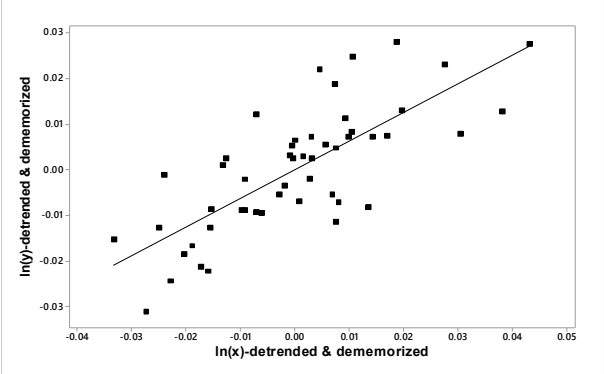


Fig. 15.33: Scatterplots of $(\ln x_t, \ln y_t)$ detrended and dememorized

If one were to compare figures 15.24-25 with 15.30–31, it becomes clear that the logarithm has acted as a variance stabilizing transformation because the t-varying variances in the latter disappear; see Spanos (1986). In addition the scatter plot in fig. 15.32 associated with the data in figures 15.30-31, indicates no departures from the elliptical shape we associate with the bivariate Normal distribution. Imposing the Reduction assumptions:

(D) **Distribution**	(M) **Dependence**	(H) **Heterogeneity**
Log-Normal	Markov	mean-heterogeneous Separable Heterogeneity?

on the joint distribution: $f(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n; \boldsymbol{\phi})$ where $\mathbf{Z}_t{:=}(y_t, x_t)$, $y_t{:=}\ln Y_t, x_t{:=}\ln X_t$ gives rise to the reduction:

$$f(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n; \boldsymbol{\phi}) \overset{\text{M}}{=} f_t(\mathbf{z}_1; \boldsymbol{\psi}_1) \prod_{t=2}^{n} f_t(\mathbf{z}_t|\mathbf{z}_{t-1}; \boldsymbol{\varphi}_t) \overset{\text{SH}}{=} f_t(\mathbf{z}_1; \boldsymbol{\psi}_1) \prod_{t=2}^{n} f(\mathbf{z}_t|\mathbf{z}_{t-1}; \boldsymbol{\varphi}).$$

Reducing $f(\mathbf{z}_t|\mathbf{z}_{t-1}; \boldsymbol{\varphi})$ further by conditioning on $X_t{=}x_t$, yields:

$$f(\mathbf{z}_t|\mathbf{z}_{t-1}; \boldsymbol{\varphi}){=}f(y_t|x_t, \mathbf{z}_{t-1}; \boldsymbol{\varphi}_1){\cdot}f(x_t|\mathbf{z}_{t-1}; \boldsymbol{\varphi}_2),$$

with $f(y_t|x_t,\mathbf{z}_{t-1};\boldsymbol{\varphi}_1)$ is the distribution underlying the Dynamic Linear Regression [DLR(1)] model with a statistical GM (table 15.10):

$$y_t=\delta_0+\delta_1 t+\alpha_1 x_t+\alpha_2 y_{t-1}+\alpha_2 x_{t-1}+u_t,\ t\in\mathbb{N}, \tag{48}$$

which can be thought of as a hybrid of the LR and AR(1) models.

Table 15.10: Normal, Dynamic Linear Regression model

Statistical GM:		$y_t=\delta_0+\delta_1 t+\alpha_1 x_t+\alpha_2 y_{t-1}+\alpha_3 x_{t-1}+u_t,\ t\in\mathbb{N}.$	
[1]	Normality:	$(y_t\|x_t,\mathbf{Z}_{t-1})\backsim \mathsf{N}(.,.),$	$t\in\mathbb{N}.$
[2]	Linearity:	$E(y_t\|x_t,\mathbf{Z}_{t-1})=\delta_0+\delta_1 t+\alpha_1 x_t+\alpha_2 y_{t-1}+\alpha_3 x_{t-1},$	
[3]	Homoskedasticity:	$Var(y_t\|x_t,\mathbf{Z}_{t-1})=\sigma_0^2,$	
[4]	Markov:	$\{(y_t\|x_t,\mathbf{Z}_{t-1}),\ t\in\mathbb{N}\}$ indep. process,	
[5]	t-invariance:	$(\delta_0,\delta,\alpha_1,\alpha_2,\alpha_3,\sigma_0^2)$ are *not* changing with t.	

Example 15.10 (contiued). Estimating the DLR model in (48) yields:

$$\begin{aligned}\ln Y_t=&\underset{(.272)}{.912}+\underset{(.001)}{.005}t+\underset{(.069)}{.708}\ln x_t+\underset{(.108)}{.565}\ln Y_{t-1}-\underset{(.097)}{.413}\ln x_{t-1}+\widehat{\varepsilon}_t,\\ &R^2=.904\ [\text{reported: } R^2=.9997],\ s=.0085,\ n=51,\end{aligned} \tag{49}$$

Estimating the two auxiliary regressions for the first two conditional moments yields:

$$\widehat{u}_t=\underset{(2.51)}{.30}+\underset{(.056)}{.028}t-\underset{(.631)}{.022}\ln x_t-\underset{(.128)}{.338}\ln Y_{t-1}+\underset{(.223)}{.308}x_{t-1}+\underset{(.014)}{.007}t^2-\underset{(.041)}{.001}(\ln x_t)^2+\underset{(.297)}{.446}\widehat{u}_{t-1}+\widehat{v}_{1t},$$

$$(\widehat{u}_t/s)^2=\underset{(13.1)}{-13.9}-\underset{(.138)}{.155}t-\underset{(.287)}{.324}(x_t^2/1000)+\underset{(.153)}{.036}(\widehat{u}_{t-1}/s)^2+\widehat{v}_{2t},$$

which indicates no departures from the probabilistic assumptions of the underlying DLR model in table 15.10.

Comparing Keynes' AIH with the *statistically adequate model in* (49) we can *infer* that the substantive model is clear false on the basis of this data.

5 Summary and Conclusions

Approximations, limited data and various forms of uncertainty lead to the use of **statistical models** in learning from data $\mathbf{x}_0$ about phenomena of interest when such data exhibit chance regularity patterns. All statistical methods (**frequentist, Bayesian, nonparametric**) rely on a prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ as the primary basis of inference. The sound application and the objectivity of their methods turns on the **validity** of the probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ vis-a-vis data $\mathbf{x}_0$.

Fundamental aim of modeling: How to *specify* and *validate* statistical models in light of substantive information pertaining to the phenomenon of interest.

Unfortunately, both the *specification* and *model validation* have been serously neglected in empirical modeling during the last 100 years of modern statistics. Statistical modeling and inference is primarily viewed as curve-fitting guided by goodness-of-fit ignoring the fact that the vary notion of goodness-of-fit is ambivalent when the probabilistic assumptions imposed on one's data are invalid. Undue reliance of asymptotic arguments of the form 'as $n \to \infty$' stems from inadequate understanding of how such arguments "are logically devoid of content about what happens at any particular n." (Le Cam, 1986).

At best, one can find more of a **grab-bag of techniques** than a systematic account of all facets of modeling: specification, M-S testing and respecification with view to establish a statistically adequate model that accounts for all the systematic information in one's data.

Error statistics attempts to remedy that by proposing a coherent account of both modeling (specification, M-S testing, respecification) and inference (estimation, testing, prediction) facets that puts the entire process of learning from data about stochastic phenomena of interest on a more sound philosophical footing (Spanos, 1986, 1999, 2000, 2010; Mayo and Spanos, 2004).

Crucial strengths of frequentist **error statistical** methods in this context:

▶ There is a clear goal to achieve: the statistical model is sufficiently adequate so that the optimality of the inference procedures is secured and the *actual* error probabilities approximate closely the *nominal* ones. This is needed if any learning from data takes place.

▶ It supplies a trenchant battery of Mis-Specification (M-S) tests for model-validation (non-parametric and parametric) with a view to minimize both types of errors and generate a **reliable diagnosis** thru self-correcting testing procedures.

▶ In offers a **seemless transition** from model validation to subsequent use in the sense that the same error statistical reasoning is used. The focus is on the question: What is the nature and warrant for frequentist error statistical model specification and validation?

Failing to grasp the correct **rationale of M-S testing** has led many to think that merely finding a statistical model that 'fits' the data well in some sense is tantamount to showing it is statistically adequate. **Excellent goodness-of-fit is neither necessary nor sufficient for reliable inferences and trustworthy evidence!**

Minimal Principle of Evidence: if the procedure had no capacity to uncover departures from a hypothesis or claim H, then not finding any is poor evidence for H.

Failing to satisfy so minimal a principle leads to models which, while acceptable according to its own self-scrutiny, such as goodness-of-fit, are in fact inadequate and usually give rise to untrustworthy evidence.

6 Appendix A: M-S testing-methodological issues

In this section we discuss some of the key criticisms of M-S testing in order to bring out some of the confusions they conceal.

6.1 The multiple testing (comparisons) issue

The multiple testing (comparisons) issue arises in the context of joint N-P tests because the overall significance level does not coincide with that of the individual hypotheses. Viewing the auxiliary regressions in (29) and (30) as providing the basis of two joint N-P tests, and choosing a specific significance level, say .025 associated with testing each individual assumptions, such as δ_4=δ_5=0 for departures from [4], implies that the overall significance level α of the F-test for H_0 will be greater than .025. How are these two thresholds related?

Let us assume that we have m individual null hypotheses (linear in $\boldsymbol{\theta}$), $H_0(\theta_i),\ i$=1, ..., m, pertaining to a prespecified statistical model:

$$\boxed{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,\ \text{for } \boldsymbol{\theta}\in\Theta\subset\mathbb{R}^m,\ m<n,} \tag{50}$$

, saysuch that $H_0\,(\boldsymbol{\theta})=\cup_{i-1}^{m}H_0(\theta_i)$, and the overall F-test rejects H_0 when the smallest p-value, $p_i(\mathbf{x}_0)$ associated with each $H_0(\theta_i)$, is less than α. This can be framed in the form of:

$$\{\min_{1\leq i\leq m}(p_1(\mathbf{x}_0),\ \ldots\ ,p_m(\mathbf{x}_0))<\alpha\}=\bigcup_{i=1}^{m}\{p_i(\mathbf{x}_0)<\alpha\}.$$

To evaluate the probability associated with this rejection rule, one can use the sampling distribution of the p-value under H_0 when $p(\mathbf{X};\boldsymbol{\alpha})$ is viewed as a function of the sample and α, known to be Uniform: $p(\mathbf{X};\alpha)\backsim\mathsf{U}(0,1)$, for $\alpha\in(0,1)$,
i.e. $\mathbb{P}(p(\mathbf{X};\alpha)<\alpha)=\alpha$; see Cox and Hinkley (1974). Using Boole's inequality (chapter 2) we can deduce that:

$$\mathbb{P}\left(\bigcup_{i=1}^{m}\{p_i(\mathbf{x}_0)<\alpha\}\right)\leq\sum_{i=1}^{m}\{p_i(\mathbf{X};\alpha)<\alpha\}=m\alpha.$$

That is, the overall significance level of the joint test for H_0 is $m\alpha$.

In light of that, a simplistic rule of thumb for controling the overall (joint) significance level at α is to use $\frac{\alpha}{m}$ for the individual hypotheses $H_0(\theta_i),\ i$=1, ..., m. That is, the rejection rule for the individual hypotheses should be:

$$\text{Reject } H_0(\theta_i) \text{ when } p_i(\mathbf{x}_0)<\tfrac{\alpha}{m}.$$

This is known as the *Bonferroni rule*; see Lehmann and Romano (2005).

In the case of the M-S tests relating to the auxiliary regressions in (29) and (30), $m\leq 3$, and one could use this rule to avoid over-rejection. The relevant significance level for infering that H_0 is false, i.e. $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is misspecified, is that of the joint test, but the tests for the individual assumptions can shed light on how to respecify the model. Having said that, a more serious problem for the choice of α stems from large n problem as it relates to the p-values and accept/reject H_0 rules; see chapter 13.

CAUTION: the multiple hypotheses problem is often misleadingly defined more broadly as applying too many tests to the same data $\mathbf{x}_0$, insinuating that such a large number of inferences must be illegitimate; an unwarranted claim since multiple tests of the same assumptions correct each other.

6.2 Securing the effectiveness/reliability of M-S testing

There are a number of strategies designed to enhance the effectiveness/reliability of M-S probing thus render the diagnosis more reliable.

■ A most efficient way to probe $[\mathcal{P}(\mathbf{x})-\mathcal{M}_\theta(\mathbf{x})]$ is to construct M-S tests by modifying the original tripartite **partitioning** that gave rise to $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in directions of educated departures gleaned from Exploratory Data Analysis. This gives rise to encompassing models or directions of departure, which enable one to eliminate an *infinite* number of alternative models at a time; Spanos (1999). This should be contrasted with a most inefficient way to do this, that involves probing $[\mathcal{P}(\mathbf{x})-\mathcal{M}_\theta(\mathbf{x})]$ *one model at a time* $\mathcal{M}_{\boldsymbol{\varphi}_i}(\mathbf{x})$, $i=1,2,...$ This is a hopeless task because there is an infinite number of such alternative models to probe for and eliminate.

■ **Judicious combinations** of omnibus (non-parametric), directional (parametric) and simulation-based tests, probing as broadly as possible and upholding dissimilar assumptions. The interdependence of the model assumptions, stemming from $\mathcal{M}_\theta(\mathbf{x})$ being a parametrization of the process $\{X_t,\ t\in\mathbb{N}\}$, plays a crucial role in the self-correction of M-S testing results.

■ **Astute ordering** of M-S tests so as to exploit the interrelationship among the model assumptions with a view to 'correct' each other's diagnosis. For instance, the probabilistic assumptions [1]-[3] of the Normal, Linear Regression model (table 8) are interrelated because all three stem from the assumption of Normality for the vector process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$, where $\mathbf{Z}_t:=(Y_t, X_t)$, assumed to be NIID. This information is also useful in narrowing down the possible alternatives. It is important to note that the Normality assumption [1] should be tested last because most of the M-S tests for it assume that the other assumptions are valid, rendering the results questionable when that clause is invalid.

■ **Joint M-S tests** (testing several assumptions simultaneously) designed to avoid 'erroneous' diagnoses as well as minimize the maintained assumptions.

The above strategies enable one to argue with **severity** that when no departures from the model assumptions are detected, the model provides a reliable basis for inference, including appraising substantive claims (Mayo & Spanos, 2004).

6.3 The infinite regress and circularity charges

The *infinite regress* charge is often articulated by claiming that each M-S test relies on a set of assumptions, and thus it assesses the assumptions of the model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ by invoking the validity of its own assumptions, trading one set of assumptions with another *ad infinitum*. Indeed, some go as far as to claim that this reasoning is

often *circular* because some M-S tests inadvertently assume the validity of the very assumption they aim to test!

A closer look at the reasoning underlying M-S testing reveals that both charges are misplaced.

■ **First**, the scenario used in evaluating the type I error invokes no assumptions beyond those of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, since every M-S test is evaluated under:

$$\boxed{H\text{: } \textit{all} \text{ the probabilistic assumptions of } \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) \text{ are valid.}}$$

Moreover, when any one (or more) of the model assumptions is rejected, the model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, as a whole, is considered misspecified.

Example. In the context of the simple Normal model (table 6), the *runs test* is an example of an omnibus M-S test for assumptions [2]-[4]. The original data, or the residuals, are replaced with a + when the next data point is an up and with a − when it's a down. A *run* is a sub-sequence of one type (+ or −) immediately preceded and succeeded by an element of the other type.

For $n \geq 40$, the type I error probability evaluation is based on:

$$Z_R(\mathbf{X}) = \frac{R-([2n-1]/3)}{\sqrt{[16n-29]/90}} \overset{[1]\text{-}[4]}{\backsim} \mathsf{N}(0,1).$$

It is important to emphasize that the runs test is *insensitive* to departures from Normality, and thus the effective scenario for deriving the type I error is under assumptions [2]-[4].

■ **Second**, the power for any M-S test, is determined by evaluating the test statistic under certain forms of departures from the assumptions being appraised [no circularity], but retaining the rest of the model assumptions.

For the runs test, the evaluation of power is based on:

$$Z_R(\mathbf{X}) \overset{[1]\&\overline{[2]}\text{-}\overline{[4]}}{\backsim} \mathsf{N}(\delta, \tau^2),\ \delta \neq 0,\ \tau^2 > 0,$$

where $\overline{[2]}$-$\overline{[4]}$ denote specific departures from these assumptions considered by the test in question. However, since the test is insensitive to departures from [1], the effective scenario does not have any retained assumptions. One of the advantages of nonparametric tests is that they are insensitive to departures from certain retained assumptions.

Bottom line: in M-S testing the evaluations under the null and alternative hypotheses invoke only the model assumptions; no additional assumptions are involved. Moreover, the use of joint M-S tests aims to minimize the number of model assumptions retained when evaluating under the alternative.

6.4 Illegitimate double-use of data charge

In the context of the error statistical approach it is certainly true that the same data $\mathbf{x}_0$ are being used for two different purposes:

► (a) to test primary hypotheses in terms of the unknown parameter(s) $\boldsymbol{\theta}$, and

► (b) to assess the validity of the probabilistic assumptions comprising the pre-specified model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$,

but 'does that constitute an illegitimate double-use of data?'

Mayo (1981) answered that question in the *negative*, arguing that the original data $\mathbf{x}_0$ are commonly *remodeled* to $\mathbf{r}_0=\mathbf{G}(\mathbf{x}_0)$, $\mathbf{r}_0\in\mathbb{R}^k$, $k\leq n$, and thus rendered distinct from $\mathbf{x}_0$ when testing $\mathcal{M}_{\theta}(\mathbf{x})$'s assumptions:

"What is relevant for our purposes is that the data used to test the probability of heads [primary hypothesis] is distinct from the data used in the subsequent test of independence [model assumption]. Hence, no illegitimate double use of data is required." (Mayo, 1981, p. 195).

Hendry (1995), p. 545, interpreted this statement to mean:

"... following Mayo (1981), diagnostic test information is effectively independent of the sufficient statistics, so 'discounting' for such tests is not necessary."

Combining these two views offers a more formal answer.

First, (a) and (b) pose very different questions to data $\mathbf{x}_0$. M-S testing is asking the question: ► are data $\mathbf{x}_0$ a truly typical realization of the stochastic mechanism specified by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$? That is: ► could the generating mechanism $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ have generated data $\mathbf{x}_0$?

On the other hand, all forms of statistical inference presuppose that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ could have generated data $\mathbf{x}_0$ and they pose questions that will reduce the original parameter space in order to learn from data $\mathbf{x}_0$ about the true statistical Data-Generating Mechanism (DGM):

$$\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\},\ \mathbf{x}\in\mathbb{R}^n_X.$$

Second, the M-S testing is probing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and N-P testing and estimation procedures are probing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

Indeed, one can go further to argue that the answers to the questions posed in (a) and (b) **rely on distinct information** in data $\mathbf{x}_0$.

Under certain conditions, the sample can be split into two components:

$$\mathbf{X}\rightarrow(\mathbf{S}(\mathbf{X}),\ \mathbf{R}(\mathbf{X}))$$

that induce the following reduction in $f(\mathbf{x};\boldsymbol{\theta})$:

$$f(\mathbf{x};\boldsymbol{\theta})=c\cdot f(\mathbf{s};\boldsymbol{\theta})\cdot f(\mathbf{r}),\ \forall\,(\mathbf{s},\mathbf{r})\in\mathbb{R}^m_s\times\mathbb{R}^{n-m}_r,$$

where $c=|J|$ is the Jacobian of: $\mathbf{X}\rightarrow(\mathbf{S}(\mathbf{X}),\ \mathbf{R}(\mathbf{X}))$,

$\mathbf{S}(\mathbf{X}):=(S_1,...,S_m)$ is a *complete sufficient* statistic,

$\mathbf{R}(\mathbf{X}):=(R_1,...,R_{n-m})$, a *maximal ancillary* statistic, and

$\mathbf{S}(\mathbf{Z})$ and $\mathbf{R}(\mathbf{Z})$ are independent.

What does this reduction mean?

$$f(\mathbf{z};\boldsymbol{\theta})=c\cdot\overbrace{f(\mathbf{s};\boldsymbol{\theta})}^{\textbf{inference}}\cdot\overbrace{f(\mathbf{r})}^{\textbf{model validation}} \tag{51}$$

► [a] all primary inferences are based exclusively on $f(\mathbf{s};\boldsymbol{\theta})$, and

► [b] $f(\mathbf{r})$ can be used to validate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ using error probabilities that are free of $\boldsymbol{\theta}$.

Example. For the simple Normal model (table 4) holds for $\mathbf{S}{:=}(\overline{X}_n, s^2)$:

$$\overline{X}_n = \tfrac{1}{n}\sum_{k=1}^{n} X_k, \qquad s^2{=}\tfrac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2,$$

as the minimal sufficient statistic, and $\mathbf{R}(\mathbf{X}){=}(\widehat{v}_3,..,\widehat{v}_n)$, where:

$$\widehat{v}_k{=}\ (\sqrt{n}(X_k{-}\overline{X}_n)/s) \backsim \mathsf{St}(n{-}1),\ k{=}1,2,..,n,$$

known as the **studentized residuals**, the maximal ancillary statistic.

► This explains why M-S testing is often based on **the residuals**, and confirms Mayo's (1981) claim that $\mathbf{R}(\mathbf{X}){=}(\widehat{v}_3,..,\widehat{v}_n)$ provides information distinct from $\mathbf{R}(\mathbf{X})$ upon which the primary inferences are based.

The crucial argument for relying on $f(\mathbf{r})$ for model validation purposes is that the probing for departures from $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is based on error probabilities that do not depend on $\boldsymbol{\theta}$.

Generality of result in (51). This result holds for almost all statistical models routinely used in statistical inference, including the simple Normal, the simple Bernoulli, the Linear Regression and related models and all statistical models based on the (natural) Exponential family of distributions, such as the Normal, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, Poisson, Wishart, geometric, Laplace, Levy, log-Normal, Pareto, Weibull, binomial (with fixed number of trials), multinomial (with fixed number of trials), and negative binomial (with fixed number of failures) and many others. Finally, the above result in (51) holds 'approximately' in all cases of statistical models whose inference relies on asymptotic Normality.

6.5 Revisiting the pre-test bias argument

Most traditional econometric textbooks indicate that Mis-Specification (M-S) testing and respecification are vulnerable to the pre-test bias charge.

To discuss the merits of the **pre-test bias charge**, consider the Durbin-Watson test, for assessing the assumption of no autocorrelation for the linear regression errors, based on (see Greene, 2000):

$$H_0 : \rho = 0, \text{ vs. } H_1 : \rho \neq 0,$$

Step 1. The pre-test bias perspective interprets this M-S test as equivalent to choosing between two models:

$$\begin{aligned} \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}) :&\quad y_t = \beta_0 + \beta_1 x_t + u_t, \\ \mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z}) :&\quad y_t = \beta_0 + \beta_1 x_t + u_t,\ u_t{=}\rho u_{t-1} + \varepsilon_t. \end{aligned} \tag{52}$$

Step 2. This is formalized in decision-theoretic language into a choice between two estimators of β_1, conceptualized in terms of the *pre-test estimator*:

$$\ddot{\beta}_1{=}\lambda\widehat{\beta}_1 + (1{-}\lambda)\widetilde{\beta}_1,\ \lambda{=}\begin{cases} 1, & \text{if } H_0 \text{ is accepted} \\ 0, & \text{if } H_0 \text{ is rejected;} \end{cases} \tag{53}$$

$\widehat{\beta}_1$ is the OLS estimator under H_0, and $\widetilde{\beta}_1$ is the GLS estimator under H_1.

Step 3. This perspective claims that the relevant error probabilities revolve around the MSE $E(\ddot{\beta}_1-\beta_1)^2$, whose sampling distribution is usually non-Normal, biased and has a highly complicated variance (Leeb and Pötscher, 2005).

▶ The pre-test bias argument, based on (53), is **highly questionable** primarily because it ***ignores the relevant error probabilities***.

First, it misinterprets M-S testing by recasting it as a decision-theoretic estimation problem. As argued discerningly by Hacking (1965), pp. 31:

"Deciding that something *is* the case differs from deciding to *do* something."

M-S testing asks whether $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ *is* statistically adequate, i.e. it accounts for the chance regularities in data $\mathbf{z}_0$ or not.

▶ It is not concerned with selecting between two models come what may.

Second, even if one were to frame an M-S testing inference as concerned with a comparison between $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ and a broader alternative model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ arising from narrowing $[\mathcal{P}(\mathbf{z})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$, one cannot ignore the relevant errors:

(i) the selected model is inadequate but the other model is adequate, or

(ii) both models are inadequate.

In contrast, $E(\ddot{\beta}_1-\beta_1)^2$ evaluates the **expected loss** for each $\beta_1\in\mathbb{R}$, resulting from the modeler's supposedly tacit intention to use $\ddot{\beta}_1$ as an estimator of β_1.

▼Is there a connection between $E(\ddot{\beta}_1-\beta_1)^2$, for all $\beta_1\in\mathbb{R}$, and the errors (i)-(ii)?

The short answer is none. The former evaluates the expected loss stemming from one's (misguided) *intentions*, but the latter pertain to the relevant error probabilities (type I & II) associated with the inference that one of the two models is statistically adequate. The latter errors are based on hypothetical (testing) reasoning, but the former are risk evaluations based on an arbitrary loss function.

▶ How does one evaluate the 'loss' arising from a statistically misspecified model? The only relevant discrepancy is that between the relevant actual and nominal error probabilities; *not* some discrepancy between two models.

Third, the case where an M-S test supposedly selects the alternative $(\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z}))$, the implicit inference is that $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ is statistically adequate. This constitutes a classic example of **the fallacy of rejection** [evidence *against* H_0 is misinterpreted as evidence *for* H_1]. The validity of $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ needs to be established separately by thoroughly testing its own assumptions. Hence, in a M-S test one should *never* accept the alternative without further testing; see Spanos (2000).

Fourth, the case where a M-S test supposedly selects the null $(\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}))$, the implicit inference is that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is statistically adequate.

This inference is problematic for two reasons.

▼ *Firstly*, given the multitude of assumptions constituting a model, there is no single comprehensive M-S test based on a parametrically encompassing model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$, that could, by itself, establish the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$.

▼ *Secondly*, the inference is vulnerable to **the fallacy of acceptance** [*no* evidence

against H_0 is misinterpreted as evidence *for* it]. It is possible that the particular M-S test did not reject $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ because it had very low power to detect an existing departure.

In practice this can be remedied using additional M-S tests with higher power to cross-check the results, or/and use a post-data evaluation of inference to establish the warranted discrepancies from H_0.

■ **To summarize**, instead of devising ways to circumvent the fallacies of rejection and acceptance and avoid erroneous inferences in M-S testing, the pre-test bias argument embraces these fallacies by recasting the original problem (in step 1), formalizes them (in step 2), and evaluates risks (in step 3) that have no bearing on erroneously inferring that the selected model is statistically adequate.

▼ The pre-test bias charge is ill-conceived because **it misrepresents model validation as a choice between two models come what may**.