Statistical Model Validation: Mis-Specification (M-S) Testing and Respecification

Aris Spanos [March 2023]

1. Introduction

Why model validation is crucially important

2. Model-based frequentist inference: a bird's eye view

Parametric statistical models and sampling distributions

3. Statistical misspecification and unreliable inferences

How statistical misspecification affects all approaches to inference:
frequentist, Bayesian, nonparametric and Akaike-type selection procedures

4. Securing the adequacy of a statistical model

Separating the modeling from the inference facet
Mis-Specification (M-S) vs. N-P testing

5. Securing the adequacy in practice: empirical examples

6. Conclusions

1 Introduction

All approaches to empirical modeling using statistical analysis involve three basic components.
(1) Questions of substantive interest (however vague or specific),
(2) the **relevant data** $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$ selected to shed light on these questions, and
(3) a set of **probabilistic assumptions** invoked by the inference procedures employed to learn from data about these questions.

The totality of these probabilistic assumptions comprise the (implicit) **statistical model** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ which provides the **inductive premises** of inference.
All forms of statistical analysis, including descriptive statistics, depend crucially on the **statistical adequacy** (approximate validity)**,** vis-a-vis data $\mathbf{x}_0$, of the assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.
The foundational issue on **how to secure the statistical adequacy of the inductive premises** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ has bedeviled statistical inference since the 1930s. Its importance stems from the fact that statistical adequacy provides the **key** for securing the **reliability of inference and the trustworthiness of em-**

pirical evidence.

Although the untrustworthiness of empirical evidence has dominated current discussions in the **replication crisis** literature for more than a decade, establishing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ has been totally ignored by this literature**.** Instead, its has been on **several abuses of frequentist testing** as being the primary contributors to the untrustworthiness of published empirical evidence. My view is that the replication crisis literature is **missing the forest by focusing on a few trees**! Why?

The problem of establishing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is multifaceted and a lot more intricate than **foisting an a priori postulated model** $\mathcal{M}_{\varphi}(\mathbf{x})$ on the data and claiming statistical significance/non-significance of key parameters of interest on the basis of a computer software output.

Regrettably, as Rao (2004) points out, the **statistical adequacy** of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is often ignored in statistics courses: "They teach statistics as a deductive discipline of deriving consequences from given premises $[\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$. The need for examining the premises, which is important for practical applications of results of data analysis, is seldom emphasized." ... "The current statistical methodology is mostly model-based, without any specific rules for model selection or validating a specified model." (p. 2)

2 Model-based frequentist inference: a bird's eye view

Model-based frequentist statistical inference, pioneered by Fisher (1922), revolves around a prespecified parametric statistical model, whose generic form is:

$$\boxed{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,\ \text{for } \Theta\subset\mathbb{R}^m,\ m<n.} \tag{2.0.1}$$

where $f(\mathbf{x};\boldsymbol{\theta})$ denotes **the (joint) distribution of the sample $\mathbf{X}$**$:=(X_1, X_2, ..., X_n)$, $\mathbb{R}^n_X$ denotes the sample space, and Θ the parameter space; see Spanos (2019).

From a purely probabilistic perspective, $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ can be viewed as **a particular parameterization** $(\boldsymbol{\theta}\in\Theta)$ of the observable process $\{X_k,\ k\in\mathbb{N}:=(1,2,...,n,...,)\}$, underlying data $\mathbf{x}_0$.

$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ provides a description of a **statistical mechanism** which is selected on the basis that it *could have* given rise to data $\mathbf{x}_0$, or equivalently, data $\mathbf{x}_0$ constitute a 'typical' realization of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$'.

Example 1. Consider the case where the question of interest relates to the mean $(E(X_t)=\mu)$ of a stochastic process $\{X_t,\ t\in\mathbb{N}\}$, and the relevant $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is a simple Normal model (Table 1).

Table 1: The simple Normal model

Statistical GM:		$X_t=\mu+u_t,\ t\in\mathbb{N}:=(1,2,...,n,...)$
[1]	Normal:	$X_t \backsim \mathsf{N}(.,.),\ x_t\in\mathbb{R},$
[2]	Constant mean:	$E(X_t)=\mu,\ \mu\in\mathbb{R},\ \forall t\in\mathbb{N},$
[3]	Constant variance:	$Var(X_t)=\sigma^2>0,\ \forall t\in\mathbb{N},$
[4]	Independence:	$\{X_t,\ t\in\mathbb{N}\}$ is an independent process.

The distribution of the sample $f(\mathbf{x};\boldsymbol{\theta})$, $\boldsymbol{\theta}:=(\mu,\sigma^2)$, encapsulates all the probabilistic assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ since:

$$f(\mathbf{x};\boldsymbol{\theta})\overset{[4]}{=}\prod_{k=1}^{n}f_k(x_k;\boldsymbol{\theta}_k)\overset{[2]\text{-}[3]}{=}\prod_{k=1}^{n}f(x_k;\boldsymbol{\theta})\overset{[1]}{=}(\tfrac{1}{\sqrt{2\pi\sigma^2}})^n\exp\left\{-\tfrac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2\right\},\ \mathbf{x}\in\mathbb{R}^n.$$

The *main objective* of model-based frequentist inference is to '**learn from data** $\mathbf{x}_0$**' about** $\boldsymbol{\theta}^*$ using statistical 'approximations' relating to the neighborhood of $\boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ denotes the 'true' value of $\boldsymbol{\theta}$ in Θ.

That is shorthand for saying that there exists a $\boldsymbol{\theta}^*\in\Theta$ such that $\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\}$, $\mathbf{x}\in\mathbb{R}^n_X$, could have generated $\mathbf{x}_0$.

The cornerstone of frequentist inference is the concept of a **sampling distribution**, $f(y_n;\boldsymbol{\theta})$, for all $(\forall)$ $y\in\mathbb{R}_Y$, of a statistic $Y_n=g(X_1, X_2, ..., X_n)$ (estimator, test, predictor), which is derived directly from $f(\mathbf{x};\boldsymbol{\theta})$ via:

$$F_n(y)=\mathbb{P}(Y_n\leq y)=\underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\ g(\mathbf{x})\leq y\}} f(\mathbf{x};\boldsymbol{\theta})d\mathbf{x},\ \ \forall y\in\mathbb{R}_Y, \tag{2.0.2}$$

where $f(y_n;\boldsymbol{\theta})=dF_n(y)/dy$.

It is important to emphasize that $\boldsymbol{\theta}$ in (2.0.2) is prespecified using **two different forms of reasoning**:

(i) **factual** (estimation and prediction): presuming that $\boldsymbol{\theta}=\boldsymbol{\theta}^*$, and

(ii) **hypothetical** (hypothesis testing): H_0: $\boldsymbol{\theta}\in\Theta_0$ (presuming $\boldsymbol{\theta}\in\Theta_0$) vs. H_1: $\boldsymbol{\theta}\in\Theta_1$ (presuming $\boldsymbol{\theta}\in\Theta_1$); see Spanos (2019).

The **sampling distribution**, $f(y_n;\boldsymbol{\theta})$, $\forall y\in\mathbb{R}_Y$, frames the uncertainty relating to the fact that data $\mathbf{x}_0$ constitutes a single realization (out of $\forall\mathbf{x}\in\mathbb{R}^n_X$) of $\mathbf{X}$.

$f(y_n;\boldsymbol{\theta})$, $\forall y\in\mathbb{R}_Y$ is used to **calibrate the capacity** (optimality) of the inference procedure in terms of the relevant error probabilities (type I, II, power).

The derivation of $f(y_n;\boldsymbol{\theta})$, $\forall y\in\mathbb{R}_Y$, in (2.0.2) presumes the validity of $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$, and thus Fisher (1922) calls for establishing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$:

"For empirical as the specification of the hypothetical population $[\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts." (p. 314).

Fisher goes on to emphasize the importance of model validation and the crucial role of Mis-Specification (M-S) testing to **provide an empirical justification for statistical induction**:

"The possibility of developing complete and self-contained tests of goodness of fit deserves very careful consideration, since therein lies our justification for the free use which is made of empirical frequency formulae." (p. 314)

Statistical induction differs from other variants of induction in so far as its justification is empirical, stemming from the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and not from any a priori stipulations.

The **statistical adequacy** (approximate validity) of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ plays a crucial role in securing the **reliability of inference** because it assures 'control' of the relevant probabilities.

When $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is **statistically misspecified**:

(a) $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$, is erroneous, and thus the likelihood function, $L(\boldsymbol{\theta};\mathbf{x}_0)\propto f(\mathbf{x}_0;\boldsymbol{\theta})$, $\boldsymbol{\theta}\in\Theta$, is invalid, rendering unreliable both, the **Akaike-type** model selection procedures, and **Bayesian inference** based on the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x}_0)\propto\pi(\boldsymbol{\theta})\cdot L(\boldsymbol{\theta};\mathbf{z}_0),\ \boldsymbol{\theta}\in\Theta$; see Spanos (2010).

(b) An invalid $f(\mathbf{x};\boldsymbol{\theta})$ **distorts** the sampling distribution $f(y_n;\boldsymbol{\theta})$ in (2.0.2).

(c) A distorted $f(y_n;\boldsymbol{\theta})$ gives rise to '**non-optimal**' inference procedures and **sizeable discrepancies** between the **actual** and **nominal** error probabilities.

(d) A distorted $f(y_n;\boldsymbol{\theta})$ renders the inference results **unreliable** and the ensuing **evidence untrustworthy**. Applying a .05 significance level test when the actual type I error probability is closer to .97 (see Table 2), will yield spurious and untrustworthy evidence (Spanos and McGuirk, 2001).

3 Misspecification and inference: a first view

Statistical adequacy ensures that the **notional** (derived presuming $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is valid) **optimal properties of estimators and tests are actual**, and secures the reliability and precision of inference by guaranteeing that the relevant **actual error probabilities approximate closely the nominal ones**.

3.1 Actual vs. nominal error probabilities

Simulation example (Spanos and McGuirk, 2001). Consider a simulation with $N{=}10,000$ replications with $n{=}100$ based on the following two scenarios.

S1. The **estimated LR model is statistically adequate**. That is, the probabilistic assumptions of [1]-[5] (Table 4) are valid for data $\mathbf{z}_0$; the true and estimated models coincide: $Y_t{=}\beta_0{+}\beta_1 x_t{+}u_t$.

S2. $Y_t{=}\beta_0{+}\beta_1 x_t{+}u_t$ is **estimated**, but the **true model** is $Y_t{=}\delta_0{+}\delta_1 t{+}\beta_1 x_t{+}u_t$. This renders invalid assumption [5] since $\beta_0(t){=}\delta_0{+}\delta_1 t$.

Table 4: Normal, Linear Regression (LR) model

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

$$\left.\begin{array}{lll}
[1] & \text{Normality:} & (Y_t|X_t{=}x_t) \backsim \mathsf{N}(.,.), \\
[2] & \text{Linearity:} & E\left(Y_t|X_t{=}x_t\right){=}\beta_0 + \beta_1 x_t, \\
[3] & \text{Homosk/city:} & Var\left(Y_t|X_t{=}x_t\right){=}\sigma^2, \\
[4] & \text{Independence:} & \{\left(Y_t|X_t{=}x_t\right),\ t{\in}\mathbb{N}\}\ \text{independent process}, \\
[5] & \text{t-invariance:} & \boldsymbol{\theta}{:=}\left(\beta_0, \beta_1, \sigma^2\right)\ \text{are } \textit{not} \text{ changing with } t,
\end{array}\right\}\ t{\in}\mathbb{N}.$$

Table 5: Linear Regression (LR) and misspecification

	S1: Adequate LR model				**S2: Misspecified LR model**			
Replications: N=10000	True: Y_t=1.5+0.5x_t+u_t, Estim: Y_t=β_0+$\beta_1 x_t$+u_t,				True: Y_t=1.5+$\underline{.13t}$+.5x_t+u_t Estim: Y_t=β_0+$\beta_1 x_t$+u_t,			
	n=50		n=100		n=50		n=100	
Parameters	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$[\beta_0$=1.5$]$ $\widehat{\beta}_0$	1.502	.122	1.500	.087	0.462	.450	0.228	.315
$[\beta_1$=.5$]$ $\widehat{\beta}_1$	0.499	.015	0.500	.008	1.959	.040	1.989	.015
$[\sigma^2$=.75$]$ $\widehat{\sigma}^2$	0.751	.021	0.750	.010	2.945	.384	2.985	.266
$[\mathcal{R}^2$=.25$]R^2$	0.253	.090	0.251	.065	0.979	.003	0.995	.001
t-statistics	Mean	α=.05	Mean	α=.05	Mean	α=.05	Mean	α=.05
$\tau_{\beta_0}=\frac{\widehat{\beta}_0-\beta_0}{\hat{\sigma}_{\beta_0}}$	0.004	.049	0.015	.050	-1.968	0.774	-3.531	0.968
$\tau_{\beta_1}=\frac{\widehat{\beta}_1-\beta_1}{\hat{\sigma}_{\beta_1}}$	-.013	.047	-.005	.049	35.406	1.000	100.2	1.000

When the estimated LR model is statistically adequate:
(i) the empirical means of the N point estimates are *highly accurate* and the empirical (actual) type I error probabilities associated of the t-tests are very close to the nominal (α=.05) even for a sample size n=50, , and
(ii) their accuracy improves as n increases.

When the estimated model is misspecified:
(i)* the empirical overall means based on the N estimates are *highly inaccurate* (inconsistency) and the actual type I error probabilities are *much larger* ($>$ .77) than the nominal (α=.05), and
(ii)* as n increases the inaccuracy of the estimates increases and the actual type I error probabilities approach 1!

This undermines the claim that **for n large enough** one can **ignore statistical misspecification**. The only criterion for evaluating the *reliability* of frequentist inferences is: **actual error probabilities$\simeq$ nominal ones**.

This is particularly relevant for a wide variety of robustness claims currently invoked by textbooks. **There are no real robustness results** for generic departures from IID; see Spanos (2002). It should also be emphasized that in table 5 only assumption [5] is invalid, but in practice there are often several invalid assumptions.

4 Securing the adequacy of a statistical model

Why is the problem of establishing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ multifaceted and intricate?

One needs to address several related foundational problems, including:

1. The respective roles of the **substantive subject matter information** and the **statistical information** – the probabilistic assumptions imposed on data $\mathbf{x}_0{:=}(x_1,...,x_n)$ – invoked (implicitly or explicitly) by statistical inference procedures.

2. What constitutes a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and how can one specify it in terms of **complete, internally consistent and testable set of probabilistic assumptions**. Attaching white noise error terms on mathematical equations will not do; see McCullagh (2002).

4.1 Modeling: specification, M-S testing and respecification

Ensuring the reliability of statistical inference based on $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, calls for certain distinctions to be made with a view to elucidate the problem.

(1) Distinguish between the **statistical** $[\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ and **substantive** $[\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})]$ **models**, where $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ comprises solely the probabilistic assumptions imposed on the observable process $\{X_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$.

The questions of interest could be as vague as tentative conjectures, or a formal substantive model:

$$\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\varphi}),\ \boldsymbol{\varphi}\in\boldsymbol{\Phi}\},\ \mathbf{x}\in\mathbb{R}^n_X,\ \text{for } \boldsymbol{\varphi}\in\boldsymbol{\Phi}\subset\mathbb{R}^p,\ p\leq m,$$

where $\boldsymbol{\varphi}$ denotes the **substantive parameters of interest**.

(2) Separate the **modeling** (specification, M-S testing and respecification of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$) from the **inference facet** so that one can establish the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ before any inferences relating to $\boldsymbol{\theta}$ or $\boldsymbol{\varphi}$ are drawn.

(a) The **specification** (initial selection) of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ should be based on:

(i) Choosing the **appropriate probabilistic assumptions** relating to $\{X_k,\ k\in\mathbb{N}\}$ to account for all the *statistical systematic information* in data $\mathbf{x}_0$ in the form of **chance regularity patterns** exhibited by $\mathbf{x}_0$ that can be accounted for using probabilistic assumptions from three broad categories: **Distribution, Dependence and Heterogeneity;** see Spanos (1986).

$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ should be specified in terms of **a complete, internally consistent**

and testable set of probabilistic assumptions pertaining to the **observable** stochastic process $\{X_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$. Attaching unobservable white-noise error terms $\{u_t,\ t\in\mathbb{N}\}$ on mathematical equations will NOT do; see McCullagh (2002).
(ii) Selecting a particular statistical parametrization $\boldsymbol{\theta}\in\Theta$ that enables one to pose the substantive questions of interest to data $\mathbf{x}_0$ by ensuring that the statistical ($\boldsymbol{\theta}$) and substantive ($\boldsymbol{\varphi}$) parameters are interrelated via a set of restrictions, say $\mathbf{g}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}$. When the modeling begins with an a priori postulated (substantive) model, the implicit statistical model should be unveiled.

(b) Frame a thorough **Mis-Specification (M-S) testing** procedure to probe adequately the different ways the probabilistic assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ might be invalid for data $\mathbf{x}_0$.

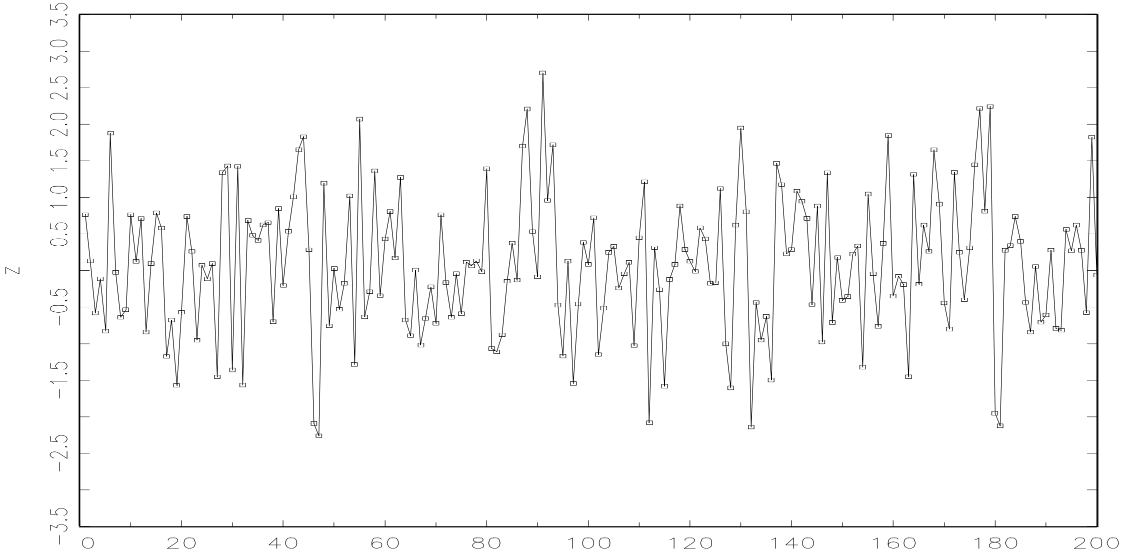
(c) When any of these assumptions are found to be invalid, $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ should be **respecified** with a view to reach a statistically adequate model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})$ that accounts for the systematic information in data $\mathbf{x}_0$.

4.2 Chance regularity patterns in observed data: t-plots

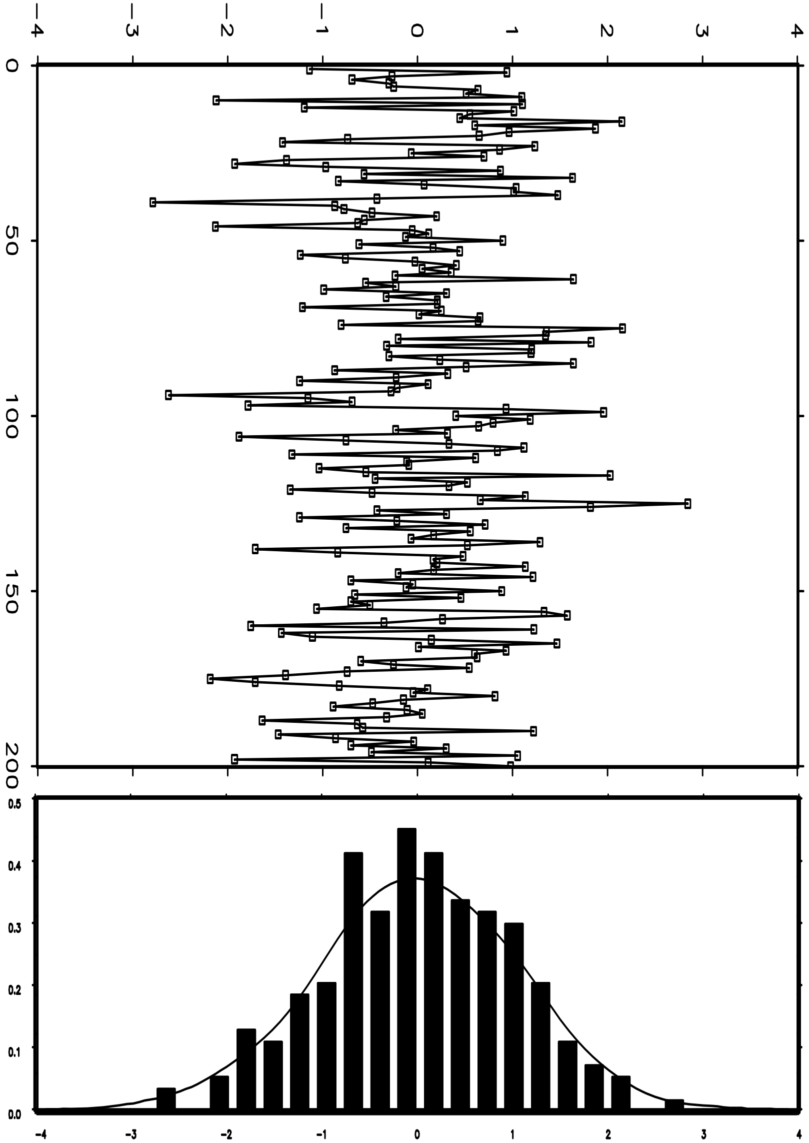
Probabilistic assumptions: (D) Distribution, (M) Dependence, (H) Heterogeneity.

4.2.1 Assessing *distribution* assumption for IID data

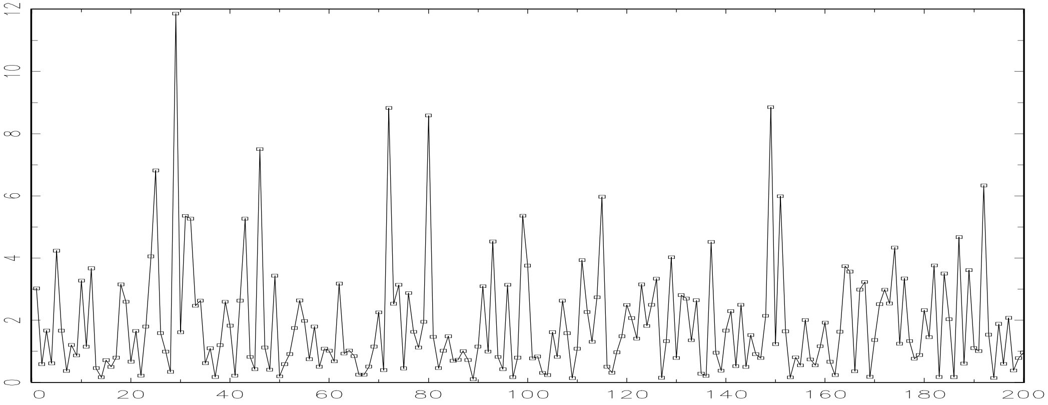
t-plots $\{(t, x_t),\ t{=}1, 2, ..., n\}$ as they relate to the density function via the histogram.



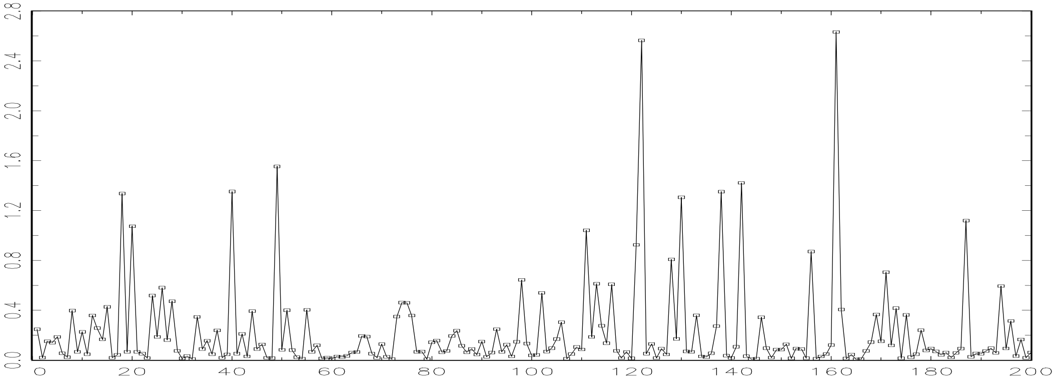
A typical realization of a NIID process



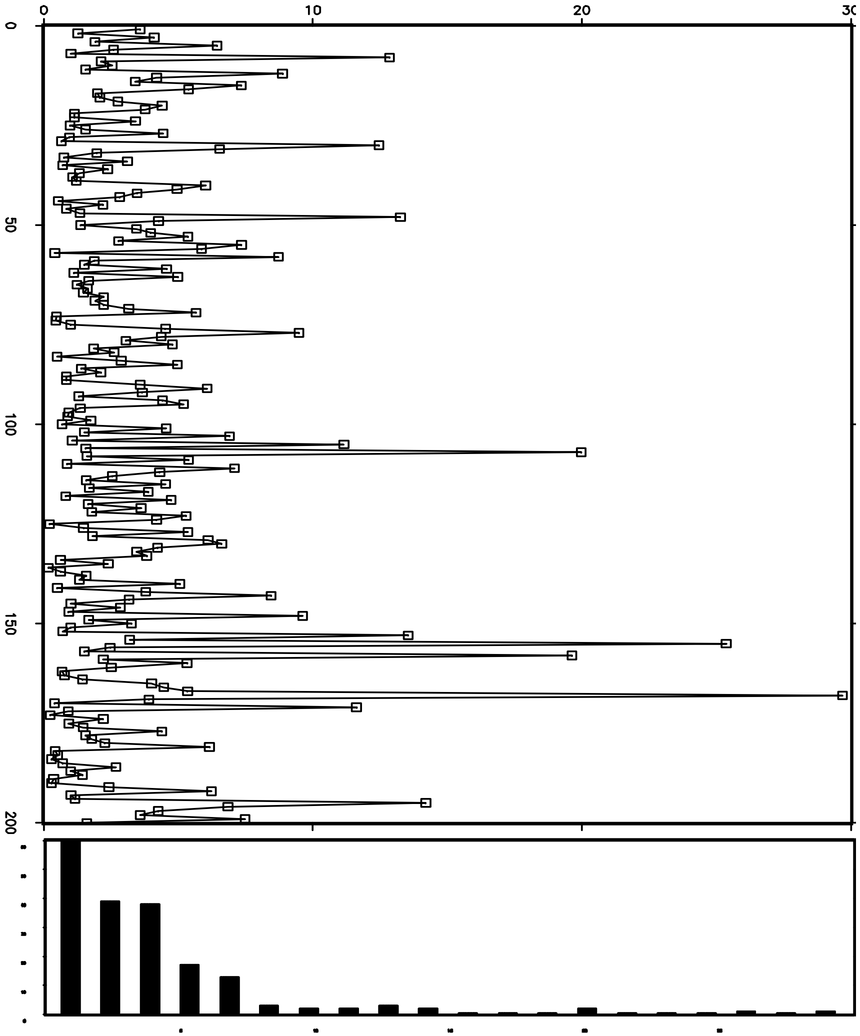
Smoothing the histogram using a kernel



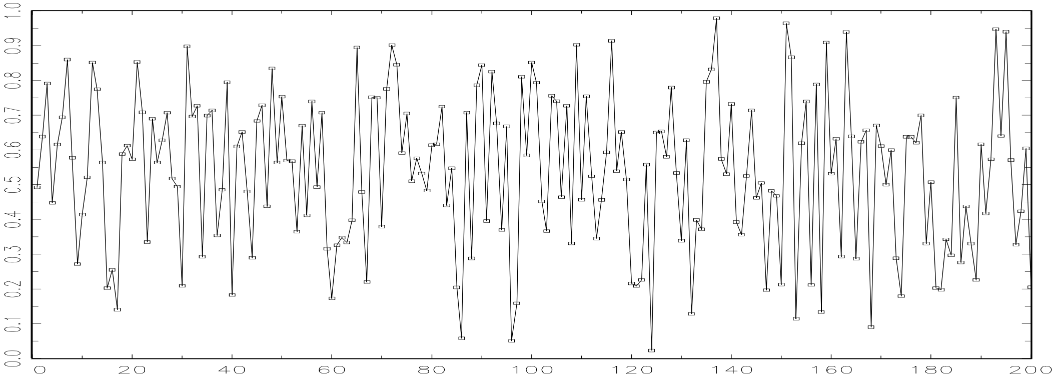
A typical realization of an Expon. IID process



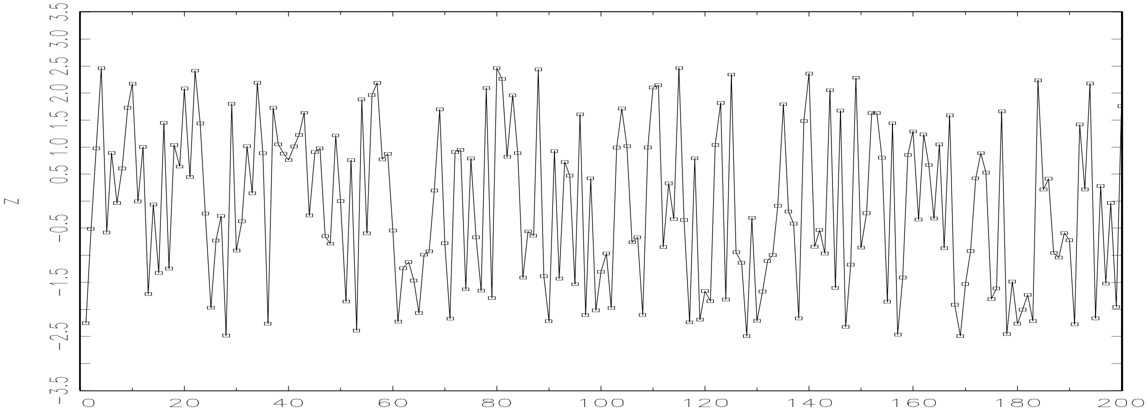
A typical realization of a log-Normal IID process



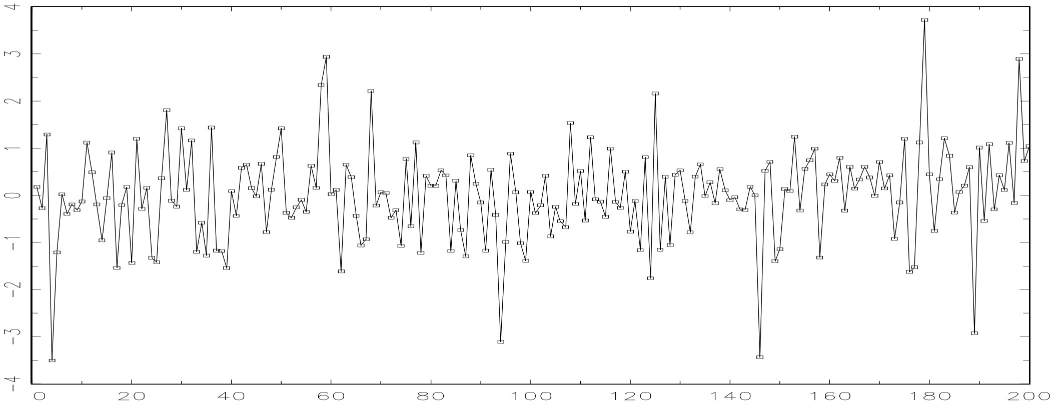
Assessing the Distribution assumption



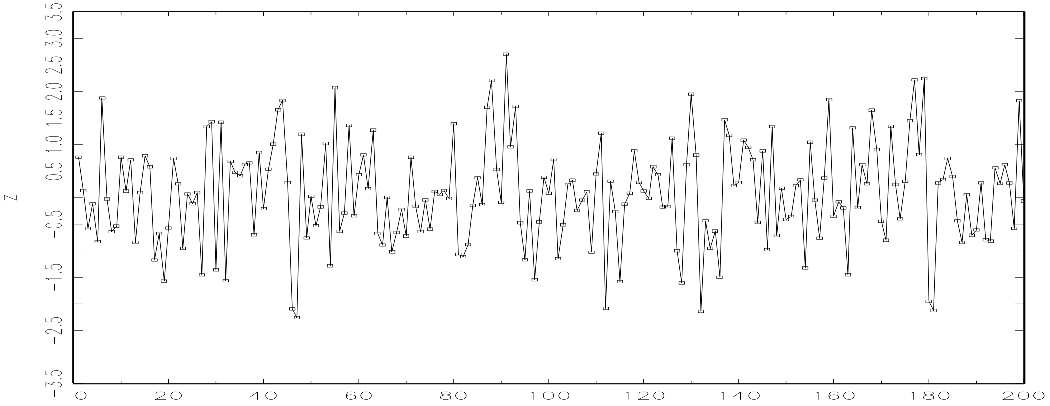
A typical realization of a Beta IID process



A typical realization of a Uniform IID process



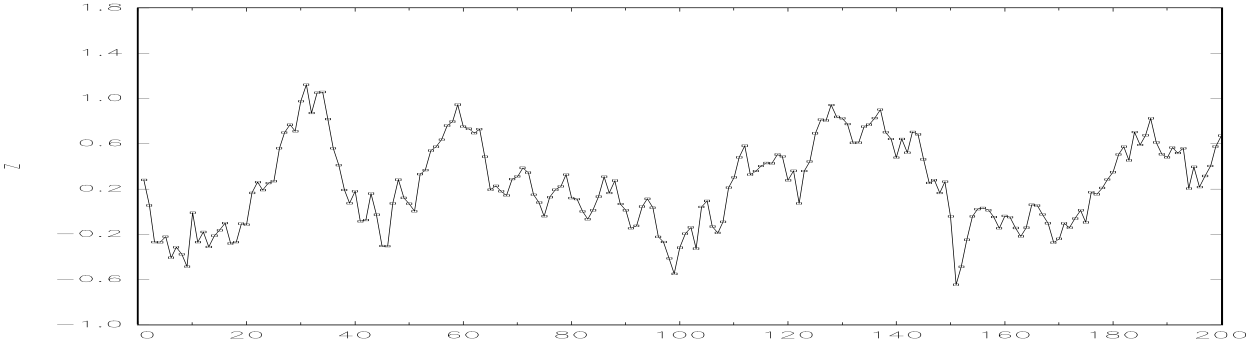
A typical realization of a Student's t IID process



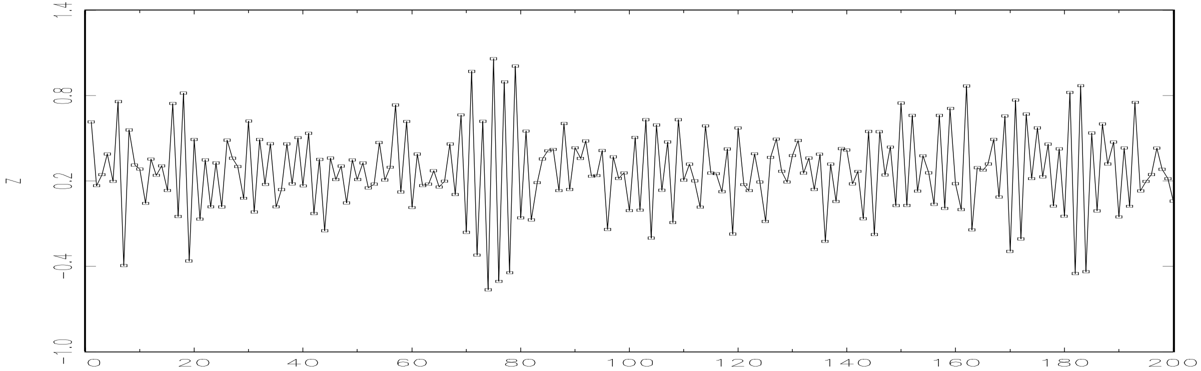
A typical realization of a NIID process

4.2.2 Assessing the *Independence* assumption

Irregular cycles indicate the presence of temporal **positive dependence**.

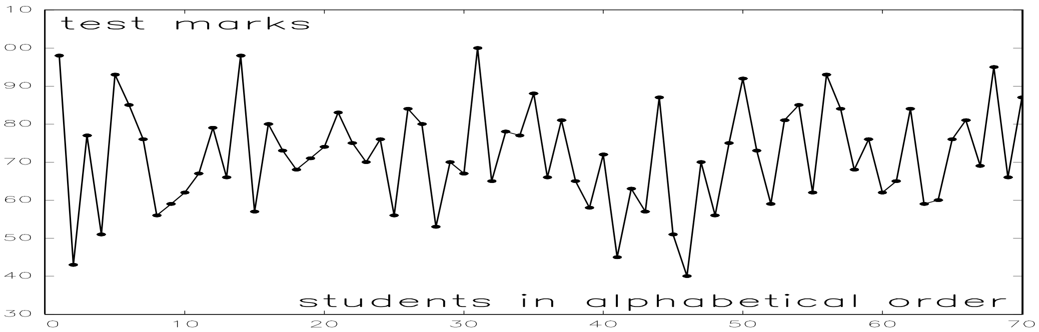


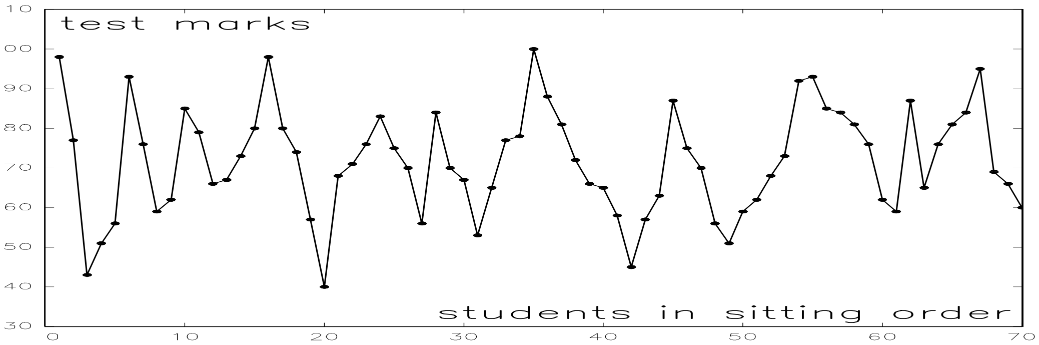
A typical realization of a positively dependent process



A typical realization of a negatively dependent process

Flip-flopping of ups and downs indicate the presence of **negative dependence**.

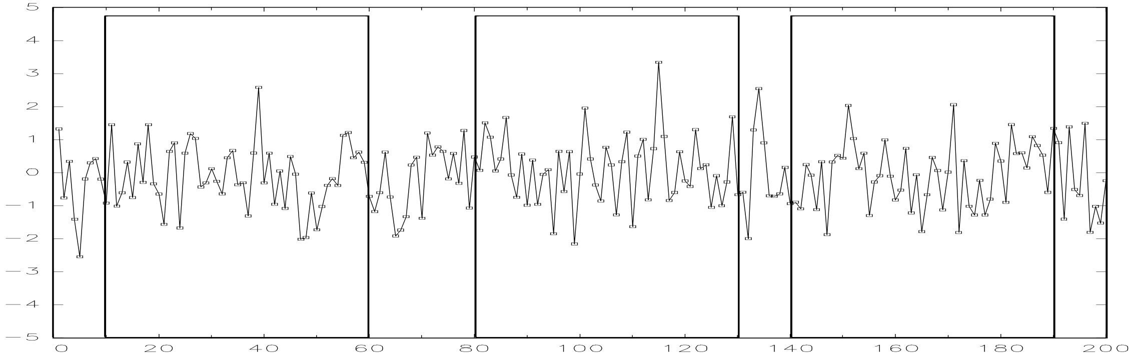




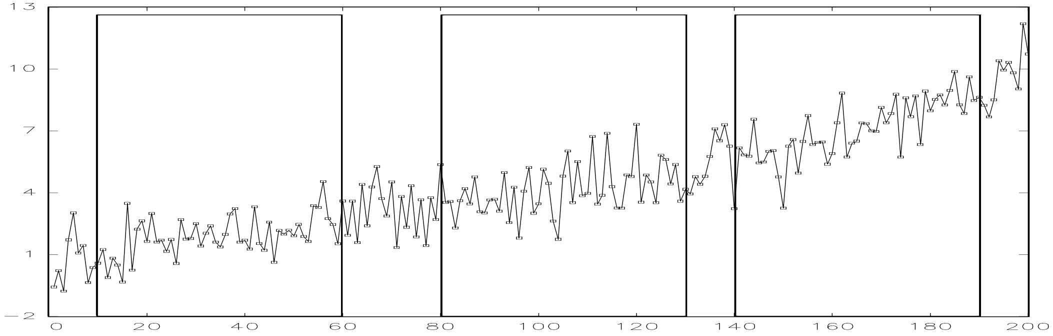
The exam scores data illustrates the presence of ‘spatial’ dependence because when the observations are plotted according to alphabetical order they seem IID, but when plotted according to the sitting arrangement they exhibit irregular cycles, indicating positive spatial dependence; cheating!

4.2.3 Assessing the *Identically Distributed* assumption

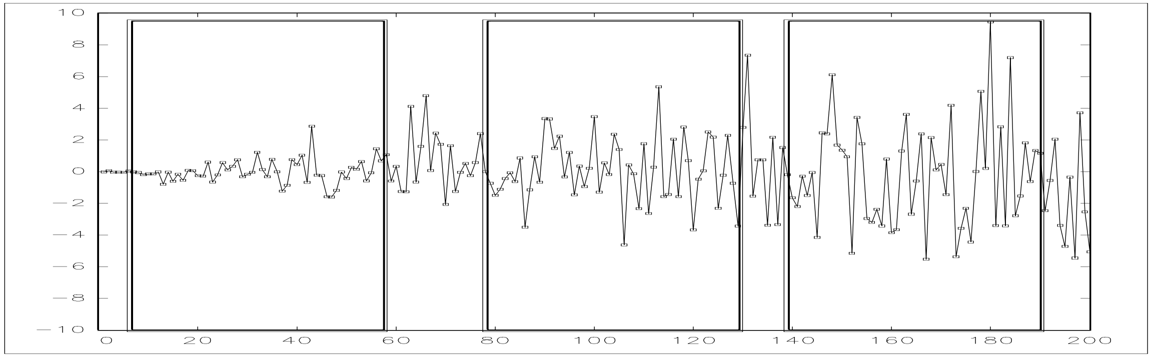
Trends and shifts in mean and variance indicate departures from ID.



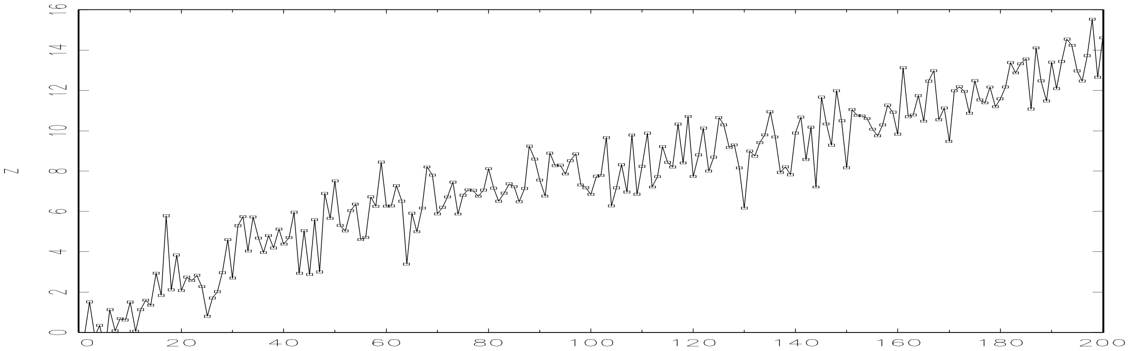
A typical realization of a NIID process



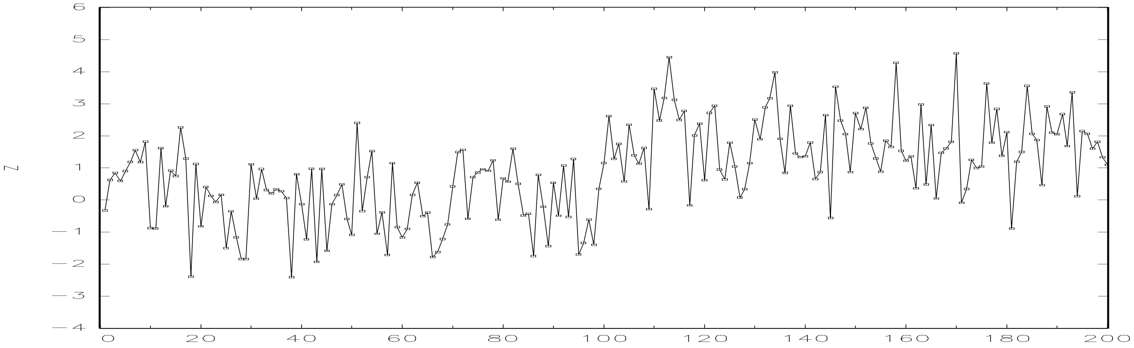
A typical realization of a mean-heterogeneous process



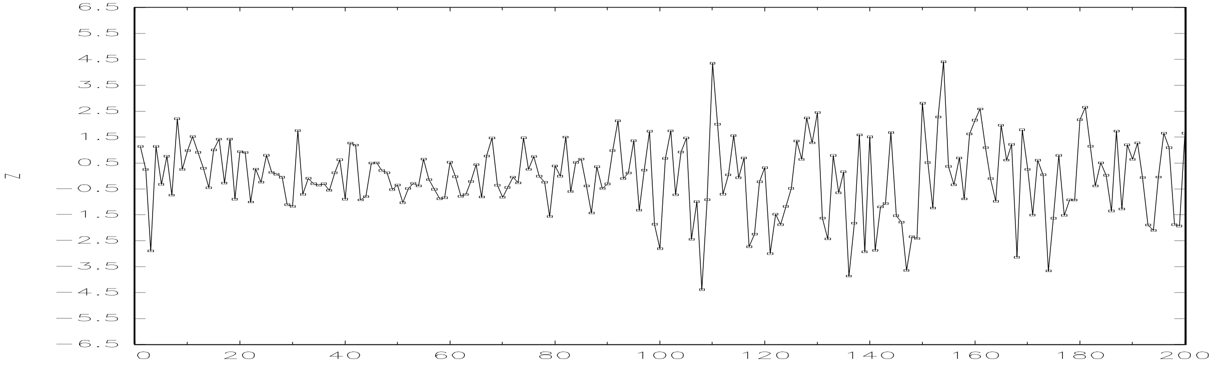
A typical realization of a variance-heterogeneous process



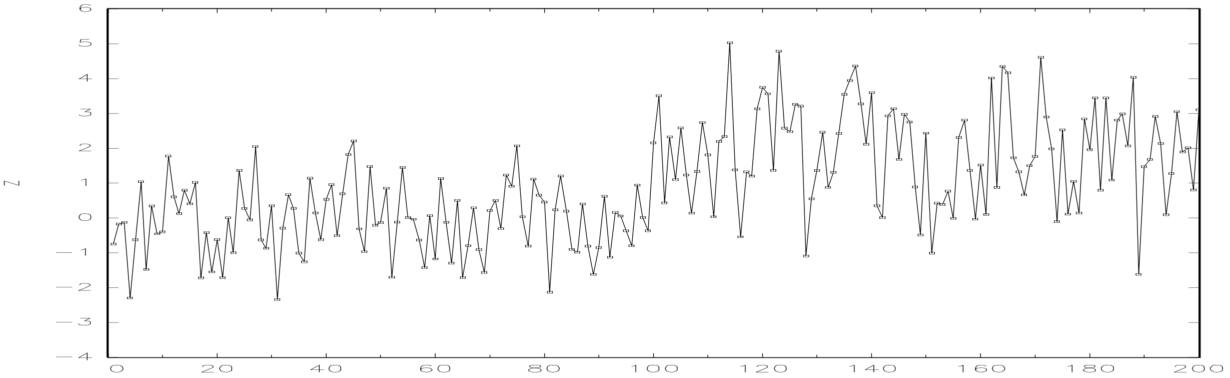
A typical realization of a mean-heterogeneous process



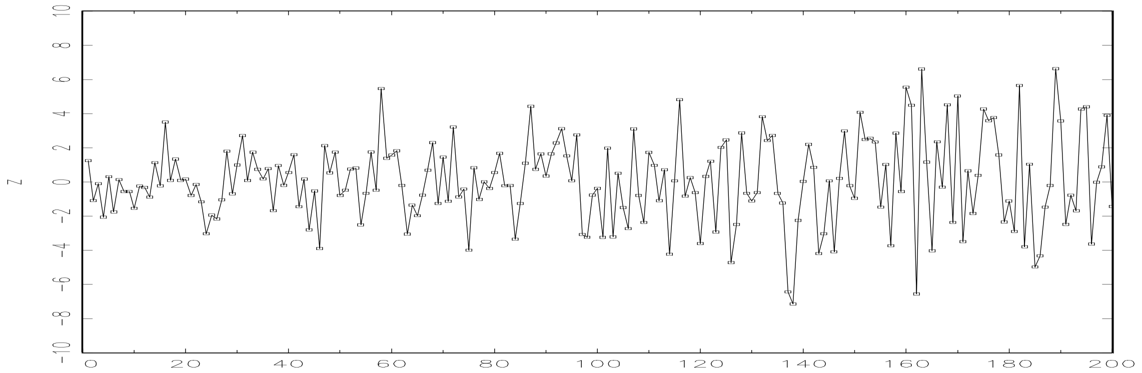
A typical realization of a mean-heterogeneous process



A typical realization of a variance-heterogeneous process

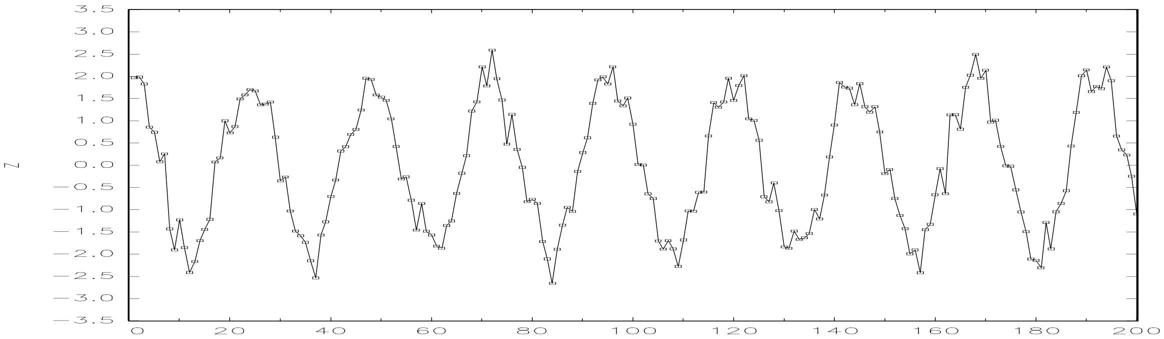


A typical realization of a mean&variance-heterogeneous process

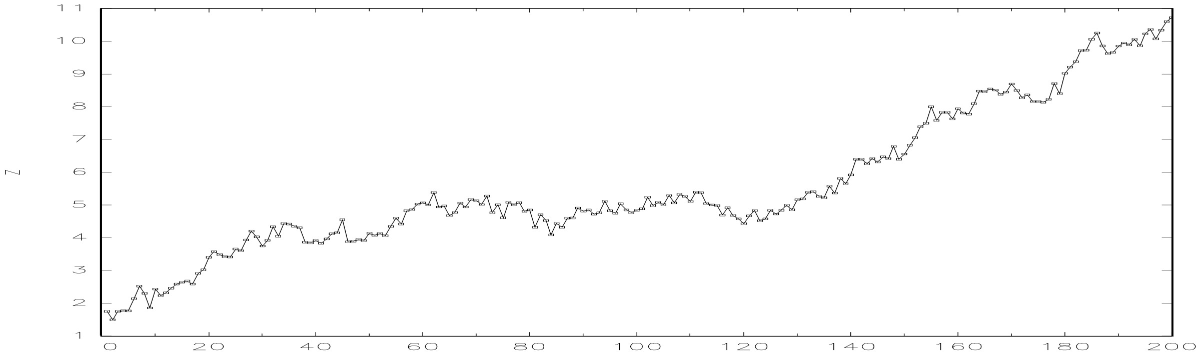


A typical realization of a variance-heterogeneous process

In contrast, the presence of regular cycles indicates the presence of mean heterogeneity.



A typical realization of a mean-heterogeneous (cyclical) process



A typical realization of a positively dependent & heterogenous process

4.3 M-S testing vs. N-P testing

The distinguishing characteristic between hypothesis testing proper and M-S testing is that the former probes within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and the latter outside its boundary; see figures 1-2.

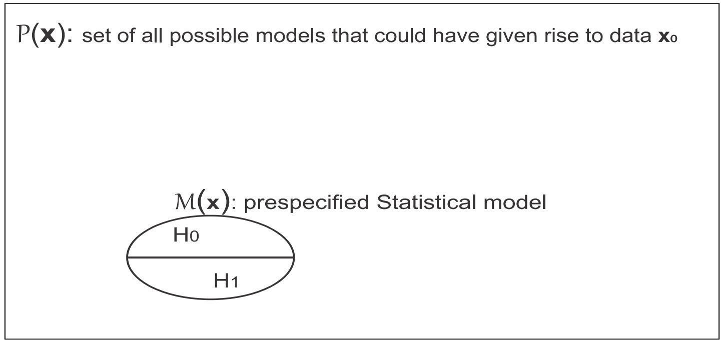

P(x): set of all possible models that could have given rise to data x₀
M(x): prespecified Statistical model
H0
H1

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

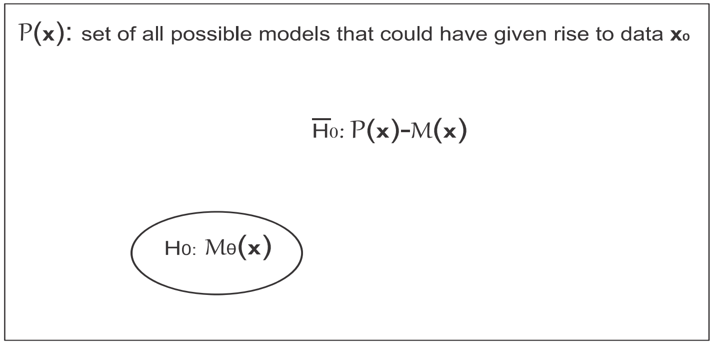

P(x): set of all possible models that could have given rise to data x₀
H̄0: P(x)-M(x)
H0: Mθ(x)

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

Mis-Specification (M-S) vs. N-P testing.

(a) The fact that N-P testing is probing *within* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and M-S testing is probing outside, i.e. $[\mathcal{P}(\mathbf{x}) - \mathcal{M}_\theta(\mathbf{x})]$, renders the latter more vulnerable to the fallacy of rejection. Hence in practice one should *never* accept the particularized H_1 without further probing.

(b) In M-S testing the **type II error** is the **more serious** of the two errors. One will have another chance to correct for **the type I error** at **the respecification stage**. Hence, M-S testing is vulnerable to **the fallacy of acceptance**: misinterpreting 'accept H_0' a evidence *for* H_0.

(c) Since $\overline{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})}$ cannot be explicitly operationalized, and thus M-S testing depends on how one renders probing $[\mathcal{P}(\mathbf{x}) - \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ **operational using parametric and nonparametric tests** for detecting possible departures from $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; see Spanos (2018). Two effective way to particularize $\overline{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})}$ are given below. The first encompasses $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in a broader model. The second probes in **several directions of departure** from specific assumptions using auxiliary regressions, without concerns on whether these directions comprise a formal statistical model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$.

(d) The only legitimate inference one can draw on the basis of M-S testing is whether $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is **statistically adequate** (its probabilistic assumptions are approximately valid) or **misspecified**. It provides no grounds for inferring that the particular encompassing model is valid!! It's highly vulnerable to **the fallacy of rejection:** misinterpreting 'reject H_0' a evidence *for* H_1.

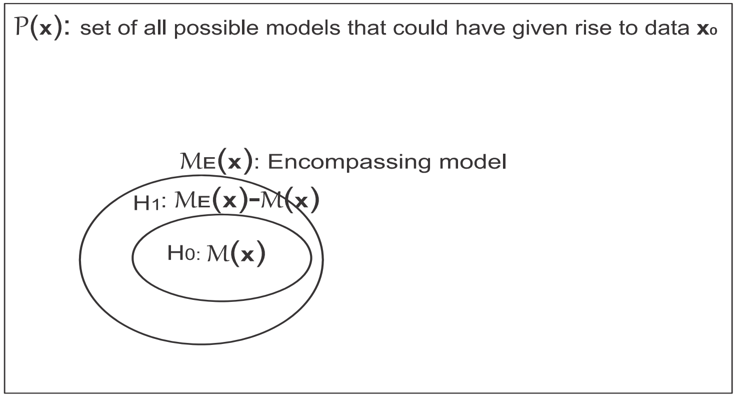

$\mathcal{P}(\mathbf{x})$: set of all possible models that could have given rise to data $\mathbf{x}_0$
$\mathcal{M}_E(\mathbf{x})$: Encompassing model
H_1: $\mathcal{M}_E(\mathbf{x})$-$\mathcal{M}(\mathbf{x})$
H_0: $\mathcal{M}(\mathbf{x})$

Fig. 6: M-S testing using an encompassing model

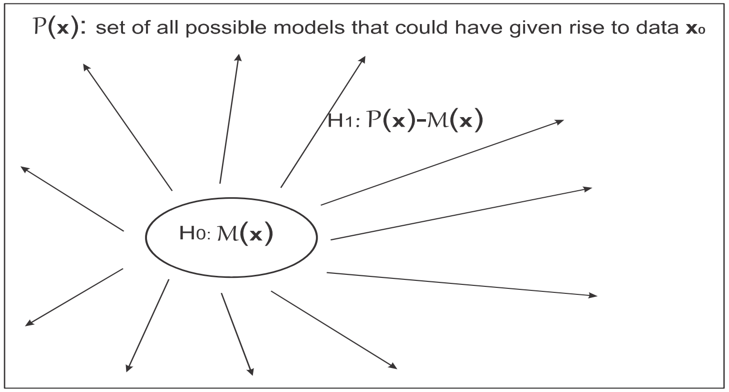

$\mathcal{P}(\mathbf{x})$: set of all possible models that could have given rise to data $\mathbf{x}_0$
H_1: $\mathcal{P}(\mathbf{x})$-$\mathcal{M}(\mathbf{x})$
H_0: $\mathcal{M}(\mathbf{x})$

Fig. 7: M-S testing using directions of departures

5 Establishing statistical adequacy in practice

An empirical study is said to be **replicable** if its statistical results can be:
(i) independently confirmed with very similar or consistent results by other researchers,
(ii) using akin data and (iii) studying the same phenomenon of interest.

It is currently argued that in several scientific fields most empirical results are **non-replicable,** and thus **untrustworthy**!

Unfortunately, replicability does not secure the trustworthiness of the evidence.

A necessary condition for that is the **statistical adequacy** of the invoked $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ since **untrustworthy evidence** can be easily replicated when practitioners use akin data and the same uninformed implementation of their statistical analysis, which would not ensure trustworthiness; see Spanos (1995) for an empirical example.

The following example illustrates the **dangers of ignoring statistical adequacy**.

Example #. Consider two studies based on akin data sets $\mathbf{x}_1$, $\mathbf{x}_2$, with n=100, and the same $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the simple Normal in (**??**), giving rise to the following results:

$$X_{1t}= \underset{(.103)}{3.208}+\widehat{u}_{1t},\ s_1{=}1.029, \quad X_{2t}= \underset{(.118)}{3.277}+\widehat{u}_{2t},\ s_2{=}1.185, \qquad (5.0.3)$$

with the standard errors in brackets, and $\widehat{u}_{it}$, i=1, 2, t=1, ..., n, denoting the residuals.

Table 6 compares their results in (5.0.3) together with the resulting CIs and N-P testing whose p-values are given in square brackets. Testing the difference between the two means, H_0: $\mu_1-\mu_2=0$ vs. H_1: $\mu_1-\mu_2\neq0$, yields $\tau(\mathbf{x}_1,\mathbf{x}_2)=.43[.664]$, rendering the inference results of the two studies almost identical, statistically speaking.

Table 6: Statistical Inference Results

	Point estimates	Observed CIs	H_0: $\mu_i=3.2$, $i=1,2$
$\mathbf{x}_1$:	$(\overline{x}_1=3.208,\ s_1=1.029)$,	$(3.006,\ 3.410)$,	$\tau(\mathbf{x}_1)=.078[.938]$
$\mathbf{x}_2$:	$(\overline{x}_2=3.277,\ s_2=1.185)$,	$(3.045, 3.509)$,	$\tau(\mathbf{x}_2)=.653[.515]$

Despite the nearly identical inference results, it's not obvious how an arbiter can decide with any confidence whether: (i) the study with data $\mathbf{x}_2$ constitutes a successful replication with trustworthy evidence of the study with data $\mathbf{x}_1$, and (ii) the two data sets $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated by the same statistical mechanism. In contrast to the simulation in example 6 (Appendix), the preconditions (a)-(c) do not hold with actual data such as $\mathbf{x}_1$ and $\mathbf{x}_2$. Hence, the only way to draw reliable conclusions is to test the statistical adequacy of the estimated models in (5.0.3). As argued above, any departures from the NIID assumptions are likely to induce sizeable *discrepancies* between the nominal and the *actual* error probabilities rendering their inference results untrustworthy.

The statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ can be established using informal t-plots of the data in conjunction with formal Mis-Specification (M-S) tests. The t-plot for data $\mathbf{x}_1$ (figure 1) exhibits no obvious departures from the NIID assumptions, but the t-plot of $\mathbf{x}_2$ (figure 2) exhibits a likely departure from 'Independence' in the form of 'irregular cycles'; see Spanos (2019), ch. 5.

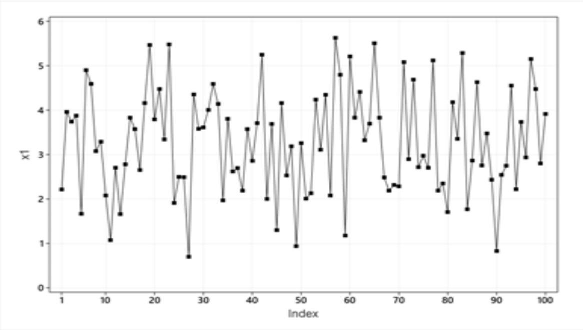


Fig. 1: t-plot of data $\mathbf{x}_1$

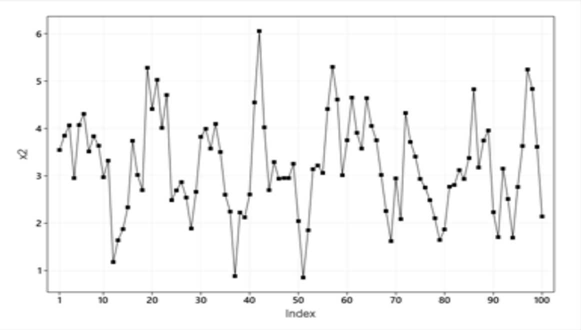


Fig. 2: t-plot of data $\mathbf{x}_2$

The IID assumptions can be formally tested using the runs test:

$$d_R(\mathbf{X})=\frac{[R-E(R)]}{\sqrt{Var(R)}} \overset{\mathsf{IID}}{\approx} \mathsf{N}(0,1),\ d_R(\mathbf{X}) \overset{\text{non-IID}}{\backsim} \mathsf{N}(\delta_n, \tau_n^2),\ \ \delta_n{\neq}0,\ \ \tau_n^2{>}0, \qquad (5.0.4)$$

where R-the number of runs, $E(R)=\frac{2n-1}{3}$, $Var(R)=\frac{16n-29}{90}$, δ_n and τ_n^2 depend only on n. For testing Normality, one can apply the Anderson-Darling (A-D) test; see Spanos (2019). The M-S testing results in Table 7 confirm that the simple Normal model in (**??**) is statistically adequate for data $\mathbf{x}_1$, but misspecified for data $\mathbf{x}_2$ since the 'Independence' assumption is invalid.

The M-S testing results in Table 7 indicate that the inference results in Table 6 based on $\mathbf{x}_2$ will be untrustworthy due to sizeable discrepancies between the actual and nominal error probabilities; see Spanos (2009). Seeking to invoke generic robustness in such cases is usually forlorn since no robustness results pertaining to generic departures from the IID assumptions are available; see Spanos (2002). Hence, the study based on data $\mathbf{x}_2$ does not represent a successful replication (with trustworthy evidence) of the study based on data $\mathbf{x}_1$.

Table 7: Simple M-S tests for IID

Data	Runs test for IID	Normality test
$\mathbf{x}_1$:	$d_R(\mathbf{x}_1)=\frac{(50-51)}{4.178}=-.239[.918]$,	$A\text{-}D(\mathbf{x}_1)=.414[.330]$,
$\mathbf{x}_2$:	$d_R(\mathbf{x}_2)=\frac{(32-51)}{4.178}=-4.548[.000]$,	$A\text{-}D(\mathbf{x}_2)=.177[.918]$,

These tests are adequate to establish that a simple model, based on NIID is statistically misspecified, but for more elaborate models one needs more powerful tools for M-S testing such as auxiliary regressions probing in all the different directions the probabilistic assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ can be false!

Respecifying the original simple Normal model in Table 1 takes the form of selecting a different statistical model $\mathcal{M}_{\boldsymbol{\vartheta}}(\mathbf{x}_2)$ which could account for the irregular cycles in figure 2. An obvious choice will be to replace the Independence assumption with some form of dependence, say Markov. This give rise to an Autoregressive [AR(1)] model (Spanos, 2019):

$$(X_t|X_{t-1})\backsim\mathsf{N}(\alpha_0{+}\alpha_1 X_{t-1},\ \sigma_0^2),\ (\alpha_0{,}\alpha_1,\sigma_0^2){\in}\mathbb{R}\times(-1,1){\times}\mathbb{R}_+,\ x_t{\in}\mathbb{R},\ t{\in}\mathbb{N}. \tag{5.0.5}$$

Example #. Consider the t-plots of simulated data $\mathbf{Z}_0{:=}\{(x_t,y_t),\ t{=}1,2,...,n\}$, in figures 1 and 2. Both t-plots indicate that the data exhibit mean heterogeneity (trending mean) as well as irregular cycles that reflect the presence of positive dependence, suggesting that the IID are likely to be invalid. To get a better diagnosis of the dependence, figures 3 and 4 depict the t-plot of the detrended data.

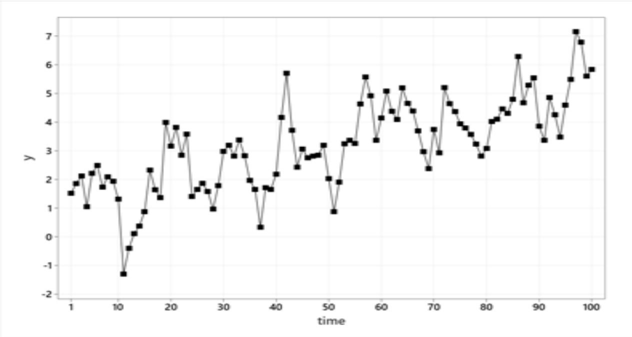


Fig. 1: t-plot of y_t, $t{=}1, 2, .., n$

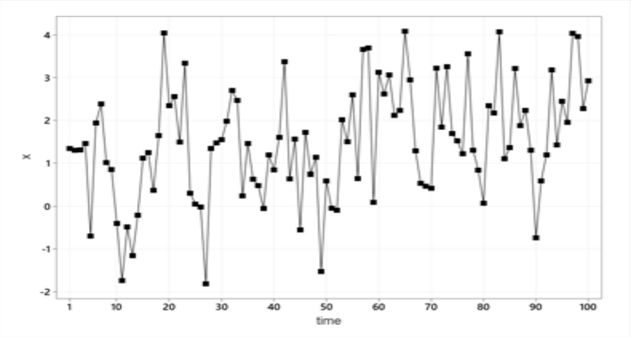


Fig. 2: t-plot of x_t, $t{=}1, 2, .., n$

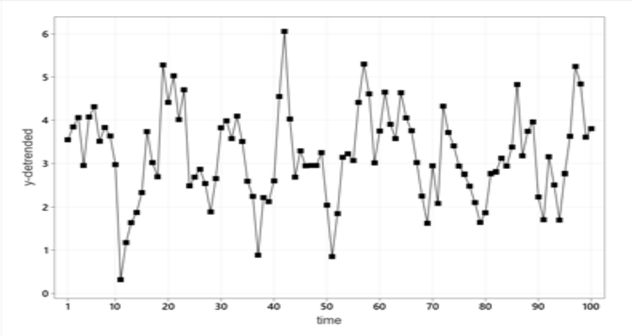


Fig. 3: t-plot of of detrended y_t

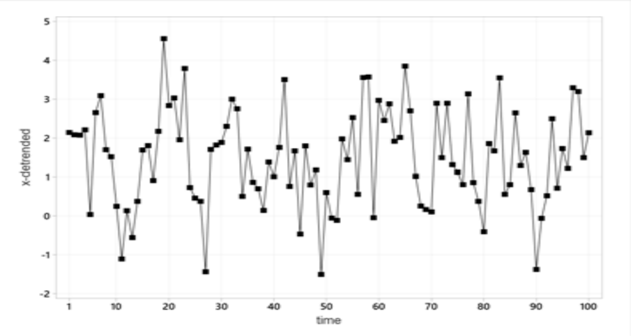


Fig. 4: t-plot of detrended x_t

The distinct irregular cycles in figures 3 and 4 indicate the presence of positive dependence, and the bell-shape symmetry indicates no serious departures from

Normality; see Spanos (2019), ch. 5. Taken together, these chance regularity patterns suggest that an appropriate statistical model might be the Dynamic Linear Regression (DLR) with mean heterogeneity (Table 3).

M-S testing. To illustrate M-S testing and respecification, let us return to the data in figures 1 and 2 and consider the scenario where a practitioner begins by estimating the LR model (Table 1) to yield:

$$Y_t{=}\underset{(.179)}{2.04} + \underset{(.09)}{.783}x_t + \widehat{u}_t,\ \ s{=}1.204,\ \ R^2{=}.436,$$

where the standard errors (SEs) of the estimated coefficients are shown in brackets. If the practitioner takes these estimates at face value, will conclude that there is a linear relationship between Y_t and X_t, since both coefficients are significantly different from zero with tiny p-values. These inferences are called into question since the following M-S testing results based on the auxiliary regressions (**??**)-(**??**), indicate that assumptions [4] and [5] are invalid:

$$\widehat{v}_t = \overbrace{\underset{(4.23)}{-.355}\ \underset{(.091)}{-.388}\, x_t}^{\text{original specification}} + \overbrace{\underset{(.026)}{.026}x_t^2}^{\neg[2]} - \overbrace{\underset{(.164)}{.918}t_s}^{\neg[5]\blacklozenge} + \overbrace{\underset{(.064)}{.161}x_{t-1} + \underset{(.080)}{.187}Y_{t-1}}^{\neg[4]\blacklozenge},$$

$$\widehat{v}_t^2 = \underset{(.212)}{\overbrace{.841}^{\text{original spec.}}} - \underset{(.274)}{\overbrace{.187t_s}^{\neg[5]}} - \underset{(.028)}{\overbrace{.035x_t^2}^{\neg[3]}} - \overbrace{\underset{(.033)}{.048x_{t-1}^2} + \underset{(.020)}{.037Y_{t-1}^2}}^{\neg[4]}. \tag{5.0.6}$$

Respecification involves seeking a different statistical model $\mathcal{M}_{\boldsymbol{\vartheta}}(\mathbf{z})$ that could potentially account for all the statistical systematic information in $\mathbf{Z}_0$.

As argued above, both the t-plots and the above M-S testing results suggest that the respecified model $\mathcal{M}_{\boldsymbol{\vartheta}}(\mathbf{x})$ should be the DLR model in Table 3.

Table 3: 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:	$(Y_t \mid X_t, \mathbf{Z}_{t-1}) \backsim \mathsf{N}(.,.)$, where $\mathbf{Z}_t := (Y_t, X_t)$	$t\in\mathbb{N}$.
[2]	Linearity:	$E\left(Y_t \mid X_t, \mathbf{Z}_{t-1}\right) = \delta_0 + \delta_1 t + \alpha_1 x_t + \alpha_2 Y_{t-1} + \alpha_3 x_{t-1}$,	
[3]	Homoskedasticity:	$Var\left(Y_t \mid X_t, \mathbf{Z}_{t-1}\right) = \sigma_0^2$,	
[4]	Markov dependence:	$\{(\mathbf{Z}_t \mid \mathbf{Z}_{t-1}),\ t\in\mathbb{N}\}$ is a Markov process,	
[5]	t-invariance:	$\left(\delta_0, \delta, \alpha_1, \alpha_2, \alpha_3, \sigma_0^2\right)$ are *not* changing with t.	

Estimating the DLR model, $\mathcal{M}_{\boldsymbol{\vartheta}}(\mathbf{x})$, yields the following results:

$$\mathcal{M}_{\widehat{\boldsymbol{\vartheta}}}(\mathbf{z}):\; Y_t = \underset{(.240)}{1.584} + \underset{(.197)}{1.088} t_s + \underset{(.082)}{.404} x_t + \underset{(.075)}{.181} x_{t-1} + \underset{(.096)}{.238} Y_{t-1} + \widehat{\varepsilon}_t, \tag{5.0.7}$$

where $s{=}.729,\;\; R^2{=}.80,\;\; n{=}99$. The analogous M-S testing auxiliary regressions to :

$$\widehat{u}_t = \overbrace{\underset{(.305)}{.157} + \underset{(.346)}{.242} t_s + \underset{(.113)}{.033} x_t + \underset{(.087)}{.029} x_{t-1} + \underset{(.129)}{.036} Y_{t-1}}^{\text{original specification}} - \overbrace{\underset{(.037)}{.018} \widehat{Y}_t^2}^{\neg[2]} + \overbrace{\underset{(.303)}{.120} t_s^2}^{\neg[5]} + \overbrace{\underset{(.081)}{.067} x_{t-2} - \underset{(.10)}{.09} Y_{t-2}}^{\neg[4]}$$

$$\widehat{u}_t^2 = \overbrace{\underset{(.212)}{.168}}^{\text{original spec.}} - \overbrace{\underset{(.212)}{.024} t_s}^{\neg[5]} + \overbrace{\underset{(.021)}{.027} \widehat{Y}_t^2}^{\neg[3]} - \overbrace{\underset{(.021)}{.018} x_{t-1}^2 - \underset{(.016)}{.021} Y_{t-1}^2}^{\neg[3]-[4]},\;\; A\text{-}D(\mathbf{v}){=}.383[.391],$$

where $\widehat{Y}_t$ denotes the fitted values, A-D denotes the Anderson-Darlin Normality test, indicate no departures from its assumptions [1]-[5] (Table 3It).is important to emphasize that the appropriate form of heterogeneity (1st degree trend polynomial) and dependence (one lag) for $\{\mathbf{Z}_t,\; t{\in}\mathbb{N}\}$ in (5.0.7) can only be justified in terms of the statistical adequacy of $\mathcal{M}_{\boldsymbol{\vartheta}}(\mathbf{z})$, i.e. its own assumptions [1]-[5] are valid for $\mathbf{Z}_0$, and not on the basis of the auxiliary regressions of the misspecified model in (5.0.6).

5.1 Reluctance to test the validity of model assumptions

The crucial importance of securing statistical adequacy stems from the fact that **no trustworthy evidence** for or against a substantive claim (or theory) can be secured on the basis of a statistically misspecified model.

To paraphrase Box (1979): "All models are crude approximations of reality, but statistically misspecified models are useless for inference!"

▶ **Why is there such reluctance** to secure **statistical adequacy**?

The empirical modeling literature appears to **underestimate** the potentially devastating effects of statistical misspecification on the reliability of inference. This misplaced underestimation stems from a number of **questionable arguments**.

(i) **Curve-fitting vs. modeling.** Empirical modeling is often misleadingly viewed as curve-fitting a **substantive model** $\mathcal{M}_{\varphi}(\mathbf{x})$ guided by **goodness-of-fit measures**. This stems viewing $\mathcal{M}_{\varphi}(\mathbf{x})$ as established knowledge, instead of tentative conjunctures to be tested against data.

(ii) **Conflating statistical with substantive adequacy**.

[a] **statistical adequacy**: does $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ account for the chance regularities in

$\mathbf{x}_0$? $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ relates to the statistical information contained in data $\mathbf{x}_0$.

[b] substantive adequacy: does the model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ adequately captures (describes, explains, predicts) the phenomenon of interest? Substantive inadequacy arises from flawed *ceteris paribus* clauses, missing confounding factors, systematic approximation error, etc. This confusion also permeates the slogan "**all models are wrong, but some are useful**" (Box, 1979)

(iii) **Misplaced asymptotics**. The current undue reliance on **asymptotic procedures** based on probabilistic **assumptions that are non-testable**! As argued by Le Cam (1986a, p. xiv): "... limit theorems "as n tends to infinity" are logically devoid of content about what happens at any particular n."

(iv) **Robustness**. There is also the undue reliance on **vague 'robustness' results** whose generality and applicability is often greatly overvalued. On closer examination the adjustments used to secure robustness do nothing to alleviate the problem of sizeable discrepancies between actual and nominal error probabilities.

(v) **Descriptive vs Inferential statistics**. Amrhein et al. (2019), in a paper provocatively entitled "Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication", contends that one can evade the need to secure the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ upon which the frequentist inferential results are grounded by re-interpreting them as mere '**descriptive statistics**'. The truth is that the **validity of the same probabilistic assumptions** that underwrite the reliability of the sampling distribution of any statistic Y_n, $f(y_n;\boldsymbol{\theta})$, $y_n{\in}\mathbb{R}$, also ensure the **'appropriateness' of any descriptive statistic** y_n to provide a **pertinent statistical summary** of data $\mathbf{x}_0$.

(vi) **The hybrid-model alibi**. Empirical modeling in most disciplines is dominated by **hybrid-models**: a blend of probabilistic assumptions and 'background assumptions'.

Greenland & Rafi (2021): "All scientific interpretations of statistical outputs depend on background (auxiliary) assumptions that are rarely delineated or checked in full detail. These include not only the usual modeling assumptions, but also deeper assumptions about the data-generating mechanism that are implicit in conventional statistical interpretations yet are unrealistic in most health, medical etc."

Such 'hybrid-models' evoke **Duhem's (1914) conundrum** where 'no primary hypothesis can be tested separately from the set of "auxiliary assumptions" invoked in empirical testing'. As a result, their validation can be ignored by invoking the **undelineated assumptions** as an alibi.

6 Summary and conclusions

Statistical misspecification of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ forfeits **control** of the relevant **error probabilities** since:

$$\boxed{\textbf{actual} \neq \textbf{nominal error probabilities}}$$

which **undermines** the reliability of inference and the trustworthiness of evidence. To task of securing the **statistical adequacy** of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is multifaceted and raises several related problems that need to be addressed.
The first is the formal framing of a **statistical model** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in terms of a **complete, internally consistent and testable set of probabilistic assumptions** relating the observable process $\{X_k,\ k\in\mathbb{N}\}$, underlying data $\mathbf{x}_0$.
The second relates to **two crucial distinctions** to be deployed in the context of Fisher's model-based statistical induction:

(A) the **statistical vs. the substantive** information/assumptions/models, and

(B) the **modeling vs. the inference facet** in learning from data.

■ The **statistical vs. the substantive** distinction is used to address the **Duhem conundrum** by ensuring that the 'auxiliary assumptions' are only those comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ whose validity can be assessed using M-S testing based on ancillary statistics usually the residuals.

■ Separating the **modeling** from the **inference facet** plays a crucial role in establishing the **statistical adequacy** of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ at the **modeling** facet, and securing the **actual** optimality and reliability of inference with a view to ensure the trustworthiness of evidence.

The modeling and inference facets can be formally separated when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ belongs to the **exponential family** of distributions (Fisher, 1934), since:

$$f(\mathbf{x};\boldsymbol{\theta})=|J|\cdot f(\mathbf{s},\mathbf{r};\boldsymbol{\theta})=|J|\cdot f(\mathbf{s};\boldsymbol{\theta})\cdot f(\mathbf{r}),\ \forall\,(\mathbf{s},\mathbf{r})\in\mathbb{R}_s^m\times\mathbb{R}_r^{n-m}, \qquad (6.0.8)$$

where $\mathbf{S}(\mathbf{X}){:=}(S_1,...,S_m)$ is a **complete sufficient** statistic and $\mathbf{R}(\mathbf{X}){:=}(R_1,...,R_{n-m})$, a **maximal ancillary** statistic, with $\mathbf{S}(\mathbf{X})$ and $\mathbf{R}(\mathbf{X})$ independent. The separation in (6.0.8) means that all primary inferences can be based exclusively on $f(\mathbf{s};\boldsymbol{\theta})$, and $f(\mathbf{r})$ (free of $\boldsymbol{\theta}$) can be used to validate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Often $\mathbf{R}(\mathbf{X})$ is function of the residuals!

7 Appendix: Misplaced criticisms of M-S testing

Criticisms of M-S testing. M-S testing has been criticized on several grounds, including: (a) data-mining/snooping, (b) double-use of data, (c) infinite regress/circularity, (d) pre-test bias, (e) multiple testing issues, and (f) erroneous diagnoses; see Appendix for a summary.

7.1 The infinite regress and circularity charges

The *infinite regress* charge is articulated by claiming that each M-S test relies on a set of assumptions, and thus it assesses the assumptions of the model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ by invoking the validity of its own assumptions, trading one set of assumptions with another *ad infinitum*. This reasoning is often *circular* because some M-S tests unwittingly assume the validity of the very assumption under test!

► A closer look at M-S testing reasoning reveals that both charges are misplaced. An M-S test is just a combination of a 'distance' function and a rejection region whose relevant error probabilities are evaluated under *hypothetical scenarios* that involve *only* the probabilistic assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$.

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

H: *all* the probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ are valid.

Example. The *runs test* is an example of an omnibus M-S test for assumptions [4]-[5] (table 2), whose distribution under the null, for $n \geq 40$, is based on:

$$Z_R(\mathbf{Z}) = [R - E(R)] \, / \sqrt{Var(R)} \overset{\text{[1]-[5]}}{\backsim} \mathsf{N}(0,1).$$

NOTE that the runs test is *insensitive* to departures from Normality.

■ **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, or choose tests which are insensitive to departures from the retained assumptions:

$\overline{H}$: particular departures from the assumption(s) being tested, but the rest of the assumption(s) of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ hold for data $\mathbf{z}_0$.

For the runs test, the evaluation under the alternative takes the form:

$$Z_R(\mathbf{Z}) \overset{[1]-[3]\&\overline{\text{[4]}}\text{-}\overline{\text{[5]}}}{\backsim} \mathsf{N}(\delta, \tau^2),\ \ \delta \neq 0,\ \ \tau^2 > 0,$$

where $\overline{[4]}$ and $\overline{[5]}$ denote specific departures from the assumptions tested.
Alternative scenarios in the M-S testing will affect the power of the test in a variety of ways, and one needs to apply a battery of different M-S tests to ensure broad probing capacity and self-correcting in the sense that the effect of any departures from the maintained assumptions is also detected.
▶ When a departure is detected by an M-S test with *low power*, it provides better evidence for its presence than a more powerful test.

7.2 Illegitimate double-use of data

In the context of the error statistical approach it is certainly true that the same data $\mathbf{z}_0$ 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 prespecified model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$,
but 'does that constitute an illegitimate double-use of data?'
The short answer is *no* for two interrelated reasons.
First, (a) and (b) pose very different questions to data $\mathbf{z}_0$, and
second, the probing takes place within vs. outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, respectively.

Neyman-Pearson (N-P) testing assumes that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is adequate, and poses questions within its boundaries.

In contrast, the question posed by M-S testing is whether or not the particular data $\mathbf{x}_0$ constitute a '*truly typical realization*' of the stochastic mechanism described by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, and the probing takes place outside its boundaries, i.e. in $[\mathcal{P}(\mathbf{z}) - \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$; see Spanos (2000).

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

Prompted by a remark by Hendry (1995), crediting Mayo (1980) with showing that M-S testing does not involve illegitimate double use of data, Spanos (2007) demonstrated the following result.

For many statistical models, the distribution of the sample $f(\mathbf{z};\boldsymbol{\theta})$ simplifies to:

$$f(\mathbf{z};\boldsymbol{\theta}) = |J| \cdot f(\mathbf{s},\mathbf{r};\boldsymbol{\theta}) = |J| \cdot f(\mathbf{s};\boldsymbol{\theta}) \cdot f(\mathbf{r}), \tag{7.2.9}$$

for all $(\mathbf{s},\mathbf{r}) \in \mathbb{R}_s^m \times \mathbb{R}_r^{n-m}$, where $|J|$ denotes the Jacobian of the transformation:

$$\mathbf{Z} \rightarrow (\mathbf{S}(\mathbf{Z}),\ \mathbf{R}(\mathbf{Z})), \tag{7.2.10}$$

$\mathbf{R}(\mathbf{Z}) := (R_1, ..., R_{n-m})$, is a *complete sufficient* statistic and $\mathbf{S}(\mathbf{Z}) := (S_1, ..., S_m)$ a *maximal ancillary* statistic, and $\mathbf{S}(\mathbf{Z})$ and $\mathbf{R}(\mathbf{Z})$ are independent.

The separation in (7.2.9) means that all primary inferences can be based exclusively on $f(\mathbf{s};\boldsymbol{\theta})$, and $f(\mathbf{r})$ (free of $\boldsymbol{\theta}$) can be used to validate the statistical model in question.

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

Example. For the simple Normal model (table 2), (7.2.9) holds with the minimal sufficient statistic being $\mathbf{S}{:=}(\overline{X}_n, s^2)$:

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

and the maximal ancillary statistics being $\mathbf{R}(\mathbf{X}){=}(\widehat{v}_3, .., \widehat{v}_n)$, where:

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

are known as the *studentized* residuals. This result explains why it's no accident that the majority of M-S tests rely on the residuals.

Example. This result also holds for the Normal/Linear Regression model, where the one-to-one transformation in (7.2.10) takes the form:

$$(y_1, y_2, ..., y_n) \longleftrightarrow (\widehat{\boldsymbol{\beta}}, s^2, \widehat{v}_{k+2}, .., \widehat{v}_n), \tag{7.2.11}$$

where $\mathbf{S}{:=}(\widehat{\boldsymbol{\beta}}, s^2)$: $\widehat{\boldsymbol{\beta}}{=}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, $s^2{=}\frac{1}{T-k}\widehat{\mathbf{u}}^\top\widehat{\mathbf{u}}$, is the minimal sufficient statistic,
and the maximal ancillary statistic is the studentized residuals $\mathbf{r}{:=}(\widehat{v}_{k+2}, .., \widehat{v}_n)$:

$$\widehat{v}_t = \left(\frac{(y_t - \mathbf{x}_t^\top\widehat{\boldsymbol{\beta}})}{s\sqrt{(1-h_{tt})}}\right) \backsim \mathsf{St}(n-k), \; t = 1, 2, .., n,$$

where h_{tt} denotes the t-th diagonal element of $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

How general is this result? In addition to the simple Normal and the Normal/Linear Regression model, the result:

$$f(\mathbf{z}; \boldsymbol{\theta}) = |J| \cdot f(\mathbf{s}, \mathbf{r}; \boldsymbol{\theta}) = |J| \cdot f(\mathbf{s}; \boldsymbol{\theta}) \cdot f(\mathbf{r}), \; \forall\, (\mathbf{s}, \mathbf{r}) \in \mathbb{R}_s^m \times \mathbb{R}_r^{n-m}, \tag{7.2.12}$$

can be extended to the analogous simple and regression models associated with univariate and multivariate Exponential family of distributions; see Spanos (2007). Moreover, all statistical techniques in econometrics that rely on asymptotic Normality, invoke this result 'approximately'.