SOAS Department of Economics

Redirect the Probability Approach in Econometrics Towards PAC Learning, Part II

Duo QIN

Working paper

No. 256

March 2023

The SOAS Department of Economics Working Paper Series is published electronically by SOAS University of London.

ISSN 1753 – 5816

Suggested citation
Qin, Duo, (2023), "Redirect the Probability Approach in Econometrics towards PAC Learning, Pt II", SOAS Department of Economics Working Paper No. 256, London: SOAS University of London.

Department of Economics
SOAS University of London
Thornhaugh Street, Russell Square, London WC1H 0XG, UK
Phone: + 44 (0)20 7898 4730
Fax: 020 7898 4759
E-mail: economics@soas.ac.uk
http://www.soas.ac.uk/economics/

Redirect the Probability Approach in Econometrics Towards PAC Learning, Part II

Duo Qin[*]

Department of Economics, SOAS, University of London, UK

March 2023

Abstract

Infiltration of machine learning (ML) methods into econometrics has remained relatively slow, compared with their extensive applications in many other disciplines. The bottleneck is traced to two key factors – a communal nescience of the theoretical foundation of ML and an outdated probability foundation. The present study ventures on an overhaul of the probability approach by Haavelmo (1944) in light of ML theories of learnibility, centred upon the notion of *probably approximately correct* (PAC) learning. The study argues for a reorientation of the probability approach towards assisting decision making for model learning and selection purposes. The second part of the study comprises two chapters.

Keywords: probability; machine learning; hypothesis testing; estimation, model generalisability.

JEL classification: C10, C18, B40.

Acknowledgements: I am greatly indebted to Xi-Yu Jiao, Ruben Lee, Xue-Lin Liu, Shan Lu, Lina Song, Sophie van Huellen, Qing-Chao Wang, Chris Watkins, Wei-Ying Zhang and also posthumously Olav Bjerkholt for invaluable discussion, encouragement, comments and suggestions I have received.

[*] Department of Economics, SOAS University of London. Russell Square, London WC1H 0XG, UK. Email: dq1@soas.ac.uk

Drafts of the first three chapters are in SOAS Economics Working Paper No. 249 (2022)

4. Roles of Hypothesis Testing and Economic Model Formulation

Hypothesis testing is the means to verify economic theories against data in Haavelmo's monograph. This ideological pillar gives rise to the motto of 'measurement with theory' (Koopmans, 1947), a motto which has played a key role in shaping how econometrics is organised and consolidated under Haavelmo's probability approach. Estimation and hypothesis testing, therefore, form the core of present-day econometrics.

However, there is an unsurmountable gap between the reality of positive economic issues in general and the setting of hypothesis testing in classical statistics. Hypothesis testing was originally set up for analysing randomised data samples. Sample collection is guided by and targeted at single parameters. Their expected measurements are formulated as null hypotheses. The null is taken as a *plausible* norm and its rejection is expected to be a surprise. Such surprises are regarded as random events with small probabilities and classified into the alternative hypothesis. This dichotomous formulation facilitates discriminative decision making by means of probability-based thresholds. But more importantly, hypothesis testing serves primarily as a *confirmative* tool oriented towards the null. Under the motto of 'measurement with theory', the economic theories to be verified are treated as the null. However, the economic reality differs substantially from such a stringent formulation. This reality is best conceived as what Valiant described as a theoryless situation and causal claims that compose economic theories are effectively common sense based. As such, theoretical models deductively formulated *a priori* are not data-congruous in general, and uncertainty in the formulation is not ignorable. Consequently, verification or confirmation of causal claims in economics cannot be simply equated to hypothesis testing of given parameters in theoretical models. The claims are generally too complex to be formulated uniquely into single parameters to be tested following the confirmative statistical approach. The task entails judgments based on comprehensive analyses of empirical results from elaborate modelling experiments. Mechanical applications of hypothesis testing in such a context are bound to lead to mistakes. The innate flexibility and uncertainty of models apparently offers a fertile ground for inventions of hypothesis tests and their applications. But the hypotheses in these tests are generally statistical features being derived from fitting prior chosen models to data. The features are mostly indirect symptoms of the formulated causal claims of interests, making the test results inconclusive for theory verification. Meanwhile, researchers are likely to be induced into a confirmative bias trap by the restrictive framework of hypothesis testing.

Section 4.1 illustrates, by means of two basic cases, how confirmative bias can lead researchers astray in both applied and theoretical research. The first case is the use of endogeneity tests. The presence of endogeneity bias in econometric models is widely taught as a norm. As a result, endogeneity tests are performed routinely in econometric studies with the aim of verifying specific economic causal claims. Unfortunately, as the object of the test is a non-uniquely defined and model-dependent statistical feature, this implicitly encourages applied modellers to actively practise *p*-hacking research. The second case is the design of tests on single time-series variables for stochastic grouping purposes. The design is based on the critical assumption described in the last chapter, namely that stochastic data generation processes of single variables are the very foundation of jointly distributed variables. Since many economic variables exhibit dynamic features dissimilar to what typical stationary processes exhibit, the topic has become part of the frontiers in econometric research. To verify the initially assumed stand, test designers are led to extending single-time-series based models to accommodate recalcitrant dynamics of observed single variables. From the viewpoint of philosophy of science, their endeavours fall into 'HARKing' (hypothesizing after the results are known) and defective accommodationism, since what they are pursuing are merely data symptoms rather than the underlying economic causes.

Use of hypothesis testing is bound to lead towards diagnostics in aid of model learning and selection, once the learning step is embraced in econometrics. Section 4.2 turns to the diagnostic role of hypothesis testing and highlights the limitations of the hypothesis testing framework. Its confirmative stance determines how limited these tests are when modellers face various exploratory learning tasks, especially when considerable uncertainty is present about what alternative hypotheses should be. This explains why information-based statistics, cross validation techniques and feature selection procedures occupy a relatively central place in machine learning. It is only after data-congruent models have been learnt that confirmative testing on parameters in the models becomes empirically applicable.

In his "*The Testing of Hypotheses*" chapter, Haavelmo devotes a full section to the 'Meaning of the Phrase "Formulate theories by looking at data"' (section 17). Interestingly, he concedes that such formulation belongs to 'inductive science' and 'involves risk of failures, which are beyond our control'. This concession indicates that Haavelmo's conception of hypothesis testing is broadly in accordance with the idea of statistical learning described in section 4.2. The limits of history, however, confine him to the narrow framework of classical statistics, because 'discussion of "right" or "wrong" in connection with such empirical, inductive processes would take us into metaphysics' (p.82). This statement is now clearly outdated with the advent of machine learning. As shown in section 4.2, faithful formulation of economic causal claims is indeed a far more complex issue than what the hypothesis testing apparatus can tackle. Section 4.3 expounds this point further through two examples, one on the problem of selection bias in microeconometrics and the other time-series aggregate index making in macroeconometrics. These examples demonstrate how easily mechanical use of the hypothesis testing framework is prone to Type III error, and how crucial it is that the search for best possible data-congruous models is done via careful infusion of logical reasoning and inductive learning. Consequently, the formulation task should be placed at the core of econometric research.

4.1 Confirmative Bias: Pitfalls in Blind Adoption of Hypothesis-Testing Framework

The statistical tools advocated by Haavelmo are primarily for testing the statistical significance of single parameters of theoretical interest. His initial focus is on how compatible these tools are for individual parameters in SEMs, as these models are deemed the most fundamental type. His derivation of the OLS inconsistency in SEMs, now widely referred to as endogeneity bias, has set the core econometric research path on estimation and testing of *a priori* defined structural parameters. Therefore, the first case to be considered here is endogeneity tests on single parameters.

Case one: endogeneity tests

Although SEMs should be highly multivariate in general, Haavelmo's derivation of endogeneity bias is based on a bivariate static SEM (1943, 1944: ch. 5), which is an extreme oversimplification of reality:

$$(4.1.1) \qquad \begin{array}{l} y = \beta_1 x + \varepsilon_1 \\ x = \beta_2 y + \varepsilon_2 \end{array}$$

It is easy to derive from (4.1.1) that $cov(x\varepsilon_1) \neq 0$ and $cov(y\varepsilon_2) \neq 0$ when x and y are assumed to be jointly distributed random variables. The correlations make the OLS inconsistent for (4.1.1) because of its single-equation based optimisation criterion. For example, the OLS of the upper equation in (4.1.1) amounts to:

$$(4.1.2) \qquad y = \beta_{yx} x + \varepsilon_y, \qquad \text{with } cov\left(x\varepsilon_y\right) = 0.$$

The inconsistency implies $\beta_{yx} \neq \beta_1$. The inequality is referred to as 'simultaneity bias'. Since the explanatory variable, x, in (4.1.2) is also an explained variable in (4.1), the bias becomes widely known as endogeneity bias.

Nowadays, concerns over endogeneity bias are widely seen in the context of single-equation models. The instrumental variable (IV) method is prescribed to circumvent the bias in this context (see section 5.1 for further discussion). The IV route effectively refutes (4.1.2) and replaces it with the following conditional model involving a set of IVs, denoted by V:

$$(4.1.3) \quad y = \beta_{yx^V} x^V + \varepsilon_y^V, \text{ with } x^V = V\hat{\gamma}_{xV}, \hat{\gamma}_{xV} = (V'V)^{-1}V'x, \text{ and } cov\left(x^V \varepsilon_y^V\right) = 0.$$

(4.1.2) and (4.1.3) are two non-nested rival conditional models under the premise of $x \not\approx x^V$. In Haavelmo's classification, the prior belief in the presence of endogeneity bias amounts to the assertion that x is an observable variable with non-negligible measurement errors whereas x^V is the theoretical variable. This assertion underpins endogeneity tests. The null of the tests, H_0, is posed as: $plim\left(\beta_{yx^V} - \beta_{yx}\right) = 0$, versus the alternative, H_1: $plim\left(\beta_{yx^V} - \beta_{yx}\right) \neq 0$. Arguably the most popular endogeneity test is the Hausman test.

Now, attend to the way that endogeneity tests are formed, as it deviates from the norm of hypothesis testing. What economists expect to be widely confirmed, i.e. $\beta_{yx} \neq \beta_1$, is formulated in the alternative instead of the null. The formulation thus makes these tests look like diagnostic tests, which are discussed in the next section. But what is more important to note is that the option of (4.1.3) against (4.1.2) guided by endogeneity tests is in conflict with the criterion of optimising model fit. It is commonly known that the mean squared errors of ε_y^V are usually larger than those of ε_y, leaving (4.1.3) with inferior fit as compared to (4.1.2). Meanwhile, by involving IVs, (4.1.3) is a more complicated model by construction than (4.1.2). According to the principles of PAC learnability, rejection of a relatively simple model in favour of more complicated ones should be based on the evidence of the latter enjoying higher degrees of fit and generalisability than the former.[1] The conflict results from the fact that endogeneity tests actually derive their discriminative power from single-parameter significance tests. This becomes evident from the following model:

$$(4.1.4) \qquad y = \beta_{yxx^V}x + \beta_{yx^Vx}x^V + \varepsilon_y^{V'}$$

A significance test, namely $\beta_{yx^Vx} \neq 0$, is now the equivalent of the endogeneity test of (4.1.3) against (4.1.2).[2] Given the relative ease to find $x^V \not\approx x$, it is no wonder that inferior models in terms of fit are frequently selected. There is a widely used rhetoric to defend the choice of (4.1.3) against (4.1.2) despite the latter enjoying a better fit: Optimal model prediction is irrelevant for empirical studies whose tasks are explicitly set to verify causal claims of specific interest. The single-parameter-based significance test in (4.1.4) is a precise embodiment of this rhetoric. Its firmly confirmative biased position makes endogeneity tests fundamentally different from diagnostic tests for function learning purposes.

The essence of a significance test on $\beta_{yx^Vx} \neq 0$ in (4.1.4) plus a diagnostic veneer helps explain why endogeneity tests and IV estimation has won such widespread popularity among applied economists when they are engaged in confirmative driven causal investigations. Since IVs are, by definition, outside of the variable set of an *a priori* formulated model, it is usually not difficult to select, via experiments with various IV combinations, a desirable IV set whose correlation with the model variable set enables IV estimates of the structural parameters to meet the prior expectation. The IV route thus spares them from the pain of a thorough model learning and selection. However, such concerted efforts to search for desirable *p*-values are dubbed '*p*-hacking' in applied statistics, and use of the practice solely for the sake of achieving journal publications is described as 'star wars'.[3]

One telling case of IV-induced *p*-hacking is an investigation by Brodeur *et al* (2020). This paper aims at evaluating the intensity of *p*-hacking effects on the topic of PEMs. The evaluation is carried out on 13,440 hypothesis-testing results of PEMs reported in 308 papers, which are selected from 2015 publications in 25 top-ranking journals. These results are obtained by means of four methods: randomized control trials, the IV, the difference-in-difference method and regression discontinuity design. Brodeur *et al* (2020) find that the IV method is one of the most effective for *p*-hacking. This outcome is actually predictable from (4.1.4). So long as non-unique choices of IVs provide modellers with adequate 'wiggle room', it should be relatively easy to find certain x^V which satisfy $x \not\approx x^V$ and meet the criterion of $\beta_{yx^Vx} \neq 0$. Obviously, *p*-hacking activities tarnish the credibility of the subsequent statistical results. Growth of the practice of *p*-hacking reflects an overreliance among applied modellers on *p*-values as *the* measure of uncertainty relevant to testing the hypotheses of interest. An extensive exposition of the incapacity of *p*-values as exclusive uncertainty measures can be found in the 2019 special issue of American Statisticians entitled

[1] One should not confuse the IV case with overfitting. Overfitting describes the situation where although model fit keeps rising with model complexity within training samples, model prediction deteriorates outside training samples. In the IV case, model fit worsens when (4.1.3) is compared with (4.1.2), irrespective of whether it is inside or outside training samples.

[2] This expedient route is described in various textbooks, e.g. see Berndt (1991: ch. 8).

[3] Brodeur *et al* (2016) come up with this description because of the widely established practice in journals to use different numbers of '*' to indicate different degrees of statistical significance, e.g. * indicates 5% and ** 1%. As for *p*-hacking, see Hirschauer *et al* (2016) for further discussion.

‘Statistical Inference in the 21st Century: A World Beyond p < 0.05’.[4] Among other things, the issue warns modellers against interpreting p-values in hypothesis testing of single causes as the empirical risk, as those values do not reflect any model uncertainty.

Case two. Single time-series variable tests for classification purposes

This research topic has primarily been led by development of unit-root tests, as described in the context of marginal-distribution-based time-series models in section 3.1. The topic derives its significance mainly from (a) a widespread conviction of the critical assumption, described at the beginning of Chapter 3, that data generation processes of single variables are the very foundation of jointly distributed variables; (b) observations of slow dynamic features in many time-series variables, features which apparently exceed characterisation of stationary processes; (c) the unavailability of hypothesis testing techniques for nonstationary variables in classical statistics.

The initial research is based on AR models. Take the AR(1) model (3.1.2) in Chapter 3 for example. Unit-root tests for that model are formulated as H_0: $\rho = 1$ versus H_1: $|\rho| < 1$, such as the Dicky-Fuller (DF) test, arguably the most popularly used unit-root test in practice. The formulation reflects a common expectation that the presence of unit roots should be the norm, namely that economic variables to be tested should be nonstationary by and large. The formulation, however, poses serious technical challenges for theoretical econometricians as it falls outside the apparatus of classical hypothesis testing. It turns out that more complex mathematics is required for tackling non-stationarity than is needed for stationarity, in order to try and minimise the size distortion and maximise the power of the tests by the standard of optimal test designs.[5] Meanwhile, recalcitrant dynamics observed in reality compels augmentation of single time-series variable models underlying the tests. In the augmented DF test, for instance, (3.1.2) is augmented into an AR(n):

(4.1.5) $$y_t = \sum_{i=1}^{n} \rho_i y_{t-i} + u_t$$

Unsurprisingly, the AR(n) model class is still inadequate in practice. One generalisation route is to introduce a deterministic component, μ_t, into the model class, e.g.:

(4.1.6) $$y_t = \mu_t + z_t,\ \ z_t \sim AR(n)$$

A great deal of flexibility is built in (4.1.6) via different formulations of μ_t. One popular version is: $\mu_t = d_t'\delta$, where d_t' denotes a vector of *a priori* given deterministic variables, and δ the corresponding parameter vector. For instance, a time trend, t, can be included in d_t' to accommodate the proposition that the variable to be tested evolves with time in a deterministic way; and when the variable is known to have experienced ‘structural’ breaks, dummy variables representing the breaks can be added into d_t'. To account for the latter phenomenon more realistically than the dummy specification, the breaks can also be represented stochastically, say by a Bernoulli process, π_t, in the following formulation: $\mu_t = \mu_{t-1} + \pi_t \varepsilon_t$, $\varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$. Clearly, different formulations evoke different optimal test designs. Furthermore, in the case that μ_t is composed of unknown parameters, such as δ, their optimal estimation has to be considered in the test designs. The combined consideration

[4] Applications of significance tests are a common practice in so many disciplines that this topic has attracted wide attention, e.g. see a detailed analysis by Mayo (2018) from the perspective of the philosophy of science.

[5] For detailed surveys and reviews on this topic, see Part IV, Volume 1, Palgrave Handbook of Econometrics, edited by Mills and Patterson (2006).

scales up mathematical complexity. It is no wonder that the research topic has earned the state-of-art reputation.

Contrary to the impressive theoretical achievements of single time-series variable tests, their empirical efficacy remains rather limited. For most economists, the rigorous classification of single variables into different types of stochastic processes bears too little relevance to their direct research concerns. Outlook differences between the test developers and economists are particularly telling with regard to structural breaks. While the breaks are regarded as sporadic contaminations to stochastic processes following the general form of (4.1.6), what really concerns economists are model forecasting failures (breaks) in the wake of unexpected shocks, and also how to make it possible to predict upcoming foreseeable shocks and their possible impacts. Behaviour based 'structural' models in economics exclude the single time-series variable model type. In fact, observations that dynamic patterns of single variables are approximately characterizable by certain stochastic processes do not warrant the deduction that these variables are intrinsically generated by those processes. It is an epistemic pitfall to equate observable data features with data generation mechanisms. When the object of hypothesis testing is on superficial symptoms rather than causes, the resulting tests cannot have much substantial efficacy in practice. What the tests can do is merely to reinforce the confirmative bias, namely the belief that economic data are fundamentally generated by single-variable DGPs. The research path of model augmentation from (3.1.2) towards (4.1.6) falls into the defective methodology, referred to sometimes as accommodationism.[6] A distinct feature of the methodology is to produce models which overfit, because overfitting models facilitate confirmations of hypotheses which are otherwise rejected. When research is motivated solely by *a priori* resolve to confirm a certain hypothesis being the norm, it is usually not difficult to find models which are flexible enough to render the expected outcome. The research practice is also referred to as 'HARKing', a term coined by Kerr (1998) for the situation when *post hoc* hypotheses are reported in research as if they were *a priori* formulated. Kerr has also summed up a list of costly consequences of HARKing activities in his reflections on scientific research methods under the logical empiricist approach.

It should be noted that the practice of HARKing should not be deemed undesirable *per se*. The fact that the tendency to Hark is widespread indicates how complex the investigative process of societal causal relationship is in an open world. The costly consequences of HARKing activities are effectively manifestation how the above process cannot be confined within the classical hypothesis testing framework.[7] The incompatibility of the framework with social science research issues is caused, according to Deming (1975), by confusion of two different types of research questions involving statistical analyses – 'enumerative' and 'analytic'. An enumerative research question is definable within a known universe/frame, such that data samples pertinent to the question are randomly collectable under controlled experiments. Furthermore, post-research targetable actions depend vitally on the estimation results of the statistical analysis. Social science research questions, on the other hand, are generally analytic by nature. Such questions usually concern certain causal systems and are

[6] See Hitchcock and Sober (2004).

[7] See Prosperi *et al* (2019).

posed with the intention to improve their future performance. The post-research targetable actions depend on conditional predictions of the causal systems under uncertain circumstances. Additional constraints are thus needed to clarify the uncertainty before hypothesis testing is applicable.[8] Under the circumstances, HARKing is effectively to try and accommodate the stringent hypothesis-testing setting to reality. Sadly, such an attempt is often made at the expense of neglecting the need for statistical prediction in an uncertain universe, a crucial task in analytical studies. Further discussion on this point is in Chapter 6.

Positive economic research is essentially data-analytic. Tackling data-analytic questions entails designing and selecting models which not only best fit the desired purposes but also are data-congruent. As argued in the previous chapters, *a posteriori* model learning is a prerequisite for that task. The design of effective procedures of data-aided model selection is essential for the success of such learning. Since the selection involves setting statistical criteria, their assessment evokes the use of hypothesis-testing tools. It should be stressed that these tools do not serve to verify subject-matter based causal claims. Rather, they are used to check if various model-derived statistical features, which are formulated into relevant hypotheses, meet with the expected statistical criteria for model selection. It is based on this distinction that these hypotheses are often referred to as auxiliary hypotheses, and the resulting tests as diagnostic tests.[9]

4.2 Diagnostic Testing and Model Selection

Tests on model residual distributions against the classical assumption, namely that the residuals are identically independent distributed (iid), are the most basic and popular type of diagnostic tests. The assumption is embodied in a number of null hypotheses in those tests, for example, the residuals are serially uncorrelated, homoskedastic and fit into a normal distribution. Apart from residual distributional features, there are two other areas where diagnostic tests are often targeted. One is to test for the constancy of parameter estimates, such as the Chow test and the CUSUM (cumulated sum of the residuals) test. These tests are essential for model-based prediction. The other area is to assess the relative performativity of different models. For example, tests on parameter zero-value restrictions against omitted variables in nested models of different sizes, and likelihood-ratio based tests to discriminate between non-nested models, especially theoretically rival models. Diagnostic tests targeting non-nested models have evolved into quite a sizable group, and are commonly known as encompassing tests.[10] The statistical features of the residuals form the ultimate testing object, despite the differently formulated hypotheses in parameter constancy tests or encompassing tests.

By convention, the null hypotheses of diagnostic tests represent the desired data features of the models under design. Rejections of the null indicate incompatibility of model against data. Therefore, diagnostic tests are also referred to as mis-specification tests, so as to

[8] See Hahn and Meeker (1993).

[9] Caution is needed when ‘auxiliary’ is used to classify hypotheses. All statistical hypotheses predicated on econometric models are auxiliary, from the standpoint of the hypothetical, economic causal claims, as long as the claims are not directly verifiable without models.

[10] Cox (1961, 1962) is acknowledged as pioneer of non-nested tests; see Pesaran (1987), Mizon and Richard (1986) for early development of encompassing tests.

differentiate them from specification tests, which are significance tests of structural parameters. Since statistical properties of the residuals are the focus of diagnostic tests, subsequent model revisions are rationalised as an error-statistical approach.[11] The approach effectively promises a route towards appropriate HARKing. In econometric practice, rejections of the null are prevalent because *a priori* formulated models are rarely data-congruous. The question of how to prescribe appropriate remedial actions after such diagnoses remains a major weakness in research. Under the assumption that structural models are *a priori* known to be correct globally, the textbook strategy is to prescribe optimal estimator choices as the core route, following Haavelmo's lead. Take tests of residual autocorrelation for example. Residual autocorrelation is commonly diagnosed from static models applied to time-series and/or dynamic panel data. It is not difficult to prove that the OLS, the default estimator, is inefficient in the presence of residual autocorrelation, and that the inefficiency can be overcome by the generalised method of moments (GMM) estimators which incorporate the autocorrelation as weights (see the second case in section 5.1). From the viewpoint of model revision, the efficiency gain in these GMM estimators comes essentially from an implicit model augmentation – augmenting static models into dynamic ones through the incorporation of a residual autocorrelation component. Hence, the practice of estimator-centred prescriptions amounts to implicitly accommodationist HARKing. A clear manifestation of this is the case of DSGE models discussed in section 3.2 of the last chapter. In contrast, the incorporation of 'structural' breaks into the AR models in the second case of the previous section amounts to explicit HARKing. It is debatable if either way of HARKing is appropriate.

An important question thus arises: What are the pitfalls in prescriptions following the alternative when the null has been rejected? Reflection on the question leads to two interrelated key factors: the limitation of the confirmative hypothesis-testing setting and the nature of derivative residuals. Most of the diagnostic tests follow the standard format of dichotomous hypothesis tests, where the null is expected to be the norm and its rejection is a small chance/probability event. In econometric reality, however, rejections in diagnostic tests are so common that they serve merely as a formal rejection of the expectation. Consequently, model revisions are a must whether done explicitly or implicitly. However, since the direct objects of testing are the symptomatic features of the model residuals, it is beyond the capacity of diagnostic tests to clearly specify model mis-specification causes. This incapacity implies that it is indeterminate whether the null rejections from different diagnostic tests are due to the same cause or different causes. For example, an economic crisis can be reflected in rejections of homoscedasticity and normality of the residuals as well as parameter inconstancy when the model does not take the crisis into explicit consideration. It is obviously erroneous to adopt and incorporate every single alternative literally in the treatment prescription under the circumstances. Consequently, rigorous model selection procedures under a systematic approach are called for. The LSE dynamic specification approach is a good example of such an endeavour. To avoid the above pitfalls, applied researchers are advised to start modelling from dynamically the most general model class, so that the basic requirement of the null being the norm is most likely to be satisfied. Once the limitation of

[11] See Spanos (2010, 2018) for more discussion.

diagnostic hypothesis testing is avoided in the first place, it becomes possible to carry out stepwise model reduction, with the help of diagnostic tests, to select the most parsimonious models. The LSE general-to-specific approach is essentially the result of summing up all the dynamic mis-specification problems, which commonly occur when *a priori* conventionally formulated models are fitted against time-series data. In comparison, it is a far more challenging task in micro-econometric research to sum up all the commonly seen mis-specification problems, especially in view of the ever-growing sample sizes and data varieties. In the absence of general model classes which would satisfy the requirement of the null being the norm, it is no wonder that diagnostic tests have not been used much in micro-econometric practice.

The limitations of diagnostic tests become increasingly apparent when multiple plausible alternative attributes are specifiable. Large-scale hypothesis testing tools are needed under such circumstances, tools which take into consideration the appropriate allocation of p-values corresponding to those alternative attributes.[12] But when the number of possible attributes is *a priori* unknown, the set of alternative attributes cannot be closed, p-value based diagnostic tests are devoid of their base. This explains why discriminative tests based on thresholds other than p-values are used widely during the function estimation process in machine learning. In short, exploratory research tasks entail far more versatile tools than what the confirmative statistical toolbox can offer.

Let us now reflect on the role of diagnostic tests during the process of model design and selection, from the perspective of machine learning theory. Specifically, the discussion is anchored in the context of two equations, (2.1.2) and (3.3.1) from the previous two chapters. The anchor enables us to deconstruct model selection problems into the following two aspects:

Q1. Which variables in the two candidate sets, $\{x_i \cdots x_n\}$ and $\{z_1 \cdots z_m\}$, of (2.1.2) regularly play a significant explanatory role for the target variable, y?

Q2. Of those selected input variables in (3.3.1), what functional forms are best representative of causally interpretable relationships between them and the target variable?

In order to tackle each question, multiple decision criteria and rules ought to be considered in the learning algorithm. A systematic model selection strategy is essential to ensure that the selected criteria and rules can be executed efficiently, either in a sequential or iterative manner. Above all, the algorithm design should follow the principle of SRM described in Chapter 2. SRM delivers accuracy and generalisability, two basic model selection rules. Accuracy is further embodied in ERM while generalisability is captured in the notions of parsimony and stability. Given a model class for a certain analytical question, the primary goal is to select models which neither underfit nor overfit data within the class. To realise this goal, we need to divide data into a training set and a validation set: $D = D_{train} \cup D_{validation}$ with $D_{train} \cap D_{validation} = \emptyset$, so that the goal can be translated into a search for the best possible bias-variance or bias-complexity trade-off. The criterion of ERM is clearly inadequate for the search and prone to selecting models which overfit, as it aims only at

[12] See Efron and Hastie (2016: ch. 15) for further discussion.

minimising the training errors, ε_{train}. To balance ERM with parsimony, a set of statistics are devised to penalise model fit by model size relative to sample size. The adjusted R^2 is arguably the most popularly used statistic in econometric practice.[13] A group of popularly used statistics is known as information criteria, including the Akaike information criterion (AIC), the Bayesian information criterion (BIC) known also as Schwartz criterion, and the Hannan and Quinn criterion (HQ). Their differences are the result of different formulations of penalty measures for model complexity. Theoretically, the differences reflect different, sometimes contentious, viewpoints about which rules should be the most appropriate. For example, AIC is motivated by the rule of predictive accuracy or asymptotic efficiency, whereas BIC is based on asymptotic consistency. In practice, BIC may guide modellers to more parsimonious models than what AIC indicates, because the weight of penalty on model complexity is heavier in BIC than in AIC, namely $dln(n)$ versus $2d$ (where d is model size and n sample size). Hence, AIC can be more sensitive in signalling underfitting cases than BIC, and vice versa for overfitting cases.[14] Information criteria are easily applicable to distribution-free function learning situations, as they are free of p-value based thresholds. Moreover, they can be combined into regularised loss functions to facilitate model-size constrained estimation (see section 5.2 of the next chapter for further discussion).

The diagnostic tools discussed in the preceding paragraph are primarily in response to Q1. As for Q2, the key standard or criterion is economic explicability, or more broadly, subject-matter potency. Specifically, this criterion demands for intelligent parameter designs in (3.3.1). It was pointed out in section 3.3 of the last chapter that meaningful parameter designs, referred to as feature learning, cannot be accomplished *a priori*. Diagnostic tests obviously play an important role in feature learning. Clearly, confirmations of diagnostic tests on model residuals are a prerequisite for conducting statistical inference on individual parameters. Moreover, tests on the constancy of parameter estimates are an indispensable part of feature learning. After all, parameter constancy is a key manifestation of structural invariance, and hence generalisability, of models. When models are fed by time-series data, recursive and/or rolling estimation can be exploited to construct sequences of parameter constancy tests, particularly for situations where sample sizes are rather limited. In the event of cross-section-data-based models, adequately designed cross-validation experiments are essential for testing parameter constancy and model stability.[15] Finally, various encompassing tests form another set of useful tools for model selection, complementary to information criteria.

It is clear from the above discussion that statistical tests function primarily as a discriminative toolbox for learning algorithm designs, and that it is premature to conduct specification tests before best possible generalisable models are successfully learnt. Correspondingly, the primary function of probability is also discriminative. Its function to

[13] A mirror statistic to the adjusted R^2 is Mallows' C_p. But this statistic is rarely used in econometrics.

[14] For further discussion on different choices of penalty measures in information criteria, see Kadane and Lazar (2004), Ding *et al* (2018). For detailed discussion on the AIC-BIC debates and related issues, see Yang (2005), and also Dziak *et al* (2020).

[15] For a comprehensive survey of cross validation methods, see Arlot and Celisse (2010).

assist statistical inference in specification tests is secondary, since these tests are predicated on having found interpretable parameters of well learnt models.

For most data-analytical questions, prior knowledge can merely offer a rough and inconclusive set of possible causal factors. The task of factor selection entails appropriate sequencing of the possible variables and features to be filtered. There are two sequencing procedures in machine learning: forward selection and backward elimination.[16] The latter procedure effectively underpins the LSE general-to-specific dynamic specification approach.[17] The backward selection procedure is mostly applicable for cases where the number of variables to be filtered is limited, and there are notable correlations between them, for example, in the case of lagged variable selection in time-series modelling research. When there are enormously many variables to be filtered and evidence of causally substitutive effects among them is weak, forward selection is the viable route. When modellers are confronted with the compound task of variable selection and input feature design, algorithms allowing for iterative experiments are usually needed.

4.3 "Formulate Theories by Looking at the Data": Modelling via Data Deconstruction

Section 17 in Haavelmo's chapter on The Testing of Hypotheses is entitled 'The meaning of the phrase "to formulate theories by looking at the data"'. Here, Haavelmo emphasises the importance of '"data" in the broad sense of empirical knowledge' (p.81) in formulating theories. Such a formulation is effectively an acquiescence of HARKing activities and the limited applicability of the hypothesis-testing framework. In machine learning, nowadays, the process of model formulation through 'looking at data' has been largely computerised, and the model learning process from an open and theoryless environment has been proven far more complicated than formulation of null hypotheses in confirmative statistics. A key contributor to the complexity is the need to substantiate theoretical claims through inductive learning, or the need to 'ground' and 'robustify' these claims in Valiant's parlance (2000). This need requires researchers to take into careful consideration both the theoretical claims and data-related specific circumstances during model formulation.[18] The formulation is, after all, a higher-level research issue than statistical testing.

In the last section, the discussion about model design and selection following the machine learning approach is centred on two model classes, (2.1.2) and (3.3.1). Both are broad functional representations of many applied issues. The discussion is thus too general to convey adequately the aforesaid complexity. Two examples are given below to rectify this inadequacy.

Case one: Causal inference of supply-side wage effects using incomplete data samples

Consider household surveys, where the wage rate entry is missing for those who are unemployed. This poses the question of whether models regressed with subsamples excluding

[16] For more discussion on these procedures, see Hastie *et al* (2009: ch. 3), Shalev-Shwartz and Ben-David (2014: ch. 25).

[17] This is clearly embodied in the designs of the software applications, PcGive and PcGet, see Hendry and Krolzig (2003), Hendry and Doornik (2014).

[18] For further discussion on the importance of engaging vernacular knowledge in applied economic analyses, see Swann (2006).

those missing observations are valid for inference. In textbook econometrics, the case is framed as a truncated or censored data complication, which induces selection bias, i.e. inconsistency, in the OLS based regression, because the corresponding model is formulated as a limited dependent variable model.[19] This is, unfortunately, a misconceived approach. Any assessment of estimator consistency assumes known population characteristics. But these characteristics are unknown in the present case as they hinge on counterfactual speculation. The limited dependent variable model representation is far too simplistic to support such speculation. Hence, *a posteriori* predictive estimates of those characteristics are needed, which requires, in the first place, inference of counterfactual labour supply behaviour of those unemployed at the individual level. From-specific-to-specific inferential issues are classified as transduction in machine learning. Let us go into more details of how the case should be formulated based on transductive learning.

Suppose that a household survey sample, $X = \{h, w, Z\}$, adequately represents a certain target population. In the sample, h denotes hours of work, w hourly wage rate, and Z an input variable set for both h and w ($Z = Z_h \cup Z_w$, and $Z_h \cap Z_w \neq \emptyset$). For respondent i who reports $h_i = 0$, w_i is missing. Construct a missing data indicator: l with $l_i = 0$ if w_i is missing and $l_i = 1$ if w_i is observed (the subscript of individual respondents is mostly omitted hereafter for simplicity). Variable l is commonly referred to as labour force participation (LFP). Subdivide the sample by l: $X = \{X_0, X_1\}$, where X_1, known also as the complete-record sample, is much larger than X_0, but the size of X_0 is not trivial. Write the labour supply model for the wage effects using X_1 as:

(4.3.1) $$h_1 = f_{h_1}\left(w_1, Z_{h_1}; \boldsymbol{\beta}_1\right) + \varepsilon_{h_1}$$

where the parameter corresponding to w_1 in set, $\boldsymbol{\beta}_1$, is the focus of inference. The research question can now be phrased as whether (4.3.1) is valid for inference concerning the group classified by l_0, had the whole group decided to work. Note that the question designates h_0 as counterfactually missing instead of being zeros as the survey has recorded.

A textbook claim for group l_0 to make the none-LFP decision is that potential wages, w_0, are non-market clearing. The claim implies that the missing wage data pattern is not random. The implication fuels the derivation of selection bias, namely inconsistency of the OLS estimates of $\boldsymbol{\beta}_1$ for the inference task. The pivotal role of the missing wage data pattern places transductive learning of plausible w_0 a first priority. A principled approach to tackle the learning task is Ruben's multiple imputation (MI) analysis (1987). What MI essentially does is to run stochastic predictions of individuals in w_0 via exploiting data similarities between Z_1 and Z_0.[20] In MI analysis, data missing mechanisms are categorised into three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Write a general human capital model as:

(4.3.2) $$w_1 = f_{w_1}\left(Z_{w_1}; \boldsymbol{\alpha}_1\right) + \varepsilon_{w_1}$$

[19] For more detailed descriptions, see Maddala (1983), Cameron and Trivedi (2005: ch. 16), and Kennedy (2008: ch. 17).

[20] See Carpenter and Kenward (2013) for a detailed description on MI analysis and related techniques.

The three types of mechanisms can be denoted as: (a) MCAR: $Pr(l_i|w_0, Z_w) = \Pr(l_i)$, (b) MAR: $Pr(l_i|w_0, Z_w) = \Pr(l_i|w_1, Z_w)$ and (c) MNAR: $Pr(l_i|w_0, Z_w) \neq \Pr(l_i|w_1, Z_w)$. The above non-random claim is formalised in the last type. MNAR cases can be simulated on the basis of MI results, which are attainable using (4.3.2) under the MAR assumption. Such simulation is known as sensitivity analysis. Specifically, denoted imputed w_0 under MAR as w_0^{MAR} and concatenate it with w_1 to form $w_{0+1}^{MAR} = \begin{pmatrix} w_0^{MAR} \\ w_1 \end{pmatrix}$. Since the following the LFP model holds:

(4.3.3) $$l = f_l(Z, w_{0+1}^{MAR}; \boldsymbol{\gamma}^{MAR}) + \varepsilon_l = f_l(Z; \boldsymbol{\gamma}^{MAR}) + \varepsilon_l \quad \text{with } \gamma_w^{MAR} = 0,$$

we can impose systematic shifts on w_0^{MAR} to produce w_{0+1}^{MNAR} such that (4.3.3) becomes:

(4.3.3') $$l = f_l(Z, w_{0+1}^{MNAR}; \boldsymbol{\gamma}^{MNAR}) + \varepsilon_l \quad \text{with } \gamma_w^{MNAR} \neq 0.$$

Simulations with different plausible shifts guided by prior knowledge provide us with a range of w_0^{MNAR} scenarios.

Once w_0^{MAR} and w_0^{MNAR} are obtained, we are in a position to impute their corresponding counterfactual h_0 using (4.3.1), because h_0 is missing in monotone with w_0,[21] and MAR according to prior economic reasoning, i.e. $Pr(l_i|h_0, w, Z_h) = \Pr(l_i|h_1, w, Z_h)$. The MI analysis also tells us that OLS regressions with missing response variables under MAR are not inconsistent when complete-record samples are used, and that inconsistency arises only with MNAR response variables, provided that the regression models in application are correctly specified.[22] Accordingly, selection bias should be irrelevant for $\boldsymbol{\beta}_1$ in (4.3.1) and should be relevant only for $\boldsymbol{\alpha}_1$ in (4.3.2) under MNAR.

However, the prerequisite of known correct models is untenable in practice. From a pragmatic viewpoint, parameter constancy between (4.3.1) and the model applied to whatever counterfactual situations under concern is the basic requirement of the present case. In comparison, consistency is an imprecise criterion because it implicitly assumes that the population of interest is *a priori* known for certain. Since MI analysis results in a plausible range of simulated complete-data samples $X_{0,MI}$, we can use them to assess parameter constancy empirically, i.e. regressing (4.3.1) with $X_{0,MI}$ to test H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{0,MI}$. Likewise, we can regress (4.3.2) with $X_{0,MI}$ to test H_0: $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_{0,MI}$. Both nulls are the expected norm except for (4.3.2) under MNAR. In other words, unexpected test rejections, particularly rejections of $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{0,MI}$, can only indicate model generalisation inadequacy rather than selection bias. This diagnosis is corroborated by undesirable statistical properties of the residuals from the two-subsample regressions.[23]

The above formulation shows how misconceived the selection bias claim is. It is also worthwhile noting that the primary weakness in confining the research realm within the supervised learning model boundary is actually not estimation inconsistency, but inefficiency, because idiosyncratic data information in Z_0 is indispensable for predictive estimation of

[21] For the definition of monotone missing, see Carpenter and Kenward (2013: ch. 3).

[22] See Carpenter and Kenward (2013: ch. 1) and Carpenter and Smuk (2021) for further discussion.

[23] In practice, f_{h_1}, f_{w_1} and f_l are commonly cast in parametrically linear forms. See Qin *et al* (2019) for an exploratory investigation of this example using the US data on married women labour supply.

counterfactual simulations. This point becomes transparent when research questions are led by specific policy targets, as in the case of the PEM-based average treatment effect research. The set policy targets have effectively fixed the population characteristics of interest. The validity and effectiveness of policy treatment evaluations is crucially predicated on careful *a posteriori* selection of appropriate control groups, frequently by transductive inference using matching techniques. The selection amounts to filtering l_1 for subgroups which are most comparable to specified policy targeted subgroups in l_0 in the present case.

Case two: Model-based leading indicator construction

This research question has remained a serious challenge in empirical business cycle studies for nearly a century. A well-known example is the factor-analysis based aggregation modelling approach.[24] Unfortunately, progress on this issue is disproportionately slow, especially compared to the amount of research inputs. The progress has been hindered by two related problems: lack of invariance in the loading structure of indicators and lack of adequate validation evidence of their predictive power beyond the training samples. From the viewpoint of machine learning, these problems manifest poor formulation of an analytical research question. The intended aim of leading indicator construction is to improve forecasting of certain macro targets, as implied by the attribute, 'leading'. The aim dictates that the research issue belongs ultimately to a supervised learning task. However, factor analysis falls into the unsupervised dimension reduction category.[25] Moreover, from a causal modelling perspective, observed disaggregate variables in factor analysis are a manifestation of certain latent common causes. In composite index making, which is what the present research amounts to, none of the constructed indicators can possibly be conceived as a common cause of the related observed disaggregate variables.

Let us demonstrate how the above research question should be reformulated using the case of constructing financial conditions indicators (FCIs) for macro forecasting.[26] Denote the macro forecasting variable of interest by y_t and suppose that its conventional macro model is of the ARDL class:

$$(4.3.4) \qquad y_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i y_{t-i} + \sum_{i=0}^{n} B_i X_{t-i} + e_t$$

where $X = \{x_1, \cdots, x_k\}$ denotes an input variable set including financial variables such as interest rates and monetary aggregates, and B is the associated parameter vector. Now, numerous economists have stated that those macro financial variables are inadequate to represent fully and speedily the vast financial markets. Their concern can be formulated as a hypothesis of omitted latent leading variables in (4.3.4), which can be approximated by FCIs. Suppose that a small number of FCIs, say $F = \{f_1, \cdots, f_s\}$, can be derived from a large set of observed financial market variables. The hypothesis amounts to extending (4.3.4) into:

$$(4.3.5) \qquad y_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i y_{t-i} + \sum_{i=0}^{n} B_i X_{t-i} + \sum_{i=0}^{n} \Gamma_i F_{t-i} + \varepsilon_t$$

[24] See Sargent and Sims (1977), Stock and Watson (1989) for further detailed discussion.

[25] Factor analysis is closely related to principal component analysis. Both are categorised as dimension reduction tools for unsupervised learning tasks. In measurement theory, models underpinning these tools are described as *reflective* models, whereas models for composite index making purposes are referred to as *formative* models. The latter type is more complicated and entails specification of more criteria than the former, e.g. criteria involving the target variables of prediction, see Markus and Borsboom (2013: Part II).

[26] See Qin *et al* (2022a) and van Huellen *et al* (2022) for a detailed account, and illustrations of this case.

Verification of the hypothesis lies in the search for evidence whether there are selected models in (4.3.5) which parsimoniously encompass the selected model in (4.3.4).

The above formulation embeds the FCI construction in a supervised learning framework. Hence, the following partial regression measurement model class is chosen, taken into consideration the 'leading' attribute:

(4.3.6) $$y_t = \sum_{i=1}^{n} \omega_{ji} I_{jt-i} + \vartheta_{jt}, \;\; j = 1, \cdots, m$$

where $\{I_{jt}, j = 1, \cdots, m\}$ denotes a large set of observable financial variables. The weights, ω_{ji}, can be estimated by the partial least-squares principle. Subsequently, the first few indices, F, can be selected for experiments using (4.3.5). According to the measurement theory, any measurable indices in F should be time-wise concatenable.[27] To validate this fundamental criterion, recursive estimation and tests of parameter constancy form a key part of the model selection process with respect to (4.3.6).

By putting the research task into a supervised learning framework, we can also adopt more target specific designs into the FCI construction than the basic case where factor analysis is used for the measurement model. This is best reflected in two dynamic aspects in the design. Specifically, dis-synchronised dynamics among observable financial variables from various markets is taken into the design of (4.3.6) such that the surviving variables through model selection can be dynamically heterogenous, namely in different lag orders. As for (4.3.5), we can also transform it into an ECM, in view of the popular transformation of ARDL models into ECMs, so as to decompose the target into a short-run variable and a long-run disequilibrium component. Consequently, long-run and short-run FCIs can be separately constructed for better results in terms of 'separate factors of variation'. Furthermore, the supervised learning step of (4.3.6) can be extended by a preliminary step of unsupervised dimension reduction when there is evidently redundant information among financial variables, for example, with similar variables from the same type of markets but geographical different locations.

It is worth reiterating that all the equations in the above two cases are model classes and that function estimation or learning in each class is essential. Yet the learning itself may still be inadequate for faithful formulation of economic hypotheses of interest. The formulation task goes well beyond the stringent hypothesis-testing framework. These cases therefore reinforce the key arguments of the last two chapters. Specifically, learning is indispensable for the formulation of practically robust economic models. The formulation process involves a close and principled interplay of subject knowledge and data through numerous experiments iteratively. Hence, the dichotomy of measurement with theory vis-a-vis measurement without theory should be discarded. This viewpoint actually reinforces Zellner's appeal, nearly half century ago, in his concluding remark: '"Theory without measurement" and "measurement without theory" are extremes to be avoided' (1979). Here, the 'extremes' are based on an oversimplistic way to classify models, namely to label models as 'theoretical' because they are *a priori* deduced mathematically, and models as 'empirical' when they are developed *a posteriori* through data analysis. The rise of AI and machine

[27] See Markus and Borsboom (2013: ch. 2) for further discussion.

learning has made this classification effectively obsolete. From the perspective of PAC learning, the interplay is an indispensable way to circumvent the brittleness and lack of targeted real-world focus of models produced by the hypothetical-deductive route. What is dispensable, however, is the imposition of a probability space on the distribution of the variables of interest. Probability is primarily and predominantly used in diagnostic tools for assisting decision making. The role of statistical inference on interpretable parameters becomes subordinate, albeit while still remaining an integral part of model selection during the formulation process.

5. Problems and Potential of Estimation

Optimal estimation of structural parameters in *a priori* deductively formulated theoretical models forms the core of present-day econometrics. This core is well established after decades of developments and extensions of Haavelmo's derivation of endogeneity bias. A recent summary of this core can be found in the discussion around the concept of 'estimand' by Athley and Imbens (2019: section 2.1). It is important to reiterate here that the assessment of statistical optimality of estimands is predicated upon the assumption of *known* correct models. The previous chapters have exposed how untenable this assumption is, no matter how elaborate and rigorous the deductive process is of those models. The flaw in this assumption thus undermines the core of present-day econometrics.

The primary task of this chapter is to lay bare the methodological defects in blindly seeking optimal estimators for *a priori* postulated structural parameters. Algebraic proofs of non-optimality of the OLS are based on diagnoses that residuals of the estimated models fall short of the classical assumptions in a default regression setting. More often than not, however, such diagnoses reflect the inadequacy of *a priori* postulated models for practical purposes. But if the global correctness of these models is taken for granted, the inadequacy can only be deemed as estimation defects. Their treatments result inevitably in implicit model modifications, and sometimes deterioration in model fitting or generalisability. Hence, the estimator treatment route effectively camouflages model formulation problems, offering systematically false confirmations for inadequately formulated models. This pitfall is demonstrated in Section 5.1 through an extension of the two cases discussed in Section 4.1. The discussion shows it is evidently premature to consider the task of parametric inference before model formulation problems have been empirically ironed out.

Once model learning and selection is prioritised, the primary role of estimation becomes the assistance of model selection. Section 5.2 examines how estimation fulfils this role as part of the learning algorithms. Since the algorithm design is ruled by SRM, choice of the basic estimation principle is driven by loss function minimisation subject to SRM, and the estimation results are used as indirect data evidence for model assessment. Formalisation of this role results in extension of estimators by a regularised loss minimisation constraint embodying the desired model selection principle. From the applied perspective, it is crucial *not* to interpret the resulting estimators in the same way as conventional estimators for optimal parametric inference purposes. This is because the rule of loss function minimisation is assessed against validation samples whereas conventional estimators are designed in a dichotomous sample-population setting, which ignores the need for dividing available data into training and validation samples.

Section 5.3 delves into the topic of post-selection inference. Once detectable non-optimal problems from model residuals are ironed out, the estimation problems presently described in econometric textbooks should generally disappear. The key task now falls on model reparameterization in feature design or representation learning, namely design and selection of parametric representations which not only faithfully translate the original economic causal claims but also capture the smallest possible separable input factors of data variations. Inferability of the resulting parameters touches on the question of whether classical tools are directly applicable to post-selection models. Contentions over this question are briefly discussed in this section. In short, the present chapter reinforces the main conclusion of the last chapter, namely that the framework of classical statistics is too narrow and thus impotent for empirical verification of economic theories.

5.1 Problems of Estimation Prior to Model Selection

As already discussed in Chapter 4, there is an unsurmountable gap between the setting of classical statistics and that of positive economic issues in general. However, since current econometrics is established within the framework of confirmative statistics, which assumes *a priori* models as the correct norm, estimation becomes the only gap-filling outlet when statistically undesirable signs reflecting the gap emerge from model residuals, either through logical deduction or data experiments. These signs are diagnosed as non-optimality of the OLS estimator, since OLS-based regression models form the key statistical apparatus initially adopted in econometrics. Consequently, econometric research is predominantly focused on the rectification of the prescribed non-optimality by alternative estimators. Sadly, this focus is strategically flawed. It mistakes fundamentally analytical research questions as enumerative ones. Pitfalls of this estimator-centred approach are demonstrated below through the two cases described already in Section 4.1.

Case one: OLS inconsistency based on *a priori* deduced endogeneity bias

The deduction and the estimation-centred solution of the bias is originally offered by Haavelmo in the context of a bivariate static SEM. Subsequent emulations of his analysis lead to extension of the endogeneity bias label to problems, such as omitted variable bias (OVB) and selection bias, in the context of single-equation models. Since these problems occur frequently in practice, fear and confusion over endogeneity bias has plagued econometric practice, and its IV treatment has effectively promoted widespread *p*-hacking.

To uproot the conceptual flaws surrounding the diagnosis and treatment of endogeneity bias, we adopt the separate notations in machine learning for error terms, ε_{in} and ε_{out}, to differentiate in-sample and out-of-sample errors. Continuing with the bivariate SEM of (4.1.1) in the last chapter, let us focus the examination on the upper equation. Denote the OLS estimation of that equation as:

(5.1.1) $y = \hat{\beta}_1 x + \varepsilon_1, \qquad \hat{\beta}_1 = (X'X)^{-1}X'Y.$

Simultaneity of y and x in (4.1.1) implies $corr\left(x\varepsilon_{1,out}\right) \neq 0$, whereas the OLS is predicated on $corr\left(x\varepsilon_{1,in}\right) = 0$. Hence, inconsistency of $\hat{\beta}_1$ with the SEM formulation is easily deducible prior to any data experiments. The inconsistency can be readily extended to single-equation models where $E(y|x)$ underlying (5.1.1) embodies causal claims based on a highly partial equilibrium principle. In such circumstances, omitted other input variables are bound to be present in passively observed samples. Moreover, this inevitable presence is accompanied by uncertainty in identifying fully what these omitted variables are. Denote such an *a priori* unspecifiable omitted variable as z. When prior reasoning cannot rule out $corr(xz) \neq 0$, the problem of OVB must be attended. Starting from (5.1.1) and following Haavelmo's step, the problem is readily reduced to $corr\left(x\varepsilon_{1,out}\right) \neq 0$ in general. As OVB is added as another key source of endogeneity bias, the OLS is widely held up to infamy for not being an asymptotically consistent estimator in econometrics.

Standard treatment of endogeneity bias is to replace the OLS by IV estimators. Denoting an IV set by V, the IV estimator can be expressed as the generalised least squares (GLS) in correspondence to the OLS in (5.1.1):

$$(5.1.2) \qquad \hat{\beta}_1^{IV} = (X'V(V'V)^{-1}V'X)^{-1}X'V(V'V)^{-1}V'Y$$

The IV estimator can also be represented in terms of a two-stage least squares (2SLS) procedure. The 1st-stage is a regression model based on $E(x|V)$, and the 2nd-stage is to replace x in (5.1.1) by $x^V \equiv E(x|V)$.[28] The corresponding 2nd-stage model is:

$$(5.1.3) \qquad y = \hat{\beta}_1^{IV}x^V + \varepsilon^V, \quad \hat{\beta}_1^{IV} = \left(X^{V\prime}X^V\right)^{-1}X^{V\prime}Y.$$

Evidence of $\hat{\beta}_1^{IV} \neq \hat{\beta}_1$ is widely accepted as empirical verification of endogeneity bias, but disquiet remains on the choice of corresponding IVs as well as the interpretation of $\hat{\beta}_1^{IV}$. In order to get $\hat{\beta}_1^{IV} \neq \hat{\beta}_1$, one must have $x^V \not\approx x$. This prohibits modellers from taking the 1st-stage modelling task seriously in the sense that they must try *not* to fit x as best as possible for the sake of verifying $\hat{\beta}_1^{IV} \neq \hat{\beta}_1$. This rule of non-optimality determines the non-uniqueness in IV choice. It reveals the model revision function of the IV estimator: tacit replacement of the *a priori* assumed 'endogenous' input variable by an IV-generated variable. From this viewpoint, the GLS of (5.1.2) can be regarded as the OLS on certain IV-weighted composites of x, and the choice of the IV estimator over the OLS is equivalent to one of (5.1.3) over (5.1.1). This brings us to the interpretation aspect. When $x^V \not\approx x$, (5.1.1) and (5.1.3) are *de facto* non-nested rival models. Since there are multiple ways of generating the IV-compounded x^V in (5.1.3), it lacks clarity, if not credibility, to interpret $\hat{\beta}_1^{IV}$ as measuring the causal effect of $x \rightarrow y$ rather than $x^V \rightarrow y$. From the perspective of two non-nested rival models, the *a priori* choice of IV estimation amounts to rejecting the original causal claim in (5.1.1) by implicitly replacing it with (5.1.3). In other words, the IV estimator achieves the removal of the diagnosed $corr\left(x\varepsilon_{1,out}\right) \neq 0$ through modifying the key conditional variable of interest, namely from x to x^V. Considering the multitude of possible

[28] It should be noted that it is incorrect to describe the IV estimator by a chain rule, i.e. $V \rightarrow x \rightarrow y$, following the 2SLS representation. For more detailed methodological discussion on the IV estimation and endogeneity bias, see Qin (2015, 2019).

x^Vs, it is unsurprising that IV-modified models are unlikely to gain unanimous trust as faithful representations of the original causal models.[29]

The predicament of the IV remedy brings us back to $corr(x\varepsilon_{1,out}) \neq 0$ and, specifically, the presumption upon which it is based, namely that given theoretical models are globally correct under normal circumstances. The general untenability of this presumption have been discussed before. Let us examine here how logically unsound the specific model setting underpinning the diagnosis of endogeneity bias is from two respects: formulation of the common-sense rule of interdependency between variables into simultaneous relations, and the nature and roles of the error term as model residuals. It is now a well-known fact that symmetrically formulated SEMs are not operational statistically, and that statistical operationality entails asymmetrically formulated conditional models.[30] Model representation of interdependency in the real world involves formulating complex feature representations of dynamic interactions between variables. Hence, static SEMs, as depicted by (4.1.1), are not only inadequate but simply unusable. When simultaneity is embedded in dynamically adequately formulated models, endogeneity bias becomes practically negligible, as formally proved in Wold's 'proximity theorem' (Wold and Juréen, 1953: 37-8) long ago. It should be noted that model adequacy in Wold's 'proximity theorem' is embodied by assumed white-noise model residuals. This assumption is expressed as consistency of models by Cox (2006: Appendix B). A discussion on how the notion of consistency relates to model learnability in machine learning can be found in Shalev-Shwartz and Ben-David (2014: section 7.4), as described in Chapter 2. Clearly, any consideration of the issue of how parameters in a given model should be optimally (including consistently) estimated should be predicated upon the verified condition that the model is data congruent and generalizable.

Assessment of model generalisability brings us to the nature and roles of the error term. It is crucial here to clarify the difference between ε_{in} and ε_{out}. Given data samples, ε_{in} is *known* unknowns in the sense that it is the residuals of predefined models and chosen estimators. As such, statistical properties of ε_{in} are derivative of both the models and the estimators. In contrast, ε_{out} is *unknown* unknowns. Its properties are not fully foreseeable even from a good combination of prior knowledge and available data information. Concerns over this predictive uncertainty lead to the focus on E_{out} as the object of assessment for model generalisability, as shown in the description of PAC learnability in Section 2.2. Division of available data into training and validation samples, and attentive examination of the latter by means of cross-validation techniques are manifestations of this focal quest. Sadly, the division is absent in textbook econometrics. It turns out that cross-validation or model-projection checks are the key to best exposing the spurious nature of the alleged

[29] Wold is among the first to argue that the essence of endogeneity bias is not an estimation issue but a causal modelling issue, see Wold (1954, 1956). He also highlights the inadequacy of a deductive modelling route, see Strotz and Wold (1960). In microeconometrics, an interesting case where the function of model modification of the IV route is effectively recognised is the redefinition of the average treatment effect in the PEM into 'local average treatment effect (LATE)', see Angrist *et al* (1996). Obviously, the term 'local' differentiates the IV-based PEMs from those *a priori* postulated PEMs.

[30] Evaluation of statistical operationality has been formalised into identification conditions in the context of SEMs. Unfortunately, these conditions are unrelated to a general connotation associated with 'identification', namely how models are verified to correspond closely to reality via data experiments, see Qin (1993: ch. 4).

consistency of the IV treatment. It was already well-known in macro-econometrics prior to the rise of VAR models that model projections using SEMs which were estimated by the IV route were unequivocally worse than those by the OLS.[31] In the context of cross-section sample-based studies, cross-validation experiments reported by Young (2017) and also van Huellen and Qin (2019) show no evidence of asymptotic convergence in IV-based estimation results. Specifically, their cross-validation results show that model (5.1.3) demonstrates generally superior asymptotic convergence to that of (5.1.1), indicating the former having stronger generalisability than the latter. Their findings imply that x is more often a valid conditional variable than its IV-compounded substitutes. Deduction of endogeneity bias from inadequate and unvalidated models is merely a delusion. Endogeneity bias should thus not be used as the reason for modifying arbitrarily key causing variables postulated by economists on the basis of prior common-sense knowledge.

Case two: OLS inefficiency due to observed autocorrelation in $\{\widehat{\varepsilon_{in}}\}$ when *a priori* formulated models are fitted to time-series data samples

Treatment of the problem by estimators involves utilising the estimated autocorrelation coefficients of $\widehat{\varepsilon_{in}}$ as weights added to the OLS.[32] The resulting estimator is again a GLS. In the simplest case of 1st-order error autocorrelation of a bivariate static model:

(5.1.4) $$y_t = \beta x_t + \varepsilon_t,\ \varepsilon_t = \rho\varepsilon_{t-1} + \nu_t,$$

the GLS can be of the general form: $\hat{\beta}^{\Omega} = (X'\Omega(\rho)^{-1}X)^{-1}X'\Omega(\rho)^{-1}Y$, where $\Omega(\rho)$ denotes the weight matrix in ρ. Now, substituting $\varepsilon_t = y_t - \beta x_t$ away from (5.1.4), we get:

(5.1.5) $$(y_t - \rho y_{t-1}) = \beta(x_t - \rho x_{t-1}) + \nu_t$$

Hence, the efficiency gain of the GLS over the OLS is achieved by an implicit modification of the static (5.1.4) into a dynamic model. Notice that the modification imposes a specific restriction via ρ on the dynamic pattern of both x_t and y_t. (5.1.5) is thus referred to as a COMFAC (common factor) model.[33] Reparameterise (5.1.5) into an ECM:

(5.1.6) $$\Delta y_t = \beta\Delta x_t + (\rho - 1)[y - \beta x]_{t-1} + \nu_t\ .$$

(5.1.6) reveals that the common factor restriction amounts to imposing equality of the short-run and long-run impacts of x_t on y_t. This is clearly an unintended restriction as far as the original causal claim is concerned. Numerous empirical results have shown that (5.1.6) misrepresents dynamic interactions between macro time-series variables in an oversimplistic way. In fact, the rise of VAR modelling in macroeconomics has made the above GLS treatment redundant, as residual autocorrelations are absent generally in VAR models.

Modelling experience with VARs tells us that observed residual autocorrelations indicate model inadequacy in dynamic formulation, a specific OVB problem where what have been omitted are informative lagged variables. The GLS treatment based on (5.1.4) offers an expedient way of removing the symptom. But this estimation treatment route contains serious

[31] See Qin (2013a: ch. 1) for detailed description.

[32] The estimation treatment originates from Orcutt's investigation of single time-series (1948), and the 1st-order error-term autoregression based estimator is widely taught as the Cochrane-Orcutt estimation procedure (Cochrane and Orcutt, 1949). An early textbook to formalise the procedure into the GLS is Malinvaud (1970).

[33] See Hendry (1995: ch. 7) for detailed exposition.

side effects. Practically, it is prone to bias in the estimated long-run steady-state parameters of theoretical interest due to the common factor restriction. Epistemically, it can easily misguide armchair economists to generalise this observed undesirable symptom from $\widehat{\varepsilon_{in}}$ to a global property of ε_{out}, and to use the property as a panacea for possible dynamic formulation inadequacy, as shown in the case of DSGE model research discussed in Section 3.2 of Chapter 3. This epistemic pitfall further highlights the rudimentary weakness of applying directly confirmative statistics apparatus to economic issues.

It is evident from both examples that what the estimator treatment route does is to modify inadvertently original *a priori* models, giving rise to the phenomenon of polysemy of structural parameters described in Qin (2013a: ch. 7). As a result, model generalisability is likely to deteriorate and/or parametric interpretability to become obscured. The solution is to deem statistically undesirable properties of the error term as symptoms of model formulation inadequacy and tackle the inadequacy following the PAC learning principle.

5.2 Potentials of Estimation in Serving Model Selection

For the task of learning models with the best possible generalisability, the primary role of estimation is to assist model selection as efficiently as possible, instead of deciding what the optimal estimator should be for individual parameters. The previous chapters have explained that SRM is the fundamental rule of this learning task. Moreover, most positive economic issues can be classified as supervised learning issues, and economic behaviour rules associated with those issues can be formalised as convex learning problems within the linear model class, which are solvable through minimisation of certain quadratic loss functions. A basic criterion of the minimisation is the least-squares principle.[34] It is worthwhile reiterating here that this least-squares principle is evoked in a discriminative or distribution-free manner during the model learning process. As such, the principle should be evaluated and interpreted in the context of optimal control algorithms rather than that of parametric inference. In other words, practitioners should be fully aware that the least-squares based minimisation is more versatile than an estimator for the latter purpose. Different evaluation criteria stem from different roles. Let us extend this point by looking at two issues: interactive uses of conventional estimators together with diagnostic tests, and the augmentation of conventional estimators.

Estimation primarily forms part of model selection algorithms. It is used interactively with various tests in an iterative manner. The essential goal of the algorithms is SRM. This goal entails an additional filter on those in-sample quadratic-loss-function-minimisation based estimation results. They should be filtered by the criterion of out-of-sample error minimisation, a requirement embodied in the VC generalisation bound for regression models, as described in Section 2.2. Define multiparameter regression model h by:

(5.2.1) $$h\colon y = X_d\beta + \varepsilon,$$

[34] In the case of discrete choice models using the logistic function specification, the minimisation results in a principle equivalent to the maximum likelihood (ML) principle. From the decision theory perspective, the maximisation can be treated as an optimisation issue of convex functions, without the need for any distributional assumptions, e.g. see Abu-Mostafa *et al* (2012: ch. 3), Shalev-Shwartz and Ben-David (2014: ch. 9).

where d denotes the dimension of X, its VC generalisation bound can be expressed by the following:[35]

(5.2.2) $$E[\varepsilon^2]_{out}(h) = E[\varepsilon^2]_{in}(h) + O\left(\frac{d}{N}\right),$$

where $E[\varepsilon^2]_{out}(h)$ and $E[\varepsilon^2]_{in}(h)$ are out-of-sample and in-sample mean squared errors respectively, and the term, $O(\cdot)$, means that its absolute value is asymptotically smaller than a constant multiple of the ratio, $\frac{d}{N}$, with N denoting the sample size. $O(\cdot)$ shows us the vital role of this ratio in minimising the generalisation error, $E[\varepsilon^2]_{out}(h) - E[\varepsilon^2]_{in}(h)$. Specifically, for models with equal $E[\varepsilon^2]_{in}(h)$, it is the model with smallest ratio, $\frac{d}{N}$, whose generalisation error should be the smallest. This not only formally corroborates with the conventional wisdom of going for the simplest possible models, but also clarifies why estimation results during model selection should be summarily assessed jointly with various information criteria, such as AIC, BIC or simply adjusted R^2, as described in Section 4.2. All these statistics penalise model complexity via variants of the ratio, $\frac{d}{N}$. This explains why information criteria are extensively used at the primary model selection stage. In a multivariate regression model context, the selection rule amounts roughly to model reduction via excluding those *X*s whose parameter estimates are statistically insignificant.

An alternative way is to augment estimators by merging the above interactive use of estimation and tests into their mathematical expressions. The augmentation amounts to adding a constraint, for example, via subjecting the commonly used $E[\varepsilon^2]_{in}(h)$ minimising principle to a constraint of minimising $E[\varepsilon^2]_{out}(h)$. The resulting principle is referred to as regularised loss function minimisation, or simply regularised loss minimisation (RLM). This route is thus known as 'regularisation'. In the context of regression models, h, estimation by the principle of quadratic loss function minimisation can be expressed summarily as:

(5.2.3) $$\hat{\beta}_{OLS} = \text{argmin}\{\|y - X\beta\|^2\} = (X'X)^{-1}X'y$$

where argmin$\{\cdot\}$ means obtaining β by minimising $\|y - X\beta\|^2$, a quadratic norm which denotes the residual sum of squares (RSS), namely a quadratic loss function, of model h. However, this optimisation solution only targets at $E[\varepsilon^2]_{in}(h)$. To incorporate the target of $E[\varepsilon^2]_{out}(h)$ as well, argmin$\{\cdot\}$ is extended by a regularisation term on β. Specifically, write the extended function as:

(5.2.4) $$\hat{\beta}_{RLS} = \text{argmin}\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\} = (X'X + \lambda D)^{-1}X'y$$

where $\lambda > 0$ is a tuning parameter and D denotes a diagonal matrix. Clearly, $\hat{\beta}_{OLS}$ is a special case of $\hat{\beta}_{RLS}$ with $\lambda = 0$. This modified target function in (5.2.4) produces the ridge estimation principle. Aiming at minimising $E[\varepsilon^2]_{out}(h)$ via its bias-variance trade-off relation, the modification is carried out through tuning down the variance part, by means of the regulation term, $\lambda\|\beta\|^2$, as long as the variance decrease is greater than offsetting the corresponding increase of the bias part. Essentially, the possibility of shrinking $E_{out}(h)$ via balancing the two components lies in the potential of model dimension reduction. This is

[35] For detailed explanation, see Abu-Mostafa *et al* (2012: ch. 3), Hastie *et al* (2009: ch. 7.9) and Shalev-Shwartz and Ben-David (2014: ch. 11).

clear if we consider the case of comparing the systematic parts of two models of different sizes, say $y = X_p\beta$ and $y = X_d\beta$, where $p > d$. Since $\sum_{i=1}^{p} \beta_i^2 > \sum_{i=1}^{d} \beta_i^2$, $\lambda||\beta||^2$ penalises larger models more than smaller models *ceteris paribus*, though there is no explicit alteration of the dimensions. Hence, this approach is described as exerting a certain soft cut-off threshold on model complexity, as opposed to a hard cut-off threshold, *d*, under the previous approach.

The controlling effect of the regularisation term on model complexity becomes more explicit when the quadratic specification of the term is replaced by an absolute value specification. The resulting estimator is known as *lasso*:

(5.2.5) $$\hat{\beta}_{Lasso} = \text{argmin}\{||y - X\beta||^2 + \lambda \sum|\beta|\}, \text{ or: } \min_{\beta}\{RSS\} \text{ subject to } \sum|\beta| \leq C$$

The equivalence of $\lambda \sum|\beta|$ to setting a threshold, $\sum|\beta| \leq C$, reveals that the previous approach of setting hard cut-off thresholds can be seen as a special case of *lasso* – when C is such that all insignificantly small parameters are cut off. If C is too large, the model risks overfitting whereas if C is too small, the model suffers underfitting. Regularisation via *lasso* effectively formalises the mixed use of estimators and tests by condensing the goal of minimising $E[\varepsilon^2]_{out}(h)$ into the estimation principle.[36]

Pivotal to the success of regularisation is the step of calibrating λ or C. By balancing between low empirical risk and possibly simplest models, the learning rule of RLM determines model complexity at its best possible parsimonious point, the point where the selected model is free of underfitting and overfitting. What underlies this balancing act of the calibration process is the division of data into training versus testing samples. Hence, the calibration implies cross-validation checks, and RLM is consistent with the rule of SRM. Moreover, model stability is proved to be a key condition for the success of regularisation, and thus the calibration.[37]

Overall, the role of estimation during the model learning process well exceeds the framework of statistical inference. It is an epistemic mistake to mix this approach of predictive estimation with the conventional role of estimation as a purely enumerative means. In particular, it is important to see that RLM-based estimators are merely a mathematically succinct way to convey the need for taking both E_{in} and E_{out} into consideration for model selection purposes. Therefore, those estimators must *not* be judged by the criteria used for estimators in the context of optimal parameter inference.[38]

The embodiment of SRM in the RLM rule sheds light on the practical efficacy of the parsimonious encompassing principle of the LSE approach in econometrics.[39] A compelling

[36] Regularisation is explained in detail in many machine learning textbooks, see Abu-Mostafa *et al* (2012: ch. 4), James *et al* (2013: ch. 6), Efron and Hastie (2016: chs. 7 & 16). The last book describes ridge regression from the James-Stein theorem, which proves the superiority of ridge estimators over ML estimators on the basis of distributor-free decision theory.

[37] For further discussion, see Shalev-Shwartz and Ben-David (2014: ch. 13).

[38] One example of such a misconception is the incorrect use of the criterion of unbiasedness to assess $\hat{\beta}_{RLS}$. Clearly, when (5.2.3) is unbiased, (5.2.4) must be biased and hence not optimal. This misuse may explain why the approach of predictive estimation has not been popularised in applied econometric research.

[39] See Hendry (1995: ch. 14) for a detailed exposition.

example of the efficacy is the estimation of latent long-run relations of multivariate time-series models via ECMs. Essentially, that principle is built on the same criteria of selecting models with the best possible fitting-stability trade-off. Although conceptually absent, model learnability is implicit in the aim of outperforming the extant rival models. Estimation and statistical tests are used in combination, coined as 'testimation' (Trivedi, 1984), during the selection process. The task of parameter inference is deferred to the end of the process, making fruitful model selection search a prerequisite of the inference. On reflection, it is more precise to call $\hat{\beta}_{RLS}$ and $\hat{\beta}_{Lasso}$ 'testimators' than estimators.

5.3 Post Selection Inference

Once the model learning process is over and the best possible parsimonious and data-consistent model selected, the issue arises as to whether optimal estimation and inference on parameters of the chosen model is now feasible. This issue is examined here from two angles. The first concerns reparameterization, and the second the caution and consideration needed for applying classical estimators after reparameterization.

Let us discuss reparameterization in the context of equations (2.1.1) and (2.1.2) in Chapter 2, namely a setting where the *a priori* theoretical model is simpler than the *a posteriori* model. The parameters postulated in theoretical models are generally weak as far as their statistical separability from data variation is concerned. This is because the causal relations they represent are mostly partial equilibrium states and the economic variables involved are flow and stock variables. From the viewpoint of (2.1.1), this weakness is commonly referred to as collinearity among x_i in econometrics. But from the viewpoint of (2.1.1) against (2.1.2), it is referred to as OVB of x_i with respect to z_i. Since the bias is resolved via model learning and the selected model is of the type like (2.1.2), the main goal of reparameterization is to circumvent collinearity as much as possible. An excellent example is the reparameterization of ARDL models into ECMs in the context of time-series modelling. Transformation of selected ARDL models into ECMs enables minimisation of collinearity in the former due to observable inertia in variables. However, the minimisation cannot resolve significant collinearity among x_i when the long-run relation embedded in the error-correction term is multivariate. Under these circumstances, calibration experiments are needed during the ECM estimation in order to find the best parametrically interpretable models. This may explain why calibration plays such an important role in DSGE modelling research. Effective model learning from cross-section data samples is a more complicated task than that from time-series samples, making the machine learning approach indispensable. Under this approach, the issue of reparameterization is widely discussed under 'feature design' or 'feature representation learning' in machine learning. It is noteworthy that Haavelmo's discussion on 'autonomy' in his Chapter 2 crystallises finally in feature representation learning, and its pursuit is no longer a supposition but empirically operational. Clearly, feature learning entails good use of subject-matter knowledge. As far as feature learning in econometric research is concerned, careful consideration of both economic rules and statistical criteria is essential in successful designs of algorithms, and the learning commonly involves heavily iterative steps. The core research attention of applied economists should therefore be focused on this challenging task.

The issue of optimal estimation is now on the agenda for parameters which are not gravely affected by collinearity after reparameterization or feature learning. Estimator choice for these parameters should be a relatively easy task at the post-selection stage. Since problems detectable from model residuals have been effectively ironed out, textbook reasons for more complicated estimators than the OLS have largely evaporated. The maximum likelihood (ML) principle is known to produce equivalent estimators to the LS-based estimators for well-specified regression models, and the principle serves as the linchpin of model-based inference in accord with classical statistics.[40]

When it comes to conducting post-section parametric inference, an issue of great concern is the accuracy or validity of the classical method of confidence interval (CI) calculation. Specifically, there is notable scepticism about whether such intervals are asymptotically correct with *p*-values appropriately controlling type-I error. The scepticism stems from a lack of consensus on what the appropriate probability space should be from which those *p*-values can be drawn for statistically learnt models with explicit uncertainty. On reflection, the scepticism is deeply rooted in the prejudice against data mining activities of the pre-computer era, and fuelled by fears of 'double dipping', an easily committed mistake of circular analysis during those activities. Circularity arises from double counting of data evidence, because available data have already been used for model selection. A simple and safe strategy against circularity is to conduct parametric inference only with 'undipped' data, via splitting available data samples to enable the inference being based on undipped samples only.[41] Admittedly, this practice is not always viable when sample sizes are limited. However, since the strategy underlies all the cross-validation techniques based on the division of training and testing sets, the resulting models thus selected should outperform in general those *a priori* deductively formulated models, especially considering the impossibility of obtaining data-congruent models by the deductive route alone in an open world milieu. Therefore, models statistically learnt following the PAC learnability principle fit relatively closest to the required setting of confirmative statistics, and are definitely better and hence more ready for CI calculation than models built solely by the *a priori* deductive route.[42] In other words, even when post-selection estimation apparently falls foul of double dipping, the resulting CIs of estimated parameters in models which have survived the PAC learning process should be far more reliable and credible than those in models without having undergone the process.

Nevertheless, the explicit acknowledgement of uncertainty in the PAC learnability principle remains a deterrent to some modellers in treating post-selection models as certainty equivalent units. The development of selective inference techniques is a concentrated endeavour to modify, in order to account for model selection uncertainty, the probability space within which parametric accuracy is measured.[43] The basic idea is as follows. Write the conventional CI formula for β_j of an *a priori* known model, M^h, with the desired significance level, α, as:

[40] See, for example, Shalev-Shwartz and Ben-David (2014: ch. 24).

[41] The method of data sample splitting in statistics is attributed to Cox (1975). In econometrics, scepticism against data mining has been long debated, see Qin (2013a: ch. 9).

[42] This point is effectively demonstrated in formalised theoretical arguments by Zhao *et al* (2021).

[43] See Taylor and Tibshirani (2015), Efron and Hastie (2017: ch. 20) for further explanation.

$$(5.3.1) \qquad \mathbb{P}\left(\beta_j \in \mathrm{CI}_j(\alpha)\right) = \mathbb{P}\left(\beta_j \in \mathrm{CI}_j(\alpha) \middle| j \in M^h\right) \geq 1 - \alpha$$

Denote a statistically selected model by $\widehat{M}$. Since the uncertainty in selecting $\widehat{M}$ is intimately related to d, a parameter of the selected model size defined in equation (5.2.2), an augmented CI which explicitly takes the uncertainty into consideration should be:

$$(5.3.2) \qquad \mathbb{P}\left(\beta_j \in \mathrm{CI}_{j \cdot \widehat{M}_d}(\alpha) \middle| j \in \widehat{M}\right) \geq 1 - \alpha$$

The above interval formula is developed in the context of simultaneous inference.[44] The idea of simultaneous inference arises from concerns over the lack of account, in individual confidence intervals, of the uncertainty effects due to other related input variables in a multivariate model setting. A classic case of simultaneous confidence interval is Scheffé's interval, which is based on *F* distribution rather than *t* distribution. The resulting intervals are usually notably wider than the conventional ones due to widening of the underlying probability space. However, Scheffé's interval is not designed for tackling model selection related uncertainty.

Should (5.3.2) be adopted as a standard CI formula for post-selection estimation? To find answers, we need to clarify the question of whether CIs calculated respectively from (5.3.1) and (5.3.2) are indeed comparable. On reflection, what causes $\left|\mathrm{CI}_{j \cdot \widehat{M}_d}\right| > \left|\mathrm{CI}_j\right|$ is the different conceptualisation of uncertainty in the generalisability of $\widehat{M}$ versus M^h. Based on the premise of M^h being globally correct, the possibility of any model generalisation uncertainty is totally disregarded in (5.3.1). But the uncertainty of $\widehat{M}$ is included in (5.3.2), making $\widehat{M}$ less certain than M^h from the outset. Since the premise is clearly at odds with reality, it is unreasonable to use comparison of the two types of intervals as evidence for assessing post-selection effects, and/or for discrediting data mining with the 'double dipping' claim.

Mounting evidence from various research fields confirms beyond doubt now that machine-learnt models following the PAC learnability principle outperform, in general, models built solely by *a priori* mathematical derivation. The evidence contradicts sharply with $\left|\mathrm{CI}_{j \cdot \widehat{M}_d}\right| > \left|\mathrm{CI}_j\right|$. The contradiction reinforces our earlier discussion on how untenable it is to equate naïvely the task of verification of economic theories to that of statistical estimation and hypothesis testing, since the task entails a complicated process of learning and selecting $\widehat{M}$ from data. In particular, the process involves multiple optimal control criteria in combined use with a host of computational and statistical tools. As a result, CIs of the parameters of interest should not be treated as the sole evidence for theory verification, and information gathered during the model selection process should also be taken into consideration as indirect evidence. In other words, reports of applied economic studies should faithfully cover all the evidence upon which the post-section CIs of the parameters of interest are predicated. To a great extent, such a broad coverage of evidence together with the conventional CIs is practically a more viable strategy than the approach of trying to summarise everything into one CI formula.[45]

[44] For detailed formulation and discussion, see Berk *et al* (2013), Efron and Hastie (2017: ch. 20).

[45] See Holmes (2018) for a detailed discussion.

From an epistemological perspective, disputes over, and discussions on, the issue of post-selection inference remind us again of the limited capacity of probability as a practical measure of uncertainty. In the PAC learning principle, the probability notion for model generalisation uncertainty is merely a formal and explicit representation of our limited capacity of learning and cognitive power. In other words, the concept is analytical rather than enumerative. Any attempts to measure that uncertainty through selected model sizes are therefore prone to committing unignorable measurement errors. Essentially, statistically measurable probabilities of the uncertainty of specific research objects are predicated on clearly defined probability spaces with absolute certainty. Such definition is impossible to materialise in the context of model learning. It should also be noted that although it is well-known that probability measures across differently closed spaces are not comparable, mistakes in making such comparisons are widespread in practice. Furthermore, the higher the dimension of the defined probability space, the harder it is to interpret the resulting probability as the measure of specific uncertainty of interest, not to mention the rising technical complexity in keeping track of the optimality of the measurement.[46]

[46] In relation to the rational way of applying probability, Holmes (2018) tells the following story about Einstein. He said to his students, ‘Life is finite. Time is infinite. The probability that I am alive today is zero. In spite of this, I am now alive. Now how is that?’ None of his students had an answer. After a pause, Einstein said, ‘Well, after the fact, one should not ask for probabilities.’

第四章 假设检验的用处与经济假说的模型构述

在哈维尔莫的概率论方法中，假设检验概念被置于连接理论与数据桥梁的关键点。后继的科普曼斯将哈维尔莫的理念凝括为“有理论的测度” 模式 (Koopmans, 1947)。该模式强化了哈维尔莫为计量学研究的划界，巩固了学科将为先验给定的理论模型做假设检验与参数估计作为技术主体的格局。

统计学的假设检验技术是为了分析随机试验数据场景而设计的。数据样本的设计和采集一般是以测度某给定的单个参数为目标的。研究者把对参数的预期测度设为原假设。通常，原假设为反映常态的特征，其证伪应是偶发的小概率随机事件。因此，这些小概率随机事件被统归入备择假设。这种原假设与备择假设的二分法模式，使得概率尺度得以成为验证前定假说真伪度的简便判别标准。而假设检验的用途则定位在以原假设为重心的*验证性*分析上。在“有理论的测度” 模式中，待验证的经济理论被置于原假设的位置。不幸的是，实证经济学研究的场景及应用目的与统计假设检验框架之间有着不可逾越的鸿沟。上两章业已阐明，实证经济学研究的场景属于 Valient 所描述的理论贫乏场景，计量模型中的经济学假说基本属于常识性知识。由先验演绎推导形成的理论模型普遍缺乏数据相合性，具有不可忽略的不确定性。换句话说，需被验证的经济学命题往往没有可由作为假设检验对象的单一参数精准表出的形式，其验证任务需通过对复杂的模型试验结果综合判断来完成。因此，对经济学命题的验证，一般是不可能直接用对理论模型中给定参数做统计假设验证的方法来实现的。这意味着，将假设检验框架机械套用在对实证经济学命题的验证上就会引发诸多问题。模型固有的可塑性和不确定性为构造以检验各种数据统计表象特征为对象的假设提供了丰富的拓展空间。这些表象特征仅是前定模型拟合数据样本的衍生物，并不能构成验证经济学命题的直接确凿佐证。同时，假设检验框架的局限视角，也极易使研究者误入验证性偏差的陷阱。

第 4.1 节通过两个基础案例来展示验证性偏差对计量学应用和理论研究的迷惑力。第一个例子是内生性检验。在以验证特定前定因果关系为目标的计量模型研究中，内生有偏性被广泛视为一个最常见、最基础的问题。因此，检验理论关注参数的内生有偏性被程式化为一必经步骤。但是，由于内生性检验所设的直接对象是由模型衍生得来的、不具唯一性，这一检验其实助长了 p 值操纵之风的盛行。第二个例子是时序单变量的数据特征归类检验，归类以甄别变量是否是平稳或非平稳过程为主旨。检验的源动力是上章所述的单变量的随机生成机制是经济变量联合概率分布基础元素的信念。由于不少经济变量的动态表征会超出经典统计学所设定的平稳随机时序过程的范围，本题便成为计量学基础理论研究的一大热题。为了维持上述信念，检验研发者不得不将单变量呈现出的各种复杂动态表征考虑进检验所依存的模型框架中。从科学哲学的视角看，这种扩展模型的做法属于事后假设 (“HARKing” (hypothesizing after the results are known))，误入了偏执的假说迁就主义 (accommodationism) 歧途。

一旦学界达成共识，将统计学习视为计量学建模所必行之路，假设检验的首要用途就自然会转向协助模型学习和选择的方向。第 4.2 节综述假设检验在模型学习过程中的诊断性功能，并进一步揭示假设检验模式对于非确定性假设选择的模型学习任务的不适合性。从机器学习的视角出发，假设检验框架之外的、以数据探索性分析为目

标的统计学习手段，才是模型的设计和选择中需首选的必备工具，如基于信息准则的统计量、基于样本分解的交叉验证法、以及对输入变量做系统排序筛选的步骤。只有在模型的学习选择完成之后，统计检验框架才有对模型内表述经济假说的参数做验证性分析的适用场景。

哈维尔莫在他的《假设检验》一章中，用了一节的篇幅强调了“从数据观察中形成理论”的重要性（第 17 节）。他承认，构建能概述经验现象的理论模型类任务，实属“归纳性科学”的范畴，存在超出建模者可控的“失败风险”。这意味着，他所向往的假说检验概念趋同于我们在第 4.2 节所述的机器学习思路。但在历史局限下，他把如何评判这种“经验式归纳过程‘正确与否’的问题”归入“形而上学”的讨论（原文第 82 页）。显而易见，第 4.2 节有关机器学习方法的讨论，为如何将他“从数据观察中形成理论”的理念付诸实施描述了一条可操作路径。第 4.3 节进而通过两个实例，具体说明统计学习手段在对经济问题的模型构述过程中的必要性。第一例解析微观计量学中的选择有偏性问题，第二例考察宏观计量学中的时序加总指标建模问题。从这两例不难看出，针对具体研究问题及其应用场景，缜密构述理论假说适用的模型类，并且充分利用机器学习及其他探索性统计手段来系统拆析数据、选择模型，避免犯第 2.4 节所述的第三类错误，才应是经济计量学研究的核心任务。

4.1 验证性偏差：机械套用统计假设检验框架的困误

哈维尔莫引入的统计学技术是以验证单个假设和估计单个参数为对象的。面对经济学多变量联立模型的基础构思，他首先关注和考察的问题是统计估计技术对联立模型中单个变量的适用性。他对于普通最小二乘法用于联立模型中的单个参数估计时缺乏一致性的推证，即当今学界通称的内生有偏性，为计量学研究重心奠定了的方略：研制并确保先验给定模型中结构参数的统计最优估计（详述见下章第 5.1 节）。本节便以单一参数的内生性检验为首例。

例一：内生性检验

虽然联立模型的一般形式是多方程的多变量回归模型，哈维尔莫对参数内生有偏性的推证 (1943; 1944, 第 5 章)，却是建立在最简单的、对现实过度简化的双变量静态联立模型之上的：

(4.1.1)
$$\begin{aligned} y &= \beta_1 x + \varepsilon_1 \\ x &= \beta_2 y + \varepsilon_2 \end{aligned}$$

在假定(4.1.1)中的变量服从联合分布的前提下，不难得出$cov(x\varepsilon_1) \neq 0$ 及$cov(y\varepsilon_2) \neq 0$的结果。这时，仅采用普通最小二乘法估计单一回归模型，如：

(4.1.2)
$$y = \beta_{yx} x + \varepsilon_y,$$

对照 (4.1.1)，就会有 $\beta_{yx} \neq \beta_1$。这是因为(4.1.2) 中的β_{yx}是以$cov\left(x\varepsilon_y\right) = 0$为前提条件的，该条件与(4.1.1) 隐含的误差项前提条件相矛盾。由该矛盾引致的参数偏差被称为“联立有偏性”。由于(4.1.2)中的解释变量x在模型(4.1.1)中也是被解释变量，因此上述偏差也被通称为内生有偏性。

目前学界对内生有偏性的关注主要出现在基于单方程式的应用计量模型场景。这时通用的偏差矫正法是工具变量估计法。以(4.1.1)中的第一式为例，对该式实施工具变量法估计相当于拒绝(4.1.2)，并以下述模型取而代之：

(4.1.3) $y = \beta_{yx^V} x^V + \varepsilon_y^V$，其中 $x^V = V\hat{\gamma}_{xV}, \hat{\gamma}_{xV} = (V'V)^{-1}V'x$, 且 $cov\left(x^V \varepsilon_y^V\right) = 0$。

式中的V表示工具变量集。在$x^V \napprox x$的前提下，(4.1.3)与(4.1.2) 实为两个互不嵌套的条件对立模型(non-nested rival conditional models)。按照哈维尔莫的变量分类，内生有偏性的推论相当于将x^V设为理论变量、并将x设为含有不可忽略的测度误差的可测变量的推论。内生性检验便是针对两者的取舍决策而设计的。检验的原假设和备择假设分别为：$H_0: plim\left(\beta_{yx^V} - \beta_{yx}\right) = 0$；$H_1: plim\left(\beta_{yx^V} - \beta_{yx}\right) \neq 0$。Hausman 检验便是最常用的内生性检验。

必须注意的是，经济学家通设和预期的内生性常态，在内生性检验中并不是由原假设表述的，而是由备择假设表述的。这种一反验证性假设检验设计惯例的做法，使得内生性检验更貌似于下一节所讨论的诊断性检验。更值得注意的是，依据内生性检验取(4.1.3)而舍(4.1.2)的决策是不考虑、而且通常是排斥模型拟合最优化选择标准的。在绝大多数的应用案例中，用工具变量法估计的模型残差之标准差要大于用普通最小二乘法的标准差，这时若按模型拟合程度评判，(4.1.3)是劣于(4.1.2)的模型。同时，由于 (4.1.3)引入了工具变量，其构成其实要比(4.1.2) 更为复杂。根据第二章综述的机器学习理论，拒绝相对简单的模型而选取相对复杂的模型，至少要有后者相对于前者有更高的拟合度和泛化度为实据[1]。另外，透过现象看本质，内生性检验的决策标准，实为检验单个参数显著性的标准，这一点可由以下模型表出：

(4.1.4) $y = \beta_{yxx^V} x + \beta_{yx^V x} x^V + \varepsilon_y^{V'}$

这时内生性检验等价于对上式中 $\beta_{yx^V x} \neq 0$ 的显著性检验[2]。正是这个单一的评判标准，加上构造$x^V \napprox x$的多种可能性，导致了应用学界众多的择差拒优模型结果。值得一提的是，应用学界对于这类选择决策泛用如下说辩：对于以验证理论假说所关注的单个结构参数为目的的研究课题而言，模型的最优化拟合及最优预测都不构成模型研究的必备标准。$\beta_{yx^V x} \neq 0$ 这一标准不过是对上述说辩的具体体现。如此彰显的验证性偏差视角决定了，内生性检验从本质上是有别于那些协助模型学习选择过程的诊断性检验的。

内生性检验的显著性检验本质，加之其貌似于诊断性检验的表象，为应用模型研究者通过工具变量的选取来维持前定预期因果模型，提供了一条便利途径。由于工具变量在定义上是在结构模型所包括的变量集之外的，通过试验不同工具变量的组合、利用它们与前定疑似内生变量的相关性，取得统计显著的β_{yx^V}或者$\beta_{yx^V x}$估值，往往并非难事。因此，工具变量法不但免除了应用模型研究者通过数据来学习选择最优模型的麻烦，又有助于他们寻取所期待的结构参数估值。值得一提的是，在统计模型的应

[1] 模型(4.1.3)比(4.1.2)拟合程度差的情形不同于过拟合的情形。后者指的是模型在样本内的拟合程度随着模型设定复杂度而提高、但模型在样本外的预测精度随之下降的情形。而基于工具变量的模型相对于不使用工具变量的模型，两者均下降。

[2] 这种实用检验方式在教科书中多有介绍，如参见 Berndt (1991) 的第 8 章。

用研究中，这类将谋求参数的统计显著性视为主旨的研究行为被简称为 p 值操纵；通过获得统计显著的参数估值来追求研究论文发表的学风，则被描述为“星级大战”[3]。

有关工具变量法的 p 值操纵程度，Brodeur *et al* (2020) 的案例分析颇具说明力。该分析以比较项目评价模型所通用的四种估计法的 p 值操纵程度为主题。这四种估计法是：数据对照随机试验法、工具变量法、倍差法和断点回归设计法。分析的案例来自从 25 个经济学顶级期刊在 2015 年发表的 308 篇论文，分析的样本是这些论文发表的 13440 个有关项目因果参数的估值。分析比较结果是，工具变量法的 p 值操纵程度相对最为显著。这一结果其实是 (4.1.4) 的必然产物。只要工具变量有一定的选择周旋空间，建模者便很可能找到满足 $x \not\approx x^V$ 的x^V，实现验证$\beta_{yx^{V_x}} \neq 0$ 的目标。p 值操纵是验证性偏差的典型表现。显然，享有 p 值操纵空间的科研方法是有损于其客观可信度的。这种学风的滋生反映出科研者对统计假设检验框架之局限性缺乏正确认识，没有看到在大多数现实情形下，由假设检验得出的 p 值远远不能作为测度不确定性的唯一指标。鉴此，《美国统计学家》学刊 2019 年特发了题为 “21 世纪的统计推断：p 值 <0.05 的彼岸”的专刊[4]，警醒统计模型应用研究者慎重使用统计推断技术，周全考虑统计试验的不确定因素。

例二：对时序单变量是否归于非平稳过程的分类性检验

本题成为计量学理论研究热题的主要原因有三：(a) 单个变量的随机生成机制是经济变量联合分布基础元素的信念；(b) 许多经济单变量明显超出了简单平稳过程可描述范围的表象；(c) 非平稳随机过程在经典假设检验中的空白。本题的研究初始于基于自回归模型前提上，区分变量是否含单位根的检验。以第 3.1 节中的一阶自回归概率模型 (3.1.2)为例，假设检验的设定为 $H_0: \rho = 1$，$H_1: |\rho| < 1$，如见最常用的 Dickey-Fuller (DF)检验。鉴于大量宏观经济变量呈现出的缓慢动态表征，DF 检验将单位根设定为反映常态的原假设。这一设定却超出了以平稳过程为基础的统计学假设检验框架。非平稳随机过程的分布函数形式给构造设计满足统计学最优检验标准的检验统计量带来巨大技术挑战[5]。为了尽量减小检验显著性水平的畸变，并尽量提高检验的功效，检验设计者需要克服比传统检验复杂许多的数学难题。同时，为了提高检验的适用范围，还需要不断扩展检验所基于的单变量模型，如从 (3.1.2) 的$AR(1)$向高阶自回归模型的扩展：

(4.1.5) $$y_t = \sum_{i=1}^{n} \rho_i y_{t-i} + u_t,$$

扩展的 DF 检验就是基于 (4.1.5) 的。然而，$AR(n)$ 模型类仍不足以充分概述现实中单个经济变量的常态动态表征。一种更广义化的形式是含确定性要素 μ_t 的时序模型：

(4.1.6) $$y_t = \mu_t + z_t,\ z_t \sim AR(n)$$

[3] 由于在各学术界发表的论文中，通用不同个星号“*”表示参数估值的统计显著程度，Brodeur *et al* (2016) 便把这种注重追求统计显著的参数估值学风描述为论文发表的“星级大战”。有关 p 值操纵的详述，则可见 Hirschauer *et al* (2016)。

[4] 鉴于统计显著性的广泛应用，这一统计学争论也得到了许多学科的关注，如见 Mayo (2018) 从科学哲学角度的反响阐述。

[5] 有关本题研究的详细综述，可参见 Mills and Patterson (2006) 主编的 Palgrave 经济计量学手册第一卷《经济计量学理论》中的第 Ⅳ 部分。

式中的 μ_t 成为扩展模型涵括各种可能假设的灵活项，如 $\mu_t = d_t'\delta$ （d_t'表示先验给定的确定性变量向量，δ为相应的待测参数向量）。针对那些认为变量是随时间而变化的假说，就可在 d_t' 中设入时间变量 t，即将时间设为一个确定性因素；而针对变量不时会呈现出的"结构"断裂现象，则可将这些断裂点由哑变量表出，包含到d_t' 中。为了体现结构断裂的随机偶然性，还可设定 $\mu_t = \mu_{t-1} + \pi_t\varepsilon_t, \varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$，其中的 π_t为代表非确定性结构断裂事件的伯努利随机过程。显然，对于μ_t的不同设定都会影响相应假设检验的优化设计。当 μ_t的设定中涉及待测参数时，假设检验的设计还必须考虑进模型参数的最优估计问题。鉴于所涉数学推导的复杂性，加之对计量学主体范畴的覆盖性，该课题被视为引领计量学理论研究前沿技术水平的热题也就顺理成章了。

不过，尽管单变量时序归类检验理论设计的成果云集，它们的应用功效却颇为有限。对于绝大多数经济学家而言，精确推断单变量所属的随机过程归类与他们关注的现实课题相差甚远。最能反映这一差距的便是对结构断裂截然不同的解释。在以 (4.1.6) 为基点的理论研究中，结构断裂被处理为随机数据生成过程中的偶然污染点。但在现实中，应用模型在偶发意外事件冲击下的承受力，以及模型对有可能预测的突变事件的预警能力，则是大多数经济学家所关注的基本问题。他们也从不把单变量模型视为结构模型。归根结底，式(4.1.6)不过是一个对数据时序表征的描述性泛式。它与经济学家所关注的、由数个因果变量构成的经济机制差异明显。因此，从时序变量中观测到一定的规范性随机表征，并不能证实这些表征一定是变量的数据生成机制。不言自明，当假设对象是表象而不是本质时，检验结果不具实质性功效是必然的。这种验表不验本的假设检验只能起到维持验证性偏差的作用。从 (3.1.2) 到 (4.1.6)的广义化过程，都是为了维持单个变量的数据生成机制之基础性的信念。正是出于这种认知的验证偏差，数据中不断反映出的单变量随机特征的脆弱性才被纳入模型需扩展的范围之内，加以正规化。Hitchcock and Sober (2004) 从科学哲学的角度出发，将这种一味增加假设模型可塑性的做法归为认识论上的迁就主义。迁就主义会激励过拟合模型的研制，使原本不被实例证实的假说成为不可被数据拒绝的统计假设。Kerr (1998) 在反思逻辑经验主义的方法论时，构撰了"HARKing"一词，以描述这类依已知结果设定事后假设、又将它们表述为事前假设的行为。他还归纳了 HARKing 容易导致的各种科研弊端。

不过需要阐明的是，HARKing 行为本身并非必不可取，它的广泛存在其实反映着现实中的因果推理过程的复杂性。HARKing 的负面效应的主要致因，是不能满足该过程所需的经验学习成分的统计假设检验框架。也就是说，对于需要 HARKing 的课题而言，直接用假设检验的套路做经验研究是不适宜的[6]。这种不适宜性，其实是源于经典统计学的研究对象与大多数应用社会研究对象的不匹配。为了明示这两种对象性质上的不同，Deming (1975) 将统计学的研究对象归为枚举类(enumerative)问题，以区分于应用社会研究的解析类(analytic)问题。统计假设检验针对的枚举类问题的前提假说形式简单，总体划界明确，数据样本的采集控制条件严明。由于问题分析的实效完全取决于从样本得出的测度结果，因此统计分析的焦点就是参数最优估计。解析类问题则不同，其研究对象通常是社会中的某种因果关系，研究旨在探究如何改善该关系的未来性能的切实途径。因此，对该关系在总体不具确定性的未来做出条件预测，才是问题研究的重心。采用假设检验手段来处理解析类问题，必然会引致对通用的简单假设

[6] 详述可见 Prosperi *et al* (2019)。

构述形式附加更多的条件约束的做法。这种做法其实是想通过扩展理论假说形式的可塑性途径，来尽量挽救假设检验框架的适宜性[7]。这便是体现迁就主义的 HARKing 行为的实质所在。遗憾的是，这种方略忽视了解析类问题需要在开放式场景下做条件预测这一关键目标。有关这一点的详细讨论见第 6 章。

不言而喻，实证经济学的研究对象大都为解析类问题。如前所述，处理这类问题的关键在于如何根据研究任务设计和选择出与数据信息相合的模型。也就是说，为了用假设检验工具来实证解析类问题，首先必须有实证可足以信赖的模型。这种模型的学习过程也会用到假设检验工具。但这时的检验并不直接服务于实证先验关注的因果假说命题的目标需求，而是为了分辨模型的各种数据统计表征是否符合最优模型的标准统计特征，以佐助最优模型的选择过程。为了区别这类检验的不同用途，在计量学教科书中，通常把这类检验称为诊断性检验，并把检验中的假设统称为辅助性假说[8]。

4.2 诊断性检验与模型选择

经济计量学中最常见的诊断性检验莫属对模型残差项之随机分布特征的检验了。这些检验均以残差独立同分布这一经典假定为基准。该假定由若干原假设表出，如残差项服从正态分布、残差项同方差、残差项序列不相关等等。除了检验残差之外，诊断性检验还涉及两个方面，一是对模型参数估值的常数不变性检验，如邹检验及残差的累积和(CUSUM)检验。这类检验是任何用于预测的模型所必做的。另一方面是比较分辨不同模型功能的检验，如在可嵌套模型范围内针对遗漏变量风险的参数零约束检验、以及比较互不嵌套模型的似然比检验。针对模型比较的检验被统称为包容检验[9]，检验的假设形式多样。值得一提的是，无论是对参数估值的常数不变性检验还是针对模型的包容检验，检验的根本对象仍然少不了模型残差的统计表征。

诊断性检验的原假设以描述研究者对于模型所预期的理想统计特征为对象。原假设被拒的检验结果反映着模型与数据之间的不匹配，因此诊断性检验也被称为模型误设检验，以别于针对检验理论假说的参数显著性的模型设定检验。另外，由于诊断性检验的最终对象仍为残差项的统计表征，因此针对诊断结果修正模型设定的做法被称为误差统计方法[10]。误差统计方法其实是为如何合理地 HARKing 寻求路径。现实中，先验构造的模型与具体应用数据场景匹配的情形是小概事件。因此，诊断性检验被拒属于应用计量模型研究的常态。这时该如何做出适当的治理方案，是计量学研究中的一大软肋。在哈维尔莫定位的先验理论模型设定无误这一前提假定约束下，教科书把选择参数最优估计法立为治理非理想诊断性检验结果的通用之方。就以残差项序列不相关的检验为例。当用静态模型拟合时序样本时，通常会检验出残差的自相关表征。这时不难证明，通用估计模型的普通最小二乘法丧失有效性。弥补这一有效性缺失的主流计量学方法，是采用包含误差自相关系数信息的广义距法，取代普通最小二乘法。广义距估计方法被普遍视作诊治残差自相关的便利方法，这尤其体现在使用动态面板数据方面的模型研究中。从模型设定角度看，这种广义距法对估计有效性的提

[7] 详述可见 Hahn and Meeker (1993)。

[8] 有必要指出，用“辅助性”这一概念对检验的假设分类有着局限误导性。从验证经济理论都需要通过模型来间接实现的认知广角看，计量模型中使用的假设检验其实都属于辅助性的检验。

[9] 包容检验源于 Cox (1961; 1962)； 在计量学中的早期发展可参见 Pesaran (1987)、以及 Mizon and Richard (1986)。

[10] 详述可见 Spanos (2010; 2018)。

高，其实是根据备择假设，通过在估计量中引入对误差项的自相关设定，将静态模型隐蔽地改变成动态模型来实现的。也就是说，这一诊治方法属于隐性的迁就主义 HARKing。这点从上章 3.2 节中所举的 DSGE 模型例子中反映得最明显。上节例二中对于含结构断裂单变量广义模型的设定，则不过是将这种隐性行为显性化了。

那么，在诊断性检验被拒时，直接按备择假设修订模型的处方之弊端何在？追根求源，引致处方背后的认知失误有两个相互关联的因素：假设检验框架之局限性和模型残差之衍生性。诊断性检验受到验证性假设检验框架的制约，将预期常态的原假设被拒的情形设为小概率事件。应用模型研究大量结果的事与愿违，等于是证伪了上述预期，使得根据检验结果隐蔽或公开修正模型的 HARKing 行为成为必行之路。遗憾的是，由于检验的直接对象只是某种统计表征，当原假设被拒时，检验所能告诉我们的只是样本数据测出该表征不达标的现象，而不能告诉我们不达标现象的致因。判定模型误设的原因已超出了检验的设计范围。就针对模型残差的诊断性检验而言，由于残差是模型的附属衍生品，不达标的检验结果只不过是一种对模型不达预期理想状态的表述。这时若简单依备择假设来制定 HARKing 方案，就只能是治表不治本。例如，当我们用常态模型来拟合的数据样本包含某种突发经济危机事件时，残差的多项诊断性检验之原假设都可能被拒绝，如残差项非正态分布、异方差等等。这时，若参考每个备择假设、通过采用复杂的误差项分布函数来治理问题，显然会弄巧成拙。可见，就模型学习和选择任务而言，诊断性检验工具还远不够用，需要有严格缜密的系统建模策略和步骤。其实，伦敦经济学院派的动态建模路径正是应此而生的。该路径总结了先验模型动态设定不足的通例，采纳一般动态模型类为模型选择的起点。由于从预期上该模型类应能使诊断性检验之原假设的验证成为常态，这样该建模路径就从起点上基本避免了假设检验的局限性。建模者还可以利于各种诊断检验工具对该模型类做逐步约化，直到筛选出数据吻合的最简洁模型为止。相比之下，在基于横截面数据样本的微观计量模型研究中，先验模型构述不足的原因错综复杂，要想总结出能满足使原假设的验证成为常态这一前提条件的模型类绝非易事。这也许就是诊断性检验在微观计量模型研究中一直不受待见的一个重要原因吧。

必须看到，探索性建模过程很难有唯一的一条通用系统路径可循。就诊断性检验工具的适用性而言，当备择假设的设定面临多种可选项时，基于二择取一的检验框架就不再适用，需要构造考虑到多重可行假设的 p 值分布的大规模假设检验手段[11]。而当备择假设的多种可选项个数是先验无法确定、亦即备择假设集不属于闭集的时候，以概率 p 值为临界值的诊断性检验就失去了适用性。这就是为什么在机器学习中，协助模型的函数学习任务的判别性检验大都不是以 p 值为临界值的统计量的原因。无需多言，在探索性研究任务面前，验证性统计工具的实用性是相当有限的。

下面，让我们从机器学习的模型设计和选择原则角度出发，来讨论一下诊断性检验在选模过程中的位置和用途。为了便于讨论，我们以第二章中的关系式(2.1.2)和第三章中的关系式(3.3.1)为对象，将模型选择问题分解如下：

A. 就 (2.1.2)中的两个输入变量集$\{x_i \cdots x_n\}$和$\{z_1 \cdots z_m\}$而言，有哪些变量对目标变量起着显著而且有规律的解释作用？

[11] 详述可见 Efron and Hastie (2016, Ch15)。

B. 就 (3.3.1) 中那些被选入的输入变量而言，如何构述和选择它们相应的输入函数形式，才能最好地体现它们与目标变量的经济因果关系？

据机器学习理论，每个问题的处理都需要借助多个决策标准以及比较尺度，并需要在一个系统建模策略下通过构造有效的算法和迭代步骤来实施。算法和迭代步骤背后的最基本原则，应当是第二章所述的 SRM 准则。该准则含有两个选模基准：模型的精确性和泛化性。精确性体现在模型的 ERM 上，而泛化性则要求模型既简洁又稳定不变。这意味着，在以给定的模型类为出发点时，选模的关键是找到既不欠拟合又不过拟合的模型。为此，我们需要将数据样本分解为模型训练和测试两个部分，即 $\mathfrak{D} = \mathfrak{D}_{train} \cup \mathfrak{D}_{validation}$（$\mathfrak{D}_{train} \cap \mathfrak{D}_{validation} = \emptyset$），通过搜索偏差-复杂度之最优权衡来实现模型选择。就统计量而言，辨别模型欠拟合或过拟合的最常用统计量莫属集模型拟合精度和规模大小为一身的调整R^2了[12]。除此之外，常用标准还包括一组称为信息准则的统计量，如赤池信息准则 (AIC)、贝叶斯信息准则 (BIC)、汉南-奎因信息准则 (HQ) 等等。这些准则的差异源于它们对模型规模扩展的惩罚函数形式的不同选择。从理论角度看，惩罚函数形式的不同反映着选模标准的差异。例如，AIC 统计量是以模型的预测精度、即渐近有效性作为构造惩罚函数的基准，而 BIC 统计量则是以模型的渐近一致性作为构造惩罚函数的基准。从应用角度看，由于 BIC 对于模型规模扩展的惩罚程度要严于 AIC，按 BIC 选取的模型很可能会比按 AIC 选取的模型要简洁些。也就是说，AIC 对于模型欠拟合的情形比较敏感，BIC 则对于模型过拟合的情形比较敏感[13]。与以概率作评判基准的假设检验相比，基于信息准则的统计量不仅避免了选择适当概率临界值的麻烦和争议、更易操作，而且还可被灵活融入以控制模型规模为目标的正则化最小二乘估计法中，作为针对ε_{out}的有约束优化算法的校准项（详见下章第 5.2 节的讨论）。

以上讨论的检验手段及步骤，主要是针对问题 A 的。对于 B，需要考虑的一个重要评判标准是经济学解释性、亦即模型对于关注议题的说服力。就 (3.3.1)而言，该式中的参数就是上述标准的评判对象。我们在上章 3.3 节业已阐明，设计对议题有意义的参数不是单靠先验推理就能实现的，该设计也是一项后验学习任务，与函数学习任务密切相关，通称为特征学习。在特征学习过程中，诊断性检验显然是不可缺少的工具。首先，模型误差项顺利通过各种常规诊断性检验，无疑是评价参数估值之统计特征性质的必备前提。其次，对模型参数不变性的检验也是特征学习中的必备环节。毕竟，不断寻求泛化性更强的模型是科学研究的共同目标。泛化性意味着模型结构之不变性，该不变性又是从参数估值之不变性集中反映出来的。对于基于时序样本的模型来说，递归估计法原理是对样本做训练与测试部分划分的有效手段，相应的各种递归式参数估值不变性检验是验证模型泛化性的便利工具，在样本量相对有限的情形下尤为如此。对于基于大样本微观数据的模型来说，机器学习中的各种交叉验证法则是验证参数不变性及模型稳定性的必备工具[14]。另外，各种包容性检验也是模型选择过程中的一组便利工具包，对各种信息准则统计量起补充作用。

[12] 统计学中与调整R^2相对应的还有马洛斯 (Mallows) 的C_p。不过这一统计量在计量学教科书中很少介绍。

[13] 有关信息准则中不同惩罚函数选择的讨论，详述可见 Kadane and Lazar (2004) 及 Ding *et al* (2018)。有关模型选择过程中采用 AIC 与 BIC 准则之争的讨论，可参见 Yang (2005)、以及 Dziak *et al* (2020)。

[14] 有关交叉验证的综述讨论，可参见 Arlot and Celisse (2010)。

上述分析表明，在开放世界中的解析类问题面前，统计检验手段的首需功能是判别功能。只有在判别学习到了具有泛化力的最简洁模型之后，计量学采用验证性假设检验框架、对表示经济理论假说的参数做模型设定检验的时机才成熟。这意味着，支撑统计检验的概率论的首要作用也是其判别决策功能。相比之下，概率对于具有经济学含义的参数的统计推断功能仅是第二位的。

还需强调的是，在不少解析类问题中，先验假说只能给出因果解释要素的大致可能范围，而不能明确划定一个解释变量闭集。鉴别哪些变量确实对目标变量起着规律性的作用，是探索性后验学习中的一个要点。这时，除了利用信息准则等统计量协助变量选择之外，如何排序变量选择过程也是建模者需要考虑的。对于这一点，在机器学习中有着正向选择 (forward selection) 和反向排除 (backward elimination) 两个排序途径[15]。计量学中的伦敦经济学院从一般到具体的动态建模法，其实就是用的反向排除法选择原则[16]。该原则适用于被选择变量个数有限、但变量之间存在显著相关性的情形，如对若干个时序变量滞后项的筛选。正向选择途径通常适用于被选择变量集比较大、而变量间的因果替代效应不强的情形。而当研究课题中的解释变量及其输入形式都存在很大的不确定性时，建模者往往需要迭代循环使用函数学习与特征学习的算法反复试验，才有望实现应用模型的最终设计和选择。

4.3 “从数据观察中形成理论”：从数据拆析到模型构述

哈维尔莫在他的《假设检验》一章中，着重解析了“从数据观察中形成理论”（原著第 17 节标题）的含义。他强调，数据中的“经验知识”是理论建模必须参考的知识（第 81 页）。这相当于对 HARKing 行为的某种认可，也是对假设检验框架缺乏直接适用性的默认。如今，利用电脑总结提取数据经验知识，无疑是对人脑的强大扩充。不仅如此，我们还可将“从数据观察中形成理论”的过程电脑化、系统化，这便是机器学习的真谛。显然，参考数据信息后验形成理论的过程要比单纯依靠先验推理构述模型的过程复杂。导致该复杂性的一个不容忽略的主因是：适于应用的理论模型不仅要包含具有全局普遍性的理论假说成分，而且还要包括反映具体研究场景中不可忽略的个性化特征。对这两个部分的适当权衡和结合，是实证模型研究成功的必备前提[17]。毕竟，模型构述任务是比统计检验层次更高的研究任务。

上节沿机器学习思路讨论的模型选择过程是围绕着(2.1.2) 和 (3.3.1) 这两个模型类的。这两式只是对许多实证问题的广义函数表示，并不能充分体现根据特定研究课题、通过数据拆析实现模型构述的必要过程。下面就用两个课题案例来补充这一点。

<u>案例一：采用数据有缺失的样本对劳动力供给的工资效应做估计推断</u>

在用住户调查样本研究劳动力供给行为时，由于样本中无业应答人的工资一项空缺，就造成了样本的部分数据缺失。这时如仅用从业人员的子样本做工资供给效应的估计，其结果是否具有总体推广性呢？计量学教科书将上述问题构述为采用截尾 (truncated) 或删失 (censored) 数据样本的限值因变量模型 (limited dependent variable

[15] 详解可参见 Hastie *et al* (2009, Ch3) 和 Shalev-Shwartz and Ben-David (2014, Ch25)。

[16] 这一原则贯穿了 PcGive 和 PcGet 这两个时序计量建模软件包的设计理念，详述见 Hendry and Krolzig (2003)、及 Hendry and Doornik (2014)。

[17] 这里可参见 Swann (2006) 对注重考察分析具体经济特征之重要性的讨论。

model)，并将问题的处治集中在一致估计法的选择上[18]。这一定位的推理依据是，数据缺失是源于应答人的选择行为，致使通用的回归模型被结尾，使得普通最小二乘法失去一致性。因此，这种欠一致性被称为选择有偏性。然而，上述推理存在着认知偏误。估计量的一致性是针对已知总体特征而论的。而本例所涉的推断总体属于**反事实**总体，即假若那些无业人员也去就业的虚构情形。因此，问题的答案取决于我们对该虚构情形总体特征的构述， 这一构述是限值因变量模型类所不能满足的。显然，解决问题的起点是对每个无业人员的反事实劳力个体供给行为做出预测估算，这样才能对先验未知的反事实总体特征及其可行疆界做出后验预测估算。前面的第 2.3 节业已提过，针对个体或特例的统计推断问题属于传导学习范畴。

下面就细述一下我们应如何从传导学习理念出发来构述本课题。用 $X = \{h, w, Z\}$ 表示针对某现实总体有着足够代表性的住户调查样本。其中的 h 为劳力的工作小时量，w 为小时工资率，Z 是与h 和 w 有关的解释变量集 ($Z = Z_h \cup Z_w$, 且 $Z_h \cap Z_w \neq \emptyset$)。当应答人$h_i = 0$ 时，样本中的 w_i 缺失。构造一个缺失数据指数 l：当w_i 缺失时，$l_i = 0$，当w_i 不缺失时， $l_i = 1$ 。在微观经济学中，通称 l 为劳动力参与率(labour force participation (LFP))（为简化起见，下文中尽量删去了标注个体的下角标 i）。依 l 将样本分为两个子样：$X = \{X_0, X_1\}$。通常，X_1的样本量要明显大于X_0的样本量，但后者也非小得可忽略不计。现将使用 X_1刻化劳力工资供给效应的模型记为：

$$h_1 = f_{h_1}\left(w_1, Z_{h_1}; \boldsymbol{\beta}_1\right) + \varepsilon_{h_1} \tag{4.3.1}$$

参数集$\boldsymbol{\beta}_1$中对应于w_1的是理论最关注的参数。这时我们关注的议题是：由 (4.3.1) 得出的结果是否可用于推断假想 l_0 中所有人都就业的虚拟行为特征？必须看清，此议题下的h_0 已经不是原样本中的零值变量，而是反事实场景中的缺失变量。

经济学对 l_0组人群不就业选择的一个通用解释是：该人群的潜在劳动力价格w_0是无法实现劳动力市场清零的价格。这一假说解释判定工资数据缺失的形式是非随机的。于是便有了对使用普通最小二乘法估计 $\boldsymbol{\beta}_1$是选择有偏性的推论。鉴于工资数据缺失形式的关键作用，我们就先从传导学习w_0入手。为此，我们采用 Ruben (1987) 的多重插补法 (multiple imputation) 来系统分析缺失数据的可行形式及其后果。本质上看，多重插补法就是利用Z_1与 Z_0之间的相似可比信息对每个缺失值的随机预测[19]。在多重插补法分析中，数据缺失机制被分为三个种类：完全随机缺失 (missing completely at random (MCAR)) 、随机缺失 (missing at random (MAR)) 和非随机缺失 (missing not at random (MNAR))。将通用的人力资源模型记为：

$$w_1 = f_{w_1}\left(Z_{w_1}; \boldsymbol{\alpha}_1\right) + \varepsilon_{w_1} \tag{4.3.2}$$

上述三种机制便可记为： (a) MCAR: $Pr(l_i|w_0, Z_w) = \Pr(l_i)$， (b) MAR: $Pr(l_i|w_0, Z_w) = \Pr(l_i|w_1, Z_w)$ 及 (c) MNAR: $Pr(l_i|w_0, Z_w) \neq \Pr(l_i|w_1, Z_w)$。不难看出，MNAR 涵盖了经济学有关非随机工资数据缺失的解释。对于 MNAR 情形，我们可以通过对多重插补值做敏感性分析来试验考察，其基本步骤如下：利用 (4.3.2) 在假定 MAR 情形下对 w_0做

[18] 详述见 Maddala (1983)、 Cameron and Trivedi (2005, Ch16) 及 Kennedy (2008, Ch17)。

[19] 有关多重插补法的详细讲解，可参见 Carpenter and Kenward (2013)。

多重插补，记为 w_0^{MAR}并将其与w_1串联，构成$w_{0+1}^{MAR}=\begin{pmatrix} w_0^{MAR} \\ w_1 \end{pmatrix}$。由于在 MAR 前提下，以下的 LFP 模型必然成立：

(4.3.3) $$l=f_l(Z,w_{0+1}^{MAR};\boldsymbol{\gamma}^{MAR})+\varepsilon_l=f_l(Z;\boldsymbol{\gamma}^{MAR})+\varepsilon_l \quad \text{with } \gamma_w^{MAR}=0,$$

我们便可以通过对w_0^{MAR}施加一定的系统位移，生成w_0^{MNAR}以及$w_{0+1}^{MNAR}=\begin{pmatrix} w_0^{MNAR} \\ w_1 \end{pmatrix}$，代入上式，使之被下式替代：

(4.3.3’) $$l=f_l(Z,w_{0+1}^{MNAR};\boldsymbol{\gamma}^{MNAR})+\varepsilon_l \quad \text{with } \gamma_w^{MNAR}\neq 0$$

依据先验知识设计一系列现实可行的系统位移，我们就可以通过上述敏感性分析模拟出一系列的w_0^{MNAR}。

得到了w_0^{MAR}和w_0^{MNAR}之后，我们就可以利用 (4.3.1) 对反事实的h_0做相应的多重插补。据多重插补分析法，这时的h_0属于随 w_0缺失的单调缺失[20]。同时据经济学常理，其缺失机制应为 MAR，即 $Pr(l_i|h_0,w,Z_h)=\Pr(l_i|h_1,w,Z_h)$ 。据多重插补分析法的推证，在研究对象的正确统计模型已知的前提条件下，只有在有数据缺失的变量是响应变量、而且其缺失机制是 MNAR 时，用普通最小二乘法回归无数据缺失子样本的参数估值才丧失一致性[21]。这意味着，(4.3.1) 中的$\boldsymbol{\beta}_1$是不受选择有偏性之扰的。选择有偏性仅在w_0^{MNAR}情形下会干扰 (4.3.2) 中的$\boldsymbol{\alpha}_1$。

不过，实际应用场景通常不满足正确统计模型已知这一前提条件。而且，我们所关注的现实问题，其实是采用无数据缺失子样本的劳动力供给模型对于特定反事实场景的参数不变性。将该不变性表述为估计量的一致收敛性是不准确的，因为一致收敛性隐含假定相关总体是先验确定已知的。使用多重插补法得到w_0^{MAR}和各种w_0^{MNAR}及其相对应的h_0^{MAR}之后，我们就有了一系列的无缺失数据的模拟子样本$X_{0,MI}$，这些样本构成我们推测反事实可行总体的疆界。利用这些$X_{0,MI}$回归 (4.3.1)，我们就可以通过常数不变性检验，即 H_0： $\boldsymbol{\beta}_1=\boldsymbol{\beta}_{0,MI}$来得出本题的经验答案。同理，我们也可通过回归 (4.3.2) 来检验 H_0：$\boldsymbol{\alpha}_1=\boldsymbol{\alpha}_{0,MI}$。上述 H_0 都应该属于预期的常态，除了后一检验用于$X_{0,MI}$中含有w_0^{MNAR}的情形。也就是说，选择有偏的推断仅适用于 (4.3.2) 在 MNAR 场景下得到 $\boldsymbol{\alpha}_1\neq\boldsymbol{\alpha}_{0,MI}$的检验结果。原假设被拒的其它反预期检验结果，特别是$\boldsymbol{\beta}_1\neq\boldsymbol{\beta}_{0,MI}$的结果，其导因只能是模型的泛化性不足，而不是选择有偏性。这一诊断通常有模型对两部分子样本回归残差呈现出的各种不满足统计理想分布的特征作为佐证[22]。

以上分析充分暴露了关于 (4.3.1) 选择有偏推理的认知极其模型构述之误导性。由于如何采用后验手段来模拟可行反事实总体特征是研究本例的首要和基本问题，这个问题的处理超出了有监督学习任务类。采用有监督学习模型类来处理问题的首要弱点，并不是估计量欠一致性或选择有偏性，而是由忽略Z_0中个体特征信息所致的欠有效性。

[20] 有关单调缺失的定义，详见 Carpenter and Kenward (2013, Ch3)。

[21] 详述见 Carpenter and Kenward (2013, Ch.1) 及 Carpenter and Smuk (2021)。

[22] 应用中，通常将 f_{h_1}、 f_{w_1}及 f_l 设定为参数线性型函数。有关本例的详细阐述及探索性实证案例分析，可参见 Qin *et al* (2019)。

这一点从使用项目评价模型测度政策评价处理效应的研究中可以更清晰地体会到。特定的政策目标相当于锁定了课题所关注的总体特征疆界。这时如何根据对照样本匹配原则后验选择"控制"人群组的问题，便成为估计政策处理效应模型研究之合理和有效性的关键前提。这种选择相当于根据政策目标来过滤本题中的 l_1，以选出与政策目标最匹配的l_0中的子样群。对照匹配的选择原则是等价于传导分析原则的。

案例二：用模型综合多项经济指标生成前导指数

在大约一个世纪的经验式商业周期研究中，发掘和构造可用于经济预测的前导指数一直是经济学家试图攻克的一项难题。在利用模型生成综合前导指数方面，经典的首创性研究有 Sargent and Sims (1977)和 Stock and Watson (1989)，其模型基础是统计因子分析法 (factor analysis)。遗憾的是，基于因子分析模型的综合指数的应用结果一直不理想，主要问题有：因子负荷系数普遍欠常数性；对建模关注目标的预测成效尚缺乏足够确凿的证据。

从机器学习的视角看，用因子分析模型构述综合前导指数的做法也犯了第三类错误，认知失误主要在两方面。其一，前导预测是指数构造的主旨，针对待预测的目标变量而言，该模型研究应属于有监督学习的范畴，而因子分析法原本属于无监督学习的变量降维范畴[23]。其二，从模型的因果关系构述角度看，因子分析法假定多个可测指标变量是其内涵潜变量、即共性因子的表象；也就是说，假定潜在共性因子是多个可测指标变量的共有自变量。而利用模型生成综合指数的议题属于综合指数的构造问题，可测的个体指标变量不是模型生成的综合指数的表象，即后者不是前者的共有自变量。

下面，我们就以金融状况指数 (financial conditions indicators (FCIs))为例，来讨论一下如何根据待预测的宏观变量来构述综合前导指数的测度模型问题[24]。设我们所关注的宏观目标为y_t，解释该变量的常规宏观模型为上章介绍的 ARDL 模型类，如：

$$y_t = \alpha_0 + \textstyle\sum_{i=1}^{n} \alpha_i y_{t-i} + \sum_{i=0}^{n} B_i X_{t-i} + e_t \tag{4.3.4}$$

式中的解释变量集$X = \{x_1, \cdots, x_k\}$ 包括传统的宏观金融状况变量，如利率及货币总量或其变动率等等，B是相应的参数向量。但是，由于这类变量来自银行系统，很可能对发达的金融市场全貌信息反映不足或者不及时。这一可能性假说相当于假定 (4.3.4) 存在遗漏前导金融潜变量问题。为了检验这一假说，我们设存在若干几个能够提高 (4.3.4) 之预测力的金融市场总体状况综合指数$F = \{f_1, \cdots, f_s\}$，亦即式 (4.3.4) 应被扩展为：

$$y_t = \alpha_0 + \textstyle\sum_{i=1}^{n} \alpha_i y_{t-i} + \sum_{i=0}^{n} B_i X_{t-i} + \sum_{i=1}^{n} \Gamma_i F_{t-i} + \varepsilon_t \tag{4.3.5}$$

[23] 因子分析法是与主成分分析法密切相关的，它们在机器学习教科书中都被归类于无监督学习的降维工具。而在测度理论中，基于因子分析和主成分分析的降维模型被解释为反映性测度模型 (Reflective Measurement Model)，用于刻化可观测的变量与它们共同反映的潜在自变量间的因果关系；构造综合指数的降维模型则被解释为形成性 (Formative) 或组合 (Composite) 测度模型，这类模型所需的降维标准要比反映性测度模型更复杂，通常要引入目标变量，参见 Markus and Borsboom (2013, Part II)。

[24] 有关本例的详细阐述和案例分析，可参见 Qin *et al* (2022) 和秦朵等 (2021)。

倘若通过模型选择约化之后，由(4.3.5)类得出的预测结果的确优于由(4.3.4)类得出的结果，那么上述假说则被验证。

上述模型框架把 FCI 的构造问题明确定位在有监督学习的变量降维范畴之内。令$\{I_{jt}, j = 1, \cdots, m\}$表示由各个金融市场搜集来的一个数量可观的可观测变量集，这些市场涵盖股票市场、债券市场、货币市场及衍生品市场。这时，我们就可根据"前导"的要求设定如下的偏回归测度模型类：

(4.3.6) $$y_t = \sum_{i=1}^{n} \omega_{ji} I_{jt-i} + \vartheta_{jt}, \;\; j = 1, \cdots, m$$

式中权数ω_{ji}由偏最小二乘法估计法得出。将得出的最初几个F用于(4.3.2)类，做上述两个模型的约化比较，以实现假说验证。这里必须强调，测度理论的一项基本准则是，任何综合指标必须满足时序上的毗连性(time-wise concatenable)[25]。这意味着，在从(4.3.6)选择约化出的具体测度模型过程中，必须做足够的递归估计试验及参数的常数性试验。

一旦采纳了有监督学习的建模框架，综合生成前导指数的构述设计就有了针对性，这也就增加了为具体问题具体设定的空间。这是基于因子分析法的 FCI 建模框架所无法相比的。这一模型构述个性化的特征可由动态特征设计上的两个侧面来展示。其一，为了体现来自不同金融市场可测变量间的动态非同步性，我们选择采用滞后分布模型形式作为测度模型类 (4.3.6)。其二，考虑到宏观目标变量与金融指数的动态特征的复杂匹配问题，我们可利用从 ARDL 模型类向 ECM 模型类的等价转换关系 (见第 3.3 节)，先将(4.3.5)转换为 ECM 类，再将有监督学习的目标分解为两个：一个为短期变量，另一个为长期非均衡复合变量。相应地，前导指数的构建也可以被分解为短期和长期两种，以便更好地满足对于数据变动之独立可分离性的设计理念。另外值得一提的是，当可测金融变量集包括大量来自临近地区或相似经济体同一金融市场的相似指标时，为了控制这些指标间的信息重复性，我们还可以在通过 (4.3.6) 做有监督学习的变量降维之前，增加一个聚合这些指标的非监督学习的降维步骤。

值得重复强调的是，上述两个案例中的各经济关系式都是模型类，从各类选择出具体模型函数式，无疑是实现理论假说验证目标的一个基本环节。但仅此一环节并不足以实现对所关注理论假说准确到位、且具有实用功效的模型构述。这两个案例不仅是对统计假设检验框架远不能胜任该构述任务的现身说法，也是对前几章中关键论点的现身说法。具体地，后验学习是构建实用并稳健模型的必备环节。只有缜密设计理论知识与数据信息得以密切相互作用的系统试验空间，通过进行大量的数据迭代试验，我们才有望找到最能满足 PAC 可学性标准的模型。鉴此，我们必须摒弃将计量学研究模式简单分为有理论测度与无理论测度的二分法传统理念。其实，早在近半个世纪前，Zellner 就提出过类似观点。在他 (1979) 综述有关计量学的因果关系研究中的最后结论就是："'有理论测度'与'无理论测度' 属于必须要避免的两种极端观念"。从目前通用的理论模型和经验模型划分标准看，上述极端观念是建立在简单地依照模型的构述形式来区分模型种类的。凡单纯靠先验推导演绎得出的模型就赋予"理论"标签，而需要经后验数据分析得出的模型便只能是"经验"模型。这种分类理念的明显落伍程度

[25] 详述见 Markus and Borsboom (2013, Ch2)。

已被人工智能和机器学习的崛起而暴露无遗。从 PAC 可学性的视角看，后验学习是克服纯先验推理模型内在的脆弱性和对实际目标缺乏针对性的唯一可行途径。同时，由于初始模型类的选择大都与变量的随机分布函数无关，在模型构建过程中假设所涉变量是取自某随机分布空间的传统主流做法，其实是一个冗余环节。概率测度的主要和第一功能是作为协助模型选择学习过程的诊断决策工具。另外，有关可解释参数的统计推断任务，在模型学习和选择过程中，应被排序为一个相对后续的次要环节。

第五章 估计方法的问题与潜能

如何对先验理论模型中给定的结构参数做出最优估计？这是经济计量学中的核心问题。这一问题的原始定位，便是哈维尔莫对内生有偏性的推证。继哈维尔莫推证之后，学界内对各种参数形式的最优估计问题的集体讨论和研究扩展，使得该问题的核心位置在现今学科中稳固无疑。Athley and Imbens (2019, 第 2.1 节) 一文中对"真实值的估计 (estimand)"概念的讨论，是反映这一现状的一例。有必要重申，对于最优估计方法的甄别，是以先验给定理论模型为普遍正确的模型为假定前提条件的。前面的章节已充分揭示，无论理论模型的先验推导过程多么严格和复杂精妙，该假定前提都无法成立。一旦学界公认这一事实，摒弃该假定前提，最优估计核心位置就必然被动摇。

充分揭示以参数最优估计为核心的研究方法缺陷是本章的首要任务。统计学从简单回归模型场景出发，把普通最小二乘法 (OLS) 作为通用的最优估计方法。在计量学中，有关 OLS 丧失最优性的数学推证，其理论模型场景都超出了简单回归模型的场景。推证的直接依据则是模型之残差项不能满足理想的经典统计假设条件。我们在上章业已阐明，由模型残差测出的各种统计上欠理想的表征，大都反映着先验模型构述不足的问题。这时将模型的普遍正确性置于无可置疑的前提之上，由模型残差暴露出的问题便只有被推论为估计方法欠最优的问题了。不幸的是，采用改变估计方法来诊治残差问题，是一个治标不治本的研究策略。它不仅会掩盖先验模型构述不足之本，还会隐性更改模型所表述的理论假说之本意。另外，估计结果还可能使模型的拟合度或泛化度比不改变方法的结果更为糟糕。因此，以估计方法的选择作为核心的研究策略，实质上是一个粉饰模型构述问题、为先验构述不足的模型系统提供虚谬实证的策略。我们在 5.1 节沿用第 4.1 节的两个例子来说明这一点。这两例的讨论表明，在消除先验模型中的各种构述不足问题之前，考察参数最优估计问题的条件尚未成熟、为时过早。

一旦计量学把通过统计学习来选择模型的研究任务置于首位，辅助模型选择就成为估计的首要职能。讨论这一职能便是 5.2 节的主题。在模型学习过程中，估计作为统计学习算法的一部分，起着从样本中提取可用于评价模型功能的间接数据证据的作用。由于结构风险最小化是模型学习的基本目标，估计法的通用选择标准便是对目标损失函数的最小化。将这一选择标准数学正规化，就形成了一类增广估计量。增广估

计量中增加的是一附加约束条件：即体现模型选择目标的正则化 (regularized) 损失函数。必须明确，从应用的角度出发，我们需从认知上将这类估计量与为统计推断服务的传统估计量区别看待。这是因为处理模型学习问题的关键数据场景是测试样本，而经典统计学中的估计问题是建立在样本与总体二分法的设计理念之上的，没有考虑样本测试这一环节。

第 5.3 节的议题是模型选择之后的参数推断问题。由于在有序的模型学习过程中，从残差项暴露出来的问题大都得到解决，目前教科书中广泛讨论的各种复杂估计问题便不复存在了。这使选择参数之最优估计量的任务简化了许多。相比之下，更重要和艰巨的任务是模型的再参数化，亦即如何通过模型变形的设计实现如下双重目标：(a) 输入因素相对简洁和独立可分离性；(b) 参数的理论解释可信赖化。前面的第 3.3 节已经在特征设计或表述学习的议题下讨论了相关任务。这里需明确的是，模型选择之后的参数推断是统计学习与经典统计学的交接点。目前，统计学和机器学习学界就经典统计学工具对于统计学习得来的模型之适用性问题的讨论仍未达成共识。本节将简单概述有关讨论要点。总之，本章进一步强化了上章的结论：针对经济学假说的实证分析研究的任务而言，经典统计学的框架实在过于狭窄和简单，无法作为有效完成任务的核心工具源。

5.1 先于模型选择的参数估计问题

第四章已阐明，实证经济学问题的研究场景与经典统计学的随机试验控制场景之间有着不可逾越的鸿沟。建立在实证性统计学框架内的经济计量学，是以先验理论模型的常态正确性为起点的。当模型残差显露出该鸿沟的任意征兆时，用改变估计方法来诊治征兆、填补鸿沟便成了计量学的唯一方略。这些征兆有的是从先验推理中发现的，有的是从后验试验中发现的。由于计量学从统计学引进的模型以基于 OLS 的回归模型为主体，计量学对估计问题的讨论便集中体现在对 OLS 估计量缺乏最优性质的推论上了。研制和使用可矫正 OLS 的欠最优性的其它估计方法，就成为经济计量学的核心关注点。不幸的是，这种以估计方法为核心的研究策略犯有方向性错误。本节通过延续讨论上章 4.1 节中的两个例子来揭示其认知陷阱。

例一：由先验推证的内生有偏性而对 OLS 欠一致性的推论

前面已经提到，上述推论以及诊治思路源于哈维尔莫对联立方程模型估计问题的讨论。后续的效仿研究将他的讨论思路推衍到了单方程模型中。于是，内生有偏性便成为对遗漏变量有偏问题以及样本的总体代表有偏性问题的综合诊断。由于这类问题是单方程模型研究中普遍关注的问题，使得学界内充斥了对内生有偏性的恐惧和认知混淆。同时，通用诊治该有偏性的工具变量法的普及又助长了 p 值操纵的盛行。

为了明示诊治内生有偏性的认知缺陷根源，我们采用机器学习中将误差项分记为样本内 ε_{in}与样本外 ε_{out} 的做法。现沿用上章的二元联立模型 (4.1.1)，从模型中的第一式来考察内生有偏性。将该式的 OLS 估计记为：

(5.1.1) $$y = \hat{\beta}_1 x + \varepsilon_1, \qquad \hat{\beta}_1 = (X'X)^{-1}X'Y$$

模型 (4.1.1) 中的y 与 x 的联立设定意味着 $cov(x\varepsilon_{1,out}) \neq 0$，但是 OLS 估计的前提条件是$corr(x\varepsilon_{1,in}) = 0$。因此，在未做数据分析前就不难发现，$\hat{\beta}_1$与模型的联立设定不相匹配。对于 OLS 这种欠一致性的推理目前被广泛沿用到单方程模型的情形。这是因为，单方程模型的因果关注点通常是类似 (5.1.1) 中的$E(y|x)$ 这种高度局部的均衡理论，而现实中经济变量间的相互影响又是普遍存在的。这意味着，在被动观测的数据面前，遗漏变量问题是不可避免的。现将遗漏变量记为 z。这时若先验知识不具排除 $cov(xz) \neq 0$ 可能性的理由，研究者就必须考虑应对遗漏变量有偏性的措施。从 (5.1.1) 出发，套用哈维尔莫的分析思路就不难得出，遗漏变量有偏性同样会导致 $corr(x\varepsilon_{1,out}) \neq 0$。于是，OLS 欠一致性的逻辑推证便在计量学中被广义化，使之在学界内声名狼藉。

计量学教科书推广的诊治内生有偏性的处方是用工具变量(IV)估计法替代 OLS。设V为一个 IV 集，我们可以把 IV 估计量表示为对应于(5.1.1)中 OLS 的广义最小二乘（GLS）估计量：

(5.1.2) $$\hat{\beta}_1^{IV} = (X'V(V'V)^{-1}V'X)^{-1}X'V(V'V)^{-1}V'Y$$

表述 (5.1.2) 的另一种方式是两阶段最小二乘法（2SLS）。第一阶段是一个基于$E(x|V)$的回归模型[26]，第二阶段是用变量$x^V \equiv E(x|V)$来代替 (5.1.1) 中的x，亦即：

(5.1.3) $$y = \hat{\beta}_1^{IV} x^V + \varepsilon^V, \qquad \hat{\beta}_1^{IV} = (X^{V'}X^V)^{-1}X^{V'}Y$$

$\hat{\beta}_1^{IV} \neq \hat{\beta}_1$便是学界广泛认可的内生有偏性的经验证据。但是，在工具变量选择问题上，以及连带有关 $\hat{\beta}_1^{IV}$的经济解释方面，学界中的争论迟久不休。追根求源，争论不休不无道理。先从 IV 的选择看，为获取 $\hat{\beta}_1^{IV} \neq \hat{\beta}_1$，就必须先有$x^V \not\approx x$。这意味着在 2SLS 中第一阶段的回归时，必须放弃通用的对被解释变量 x 实现最优拟合逼近的目标。也就是说，为了得到$\hat{\beta}_1^{IV} \neq \hat{\beta}_1$，就必须容许在第一阶段采用非最优回归的选择规则，这就决定了 IV 选择的非唯一性。再从$\hat{\beta}^{IV} \neq \hat{\beta}_1$对$x^V \not\approx x$的依存关系看，选用 IV 估计法实质上是对模型所关注的自变量做了隐性修改，即用 IV 生成的变量x^V替换掉理论假说中被先验判定为“内生”的条件变量x。从变量修改的角度看，我们可以把 (5.1.2) 中的 GLS 看作是由 IV 对x做了某种加权组合后的 OLS；于是，选择 IV 估计量而不用 OLS，就等价于选择模型 (5.1.3) 而不用 (5.1.1)。从这两个模型选择的视角看问题，我们就能明白为什么关于 IV 的选择争论会涉及到模型解释的层面了。在$x^V \not\approx x$的前提下，(5.1.1) 与 (5.1.3) 属于两个互不嵌套的条件对立模型。由于从(5.1.3)生成x^V的 IV 组合多种多样，若无视其刻化的$x^V \longrightarrow y$之因果关系，把β^{IV}作为测度(5.1.1)中的$x \longrightarrow y$之因果关系的效应的解释，则必然既欠缺明确性又匮乏可信度。从这两个非嵌套对立模型的角度看，IV 估计量的选择相当于拒绝了(5.1.1)中的$E(y|x)$，而支持(5.1.3)中的$E(y|x^V)$。换言之，IV 估计法的实质就是通过修改研究者所关注的条件变量，来治理先

[26] 有必要指出，将 IV 估计量描述为链式法则，即将 2SLS 表示为$V \rightarrow x \rightarrow y$的形式，是逻辑有误的。关于 IV 估计和内生有偏性更详细的讨论可见 Qin (2015; 2019)。

验得出的 $corr(x\varepsilon_{1,out}) \neq 0$ 的推论。而正是由于 IV 修改方案存在多种可能性，将修改的模型完全等同于原理论假说模型的解释法，显然是不能说服所有研究者的[27]。

上述困境促使我们重新审视 $corr\left(x\varepsilon_{1,out}\right) \neq 0$ 这一推论，特别是支撑该推论的假定前提：即给定理论模型在常态下的普遍正确性。前面的分析已经阐明和重申了该假定前提普遍不成立的现实。这里来从两个方面进一步揭示，内生有偏推论所基于的特定模型构述是多么不切实际的构述。第一个方面是有关对变量间相互影响关系的联立模型构述问题，第二个方面是有关模型残差项的本质与作用问题。首先，完全对称的联立模型构述是统计上无法操作的，凡统计上可操作的模型必须是变量关系非对称的条件模型[28]，这点在第三章业已阐述过了。对现实中变量间相互影响关系逼真的模型构述，必须充分体现出变量间复杂动态关系的特征。纯静态的联立模型，如 (4.1.1)，不仅缺少对这种起码的动态特征的构述，而且是不可操作的。当将静态联立模型动态扩展，使其足以反映现实数据中的动态特征，这时由联立设定所致的内生有偏性便会降低到可被忽略不记的程度。这一结论的正式表述是 Wold’s 多年前所推证的“近似定理” (proximity theorem) (见 Wold and Juréen, 1953, 第 37-8 页) 。定理中采用残差分布满足白噪声来表述模型足以反映数据中的动态特征这一条件。值得注意，在 Cox 的书中 (2006, 附录 B) ，类似的先决条件被用作描述模型一致性的准则。机器学习中有关模型一致性与可学性的讨论，则可见 Shalev-Shwartz and Ben-David (2014, Section 7.4)，以及前面第二章中的有关简述。对模型做一致性的评判，实质上是强调了一个基本逻辑排序：在考虑参数的估计方法是否最优（包括是否与模型设定一致）的问题之前，必须首先具备已被验证是具有泛化能力的模型。

模型泛化能力的验证必然涉及到模型残差项的本质与作用。我们来看看这一方面的认知问题。在开放世界的场景下，认知的关键是将误差项ε_{in}和ε_{out}两者间的本质差别区分开来。ε_{in}是在给定数据样本后由模型生成的，除了所设模型之外，其“未知”的随机统计特性还取决于所选的估计方法。因此，ε_{in}属于“已知的未知量 (*known* unknowns)”。相比之下，ε_{out}则是“未知的未知量 (*unknown* unknowns)”。即使有了数据样本，并且所选模型是综括先验知识和后验数据信息的有效模型，我们也无法对ε_{out}的统计特性做出确定无疑的预判。正因如此，机器学习才把对E_{out}不确定性的考察作为考察模型泛化性能的重心。这从第 2.2 节简述的 PAC 可学性，以及将现有数据划分为训练样本和测试样本、用后者模拟分析ε_{out}的各种交叉测试技术都能体会到。不幸的是，以经典统计学为基础的计量学没有区分这两种误差，因此不倡导做训练样本和测试样本的划分试验。而能够彻底暴露工具变量法之伪一致性的有效途径，正是基于

[27] Wold 是最早指出内生有偏性的本质不是估计问题、而是因果模型构述问题的学者之一，如见 Wold (1954; 1956)。他还对纯靠演绎推理建模的认知路径的不足之处提出质疑，如见 Strotz and Wold (1960)。在微观计量学中，对 IV 路径改变变量涵义的最明显认同的一例，便是在项目评价模型中将表示“平均处理效应”的参数重新解释为“局部平均处理效应”的做法，如见 Angrist *et al* (1996)。这里，“局部”一词显然是对先验定义的理论关系之普遍存在的重大解释变更。

[28] 在经济计量学的正规化过程中，评价联立模型是否具备统计上可操作性的条件被统称为“识别条件”。值得注意的是，“识别”这一用词含有通过数据试验辨别模型是否符合现实的内涵，而计量学中的识别条件却与该内涵毫无关系，有关讨论可参见 Qin (1993, Ch4)。

上述划分的交叉测试结果，亦即模型预测检验结果。早在 VAR 动态模型崛起之前，宏观计量模型研究中的大量反复模型预测检验结果就已表明，采用 IV 估计法的联立模型的预测检验结果要明显劣于采用 OLS 估计法的结果[29]。就基于横截面数据样本的模型而言，Young (2017) 及 van Huellen and Qin (2019) 采用交叉验证手段来考察 IV 统计量的“样本外”渐近收敛性，所得结果都是对先验断言的 IV 一致性的证伪。他们的试验表明，模型 (5.1.3) 的样本外渐近收敛性普遍要比 (5.1.1) 差得多，即前者的泛化性能比后者高。这意味着，x 通常要比那些由 IV 组合生成的替代变量更具备做条件自变量的资格。内生有偏性不过是基于既先验构述不足又得不到后验测试验证的模型之上的臆断，不应被作为随意修改经济学家根据先验知识提出的因果关系中的自变量定义的理由。

例二：在给定模型中，据由时序数据后验得出$\widehat{\varepsilon_{in}}$的自相关对 OLS 欠有效性的诊断

教科书对上述欠有效性的诊治策略是将估计得出的$\widehat{\varepsilon_{in}}$自相关系数作为权数，形成 GLS 来取代 OLS。以最简单的二元静态模型误差项一阶自相关为例，则有：

(5.1.4) $$y_t = \beta x_t + \varepsilon_t,\ \varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$$

β的 OLS 估计如(5.1.4)所示；它的 GLS 估计可记为：$\hat{\beta}^{\Omega} = (X'\Omega(\rho)^{-1}X)^{-1}X'\Omega(\rho)^{-1}Y$，其中的$\Omega(\rho)$表示基于$\rho$的加权矩阵[30]。不难推导，GLS 相对于 OLS 的有效性改善，其实源于该估计方法对(5.1.4)中静态模型的隐性动态扩展。不过，该隐性扩展是附带严格参数约束的。由 (5.1.4) 可推出：$\varepsilon_t = y_t - \beta x_t$；将其代入静态模型便可得出：

(5.1.5) $$(y_t - \rho y_{t-1}) = \beta(x_t - \rho x_{t-1}) + \nu_t$$

也就是说，ρ的作用是对两个变量的动态形式隐性施加了一个共享特征约束，因此通称 (5.1.5) 为“共因子”模型[31]。对 (5.1.5) 做参数变换，把它转换为一个误差修正模型：

(5.1.6) $$\Delta y_t = \beta\Delta x_t + (\rho - 1)[y - \beta x]_{t-1} + \nu_t$$

我们就可发现，共因子的约束相当于约束x_t 对于 y_t的长期和短期效应是大小相同的。大量应用动态建模研究的结果表明，这样强的一个约束通常不吻合经济时序变量间的动态特征。这些动态特征往往需要采用更具一般性的动态模型来刻化。这也是第三章所述的 VAR 模型类在宏观经济学兴起的一个重要因素。

VAR 模型类在宏观应用研究中的普及充分表明，后验观测到的残差项自相关现象实为模型动态设定不足所致。这种不足也属于一种遗漏变量有偏问题，被遗漏的是与现期变量相关的滞后或延迟变量，反映着时序变量的动态惯性，因此对于我们测度变

[29] 详述可见 Qin (2013a, Ch1)。

[30] 这种估计方法最初是由 Orcutt (1948)提出的。鉴于一阶残差项自回归的估计法源于 Cochrane and Orcutt （1949)一文，教科书中通常把该方法称为 Cochrane-Orcutt 估计步骤。Malinvaud (1966)是将该步骤推广介绍为一般的 GLS 估计量的早期教科书。

[31] 详述可见 Hendry (1995, Ch7) 或韩德瑞和秦朵 (1998, 第 7 章)。

量间长期均衡关系起着举足轻重的作用。而仅凭先验数学推导是无法准确得出与数据相吻合的动态模型的，这一点在前面章节已经讨论过了。不幸的是，从采用 GLS 估计法处理 (5.1.4) 的视角出发，就会引致把残差项解释为对先验模型的动态冲击扰动变量的认知。这种认知把先验构述不足的模型经后验样本得出的$\widehat{\varepsilon_{in}}$自相关症状泛演成为$\varepsilon_{out}$的广义属性，并把$\varepsilon_{out}$视作一个动态扰动潜变量引入联立模型，作为解释目标变量的基础动态源，如见第 3.3 节所述的 DSGE 模型例子。这一认知陷阱再次暴露了将验证性统计学直接套用在处理实证经济学研究问题之上的方法论弊端。

上述两例的讨论充分表明，由模型残差项诊断出来的非统计最优症状，其实反映着先验模型构述本身的缺陷。这时，在原模型基础上采用改变估计方法来诊治这些症状，其实是一种隐性改变原模型构述的方法。由于被改变的通常涉及所关注参数，从而引发有关结构参数的多重定义问题[32]。这种以改变估计量为核心的策略不仅会造成学界就模型解释问题的争论不休，还很可能会导致模型泛化功能的退化。走出上述困境的唯一可行途径是，正视先验模型普遍存在的构述不足问题，遵循 PAC 可学性原理，系统解决由残差项显露出来的建模问题。

5.2 以模型选择为职能的估计法

就选择泛化性能相对最优的模型学习任务而言，估计的职责首先是如何尽可能有效地协助该学习任务，而不是如何对单个参数做最优估计推断。模型学习任务的基本目标是结构风险最小化。按机器学习的分类，实证经济学研究的问题大都可被归为有监督的学习问题，其基本行为规则可用线性模型类中的凸函数形式来构述，通过对某个线性预测模型的二次损失函数求最小化来实现。如前所述，最小二乘法是实现该最小化的基本法则[33]。再次强调，这里采用的最小二乘法是以判别式建模任务为出发点的。因此，我们应该从最优控制算法的视角来认识和评判该估计法，而不是从统计推断的尺度来认识或评判它。换句话说，统计推断不仅不是评判最小二乘法的唯一尺度，也不是针对模型学习任务来评判最小二乘法的适当尺度。认清由于不同职能引致的不同评判标准，对于我们正确理解在建模过程中由最小二乘法引发的估计方法扩展尤为重要。下面我们就从两方面细述这一点：一是对传统估计量与检验方法结合使用的方面，二是对传统估计量的数学扩展方面。

在执行模型选择任务的算法中，估计步骤通常是与其他检验及评判步骤结合并且迭代使用的。其大致流程是：针对一个初选的模型类，采用以样本内二次损失函数的最小化为目标的估计法进行拟合，然后为模型筛选对估计结果做评判。评判中最为关键的基本准则是估计样本外的误差最小化。第二章业已阐明，该准则体现了回归模型的 VC 泛化界限。将多变量的回归模型h记为：

(5.2.1) $$h: y = X_d\beta + \varepsilon,$$

[32] 详述见 Qin (2013a, Ch7)。

[33] 对于采用对数函数形式的离散选择模型而言，相应的损失函数最小化问题是与似然函数的最大化等价的。从决策理论出发，该最大化问题也可以从求解凸优化问题导出，无需任何统计分布前提假设，如参见 Abu-Mostafa et al (2012, Ch3)、及 Shalev-Shwartz and Ben-David (2014, Ch9)。

其中的 d 表示变量集 X 的维度。模型h的 VC 泛化界限可由下式表出[34]:

$$E[\varepsilon^2]_{out}(h) = E[\varepsilon^2]_{in}(h) + O\left(\frac{d}{N}\right), \tag{5.2.2}$$

式中的$E[\varepsilon^2]_{out}(h)$和$E[\varepsilon^2]_{in}(h)$分别表示估计样本外和样本内的均方误差，末项$O(\cdot)$则表示该项的绝对值在数量级上渐近小于包含 $\frac{d}{N}$ 的某乘数，该比率中的 N 表示样本量。不难看出，对泛化误差 $[E[\varepsilon^2]_{out}(h) - E[\varepsilon^2]_{in}(h)]$ 的最小化目标而言，比率 $\frac{d}{N}$ 起着关键作用。例如，对于一组$E[\varepsilon^2]_{in}(h)$相同的模型来说，其中哪个模型的 $\frac{d}{N}$ 越小，它的泛化误差也就越小。可见，上述泛化界限其实是对尽量寻求简洁模型的传统做法的抽象概括。回顾第四章提及的 AIC 及 BIC 等信息准则，这些统计准则都是通过与 $\frac{d}{N}$ 比率相关的量来惩罚模型的复杂程度的。上述泛化界限也从理论上表明了，为什么各种信息准则通常是探索式建模者过滤评判模型估计结果的初选。在多变量模型选择的实践中，采用信息准则控制模型规模，大致等同于应用研究中尽量删除系数估值不显著的输入变量的做法。

从数学表达上，我们也可以把上述的模型规模过滤评判步骤与估计法合二为一。基本的思路是，把寻求泛化误差最小化的目标作为一条附加约束条件，包括进估计量来，亦即将$E[\varepsilon^2]_{in}(h)$最小化原则约束于$E[\varepsilon^2]_{out}(h)$最小化的条件之下，生成一扩展的估计量。此举的数学术语是对损失函数最小化的“正则化”，或正则化的损失函数最小化 (RLM)。现将在回归模型h的情形下，以其损失函数最小化为目标的 OLS 记为:

$$\hat{\beta}_{OLS} = \text{argmin}\{\|y - X\beta\|^2\} = (X'X)^{-1}X'y \tag{5.2.3}$$

其中的argmin{·}表示，β是由二次范数 $\|y - X\beta\|^2$ 的最小化求解得出的。该二阶范数即h的二次损失函数，也是模型的残差平方和。显然，式(5.2.3) 仅以$E[\varepsilon^2]_{in}(h)$为最小化目标。为了将$E[\varepsilon^2]_{out}(h)$最小化的目标也考虑进来，我们可以在式中的argmin{·}加一个二次项，作为对β做正则化的过滤:

$$\hat{\beta}_{RLS} = \text{argmin}\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\} = (X'X + \lambda D)^{-1}X'y \tag{5.2.4}$$

上式中的λ为一个调节参数，有$\lambda > 0$；D为对角阵。从比较 (5.2.3) 与 (5.2.4) 的视角看，我们可把$\hat{\beta}_{OLS}$视为$\lambda = 0$时$\hat{\beta}_{RLS}$的特例。统计学中，通常将遵循(5.2.4)中优化原则得出的估计法称为岭估计 (ridge estimation)。让我们来考察上式argmin{·}中的两个二次范数。第一个范数代表估计偏差、第二个代表估计方差。从统计学习理论中所强调的偏差-方差权衡关系的视角看，$E[\varepsilon^2]_{out}(h)$ 的最小化目标，可由调节参数 λ 从而调节偏差与方差权衡关系来实现。具体地，不断调节λ来减小$\lambda\|\beta\|^2$，直到其下降已经不能抵消偏差项的相应增大为止。通过权衡偏差与方差两项来尽量减小$E[\varepsilon^2]_{out}(h)$，实质上就是通过尽量控制缩小模型规模来优化模型选择。我们可以从比较两个规模不同的模型来明晰这一点。设有两个主体部分为$y = X_p\beta$ 和 $y = X_d\beta$的模型，其中 $p > d$。显然，$\sum_{i=1}^{p}\beta_i^2 > \sum_{i=1}^{d}\beta_i^2$。在其他条件不变的前提下，$\lambda\|\beta\|^2$对于规模相对大的模型的正

[34] 有关 VC 泛化界限的具体讨论，可参见 Abu-Mostafa *et al* (2012, Ch3)、Hastie *et al* (2009, Ch7.9) 以及 Shalev-Shwartz and Ben-David (2014, Ch11)。

则化约束就相对更强，即使表面看来该项对模型规模没有做明确的削减。正因如此，通过调节λ来控制模型规模的方法被称为软截止阈值法，而把前述根据一定的信息准则来选定模型规模 d 的方法称为硬截止阈值法。

如果我们把(5.2.4)中正则化一项的范数设定改为绝对值函数，通过正则化寻求$E[\varepsilon^2]_{out}(h)$的最小化与控制模型规模之间的密切关系就更为直接明了了：

(5.2.5) $\hat{\beta}_{Lasso} = \text{argmin}\{\|y - X\beta\|^2 + \lambda\sum|\beta|\}$，或在$\sum|\beta| \leq C$条件下求 $\min_{\beta}\{RSS\}$

式中的 RSS 表示残差平方和。通称由 (5.2.5) 得出的估计法为套索 (lasso) 回归。由于$\lambda\sum|\beta|$等价于设定一个阈值约束：$\sum|\beta| \leq C$，我们就可把前述根据信息准则来选定 d 的硬截止阈值法视为套索回归的一个特例，如选定C为排除所有统计不显著参数估值的截止阈值。显然，C的选值越大，模型过拟合的风险就会增大；而C的选值越小，模型欠拟合的风险就会上升。可见，通过套索回归实现正则化，只不过是从数学形式上，把前述样本内估计后利用信息准则对模型做过滤评判的步骤并入了估计公式罢了[35]。

在上述估计式中，我们对参数λ或者C的校准选择是利用正则化成功选取最优模型的关键。校准参数的目的是，通过平衡尽量降低经验风险和模型复杂度两个目标，来找到既不过拟合又不欠拟合的模型。而这一平衡选择的前提是对数据样本做训练样本与测试样本的划分。也就是说，校准参数隐含着交叉验证的步骤。因此，RLM 是与第二章上述结构风险最小化原则相一致的。第二章已经提过，寻求偏差-方差关系之最优权衡点其实是等价于寻求拟合-稳定性关系之最优权衡点的。相应地，校准参数选择成功与否的一条关键条件是模型在估计样本外的稳定性[36]。

无论是对传统估计量与检验方法的结合使用，还是对传统估计量的数学扩展，以模型学习为目标的估计功能都远远超出了验证性统计推断的框架。若将这种以模型泛化或预测能力为主旨的估计路径与为枚举类研究课题目标服务的传统估计路径混为一谈，就犯了严重的认知错误。这里特别需要重申的是，基于 RLM 的估计量只是从数学上对模型选择过程中必须综合考虑E_{in}与E_{out}这一基本条件的高度精炼表达形式。因此，我们切不可用评判传统最优估计量的通用标准来评判包含正则化原则的估计量[37]。

上述对 RLM 法则的诠释也有助于我们进一步认识伦敦经济学院经济计量学派所推举的简洁包容原则[38]。在第三章描述的基于多变量误差修正模型所得出的长期关系

[35] 机器学习教科书大都对正则化有详细讲解，如见 Abu-Mostafa *et al* (2012, Ch4)， James *et al* (2013, Ch6)，Efron and Hastie (2016, Chs 7 and 16)。值得提及的是，Efron and Hastie (2016)一书是从 James-Stein 定理的角度来描述岭回归的。而该定理是在不含任何分布假定前提的决策论基础上，证明岭回归要优于 ML 估计量的。

[36] 详述可见 Mukherjee *et al* (2006) 以及 Shalev-Shwartz and Ben-David (2014, Ch13)。

[37] 有关这种认知错误的一个明显例子是用评判传统估计量无偏性的标准来评判 $\hat{\beta}_{RLS}$。显然，当(5.2.3) 中的估计量满足无偏性时， (5.2.4)中的估计量必然有偏，因此后者不是传统意义下的最优估计法。这种评判理念应当是岭估计法一直不受经济计量学界青睐的一个重要原因。

[38] 详述见 Hendry (1995, Ch14)、或韩德瑞和秦朵 (1998, 第 14 章)。

参数估值的成功案例，便是简洁包容原则的一个最好佐证。不难看到，简洁包容原则其实就是通过寻求拟合-稳定性关系之最优权衡点来选择模型的。虽然简洁包容原则未曾明确提及统计学习建模的概念，它对模型选择需要包容和超越现有对立模型的要求本身就体现着统计学习建模的主旨。在寻求简洁包容模型过程中，估计与检验方法交叉并用。为了描述这一并用特征， Trivedi (1984)构撰了一个合并词—"检验估计"(testimation)。鉴于"检验估计"强调了估计法在模型选择过程中的辅助功能，我们也许该把基于 RLM 法则的估计量，如$\hat{\beta}_{RLS}$ 及 $\hat{\beta}_{Lasso}$，统称为"检验估计" 更为贴切。

5.3 模型选择之后的估计推断

一旦模型学习阶段结束，有了与数据吻合的简洁模型，模型内参数的最优估计推断问题便升到议事日程上来了。本节将这个问题分为两个方面来讨论：对学习所选模型的再参数化以及参数化后做估计推断所需注意考虑的问题。

对于第一个方面的讨论，这里沿用第二章的式 (2.1.1) 与 (2.1.2) 所勾画的场景，即先验理论假说模型要比后验模型更简单的情形。第三章业已提及，由于被关注的经济变量大都是流量型或存量型变量，先验理论构述它们之间的因果关系又通常是某种局部均衡关系，因此模型中的有关参数一般缺乏从数据变动中独立可分离的性质。在计量学中，若从(2.1.1) 的视角出发，上述现象一般被描述为x_i的多重共线问题。而若从(2.1.1) 与 (2.1.2) 中两类变量的关系视角来看，它又被称为x_i相对于 z_i的遗漏变量有偏问题。鉴于模型学习的结果一般克服了遗漏变量有偏问题，所选模型属于类似 (2.1.2) 的情形。于是，如何最大限度克服模型中的共线性问题，便是需对所选模型做再参数化的主要缘由。第 3.3 节中的由 ARDL 模型类向 ECM 的再参数化的例子便是在使用时序数据进行模型学习的一个最好范例。一旦从 ARDL 学习约化到模型的具体动态形式，我们就能通过向 ECM 的参数转化尽量减小 ARDL 模型中由动态特征引致的变量间共线性。有必要重申，当理论因果关系含多个输入变量时，即由 ECM 再参数化得出的长期项内含多个x_i时，它们参数间的显著共线性通常是无法避免的了。这时需要采用对长期参数的校准试验来完成对 ECM 的估计，得到更适合经济解释的模型参数。这也是为什么校准试验法在 DSGE 模型研究中占据重要位置的原因。第 3.3 节还指出，针对横截面数据的模型学习任务更为复杂，因此更有必要借助机器学习的各种手段和分析思路。在机器学习中，有关模型参数学习的研究讨论被归入"特征设计"和"特征表示学习"的范畴。值得提及的是，哈维尔莫在他第二章专节讨论的"自律性"概念终于在有关特征表示学习过程中得到了可实验操作的具体表达。显然，专业知识是完成特征学习所必不可少的要素。由于经济计量模型参数学习需要综合考虑经济学与统计学两方面的要求标准，因此这一任务应该是建模研究中最具挑战性的一环了，只有应用经济学家才可能胜任这一挑战。任务的完成往往需要对算法的不断调整和多次的迭代试验，需要经济学家反复判别和比较不同再参数化形式，从中选出最能如实准确地将先验知识转化为实验可测的模型参数形式。可见，特征学习其实是应用经济学家建模研究任务的重心。

在完成了再参数化或者特征学习之后，对于那些不受共线性严重困扰的参数来说，我们就能考虑参数的最优估计问题了。由于可能从残差项显露出来的诸多问题在模型选择过程中都被排除和解决了，这将大大减少目前教科书中所列举的那些可能影响参数最优估计的因素，简化许多有关估计量选择问题讨论的必要性。对于构述充分的回归模型来说，最小二乘估计是等价于极大似然估计的。极大似然估计原理则是从无分布假设的数据学习过渡到以分布假设为前提的经典统计推断的关键连接点[39]。

实践中引起争议的一个主要问题是：经典统计学中的置信区间估计法是否能适用于经后验选择的模型？或者说，由经典统计学选择控制第一类错误的 *p*-值得出的置信区间对于选择后的模型是否仍保持其渐近一致性？争议的实质是质疑测度参数估值不确定性之概率空间的定界。追根求源，与争议密切相关的是在机器学习崛起之前，学界中长期存在着的对于探索性数据发掘研究可靠度的怀疑和偏见。这些怀疑和偏见牵扯到在数据发掘中一个最忌讳易犯的错误：双重浸渍(double-dipping)，即不恰当地重复使用数据证据、夸大估计结果成效的错误。由于模型选择过程已经参考使用了参数估值的信息，那么在选定模型后，若仍使用传统的参数估值置信区间，显然是对数据证据的重复使用。为了防范双重浸渍，数据发掘研究者大都采用未被“浸渍”的样本来计算参数估值的置信区间。这种做法的基准其实就是交叉验证法基于的数据分割原则[40]。必须看到，模型选择过程也使用了基于数据分割的交叉验证法。在许多数据量有限的情形下，为了统计推断而留用足够的未被“浸渍”的子样本是难以实现的。不过在开放世界场景下，单凭先验演绎推导的模型一般是不具数据相合性的。与这些模型相比，根据 PAC 可学性原理、通过交叉验证学习来的泛化性能相对最优的模型，应当是最为接近经典统计推断所要求的模型条件的，即模型中待检验推断的零假设应是常态下最可能被验证的假设。鉴此，所选模型也就应相对最满足传统置信区间应用的前提条件[41]。这时，即使将已被模型选择过程 “浸渍”的样本用作参数估值的做法犯了双重浸渍，但与那些未经模型选择筛选的欠拟合或过拟合模型得出的参数估值相比，来自选择后模型的参数估值置信区间仍要比未经选择模型的更可靠。这一逻辑推理恰恰相悖于所谓的夸大效应之论点。

无需讳言，PAC 可学性理论明确包含了学习之不确定性因素。因此，是否应该将该因素也明确引入选择后模型的参数推断阶段，便成为一个统计学和机器学习学界仍商榷不休的议题。对于选择后模型直接使用传统的置信区间，相当于忽略了可学性理论中的不确定性因素，假定选择后的模型是具有确定性等价的模型。如要避免这一假定，我们似乎就应在被选模型后的推断时把模型选择的不确定性也考虑进来，即扩展测度参数估值之不确定程度的概率空间。所谓“选择推断技术”就是沿此思路发展起来

[39] 详述可见 Shalev-Shwartz and Ben-David (2014, Ch24)。

[40] 统计学界通常把 Cox (1975)作为数据分割原则的初源。至于数据发掘研究，主流经济计量学界是长期持怀疑和抵制态度的，可参见 Qin (2013a, Ch9.4)。

[41] 这一论点也是 Zhao *et al* (2021)通过数学公理形式而表达的。

的[42]。选择推断技术的核心是增加传统的置信区间算法中的成分。具体地，令β_j表示某先验已知模型M^h的关注参数，其预期的显著性水平为α。传统的置信区间则为：

(5.3.1) $$\mathbb{P}\left(\beta_j \in \mathrm{CI}_j(\alpha)\right) = \mathbb{P}\left(\beta_j \in \mathrm{CI}_j(\alpha) \middle| j \in M^h\right) \geq 1 - \alpha$$

现记选择后的模型为$\widehat{M}$。由于选择$\widehat{M}$的不确定性与模型规模—亦即式(5.2.2)中的参数*d*—密切相关，因此在计算选择推断的置信区间时，可用这个因素来代表上述不确定性：

(5.3.2) $$\mathbb{P}\left(\beta_j \in \mathrm{CI}_{j\cdot\widehat{M}_d}(\alpha) \middle| j \in \widehat{M}\right) \geq 1 - \alpha$$

上式的基本思路来自多变量模型中，以联立推断为目标的单个参数估值置信区间[43]。联立推断法的初源是 Scheffé 置信区间，该区间是为了将模型中其他相关变量的不确定性也考虑进来而设计的，其显著性水平来自F分布，而不是t分布。依联立推断法计算的置信区间一般宽于传统的置信区间。这是因为前者的概率空间设定不同于后者的，亦即两者所测度的不确定性之范围不同。不过需要明确，Scheffé 置信区间并不是针对模型选择而设计的。

那么，我们是否该用 (5.3.2) 而不是 (5.3.1)来计算选模后参数估计的置信区间呢？为了回答这一问题，我们必须首先澄清 (5.3.2) 与 (5.3.1) 是否具有可比性的问题。比较两式便可看到，导致$\left|\mathrm{CI}_{j\cdot\widehat{M}_d}\right| > \left|\mathrm{CI}_j\right|$的原因是我们对 M^h和$\widehat{M}$两个模型之泛化确定性的不同划界。式(5.3.1)是以 M^h之总体泛化性确定无疑作为假定前提的，因此在CI_j中没有考虑M^h的不确定性。而正是由于式(5.3.2)的$\mathrm{CI}_{j\cdot\widehat{M}_d}$中考虑了$\widehat{M}$的不确定性，才导致了计算的置信区间相对更宽的结果。显然，假定M^h之总体泛化性确定无疑的前提条件在现实中是不成立的。因此，以模型相对于数据样本的独立性 (即数据是否被“浸渍”) 为理由，对于未经后验选择的模型采用式(5.3.1)计算参数估值的置信区间，而用式(5.3.2)计算经后验选择的模型之参数估值的置信区间，通过比较这两种置信区间来评价选择后模型之可靠性程度，其实是不合理的。

现实中的大量不容置疑的实证表明，据 PAC 可学性原理机器学习来的模型要明显优于单靠先验数学演绎推导来的模型。这一事实与$\left|\mathrm{CI}_{j\cdot\widehat{M}_d}\right| > \left|\mathrm{CI}_j\right|$的明显相悖再次警示我们，将在理论贫乏场景下假说验证的任务简单置于经典统计估计检验框架之内是个多么严重的认知失误。前面的分析还表明，从常识性先验假说出发到统计学习得出的$\widehat{M}$，需要使用多重最优控制标准和一系列的计算数学和统计学工具。在模型学习过程中得到的间接验证数据，都是后续得出的参数估值置信区间的必要铺垫。因此我们在讨论验证理论实据时，不能仅限于或者集中在关注参数估值的置信区间上，而需要明确和如实地将得到该区间的必要前提数据实证也概括进来。换句话说，从应用经济学家的角度考虑，用多种相关数据实证来完善经济假说的验证分析，应当要比用一个类

[42] 详述可见 Taylor and Tibshirani (2015) 及 Efron and Hastie (2016, Ch20)。

[43] 有关联立推断的具体讨论，可参见 Berk *et al* (2013)以及 Efron and Hastie (2016, Ch20)。

似选择推断的置信区间这样相对复杂的统计量，试图集中概括建模过程中所有的不确定性的思路要明确易懂得多[44]。

从认识论的角度看，PAC 可学性理论中代表模型泛化功能不确定性的概率参数，只是为表示数据学习空间是不可完全封闭的一个抽象概念，是对我们认知局限的一个公认形式。因此，这一概念的应用场景属于解析类而不属于枚举类范畴。任何试图通过某种与模型规模有关的参数来具体测度这种不确定性的做法，都会陷入不容忽略的测度误差的困扰。为了获得统计上可计算的描述某研究对象不确定性的概率测度，我们必须首先具有一个可以明确定义的包含该研究对象的闭空间。这一前提要求显然与探索性统计学习的主旨相悖，在模型学习情形下难以实现。众所周知，定义在不同闭空间上的概率测度是不具可比性的，但是现实中不乏犯这类比较错误的例子。还需看到，定义的概率闭空间的维度越大，不但数学推导基于该空间的概率分布形式、以及相继可用的最优统计量之复杂程度会越高，而且对这些统计量的应用结果做出明确无误的解释也就越为困难，造成认知误解的风险就越高。统计学界有关选择性推断技术的研究与争论，让我们再次领悟了概率测度的应用局限性[45]。

[44] 详述可见 Holmes (2018)。

[45] 为了阐明概率这一测度概念之相对性， Holmes (2018) 讲了一个爱因斯坦的故事。一次，爱因斯坦问他的学生："生命有限而时间无限。从无限的时间看，我今天活着的概率为零。可是我今天还活着。这该如何解释呢？" 学生们都无语。停顿片刻，爱因斯坦说："一旦有了事实，我们就不该再提概率了。"

References

Abu-Mostafa, Y. S., M. Magdon-Ismail, and H.-T. Lin (2012) *Learning From Data*, AMLBook.

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996) Identification of causal effects using instrumental variables, *Journal of the American Statistical Association*, 91(434): 444-55.

Arlot, S. and A. Celisse (2010) A survey of cross-validation procedures for model selection, *Statistics Survey*, 4: 40-79.

Athey, S. and G.W. Imbens (2019) Machine learning methods that economists should know about, *Annual Review of Economics*, 11(1): 685-725.

Berk, R., L. Brown, A. Buja, K. Zhang and L. Zhao (2013) Valid post-selection inference, *Annals of Statistics*, 41(2): 802-37.

Berndt, E.R. (1991) *The Practice of Econometrics: Classic and Contemporary*, Addison Wesley.

Brodeur, A., M. Lé, M. Sangnier and Y. Zylberberg (2016) Star wars: The empirics strike back, *American Economic Journal: Applied Economics*, 8(1): 1-32.

Brodeur, A., N. Cook and A. Heyes (2020) Methods matter: p-hacking and publication bias in causal analysis in economics, *American Economic Review*, 110(11): 3634-60.

Cameron, A.C. and P.K. Trivedi (2005) *Microeconometrics: Methods and Applications*, Cambridge: Cambridge University Press.

Carpenter, J. and M. Kenward (2013) *Multiple Imputation and Its Application*, Chichester: John Wiley & Sons.

Carpenter, J. and M. Smug (2021) Missing data: A statistical framework for practice, *Biometrical Journal*, 63: 915-47.

Cochrane, D. and G. Orcutt (1949) Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms, *Journal of American Statistical Association*, 44: 32-61.

Cox, D. (1961) Tests of separate families of hypothesis, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1: 105-23.

Cox, D. (1962) Further results on tests of separate families of hypotheses, *Journal of Royal Statistical Society*, B24: 406-24.

Cox, D. (1975) A note on data-splitting for the evaluation of significance levels, *Biometrika*, *62*(2): 441-44.

Deming, W.E. (1975) On probability as a basis for action, *The American Statistician*, 29(4): 146-52.

Ding, J., V. Tarokh and Y.-H. Yang (2018) Model selection techniques: An overview, *IEEE Signal Processing Magazine*, 35(6): 16-34.

Dziak, J.J., D.L. Coffman, S.T. Lanza, R.-Z. Li and L.S. Jermiin (2020) Sensitivity and specificity of information criteria, *Briefings in bioinformatics*, 21(2): 553-65.

Efron, B. and T. Hastie (2016) *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*, Cambridge University Press.

Haavelmo, T. (1943) The statistical implications of a system of simultaneous equations, *Econometrica*, 11: 1-12.

Haavelmo, T. (1944) The probability approach in econometrics, *Econometrica*, 12, supplement.

Hahn, G. and W. Meeker (1993) Assumptions for statistical inference, *The American Statistician*, 47(1):1-11.

Hastie, T., R. Tibshirani and J. Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition, Springer.

Hendry, D.F. (1995) *Dynamic econometrics*, Oxford: Oxford University Press.

Hendry, D.F. and H-M. Krolzig (2003) New developments in automatic general-to-specific modelling, in B.P. Stigum ed., *Econometrics and the Philosophy of Economics*, Princeton: Princeton University Press, pp. 379-419.

Hendry, D.F. and J.A. Doornik (2014) *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*, MIT Press.

Hirschauer, N., O. Mußhoff, S. Grüner, U., Frey, I., Theesfeld and P. Wagner (2016) Inferential misconceptions and replication crisis. *Journal of Epidemiology, Biostatistics, and Public Health*, 13(4): e12066-1-16.

Hitchcock, C. and Sober, E. (2004) Prediction versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science, 55*(1), 1–34.

Holmes, S. (2018) Statistical proof? The problem of irreproducibility, *Bulletin of the American Mathematics Society*, 55(1): 31-55.

James, G., D. Witten, T. Hastie and R. Tibshirani (2013) *An Introduction to Statistical Learning*, Springer.

Kadane, J.B. and N.A. Lazar (2004) Methods and criteria for model selection. *Journal of the American Statistical Association*, 99: 279-90.

Kennedy, P. (2008) *A Guide to Econometrics*, Wiley-Blackwell.

Kerr N.L. (1998) HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2(3): 196–217.

Koopmans, T.C. (1947) Measurement without theory, *The Review of Economics and Statistics*, 29(3): 161-72.

Maddala, G.S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.

Malinvaud, E. (1966) *Statistical Methods in Econometrics*. Amsterdam: North-Holland.

Markus, K. and D. Borsboom (2013) *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*, Routledge.

Mayo, D. (2018) *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Mill, T.C. and K. Patterson eds. (2006) *Palgrave Handbook of Econometrics*, vol. I Econometric Theory, Palgrave MacMillan.

Mizon, G.E. and J.-F. Richard (1986) The encompassing principle and its application to testing non-nested hypotheses, *Econometrica*, 54(3): 657-78.

Orcutt, G. (1948) A study of the autoregressive nature of the time series used for Tinbergen's model of the economic system of the United States 1919-1932, *Journal of the Royal Statistical Society Series B*, 10: 1-45.

Pesaran, M.H. (1987) Global and partial non-nested hypotheses and asymptotic local power, *Econometric Theory*, 3(1): 69-97.

Prosperi, M., Bian, J., Buchan, I.E., J.S. Koopman, M. Sperrin and M. Wang (2019) Raiders of the lost HARK: A reproducible inference framework for big data science. *Palgrave Communications*, 5: 125.

Qin, D. (1993) *The Formation of Econometrics: A Historical Perspective*, Oxford: Clarendon Press.

Qin, D. (2013a) *A History of Econometrics: The Reformation from the 1970s*, Oxford University Press.

Qin, D. (2015) Resurgence of the endogeneity-backed instrumental variable methods, *Economics: The Open-Access, Open-Assessment E-Journal*, 9(7): 1-35.

Qin, D. (2019) Let's take the bias out of econometrics, *Journal of Economic Methodology*, 26(2): 81-98.

Qin, D., S. van Huellen, R. Elshafie, Y.-M. Liu and T. Moraitis (2019) A principled approach to assessing missing-wage induced selection bias, *SOAS Economics Working Paper Series*, No. 216.

Qin D, S. van Huellen, Q.C. Wang, T. Moraitis (2022) Algorithmic modelling of financial conditions for macro predictive purposes: Pilot application to USA data. *Econometrics*, 10(2): 22. https://doi.org/10.3390/econometrics10020022

Sargent, T. and C.A. Sims (1977) Business cycle modeling without pretending to have too much a priori economic theory, in *New Methods in Business Cycle Research: Proceedings from a Conference*, Federal Reserve Bank of Minneapolis, pp 45-109.

Shalev-Shwartz, S. and S. Ben-David (2014) *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press.

Spanos, A. (2010) Theory testing in economics and the error-statistical perspective, in D. G. Mayo and A. Spanos eds., *Error and Inference: Recent exchanges on experimental reasoning, reliability and the objectivity and rationality of science*, Cambridge: Cambridge University Press.

Spanos, A. (2018) Mis-specification testing in retrospect, *Journal of Economic Survey*, 32(2): 541-77.

Stock, J.H. and M.W. Watson (1989) New indexes of coincident and leading economic indicators, *NBER Macroeconomics Annual*, 4: 351-94.

Strotz, R.H. and H.O.A. Wold (1960) Recursive vs. nonrecursive systems: An attempt at synthesis, *Econometrica*, 28: 417-27.

Swann, G.M.P. (2006) *Putting Econometrics in its Place: A New Direction in Applied Economics*, Edward Elgar.

Taylor, J. and R.J. Tibshirani (2015) Statistical learning and selective inference, *Proceedings of the National Academy of Sciences*, 112(25): 7629-34.

Trivedi, P.K. (1984) Uncertain prior information and distributed lag analysis, in Hendry and Wallis eds., *Econometrics and Quantitative Economics*, Oxford: Basil Blackwell, pp. 173-210.

van Huellen, S. and D. Qin (2019) Compulsory schooling and returns to education: A re-examination, *Econometrics*, 7(3): 36; https://doi.org/10.3390/econometrics7030036

van Huellen, S., D. Qin, S. Lu, H.-W. Wang, Q.-C. Wang and T. Moraitis (2022) Modelling opportunity cost effects in money demand due to openness, *International Journal of Finance & Economics*, 27(1): 697-744.

Wold, H.O.A. (1954) Causality and econometrics, *Econometrica,* 22: 162-77.

Wold, H.O.A. (1956) Causal inference from observational data: A review of ends and means. *Journal of Royal Statistical Society* A, 119: 28-61.

Wold, H.O.A. and l. Juréen (1953) *Demand analysis*, John Wiley & Sons.

Yang, Y.-H. (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation, *Biometrika*, 92(4): 937-50.

Young, A. (2017) Consistence without inference: Instrumental variables in practical application, *Working Paper*, London School of Economics.

Zellner, A. (1979) Causality and econometrics, *Carnegie-Rochester Conference Series on Public Policy*, 10: 9-54.

Zhao, S., D. Witten and A. Shojaie (2021) In defense of the indefensible: A very naive approach to high-dimensional inference, *Statistical Science*, *36*(4): 562-77.