Summer Seminar: Philosophy of Statistics
Lecture Notes 11: Mis-Specification (M-S) Testing

Aris Spanos [SUMMER 2019]

1 Introduction

The problem of *statistical misspecification* arises from imposing (directly or indirectly) invalid probabilistic assumptions on data $\mathbf{x}_0$:=$(x_1, ..., x_n)$ by selecting an inappropriate statistical model:

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

$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ defines the premises for statistical (inductive) inferences drawn on the basis of data $\mathbf{x}_0$, and its probabilistic assumptions are selected with a view to render $\mathbf{x}_0$ a 'typical realization' thereof. This 'typicality' is tested using Mis-Specification (M-S) testing to establish the *statistical adequacy of* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, i.e. the validity of its probabilistic assumptions vis-a-vis data $\mathbf{x}_0$. When any of the model assumptions are invalid, both $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$ and the likelihood function:

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

are invalidated. This, in turn invalidates and distorts the sampling distribution $f(y_n;\boldsymbol{\theta})$ of any statistic Y_n=$g(\mathbf{X})$, where $\mathbf{X}$:=$(X_1, X_2, ..., X_n)$, used for inference since $f(y_n;\boldsymbol{\theta})$ is derived via:

$$F(Y_n\leq y)=\underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\ g(\mathbf{x})\leq y\}} f(\mathbf{x};\boldsymbol{\theta})d\mathbf{x},\ \forall y\in\mathbb{R}, \tag{1}$$

This, in turn, will undermine the reliability of any inference procedure based on Y_n=$g(\mathbf{X})$ by derailing its optimality, e.g. rendering an estimator inconsistent or/and inducing sizeable discrepancies between the actual error probabilities (type I, II, p-values, coverage) and the nominal (assumed) ones – the ones derived by invoking the model assumptions. Applying a .05 significance level test, when the actual type I error is closer to .9, will lead an inference astray.

It is important to point out that statistical misspecification also undermines non-parametric inferences that rely on broader statistical models $\mathcal{M}_F(\mathbf{x})$ that include dependence and heterogeneity assumptions. Due to the reliance on the likelihood function, Bayesian inference is equally vulnerable to statistical misspecification since the posterior distribution is:

$$\pi(\boldsymbol{\theta}|\mathbf{x}_0)\propto\pi(\boldsymbol{\theta})\cdot L(\boldsymbol{\theta};\mathbf{z}_0),\ \boldsymbol{\theta}\in\Theta.$$

This is also true for Akaike-type model selection procedures relying on $L(\boldsymbol{\theta};\mathbf{z}_0)$; see Spanos (2010b).

Mis-Specification (M-S) testing plays a crucial role in empirical modeling because it evaluates the validity of the model assumptions; the soundness of the premises of inductive inference. Its usefulness is twofold:

(i) it can alert a modeler to potential inference unreliability problems, and

(ii) it can shed light on the nature of departures from the model assumptions that could help with the model respecification: selecting a another statistical model with a view to account for the chance regularity patterns exhibited by the data.

M-S is a crucial facet of modeling because it can be used to secure the reliability and precision of inference, giving rise to trustworthy evidence.

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 15.1-15.2.

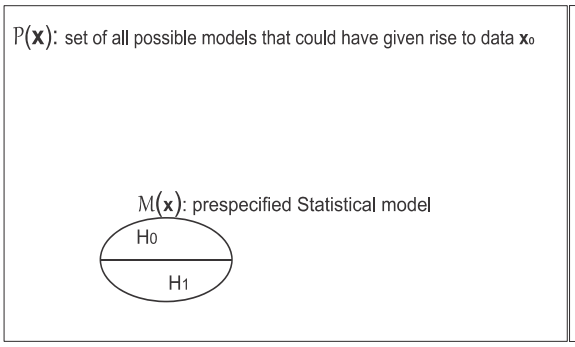


Fig. 15.1: Testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: N-P

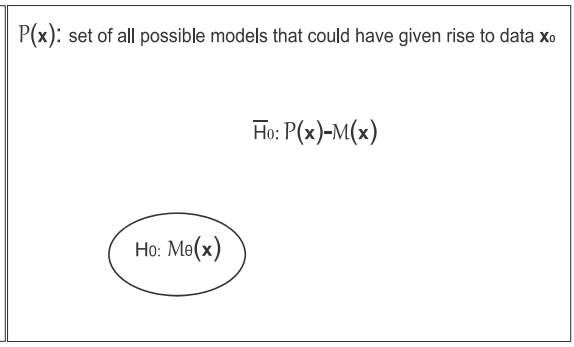


Fig. 15.2: Testing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: M-S

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 M-S testing probes outside those boundaries. N-P testing partitions the assumed model using the parameters as an index. In contrast, 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 $\overline{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})}=[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$. However, $\overline{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})}$ cannot be explicitly operationalized, and thus M-S testing is more open-ended than N-P testing, depending 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).

The current neglect of statistical adequacy is mainly due to the fact that the current statistical modeling is beclouded by conceptual unclarities stemming from the absence of a coherent empirical modeling framework that delineates the different facets of modeling and inference left a lot of unanswered questions concerning its nature and role.

- ▶ How does M-S testing differ from other forms of testing?
- ▶ How would one establish the validity of the model assumptions in practice?
- ▶ What would one do next when certain assumptions are found wanting?

A pioneer of 20th century statistics, a student of Fisher, attests to that by acknowledging the absence of a systematic way to validate statistical models: "The current statistical methodology is mostly model-based, without any specific rules for model selection or validating a specified model." (Rao, 2004, p. 2)

2 Misspecification and inference: a first view

Statistical adequacy ensures that the optimal properties of estimators and tests are real and not notional, and secures the reliability and precision of inference by guaranteeing that the relevant actual error probabilities approximate closely the nominal ones. In contrast, the presence of statistical misspecification induces a discrepancy between these two error probabilities and undermines the reliability of inference.

2.1 Actual vs. nominal error probabilities

Simple Normal model. Consider a simple (one parameter – σ^2 is assumed known) Normal model in table 15.1.

Table 15.1: 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:	$Var(X_t)=\sigma^2,\ \forall t\in\mathbb{N}$,
[4]	Independence:	$\{X_t,\ t\in\mathbb{N}\}$ is an independent process.

In chapter 13, it was shown 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: $T_\alpha:=\{\kappa(\mathbf{X}), C_1(\alpha)\}$:

$$\kappa(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma},\quad C_1(\alpha)=\{\mathbf{x}\text{: } \kappa(\mathbf{x})>c_\alpha\}, \tag{3}$$

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

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

the *type I error probability* (significance level) α is: $\mathbb{P}(\kappa(\mathbf{X})>c_\alpha;\ H_0 \text{ true})=\alpha$.
To evaluate the *type II error probability* and the power of this test:

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

the relevant sampling distribution is:

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

What is often insufficiently emphasized in statistics textbooks is that when any of the assumptions [1]-[4] are invalid for data $\mathbf{x}_0$, the above nominal error probabilities are likely to be significantly different from *actual* error probabilities, rendering inferences based on (3) *unreliable*.

Example 15.1. To illustrate how the nominal and actual error probabilities can differ when any of the assumptions [1]-[4] are invalid, let us consider the case where the *independence assumption* [4] is invalid. Instead, the underlying process $\{X_t,\ t\in\mathbb{N}\}$ is correlated:

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

For a similar example where the dependence is Markov, $Corr(X_i,X_j)=\rho^{|i-j|}$, see Spanos (2009). How does $\rho\neq 0$ affect the reliability of test T_α? The *actual* sampling distributions of $\kappa(\mathbf{X})$ are now:

$$\begin{aligned}&\text{(i)}^*\ \kappa(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}\overset{\mu=\mu_0}{\backsim}\mathsf{N}\left(0,d_n(\rho)\right),\\ &\text{(ii)}^*\ \kappa(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}\overset{\mu=\mu_1}{\backsim}\mathsf{N}\left(\tfrac{\sqrt{n}(\mu_1-\mu_0)}{\sigma},d_n(\rho)\right)\\ &\qquad d_n(\rho)=(1+(n-1)\rho)>1 \text{ for } 0<\rho<1 \text{ and } n>1.\end{aligned} \tag{7}$$

To trace how $\rho\neq 0$ induces discrepancies between the nominal and actual error probabilities, consider the case where $\alpha=.05$ ($c_\alpha=1.645$), $\sigma=1$ and $n=100$. To find the actual type I error probability we need to evaluate the tail area of the distribution in (i)* beyond $c_\alpha=1.645$:

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

where $Z\backsim\mathsf{N}(0,1)$. The results in table 15.2 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 the actual type I error probability increases with the value of ρ as shown in table 15.2.

Table 15.2: Type I error of T_α when $Corr(X_i,X_j)=\rho$

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

Table 15.3: Power $\pi^*(\mu_1)$ **of** T_α **when** $Corr(X_i,X_j)=\rho$

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

Similarly, the *actual power* for $\rho \neq 0$ should be evaluated using:

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

giving rise to the results in table 15.3. Looking at these results closely it is clear for values of μ_1 close to the null $(.01, .02, .05, .1)$, the power *increases* as $\rho \to 1$, but for values of μ_1 away from the null $(.2, .3, .4)$, the power *decreases*. The conventional wisdom often believes that an increase in power is always a good thing. It is not when μ_1 is very close to the null because this destroys the 'probativeness' of a test by rendering it vulnerable to the fallacy of rejection! The test 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; see Mayo (1996).

The Linear Regression model. A more realistic example is the case where the prespecified statistical model is the Normal, Linear Regression (LR) model in table 15.4. When estimating the LR model, it can happen that the modeler ignores mean heterogeneity issues. To illustrate how that can devastate the reliability of inference let us compare two scenarios.

Table 15.4: Normal, Linear Regression model

Statistical GM:	$Y_t=\beta_0+\beta_1 x_t+u_t, \quad t\in\mathbb{N}:=(1,2,...,n,...)$		
[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:	$(\beta_0,\beta_1,\sigma^2)$ are *not* changing with t,	
$\beta_0=(\mu_1-\beta_1\mu_2)\in\mathbb{R}$, $\beta_1=\left(\frac{\sigma_{12}}{\sigma_{22}}\right)\in\mathbb{R}$, $\sigma^2=(\sigma_{11}-\frac{\sigma_{12}^2}{\sigma_{22}})\in\mathbb{R}_+$.			

Scenario 1. The estimated LR model is statistically adequate; assumptions [1]-[5] are valid for the particular data; the true and estimated models coincide: $Y_t=\beta_0+\beta_1 x_t+u_t$.

Scenario 2. The modeler estimates the LR $Y_t=\beta_0+\beta_1 x_t+u_t$, but the true model is $Y_t=\delta_0+\delta_1 t+\beta_1 x_t+u_t$. This renders the estimated model statistically misspecified because part of assumption [5] is invalid; β_0 is not t-invariant, instead $\beta_0(t)=\delta_0+\delta_1 t$. As shown in chapter 7, such a case can easily arise in practice when the data exhibit mean heterogeneity; see Example 7.22.

Example 15.2 (Spanos and McGuirk, 2001). To illustrate how the misspecification of ignoring the trend in the a LR model will seriously undermine the reliability of inference, let us use simulation for the above two scenarios using $N=10,000$ replications. As can be seen from the simulation results in table 15.5, when the estimated LR model is *statistically adequate*: (i) the point estimates are *highly accurate* and the empirical 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 to n=100. It is worth noting that when the estimated statistical model is statistically adequate, the estimates are close to the true parameter values, and the empirical error probabilities are very close to the nominal ones.

Table 15.5: Linear Regression (LR) and mean heterogeneity

	Adequate LR model				**Misspecified LR model**			
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+.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	Std	Mean	Std	Mean	Std	Mean	Std
[β_0=1.5] $\hat{\beta}_0$	1.502	.122	1.500	.087	0.462	.450	0.228	.315
[β_1=.5] $\hat{\beta}_1$	0.499	.015	0.500	.008	1.959	.040	1.989	.015
[σ^2=.75] $\hat{\sigma}^2$	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
$\tau_{\beta_0}=\frac{\hat{\beta}_0-\beta_0}{\hat{\sigma}_{\beta_0}}$	0.004	.049	0.015	.050	-1.968	0.774	-3.531	0.968
$\tau_{\beta_1}=\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\beta_1}}$	-.013	.047	-.005	.049	35.406	1.000	100.2	1.000

In contrast, when the *estimated model is misspecified* [assumption [5] is invalid]: (iii) the point estimates are *highly inaccurate* (a symptom of inconsistent estimators) and the empirical type I error probabilities are *much larger* than the nominal (α=.05), and (iv) as n increases the inaccuracy of the estimates increases (they get further and further away from the true values as n increases) and the empirical type I error probabilities approach 1! This brings out the folly of the widely touted assertion that when n is large enough the modeler does not need to worry about statistical misspecification.

It is important emphasize that these examples are only indicative of an actual situation facing a practitioner. More often than not, more than one of the model assumptions are invalid. This renders the reliability of inference a lot more dire in practice than these examples might suggest; see Spanos and McGuirk (2001).

2.2 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. The question that naturally arises:

► In light of the dire consequences of statistical misspecification, why is there such reluctance in most applied fields to secure statistical adequacy [validate the statistical premises] using thorough Mis-Specification (M-S) testing?

A crucial reason for this neglect is that the empirical modeling literature appears to **seriously underestimate the potentially devastating effects of statistical misspecification on the reliability of inference**. This misplaced confidence in this practice stems from a number of questionable arguments, including the following.

(i) Empirical modeling is misleadingly viewed **as a curve-fitting exercise guided by goodness-of-fit measures** where a substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ is foisted on the data. This stems from presuming that $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ valid on *a priori* grounds by viewing it as established knowledge, instead of tentative conjunctures to be confronted with data.

(ii) **Confusing statistical with substantive adequacy**.The statistical premises are inadvertently blended with the substantive premises of inference, and the empirical literature conflates two very different forms of *misspecification*: statistical and substantive. Hence, it should come as no surprise that econometrics textbooks consider 'omitted variables' as the most serious form of statistical misspecification (omitted variables bias and inconsistency; see Greene, 2012), when in fact it has nothing to do with the statistical assumptions; it is an issue relating to substantive adequacy (Spanos, 2006c).

(iii) The confusion between statistical and substantive misspecification also permeates the argument attributed to George Box (1979) that "**all models are wrong, but some are useful**". A careful reading of Box (1979), p. 202, however, shows that he was talking about 'substantive' inadequacy and not statistical. It is one thing to claim that one's model is not 'realistic enough', and quite another to turn a blind eye to the problem of imposing invalid probabilistic assumptions on one's data, and then proceed to use the inference procedures that invoke these assumptions as if the invalid assumptions do not matter. Indeed, in an even earlier publication Box (1976) argued for a balanced approach between theory and data: "One important idea is that science is a means whereby learning is achieved, not by mere theoretical speculation on the one hand, nor by the undirected accumulation of practical facts on the other, but rather by a motivated iteration between theory and practice." (p. 791).

(iv) The current undue reliance on asymptotic procedures for learning from data. Such undue reliance ignores the fact that limit theorems invoked by Consistent and Asymptotically Normal (CAN) estimators and associated tests, also rely on probabilistic assumptions which are usually non-testable, rendering the reliability of the resulting inferences dubious. Worse, 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 (1986a, p. xiv): "... limit theorems "as n tends to infinity" are logically devoid of content about what happens at any particular n."

Believing that the validity of asymptotic inference results stemming from invoking the heuristic 'as $n \to \infty$' is just an illusion. The trustworthiness of any inference results invoking a CAN estimator relies solely on the approximate validity of the probabilistic assumptions imposed on $\mathbf{x}_0$ for the specific n, and nothing else.

(v) There is also the undue reliance on **vague 'robustness' results** whose gener-

ality 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; see Spanos and McGuirk (2001), Spanos and Reade (2015).

(vi) Finally, the neglect of M-S testing to secure the statistical adequacy of an estimated model is often explained away using ill-thought out **methodological criticisms against M-S testing**, 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 Spanos (2000; 2010a; 2018) for a rebuttal of these criticisms; see Appendix for a summary.

The Probabilistic Reduction (PR) perspective. Distinguishing between the 'statistical' and 'substantive' information, ab initio, and viewing a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in purely probabilistic terms, enables one to address a number of problems associated with several conceptual and practical problems in establishing statistical adequacy, to secure the reliability of inference.

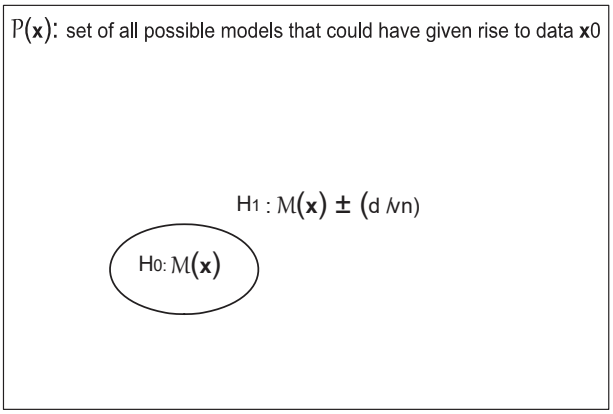
First, this distinction delineates two different questions often conflated in practice:
[a] statistical adequacy: does $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ account for the chance regularities in $\mathbf{x}_0$? $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is built exclusively on the statistical information contained in data $\mathbf{x}_0$, and acts as a *mediator* between $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ and $\mathbf{x}_0$.
[b] substantive adequacy: does the model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ adequately captures (describes, explains, predicts) the phenomenon of interest? Substantive inadequacy arises, not from invalid probabilistic assumptions, but from highly unrealistic structural models, flawed *ceteris paribus* clauses, missing confounding factors, systematic approximation error, etc. In this sense, probing for substantive adequacy is a considerably more complicated problem, which, at the very minimum, includes the validity of the overidentifying restrictions stemming from $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$, after securing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Without it, the reliability of such tests is questionable.

The second practical problem addressesed by the defining $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ as comprising all the probabilistic assumptions imposed on the particular data $\mathbf{x}_0$ is the need for a complete list of *testable* probabilistic assumptions, which is almost never available in textbooks. Usually the list of assumptions is often incomplete, specified in terms of the unobservable error term and some of the assumptions are not testable; see chapter 14, section 2.1. This undermines the effectiveness of any form of M-S testing, rendering it ad hoc and partial at best.

The third important problem addressesed by the PR perspective is the confusion between M-S and Neyman-Pearson (N-P) testing, stemming from using the same test procedures, Likelihood-ratio, Lagrange Multiplier and Wald, for both types of testing; Spanos (1986). This has led to a number of misleading claims and charges against M-S testing such as calling into question the legitimacy of the latter, including 'vulnerability to multiple testing, 'illegitimate double use of data', 'pre-test bias', 'infinite regress', etc.; see Spanos (2010b).

3 Nonparametric (omnibus) M-S tests

Why omnibus M-S tests? A crucial advantage of the **omnibus (nonparametric) tests** is that they probe more broadly around the $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ **(locally**) 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!



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

To get some idea of what M-S testing is all about, however, let us focus on a few simple tests to assess assumptions [1]-[4] of the simple Normal model (table 15.1 with σ^2 an *unknown* parameter).

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

To avoid misinterpretations, it is important to keep in mind that the description 'NIID data' is shorthand for 'the sample $\mathbf{X}$:=$(X_1, X_2, \ldots, X_n)$ giving rise to the plotted data $\mathbf{x}_0$:=$(x_1, x_2, \ldots, x_n)$ is NIID'. The chance regularity pattern of focus in this section is the *unpredictability* of the direction of change (ups and downs), and corresponds to the probabilistic concept of *independence.*

Thought experiment 1. *Imagine hiding away the plot to our right of a particular t, and trying to predict the direction (up or down) of the next few observations. When one has difficulty guessing correctly the direction of the next observation, the data exhibit independence.*

Runs up and down. A nonparametric way to look at this unpredictability of the direction of change is to **ignore the values taken by the data series altogether** and concentrate just on the **sign (direction)** of the differences (changes).

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:

$$(x_1, ..., x_n) \rightarrow \{(x_t - x_{t-1}),\ t=2, ..., n\} \rightarrow [\underbrace{--}_{1}\ \underbrace{+}_{2}\ \underbrace{-}_{3}\ \underbrace{+}_{4}\ \underbrace{--}_{5}\ \underbrace{++}_{6}\ \underbrace{-}_{7}\ \underbrace{+}_{8}\ \underbrace{--}_{9}] \tag{8}$$

Step 3: count the number of runs and their length. The runs for observations $t=1-14$ (figure 5.3) which consists of 9 runs of different lengths; the first is a run with 2 negative signs, the second run with 1 positive sign, are shown in (8). From this sequence of +'s and −'s no regular pattern to be used for guessing the next up or down can be discerned.

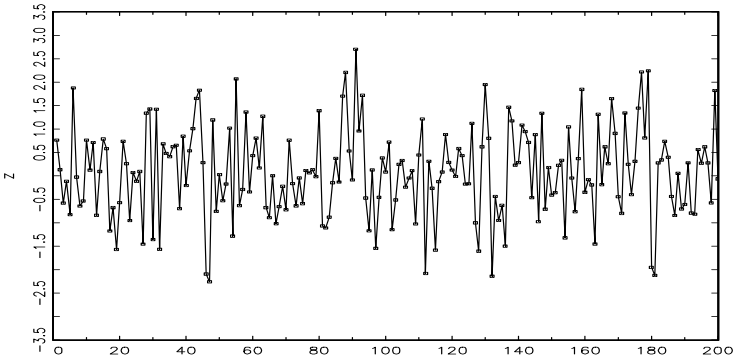


Fig. 5.3: Typical realization of NIID data

Example 5.1. In the case of figure 5.3 the number of runs up and down for observations $t=1-50$ are:

$$\{1,1,2,1,2,1,2,1,1,2,1,1,2,1,4,1,1\}^{+},\ \{2,1,2,1,1,2,1,2,1,1,1,1,2,1,1,3,1\}^{-}$$

By treating the sequence of pluses and minuses as Bernoulli trials we can use combinatorial arguments to evaluate the number of runs expected under the assumptions that the observations were IID; see Levene (1952).

Example 5.2. Defining the random variable R-number of runs of any size, combinatorial arguments can be used to show (Levene, 1952) that:

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

These two moments can be used to construct the simplest *runs test* based on comparing the *actual* number of runs R with the number of *expected* runs $E(R)$ to frame the distance function:

$$d_R(\mathbf{X}) = \frac{[R - E(R)]}{\sqrt{Var(R)}} = \frac{R - ([(2n-1)/3])}{\sqrt{(16n-29)/90}}.$$

'Large enough' values of $d_R(\mathbf{x}_0)$, positive or negative, indicate possible departures from the IID assumptions. More formally, one can show that the distribution of $d_R(\mathbf{X})$ for $n \geq 40$ can be approximated by:

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

and 'large enough' is now measured in terms of the tails of the standard Normal distribution ($\mathsf{N}(0,1)$).

The hypothesis of interest concerns any 'random' *reordering* of the sample $\mathbf{X}:=(X_1, X_2, ..., X_n)$, 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 $(i=1, 2, ..., n)$.

Runs up and down test. The *runs test,* discussed in chapter 5, compares the *actual* number of runs R with the number of *expected* runs, assuming that $\{X_t,\ t\in\mathbb{N}\}$ were an IID process, to construct the test:

$$d_R(\mathbf{X}) = \frac{[R-E(R)]}{\sqrt{Var(R)}},\ C_1(\alpha) = \{\mathbf{x}\text{: } |d_R(\mathbf{x})| > c_{\frac{\alpha}{2}}\},\ E(R) = \frac{2n-1}{3},\ Var(R) = \frac{16n-29}{90},$$

and show that the distribution of $d_R(\mathbf{X})$, for $n \geq 40$, can be approximated by:

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

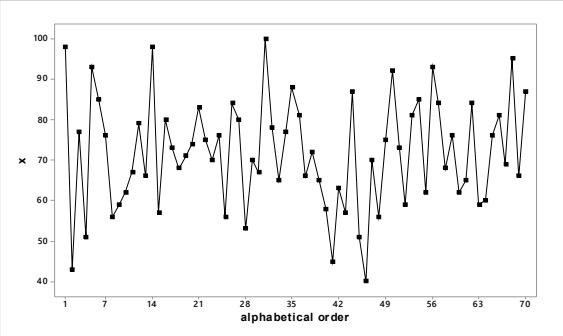


Fig. 15.3: Exam scores: alphabetical order

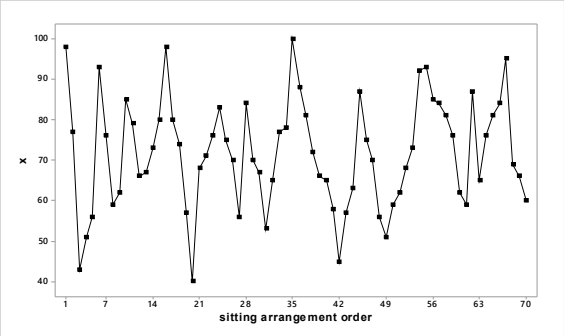


Fig. 15.4: Exam scores: sitting order

Example 15.3 (a). Consider the exam scores in figure 15.3, for which $d_R(\mathbf{x}_0) = \frac{50-46.3}{3.482} = 1.062$. Evaluating the p-value (chapter 13) yields:

$$p(\mathbf{x}_0) = \mathbb{P}(d_R(\mathbf{X}) > d_R(\mathbf{x}_0);\ H_0) = .144,$$

which indicates no departures from the IID assumptions.

Example 15.3 (b). On the other hand, the same data ordered according to the sitting arrangement (figure 15.4), yield $d_R(\mathbf{x}_0) = \frac{21-46.3}{3.482} = -7.266$, with a p-value $p(\mathbf{x}_0) = .000000$, which clearly indicates strong departures from the IID ([2]-[4]) assumptions.

3.2 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 $\mathbf{X}:=(\ X_1, X_2, ..., X_n)$,

(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:

$$H_0\text{: } F^*(x)=F_0(x),\ x\in\mathbb{R}, \tag{9}$$

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):

$$\lim_{n\to\infty}\mathbb{P}(\sqrt{n}\Delta_n(\mathbf{X})\leq x)=F_K(x)=1-2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2x^2}\simeq 1-2\exp(-2x^2),\ x>0. \tag{10}$$

Since $F_K(x)$ is known (approximated), 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 15.4. Applying the Kolmogorov test to the scores data in fig. 1.12 yields:

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

which indicates no significant departure from the Normality assumption. The P-P plot in figure 15.5 provides a depiction of what this test is measuring in terms of the discrepancies from the line to the observed points (chapter 5).

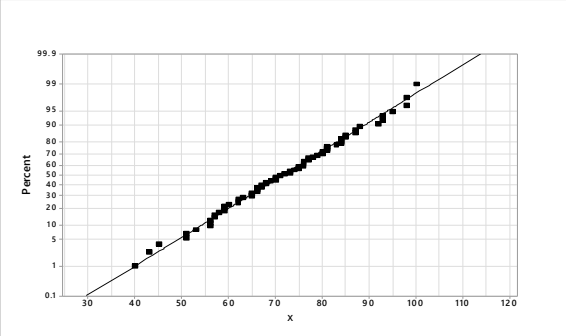


Fig. 15.5: P-P Normality plot

Note that this particular test might be too sensitive to outliers because it picks up only the biggest distance! A different distance function which is less sensitive to outliers is:

$$\mathsf{A\text{-}D}(\mathbf{X})=n\int_{-\infty}^{\infty}\frac{[\widehat{F}_n(x)-F_0(x)]^2}{F_0(x)(1-F_0(x))}f_0(x)dx, \tag{11}$$

proposed by Anderson and Darling (1952), which for the ordered sample $\mathbf{X}_{[n]}$ is:

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

In the above example the A-D test yielded the p-value:

$$\mathbb{P}(\mathsf{A\text{-}D}(\mathbf{X})>.139;H_0)=.974,$$

which confirms the result of the Kolmogorov test.

4 Parametric (directional) M-S testing

Parametric M-S tests are of 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.

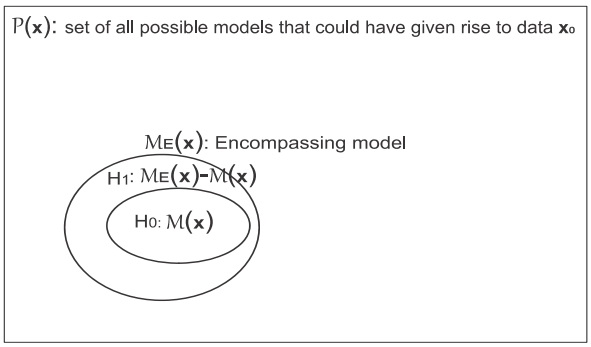


Fig. 15.6: M-S testing by encompassing

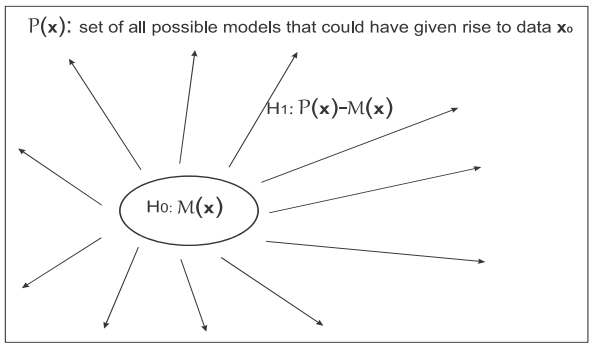


Fig. 15.7: M-S testing: directions of departures

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

In the case of the simple Normal model (table 15.1), the process $\{X_t,\ t\in\mathbb{N}\}$ is assumed to be NIID and thus the simplification of the distribution of the sample is:

$$f(x_1,x_2,...,x_n;\boldsymbol{\phi})\overset{\text{IID}}{=}\prod_{t=1}^{n} f(x_t;\boldsymbol{\varphi}),\ \forall\mathbf{x}\in\mathbb{R}^n.$$

Relaxing the IID assumptions and replacing them with Markov (M) dependence and stationarity (S), the simplification of the distribution of the sample resulting from sequential conditioning takes the form (ch. 7):

$$f(x_1,x_2,...,x_n;\boldsymbol{\phi})\overset{\text{MS}}{=}f(x_1;\varphi_1)\prod_{t=2}^{n} f(x_t|x_{t-1};\boldsymbol{\varphi}),\ \forall\mathbf{x}\in\mathbb{R}^n,$$

where the joint distribution $f(x_t,x_{t-1};\boldsymbol{\varphi})$ takes the form:

$$\begin{pmatrix} X_t \\ X_{t-1}\end{pmatrix}\backsim \mathsf{N}\left(\begin{bmatrix}\mu\\ \mu\end{bmatrix},\begin{bmatrix}\sigma(0) & \sigma(1)\\ \sigma(1) & \sigma(0)\end{bmatrix}\right). \tag{12}$$

As shown in chapter 7, this gives rise to the AutoRegressive [AR(1)] model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})$ based on $f(x_t|x_{t-1};\boldsymbol{\theta})$, whose statistical GM is:

$$\begin{aligned} &X_t=\alpha_0+\alpha_1X_{t-1}+\varepsilon_t,\ t\in\mathbb{N},\\ &\alpha_0=\mu(1-\alpha_1)\in\mathbb{R},\ \alpha_1=\tfrac{\sigma(1)}{\sigma(0)}\in(-1,1),\ \sigma_0^2=\sigma(0)(1-\alpha_1^2))\in\mathbb{R}_+.\end{aligned} \tag{13}$$

The AR(1) parametrically nests (includes as a special case) the simple Normal model, and the nesting restriction is $\alpha_1=0$. Under this restriction the AR(1) model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})$ reduces to the simple Normal model (table 15.1):

$$X_t=\alpha_0+\alpha_1X_{t-1}+\varepsilon_t\ \overset{\alpha_1=0}{\rightarrow}\ X_t=\mu+u_t,\ t\in\mathbb{N}.$$

This suggests that the nesting restriction framed as a test of the hypotheses:

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

in the context of $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})$, can be used to test assumption [4] (table 15.1).

The M-S test for assumption [4] is the *t-type* test $T_\alpha{:=}\{\tau(\mathbf{X}), C_1(\alpha)\}$:

$$\begin{aligned}
&\tau(\mathbf{X}){=}\tfrac{(\widehat{\alpha}_1-0)}{\sqrt{Var(\widehat{\alpha}_1)}} \overset{H_0}{\approx} St(n{-}2),\ C_1(\alpha){=}\{\mathbf{x}\text{: } |\tau(\mathbf{x})|>c_\alpha\},\\
&\widehat{\alpha}_1{=}\tfrac{\sum_{t=1}^n(X_t-\overline{X})(X_{t-1}-\overline{X})}{\sum_{t=1}^n(X_{t-1}-\overline{X})^2},\ Var(\widehat{\alpha}_1){=}s^2[\textstyle\sum_{t=1}^n(X_{t-1}-\overline{X})^2]^{-1},\\
&s^2{=}\tfrac{1}{n-2}\textstyle\sum_{t=1}^n(X_t-\widehat{\alpha}_0-\widehat{\alpha}_1X_{t-1})^2,\ \widehat{\alpha}_0{=}(1-\widehat{\alpha}_1)\overline{X}.
\end{aligned} \tag{15}$$

Example 15.5. For the data in figure in fig. 15.4, estimating the AR(1) model yields:

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

giving rise to the M-S t-test (15) for (14):

$$\tau(\mathbf{x}_0){=}(\tfrac{.441}{.106}){=}4.160,\ p(\mathbf{x}_0){=}.000016,$$

indicating a clear departure from assumption [4], confirming example 15.3(b).

The nesting of the original statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ within a more general statistical model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})$ gives the impression that the M-S test based on (14) has been transformed into a N-P test. Although this is technically correct, it is conceptually erroneous because there is no assumption that the nesting model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})$ is statistically adequate, as it should be the case for a reliable N-P test. The only role played by the nesting model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})$ is to provide possible directions of departure from the original model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Hence, the only clear inference one can draw from a nesting M-S test pertains exclusively to the original model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: whether $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is misspecified in the direction of departure indicated by $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})$ or not. When $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is rejected as misspecified, one cannot infer the validity of $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})$; a classic example of the fallacy of rejection; see chapter 13.

4.2 Testing Independence & mean constancy ([2],[4])

The above t-type M-S test based on the auxiliary regression (13) can be extended to provide a joint test for assumptions [2] and [4]. The nesting AR(1) model was derived in chapter 8 by replacing the stationarity assumption of $\{X_t,\ t{\in}\mathbb{N}\}$ with mean non-stationarity, that changes the $f(x_{t-1},x_t;\boldsymbol{\varphi})$ in (12) into:

$$\begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix} \backsim \mathsf{N}\left(\begin{bmatrix} \mu{+}\gamma_1 t \\ \mu{+}\gamma_1(t{-}1) \end{bmatrix}, \begin{bmatrix} \sigma(0) & \sigma(1) \\ \sigma(1) & \sigma(0) \end{bmatrix}\right). \tag{16}$$

This new $f(x_{t-1},x_t;\boldsymbol{\varphi}(t))$ gives rise to a heterogeneous AR(1) model with a statistical GM:

$$\begin{aligned}
&X_t{=}\delta_0+\overbrace{\delta_1 t}^{\overline{[2]}}+\overbrace{\alpha_1X_{t-1}}^{\overline{[4]}}+\varepsilon_t,\ t{\in}\mathbb{N},\\
&\delta_0{=}\mu{+}\alpha_1(\gamma_1{-}\mu),\ \delta_1{=}(1{-}\alpha_1)\gamma_1 t,\ \alpha_1{=}(\sigma(1)/\sigma(0)),\ \sigma_0^2{=}\sigma(0)(1{-}\alpha_1^2).
\end{aligned} \tag{17}$$

The nesting restrictions α_1=0, δ_1=0 reduce the AR(1) in (17) to the simple Normal model:

$$X_t=\delta_0+\delta_1 t+\alpha_1 X_{t-1}+\varepsilon_t \quad \underset{\delta_1=0}{\overset{\alpha_1=0}{\to}} \quad X_t=\mu+u_t,\ t\in\mathbb{N}.$$

Hence, a joint M-S test for assumptions [2],[4] (table 15.1), based on the hypotheses:

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

is a *F-type* test T_α:=$\{\tau(\mathbf{X}), C_1(\alpha)\}$ that takes the form:

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

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

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

Example 15.6. For the data in figure 15.4, the restricted and unrestricted models yielded, respectively:

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

$$x_t=\underset{(8.034)}{38.156}+\underset{(.073)}{.055}t+\underset{(0.107)}{0.434}x_{t-1}+\widehat{\varepsilon}_t,\quad s^2=144.34,\ n=69, \tag{18}$$

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 discordance with the null ([2]&[4]).

What is particularly notable about the auxiliary autoregression (18) is that a closer look at the t-ratios indicates that the source of the problem is dependence and *not* t-heterogeneity, but dependence, as the two t-ratios indicate:

$$\tau_1(\mathbf{x}_0)=\left(\tfrac{.055}{.073}\right)=.753,\ p(\mathbf{x}_0)=.226,\ \tau_2(\mathbf{x}_0)=\left(\tfrac{.434}{.107}\right)=4.056,\ p(\mathbf{x}_0)=.0000,$$

a clear departure from assumption [4], and not from [2]. This information enables one to apportion blame, which is not possible when using a nonparametric test, such as the runs test, and suggests ways to respecify the original model to account for such systematic statistical information.

Using the residuals. An alternative way to specify in (17) is in terms of the *residuals* $\widehat{u}_t$=$x_t-\overline{x}_n$=$(x_t-71.7)$, t=1, 2, ..., n, because that the auxiliary regression:

$$\widehat{u}_t=\underset{(8.034)}{-33.534}+\underset{(.073)}{.055}t+\underset{(0.107)}{0.434}x_{t-1}+\widehat{\varepsilon}_t,\quad s^2=144.34,\ n=69, \tag{19}$$

is a mirror image of (18) with identical parameter estimates, apart from the constant, which is irrelevant for M-S testing purposes.

4.3 Testing Independence & variance constancy ([2],[4])

In light of the fact that assumption [3] involves the constancy of $\sigma^2=Var(X_t)=E(u_t^2)$, we can combine it with the independence [4] assumption to construct a joint tests using the residuals squared in the context of the auxiliary regression:

$$\widehat{u}_t^2=\gamma_0+\overbrace{\gamma_1 t}^{\overline{[3]}}+\overbrace{\gamma_2 x_{t-1}^2}^{\overline{[4]}}+v_t \overset{\mathbf{x}_0}{\rightarrow} \underset{(89.43)}{\widehat{u}_t^2=295.26}-\underset{(1.353)}{1.035t}-\underset{(.014)}{.016x_{t-1}^2}+\widehat{v}_t.$$

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

4.3.1 Extending the above auxiliary regression*

The auxiliary regression (17), 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 ($\ell \geq 1$):

$$X_t=\delta_0+\textstyle\sum_{k=1}^{m}\delta_k t^k+\sum_{i=1}^{\ell}\alpha_i X_{t-i}+\varepsilon_t,\ t\in\mathbb{N}. \tag{20}$$

In practice, however, going beyond m=3 can give rise to numerical problems because ordinary trend polynomials $(t, t^2, t^3, t^4, t^5, ...)$ are likely to be *collinear*. An effective way to avoid such collinearity problems is to (a) scale the ordering t=1, 2, ..., n, using:

$$t_*=\tfrac{(2t-n-1)}{(n-1)},\ t=1,2,...,n, \tag{21}$$

and then (b) replace the ordinary with *orthogonal* polynomials $(t_o, t_o^2, t_o^3, t_o^4, t_o^5, ...)$, say the *Chebyshev polynomials,* where:

$$t_o=t_*,\ \ t_o^2=2t_*^2-1,\ \ t_o^3=4t_*^3-3t_*,\ \ t_o^4=8t_*^4-8t_*^2+1,\ \ t_o^5=16t_*^5-20t_*^3+5t_*,\ ... \tag{22}$$

4.4 The Skewness-Kurtosis test of Normality

An alternative way to test Normality is to use parametric tests probing for the validity of Normality within a broader nesting family of distributions. Such a test can be constructed using the Pearson family (chapter 10) since the skewness (α_3) and kurtosis (α_4) coefficients (chapter 3) can be used to distinguish between different members. For example, the Normal distribution is characterized within the Pearson family via:

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

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

Within the Pearson family, can frame the hypotheses of interest to be:

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

and construct the *Skewness-Kurtosis test* based on:

$$SK(\mathbf{X})=\tfrac{n}{6}\widehat{\alpha}_3^2+\tfrac{n}{24}\left(\widehat{\alpha}_4-3\right)^2 \underset{\alpha}{\overset{H_0}{\backsim}} \chi^2(2),\ \mathbb{P}(SK(\mathbf{X})>SK(\mathbf{x}_0);\ H_0)=p(\mathbf{x}_0), \tag{23}$$

where the estimated parameters (α_3, α_4) are:

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

It is important to note that the $SK(\mathbf{X})$ M-S test is particularly sensitive to outliers because it involves higher sample moments such as $\sum_{k=1}^{n}(X_k-\overline{X})^4$, which can allow one or two outlier observations to dominate the summation. For an improved version of the $SK(\mathbf{X})$ test; see D'Agostino and Pearson (1973).

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

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

$$\widehat{u}_t=\delta_0+\overbrace{\delta_1 t+\delta_2 t^2}^{\overline{[2]}}+\overbrace{\delta_3 x_{t-1}}^{\overline{[4]}}+\varepsilon_{1t},\ H_0\text{: }\delta_1=\delta_2=\delta_3=0 \text{ vs. } H_1\text{: }\delta_1\neq 0 \text{ or } \delta_2\neq 0 \text{ or } \delta_3\neq 0.$$

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

$$\widehat{u}_t^2=\gamma_0+\overbrace{\gamma_1 t+\gamma_2 t^2}^{\overline{[3]}}+\overbrace{\gamma_3 x_{t-1}^2}^{\overline{[4]}}+\varepsilon_{2t},\ H_0\text{: }\gamma_1=\gamma_2=\gamma_3=0 \text{ vs. } H_1\text{: }\gamma_1\neq 0 \text{ or } \gamma_2\neq 0 \text{ or } \gamma_3\neq 0.$$

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, such as the skewnesss-kurtosis and the Kolmogorov tests.

Example 15.7. Consider the casting of two dice data in table 1.1 (fig. 1.1). The estimated first four moments yield:

$$\overline{x}_n=\tfrac{1}{n}\sum_{k=1}^{n}x_k=7.08,\ s^2=\tfrac{1}{n-1}\sum_{k=1}^{n}(x_k-\overline{x}_n)^2=5.993,\ \widehat{\alpha}_3=-.035, \widehat{\alpha}_4=2.362$$

(a) Testing assumptions [2]-[4] using the runs test, with n=100, R=50:

$$E(R)=(200-1)/3=66.333,\ Var(R)=(16(100)-29)/90=17.456,$$

$$Z_R(\mathbf{X})=\tfrac{72-66.333}{\sqrt{17.456}}=1.356, \quad \mathbb{P}(|Z_R(\mathbf{X})|>1.356;\ H)=.175,$$

(b) Testing assumptions [2] and [4] using the auxiliary regression:

$$\widehat{u}_t=\delta_0+\delta_1 t+\delta_2 X_{t-1}+\varepsilon_{1t} \overset{\mathbf{x}_0}{\rightarrow} \widehat{u}_t=\underset{(.877)}{1.02}-\underset{(.009)}{.005}t-\underset{(.101)}{.103}X_{t-1}+\underset{(2.434)}{\widehat{\varepsilon}_{1t}},$$

indicates no departures since the F-test for the joint significance of δ_1 and α_1:

$$H_0\text{: } \delta_1=\delta_2=0, \text{ vs. } H_1\text{: } \delta_1\neq 0, \text{ or } \delta_2\neq 0,$$

$$F(\mathbf{x}_0)=\tfrac{\text{RRSS-URSS}}{\text{URSS}}(\tfrac{n-4}{2}) \quad =\tfrac{576.6-568.540}{568.540}(\tfrac{96}{2})=.680[.511],$$

and the t-tests for the significance of δ_1 and δ_2 yield:

$$\tau(\mathbf{x}_0)=\tfrac{1.02}{.877}=1.163[.248],\ \tau(\mathbf{x})=\tfrac{.103}{.101}=1.021[.312]. \tag{24}$$

(c) Testing assumptions [3]-[4] using the auxiliary regression:

$$\widehat{u}_t^2=\gamma_0+\gamma_1 t+\gamma_2 x_{t-1}^2+\varepsilon_{2t} \overset{\mathbf{x}_0}{\rightarrow} \widehat{u}_t^2=\underset{(1.84)}{6.10}-\underset{(.024)}{.021}t+\underset{(.02)}{.014}x_{t-1}^2+\widehat{\varepsilon}_{2t},$$

indicates no departures since the F-test for the joint significance of γ_1 and γ_2:

$$H_0\text{: } \gamma_1=\gamma_2=0, \text{ vs. } H_1\text{: } \gamma_1\neq 0, \text{ or } \gamma_2\neq 0,$$

$$F(\mathbf{x}_0)=\tfrac{\text{RRSS-URSS}}{\text{URSS}}(\tfrac{n-4}{2})=\tfrac{4678.67-4619.24}{4619.24}(\tfrac{96}{2})=.618=[.541],$$

and the t-tests for the significance of γ_1 and γ_2 yield:

$$\tau(\mathbf{x}_0)=\tfrac{.021}{.024}=.875[.397],\ \tau(\mathbf{x})=\tfrac{.014}{.02}=.7[.490]. \tag{25}$$

(d) In light of the fact that model assumptions [2]-]4] are valid, one can proceed to test the Normality assumption [1] using the SK test that yields:

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

The p-value indicates no discordance with the Normality assumption.

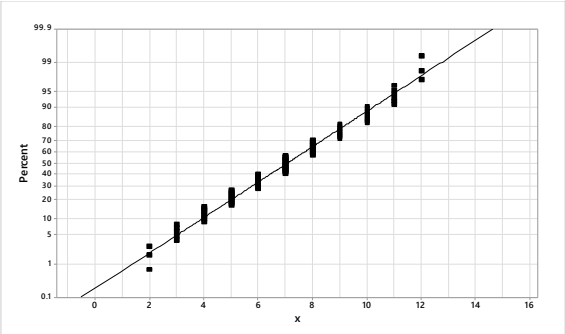


Fig. 15.8: P-P Normality plot for dice data

CAUTIONARY NOTE. This results raises an interesting question because the underlying distribution exhibited by the histogram in figure 1.3 is discrete and triangular; not Normal. What went wrong? The SK test is not a very powerful test

since α_3=0 and α_4=3 characterizes Normality within the Pearson family; it can be fooled by symmetry. A departure from Normality is revealed by the more powerful Anderson-Darling (A-D) test:

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

The P-P plot for Normality in figure 15.7 brings out the problem in no uncertain terms by pointing out the discrete nature of the underlying distribution for the dice data; notice the vertical columns of observations. This brings out the importance of graphical techniques at the modeling stages: specification, M-S testing and respecification.

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

Despite the plethora of M-S tests in both the statistics and econometrics literatures, no systematic way to apply these tests has emerged; see Godfrey (1988). The literature has left practitioners perplexed since numerous issues about M-S testing remained unanswered, including (a) the choice among many different M-S tests as well as their applicability in different models, (b) respecification: what to do when any of the model assumptions are invalid, (c) the use of omnibus (nonparametric) vs. parametric tests, and (d) the differences between M-S testing, N-P testing, Fisher's significance testing and Akaike-type model selection procedures.

5.1 Placing M-S testing in a proper context

It is argued that these issues can be addressed by having a coherent framework where different facets of modeling and inference can be delineated as articulated in chapter 10 (table 10.9) to include *M-S testing* and *Respecification* (with a view to achieve statistical adequacy).

More formally, the key difference between N-P and M-S testing is:

$$\underline{\textbf{Testing within } \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: learning from data about } \mathcal{M}^*(\mathbf{x})}$$

$$H_0\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in\mathcal{M}_0(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta_0\} \text{ vs. } H_1\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in\mathcal{M}_{\mathbf{1}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta_1\}. \tag{26}$$

$$\underline{\textbf{Testing outside } \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: probing for validity of its assumptions}}$$

$$H_0\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) \text{ vs. } \overline{H}_0\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in\overline{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})}=[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\theta}(\mathbf{x})]\,. \tag{27}$$

Note that $\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\}$ denotes the 'true' distribution of the sample.

It is important to emphasize the fact that M-S testing is a form of significance testing where null hypothesis is always defined by:

$$H_0\text{: all assumptions of } \mathcal{M}_{\theta}(\mathbf{x}) \text{ are valid for } \mathbf{x}_0, \tag{28}$$

with the particularized alternative $H_1 \subset [\mathcal{P}(\mathbf{x})-\mathcal{M}_\theta(\mathbf{x})]$ being:

$$H_1\text{: the stated departures from specific assumptions being tested, } \textit{assuming} \text{ the rest of the assumption(s) of } \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) \text{ hold for data } \mathbf{x}_0. \tag{29}$$

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** [accepting the null when false] is often the more serious of the two errors. This because one will have another chance to correct for the type I error [rejecting the null when true] at the respecification stage, where a new model aims to account for the chance regularities the original model ignored. Hence, M-S testing is also more vulnerable to the **fallacy of acceptance**.

(c) In M-S testing the objective is to probe as broadly beyond the null ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is valid) as possible, and thus tests with *low power* but broad (local) probing capacity have an important role to play. Curiously, the low local power is a blessing in M-S testing because when they indicate departures, that provides better evidence for such departures than parametric (directional) tests with very high power.

5.2 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.

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 that stems from the fact that $\mathcal{M}_\theta(\mathbf{x})$ is 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 15.6) 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 some of the other model assumptions are 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 and Spanos, 2004).

Custom tailoring M-S tests. The most effective way to ensure that M-S tests have maximum capacity to detect any potential departures from the model assumptions is to employ graphical techniques judiciously. By looking closely at various data plots, such as t-plots, scatter plots and histograms, one should be able to discern potential departures and design auxiliary regressions, by crafting additional terms, that could reveal such potential departures most effectively.

5.3 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. As shown in chapter 7, 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! As argued in chapter 7, 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 15.6: Normal, Linear Regression model

Statistical GM: $Y_t=\beta_0+\boldsymbol{\beta}_1^{\top}\mathbf{x}_t+u_t,\quad t\in\mathbb{N}:=(1,2,...,n,...)$

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

$\beta_0=(\mu_1-\beta_1\mu_2)\in\mathbb{R},\ \boldsymbol{\beta}_1=\left(\Sigma_{22}^{-1}\boldsymbol{\sigma}_{21}\right)\in\mathbb{R},\ \sigma^2=(\sigma_{11}-\boldsymbol{\sigma}_{21}^{\top}\Sigma_{22}^{-1}\boldsymbol{\sigma}_{21})\in\mathbb{R}_+.$

As argued in chapter 14, 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:

$$E\left(Y_t\mid X_t=x_t\right)=\beta_0+\beta_1x_t,\ Var\left(Y_t\mid X_t=x_t\right)=\sigma^2. \tag{30}$$

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

$$\widehat{u}_t=\delta_0+\delta_1x_t+\overbrace{\delta_2t}^{[5]}+\overbrace{\delta_3x_t^2}^{[2]}+\overbrace{\delta_4x_{t-1}+\delta_5Y_{t-1}}^{[4]}+v_{1t}, \tag{31}$$

H_0: $\delta_1=\delta_2=\delta_3=\delta_4=\delta_5=0$ vs. H_1: $\delta_1\neq0$ or $\delta_2\neq0$ or $\delta_3\neq0$ or $\delta_4\neq0$ or $\delta_5\neq0$.

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

$$\widehat{u}_t^2=\gamma_0+\overbrace{\gamma_2 t}^{\overline{[5]}}+\overbrace{\gamma_1 x_t+\gamma_3 x_t^2}^{\overline{[3]}}+\overbrace{\gamma_4 x_{t-1}^2+\gamma_5 Y_{t-1}^2}^{\overline{[4]}}+v_{2t}, \tag{32}$$

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 (30) in terms of assumptions [1]-[5]. More formally, the extra terms in (31) and (32) 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.

5.4 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 (31) and (32) 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 $\delta_4=\delta_5=0$ 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 pertaining to different model assumption, say

$$H_0(i),\ i=1,...,m, \text{ such that } H_0=\cup_{i-1}^m H_0(i),$$

and the overall F-test rejects H_0 when the smallest p-value, $p_i(\mathbf{x}_0)$ associated with each $H_0(i)$, 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:

$$p(\mathbf{X};\alpha)\backsim \mathsf{U}(0,1), \text{ for } \alpha\in(0,1),$$

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 significance level of the joint test for H_0 is $m\alpha$. Hence, a simplistic rule of thumb for controling the overall (joint) significance level at α is to use $\frac{\alpha}{m}$ for the

individual hypotheses $H_0(i)$, i=1, ..., m. That is, the rejection rule for the individual hypotheses should be:

$$\text{Reject } H_0(i) \text{ when } p_i(\mathbf{x}_0) < \tfrac{\alpha}{m}. \tag{33}$$

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 (31) and (32), two things play an important role in alleviating the problem of multiple testing. The first is that 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 M-S test**, but within that the tests for the individual assumptions can shed light on how to respecify the model. The second is that often the number of individual hypotheses being tested is $m \leq 3$, and one could use a rule, such as (33), to avoid over-rejection. 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. Moreover, over-rejection is not as serious a problem as in N-P testing, because the type I error is not the more serious of the two in M-S testing. The practitioner will have another opportunity to correct any over-rejections when respecifying to account for the particular departure(s) from the model assumptions. That is, the serious problem in M-S testing is not over-rejection but over-acceptance.

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.

5.5 Where do auxiliary regressions come from?

A strong case can be made that the best strategy to avoid 'erroneous' diagnoses, minimize the number of maintained assumptions and enhance the scope of the tests is to use *joint M-S testing*. As shown above, the model assumptions are usually interrelated, and thus testing them individually can give rise to misleading diagnoses. In the case of the LR model (table 15.4), there are natural groupings of the assumptions according to how their potential departures might change/modify the regression and skedastic functions of the original model. In addition to minimizing the error of misdiagnoses, the explicit estimation of the auxiliary regressions enables the modeler to view the statistical significance of each individual term. For instance, a practitioner can easily conceal the presence of first order autocorrelation in the residuals by using a Box-Pierce test with a high order p of lags.

The formal justification of joint M-S testing based on auxiliary regression stems from the conditional expectation property:

CE6. Regression function characterization. Consider a set of random variables defined on the probability space $(S, \mathcal{F}, \mathbb{P}(.))$ with bounded variance, including $\mathbf{Z}$:=$(y, \mathbf{X})$ (a $m \times 1$ vector) such that $E\left(|\mathbf{Z}|^2\right) < \infty$. In the context of the orthogonal decomposition used in chapter 7 to specify the statistical GM:

$$y = E(y|\sigma(\mathbf{X})) + u,$$

the orthogonality between $u=y-E(y|\sigma(\mathbf{X}))$ and any random variable with respect to $\sigma(\mathbf{X})$, which can take the form of any Borel function of $h(\mathbf{X})$ (Doob, 1953):

$$E\left([y-E(y|\sigma(\mathbf{X}))]\cdot h(\mathbf{X})\right)=0, \text{ for any Borel-function } h(\mathbf{X}). \tag{34}$$

This result can be extended to regression functions in the sense that the orthogonality:

$$E\left([y_t-g(\mathbf{X}_t)]\cdot h(\mathbf{X}_t)\right)=0, \text{ for all Borel-functions } h(\mathbf{X}_t),\ t\in\mathbb{N}, \tag{35}$$

holds if and only if: $g(\mathbf{X}_t)=E(y_t|\sigma(\mathbf{X}_t)),\ t\in\mathbb{N}$. For $u_t=y_t-E(y_t|\sigma(\mathbf{X}_t))$, the orthogonality takes the form:

$$E\left(u_t\cdot h(\mathbf{X}_t)\right)=0,\ t\in\mathbb{N}. \tag{36}$$

In light of the fact that u_t^r, $r=2,3,...$, define random variables whose mean exists, one can extend the above orthogonality to higher conditional moment functions. Of particular interest is the second, where $E(u_t^2|\sigma(\mathbf{X}_t))$:

$$E\left([u_t^2-g_2(\mathbf{X}_t)]h(\mathbf{X}_t)\right)=0,\ \text{for all Borel-functions } h(\mathbf{X}_t),\ t\in\mathbb{N}, \tag{37}$$

if and only if $g_2(\mathbf{X}_t)=E(u_t^2|\sigma(\mathbf{X}_t)),\ t\in\mathbb{N}$.

In the case of the LR model, the construction of the M-S tests will be based on seeking legitimate $D_t\subset\mathcal{F}$ for which the orthogonalities below might *not* hold:

$$\text{(i) } E\left([y_t-E(y_t|D_t)]\cdot h_1(D_t)\right)=0,\quad \text{(ii) } E\left([u_t^2-E(u_t^2|D_t)]\cdot h_2(D_t)\right)=0,\ t\in\mathbb{N}.$$

D_t is operationally legitimate for M-S testing purposes if D_t is a proper subset of the statistical universe of discourse $\mathcal{F}_Z:=\sigma\left(\mathbf{Z}_1,\mathbf{Z}_2,...,\mathbf{Z}_n\right)$. Of particular interest is the choice $\mathcal{D}_t=\sigma\left(\mathbf{X}_t,\mathbf{Z}_{t-1},\mathbf{Z}_{t-2},...,\mathbf{Z}_1\right)$ that would render $\{(u_t|\mathcal{D}_t),\ t\in\mathbb{N}\}$ a 2nd order MD process without imposing independence. Such potential non-orthogonalities can be framed in terms of using auxiliary regressions of the form:

$$u_t=\delta_1+\boldsymbol{\gamma}_1^\top\mathbf{h}_1(D_t)+v_{1t},\ u_t^2=\delta_2+\boldsymbol{\gamma}_2^\top\mathbf{h}_2(D_t)+v_{2t},\ t=1,2,...,n, \tag{38}$$

where $\mathbf{h}_r(D_t)$, $r=1,2$, denote vectors of different Borel functions relating to D_t chosen with a view to pick up different potential departures from the model assumptions. In a certain sense M-S testing based on (38) amounts to probing for departures from the process $\{(u_t|D_t),\ t\in\mathbb{N}\}$ being a Normal, Martingale Difference process.

Possible Borel functions of the original statistical information set $(\mathbf{Z}_t:=(y_t,\mathbf{X}_t),\ t=1,2,...,n)$ that can be used to define potential statistical information not accounted for by the original model, say $D_t=(\boldsymbol{\psi}_t,\mathbf{z}_{t-1},\mathbf{t})$:

$$\boldsymbol{\psi}_t:=(x_{it}\cdot x_{jt})_{i,j},\ i\geq j=2,...,k,\quad \mathbf{z}_{t-1}:=(y_{t-1},\mathbf{x}_{t-1}),\quad \mathbf{t}:=(t,t^2,..,t^p).$$

Conditioning on D_t will give rise to an alternative regression function:

$$E(y_t|D_t)=\alpha_0+\boldsymbol{\alpha}_1^\top\mathbf{x}_t+\boldsymbol{\alpha}_2^\top\boldsymbol{\psi}_t+\boldsymbol{\alpha}_3^\top\mathbf{z}_{t-1}+\boldsymbol{\delta}^\top\mathbf{t}. \tag{39}$$

As mentioned above, $(\boldsymbol{\psi}_t,\mathbf{z}_{t-1},\mathbf{t})$ are not 'omitted' but 'discounted' variables because they are already in the statistical information set: $(\mathbf{Z}_1,\mathbf{Z}_2,...,\mathbf{Z}_n)$. Comparing the original $E(y_t \mid \mathbf{X}_t{=}\mathbf{x}_t){=}\beta_0{+}\boldsymbol{\beta}_1^\top\mathbf{x}_t$ with the respecified model one can construct an auxiliary regression for testing departures from the assumptions [2], [4] and [5]:

$$\text{(GM1): } y_t{=}\beta_0{+}\boldsymbol{\beta}_1^\top\mathbf{x}_t+u_t,\ \text{(GM2): } y_t{=}\alpha_0{+}\boldsymbol{\alpha}_1^\top\mathbf{x}_t{+}\boldsymbol{\alpha}_2^\top\boldsymbol{\psi}_t{+}\boldsymbol{\alpha}_3^\top\mathbf{z}_{t-1}{+}\boldsymbol{\delta}^\top\mathbf{t}{+}v_{1t}. \tag{40}$$

Subtracting the estimated GM1, $y_t{=}\widehat{\beta}_0{+}\widehat{\boldsymbol{\beta}}_1^\top\mathbf{x}_t{+}\widehat{u}_t$, from GM2 yields an auxiliary regression in terms of the residuals of the original regression:

$$\widehat{u}_t{=}(\alpha_0{-}\widehat{\beta}_0){+}(\boldsymbol{\alpha}_1{-}\widehat{\boldsymbol{\beta}}_1)^\top\mathbf{x}_t{+}\boldsymbol{\alpha}_2^\top\boldsymbol{\psi}_t{+}\boldsymbol{\alpha}_3^\top\mathbf{z}_{t-1}{+}\boldsymbol{\delta}^\top\mathbf{t}{+}v_{1t}. \tag{41}$$

This auxiliary regression can be easily extended/modified to include higher order powers of $\mathbf{x}_t$, as well as higher order lags and/or using orthogonal polynomials in t.

IMPORTANT NOTE. In cases where the sample size n is not large enough, one can use the fitted values $\widehat{y}_t{=}\widehat{\beta}_0{+}\widehat{\boldsymbol{\beta}}_1^\top\mathbf{x}_t$ and the residuals $\widehat{u}_{t-i}$, in place of higher order functions of $\mathbf{x}_t$ and lags of $\mathbf{Z}_t$, respectively.

The F-type tests for the joint hypotheses:

$$H_0{:}\overbrace{\boldsymbol{\alpha}_2{=}\mathbf{0}}^{[2]},\ \overbrace{\boldsymbol{\alpha}_3{=}\mathbf{0}}^{[4]},\ \overbrace{\boldsymbol{\delta}{=}\mathbf{0}}^{[5]},\ \text{vs.}\quad H_1{:}\ \boldsymbol{\alpha}_2{\neq}\mathbf{0}\ \text{or}\ \boldsymbol{\alpha}_3{\neq}\mathbf{0}\ \text{or}\ \boldsymbol{\delta}{\neq}\mathbf{0}, \tag{42}$$

provides an M-S test for [2], [4] and [5], as they affect the regression function.

Analogous reasoning can be used to derive an auxiliary regression corresponding to the skedastic function $E\left(u_t^2 \mid \mathbf{X}_t{=}\mathbf{x}_t\right){=}\sigma^2$:

$$\widehat{u}_t^2{=}\gamma_0{+}\boldsymbol{\alpha}_2^\top\boldsymbol{\psi}_t{+}\boldsymbol{\alpha}_3^\top\mathbf{z}_{t-1}^2{+}\boldsymbol{\delta}^\top\mathbf{t}{+}v_{2t}, \tag{43}$$

where $\mathbf{z}_{t-1}^2$ denotes quadratic functions of $\mathbf{z}_{t-1}$, and the joint hypotheses are:

$$H_0{:}\overbrace{\boldsymbol{\alpha}_2{=}\mathbf{0}}^{[3]},\ \overbrace{\boldsymbol{\alpha}_3{=}\mathbf{0}}^{[4]},\ \overbrace{\boldsymbol{\delta}{=}\mathbf{0}}^{[5]},\ \text{vs.}\quad H_1{:}\ \boldsymbol{\alpha}_2{\neq}\mathbf{0}\ \text{or}\ \boldsymbol{\alpha}_3{\neq}\mathbf{0}\ \text{or}\ \boldsymbol{\delta}{\neq}\mathbf{0}. \tag{44}$$

The above auxiliary regressions are only indicative of factors that might be used in practice; several variations/extensions one might consider, include powers of $\widehat{y}_t$ and $\widehat{u}_{t-i}$, so long as the extra terms are functions of the original statistical information.

It is also interesting to note that the validity of assumptions [1]-[5] ensures that the error $\{(u_t|\ \mathbf{X}_t{=}\mathbf{x}_t),\ t{\in}\mathbb{N}\}$ is a Normal, MD process. The Normality assumption [1] is the only assumption that (41)-(43) do not test for. This is because the available M-S tests for [1] assume that the other assumptions are valid, rendering the results questionable when any of the other assumptions [2]-[5] is invalid. Hence, for a reliable test of Normality one should secure the validity of [2]-[5] beforehand.

5.6 Respecification

The PR perspective views respecification as a repeat of the specification facet with a view to select a new statistical model, using the tripartite partitioning of $\mathcal{P}(\mathbf{z})$. The aim is to select probabilistic assumptions that account for the chance regularities not accounted for by the original model. A discerning interpretation of a comprehensive M-S testing results and not one departure at a time, could guide the respecification by repartitioning $\mathcal{P}(\mathbf{z})$ with a view to specify a new statistical model that accounts for the statistical information the original model did not. A tentative new model is estimated and its own assumptions are tested thoroughly.

What is important to re-iterate is that in M-S testing the null is always H_0: $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is valid, but the alternative $\overline{H}_0$: $\mathcal{P}(\mathbf{z})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is non-operational. Hence, the modeler needs to select particularized alternatives or directions of departure that could never span $\mathcal{P}(\mathbf{z})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. This implies that when H_0 is rejected the specific alternative H_1, specifying a particularized subset of $\mathcal{P}(\mathbf{z})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is never an option for respecification purposes without further testing. In particular, the M-S testing based on auxiliary regressions, such as (31) and (32), could only provide information about departures from the original model assumptions. Significant coefficients indicate particular directions of departure from these assumptions. The auxiliary regressions do not provide a clear answer as to what the respecified model should look like. That is decided by the statistical adequacy of the respecified model; its assumptions are tested anew and shown to be valid. For instance, the significance of $\boldsymbol{\psi}_t$ in (31) and (32) provides only an indication that the assumptions [2]-[3] are invalid. The relevant non-linear and heteroskedastic functional forms, however, are unlikely to coincide with the polynomials used in (31)-(32). Their functional forms are determined by the joint distribution of $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$, and can be validated by M-S testing.

The PR perspective provides a broader and more coherent vantage point from that stemming from the error process $\{(u_t|\mathbf{X}_t=\mathbf{x}_t),\ t\in\mathbb{N}\}$. For instance, it views the LR model as specified in terms of the regression and skedastic functions of $D(y_t \mid \mathbf{X}_t;\boldsymbol{\theta})$:

$$E(y_t \mid \mathbf{X}_t=\mathbf{x}_t)=h(\mathbf{x}_t),\ Var(y_t \mid \mathbf{X}_t=\mathbf{x}_t)=g(\mathbf{x}_t),\ \mathbf{x}_t\in\mathbb{R}^k_X,$$

where the functional forms $h(.)$ and $g(.)$, and the relevant parameterization $\boldsymbol{\theta}$ stem from the joint distribution $D(y,\mathbf{X}_t;\boldsymbol{\varphi})$. From this perspective, departures from particular assumptions might relate to both functions. For instance, the move of retaining the Linearity and Normality assumptions, but adopting an arbitrary form of Heteroskedasticity (Greene, 2012), can easily give rise to an internally inconsistent set of probabilistic assumptions; see Spanos (1995b). The PR respecification when [1] and [3] are invalid is discussed in Spanos (1994).

5.6.1 Revisiting Yule's 'nonsense-correlations'

Example 15.8. The problem of 'spurious' associations, first noted by Pearson (1896), was high up in Yule's agenda during the first quarter of the 20th century, returning

to the 'spuriousness' problem in several papers. Yule (1926) represents the culmination of his efforts to unravel the puzzle of 'spurious' results by focusing on 'high' correlations between time series data.

He used data on the ratio of Church of England marriages to all marriages (x_t) and the mortality rate (y_t) over the period 1866-1911, to demonstrate that their estimated correlation $\widehat{\rho}_{xy}$=.9512 was both very high and statistically significant. He described this result as 'nonsense-correlation': "Now I suppose it is possible, given a little ingenuity and goodwill, to rationalize very nearly anything. And I can imagine some enthusiast arguing that the fall in the proportion of Church of England marriages is simply due to the Spread of Scientific Thinking since 1866, and the fall in mortality is also clearly to be ascribed to the Progress of Science; hence both variables are largely or mainly influenced by a common factor and consequently ought to be highly correlated. But most people would, I think, agree with me that the correlation is simply sheer nonsense; that it has no meaning whatever; that it is absurd to suppose that the two variables in question are in any sort of way, however indirect, causally related to one another." (p. 2)

Yule (1926) made a genuine attempt to link 'nonsense-correlations' to the data in question *not* being 'random series'. He could not establish a direct link between them, however, because he was missing two key components that were yet to be fully integrated into statistics. The first is the notion of a prespecified 'parametric statistical model', introduced by Fisher (1922a), comprising the probabilistic assumptions imposed on the data. The second is the theory of stochastic processes founded by Kolmogorov (1933). The vague notion of a 'random series' was formalized into a realization of an IID stochastic process, framed as a sequence of random variables 'indexed' by a particular ordering, such as time. To fully understand IID processes, however, one needs to appreciate how the assumptions of IID can be invalid in practice and how such departures can be formalized and modeled using various notions of dependence or/and heterogeneity; see Doob (1953).

Yule's reverse engineering. At the time there was no notion of a prespecified parametric statistical model, and thus Yule had to resort to 'reverse engineering':

"When we find that a theoretical formula applied to a particular case gives results which common sense judges to be incorrect, it is generally as well to examine the particular assumptions from which it was deduced, and see which of them are inapplicable to the case in point." (p. 4-5)

He proceeded to consider the formula for estimating the sample standard error and elicit the implicit probabilistic assumptions that render it a 'good' estimator of the distribution standard error. Let us emulate Yule's reverse engineering using the sample correlation coefficient, which is the focus of his paper:

$$\begin{aligned}
\widehat{Corr}(X_t,Y_t)&=\frac{\frac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})(X_t-\overline{X})}{\sqrt{\left[\frac{1}{n}\sum_{t=1}^{n}(X_t-\overline{X})^2\right]\left[\frac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})^2\right]}},\\
\overline{X}&=\tfrac{1}{n}\textstyle\sum_{t=1}^{n}X_t,\ \overline{Y}=\tfrac{1}{n}\sum_{t=1}^{n}Y_t,\ \widehat{Var}(X_t)=\tfrac{1}{n}\sum_{t=1}^{n}(X_t-\overline{X})^2,\\
\widehat{Var}(Y_t)&=\tfrac{1}{n}\textstyle\sum_{t=1}^{n}(Y_t-\overline{Y})^2,\ \widehat{Cov}(X_t,Y_t)=\tfrac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})(X_t-\overline{X}).
\end{aligned}\tag{45}$$

When is $\widehat{Corr}(X_t,Y_t)$ a 'good' estimator of the distribution (contemporaneous) correlation coefficient?

$$Corr(X_t,Y_t)=\frac{Cov(X_t,Y_t)}{\sqrt{Var(X_t)Var(Y_t)}}=\frac{E[(Y_t-E(Y_t)][(X_t-E(X_t)]}{\sqrt{E[(Y_t-E(Y_t)]^2E[(X_t-E(X_t)]^2}} \tag{46}$$

The first assumption implicit in these formulae is the *constancy* of the moments:

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

since otherwise the statistics:

$$\left(\overline{X},\overline{Y}\right) \text{ and } \left(\tfrac{1}{n}\textstyle\sum_{t=1}^{n}(X_t-\overline{X})^2,\ \tfrac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})^2,\tfrac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})(X_t-\overline{X})\right),$$

will not be 'good' estimators of:

$$(E(X_t),E(Y_t)) \text{ and } \left(E[(X_t-E(X_t)]^2,\ E[(Y_t-E(Y_t)]^2,\ E[(Y_t-E(Y_t)][(X_t-E(X_t)]\right)$$

respectively. which corresponds to a form of the *ID assumption*. Moreover, the formulae:

$$\widehat{Var(X_t)}=\tfrac{1}{n}\textstyle\sum_{t=1}^{n}(X_t-\overline{X})^2,\ \widehat{Var(Y_t)}=\tfrac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})^2,$$

implicitly assume *non-correlation* over $t\in\mathbb{N}$, since otherwise they should have included temporal covariance terms. For instance, a good estimator of $Var(X_t)=E[(X_t-E(X_t)]^2$, when $E(X_t)=\mu_2$ is constant, takes the form:

$$\widehat{Var}(X_t)=\tfrac{1}{n}\left[\textstyle\sum_{t=1}^{n}(X_t-\overline{X})^2+2\sum_{i=1}^{n}\sum_{j>i}^{n}(X_i-\overline{X})(X_j-\overline{X})\right]$$

More surprising is that Yule also sought to unveil the implicit **distributional assumption** "in order to reduce the formula to the very simple form given." (p. 5) The reason is that the sample moments are not always 'optimal' estimators of the distribution moments. For instance, the estimators in (45) will be 'optimal' under Normality, but they will be far from optimal if the distribution is Uniform; see Carlton (1946).

Table 15.7: The simple bivariate Normal model

Statistical GM:		$\mathbf{Z}_t=\boldsymbol{\mu}+\mathbf{u}_t$,	$t\in\mathbb{N}$,
[1]	Normal:	$\mathbf{Z}_t:=(y_t,X_t)^\top\backsim \mathsf{N}(.,.)$,	
[2]	Constant mean:	$E(\mathbf{Z}_t)=\boldsymbol{\mu}:=(\mu_1,\mu_2)^\top$,	
[3]	Constant covariance:	$Var(\mathbf{Z}_t)=\boldsymbol{\Sigma}:=[\sigma_{ij}]_{i,j=1}^2$,	
[4]	Independence:	$\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ is independent.	

The underlying bivariate Normal distribution takes the form:

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix}\backsim \mathsf{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}\right),$$

Given that under Normality the assumption of ID reduces to the constancy of the first two moments, and non-correlation coincides with *Independence* (I), one could make a case that the implicit parametric statistical model underlying the above formulae is the simple bivariate Normal in table 15.6. When any of the assumptions

[1]-[4] are invalid for the particular data $\mathbf{Z}_0$, the estimated correlation coefficient is likely to be 'spurious' (statistically untrustworthy). Granted, certain departures from particular assumptions, such as [2]-[4], are more serious than other departures. For instance assumptions [1] is the least problematic when the true distribution is non-Normal but bell-shape symmetric.

Recall that the statistical parameterization of the Linear Regression (LR) model in table 15.4 is:

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

and that means that the correlation (ρ_{12}) is a simple reparameterization of the the regression coefficient $(\beta_1=(\sigma_{12}/\sigma_{22}))$ since:

$$\rho_{12}=\beta_1(\frac{\sqrt{\sigma_{22}}}{\sqrt{\sigma_{11}}})=\left(\frac{\sigma_{12}}{\sigma_{22}}\right)(\frac{\sqrt{\sigma_{22}}}{\sqrt{\sigma_{11}}})=\frac{\sigma_{12}}{\sqrt{\sigma_{11}\cdot\sigma_{22}}}.$$

More formally, the NIID assumptions underlying the process $\{\mathbf{Z}_t:=(Y_t,X_t),\ t\in\mathbb{N}\}$ enable us to relate its parameters to both the bivariate distribution $f(x_t,y_t;\boldsymbol{\varphi})$ (in the context of which ρ_{12} is viewed) as well as the conditional distribution $f(y_t|x_t;\boldsymbol{\varphi}_1)$ in terms of which the LR model is specified. Formally the connection between the two distributions comes in the form of the following reduction $(\forall(x_t,y_t)\in\mathbb{R}^2)$:

$$f(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_n;\boldsymbol{\phi})\overset{\text{I}}{=}\prod_{t=1}^{n}f_t(x_t,y_t;\boldsymbol{\varphi}_t)\overset{\text{IID}}{=}\prod_{t=1}^{n}f(x_t,y_t;\boldsymbol{\varphi})=\prod_{t=1}^{n}f(y_t|x_t;\boldsymbol{\varphi}_1)f(x_t;\boldsymbol{\varphi}_2),\tag{47}$$

This suggests that one can pose the question of **statistical adequacy** in the context of the LR model, which yields:

$$y_t=\underset{(1.416)}{-10.847}+\underset{(.020)}{.419}x_t+\widehat{u}_t,\ R^2=.905,\ s=.664,\ n=46,\tag{48}$$

where the standard errors are reported in brackets below the coefficient estimates. Both coefficients (β_0,β_1) appear to be statistically significant since the t-ratios are:

$$\tau_0(\mathbf{z}_0)=\frac{10.847}{1.416}=7.660[.000],\ \tau_1(\mathbf{z}_0)=\frac{.419}{.020}=20.95[.000],$$

with the p-values given in square brackets. The implied correlation:

$$\widehat{\rho}_{12}=\widehat{\alpha}_1(\sqrt{\widehat{\sigma}_{22}}/\sqrt{\widehat{\sigma}_{11}})\rightarrow\widehat{\rho}_{12}=.419(\frac{4.854}{2.137})=.952[.000],$$

which coincides with the value in Yule (1926).

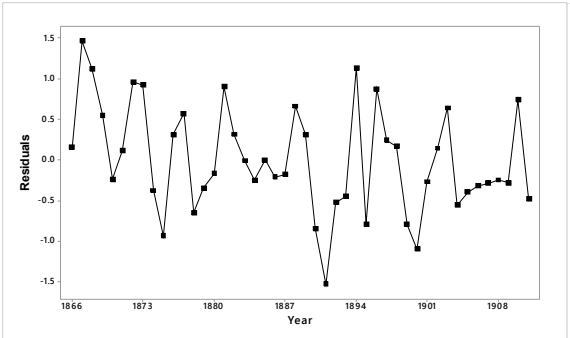


Fig. 15.9: The residuals from (48)

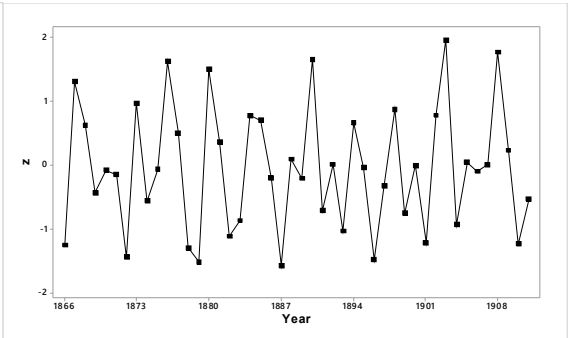


Fig. 15.10: t-plot of NIID data

However, a glance at the t-plot of the residuals (fig. 15.9) indicates that (48) is statistically misspecified since the residuals differs from that of a NIID realization (fig. 15.10) in so far as it exhibits distinct trends and cycles.

This suggests going back to the original data and taking a closer look at any departures from the IID assumptions.

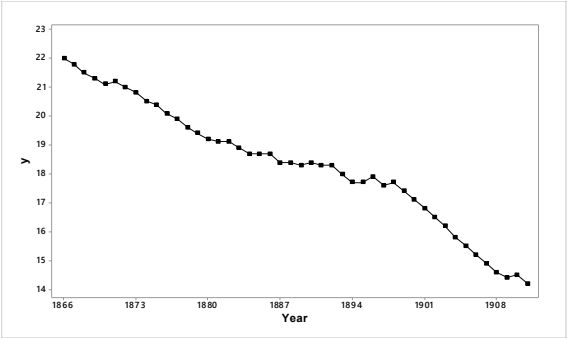


Fig. 15.11: t-plot of y_t-the mortality rate for the period 1866-1911

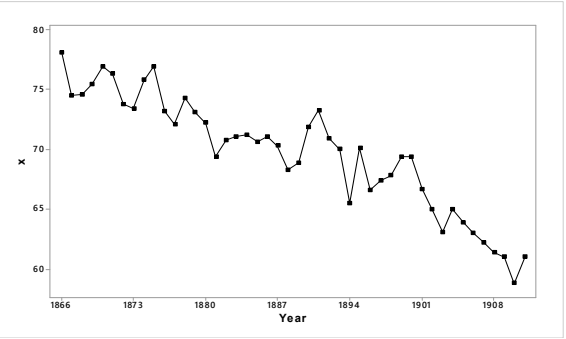


Fig. 15.12: t-plot of x_t-ratio of Church of England marriages to all marriages

A glance at the t-plots of Yule's (1926) data in figures 15.11-12 suggests that, to borrow his phrase: "Neither series, obviously, in the least resembles a random series" (aka IID); both data series exhibit mean t-heterogeneity (trending mean) and dependence (irregular cycles). To bring out the cycles more clearly one needs to subtract the trending means using a generic 2nd degree trend polynomial, as shown in figures 15.13–14, which suggest that assumptions [4]-[5] are likely to be invalid.

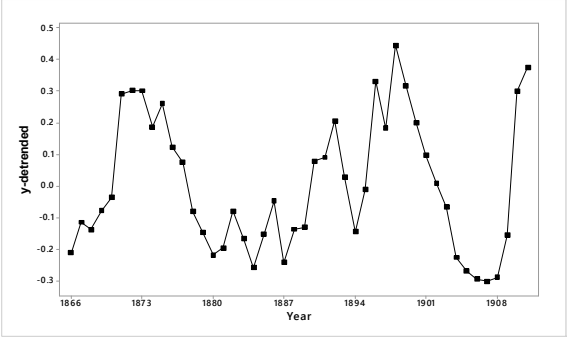


Fig. 15.13: t-plot of y_t-detrended

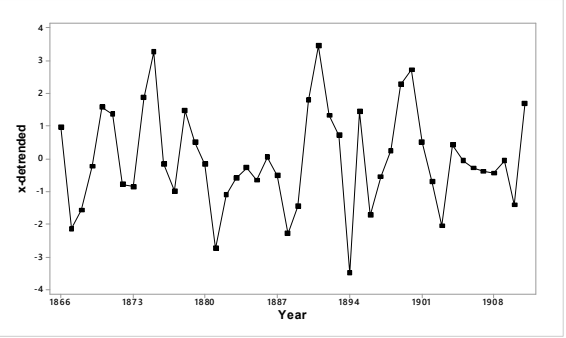


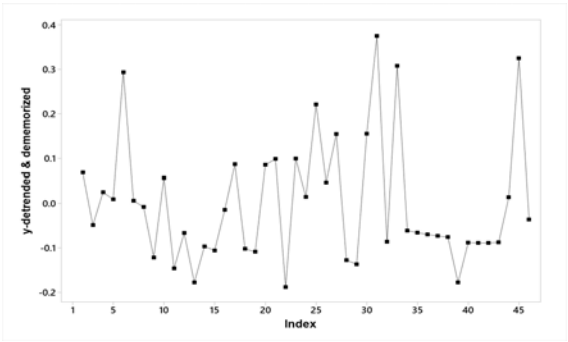
Fig. 15.14: t-plot of x_t-detrended

The 'best' auxiliary regressions used for detrending and dememorizing are:

$$x_t = \underset{(10.0)}{48.5} - \underset{(5.48)}{25.29}t - \underset{(2.25)}{5.15}t^2 + \underset{(.145)}{.293}x_{t-1} + \widehat{v}_{1t},\ R^2=.89,\ s_x=1.596,\ n=46,$$

$$y_t = \underset{(1.09)}{1.47} - \underset{(.977)}{1.678}t - \underset{(.198)}{.451}t^2 + \underset{(.060)}{.909}y_{t-1} + \widehat{v}_{2t},\ R^2=.996,\ s_y=.144,\ n=46.$$

These results show clearly that the formula in (45) will be totally inappropriate as a 'decent' estimator of any attempt to estimate the distribution correlation $Corr(X_t, Y_t)$ in (46), since it will be inconsistent!



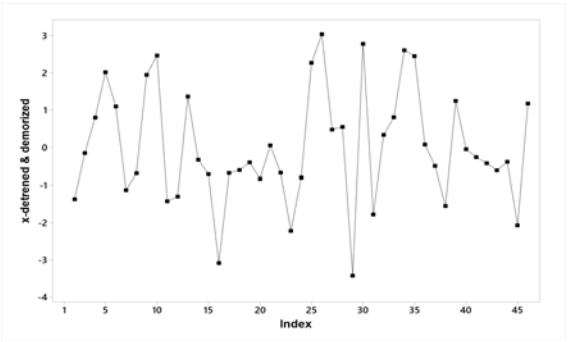


Fig. 15.15: y_t-detrended and dememorizedFig. 15.16: x_t-detrended and dememorized

What can one to estimate $Corr(X_t, Y_t)$ with a good estimator? In light of the fact that the stochastic process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ is both heterogenous and dependent, one should account for that before attempting to estimate $Corr(X_t, Y_t)$. Both departures from IID can be accounted for using the partial moments, after subtracting out the conditional means:

$$Corr(X_t, Y_t)=\frac{E[(Y_t-E(Y_t|\sigma(\mathbf{Y}_{t-1}))][(X_t-E(X_t|\sigma(\mathbf{X}_{t-1}))]}{\sqrt{E[(Y_t-E(Y_t)|\sigma(\mathbf{Y}_{t-1})]^2 E[(X_t-E(X_t|\sigma(\mathbf{X}_{t-1}))]^2}}$$

$\sigma(\mathbf{X}_{t-1})$ and $\sigma(\mathbf{Y}_{t-1})$, where $\mathbf{X}_{t-1}$:=$(X_{t-1}, X_{t-2}, ..., X_{t-m})$ and $\mathbf{Y}_{t-1}$:=$(Y_{t-1}, Y_{t-2}, ..., Y_{t-m})$ denote the $\sigma-$fields generated by the **past history** of the two process for some $m \geq 1$. In this case, this can be estimated using the residuals from the above two auxiliary regressions:

$$\widehat{Corr}(X_t, Y_t)=\frac{\frac{1}{n}\sum_{t=1}^{n}\widehat{v}_{1t}\cdot\widehat{v}_{2t}}{\sqrt{\left[\frac{1}{n}\sum_{t=1}^{n}\widehat{v}_{1t}^2\right]\left[\frac{1}{n}\sum_{t=1}^{n}\widehat{v}_{2t}^2\right]}}=.038[.804]$$

Respecification. The misspecifications detected in Yule's LR model suggest that a way to respecify it is to include trends and lags to account for the mean heterogeneity and dependence:

$$\begin{gathered} y_t=\underset{(1.27)}{1.14}-\underset{(.016)}{.0053}x_t-\underset{(.998)}{1.67}t-\underset{(.221)}{.406}t^2+\underset{(.076)}{.885}y_{t-1}+\underset{(.015)}{.006}x_{t-1}+\widehat{v}_t, \\ R^2=.996,\ s=.147,\ n=45 \end{gathered} \tag{49}$$

Note that the R^2=.996 reported above is misleading because it is based on $\sum_{t=1}^{n}(Y_t-\overline{Y})^2$ instead of $\sum_{t=1}^{n}(Y_t-\widehat{\gamma}_0-\widehat{\gamma}_1 t-\widehat{\gamma}_2 t^2)^2$; the correct value is R^2=.519. (49) turns out to be approximately statistically adequate, and it can be used to confirm that X_t and Y_t are contemporaneously $[Corr\,(X_t, Y_t)]$ and temporally $[Corr\,(X_{t-1}, Y_t)]$ uncorrelated since both coefficients of x_t and x_{t-2} are **statistically insignificant**. In contrast, both trends (t, t^2) and y_{t-1} are significant, confirming the above detected misspecifications. The results in (49) imply that a reliable estimate of $Corr\,(X_t, Y_t)$ is:

$$\widehat{\rho}_{xy}=(.0053)\left(\frac{s_{xt}}{s_{yt}}\right)=(.0053)\left(\frac{1.536}{.2119}\right)=.038[.804],$$

where (s_{xt}, s_{yt}) denote the standard errors of (X_t, Y_t) from figures 15.13-14.

These results indicate that the original Yule estimate of $Corr\,(X_t, Y_t)$:

$$\widehat{\rho}_{xy}=.952[.000],$$

constitutes *statistically untrustworthy evidence.*

5.6.2 Association reversal due to misspecification

The empirical example in this sub-section is based on cross-section data because statistical adequacy is often less well appreciated in that context.

Example 15.9 (Spanos, 2019). An economist considers the problem of evaluating the effect of education (x_k-years of schooling) on person's k income (y_k-income thousands of dollars). The data refer to n=100 working individuals from the age cohort of 30-40 years old selected using simple random sampling from a city's population. The estimated LR model (table 15.4) is:

$$y_k = \underset{(1.957)}{53.694} - \underset{(.147)}{.474}x_k + \widehat{u}_k,\ R^2=.096,\ s=3.307,\ n=100. \tag{50}$$

Both coefficients (β_0, β_1) appear to be statistically significant: $\tau_0(\mathbf{z}_0)$=27.437[.000] and $\tau_1(\mathbf{z}_0)$=3.224[.001]. The negative sign of the coefficient of x_t seems counterintuitive; additional years of education contribute negatively to one's income. He takes a closer look at the data and decides to estimate two separate LR models for male ($x_{1k},\ n_1$=50) and female ($x_{2k},\ n_2$=50):

$$\text{male:}\quad y_{1k} = \underset{(2.11)}{45.23} + \underset{(.180)}{.409}x_{1k} + \widehat{u}_{1k},\ R^2=.973,\ s=2.371,\ n_1=50, \tag{51}$$

$$\text{female:}\quad y_{2k} = \underset{(3.04)}{35.26} + \underset{(.206)}{.665}x_{2k} + \widehat{u}_{2k},\ R^2=.178,\ s=2.14,\ n_2=50. \tag{52}$$

The results in (51)-(52) indicate that (β_0, β_1) are statistically significant, and the coefficient of both x_{ik}'s, are positive. This clearly contradicts the negative sign in (50), which is usually interpreted as a case where a statistical association is reversed, often viewed as an example of Simpson's paradox; see Pearl (2009).

The estimation and testing results in (50)-(52) are trustworthy only when the model assumptions [1]-[5] (table 15.4) are valid for the particular data for each of the three estimated equations. When the original data are plotted using 'gender' (male/female) as the ordering of interest (see figures 15.14-15), it is clear that both data series exhibit a change in the mean for male (k=1, ..., 50) and female (k=51, ..., 100), suggesting that the estimated equation (50) based on the aggregated data is statistically misspecified since assumption [5] is clearly invalid.

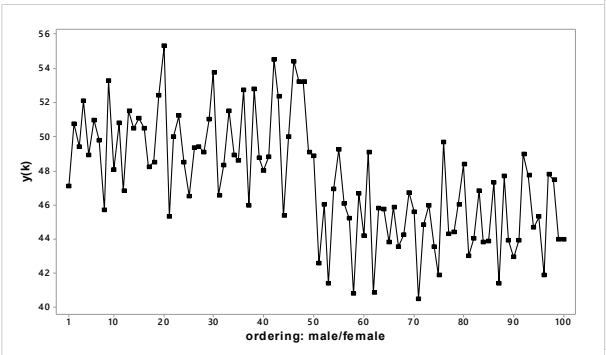


Fig. 15.15: t-plot of income (y_k)

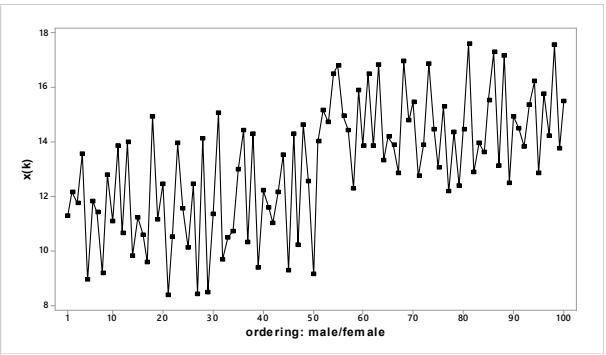


Fig. 15.16: t-plot of education (x_k)

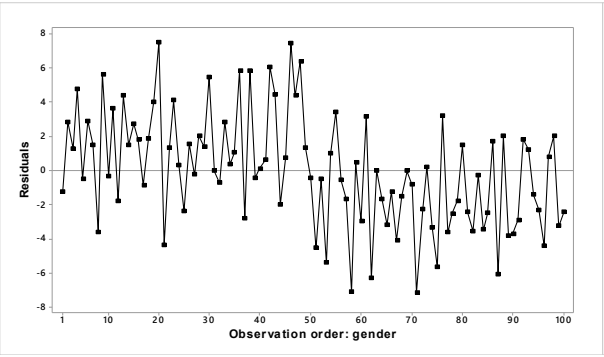


Fig. 15.17: Residuals from equation (50)

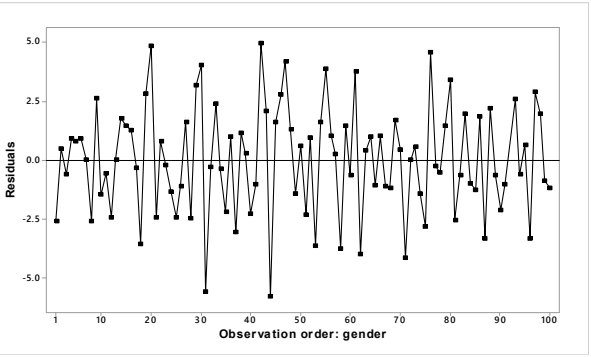


Fig. 15.18: Residuals from equation (54)

This misspecification is confirmed by the residuals of (50) in fig. 15.16 that exhibits a shfit in the sample mean; see chapter 5. The significance ($\tau_3(\mathbf{z}_0)$=10.55[.000]) of the coefficient of d_k:=(1, 1, ..., 1, 0, 0..., 0), 1-male, 0-female, confirms that using the auxiliary regression based on the residuals ($\widehat{u}_k$):

$$\widehat{u}_k = \underset{(2.00)}{-15.84} + \underset{(.135)}{.957}x_k + \underset{(.617)}{6.506}d_k,\ R^2{=}.53,\ s{=}2.272,\ n{=}100 \tag{53}$$

In light of (53), the statistical misspecification perspective explains the sign reversion between (50) and (51)-(52). Any claim that there is a 'reversal of association' is misplaced since the pooled data equation (50) is statistically misspecified, rendering the inference that the coefficient of x_k is negative and statistically significant is untrustworthy; an artifact stemming from invalid probabilistic assumptions imposed on data $\mathbf{Z}_0$. In addition, viewing 'gender' (d_k) as a missing explanatory variable is misleading because (i) the information pertaining to the ordering of potential interest is already in the original data $\mathbf{z}_0$ (figures 15.14–15), and (ii) 'gender' (d_k) is a deterministic ordering in this context; conditioning on d_k makes no probabilistic sense.

Respecification. A more pertinent explanation is that the modeler neglected by choosing to ignore (or did not plot the data) the mean heterogeneity in the data with respect to the ordering *gender* (d_k) when estimating (50). In practice such systematic information concerning heterogeneity could be modeled using dummy variables, shift functions, or/and trend polynomials in k; see chapter 5. The way to secure the reliability of inference is to respecify (50) with a view to account for the mean heterogeneity in the data (fig. 15.14–15), using the dummy variables d_k and $(1-d_k)$:

$$y_k = \underset{(3.13)}{34.8} + \underset{(3.71)}{10.5}d_k + \underset{(.170)}{.402}d_kx_k + \underset{(.211)}{.690}(1-d_k)x_k + \widehat{u}_k,\ R^2{=}.6,\ s{=}2.24,\ n{=}100 \tag{54}$$

The residuals from (54) (fig. 15.17) look like a realization of a NIID sample (chapter 5), indicating no departures from assumptions [1]-[5]; comprehensive M-S testing confirms that.

In conclusion, the sign reversal of the coefficient of x_k in (50) vs. (51)-(52) does not constitute an example of Simpson's paradox since there was never a statistically trustworthy association at the 'aggregate' data level to be reversed, due to the fact that t(50) is statistically misspecified.

6 An illustration of empirical modeling

6.1 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 $\{u_t,\ t{\in}\mathbb{N}\}$:

$$C_t{=}\alpha+\beta Y_t^D+u_t,\ u_t\backsim \mathsf{NIID}(0,\sigma^2),\ t{=}1,2,...,n. \tag{55}$$

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 (55) 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 15.10. Estimating (55) yields:

$$y_t{=}\underset{(16.930)}{-45.279}+\underset{(.007)}{.936}x_t+\widehat{u}_t,\ R^2{=}.997,\ s{=}49.422,\ n{=}52. \tag{56}$$

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

$$\tau_0(y){=}\tfrac{45.279}{16.930}{=}2.675[.004],\quad \tau_1(y){=}\tfrac{.936}{.007}{=}133.71[.000]. \tag{57}$$

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 (56) provide *evidence for* the AIH? Before the probabilistic assumptions of the underlying statistical model are tested, one cannot draw any trustworthy evidence from (56). If any of the model assumptions [1]-[5] (table 15.4) are invalid, statistical misspecification is likely render the *actual error probabilities* very different from their assumed *nominal* ones: the ones invoked in pronouncing the coefficients (α,β) significant. Moreover, the estimated model in (56) is also likely to be substantively inadequate; confounding variables, etc. An estimated model which is both statistically and substantively misspecified is typical when the modeler does not test the statistical model assumptions. Worse, one has no way to delineate the two: (i) is the theory (substantive information) false? or (ii) are the (implicit) inductive premises invalid for data $\mathbf{z}_0$?

Question (i) cannot even be posed unless the statistical adequacy of the underlying statistical model has been secured first because statistical specification errors are likely to undermine the prospect of reliably evaluating the relevant errors for primary inferences because actual and nominal error probabilities will be different. When the

underlying statistical model is statistically misspecified the above estimated coefficients, their standard errors, the t-ratios and the R^2 constitute *statistically meaningless artifacts*; see Spanos (1989b).

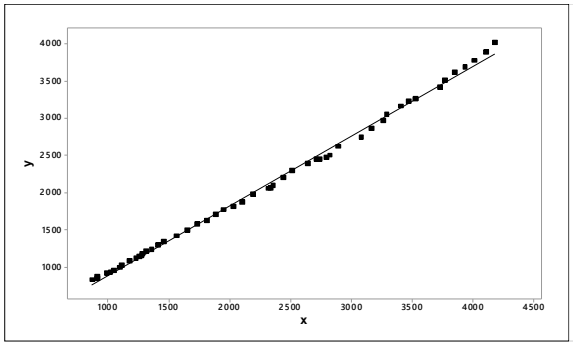


Fig. 15.19: Scatter-plot and fitted line

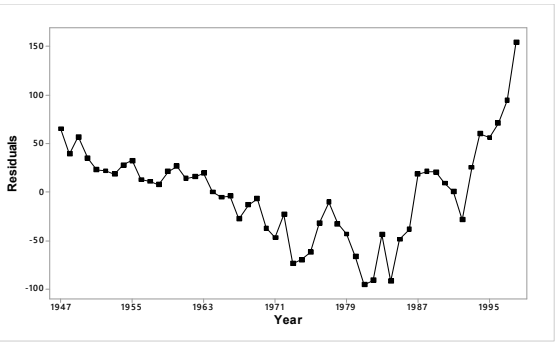


Fig. 15.20: Residuals form the fitted line

Strong hints about the serious statistical misspecifications associated with the estimated LR model in (56) 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:	SK=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]

6.2 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.

6.2.1 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.

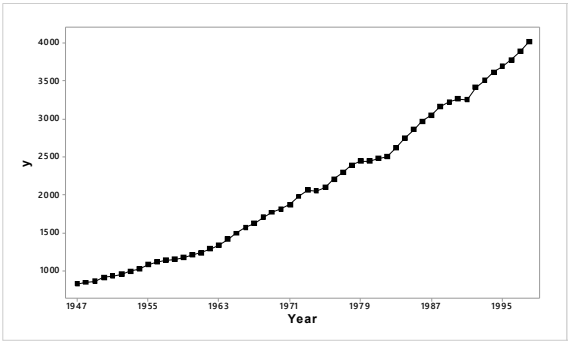


Fig. 15.21: t-plot of y_t

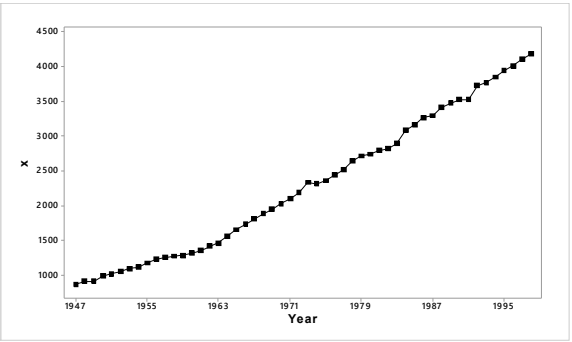


Fig. 15.22: t-plot of x_t

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

$$z_t{=}\delta_0 + \delta_1 t + \delta_2 t^2 + v_t,\ t = 1, 2, .., n, \tag{58}$$

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 (58) 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.

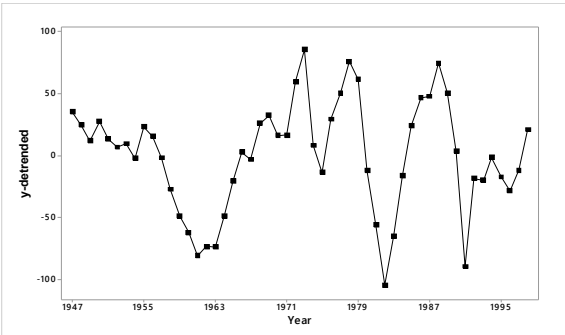


Fig. 15.23: t-plot of y_t-detrened

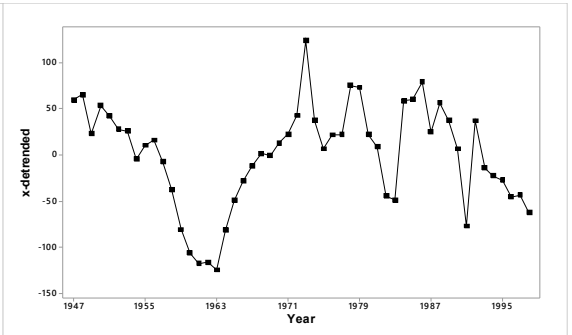


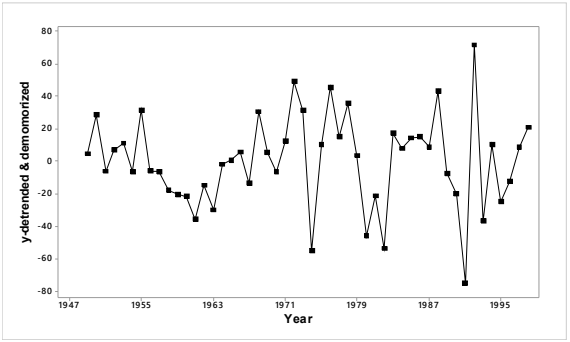
Fig. 15.24: t-plot of x_t-detrened

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

$$z_t{=}\gamma_0{+}\gamma_1 t{+}\gamma_2 t^2{+}\gamma_3 z_{t-1}{+}\gamma_4 z_{t-2}{+}\epsilon_t,\ t{=}1, 2, .., n, \tag{59}$$

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



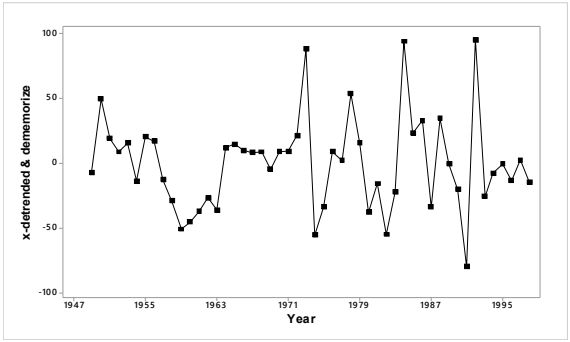


Fig. 15.25: y_t-detrended and dememorizedFig. 15.26: x_t-detrended and dememorized

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.

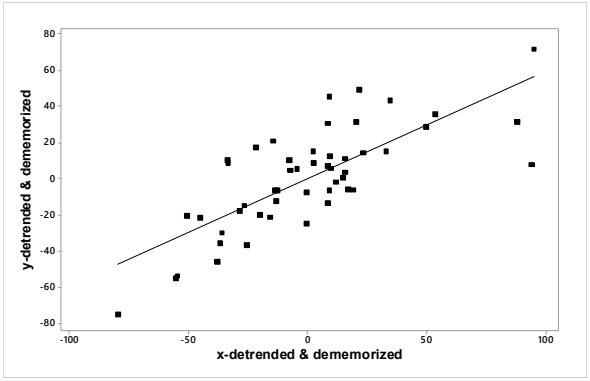


Fig. 15.27: Scatterplot of detrended and demomorized data

(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.

6.2.2 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:

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

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

$$\widehat{u}_t=\underset{(52.3)}{205.5}-\underset{(.082)}{.389}x_t+\underset{(2.61)}{5.37}t-\underset{(.0085)}{.03}\ x_t^2+\underset{(.128)}{.594}y_{t-1}-\underset{(.117)}{.384}x_{t-1}+\widehat{v}_t, \qquad (60)$$
$$R^2=.86,\ s=19.1,\ n=51$$

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

(γ_4=0): $\tau_4(\mathbf{y};45)=(\tfrac{.594}{.128})=4.641[.000015]$, ($\gamma_5$=0): $\tau_5(\mathbf{y};45)=(\tfrac{.384}{.117})=3.282[.0001]$

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:

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

$$(\widehat{u}_t/s)^2=\underset{(83.83)}{110.2}-\underset{(.043)}{.057}t-\underset{(.127)}{.252}x_t^2+\underset{(.174)}{.874}(\widehat{u}_{t-1}/s)^2+\widehat{v}_{2t}, \qquad (61)$$

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: $\overline{[5]}$ (δ_1=0): $\tau_1(\mathbf{y};45)=\left(\tfrac{.057}{.043}\right)=1.326[.194]$,

Heteroskedasticity: $\overline{[3]}$ δ_2=δ_3=0: $F(2,45)=\left(\tfrac{114.7-66.484}{66.484}\right)\left(\tfrac{45}{2}\right)=16.318[.000005]$,

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 $\tau_1(\mathbf{y};45)=(\tfrac{.042}{.014})=3.01[.004]$. This brings out the importance of joint M-S testing to avoid misdiagnosis!

In summary, the M-S testing based on the auxiliary regressions (60)-(61) 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!

6.2.3 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. In particular, let us estimate an extended regression equation with all but the **kitchen sink**, aiming to maximize the R^2:

$$\begin{gathered} y_t=\underset{(52.0)}{163.7}+\underset{(3.18)}{7.75}t-\underset{(.122)}{.159}t^2+\underset{(.105)}{.462}x_t+\underset{(.022)}{.057}x_t^2+\underset{(.130)}{.556}y_{t-1}-\underset{(.132)}{.302}x_{t-1}+\widehat{\varepsilon}_t, \\ R^2=.9997,\ s=18.957,\ n=51, \end{gathered} \tag{62}$$

Apart from the obvious fact that (62) 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.

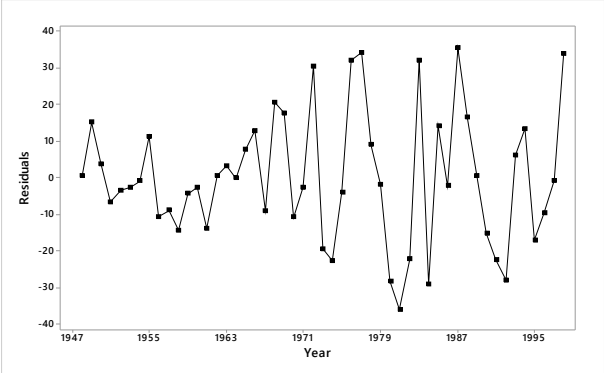


Fig. 15.28: Residuals from (62)

6.2.4 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.

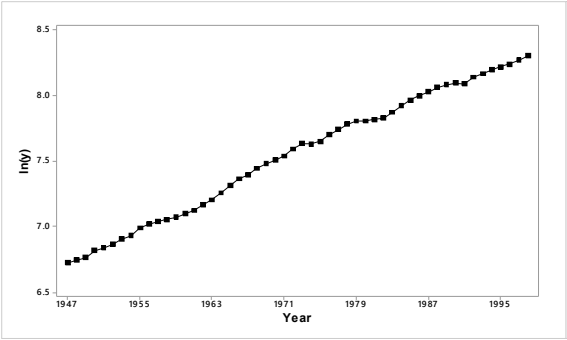


Fig. 15.29: t-plot of $\ln y_t$

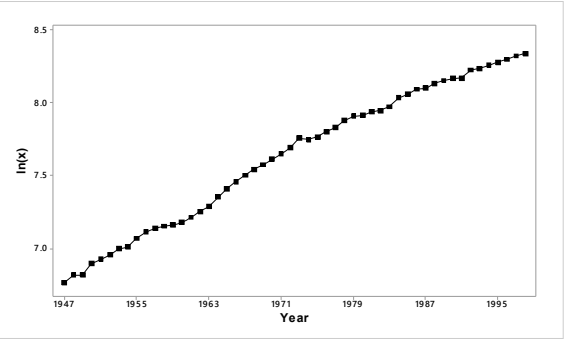


Fig. 15.30: t-plot of $\ln x_t$

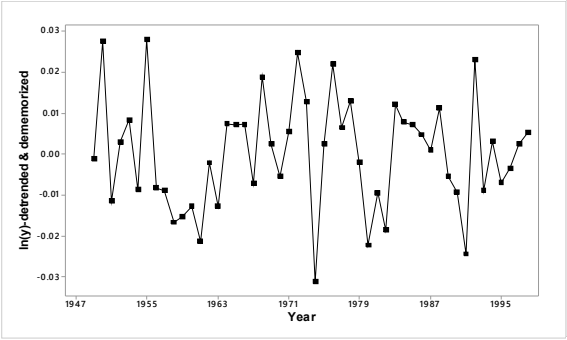


Fig. 15.31: t-plot of $\ln y_t$-detrended & dememorized

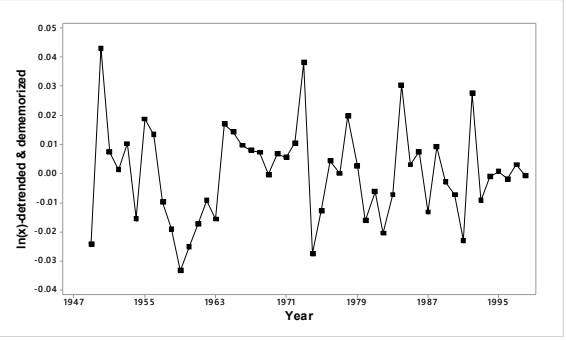


Fig. 15.32: t-plot of $\ln x_t$-detrended & dememorized

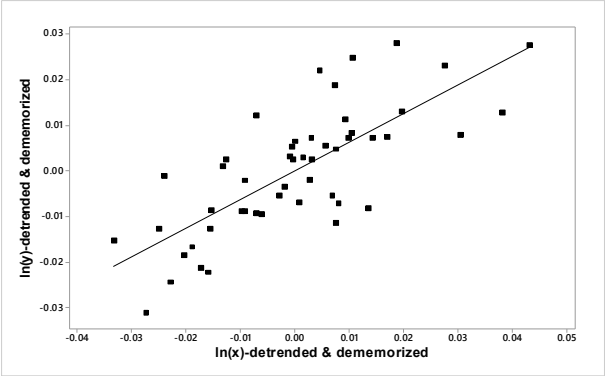


Fig. 15.33: Scatterplots of $(\ln x_t, \ln y_t)$ detrended and dememorized

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):

$$y_t=\delta_0+\delta_1 t+\alpha_1 x_t+\alpha_2 y_{t-1}+\alpha_2 x_{t-1}+u_t,\ t{\in}\mathbb{N}, \tag{63}$$

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 (63) yields:

$$\begin{gathered}\ln Y_t{=}\underset{(.272)}{.912}+\underset{(.001)}{.005}t+\underset{(.069)}{.708}\ln x_t+\underset{(.108)}{.565}\ln Y_{t-1}-\underset{(.097)}{.413}\ln x_{t-1}+\widehat{\varepsilon}_t,\\ R^2{=}.904\ [\text{reported: } R^2{=}.9997],\ s{=}.0085,\ n{=}51,\end{gathered} \tag{64}$$

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* (64) we can *infer* that the substantive model is clear false on the basis of this data.

6.2.5 What about substantive adequacy?

The statistical model (64) is statistically adequate but it does not constitute a substantive model. The trend in the estimated statistical model indicates substantive ignorance! It suggests that certain relevant (substantive) variables are missing.

Example 15.10 (continued). Returning to macro-economic theory one can pose the question of omitted relevant variables in the context of the statistically adequate model (64) and use the error reliability of the model to probe the significance of potential variables. For the sake of the discussion let us assume that the following variables might capture the relevant omitted effects: $X_{2t}-$ price level, $X_{3t}-$ consumer credit outstanding, $X_{4t}-$ short run interest rate. When these additional variables are introduced into the statistically adequate model (64), they render the trend insignificant by explaining the trending behavior, and preserve the statistical adequacy of the respecified DLR model has a statistical GM of generic form:

$$\ln y_t=\beta_0+\textstyle\sum_{j=1}^{4}(\beta_j \ln x_{jt}+\alpha_j \ln x_{jt-1})+\gamma_1 \ln y_{t-1}+u_t,\ t\in\mathbb{N}. \tag{65}$$

A detailed description of the procedure from (65), in the context of which the additional variables (x_{2t}, x_{3t}, x_{4t}) have rendered the trend term t redundant, to a new substantive model:

$$\begin{aligned}\Delta \ln y_t=&\underset{(.004)}{.004}-\underset{(.046)}{.199}(\ln y_{t-1}-\ln x_{1t-1})+\underset{(.056)}{.558}\Delta \ln x_{1t}-\underset{(.071)}{.173}\Delta \ln x_{2t}+\\ &+\underset{(.025)}{.081}\Delta \ln x_{3t}-\underset{(.011)}{.024}\Delta \ln x_{4t}+\widehat{\varepsilon}_t,\ R^2=.924,\ s=.0071,\ n=52.\end{aligned} \tag{66}$$

is beyond the scope of this chapter, but it involves imposing empirically validated restrictions with a view to find a parsimonious and substantively meaningful model. In particular, the procedure involves estimating the DLR model in (65) and securing its statistical adequacy using trenchant M-S testing for probing assumptions [1]-[5] (table 15.10), before any restrictions are imposed. In light of the fact that (65) has 11 statistical parameters and the substantive model in (66) has 7 structural parameters, going from (65) to (66) involves imposing 4 overidentifying restrictions. The overidentifying restrictions were tested using an F-type test (linear restrictions) and not rejected, and the estimated substantive model in (66) retained the statistical adequacy of (65). The latter ensures that any inferences based on (66), including forecasting and policy simulations will be statistically reliable; the relevant actual error probabilities closely approximate the nominal ones.

One might disagree with the choice of the variables (X_{2t}, X_{3t}, X_{4t}) to replace the generic trend, but the onus will be on the modeler questioning such a choice to provide a better explanation by proposing different variables that might achieve that. The statistically adequate model (64) provides a proper basis for such a discussion.

7 Summary and Conclusions

The problem of statistical misspecification arises when any assumptions invoked by a statistical inference procedure are invalid for the particular data $\mathbf{z}_0$. Departures from the invoked assumptions invalidate the sampling distribution of any statistic (estimator, test, predictor), and as a result the reliability of inference is often undermined. Mis-Specification (M-S) testing aims to assess the validity of the assumptions comprising a statistical model. Its usefulness is twofold: (i) it can alert a modeler to potential problems with unreliable inferences, and (ii) it can shed light on the nature of departures from the model assumptions.

The primary objective of empirical modeling is 'to learn from data $\mathbf{z}_0$' about observable phenomena of interest using a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ as the link between substantive information and systematic statistical information in data $\mathbf{z}_0$. Substantive subject matter information, codified in the form of a structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, plays an important role in demarcating and enhancing this learning from data when it does not belie the statistical information in $\mathbf{Z}_0$. Behind every structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ there is a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ which comprises the probabilistic assumptions imposed on one's data, and nests $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ via generic restrictions of the form $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$.

In an attempt to distinguish between the modeling and inference facets in statistical analysis the original Fisher's (1922a) framework is broaden with a view to bring out the different potential errors in the two facets and use strategies to safeguard against them. The statistical misspecification of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is a crucial error because it undermines the reliability of the inference procedures based on it. Relying on weak assumptions, combined with vague 'robustness' claims and heuristics invoking $n\to\infty$, will not circumvent this error in practice. Establishing the adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ calls for a thorough M-S testing, combined with a coherent respecification strategy that relies on changing the assumptions imposed on $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$. Joint M-S tests based on auxiliary regressions provide a most effective procedure to detect departures from the model assumptions. The traditional respecification of adopting the particular alternative H_1 used by the M-S test is fallacious.

Distinguishing between statistical and substantive inadequacy is crucial because a structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ will always be a crude approximation of the reality it aims to capture, but there is nothing inevitable about imposing invalid probabilistic assumptions on $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. When modeling with observational data, $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ will inevitably come up short in terms of substantive adequacy vis-a-vis the phenomenon of interest, but that does not preclude using a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ to answer reliably substantive questions of interest. These questions include whether (i) $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ belies the data, (ii) its overidentifying restrictions $\mathbf{G}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}$ are data-acceptable, as well as (iii) probing for omitted variables, confounding factors, systematic approximation errors, and other neglected aspects of the phenomenon of interest.

FINAL THOUGHT: the overwhelming majority of empirical results published in prestigious journals is untrustworthy, primarily because very few researchers **test the**

validity of their statistical models. Invoking vague robustness results, combined with wishful asymptotic (as $n \to \infty$) theorems based non-testable assumptions, will not secure the trustworthiness of one's empirical evidence. In contrast, securing statistical adequacy using trenchant M-S testing will address the problem of statistical misspecification, encourage a more informed implementation of inference procedures, ensure the reliability of the inference results and their warranted evidential interpretations. These will go a long way toward securing trustworthy evidence and attain the primary objective of empirical modeling and inference: learning from data about stochastic phenomena of interest.

Important concepts

Statistical misspecification, statistical adequacy, Mis-Specification (M-S) testing, default alternative in a M-S test, unreliable inference, untrustworthy evidence, facets of modeling: specification, estimation, M-S testing, respecification, nominal error probabilities, actual error probabilities, misspecification induced unreliability of inference, Probabilistic Reduction perspective, runs (up and down) test, circularity charge for M-S testing, infinite regress charge for M-S testing, Kolmogorov's test of Normality, omnibus (nonparametric) M-S tests, directional (parametric) M-S tests, M-S tests based on encompassing models, Skewness-Kurtosis test of Normality, the multiple hypotheses charge for M-S testing, joint M-S tests based on auxiliary regressions, multiple testing (comparisons), regression function characterization, Yule's nonsense correlations, traditional error-fixing strategies, Generalized Least Squares (GLS) estimator, Feasible GLS estimator, common factor restrictions, Heteroskedasticity-Consistent Standard Errors (HCSE), Autocorrelation-Consistent Standard Errors (ACSE).

Crucial distinctions

Mis-Specification (M-S) vs. Neyman-Pearson (N-P) testing, nominal vs. actual error probabilities, testing within vs. testing outside the statistical model, statistical vs. substantive adequacy, model respecification vs. error-fixing.

Essential ideas

- Statistical misspecification undermines the reliability of statistical inference and gives rise to untrustworthy evidence. Nonparametric statistics as well as Bayesian statistics are equally vulnerable to statistical misspecification. What often described as minor departures from model assumptions can have devastating effects on the reliability of inference.
- In the case of frequentist inference unreliability takes the form of biased and inconsistent estimators, as well as sizeable discrepancies between nominal (assumed) and actual error probabilities.
- Trenchant M-S testing is the most effective way to secure the reliability of inference. Joint M-S tests based on the residuals offer the most reliable way to probe for potential statistical misspecifications by taking into account the interrelationships among model assumptions.

- Yule's (1926) 'nonsense' correlations can be easily explained away on statistical misspecification grounds.
- The slogan "all models are wrong, but some are useful", provides a poor excuse to avoid validating one's statistical model by invoking a confusion between statistical and substantive inadequacy.
- Weak but non-testable probabilistic assumptions provide the most trusted way to untrustworthy evidence.
- Le Cam (1986a, p. xiv): "... limit theorems "as n tends to infinity" are logically devoid of content about what happens at any particular n."
- M-S testing differs from N-P testing in one crucial respect: the latter takes place within the boundaries of the prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, but the former takes place outside those boundaries. The null for an M-S is that the true distribution of the sample $f^*(\mathbf{x}) \in \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is valid and the default alternative is that $f^*(\mathbf{x}) \in [\mathcal{P}(\mathbf{x}) - \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$.
- When any of the assumptions of the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is rejected by the data, the only inference that follows is that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is misspecified. The next step should be to respecify it with a view to account for the systematic information in data $\mathbf{x}_0$ $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ did not. This calls seriously into question the traditional strategies based on piecemeal ad hoc 'error-fixing', that ignores the observable process $\{\mathbf{X}_t,\ t \in \mathbb{N}\}$ and focuses on the error term.
- Deterministic trends provide an generic way to account for certain forms of heterogeneity in the context of a statistical model. They are necessary to secure the reliability of inference. In the context of a substantive model, however, such terms represent ignorance and need to be replaced with proper explanatory variables.
- Be suspicious of any robustness claims made by practitioners who did not test the validity of their model assumptions. It is highly likely that on closer examination such claims fall apart. No practitioner can bluff his/her way out of dependence and heterogeneity type misspecifications, by invoking vague robustness results.

8 Appendix:Philosophical issues pertaining to M-S testing

8.1 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:

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

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:

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

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.

8.2 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:

$$\begin{array}{ll} \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}): & y_t = \beta_0 + \beta_1 x_t + u_t, \\ \mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z}): & y_t = \beta_0 + \beta_1 x_t + u_t,\ u_t = \rho u_{t-1} + \varepsilon_t. \end{array} \tag{67}$$

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*:

$$\ddot{\beta}_1 = \lambda \widehat{\beta}_1 + (1-\lambda)\widetilde{\beta}_1,\ \lambda = \begin{cases} 1, & \text{if } H_0 \text{ is accepted} \\ 0, & \text{if } H_0 \text{ is rejected;} \end{cases} \tag{68}$$

$\widehat{\beta}_1$ 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 (68), 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**.

8.3 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:

$$f(\mathbf{z};\boldsymbol{\theta}) = |J| \cdot f(\mathbf{s},\mathbf{r};\boldsymbol{\theta}) = |J| \cdot f(\mathbf{s};\boldsymbol{\theta}) \cdot f(\mathbf{r}), \tag{69}$$

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

$$\mathbf{Z} \rightarrow (\mathbf{S}(\mathbf{Z}),\ \mathbf{R}(\mathbf{Z})), \tag{70}$$

$\mathbf{R}(\mathbf{Z}){:=}(R_1,...,R_{n-m})$, 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 (69) 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), (69) holds with the minimal sufficient statistic being $\mathbf{S}{:=}(\overline{X}_n, s^2)$:

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

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 (70) takes the form:

$$(y_1, y_2, ..., y_n) \longleftrightarrow (\widehat{\boldsymbol{\beta}}, s^2, \widehat{v}_{k+2}, .., \widehat{v}_n), \tag{71}$$

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:

$$f(\mathbf{z};\boldsymbol{\theta}) = |J| \cdot f(\mathbf{s}, \mathbf{r};\boldsymbol{\theta}) = |J| \cdot f(\mathbf{s};\boldsymbol{\theta}) \cdot f(\mathbf{r}),\ \forall\,(\mathbf{s}, \mathbf{r}) \in \mathbb{R}_s^m \times \mathbb{R}_r^{n-m}, \tag{72}$$

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'.